Using Computational Linguistics to Demystify Financial Texts

Thesis submitted by

Sohom Ghosh

Doctor of Philosophy (Engineering)

Department of Computer Science & Engineering
Faculty Council of Engineering & Technology

JADAVPUR UNIVERSITY
Kolkata, India

2025

Using Computational Linguistics to Demystify Financial Texts

Thesis submitted by

Sohom Ghosh

Doctor of Philosophy (Engineering)

Department of Computer Science & Engineering
Faculty Council of Engineering & Technology

JADAVPUR UNIVERSITY
Kolkata, India

2025

Index No: 66/22/E

1. **Title of the Thesis**:
   Using Computational Linguistics to Demystify Financial Texts

2. **Name, Designation and Institution of the Supervisor**:
   Dr. Sudip Kumar Naskar
   Associate Professor
   Department of Computer Science & Engineering
   Jadavpur University, Kolkata

3. **List of Publications**

   Journals:

   (a) **Sohom Ghosh**, Ankush Chopra, Sudip Kumar Naskar, "*Learning to Rank Hypernyms of Financial Terms Using Semantic Textual Similarity*", in SN Computer Science, Springer, 4, 610 (2023), https://doi.org/10.1007/s42979-023-02134-z

   (b) **Sohom Ghosh**, Sudip Kumar Naskar, "*Recent trends in financial natural language processing research*", in Science Talks, Elsevier, Volume 8, 100270, (2023) https://doi.org/10.1016/j.sctalk.2023.100270

   (c) **Sohom Ghosh**, Sudip Kumar Naskar, "*Fincat-2: An enhanced Financial Numeral Claim Analysis Tool*", in Software Impacts, VOLUME 12, 100288, Elsevier, May 2022. https://doi.org/10.1016/j.simpa.2022.100288

   (d) **Sohom Ghosh**, Sudip Kumar Naskar, "*Detecting context-based in-claim numerals in Financial Earnings Conference Calls*", in International Journal of Information Technology, Springer, May 2022 https://doi.org/10.1007/s41870-022-00952-7

   International Conferences & Workshops:

   (a) **Sohom Ghosh**, "*Demystifying Financial Texts Using Natural Language Processing*", in Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM 2024), Boise, USA (Online) https://doi.org/10.1145/3627673.3680258

   (b) **Sohom Ghosh**, Chung-Chi Chen, and Sudip Kumar Naskar, "*Generator-Guided Crowd Reaction Assessment*" in companion proceedings of The Web Conference (WWW-2024), Singapore. https://doi.org/10.1145/3589335.3651512

(c) **Sohom Ghosh**, Arnab Majhi, Aswartha Narayana, and Sudip Kumar Naskar, "*IndicFinNLP: Financial Natural Language Processing for indian languages*", in proceedings of 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy. https://aclanthology.org/2024.lrec-main.789

(d) Neelabha Banerjee, Anubhav Sarkar, Swagata Chakraborty, **Sohom Ghosh**, Sudip Kumar Naskar, "*Fine-tuning Language Models for predicting the impact of events associated to financial news articles*", In FinNLPKDF-EcoNLP workshop of LREC-COLING 2024, Torino, Italy. https://aclanthology.org/2024.finnlp-1.25

(e) Rima Roy, **Sohom Ghosh**, and Sudip Kumar Naskar, "*Financial Argument Analysis in Bengali*", in Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE-2023) page 88–92, Panjim, India. https://doi.org/10.1145/3632754.3632763

(f) **Sohom Ghosh**, Sachin Umrao, Chung-Chi Chen, and Sudip Kumar Naskar, "*The Mask One At a Time framework for detecting the relationship between financial entities*", in Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE-2023) page 40-43, Panjim, India. https://doi.org/10.1145/3632754.3632756

(g) Harsha Vardhan, **Sohom Ghosh**, Ponnurangam Kumaraguru, and Sudip Kumar Naskar, "*A low resource framework for multi-lingual esg impact type identification*", in Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing (FinNLP), page 57-61, Bali, Indonesia, November. (2023) https://aclanthology.org/2023.finnlp-2.8/

(h) Swagata Chakraborty, Anubhav Sarkar, Dhairya Suman, **Sohom Ghosh** , Sudip Kumar Naskar, "*LIPI at the NTCIR-17 FinArg-1 Task: Using Pre-trained Language Models for Comprehending Financial Arguments*", in Proceedings of the 17$^{\text{th}}$ NTCIR Conference on Evaluation of Information Access Technologies, pp 29-36, Tokyo, Japan. (2022) https://doi.org/10.20736/0002001281

(i) Priyank Soni, **Sohom Ghosh**, Sudip Kumar Naskar, "*Detecting Issues Related to Environmental, Social, and Corporate Governance using SEC-BERT*", in proceedings of 4$^{\text{th}}$ International Conference on Data Science and Applications (ICDSA 2023), Jaipur, India https://doi.org/10.1007/978-981-99-7820-5_27

(j) **Sohom Ghosh**, Sudip Kumar Naskar, "*Using Natural Language Processing to Enhance Understandability of Financial Texts*", in proceedings of 6$^{\text{th}}$ Joint International Conference on Data Science & Management of Data (10$^{\text{th}}$ ACM

IKDD CODS and 28th COMAD), Mumbai, India https://doi.org/10.1145/3570991.3571051 [**Honourable Mention (YRS Track)**]

(k) **Sohom Ghosh**, Sudip Kumar Naskar, "*FLUEnT: Financial Language Understandability Enhancement Toolkit*", in proceedings of 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD), Mumbai, India https://doi.org/10.1145/3570991.3571067

(l) Anubhav Sarkar, Swagata Chakraborty, **Sohom Ghosh** and Sudip Kumar Naskar, "*Evaluating Impact of Social Media Posts by Executives on Stock Prices*", in proceedings of the 14th meeting of Forum for Information Retrieval Evaluation (FIRE-2022), Kolkata, India. https://doi.org/10.1145/3574318.3574339

(m) **Sohom Ghosh**, Sudip Kumar Naskar, "*LIPI at the FinNLP-2022 ERAI Task: Ensembling Transformers for assessing Maximum Possible Profit and Loss from online financial posts*", in Proceedings of the fourth workshop on Financial Technology and Natural Language Processing (collocated with EMNLP 2022), pp 111-115, Abu Dhabi, UAE (2022) https://aclanthology.org/2022.finnlp-1.13.pdf

(n) **Sohom Ghosh**, Sudip Kumar Naskar, "*Ranking Environment, Social And Governance Related Concepts And Assessing Sustainability Aspect Of Financial Texts*", in Proceedings of the fourth workshop on Financial Technology and Natural Language Processing (collocated with IJCAI-ECAI 2022), pp 243-249, Vienna, Austria (2022) https://aclanthology.org/2022.finnlp-1.33.pdf

(o) **Sohom Ghosh**, Shovon Sengupta, Sudip Kumar Naskar, Sunny Kumar Singh, "*FinRAD: Financial Readability Assessment Dataset - 13,000+ Definitions of Financial Terms for Measuring Readability*", in Proceedings of the FNP workshop of the 13th Language Resources and Evaluation Conference (LREC-2022), pp 1-9 , Marseille, France http://www.lrec-conf.org/proceedings/lrec2022/workshops/FNP/pdf/2022.fnp-1.1.pdf

(p) **Sohom Ghosh**, Sudip Kumar Naskar, "*LIPI at FinCausal 2022: Mining Causes and Effects from Financial Texts*", in Proceedings of the FNP workshop of the 13th Language Resources and Evaluation Conference (LREC-2022), pp 130–132, Marseille, France http://www.lrec-conf.org/proceedings/lrec2022/workshops/FNP/pdf/2022.fnp-1.20.pdf

(q) **Sohom Ghosh**, Sudip Kumar Naskar, "*LIPI at the NTCIR-16 FinNum-3 task: Ensembling transformer based models to detect in-claim numerals in financial*

conversations", in Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies, pp 92-94, Tokyo, Japan. (2022) http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings16/pdf/ntcir/02-NTCIR16-FINNUM-GhoshS.pdf

(r) **Sohom Ghosh**, Sudip Kumar Naskar,'*'FiNCAT: Financial Numeral Claim Analysis Tool*", in companion proceedings of TheWebConf (formerly ACM-WWW) (2022), pp 583-585, Lyon, France https://doi.org/10.1145/3487553.3524635

(s) **Sohom Ghosh**, Shovon Sengupta, Sudip Kumar Naskar, Sunny Kumar Singh, "*FinRead: A Transfer Learning Based Tool to Assess Readability of Definitions of Financial Terms*", in proceedings of 18th International Conference on Natural Language Processing (ICON-2021), pp 658-659 https://aclanthology.org/2021.icon-main.81.pdf

(t) Ankush Chopra, **Sohom Ghosh**, Term Expansion and FinBERT fine-tuning for Hypernym and Synonym Ranking of Financial Terms, In proceedings of FinNLP'21 (FinSim-3) (collocated with IJCAI-2021), pp 46-51, Montreal, Canada https://aclanthology.org/2021.finnlp-1.8.pdf

4. **List of Patents**: NA

5. **List of Presentations in National / International Conferences & Workshops**

   National / International Conferences & Workshops:

   (a) **Sohom Ghosh**, Chung-Chi Chen, Sudip Kumar Naskar, "*Generator-Guided Crowd Reaction Assessment*", in ACM Academic Research and Careers for Students (ARCS 2025), Coimbatore, India.

   (b) **Sohom Ghosh**, "*Applying Natural Language Processing on Financial Texts*", in Pingala Interactions in Computing (PIC 2025), Mysuru, India.

   (c) **Sohom Ghosh**, "*Demystifying Financial Texts Using Natural Language Processing*", in Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM 2024), Boise, USA (Online). https://doi.org/10.1145/3627673.3680258

   (d) **Sohom Ghosh**, Arnab Majhi, Aswartha Narayana, and Sudip Kumar Naskar, "*IndicFinNLP: Financial Natural Language Processing for indian languages*", in proceedings of 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy. https://aclanthology.org/2024.lrec-main.789

(e) Neelabha Banerjee, Anubhav Sarkar, Swagata Chakraborty, **Sohom Ghosh**, Sudip Kumar Naskar,"*Fine-tuning Language Models for predicting the impact of events associated to financial news articles*", In FinNLPKDF-EcoNLP workshop of LREC-COLING 2024, Torino, Italy. https://aclanthology.org/2024.finnlp-1.25

(f) Rima Roy, **Sohom Ghosh**, and Sudip Kumar Naskar, "*Financial Argument Analysis in Bengali*", in Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE-2023) page 88–92, Panjim, India. https://doi.org/10.1145/3632754.3632763

(g) **Sohom Ghosh**, Sachin Umrao, Chung-Chi Chen, and Sudip Kumar Naskar, "*The Mask One At a Time framework for detecting the relationship between financial entities*", in Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE-2023) page 40-43, Panjim, India. https://doi.org/10.1145/3632754.3632756

(h) Priyank Soni, **Sohom Ghosh**, Sudip Kumar Naskar, "*Detecting Issues Related to Environmental, Social, and Corporate Governance using SEC-BERT*", in proceedings of 4th International Conference on Data Science and Applications (ICDSA 2023), Jaipur, India https://doi.org/10.1007/978-981-99-7820-5_27

(i) **Sohom Ghosh**, Sudip Kumar Naskar, "*Using Natural Language Processing to Enhance Understandability of Financial Texts*", in proceedings of 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD), Mumbai, India https://doi.org/10.1145/3570991.3571051 [**Honourable Mention (YRS Track)**]

(j) **Sohom Ghosh**, Sudip Kumar Naskar, "*FLUEnT: Financial Language Understandability Enhancement Toolkit*", in proceedings of 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD), Mumbai, India https://doi.org/10.1145/3570991.3571067

(k) **Sohom Ghosh**, Sudip Kumar Naskar, "*LIPI at the FinNLP-2022 ERAI Task: Ensembling Transformers for assessing Maximum Possible Profit and Loss from online financial posts*", in Proceedings of the fourth workshop on Financial Technology and Natural Language Processing (collocated with EMNLP 2022), pp 111-115, Abu Dhabi, UAE (2022) https://aclanthology.org/2022.finnlp-1.13.pdf

(l) **Sohom Ghosh**, Sudip Kumar Naskar, "*Ranking Environment, Social And Governance Related Concepts And Assessing Sustainability Aspect Of Financial Texts*", in Proceedings of the fourth workshop on Financial Technology and

Natural Language Processing (collocated with IJCAI-ECAI 2022), pp 243-249, Vienna, Austria (2022) https://aclanthology.org/2022.finnlp-1.33.pdf

(m) **Sohom Ghosh**, Shovon Sengupta, Sudip Kumar Naskar, Sunny Kumar Singh, "*FinRAD: Financial Readability Assessment Dataset - 13,000+ Definitions of Financial Terms for Measuring Readability*", in Proceedings of the FNP workshop of the 13th Language Resources and Evaluation Conference (LREC-2022), pp 1-9 , Marseille, France http://www.lrec-conf.org/proceedings/lrec2022/workshops/FNP/pdf/2022.fnp-1.1.pdf

(n) **Sohom Ghosh**, Sudip Kumar Naskar, "*LIPI at FinCausal 2022: Mining Causes and Effects from Financial Texts*", in Proceedings of the FNP workshop of the 13th Language Resources and Evaluation Conference (LREC-2022), pp 130–132, Marseille, France http://www.lrec-conf.org/proceedings/lrec2022/workshops/FNP/pdf/2022.fnp-1.20.pdf

(o) **Sohom Ghosh**, Sudip Kumar Naskar, "*LIPI at the NTCIR-16 FinNum-3 task: Ensembling transformer based models to detect in-claim numerals in financial conversations*", in Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies, pp 92-94, Tokyo, Japan. (2022) http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings16/pdf/ntcir/02-NTCIR16-FINNUM-GhoshS.pdf

(p) **Sohom Ghosh**, Sudip Kumar Naskar,'*'FiNCAT: Financial Numeral Claim Analysis Tool*", in proceedings of FinWeb@TheWebConf (formerly ACM-WWW) (2022), pp 583-585, Lyon, France https://doi.org/10.1145/3487553.3524635

(q) **Sohom Ghosh**, Shovon Sengupta, Sudip Kumar Naskar, Sunny Kumar Singh, "*FinRead: A Transfer Learning Based Tool to Assess Readability of Definitions of Financial Terms*", in proceedings of 18th International Conference on Natural Language Processing (ICON-2021), pp 658-659 https://aclanthology.org/2021.icon-main.81.pdf

(r) Ankush Chopra, **Sohom Ghosh**, Term Expansion and FinBERT fine-tuning for Hypernym and Synonym Ranking of Financial Terms, In proceedings of FinNLP'21 (FinSim-3) (collocated with IJCAI-2021), pp 46-51, Montreal, Canada https://aclanthology.org/2021.finnlp-1.8.pdf

Statement of Originality

I, Sohom Ghosh, registered on 27th April 2022 (Ref. No. D-7/E/326/22, Index No. 66/22/E, Registration No. 1022204003 of 2022-2023) do hereby declare that this thesis entitled "Using Computational Linguistics to Demystify Financial Texts" contains literature survey and original research work done by the undersigned candidate as part of Doctoral studies. All information in this thesis have been obtained and presented in accordance with existing academic rules and ethical conduct. I declare that, as required by these rules and conduct, I have fully cited and referred all materials and results that are not original to this work. I also declare that I have checked this thesis as per the "Policy on Anti Plagiarism, Jadavpur University, 2019", and the level of similarity as checked by iThenticate software is 1%.

Sohom Ghosh 1st Dec 2025

Signature by Candidate with Date
Sohom Ghosh

Sudip Kumar Naskar 01/12/2025

Signature by Supervisor with Date
Dr. Sudip Kumar Naskar

ASSOCIATE PROFESSOR
Dept. of Computer Sc. & Engg.
JADAVPUR UNIVERSITY
Kolkata - 700 032

Certificate from the Supervisor

This is to certify that the thesis entitled "Using Computational Linguistics to Demystify Financial Texts"', submitted by Sohom Ghosh (Ref. No. D-7/E/326/22, Index No. 66/22/E, Registration No. 1022204003 of 2022-2023, Date of Registration: 27th April 2022), for the award of Ph.D. (Engineering) degree of Jadavpur University, is absolutely based upon his own work under the supervision of Dr. Sudip Kumar Naskar and that neither his thesis nor any part of his thesis has been submitted for any degree/diploma or any other academic award anywhere before.

SIGNATURE OF SUPERVISOR WITH DATE

Sudip Kumar Naskar

01/12/2025

Dr. Sudip Kumar Naskar
Associate Professor
Department of Computer Science & Engineering
Jadavpur University, Kolkata

ASSOCIATE PROFESSOR
Dept. of Computer Sc. & Engg.
JADAVPUR UNIVERSITY
Kolkata - 700 032

Dedicated to Maa Saraswati, the Goddess of Knowledge, Music, Art, Speech, Wisdom, and Learning.

Om Shreem Hreem Saraswatyai Namah

Acknowledgements

I would like to express my sincere gratitude for my supervisor Dr. Sudip Kumar Naskar for giving me the opportunity to work with him and helping me out in every possible way. I also want to thank my mentor Dr. Chung-Chi Chen for providing me with insightful inputs which had helped me immensely. In addition to this, special thanks to all my collaborators (Anubhav Sarkar, Swagata Chakraborty, Neelabha Banerjee, Dhairya Suman, Rima Roy, Arnab Maji, Aswartha Narayana, Harsha Vardhan, etc.), lab-mates (Atanu Mondal, Madhusudan Ghosh, etc.), friends (Pramit Bhattacharya, Nildari Prosad Lahiri, Angan Mitra, etc.) and (ex-) colleagues (Sachin Umrao, Shovon Sengupta, Ankush Chopra, DK Aswartha, etc.). Finally, thanks to my parents for always bestowing their blessings on me.

Abstract

Over the years, with the rise in digitization, people have embraced online means for investing money. This led to creation of a lot of text data relating to finance. It is also note-worthy that the global financial literacy rate is just 33%.[1]. Improving financial literacy leads to financial well-being of the people and in-turn results in economic prosperity of the world. Thus, we leverage Computational Linguistics, Natural Language Processing, and Artificial Intelligence to process financial text data for social good. Our objective is to make the process of investing: Inclusive, Improved, Impactful, and Informed. In this thesis, we cover the following themes.

- **Inclusive Investing**: Making financial texts more readable and reachable for the common people
- **Improved Investing**: Simplifying the investment process by extracting hypernyms and relation between entities
- **Impactful Investing**: Trying to balance between risk, return, and impact on environment by understanding sustainability, and ESG aspects of investment instruments
- **Informed Investing**: Filtering authentic information from false information spread across various media platforms

In addition to this, we focused specifically on helping Indian investors (**Indic Investing**) and developed several **FinNLP related tools** to ease the life of investors.

[1] https://gflec.org/wp-content/uploads/2015/11/Finlit_paper_16_F2_singles.pdf (accessed on 17th September, 2023)

Contents

List of Figures

List of Tables

Abbreviations

AE	**A**uto **E**ncoders
API	**A**pplication **P**rogramming **I**nterfaces
AE	**A**uto **E**ncoders
API	**A**pplication **P**rogramming **I**nterfaces
AR	**A**signed **R**eadability
ARI	**A**utomated **R**eadability **I**ndex
ARI	**A**utomated **R**eadability **I**ndex
ARI	**A**utomated **R**eadability **I**ndex
AUROC	**A**rea **U**nder **R**eceiver **O**perating **C**haracteristic curve
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
Bi-LSTM	**Bi-directional LSTM**
Bi-RNN	**Bi-directional RNN**
BIO	**B**egin, **I**nside, **O**utside
BSE	**B**ombay **S**tock **E**xchange
CD	**C**laim **D**etection
CLI	**C**oleman **L**iau **I**ndex
CNN	**C**onvoluted **N**eural **N**etwork
CPU	**C**entral **P**rocessing **U**nit
DCF	**D**ale-**C**hall **F**ormula
DCR	**D**ale – **C**hall **R**eadability
DRHP	**D**raft **R**ed **H**erring **P**rospectus
DSM	**D**igital **S**ingle **M**arket
EID	**E**SG **I**ssue **D**etector
ESG	**E**nvironmental, **S**ocial and **G**overnance
EPS	**E**arnings **P**er **S**hare

FD	**F**ixed **D**eposits
FIBO	**F**inancial **I**ndustry **B**usiness **O**ntology
FinNLP	**Fin**ancial **N**atural **L**anguage **P**rocessing
FinTech	**Fin**ancial **Tech**nology
FKGL	**F**lesch **K**incaid **G**rade **L**evel
FLS	**F**orward-**L**ooking **S**tatements
FLUEnT	**F**inancial **L**anguage **U**nderstandability **E**nhancement **T**oolkit
FOG	**L**insear write **Fo**rmula and **G**unning's Index
FNS	**F**inancial **N**arrative **S**ummarization
FinTOC	**F**innacial **T**able Of **C**ontent
FPO	**F**ollow-on **P**ublic **O**ffer
FRE	**F**lesch **R**eading **E**ase
FRI	**F**lesch **R**eading **I**ndex
FRI	**F**lesch **R**eading **I**ndex
FT	**F**inancial **T**exts
FW	**F**inancial **W**ords
GDP	**G**ross **D**omestic **P**roduct
GMP	**G**rey **M**arket **P**remium
GNI	**G**ross **N**ational **I**ncome
GPU	**G**raphical **P**rocessing **U**nit
GRU	**G**ated **R**ecurrent **U**nit
GUI	**G**raphical **U**ser **I**nterface
HD	**H**ypernym **D**etection
HE	**H**ypernym **E**xtraction
HNI	**H**igh **N**et worth **I**ndividual
HTML	**H**yper**t**ext **M**arkup **L**anguage
IPO	**I**nitial **P**ublic **O**ffering
ISS	**I**nstitutional **S**hareholder **S**ervices
LLM	**L**arge **L**anguage **M**odels
LSTM	**L**ong **S**hort **T**erm **M**emory
MAE	**M**ean **A**bsolute **E**rror
MAPE	**M**ean **A**bsolute **P**ercentage **E**rror
MB	**M**ain **B**oard

MD & A	**M**anagement **D**iscussion and **A**nalysis
M-BET	**M**ultilingual-**BERT**
ML	**M**aximum **L**oss
MLM	**M**asked **L**anguage **M**odelling
MOAT	**M**ask **O**ne **A**t a **T**ime
MPP	**M**aximum **P**ossible **P**rofit
MSCI	**M**organ **S**tanley **C**apital **I**nternational
MSE	**M**ean **S**quared **E**rror
NCERT	**N**ational **C**ouncil of **E**ducational **R**esearch and **T**raining
NGOs	**N**on **G**overnmental **O**rganizations
NII	**N**on-**I**nstitutional **I**nvestors
NLI	**N**atural **L**anguage **I**nference
NLP	**N**atural **L**anguage **P**orcessing
NN	**N**eural **N**etwork
NSE	**N**ational **S**tock **E**xchange
NSP	**N**ext **S**entence **P**rediction
OCR	**O**ptical **C**haracter **R**ecognition
PCA	**P**rincipal **C**omponent **A**nalysis
PDF	**P**ortable **D**ocument **F**ormat
PDF	**P**ortable **D**ocument **F**ormat
QIB	**Q**ualified **I**nstitutional **B**uyer
RA	**R**eadability **A**ssessment
RAG	**R**etrieval **A**ugmented **G**eneration
REFinD	**R**elation **E**xtraction **Fin**ancial **D**ataset
RegTech	**Reg**ulatory **Tech**nology
RHP	**R**ed **H**erring **P**rospectus
RoE	Return on Equity
RoCE	Return on Capital Employed
RQ	**R**esearch **Q**uestions
RMSE	**R**oot **M**ean **S**quared **E**rror
RNN	**R**ecurrent **N**eural **N**etworks
SA	**S**entiment **A**nalysis
SDP	**S**hortest **D**ependency **P**ath

SEBI	**S**ecurities and **E**xchange **B**oard of **I**ndia
SEC	**S**ecuritiesand **E**xchange **C**ommission
SLM	**S**mall **L**anguage **M**odel
SIS	**S**MOG **I**ndex **S**core
SIS	**S**MOG **I**ndex **S**core
SLN	**S**erial **N**umbers
SM	**S**ummarization
SME	**S**mall and **M**edium **E**nterprises
SN	**S**ustainability **A**ssessment
SVM	**S**upport **V**ector **M**achine
TFIDF	**T**erm **F**requency **I**nverse **D**ocument **F**requency
UI	**U**ser **I**nterface

Chapter 1

Introduction

1.1 Overview

Recently, the financial industry has been embracing technological advancements for improving the financial well-being of people. These technological advancements include the rapid development in the field of Artificial Intelligence, Machine Learning, and Natural Language Processing (NLP). NLP deals with computing methodologies for aiding machines to understand human language better. Traditionally the finance industry mainly used numeric, tabular data for various kinds of predictive analytics tasks like stock market prediction, fraud analytics, credit risk assessment and so on. However, over the last decade, the availability of unstructured datasets (like texts, images, and videos) has intrigued researchers into investigating how unstructured datasets can be used in the financial industry. For stock price forecasting, researchers now use market sentiment data in addition to historical stock prices. Financial Natural Language Processing (FinNLP) deals with Natural Language Processing techniques applied to text data from the financial domain. However, these financial contents are complex and not suitable for the masses. Improving financial literacy through education lays the foundation for financial inclusion schemes to be successful. Interestingly, the financial literacy rate of a nation like India with an 80% literacy rate is only 27% [1]. As per sources [2], India has the potential to be one of the most financially literate nations in the upcoming years. This holds for other developing countries as well. Improving financial literacy accelerates economic development and improves the standard of living. Globally, the financial literacy is only 33%. [3] Human beings aspire for a better life. Financial well-being enables this. However, lack of financial literacy, ever-growing wealth inequality, and misleading illicit information floating in social media inhibit one's progress towards a good fortune. In this thesis, we discuss four pillars where Natural Language Processing can help to improve financial literacy, reduce wealth disparity, ensure a sustainable future, and economic prosperity. These

[1] https://www.financialexpress.com/market/only-27-indians-are-financially-literate-sebi-garg/2134842/ (accessed on 18th September, 2023)

[2] https://www.ibef.org/blogs/india-s-growing-financial-literacy (accessed on 18th September, 2023)

[3] https://gflec.org/wp-content/uploads/2015/11/Finlit_paper_16_F2_singles.pdf (accessed on 18th September, 2023)

pillars are: **Inclusive investing, Improved investing, Impactful (green) investing, and Informed investing**. Additionally, we focus on specifically catering the Indian market (**Indic investing**) and present several resources to enhance comprehensibility of financial texts. Inclusive investing deals with enhancing the readability and reachability of financial texts. Improved investing addresses the need to simplify investors' journey by providing them with hypernyms and relations between entities. Impactful investing is associated with focusing on sustainable pathways. Informed investing is about eradicating finance-related misinformation from social media, like evaluating trustworthiness of posts by executives, detecting in-claim and exaggerated numerals, etc. In most cases, we are able to demonstrate the efficacies of our approaches by benchmarking them with existing state-of-the-art methods.

- **Inclusive Investing**: Making financial texts more readable and reachable for the common people
- **Improved Investing**: Simplifying the investment process by extracting hypernyms and relation between entities
- **Impactful Investing**: Trying to balance between risk, return, and impact on environment by understanding sustainability, and ESG aspects of investment instruments
- **Informed Investing**: Filtering authentic information from false information spread across various media platforms

In addition to this, we focused specifically on helping Indian investors (**Indic Investing**) and developed several **FinNLP related tools** to ease the life of investors.

1.2 Research Landscape

Recently, the financial industry has been embracing technological advancements for improving the financial well-being of people. These technological advancements include the rapid development in the field of Artificial Intelligence, Machine Learning and Natural Language Processing (NLP). In this section, we survey how NLP is enabling recent advancements in the field of technology enabled finance or FinTech. We present a survey of the existing tasks in Figure 1.1, compare them and mention some research questions which are worth exploring.

Over the years, the interest of people in the FinTech domain has been rising rapidly. From the year 2015, the rate of increase has escalated rapidly. NLP deals with computing methodologies for aiding machines to understand human language better. Traditionally the finance industry mainly used numeric tabular data for various kinds of predictive analytics tasks like stock market prediction, fraud analytics, credit risk assessment and so on. However, over the last decade, the availability of unstructured datasets (like texts, images and videos) has intrigued researchers to investigate how unstructured datasets can be used in the financial industry. For stock price forecasting, researchers now use market sentiment data in addition to historical stock prices. Financial Natural Language Processing (FinNLP) deals with Natural Language Processing techniques applied to text data from the financial domain. The rise in polarity of financial influencers has led to

manipulation of stock markets using schemes like "Pump & Dump". This further leads to creation of meme stocks, which affects the economy badly. Thus, there is an urgent need to promote financial literacy by spreading financial knowledge. Additionally, investors are looking for sustainable funds which have lesser effects on the environment.

Financial Natural Language Processing or FinNLP (in short) is about using computational linguistics in the financial domain. Leveraging FinNLP, we can deal with these challenges to a certain extent. Financial firms like JP Morgan and Morgan Stanley are embracing ChatGPT like AI services. Other applications of NLP in FinTech include:

- **e-KYC**: As per Know Your Customer (KYC) requirements, a customer needs to submit his/her documents to a financial institution. The institution needs to verify these documents before enabling any financial transaction. The electronic KYC (e-KYC) process can be automated by leveraging NLP powered technologies.
- **Robo-Advisor**: As personal financial advisors are not always affordable, novice investors are increasingly moving towards Robo-Advisors which are Artificial Intelligence (AI) powered financial advisors. These advisors use NLP to understand the investors' needs and advise them accordingly.
- **Chatbots**: Chatbots are an easy-to-use interface for interactions. As human beings are not capable of delivering services 24/7, NLP powered chatbots can be used to entertain customer queries.
- **Portfolio Selection and Optimization**: In the era of information explosion, it is difficult for investors to keep up with the latest happenings in the financial market. Algorithms use NLP to scan through the available financial content available and help in making decisions relating to selecting and optimizing one's portfolio.
- **Understanding opinions of customers**: For financial institutions, it is necessary to understand the pulse of their customers. Reading through every comment of the customers is difficult.
- **Content Simplification**: Annual reports, earnings calls and related financial contents are complex and difficult to understand. However, Natural Language Processing can be used to simplify these contents.

Emergence of FinTech related workshops

Recently, several workshops focused on applications of Natural Language Processing in the financial domain are being organized. A list of venues that accept FinNLP related works are presented in Figure 1.2. It started with the Data Science for Macro-Modeling (DSMM)[4] workshop at Conference on Information and Knowledge Management (CIKM)-2016. Gradually other workshops like Financial Narrative Processing (FNP), Economics and Natural Language Processing (ECONLP)[5], Financial Technology and Natural Language Processing (FinNLP workshop)[6], Knowledge Discovery from Unstructured Data in Financial Services

[4] https://dl.acm.org/doi/proceedings/10.1145/2951894 (accessed on 11 November 2022)
[5] https://aclanthology.org/venues/econlp/ (accessed on 11 November 2022)
[6] https://aclanthology.org/venues/finnlp/ (accessed on 11 November 2022)

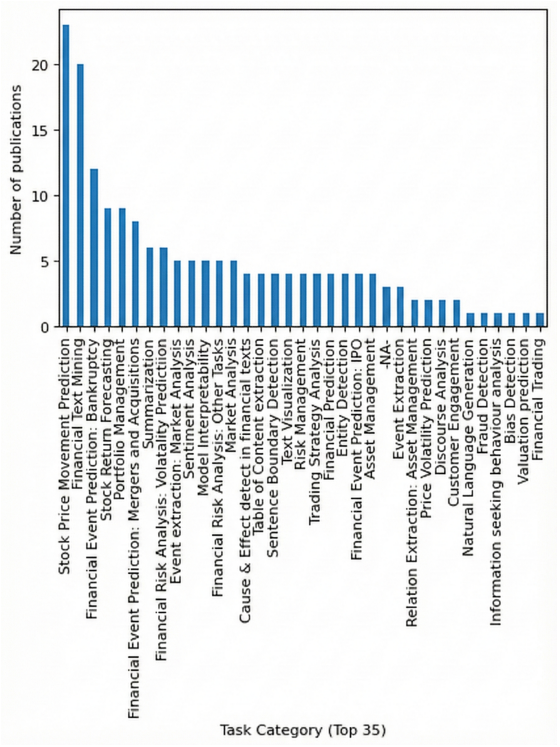

FIGURE 1.1: Task category distribution.

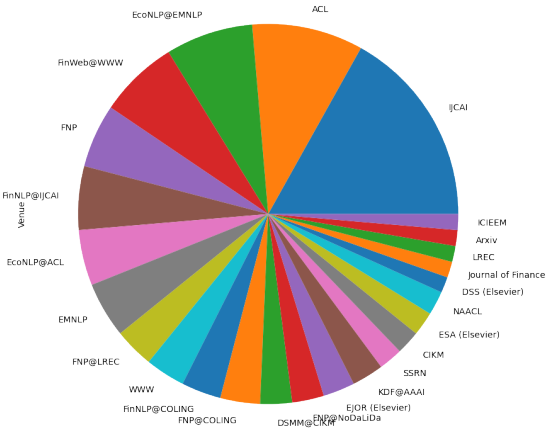

FIGURE 1.2: Distribution of venues.

(KDF)[7], FinNum[8], Multimodal AI For Financial Forecasting (MUFFIN)[9] and Modelling Uncertainty in the Financial World (MUFin)[10] came into being. Various FinNLP-themed shared tasks get hosted in these workshops. A list of these shared tasks includes Financial

[7] https://aaai-kdf.github.io/kdf2021/ (accessed on 11 November 2022)
[8] https://sites.google.com/nlg.csie.ntu.edu.tw/finnum (accessed on 11 November 2022)
[9] https://muffin-aaai23.github.io/shared_task.html (accessed on 11 November 2022)
[10] https://sites.google.com/view/w-mufin/ (accessed on 11 November 2022)

Text Summarization[11], Sentence Boundary Detection in Financial Statements [12] etc.

Moreover, ACM Economics and Computation[13] is one of the oldest and most reputed in the field of finance and economics. It is being held since 1999. International Conference on AI in Finance (ICAIF)[14] was first held in 2020. It is now being organized annually. Figure 1.3 is a timeline showing the launch of various Finance and NLP related workshops and conferences. It primarily shows events relating to the starting of major FinNLP related conferences and workshops. We see how the number of venues has grown over the last couple of years. Large Language Models specific to the financial domain are highlighted in yellow. Recently, BloombergGPT [2], a finance domain specific language model having 50 billion parameters and MultiFin [3], a financial dataset covering 15 languages were released. These recent advancements represent the ever-growing popularity of the FinNLP research domain.

Research Labs working on FinNLP

Several research groups are focusing on Financial Natural Language Processing across the globe. We mention some of the prominent ones along with some of their contributions in Table 1.1.

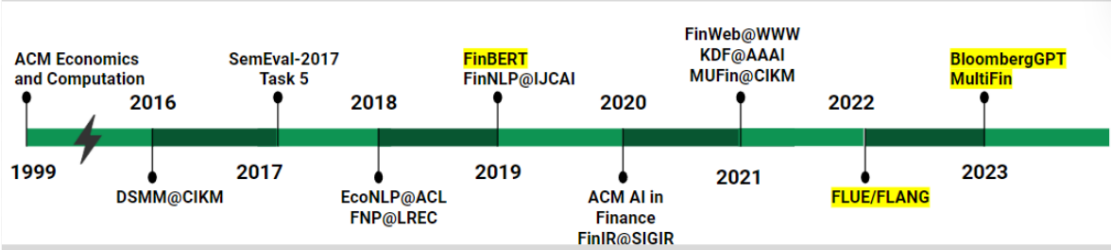

FIGURE 1.3: Timeline showing launch of various Finance and NLP related workshops and conferences.

Research Labs	Type	Location	Link
Georgia Tech. FinTech. Lab	Academic	USA	https://fintech.gatech.edu/
Corporate Financial Information Environment	Academic	UK	https://ucrel.lancs.ac.uk/cfie/
Natural Language Processing Lab.	Academic	Taiwan	http://nlg.csie.ntu.edu.tw/
London Stock Exchange Group Labs	Industrial	UK	https://www.lseg.com/about-lseg/labs
Stanford Advanced Financial Technologies Laboratory	Academic	USA	https://fintech.stanford.edu/
ISI Foundation	Academic	Italy	https://www.isi.it/en/industrial-research/fai-lab
JP Morgan AI Research	Industrial	Worldwide	https://www.jpmorgan.com/technology/artificial-intelligence
FinTech Lab, Columbia University	Academic	USA	http://fintech.datascience.columbia.edu/
BlackRock AI Labs	Industrial	Worldwide	https://www.blackrock.com/corporate/ai
FinRegLab	Industrial	USA	https://finreglab.org/ai-machine-learning/
IDRBT Artificial Intelligence & Machine Learning Lab	Academic	India	https://www.idrbt.ac.in/ai-ml-lab/

TABLE 1.1: Research labs working on FinNLP

1.3 Research Questions

The following Research Questions (RQ) are addressed in this thesis.

[11] http://wp.lancs.ac.uk/cfie/fns2022/ (accessed on 11 November 2022)
[12] https://sites.google.com/nlg.csie.ntu.edu.tw/finweb2021/shared-task-finsbd-3 (accessed on 11 November 2022)
[13] https://dl.acm.org/conference/ec (accessed on 11 November 2022)
[14] https://dl.acm.org/conference/icaif (accessed on 11 November 2022)

- **RQ-1**: How to quantify and improve readability of financial texts?
  - **Relevant Chapter**: 2
  - **Relevant Contributions**: Financial Readability Assessment Dataset (**FinRAD**) [4], A Transfer Learning Based Tool to Assess Readability of Definitions of Financial Terms (**FinRead**) [5], Financial Language Simplifier (**FinLanSer**) [Screenshot in Figure 2.8, demonstration link]
- **RQ-2**: How to ensure financial contents reach more people?
  - **Relevant Chapter**: 2
  - **Relevant Contributions**: Generator-Guided Crowd Reaction Assessment [6]
- **RQ-3**: How to improve the investment process?
  - **Relevant Chapter**: 3
  - **Relevant Contributions**: Detecting hypernyms of Financial Terms [7] [8], Extracting relationship between financial entities [9]
- **RQ-4**: How to ensure that the investments are towards betterment of the Earth and have positive impact towards the environment?
  - **Relevant Chapter**: 4
  - **Relevant Contributions**: Detecting ESG and sustainability related concepts [10], issues [11], impact type [12], and impact duration [13].
- **RQ-5**: How to safeguard investors from misinformation?
  - **Relevant Chapter**: 5
  - **Relevant Contributions**: Detecting in-claim numerals in financial texts [14] [15] [16] [17], Detecting Exaggerated Numerals in Financial Texts [15], Estimating profitability and loss from financial social media posts [18], Evaluating trustworthiness of social media posts by executives [19], Financial Argument Analysis [20].
- **RQ-6**: How to keep Indian investors informed?
  - **Relevant Chapter**: 6
  - **Relevant Contributions**: Financial Argument Analysis in Bengali [21], Financial Natural Language Processing for Indian Languages [22], Predicting success [254] and ratings [255] of Indian IPOs.

Subsequently, to enable investors to make data-driven decisions, we have developed some self-service tools [23] which are presented in Chapter 7. In [24] and [25], we summarize our contributions. After mentioning some of the limitations of the thesis, we conclude in Chapter 8.

We present the tasks related to the research questions discussed previously, along with the corresponding publications, in this section.

[15] https://huggingface.co/spaces/sohomghosh/FENCE_Financial_Exaggerated_Numeral_ClassifiEr (accessed on 17th November, 2023)

1.3.1 Inclusive Investing (RQ-1 and RQ-2)

Task-1: Given a financial text (FT), we want to assess its readability and simplify it. ([4])
Task-2: Given two FTs, we want to assess which one would reach more people. ([6])

1.3.2 Improved Investing (RQ-3)

Task-3: Given a financial jargon in a FT, we would like to retrieve its hypernym. ([7], [8])
Task-4: Given two entities in a FT, we would like to determine the relationship between them. ([9])

1.3.3 Impactful Investing (RQ-4)

Task-5: Classify a FT as Sustainable or Unsustainable ([10])
Task-6: Detect ESG Issues from FTs in English. ([11])
Task-7: Identify ESG impact type & duration from FTs. ([12], [13])

1.3.4 Informed Investing (RQ-5)

Task-8: Detect exaggerated and in-claim numerals from FTs. ([14], [17], [26])
Task-9: Evaluate the Rationale of Amateur Investors. [18]
Task-10: Evaluate the trustworthiness of Social Media Posts by Executives on Stock Prices ([19])
Task-11: Fine-grained Argument Understanding in FTs ([20])

1.3.5 Indic Investing (RQ-6)

Task-12: Financial Argument Analysis in Bengali ([21])
Task-13: Extract ESG Issues, Assess Sustainability, and Detect exaggerated numerals from FTs in Hindi, Bengali, & Telugu ([22])
Task-14: Predicting direction and under-pricing with respect to Open, High, Close prices of Indian IPOs ([254]) **Task-15**: Predicting ratings of Indian IPOs ([255])

1.3.6 Tools for FinNLP

Task-16: Develop tools for processing FTs ([23], [16], [15], [5])

In Table 1.2, we put together our approaches for solving the corresponding tasks and their performances.

Task #	Metric	Approach Summary	SOTA	Performance	New Data	Language	New Tool	Publication(s)
1	AU-ROC	FinBERT finetune	Yes	0.993	Yes	English	Yes	[5], [5]
2	F1	RoBERTa + Claude (LLM)	Yes	0.731	Yes	English	No	[6]
3	Acc.	SBERT finetune	Yes	0.967	No	English	No	[8], [7]
4	F1	SEC-BERT + Neural Network	No	0.736	No	English	No	[9]
5	Acc.	RoBERTa finetune	No	0.932	No	English	No	[10]
6	F1	SEC-BERT finetune	No	0.715	No	English	Yes	[11]
7	F1	FinBERT finetune	No	0.929 (IT)	No	English	No	[12]
7	F1	Trans-Prp + FinBERT finetune	No	0.756 (IT)	No	French	No	[12]
7	F1	Trans-Prp + FinBERT finetune	Yes	0.679 (IT)	No	Japanese	No	[12]
7	F1	Trans-Prp + FinBERT finetune	Yes	0.677 (IT)	No	Chinese	No	[12]
7	F1	Trans-Prp + PLM finetune	No	0.5882 (ID)	No	English	No	[13]
7	F1	Trans-Prp + PLM finetune	Yes	0.5616 (ID)	No	French	No	[13]
8	F1	Ensemble (FinBERT, BERT + Logistic Regression)	No	0.948	No	English	Yes	[14], [17], [16], [15]
9	MPP, ML	SBERT Chinese + Classifier, FinBERT	No	0.575 (MPP), 0.598 (ML)	No	Chinese	No	[18]
10	MAPE	Gated Recurrent Unit	Yes	0.382	Yes	English	Yes	[19]
11	F1	Cross Encoder (FinBERT Finetuned)	No	0.789	No	English	No	[20]
11	F1	Translate + Cross Encoder (SEC-BERT)	No	0.641	No	Chinese	No	[20]
12	F1	MBERT, Cross Encoder (MBERT)	No	0.721 (1st task), 0.755 (2nd Task)	Yes	Bengali	Yes	[21]
13	F1	MBERT+Classifier, Translate + RoBERTa, Translate+MBERT	Yes	0.680 (1st task), 0.950 (2nd task), 0.590 (3rd task)	Yes	Hindi	No	[22]
13	F1	MBERT+Classifier, Translate + RoBERTa, Translate+MBERT	Yes	0.650 (1st task), 0.920 (2nd task), 0.550 (3rd task)	Yes	Bengali	No	[22]
13	F1	MBERT+Classifier, Translate + RoBERTa, Translate+MBERT	Yes	0.680 (1st task), 0.920 (2nd task), 0.580 (3rd task)	Yes	Telugu	No	[22]
14	F1	McL (Num, Cat, Txt) (Classification)	Yes	0.947 (Open-MB), 0.935 (High-MB), 0.931 (Close-MB)	Yes	English	No	
14	MAE	McL (Num, Cat, Txt) (Regression)	Yes	0.167 (Open-MB), 0.193 (High-MB), 0.194 (Close-MB)	Yes	English	No	
14	F1	McL (Num, Cat, Txt) (Classification)	Yes	0.893 (Open-SME), 0.942 (High-SME), 0.911 (Close-SME)	Yes	English	No	
14	MAE	McL (Num, Cat, Txt) (Regression)	Yes	0.239 (Open-SME), 0.263 (High-SME), 0.256 (Close-SME)	Yes	English	No	
15	F1	LongFormer RoBERTa Llama 3.1 70B	Yes	0.952 (MB) 0.423 (SME)	Yes	English	No	
16	NA	Gradio (frontend)	NA	NA	NA	Various	Yes	[23], [16], [15], [5]

TABLE 1.2: Approaches and results for different tasks.
AU-ROC = Area under the ROC curve, Acc. = Accuracy, MPP = Maximum Possible Profit, ML = Maximum Loss, MAE = Mean Absolute Error, MAPE = Mean Absolute Percentage Error, NA = Not Applicable, SOTA = State-Of-The-Art, LLM = Large Language Model, MSE = Mean Square Error, MB = Main Board, PLM = Pre-trained Language Model, Trans-Prp = Translate Paraphrase, IT = Impact Type, ID = Impact Duration, McL = Machine Learning, Num = Numeric Features, Cat = Categorical features, Txt = Text Features

1.4 Contributions

- **Publications**: 8 Shared Tasks, 2 Workshops, 10 Conferences, 4 Journals
- **Venues**:**ICON-2021** Silchar (India), **FinNLP@IJCAI-2021** Montreal (Canada), **FinWeb@The Web Conference-2022** Lyon (France), **FNP@LREC** Marseille (France), **FinNLP@IJCAI-ECAI-2022** Vienna (Austria), **NTCIR-16 and 17** Tokyo (Japan), **FinNLP@EMNLP-2022** Abu Dhabi (UAE), **FIRE-2022** Kolkata (India), **CODS-COMAD-2023** Mumbai (India), **ICDSA-2023** Jaipur (India), **FinNLP@IJCNLP-AACL-2023** Bali (Indonesia), **FIRE-2023** Goa (India), **TheWebConf (WWW)-2024** Singapore, **LREC-COING 2024** Torino (Italy), **CIKM 2024** Boise (USA), **PIC 2025** Mysuru (India), **ARCS 2025** Coimbatore (India)

- **Resources**: FinRAD (dataset), FinRead (tool), FinCAT (tool), FinCAT-2 (tool), FLUEnT (tool), Executive / General Tweets (dataset), FENCE - Financial Exaggerated Numeral ClassifiEr (tool), EID - ESG Issue Detector (tool), FinLanSer (tool) , CReD (dataset), FAAB (dataset + tool), IndicFinNLP (dataset), Indian IPO Success (dataset), Indian IPO Rating (dataset)
- **Awards**: CODS-COMAD 2023 YRS Track (Honourable Mention), FinArg-1@NTCIR-17 (Rank: 2nd in Task-1 Sub-Task-2), FinNLP@IJCNLP-AACL 2023 (Rank: 1st in Chinese and Japanese), ACM Travel Grant to attend PhD Clinic co-located with CODS-COMAD 2024, FinNLP@LREC-COLING 2024 ESG Impact Duration (Rank-3 in English impact length & impact type and Rank-1 in French Impact Length), SIGIR Travel Grant to present paper at CIKM-2024, Kaggle Datasets Expert, Travel Grant IndoML-2024, ACM ARCS 2025 Travel Grant
- **Invited Talks**: 4 (XIM, Bhubaneshwar, India; Haldia Institute of Technology, Haldia, India; Brainware University, Kolkata, India; Yobe State University, Nigeria)

Chapter 2

Inclusive Investing

For improving financial literacy, it is essential to make the process of investment inclusive, i.e. financial content should be easy to comprehend and should reach as many people as possible. We address the following research questions in this chapter.

2.1 Research Questions

- **RQ-1**: How to quantify and improve readability of financial texts?
    - **Relevant Contributions**: Financial Readability Assessment Dataset (**FinRAD**) [4], A Transfer Learning Based Tool to Assess Readability of Definitions of Financial Terms (**FinRead**) [5], Financial Language Simplifier (**FinLanSer**) [Screenshot in Figure 2.8, demonstration link]
- **RQ-2**: How to ensure financial contents reach more people?
    - **Relevant Contributions**: Generator-Guided Crowd Reaction Assessment [6]

2.2 Improving Readability of Financial Texts

2.2.1 Introduction

Nowadays investors prefer to avail themselves of financial services online. This saves time as well as money. While making decisions relating to investments, they tend to read relevant content online. Not all financial content is easy to comprehend due to the presence of unknown terms. In such cases, they need to look for definitions of these terms. Interestingly, not all definitions are easy to understand. Thus, it is extremely important to aid financial content writers to assess how readable are the definitions they write are. Figure 2.1 illustrates this concept.

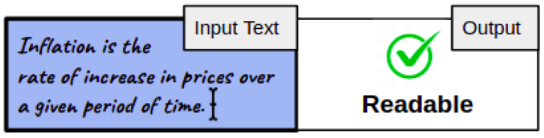

FIGURE 2.1: Readability of definition of "inflation"

We presented **FinRead**, a basic tool for demonstrating such a system in the 18th International Conference on Natural Language Processing (ICON-2021)[1] [5]. It was trained using **definitions** of 8,401 **financial terms**. In this chapter, in addition to extending this dataset to 13,112 **definitions** of **financial terms**, we have released it publicly. Subsequently, we present several enhancements to the baseline architectures.

Our contributions

- We created a dataset comprising more than thirteen thousand **definitions** of **financial terms** along with their embeddings, standard formula-based readability scores and assigned readability (**AR**) scores. We released it under the CC BY-NC-SA 4.0 license. To the best of our knowledge, we are the first to study readability in the context of short financial texts, provide the first dataset on financial terms and propose a readability measure. A sample dataset can be downloaded from here [2]
- We showed that standard rule-based readability scores (like ARI, FRI, DCF, SMOG, etc.) do not work well for financial texts.
- We proposed baseline architectures to automatically classify definitions of financial terms as readable or not.

The overall process flow is summarised in Figure 2.2. The rest of the chapter is structured as follows. Section 2.2.2 states the prior works and their connection with this work. In section 2.2.3 we describe the process we followed to collect, clean and label the data. Subsequently, we discuss various exploratory data analyses that we have performed. In

[1] http://icon2021.nits.ac.in/coloc_events.html (accessed on 18th September, 2023)
[2] https://github.com/sohomghosh/FinRAD_Financial_Readability_Assessment_Dataset (accessed on 18th September, 2023)

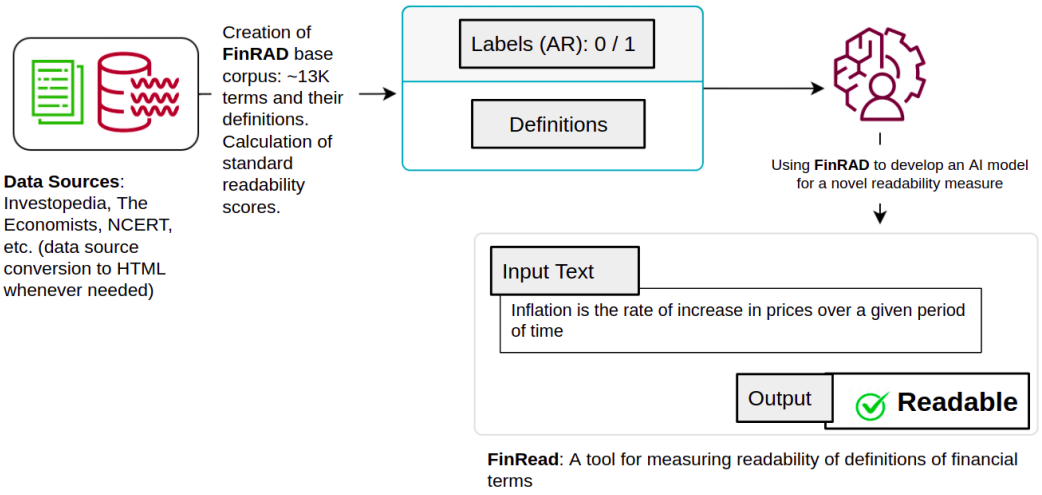

FIGURE 2.2: Overall process flow for `FinRAD`

section 2.2.4 we formally describe the task of assessing readability. We present various neural baseline architectures and their performances in section 2.2.5. In Section 2.2.6 , we present the Financial Language Simplifier (FinLanSer) tool. Section 2.2.7 concludes the chapter and provides some future directions of research.

2.2.2 Related Works

In this section, we discuss the prior works. Firstly, we describe applications of readability in general and in the context of the financial domain. We then explore some of the related works and datasets.

2.2.2.1 Readability in general

For Natural Language Processing (NLP) practitioners, understanding readability of texts has always been an active area of research. Some of the standard readability scores include: "Flesch Reading Index (FRI)" [27], "Automated Readability Index (ARI)" [28],"SMOG Index Score (SIS)" [29] and "Dale-Chall formula (DCF)" [30]. Flesch was one of the pioneers in this area. He proposed FRI which uses the ratio of total words to sentences and that of total syllables to total words as a measure of readability. Smith et al. [28] defined ARI based on characters to words and words to sentences ratio. This score was used to assign the readability of a text to one of the fourteen predefined grade levels ranging from kindergarten to college student. Another new formula SIS for calculating readability was proposed by Mc Laughlin. It comprised of calculating the ratio between the number of polysyllables and sentences. However, it was only applicable for texts having at least 30 sentences. In the paper [31], Rush criticised these scores as they only dealt with the syntactic aspect of the texts and did not consider the aspect of the reading process which was interactive. Papers that criticised these formulas include [32] and [33]. Zamanian et al. [34] presented a more detailed review of these formulas along with their advantages, and disadvantages. Some of the papers which used language models to estimate readability

include [35], [36], [37] and [38]. In his recently published study of readability of "Policy Documents on the Digital Single Market of the European Union", Ruohonen [39] argued that a PhD level education would be required to study and understand the Digital Single Market (DSM) laws and policy documents. He further observed that there were critical differences in terms of the degree of agreement in various standard readability scores. The study also demonstrated, how the readability grades across time have evolved for the laws and policy documents in DSM as well. This in turn also indicates that the existing readability scores may fail to capture domain-specific nuances for the different types of documents.

2.2.2.2 Readability in Financial Domain

Readability of financial texts has been widely explored. Most of these texts include Financial Disclosures [40], [41], Annual Reports, Management Discussions and Analysis (MD&A) [42], [43], [44], [45]. In addition to general features, Bonsall et al. [46] used the file size of 10-K documents to measure their readability. Bonsall et al. [46] proposed a new index "Bog Index" as a "plain English measure of financial reporting readability". It served as one of the standard approaches for the readability of financial reports. Loughran et al. [47] proposed a new method of measuring readability based on recommendations made by the U.S. Securities and Exchange Commission (SEC) in the year 1988. Readability scores were used for various downstream tasks like fraud detection [48], Stock Price Crash Risk prediction [49], etc. Readability of financial text books has been studied in [50], [51] and [52]. They also highlighted the limitations of these popular scores as a measure of readability due to their inherent shortcomings to deal with domain specific language and jargon. Loughran et al. [40] also highlighted the need for alternative measures of readability for the financial documents, such as disclosures. Pitler [53] proved that surface level standard readability scores do not correlate with the human assigned readability scores on the Wall Street Journal corpus. They further showed that a combination of entity coherence and discourse relations were the best features for assessing readability.

2.2.2.3 Related datasets

Related financial datasets on which readability has mostly been explored include 10-K SEC filing reports [54], disclosures [55], [56], and accounting textbooks [50], [52].

2.2.2.4 Difference with prior works

To the best of our knowledge, we are the first ones to create a dataset consisting of definitions of financial terms along with their readability scores based on their complexity. We also propose transformer based neural baselines to automatically assess the readability of such definitions.

2.2.3 Dataset

In this section, we narrate how we collected the data, cleaned and annotated it.

Tag	Source Description	AR	# Terms/Definitions
prin	*Principles of Corporate Finance* by Richard A. Brealey, Stewart C. Myers, Franklin Allen [57]	0	177
zvi	*Investments* by Zvi Bodie Alex Kane Alan J. Marcus [58]	0	492
palgrave	*The Palgrave Macmillan Dictionary of Finance*, Investment and Banking by Erik Banks [59]	0	3925
opod	*Options, Futures, and Other Derivatives*, Global Edition by John C. Hull [60]	0	527
fmi	*Financial Markets and Institutions* by Frederic S. Mishkin Stanley Eakins [61]	0	387
ncert_keec111	*NCERT Indian Economic Development Economics Class 11*[3]	1	95
ncert_kest	*NCERT Statistics for Economics Class 12*	1	53
ncert	*NCERT Introduction to MacroEconomics Class 12*	1	115
ncert_class12_econ	*NCERT Introduction to MicroEconomics Class 12*	1	41
investopedia	*Investopedia* Data Dictionary[4]	1	5946
economist	*The Economist* terms dictionary[5]	1	457
6_8_louis	*Glossary of Economics and Personal Finance Terms* from Federal Reserve Bank of St. Louis[6]	1	342
9_12_louis	*Glossary of Economics and Personal Finance Terms* from Federal Reserve Bank of St. Louis	1	188
pre_louis	*Glossary of Economics and Personal Finance Terms* from Federal Reserve Bank of St. Louis	1	36
sam	*Economics Textbook* by Paul Samuelson and William Nordhaus [62]	1	331

TABLE 2.1: Source wise distribution. AR: Assigned Readability, #: Count

2.2.3.1 Data collection

Our dataset consists of 13,112 **`financial terms`** and their **`definitions`** written by experts across multiple sources. These sources include glossaries, dictionaries from financial websites, school and graduate-level textbooks relating to economics and finance. We collected the terms from 13 different sources and removed the duplicated terms during pre-processing. Source wise distribution of the dataset is presented in Table 2.1.

2.2.3.2 Data extraction and cleaning

Only three of the data sources considered were available as web-pages which we scraped directly. They include websites of The Economist, Federal Reserve Bank of St. Louis, and Investopedia. Other datasets were available in Portable Document Format (PDF). We tried extracting the terms and definitions directly from these PDFs first. However, we found that in most of the cases we were losing out on the structure. Thus, separating the terms from the definitions was challenging. Subsequently, we converted these PDF documents to the Hypertext Markup Language (HTML) format. For this, we used various freely available online services. We removed irrelevant texts like page numbers, the word "glossary", and texts which were mistakenly identified as terms. We removed extra spaces and manually checked the final dataset to ensure that it is of high quality.

2.2.3.3 Data Annotations

Inspired by the method followed by Chakraborty et al. [63], we consulted several professional financial experts. Subsequently, we decided to assign readability scores (**AR**) to the definitions of financial terms based on their sources. This was done since readability is subjective and manually annotating the entire dataset is expensive. Definitions from the following sources were assigned a readability score of 1.

- school-level textbooks (like NCERT textbooks, economics textbooks for beginners [62])
- public websites suitable for masses (like Investopedia and The Economist).

The reason behind this is that the information from these sources is mostly consumed by beginners, school students, and by the masses. To understand the definitions which were obtained from other sources one needs to have at least undergraduate level knowledge specific to the financial domain. Thus, they were assigned a readability score of 0. This gave us 7,604 and 5,508 instances with readability scores of 1 and 0 respectively. An **AR** score of 1 represents the terms' definitions that are easily readable and 0 represents the definitions that are comparatively complex in nature or less readable. To validate this assumption we identified 112 additional terms and extracted their definitions from both kinds of sources (i.e. with **AR** = 0 and 1). We manually inspected each of the definitions and assigned them a readability score (0 or 1). In 79.91 % of the cases the manual assignment was in agreement with the assumption.

2.2.3.4 Exploratory Data Analysis

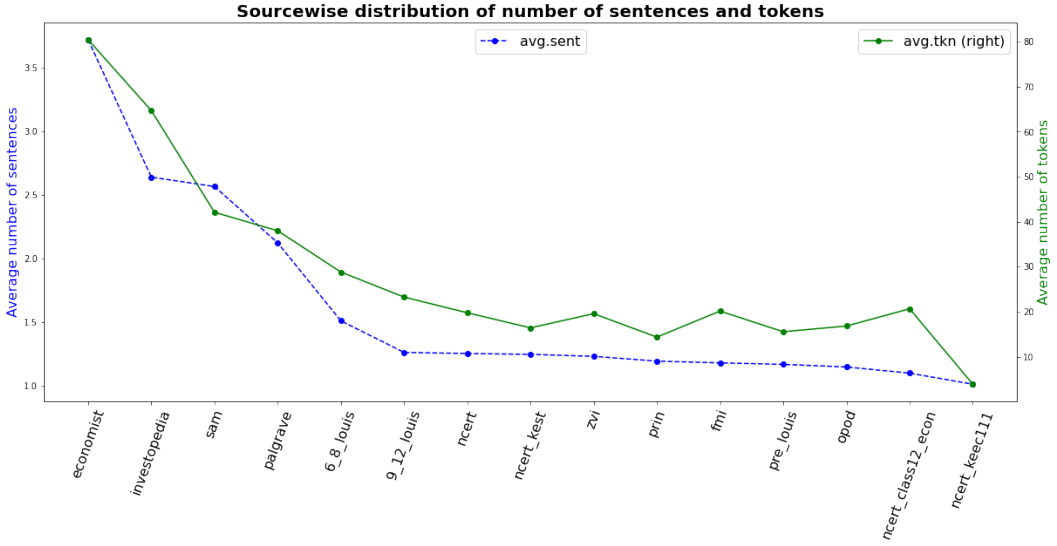

FIGURE 2.3: Source-wise distribution of the average number of sentences and tokens per definition

In this section, we present an overview of the **FinRAD** dataset and its contents. The dataset consists of 4 key fields:

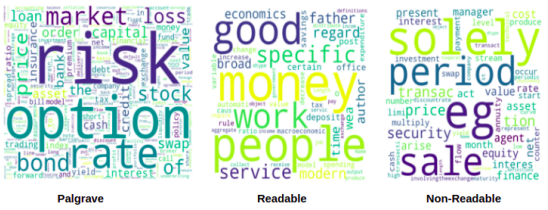

FIGURE 2.4: Word clouds of definitions from "Palgrave", readable and non-readable sources

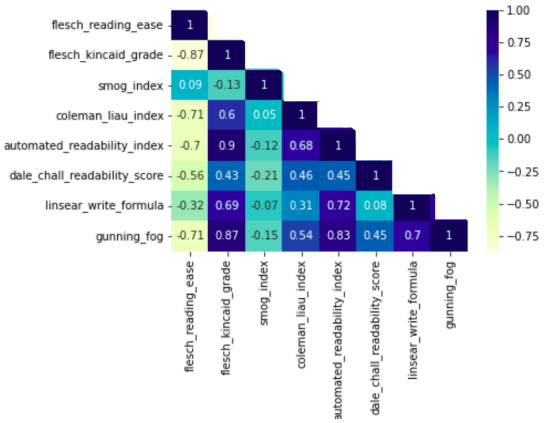

FIGURE 2.5: Correlation between standard readability scores

- **`financial terms`** (i.e. the terms that have been collected from different sources)
- **`definitions`** (i.e. the descriptions or definitions of these terms)
- **`source`** (i.e. the sources from which these terms have been collected)
- assigned readability (**`AR`** i.e. the annotated readability)

Apart from these 4 fields, the dataset also includes readability scores extracted using traditional methods. So far, 8 different scores have been provided for the definitions of the financial terms: Flesch Reading Ease (FRE) Score[27], Flesch-Kincaid Grade Level (FKGL) Score[64], SMOG Index(SI) Score[29], Coleman – Liau Index(CLI) Score[65], Automated Readability Index(ARI) Score[28], Dale – Chall Readability (DCR) Score[30], Linsear write Formula and Gunning's Fog Index (FOG) Readability Formula. For all the definitions, these scores have been calculated using the textstat[7] library.

We started by studying the distribution of the number of sentences in the definitions across different sources. Figure 2.3 summarizes the distribution of the average number of sentences per definition used to define the terms across various sources. As evident from this plot, "The Economist" have definitions with the highest average number of sentences (approximately 4 sentences). We further compared the average number of sentences per definition across assigned readability segments in Table 2.2. It is quite interesting to note that the average number of sentences per definition in the readable set is higher than that of the non-readable set. Moreover, the average sentence length (i.e. number of tokens per

[7] https://pypi.org/project/textstat/ (accessed on 18th September, 2023)

Readability type	**Avg. sentences**	**Avg. tokens**
Non-readable (0)	1.8529	32.2912
Readable (1)	2.5494	59.7701

TABLE 2.2: Average number of sentences and tokens per definition

sentence) for the readable set is 24.03 and that for the non-readable set is 17.22. This is because authors tend to use more words and shorter sentences to simplify concepts.

Subsequently, we studied the distribution of the average number of tokens present in the **definitions** across different sources. Figure 2.3 illustrates this. The average number of tokens per definition are approximately 80 and 64 for the definitions obtained from the readable sources "The Economist" and "Investopedia" respectively. This reconfirms our previous findings that authors tend to explain more to simplify concepts. In addition to this, we compared the average number of tokens across different readability segments. We observe that readable definitions have around 27 more tokens than that of non-readable ones. We provide more details and exact numbers in Table 2.2.

Word clouds are quite helpful in generating meaningful insights about text data. They offer an interesting option to visually represent the frequency of different words present in a corpus.

For ease of exposition, we have presented the word clouds of terms for one of the key sources ("Palgrave") in Figure 2.4. It accounts for almost 30% of the entire dataset of terms. Furthermore, for effective comparison we also present word clouds of non-readable and readable definitions of financial terms in the same figure. Quite evidently, the frequent terms present in the non-readable definitions (**AR**=0) are more complex than those of the readable ones (**AR**=1).

Lastly, we study the correlation between the standard readability scores and present them in Figure 2.5. Now, it is apparent that all the scores cannot be directly compared as they are generated using different mathematical principles. However, for a few scores which are comparable like Flesch Reading Ease formula [27] and The Flesch-Kincaid Grade Level [64], the positive correlation is high. Similar conclusions can be drawn for other scores as well.

2.2.4 Task

Given a set $\mathfrak{D} = \{d_1, d_2, d_3, \ldots, d_n\}$ of **definitions** of **financial terms** and a set $\mathcal{R} = \{r_1, r_2, r_3, \ldots.., r_n\}$ of readability scores where r_i is the assigned readability (**AR**) corresponding to the definitions of financial term d_i and $r_i \in \{0, 1\}$. **AR**=0 denotes non-readable and **AR**=1 denotes readable. The task is to develop a system capable of classifying a definition as readable or not. Furthermore, it should be able to automatically compute readability score r_t for **definition** of any unknown **financial term** d_t. Note: $0 \leq r_t \leq 1$.

We use Area Under the Receiver Operating Characteristic curve (AU-ROC) score as the evaluation parameter.

2.2.5 Models and Results

We divided the dataset into two parts keeping the event rate the same - the training set (67%) and the validation set (33%). Firstly, we studied how standard readability scores (like FRI, ARI, SIS, DCF, etc.) performed in a domain-specific setting like this. Most of these scores provided grade levels as outputs. We calculated the AU-ROC, F1 and Accuracy considering readability of grade level higher than 12 as 0 and rest as 1. This was done following our assumption stated in section 2.2.3.3. The performance of these standard scores in measuring readability on the validation set is presented in Table 2.3. The performance on the validation set which was calculated using these scores was not up to the mark. The best AU-ROC was only 0.4986 using the Flesch Reading Index. Thus, we trained machine learning based classifiers to assess the readability of the **`definitions`**.

We represented **`definitions`** of the terms numerically using a Term Frequency - Inverse Document Frequency (TF-IDF) matrix. We trained various machine learning based classifiers over it such as Logistic Regression, Random Forest [66] and Gradient Boosting Machine [67] and the results of these models are presented in Table 2.4. Furthermore, we experimented by replacing TF-IDF with sentence embeddings [68] created using BERT [1] and FinBERT [69]. In addition to this, we tried using other machine learning based classifiers like LightGBM [70] and XG-Boost [71]. This improved the AU-ROC on the validation set to 0.969. Finally, we fine-tuned the financial domain-specific language model FinBERT (768 dimensions) [69] for the downstream task of classifying definitions. It was trained for 20 epochs with a batch size of 256, maximum sequence length of 64, and a learning rate of 0.00002. This model out-performed all the other algorithms (**AU-ROC** = 0.9927, **Matthews Correlation Coefficient** = 0.9063, **Accuracy** = 0.9540 and **F1 Score** = 0.9610) on the validation set. The corresponding ROC curves are presented in Figure 2.6.

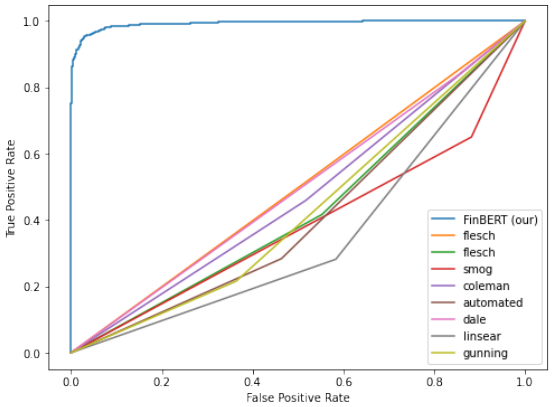

FIGURE 2.6: ROC curves

2.2.6 Financial Language Simplifier (FinLanSer)

To help the financial content writers and investors, we developed the tool **`FinLanSer`** - Financial Language Simplifier. Given a financial text, it generates paraphrases using [72], scores each of the paraphrases using the fine-tuned FinBERT model described in §2.2.5 and returns the most readable paraphrase (demonstration link).

Readability Score (RS)	RS Description	AU-ROC	F1	Accuracy
flesch_reading_ease	The Flesch Reading Ease formula [27]	0.4986	0.5516	0.5034
flesch_kincaid_grade	The Flesch-Kincaid Grade Level [64]	0.4320	0.4573	0.4296
smog_index	The SMOG Index [29]	0.3841	0.5661	0.4250
coleman_liau_index	The Coleman-Liau Index [65]	0.4710	0.4995	0.4691
automated_readability_index	Automated Readability Index[28]	0.4100	0.3494	0.3906
dale_chall_readability_score	Dale-Chall Readability Score [30]	0.4922	0.6793	0.545
linsear_write_formula	Linsear Write Formula[8]	0.3492	0.3295	0.3388
gunning_fog	The Fog Scale (Gunning FOG Formula)[9]	0.4259	0.2908	0.3936

TABLE 2.3: Performance of standard readability scores

Algorithms	Validation AU-ROC
TF-IDF vectors + Logistic Regression	0.9038
TF-IDF vectors + Random Forest	0.8866
TF-IDF vectors + Gradient Boosting Classifier	0.9116
BERT ST embeddings + Logistic Regression	0.9544
BERT ST embeddings + Random Forest	0.8801
BERT ST embeddings + Gradient Boosting Classifier	0.9063
FinBERT ST embeddings + Logistic Regression	0.9691
FinBERT ST embeddings + Random Forest	0.9434
FinBERT ST embeddings + Gradient Boosting Classifier	0.9523
FinBERT ST embeddings + Light GBM Classifier	0.9640
FinBERT ST embeddings + XGBoost Classifier	0.9626
FinBERT (fine-tuning [CLS] token)	**0.9927**

TABLE 2.4: Performance of models trained using Machine Learning

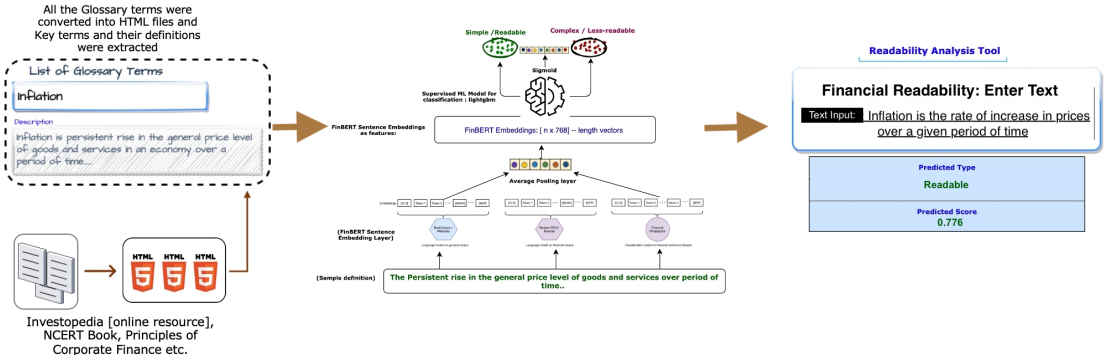

FIGURE 2.7: Financial Readability Flow Chart and Tool. It represents how the FinRAD dataset was created, and the FinRead tool was developed.

2.2.7 Conclusion

In this chapter, we presented a new dataset **`FinRAD`** for the task of evaluating the readability of **`definitions`** of **`financial terms`**. We explored the limitations of various standard formula-based readability scores which were developed to assess the readability of English texts in general. We proposed a neural architecture that outperformed all such scores in terms of AU-ROC. Finally, we developed two user friendly tools **`FinRead`** (Figure 2.7) and **`FinLanSer`** (Figure 2.8) to assess readability of financial texts and simplify them.

There are several directions in which this research can be extended in future. We present some of the research questions (RQ) here.

- **RQ1:** *Do the predicted readability scores correlate with human judgement?*
  To understand this, we need to perform a qualitative analysis of the predicted readability scores generated automatically using machine learning algorithms. This may need additional manual tagging which is subjective and expensive. If the correlation is less, it would be essential to manually tag more definitions before developing any machine learning based classifier.

- **RQ2:** *Can we have better metrics to measure the performances of the models?*
  Presently, we use the Area Under the Receiver Operating Characteristic curve (AUROC) to measure the performance of the models. An interesting direction would be to develop a new metric that correlates more with human judgement.

- **RQ3**: *Can we develop unsupervised formula-based readability scores specific to the financial domain?*
  Machine learning based supervised models are computationally expensive and need lots of data. Thus, it would be nice to explore if we can generate unsupervised formula-based readability scores specifically for the financial domain.

- **RQ4**: *Can we use Natural Language Generation methods to simplify definitions?*
  We removed duplicate terms while creating the **`FinRAD`**. A dataset consisting of readable as well as non-readable definitions for a given term would complement this. Simplification of complex definitions using Natural Language Generation techniques could be a new dimension to this research.

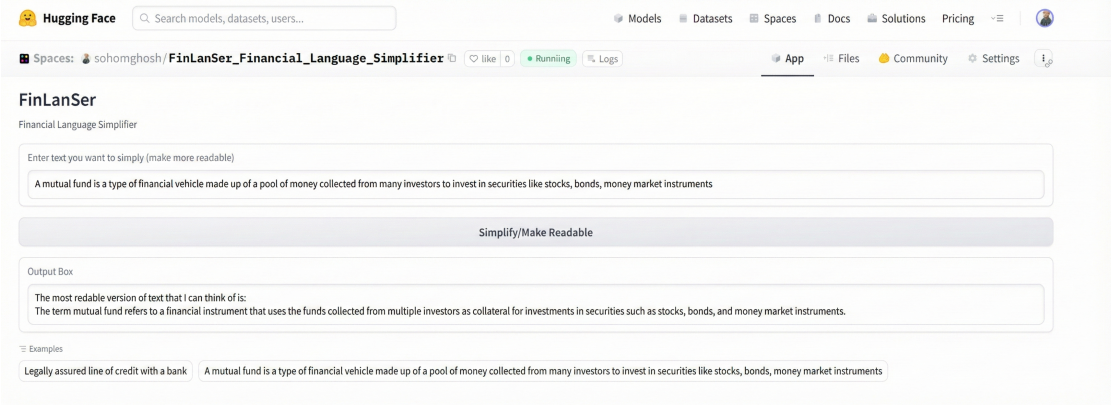

FIGURE 2.8: Financial Language Simplifier (FinLanSer)

FIGURE 2.9: Example Tweets from The White House.

2.3 Improving Reachability of Financial Texts

2.3.1 Introduction

In the swiftly evolving landscape of social media, possessing the capability to accurately anticipate the impact of a post is invaluable. Digital marketers, content creators, and organizations with the intent of influencing or interacting with their audience can strategically enhance their approach and decision-making processes if they can predict post performance prior to its publication. Nevertheless, the challenge at hand, referred to as **C**rowd **Re**action **A**ssess**M**ent (CReAM), is complicated due to the multifaceted and ceaselessly changing nature of social media engagement.

As social media platforms continue to grow, they become progressively richer data sources reflecting public sentiment and reactions. Despite this, the potential for utilizing such data to predict crowd reactions remains largely uncharted territory. There exists a pressing demand to devise and execute efficacious methods capable of accurately predicting a post's influence, be it in fostering engagement, initiating dialogue, or inspiring action. Comprehending these dynamics not only holds commercial value but also carries significant potential for influencing public discourse and democracy.

In the process of preparing a social media post, managers often draft multiple versions and select the final one based on an estimation of crowd reactions. As illustrated in Figure 2.9, when considering two tweets on an identical subject, determining which will receive a greater volume of interactions and responses is a crucial consideration for a social media manager. This chapter focuses on tweets issued by The White House and introduces a meticulously curated dataset, the Crowd Reaction Estimation Dataset (CRED). Specifically, CRED encompasses pairs of tweets paired with comparative metrics of retweet counts.

Recognizing the widespread applications of large language models (LLMs) such as ChatGPT (based on variants of GPT models [73], FLAN-UL2 [74, 75]), and Claude across various sectors,Please refer to Appendix 2.3.8.2 for details about the models. We propose employing their analytical capabilities to address the CReAM task. We suggest a Generator-Guided Estimation Approach (GGEA) that seeks LLMs' analysis on disparate tweets, utilizing the resulting enhanced information to guide classification models in making predictions. The findings from our study indicate that the proposed GGEA contributes to superior

performance compared to merely fine-tuning classification models and LLMs under one-shot settings. The most successful combination involves a fine-tuned FLANG-RoBERTa model [76] with tweet content and responses generated by Claude.

The major contributions of this chapter are as follows:

- The introduction of a novel dataset, Crowd Reaction Estimation Dataset (CRED), which simulates the decision-making process of governmental social media managers.
- The proposal of a Generator-Guided Estimation Approach (GGEA) that leverages the analytical power of LLMs to predict crowd reactions to social media posts. The demonstration of GGEA's ability to assess paraphrases of a post, providing users with different versions of their content to optimize engagement.

2.3.2 Related Work

The aspiration to maximize audience reach and reactions has been a consistent theme among social media users since the platform's inception. Several studies have focused on understanding and predicting the factors that contribute to the widespread reach of social media posts. Cha et al. [77] conducted an analysis of the activities of 52 million Twitter users, deducing that a tweet's content is a critical factor in driving retweets. Suh et al. [78] assessed 74 million tweets and explored a range of content-based and contextual features that affect the retweet count of a given tweet. They developed a Generalized Linear Model-based model to predict the number of retweets a tweet is expected to generate. Their research underscored the importance of tweet content, URLs, and hashtags as key features in predicting retweet count. Other works on cascade prediction which leverages social networks include [79] and [80].

Similarly, Bakshy et al. [81] tracked the diffusion activities of 1.6 million Twitter users, by studying the relationships between different categories of influencers and the associated cost. Beyond Twitter, research has also been conducted on the reach of Facebook posts. Bernstein et al. [82] studied the activities of 222,000 Facebook users, revealing how invisible audiences enhance the reach of Facebook posts. Tan et al. [79] analysed the effect of content of a tweet on its reach. Valkonen et al. [83] investigated how the timing of Facebook posts by Finnish consumers affects reachability. More recent studies like Gabriel et al. [84] demonstrated how the reach of news articles can be estimated using models based on GPT-2 [85] and T5 [86]. However, their approaches provide only a coarse-grained analysis and do not significantly contribute to performing a comparative analysis. They created a dataset by annotating reach levels from 1 to 5.

Our work differs from these preceding studies in several significant ways. First, we introduce a new dataset, CRED, that specifically targets the decision-making processes of social media managers in governmental contexts. Second, we propose a novel approach, the GGEA, that employs the analytical power of LLMs to predict crowd reactions to social media posts. Finally, unlike previous works, we use a T5-based paraphraser fine-tuned with ChatGPT responses [72] over GGEA for generating and assessing paraphrases of a given post, thereby allowing users to optimize engagement by exploring different versions of their content.

Topic	Avg. RT	Pairs
Business & Entrepreneurs	743.2	4,186
Fitness & Health	1010.2	628
Learning & Educational	485.1	90
Sports	728.0	6
Total		4,910

TABLE 2.5: Statistics of CRED. Avg. RT denotes the average number of retweets.

2.3.3 Dataset

Our primary task is to determine whether a tweet, denoted as $t1$, will receive more retweets than another tweet, denoted as $t2$. Given a pair of tweets, our objective is to predict the tweet with the higher retweet count accurately.

We compile our dataset from all tweets posted by the official White House Twitter handle from November 2020 to October 2022. We then proceed to curate pairwise data, ensuring each pair adheres to conditions related to temporality, topic, and crowd reaction.

We retained only those instances that fulfill the following four conditions:

- The tweets were posted during weekdays: Weekdays generally witness higher levels of social media activity, making them ideal for capturing more interactions and thus providing a more reliable measure of crowd reaction.
- The retweet counts of $t1$ and $t2$ differed by at least 10% of the less re-tweeted tweet. This condition ensures a clear distinction in the crowd reaction between the two tweets, enabling a more accurate and meaningful comparison.
- $t1$ and $t2$ were posted within 10 days of each other and the difference between the times of posting $t1$ and $t2$ was less than or equal to five hours. A short interval between posts helps control the extraneous variables such as time-specific events that might impact retweet counts.
- We utilized the Twitter Topic classification model [87] to classify tweets into topics. We considered only those tweet pairs where the model assigned the same topic to both the tweets with a probability of 0.8 or higher.

  This high probability threshold ensures that the topics of the tweets are clearly defined and comparable, which is critical for the meaningful analysis of crowd reactions to different tweets on the same topic.

Table 2.5 presents the statistics of our proposed Crowd Reaction Estimation Dataset (CRED), and the label distribution in the proposed CRED is well-balanced. The average number of retweets across different topics underscores the importance of topic control during dataset construction. Topic-wise statistics further highlight that the majority of government tweets pertain to business and entrepreneurship. For dataset separation, we use tweets from the initial 1.5 years, i.e., till April 2022, which constitutes 4,304 tweet pairs for training. The final 6 months, i.e., May 2022 to October 2022, provide 606 tweet pairs which were used for model validation.

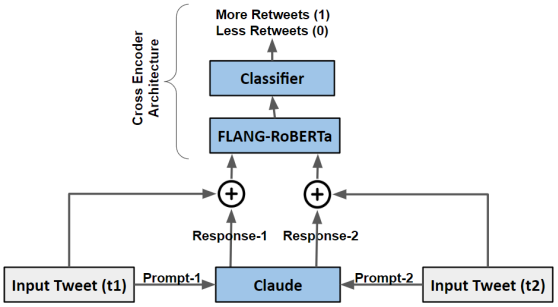

FIGURE 2.10: Architecture of GGEA.

Setup	Model	Accuracy	F1
Classifier	FLANG-RoBERTa	50.0%	66.7%
Zero-Shot	FLAN-UL2	50.2%	2.6%
	Claude	51.0%	9.2%
	React T5	52.3%	30.0%
GGEA	GGEA (ChatGPT)	67.2%	68.1%
	GGEA (FLAN-UL2)	69.6%	70.1%
	GGEA (Claude)	**71.9%**	**73.1%**

TABLE 2.6: Experimental results.

2.3.4 Methodology

We present the overall architecture of GGEA in Figure 2.10. The architecture incorporates a FLANG-RoBERTa [76] model, fine-tuned with the cross-encoder transformer architecture. The input is a concatenation of the text from a given pair ($t1$,$t2$) and the response obtained from prompting Claude to elucidate why a particular tweet in the pair is more engaging. The expected output is whether $t1$ will receive more retweets than $t2$.

To augment usability during inference, we employed an existing open-source T5 [86] based paraphraser [72] that was trained on paraphrases generated by ChatGPT. This paraphraser is utilized to generate alternatives for a given social media post. We examine two paraphrases at a time. With the assistance of our developed frameworks, the text predicted to garner maximum reactions is selected from a candidate list comprising the paraphrases and the original post. For the Claude part, two types of prompts are employed. First, we ask the LLM whether the first tweet will attract more reactions than the second. Second, we prompt the LLM with one tweet at a time, querying it to explain the reasons that make the given tweet engaging. These responses are either concatenated with the tweets or treated as independent inputs. Further details are provided in the Appendix.

2.3.5 Experiment

2.3.6 Experimental Settings

Our experimental setup compares the outcomes with two categories of baselines: supervised models and zero-shot LLMs. The first heuristic baseline fine-tunes FLANG-RoBERTa [76]

	Accuracy	F1
GGEA (FLANG-BERT)	62.7%	62.1%
GGEA (RoBERTa)	69.8%	69.5%
GGEA (FLANG-RoBERTa)	**71.9%**	**73.1%**

TABLE 2.7: Analysis of changing classifiers in GGEA (Claude).

Topic	# Instance	Accuracy	F1
Business & Entrepreneurs	570	73.0%	74.2%
Fitness & Health	34	52.9%	50.0%
Learning & Educational	2	100.0%	100.0%

TABLE 2.8: Topic-wise evaluation. We did not have any instances for the Sports category in the test set.

to make direct predictions. To evaluate the feasibility of deploying zero-shot LLMs for the proposed task, we let FLAN-UL2 [74, 75], Claude, and React T5 [84] make predictions. Going one step further, we solicit various LLMs' analysis of $t1$ and $t2$ to determine which LLM can guide the classification model to deliver enhanced performance. The performance of our models is measured using two metrics: accuracy and F1 score.

2.3.7 Experimental Results

Table 2.6 presents the experimental outcomes. Initial observations reveal that both the classification model and the LLMs struggle to excel in the proposed task. However, it is noteworthy that the proposed GGEA achieves superior performance, irrespective of the LLM consulted. We observed that the difference in performance between the GGEA (Claude) i.e. LLM augmented setup and React T5 (the best performing model in zero-shot setup), is significant (p-value$<$0.05). This result underlines the importance of collaboration between LLMs and classification models, suggesting that the generated analysis by LLMs successfully augments the information needed for models to make decisions.

Recognizing that different classifiers may influence the performance, we experimented with different classifiers under the GGEA (Claude) setting. Table 2.7 presents the experimental outcomes. These results suggest that FLANG-RoBERTa outperforms the other tested classifiers.

Finally, topic-wise evaluations are presented in Table 2.8. We observe that the Learning & Educational and Sports topics have a limited number of instances. Consequently, we compare the remaining two topics and conclude that GGEA outperforms Fitness & Health in the Business & Entrepreneurs topic, primarily due to the larger number of available training instances for the Business & Entrepreneurs topic.

2.3.8 Conclusion and Future Directions

This study provides a novel perspective on predicting crowd reactions to social media posts, introducing a new dataset and a methodology. We curated the Crowd Reaction Estimation Dataset (CRED) which encapsulates the decision-making process of governmental social

media managers through the lens of comparative retweet metrics. We also proposed the Generator-Guided Estimation Approach (GGEA), a method that harnesses the analytical prowess of large language models (LLMs) in a cooperative framework with classification models. Our experiments demonstrate that the GGEA, especially when incorporating Claude's analysis, outperforms mere fine-tuning of classification models and exhibits superior performance in predicting crowd reactions. These results highlight the potential of LLMs to augment classification models by providing nuanced analysis and further information, thereby enhancing decision-making processes.

Looking ahead, there are several intriguing avenues for further research. Firstly, GGEA could be expanded to work with other forms of social media, such as Facebook posts or Instagram captions, enabling a more comprehensive understanding of social media engagement across platforms. Secondly, GGEA could be adapted to handle multi-class problems, such as predicting whether a post will go viral or not, rather than just focusing on binary outcomes. Furthermore, the introduction of more sophisticated paraphrasing methods could potentially enhance the performance of the GGEA. Finally, an in-depth exploration of the factors contributing to the performance of different LLMs in the GGEA framework could yield valuable insights.

Limitations

While this chapter presents promising results, it is essential to acknowledge its limitations to provide a comprehensive understanding of our approach and to guide future research.

Firstly, our study relies heavily on data from a single source, specifically the official Twitter account of the White House. As a result, our model might be overfitted to the specific style, tone, and content typically associated with this account, possibly reducing its generalizability to other social media accounts, topics, or platforms. We prefer retweets to other metrics (like, quote counts) because retweets are one of the strongest means for spreading information [88]. We use tweets from only one Twitter account i.e. White House, because tweets from official handles like the WhiteHouse are authentic, unbiased, and trusted. They have a wider reach and impact the lives of people directly. Additionally, different accounts have different numbers of followers which affects the reach and engagement directly. Thus, it would be unfair to compare the number of retweets for a given tweet made by a normal person with that of an influencer. Secondly, the engagement of a social media post is influenced by various factors beyond the scope of this chapter, such as the timing of the post, the current social climate, the influence of previous posts, the visual content, and the audience demographics, among others. Our model currently does not account for these factors, which could lead to inaccuracies in its predictions. Thirdly, our study focuses solely on the metric of retweets for predicting crowd reactions. Although retweets are a significant aspect of engagement on Twitter, other elements such as likes, comments, and shares also contribute to a post's overall impact. Tweets belonging to different topics are not comparable, as their contents are different. To ensure data quality, we use instances where the tweet topic classifier [87] was highly confident. The threshold of 0.8 was decided empirically. Extending our model to incorporate these additional metrics could provide a more holistic view of a post's potential reach and influence. Furthermore, although our model makes use of large language models (LLMs) like FLAN-UL2, ChatGPT, and Claude, their performance is largely dependent on the quality of data they were trained on. As these LLMs are primarily trained on English language data, their performance might not be as effective when analysing and generating

content in other languages. ChatGPT and Claude are non-peer-reviewed proprietary large language models. This prohibits us from commenting on the reasons that lead to difference in performance.

Moving forward, we plan to address these limitations by expanding our dataset to include diverse sources, incorporating additional engagement metrics, and considering other contextual factors that influence crowd reactions. Moreover, we plan to test our approach on various platforms and languages to increase the generalizability of our model.

Appendix

2.3.8.1 Hyperparameters

Hyperparameters of all the models we trained are mentioned in Table 2.9.

2.3.8.2 Prompts

Prompt engineering is an emerging field of research. We experimented with different prompts to understand what works for us. The exact prompts are as follows:
Type 1: Prompt for FLAN-UL2. Chat-GPT, and Claude:
text-1: < `t1` >
text-2: < `t2` >
Will text-1 receive more reactions than text-2? Answer me with "yes", "no", just one word.

Type 2: Prompt for FLAN-UL2
Why is the following text so engaging?
Text: < *tweet* >
The text is engaging because

Type 2: Prompt for Chat-GPT, and Claude
Why is the following text so engaging?
Text: < *tweet* >

Following guardrails, Chat-GPT did not respond to prompts of Type 1. Details about the LLMs which were used are mentioned in Table 2.10. We observed that for Type 2 prompts, the responses from FLAN-UL2 were always shorter and crisp compared to Claude and ChatGPT. The responses from the FLAN-UL2 model consisted of maximum 24 words and 9 words on average. The Claude model provided us with point-wise answers in most cases, and it often quoted parts from the tweet while explaining why the tweet was engaging. The responses from the Claude model (297 words) were generally about twice as long as that of the ChatGPT model (128 words).

Model Type	Hyperparameters
RoBERTa FLANG -RoBERTa FLANG-BERT	Epochs=15, Train batch size =8, weight_decay=0.01, learning_rate=2e-5
Cross Encoder	Epochs=20, Train batch size = 16

TABLE 2.9: Hyperparameters of the models trained. Wherever nothing is mentioned, we use the default parameters.

LLM	Details
FLAN-UL2	Model: 20B parameters, https://huggingface.co/google/flan-ul2
Claude	Model: claude-v1, max. tokens to sample=200, https://www.anthropic.com/index/introducing-claude
ChatGPT	Model: gpt-3.5-turbo [Based on variants of GPT models [73]], https://openai.com/blog/chatgpt

TABLE 2.10: Details about the LLMs used.

Paraphraser	Output
prithivida/parrot_paraphraser_on_T5 [89]	It's time to rebuild an American economy that works for all of our families and the next
stanford-oval/paraphraser-bart-large [90]	It is time to rebuild our economy to work for all our families and for the next generation
humarin/chatgpt_paraphraser_on_T5_base [72]	The time has come to rebuild an American economy that benefits all Americans, including our families and the next generation. It's time to ensure every American has an equal chance to succeed. We need to improve our economy back on track.

TABLE 2.11: Output of different paraphrasers for the following input text. **Input:** *It's time to rebuild an American economy that works for all of our families and the next generation. It's time to ensure every American enjoys an equal chance to get ahead. It's time to build our economy back better.*

2.3.8.3 Paraphrasers

We explored three most downloaded open source paraphrasers present in the Hugging Face repository https://huggingface.co/models (accessed on 3$_{rd}$ June, 2023) which have descriptions of the underlying model architectures. We qualitatively assessed open-source paraphrasers based on their ability to generate complete and sensible paraphrases. However, we did not conduct any quantitative assessment. More details about the paraphrasers are presented in Table 2.11. We finally selected the T5-based paraphraser [72] trained using responses from ChatGPT. We use the following hyperparameters: num_beams=5, num_beam_groups=5, num_return_sequences=5, repetition_penalty=10.0, diversity_penalty=3.0, no_repeat_ngram_size=2, temperature=0.7, max_length=128.

2.3.8.4 Other Experiments

We experimented with several model architectures like vanilla RoBERTa-base (Accuracy: 0.55, F1: 0.68), fine-tuned FLANG-RoBERTa embedding using SBERT architecture then

passing the embeddings through a feedforward neural network (Accuracy: 0.69, F1: 0.67), etc. However, none of these experiments improved the performance. We experimented with FLANG-RoBERTa as the "Business & Entrepreneurs" category dominates our dataset. FLANG-RoBERTa has been specifically fine-tuned for the finance domain. FLANG-RoBERTa outperformed RoBERTa in the "Business & Entrepreneurship" category. In the "Fitness & Health" category, RoBERTa out-performed FLANG-RoBERTa. Lastly, in the "Learning & Educational" category, performance of both the models is the same.

We collected a sample of 88 instances (pairs of tweets) with a similar number of retweet count (within 10%). We call it Bucket-0. For Bucket-0 the Accuracy is 56.82% and the F1 score is 56.82%. On the test set, for different buckets constructed using the difference in number of retweets (diff), the proposed model performs as follows:
Bucket-1 (10%<=diff<60%), Accuracy: 61.2%, F1: 62.9%
Bucket-2 (60%<=diff<141.3%), Accuracy: 68.4%, F1: 68.8%
Bucket-3 (141.3%<=diff<311.5%), Accuracy: 80.67%, F1: 81.52%
Bucket-4 (diff>=311.5%), Accuracy: 77.6%, F1: 79.0%

Lastly, we used Large Language Models (like ChatGPT, Claude) as they are well known for their reasoning skills and Pre-trained Languages Models (like BERT, RoBERTa) as they are the state of the art for discriminative tasks. Our motivation was to understand if complementing the discriminative models with inputs from large language models would help. This is like leveraging the best of both worlds. The response from LLMs improved the quality of the features. Thus, the GGEA architecture outperformed others.

2.3.8.5 Example

In Figure 2.11, we present how the GGEA framework along with the paraphraser during scoring looks like. Here is an example of how GGEA framework works.

Input: *This year, our economy is projected to grow at the fastest pace in nearly 40 years. Right now, we have the opportunity to make once-in-a-generation investments in the foundations of middle class prosperity. Read more about the American Jobs Plan:*

Paraphrases generated:
1) *We are expecting the fastest growth in our economy for almost four decades this year. At present, we have the chance to invest in bolstering middle-class prosperity. Learn more about the American Jobs Plan.*
2) *The economy is expected to expand at a rate faster than it has been for almost 40 years this year. We have the opportunity to invest in the middle class's success once again. Learn more about the American Jobs Plan.*
3) *We are expecting the fastest growth in our economy for almost four decades this year. At present, we have the chance to invest in bolstering middle-class prosperity. Learn more about the American Jobs Plan: why it matters?*
4) *The economy is expected to expand at a rate faster than it has been for almost 40 years this year. We have the opportunity to invest in the middle class's success once again. Learn more about the American Jobs Plan.*
5) *Our economy is poised to grow at its fastest rate in nearly 40 years this year, making it a prime opportunity for us all to invest in the middle class's success. Learn more about*

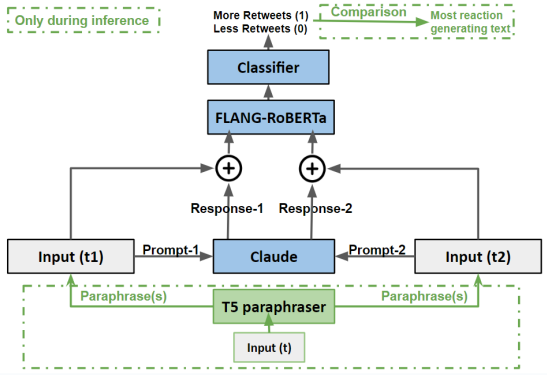

FIGURE 2.11: GGEA along with paraphraser during scoring/inference.

the American Jobs Plan:

Output:
The paraphrase which is supposed to get maximum retweets as per GGEA is:
The economy is expected to expand at a rate faster than it has been for almost 40 years this year. We have the opportunity to invest in the middle class's success once again. Learn more about the American Jobs Plan..

Chapter 3

Improved Investing

In this chapter, we discuss how Natural Language Processing can be leveraged for improving the investment journey.

3.1 Research Questions

- **RQ-3**: How to improve the investment process?
    - **Relevant Contributions**: Detecting Hypernyms of Financial Terms [7] [8], Extracting relationship between financial entities [9].

3.2 Detecting Hypernyms of Financial Terms

3.2.1 Introduction

Investors read online content (like financial reports of organizations, news) to make decisions. These contents often contain jargon unknown to the readers. The readability of these contents can be improved significantly by presenting readers with hypernyms (i.e. broad categories) corresponding to any jargon. A jargon term, being a subset, holds an "IS A" relationship with its hypernym. For example, "alternative debenture" (unknown financial term/jargon) is a kind of "bond" (hypernym). The same holds true for terms like "Bearer Bonds", "Callable Bonds" and "CoCo Bonds". This is shown in Figure 3.1. The Natural Language Processing (NLP) community has been working on methods to automatically discover hypernyms for more than a decade. Recently with the advent of shared tasks like FinSim [92] extracting hypernyms specific to the financial domain has caught the attention of this community. Inspired by the advances and contributions made by the participants in FinSim-1 [92] and FinSim-2 [93], we participated in the third edition of FinSim [94]. It comprised matching financial terms to their hypernyms. Compared to the previous two editions, the third edition consisted of larger and more diverse topics related to finance. In this chapter, we present an extension of the solutions our team LIPI developed while participating in FinSim-3 as well as the enhancements we carried out later.

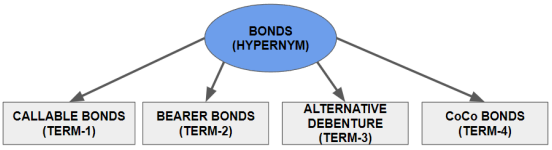

FIGURE 3.1: Terms to Hypernym relation.

The research questions we try to answer in this study are as follows.

- **RQ1:** How have the datasets and solution architectures of the FinSim challenges evolved over the years?
- **RQ2:** How to develop a system for ranking a set of hypernyms for a given financial term?
- **RQ3:** Does using domain specific embeddings improve model performance?
- **RQ4:** What is the impact of augmenting/adding data from other sources?

Our contributions in the work contained in this article are as follows.

- We review and summarize various approaches used by participants of all three editions of FinSim [92–94]. We further collate the performances of such approaches in Table 3.2.
- We explore various external financial data sources to supplement the training set.
- We propose a novel way of augmenting the training set for incorporating hierarchies that are present in the set of hypernyms.

- We develop a system capable of ranking a set of hypernyms for a given financial term.

The data set used in this chapter can be obtained from here[1]. The metadata is presented in the paper [94]. Our code base is available here [2].

This chapter is organized as follows. Section 3.2.1 introduces readers to our motivation. Section 3.2.2 briefly describes the previous works on this task. We formally define the problem statement in Section 3.2.3 and discuss the dataset used for this work in Section 3.2.4. Next, we describe our methodology, experiments and results in Section 3.2.5, 3.2.6 and 3.2.7 respectively. Section 3.2.8 concludes the chapter and Section 3.2.9 provides avenues for future work.

3.2.2 Research Landscape

In this section, we discuss the previous works in three phases. Firstly, we explore how the problem of hypernym identification have been solved in the field of computational linguistics in general. Following this, we elaborate on its applications specific to the Financial Domain. Finally, we state how our work differs from the existing work in the literature.

3.2.2.1 Hypernym Identification in NLP Literature

The task of Hypernym detection started gaining the interest of the NLP community in the early 1990s. During this time Hearst [95] did the pioneering work of automatically extracting hypernyms using lexico-syntactic patterns like "such as" followed and preceded by Noun Phrase and so on. Another pattern-based approach had been applied by Snow et al. [96]. They narrated how they extracted "dependency paths" from parse trees of sentences containing hypernyms and hyponyms using WordNet [97]. They additionally used coordinate terms i.e. terms having at least one common parent to enhance the process of hypernym identification. Sang [98] assumed that the Web contained much additional training data than any of the text corpora and developed a simple pattern-based method to extract hypernyms from the web. Furthermore, Sang et al. [99] compared two major approaches of hypernym extraction which are based on lexical (dictionary-based) and dependency patterns. Ritter et al. [100] described how they used lexico-syntactic patterns and Hidden Markov Models to identify hypernyms of noun phrases.

Caraballo [101] presented an automatic method of building a hierarchy of nouns and their hypernyms using WordNet [97]. Bottom-up clustering had been used to create the hierarchy and hypernyms had been assigned after creating a binary tree. Shinzato et al. [102] proposed a novel method of extracting hypernyms from web pages using structures of the HTML pages and other statistical features. Navigli et al. [103] introduced a novel concept of Word-Class Lattices which were learned from definitions present in Wikipedia.

[1] https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp2021/shared-task-finsim (accessed on 18th September, 2023)

[2] https://github.com/sohomghosh/FinSim_Financial_Hypernym_detection (accessed on 18th September, 2023)

They further released a Java-based tool [104] to extract hypernyms of a term and its definitions.

Recently, the use of Deep Learning Models in Computational Linguistics has gathered the interest of the NLP community. Tan et al. [105] used bi-directional Recurrent Neural Networks to extract hypernyms from definitions using Parts of Speech of constituent words. They validated this model's performance on Wikipedia as well as Stack-Overflow datasets. Liang et al. [106] studied if the property of transitivity holds in lexical taxonomies which were built automatically. They developed a supervised approach to do so. Furthermore, they used transitivity to extract new hypernym-hyponym relations.

3.2.2.2 SemEval Shared Tasks on Hypernym Detection

Problems relating to hypernym detection were provided in several editions of SemEval [107–110].

SemEval-2015 Task 17: "Taxonomy Extraction Evaluation (TExEval)" [107] dealt with extraction of hypernym-hyponym relations from texts and taxonomy construction for four different domains namely: chemicals, equipment, foods and science. Grefenstette [111] developed the best performing model using simple structure-based features like whether a term is present in a sentence and document, term and document frequencies and presence of sub-sequences.

SemEval-2016 Task 13: "Taxonomy Extraction Evaluation (TExEval-2)" [108] was the multilingual edition of TExEval [107]. It comprised corpora from several domains like environment, food and science. Different languages included English, Dutch, Italian and French. Team Taxi [112] won both the shared tasks. They used Hearest pattern and sub-string based features.

SemEval 2017 Task 10: "ScienceIE - Extracting Keyphrases and Relations from Scientific Publications" [109] dealt with extraction of important phrases (like Process, Task and Material) and relations (like hypernyms / synonyms). It was restricted to the scientific domain. Team MIT [113] achieved the first rank by creating a system using a convolutional neural network. This system used an embedding comprising relative positions, type of entity and parts of speech as input.

SemEval-2018 Task 9: "Hypernym Discovery" was introduced [110] in the year 2018. This shared task was about extracting hypernyms from corpora in three languages (English, Spanish and Italian) and two domains within English (Medical and Music). The best performing model was presented by Team CRIM [114]. This model was an ensemble of word embedding based supervised approach with a pattern based unsupervised approach.

Dash et al. [115] introduced a new neural network-based architecture, Strict Partial Order Networks (SPON) to detect hypernyms. They benchmarked it using SemEval 2018 general and domain specific hypernym discovery tasks. Very recently Bai et al. [116] proposed the use of sequential recurrent mapping models to preserve the hierarchy between terms and their hypernyms. They also performed an extensive evaluation on SemEval-2018 Task 9 datasets.

Year	Edition	Conference	#Pps	#Train	#Test	#L	#T	Acc.	MR.
2020	FinSim-1 [92]	IJCAI-PRICAI	156	100	99	8	6	0.858	1.21
2021	FinSim-2 [93]	ACM-WWW	203	614	211	10	7	0.906	1.189
2021	FinSim-3 [94]	IJCAI	211	1,050	326	17	5	0.941	1.113

TABLE 3.1: Background. #Pps is number of Prospectuses. #L, #T, Acc. and MR. denote number of Labels, Teams, Best Accuracy and Mean Rank respectively.

3.2.2.3 FinSim Shared Tasks - Hypernym Detection in Financial Texts

As mentioned earlier, the third edition of FinSim challenge [94] is the most recent one. Details relating to all editions of FinSim are mentioned in Table 3.1. These shared tasks have been organized by Fortia Financial Solutions[3]. Teams IITK [117], PolyU-CBS [118] and MXX [119] won the first, second and third editions of FinSim respectively. We narrate more details relating to the dataset of FinSim-3 in the next Section 3.2.4. Team MXX [119] used an LSTM [120] based approach over word2vec [121] embeddings to win the FinSim-3 challenge (Accuracy = 0.941, Mean Rank = 1.113). The evaluation metrics and the other aspects of the problem statement remained the same for all three editions. We organize the system descriptions of the participating teams and present them in Table 3.2. The winning entries have been highlighted in bold. Studying this table thoroughly, we observe that the Word2Vec approach remained the same for all of them. Only one of these teams MXX [119] augmented the given dataset with external data. Similarly, only one of the winning teams, PolyU-CBS used syntactic based features like Jaccard similarity. Logistic Regression emerged as the most preferred classifier. Moreover, it is interesting to note that every successive year performances of the submitted models improved significantly. Since only three teams ([122], [123], and [124]) used Knowledge Graphs, we conclude it is yet to become popular. Some of the BERT-based models like FinBERT [69], Sentence BERT [68] and RoBERTa [125] were also explored by most participants.

In recent times, Loukas [126] released the EDGAR-CORPUS comprising annual reports of listed US organizations from the year 1993 to 2020. They created word2vec [121] embeddings based on this corpus and evaluated it on the FinSim-3 dataset. They achieved an accuracy of 0.879 and a mean average rank of 1.21 using stratified 10-fold cross-validation.

3.2.2.4 Difference with Prior Works

Our work is novel in terms of the approach we used to create negative samples from the existing training dataset using the hierarchy present within the hypernyms. Unlike most others, we did not train a classifier to solve the problem of detecting hypernyms. On the other hand, we detect hypernyms by performing semantic search over fine-tuned embeddings. This makes the approach generic and robust to adding more hypernyms to the existing set.

[3] https://www.fortia.fr/ (accessed on 18th September, 2023)

3.2.3 Problem Statement

In this section, we shall narrate the problem statement and discuss the evaluation metrics.

Given a set of n financial terms (t_1, t_2, t_3, ... t_n) and their corresponding hypernyms/labels (l_1, l_2, l_3, ... l_n) where $l_i \epsilon$ {Equity Index, Regulatory Agency, Credit Index, Central Securities Depository, Debt pricing and yields, Bonds, Swap, Stock Corporation, Option, Funds, Future, Credit Events, MMIs, Stocks, Parametric schedules, Forward, Securities restrictions}. Our task is to develop a system capable of ranking all these hypernyms in order of decreasing semantic similarity for any unknown financial term.

The evaluation metrics used here are as follows:
$Accuracy = \frac{1}{n} * \sum_{i=1}^{n} I(y_i = \hat{y}_i[1])$,
$MeanRank = \frac{1}{n} * \sum_{i=1}^{n} (\hat{y}_i.index(y_i))$,
where $\hat{y}_i$ is the ranked list (with the index starting from 1) of predicted labels corresponding to the expected label y_i. I is an identity matrix. Interestingly, the organizers considered only the first three elements of the ranked list for evaluation. If any label was not present within these three elements, it was assigned rank 4.

3.2.4 Dataset

In this section, we narrate the datasets we used to perform our experiments. In addition to the data, which was provided to us by the organizing team, we explored other external datasets as well. These include Financial Industry Business Ontology (FIBO) [4], DBpedia [127], Investopedia[5], etc.

3.2.4.1 Data Description

The organizers provided us with 211 prospectuses of different companies in Portable Document Format (PDF). Furthermore, a tagged dataset comprising 1,050 financial terms and their corresponding hypernyms/labels were also provided. Out of 1,050 terms, 1,040 were distinct. We refer to this as the training set. Three of these 1,040 terms were ambiguous as they were assigned 2 different labels each. Terms with lengths of six or fewer constituted 91% of the training set. A few instances of such terms are: 'Floating Rate Note'", "Perpetual bond", etc. The number of distinct labels was 17. Their distribution is shown in Figure 3.2 and presented in Table 3.3. It is interesting to note that a hierarchy was present among these 17 labels as all of them belonged to FIBO. This hierarchy is presented in Figure 3.3. The root nodes and leaf nodes have been highlighted in yellow and grey respectively. The first child nodes have been marked in bold. Moreover, we received 326 unlabeled financial terms which constituted the test set. The hypernym "Swap" shares the same parents with "Option" whereas it does not have any relation with other hypernyms like "Future" or "Bonds".

[4] https://spec.edmcouncil.org/fibo/ (accessed on 18th September, 2023)
[5] https://www.investopedia.com/ (accessed on 18th September, 2023)

3.2.4.2 Data Augmentation

Since 91% of the training set had financial terms having only six or fewer words, we explored various ways of augmenting the dataset. Similar approach was also followed by [128] and [129] while participating in FinSim-2 and FinSim-1 respectively. This was done in three phases. Let us understand each one of them.

3.2.4.3 Acronym Expansion

Several Financial Terms were present along with their acronyms. This led to inconsistency in the training set. Keswani et al. [117] also highlighted this issue. To deal with this, we executed spaCy's[6] inbuilt acronym detector on all prospectuses. We manually investigated the outputs (i.e., a list of acronyms and their corresponding synonyms). We concluded that not all outputs were usable. We developed the following heuristics to further refine this list. We dropped records having

- expansions with number of characters lesser than that of the acronyms
- expansions with parenthesis/bracket symbols i.e., "(" or ")"
- expansions with number of characters lesser than or equal to five
- acronyms that were a valid English words including proper nouns like "bond", "England", "Germany", and so on.

The cleaned list contained 635 acronyms and their expansions. We used this cleaned list to augment our training set by replacing acronyms with their full forms wherever identified in the prospectuses.

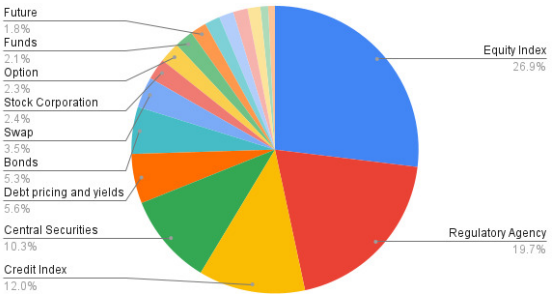

FIGURE 3.2: Distribution of labels in original training set.

[6] https://spacy.io/ (accessed on 18th September, 2023)

TABLE 3.2: Related Works - FinSim. USE: Universal Sentence Encoder, RF: Random Forest, LR: Logistic Regression, LSTM: Long Short Term Memory; NB: Naive Bayes, NN: Neural Networks, DA: Deep Attention, KG: Knowledge Graphs, SVM: Support Vector Machine, Inv: Investopedia, Ext: External. FS: FinSim

				Approach of best performing model				
Task	**Team**	**Acc.**	**MR**	**Syntactic Features**	**Classifier**	**Embeddings**	**KG**	**Ext. Data**
FS-1	Anuj [129]	0.858	1.42	Character count, Word Count etc.	SVM			Inv
FS-1	ProsperaMnet [130]	0.777	1.34			Sparse embeddings		
FS-1	FINSIM20 [131]	0.787	1.43			USE [132]		
FS-1	**IIT-K** [117]	**0.858**	**1.21**		NB	Word2Vec, BERT		
FS-2	AIAI [133]	0.877	1.278		DA	Word2Vec [121]		
FS-2	FinMatcher [122]	0.811	1.415	Word overlap	NN		RDF2vec	WordNet, Wikidata, WebIsALOD
FS-2	GOAT [128]	0.896	1.193		LR	FinBERT [69]		Inv
FS-2	L3i-LBPAM [134]	0.858	1.325			SentenceBERT [68]		
FS-2	TCSWTIM2021 [135]	0.858	1.274	Sentence extraction		TF-IDF, BERT [1]		
FS-2	JSI [123]	0.811	1.316		RF	Word2vec	FIBO	
FS-2	**PolyU-CBS** [118]	**0.906**	**1.189**	Jaccard similarity, if hypernym in term	LR	Word2Vec, BERT		
FS-3	DICoE [136]	0.904	1.162	Levenshtein distance, Upper case to lower case characters ratio	LR	Word2vec		Inv, FIBO
FS-3	MiniTrue [124]	0.865	1.315			BERT, FinBERT	RotatE	
FS-3	Lipi [7] *(Our old model)*	0.917	1.156	Acronym expansion		SentenceFinBERT		DBPedia, Inv, FIBO
FS-3	Yseop [137]	0.917	1.141		LR	FastText, SentenceRoBERTa		FIBO

Table 3.2 continued from previous page

					Approach of best performing model			
FS-3	**MXX** [119]	**0.941**	**1.113**		LSTM	Word2Vec		Inv,FIBO, NYSE, BIS

Label	Count
Equity Index	280
Regulatory Agency	205
Credit Index	125
Central Securities Depository	107
Debt pricing and yields	58
Bonds	55
Swap	36
Stock Corporation	25
Option	24
Funds	22
Future	19
Credit Events	18
MMIs	17
Stocks	17
Parametric schedules	15
Forward	9
Securities restrictions	8
Total	**1040**

TABLE 3.3: Distribution of labels in the original training set.

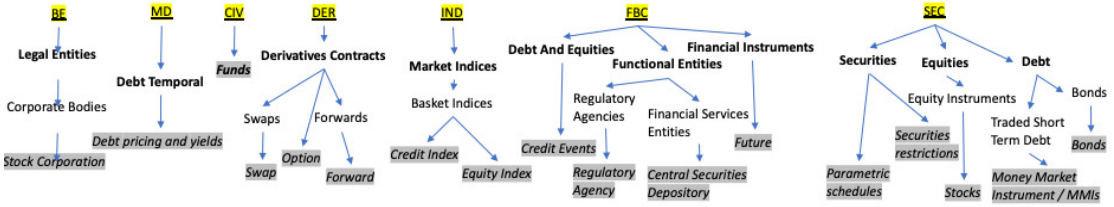

FIGURE 3.3: Hierarchy of labels as obtained from FIBO. Root nodes have been underlined and highlighted in yellow. First child nodes have been marked in bold. Leaf nodes have been italicised and highlighted in grey color. BE = Business Entities, MD = Market Data, CIV = Collective Investment Vehicle, DER = Derivatives, IND = Indices and Indicators, FBC = Financial Business and Commerce, SEC = Securities

3.2.4.4 Augmenting definitions from DBpedia

DBpedia[7] provides search Application Programming Interfaces (API)[8] which help in extracting structured information and relationships from Wikipedia[9]. Kilger [138] introduced The Linked Hypernyms Dataset which provided more specific details than DBpedia. We systematically searched DBpedia to obtain definitions of financial terms present in the training and test sets. These definitions added more context to the original terms. We present the results of invoking the search API for the term, "callable bond" in Figure 3.4. Inspecting some of these sample outputs, manually, we concluded that matching the given financial terms with the content of the "Label" tag present in the output payloads and

[7] https://www.dbpedia.org/ (accessed on 18th September, 2023)
[8] https://lookup.dbpedia.org/api/search(accessed on 18th September, 2023)
[9] https://en.wikipedia.org/(accessed on 18th September, 2023)

extract the contents of the "Description" tag. To achieve this, we pre-processed the given financial terms and the contents of the "Label" tag obtained by calling the search API for each of the terms. The pre-processing steps included conversion to lower case, punctuation and repetitive white space replacement and singularization. Furthermore, we calculated the token overlap ratio between these cleaned terms and the contents of the "Label" tag using these formulas:
$Ratio1 = length(s_1 \cap s_2)/length(s_1)$,
$Ratio2 = length(s_2)/length(s_1)$
where s_1 and s_2 represent sets of tokenized cleaned financial terms and tokenized cleaned contents of the "Label" tag respectively. After experimenting with several values, we empirically determined that $Ratio1 = 1$ and $Ratio2 <= 1.25$. This criterion enabled extraction of descriptions of the matching terms from DBpedia.

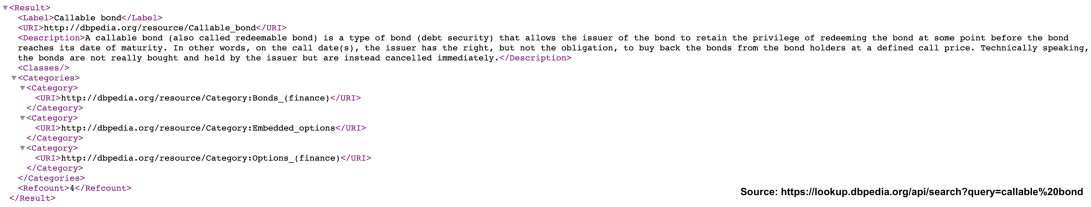

FIGURE 3.4: Result obtained by calling DBPedia Search API for the term "callable bond".

3.2.4.5 Augmenting definitions from Investopedia and FIBO

While participating in FinSim-1, Saini [129] used definitions of financial terms from Investopedia[10]. Inspired by this approach, we crawled all these definitions from Investopedia. A total of 6,261 definitions were obtained. Moreover, we obtained a glossary of 11,827 financial terms and their explanations from FIBO. We cleaned these definitions using the approach mentioned previously.

These data augmentation steps increased the size of the training set to 1,836 records, and the size of the test set to 607 records. For the financial term "callable bond" we present the result of data augmentation in Table 3.4. Table 3.5 presents the number of matches we get from different sources of data such as DBPedia, Investopedia, and so on.

3.2.4.6 Adding Data from Various External Sources

Inspired by [119], we extracted 31,748 financial terms from various other websites such as

- Bank for International Settlements [11] (for label "Regulatory Agency")
- ETF Database [12] (for label "Equity Index")

[10] https://www.investopedia.com/financial-term-dictionary-4769738 (accessed on 18th September, 2023)
[11] https://www.bis.org/regauth.htm
[12] https://etfdb.com/indexes/equity/

- Wikipedia [13] , and Wiley [14] (for label "Credit Index")
- Kaggle [15] (for label "Funds")
- ADVFN [16] & datahub[17] (for label "Stock Corporation")
- National Securities Depository Limited [18] and European Central Securities Depositories Association [19] (for "Central Securities")

We added these terms to our training set for some of the experiments we performed. Later, we discarded them as it did not result in any improvement in the model performance. This is probably because most of these terms are proper nouns as they represent names of funds, organizations, and so on.

Expanded Term/Term Definition	**Label**	**Source**
Term: Callable bond	Bonds	original, acronym expansion
Definition: Bond that includes a stipulation allowing the issuer the right to repurchase and retire the bond at the call price after the call protection period	Bonds	FIBO
Definition: A callable bond (also called redeemable bond) is a type of bond (debt security) that allows the issuer of the bond to retain the privilege of redeeming the bond at some point before the bond reaches its date of maturity.	Bonds	DBpedia

TABLE 3.4: Result obtained by data augmentation for the term "callable bond".

[13] https://en.m.wikipedia.org/wiki/Credit_default_swap_index
[14] https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119208631.app1
[15] https://www.kaggle.com/stefanoleone992/mutual-funds-and-etfs/version/3?select=MutualFunds.csv
[16] http://www.advfn.com/
[17] https://datahub.io/core/nyse-other-listings
[18] https://nsdl.co.in/related/wrld.php
[19] https://ecsda.eu/members-2/list-of-members

Data Source	**Count**
Original modelling data	1040
Acronym expansion	218
DBpedia	257
Investopedia	85
FIBO	236

TABLE 3.5: Number of matches obtained from various data sources.

	Original		Extended	
label	**# dev**	**# val**	**# dev**	**# val**
Equity Index	225	57	373	84
Regulatory Agency	159	46	260	78
Credit Index	103	21	123	27
Central Securities Depository	83	24	106	28
Bonds	49	6	110	14
Debt pricing and yields	41	17	84	34
Swap	31	5	57	9
Option	21	3	35	4
Stock Corporation	18	6	54	15
Funds	17	5	36	10
Future	16	3	29	7
Credit Events	15	3	35	6
Parametric schedules	14	1	45	3
MMIs	14	3	29	9
Stocks	12	5	23	11
Securities restrictions	7	1	28	3
Forward	6	3	13	3
TOTAL	**831**	**209**	**1440**	**345**

TABLE 3.6: Label distribution for the development and validation set before and after data augmentation.

3.2.4.7 Development, Validation and Test splits

As mentioned previously, we were provided with 1040 distinct manually tagged financial terms for training our model and 326 un-tagged instances for testing. We split the set of 1040 terms into two buckets: a development set having 831 terms (80%) and a validation set having 209 terms (20%). We performed the same operation for the augmented set having 1,836 financial terms out of which 1,785 were distinct. This resulted in a set of 1,440 distinct terms for training & validation and a set of 345 distinct terms for testing. The final output i.e., predicted ranks of the given 17 labels on the test set was to be submitted for the initial set of 326 un-tagged instances. Thus, for the augmented test set, we calculated the mean cosine similarity with each of the labels for multiple occurrences of a term. We ranked the labels based on these similarities.

The distribution of labels before ("original") and after data augmentation ("extended") is shown in Table 3.6.

3.2.5 Methodology

Our best performing model is an ensemble of two models. Each of these models has been developed in three steps.

1. negative sample creation (Reference: Algorithm 1)
2. using sentence transformers to fine-tune embeddings having 768 dimensions

3. calculating cosine similarities between terms and hypernyms.

This has been depicted in Figure 3.5. Steps 1 and 3 are common for both models. In the second step, we use FinBERT [69] embeddings for the first model and FinISH [137] embeddings for the second model.

STEP-1: In the first step, we create negative samples from the existing training set having sets of terms 'T', labels 'L', term definitions 'TT' and label definitions 'LL'. The definitions of labels and terms are obtained through data augmentation. For instances where we are not able to augment anything to a given financial term, we keep the term definition the same as the term. For each term 't' having definition 'td', its corresponding label 'l' and label definition 'ld', present in the training set we first assign a similarity score of 1.0 to the ('td', 'ld') pair. After that, we extract root node 'ln' and first child node 'lc' of 'l'. We then randomly select 10 labels and their corresponding definitions from 'L' such that none of the selected labels and their corresponding terms is the same as 'l' and 't'. For each such label 'la' and label definition 'lnd', we assign similarity scores corresponding to each of the ('td', 'lnd') pairs. This similarity score is assigned a value based on the following conditions
i) value = 2.0*k when the first child of 'la' i.e. 'lac' is the same as 'lc'
ii) value = 1.0*k when only the root node of 'la' i.e. 'lan' is same as 'ln' and its first child 'lac' is different from 'lc'
iii) value = 0.0*k when former two conditions are not met i.e. they have no ancestors in common
We present this formally in Algorithm 1. We empirically determined that keeping the value of parameter k as 0.4 gives the best result. This resulted in 63,360 instances in total out of which 49,836 had a similarity score of 0.0. We sub-sampled the instances with similarity score of 0.0. The final distribution consists of 5,760 instances with a 1.0 similarity score, 5304 instances with 0.8, 2460 with 0.4 and 550 with a similarity score of 0.0. This step is shared between both models described above.

A machine learning based classification model performs better when it is provided with more data from different classes. This motivated us to create negative samples. For example, as "Bonds" is the hypernym of "Alternate Debenture", we can safely assume that "Alternate Debenture" when paired with terms having hypernyms other than "Bonds" will constitute negative instances,

STEP-2: In the second step, for the first model we fine-tune FinBERT [69] embeddings using sentence transformer [68] architecture. For the second model, we further fine-tune the FinISH embeddings released by Yseop Labs [137]. They created this embedding by fine-tuning RoBERTa [125] on the FIBO corpus. Our objective was to minimize the multiple negative ranking loss and online contrastive loss. Multiple negative ranking loss [139] is applied only on samples which are similar to each other. This makes the embedding suitable for retrieval tasks. Online contrastive loss selects the hard cases in a batch based on the distance of separation and computes the loss only for these specific hard cases only. It tends to keep similar texts near to each other and pushes dissimilar texts away from each other in the vector space. We kept the margin parameter at 0.5. A batch size of 20, when executed for 25 epochs, gave the best result for the first model. For the second

model, a batch of 30 when executed for 45 epochs gave the best result. The sample code is available here.[20].

STEP-3: In the third step, we convert definitions of all the 17 labels/hypernyms and terms present in the validation and test set into vectors. We use the fine-tuned embeddings generated in the previous step for the same. We further calculate cosine similarity between the vectors of each of these terms with that of all the 17 hypernyms. Since we have augmented the dataset, we must aggregate this data such that we have only one record for every term. We use the mean of cosine similarities to achieve this. We do the same for the other model as well. This results in two cosine similarities for each of the terms, one obtained from the first model while the other from the second.

To ensemble, we again take the mean of the two cosine similarities we calculated for each of the terms across all the hypernyms. Finally, we rank the hypernyms in terms of decreasing order of the mean cosine similarity.

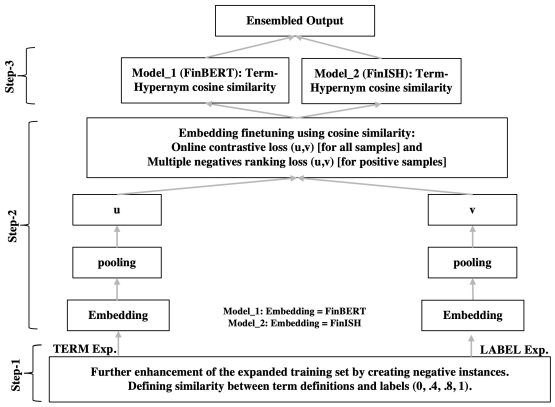

FIGURE 3.5: Methodology

3.2.6 Experimentation

In this section, we shall narrate various experiments we performed systematically to arrive at final model described in the previous section. We began by evaluating the baseline models provided.

3.2.6.1 Baselines

Let's understand the baseline solutions provided by the organizers. Kang et al. [94] trained a custom word2vec [121] model having 300 dimensions on text corpus extracted from the prospectus.

Baseline-1: In the first system, they calculate distances between terms and hypernyms based on the custom word2vec embeddings. They rank the hypernyms in increasing order of distance.

[20] https://www.sbert.net/examples/training/quora_duplicate_questions/README.html
(accessed on October 2021)

Algorithm 1 Algorithm to generate negative samples from existing training set

Require: $T > 0$ and $L > 0$ ▷ T is the augmented set of financial terms and L consists of corresponding labels i.e., hypernyms. $TT > 0$ and $LL > 0$ are the set of definitions of terms and labels respectively obtained after performing data augmentation

Require: Function FR(n) and Function FC(n) ▷ Function FR and FC returns the root node and first child node corresponding to node n respectively where n is one of the 17 labels i.e., leaf nodes/hypernyms

Ensure: $length(T) = length(TT) = length(L) = length(LL)$

$NT \leftarrow \{\}$ ▷ NT is the new set of definitions of financial terms to be created by appending negative samples

$NL \leftarrow \{\}$ ▷ NL is the new set of definitions of labels corresponding to terms in NT

$NS \leftarrow \{\}$ ▷ NS is the set of assigned similarity scores between the newly selected definitions of terms and labels in NT & NL respectively

$k \leftarrow 0.0$ ▷ 'k' is a hyperparameter. Keeping k = 0.0 gives the best result

for each term t ∈ T, term definition td ∈ TT, corresponding label l ∈ L and label definition ld ∈ LL **do**

    $NT \leftarrow NT \cup \{td\}$

    $NL \leftarrow NL \cup \{ld\}$

    $NS \leftarrow NS \cup \{1.0\}$ ▷ Assign a similarity score of 1.0 as the term and the label definition belong to the original set

    $ln \leftarrow FR(l)$ ▷ Extract root node of label 'l'

    $lc \leftarrow FC(l)$ ▷ Extract first child node of label 'l'

    R, RR ε_r L, LL where length(R)=10, length(RR)=10 ▷ Randomly select 10 labels from 'L 'and corresponding label definitions from 'LL' ensuring none of the labels are 'l' and none of their corresponding terms is 't'. This is done for creating the negative set

    for each label la ∈ R and corresponding definition lnd ∈ RR **do**

        $NT \leftarrow NT \cup \{td\}$

        $NL \leftarrow NL \cup \{lnd\}$

        $lan \leftarrow FR(la)$ ▷ Extract root node of label 'la'

        $lac \leftarrow FC(la)$ ▷ Extract first child node of label 'la'

        if $lac = lc$ **then** ▷ Check if first child nodes are the same. This implies root nodes are also the same.

            $NS \leftarrow NS \cup \{2 * k\}$

        else if $lan = ln$ **then** ▷ Check if root child nodes are same when first child nodes are different

            $NS \leftarrow NS \cup \{1 * k\}$

        else ▷ When first child nodes and root nodes are different

            $NS \leftarrow NS \cup \{0 * k\}$

        end if

    end for

end for

return NT, NL, NS

Baseline-2: The second system consists of a logistic regression-based classifier trained using custom word2vec embeddings of the financial terms as independent variables and hypernyms as the dependent variables.

3.2.6.2 Experiments

First, we removed the duplicate observations that we observed in the original dataset. We reserved 20% of the data for the unbiased validation set and the remaining 80% was used for training the models. We identified sources like DBpedia, FIBO and Investopedia which contain the definitions of many terms present in the input set. We also extracted the acronym definitions from the prospectus corpus shared by the organizers. All these sources helped us to augment the training data. The augmented dataset consisted of the original records plus augmented records where input terms were replaced with definitions and expansions. The number of instances in the original and the augmented training set was 832 and 1470. Similarly, the number of instances in the original and the augmented validation set was 208 and 366. This indicates that we were not able to get a definition or expansion for each of the terms.

We began experimentation by creating Term Frequency Inverse Document Frequency (TF-IDF) matrix, Topic Models and creating a machine learning based classifier over it. Since the performance was not appealing, we fine-tuned one of the state-of-the-art pre-trained models known as BERT [1]. We used sub-word tokenization, and followed the standard classification architecture to fine-tune the models. We took the representation from [CLS] token and passed it to the feedforward layers. The final layer of the network had 17 nodes with softmax activation. These 17 nodes provided predictions for the 17 labels mentioned previously. We did not freeze the base model parameters while training. This enabled fine-tuning of the base model for the task at hand, resulting in better performance. During the training, the error was propagated back through the transformer network. Based on the distribution of the tokenized output length, we determined that keeping the maximum input sequence length at 32. We conducted extensive hyperparameter tuning and identified that a combination of Adam optimizer with a learning rate of 0.00002 and 64 batch size yielded the best result. We trained the model for 40 epochs with an early stopping criterion based on performance on validation set. It performed the best at the 18^{th} epoch. We ordered the hypernyms in decreasing order of predicted probabilities. This performance was much better compared to the baselines.

We subsequently experimented the same fine-tuned BERT model in the augmented dataset which included the definitions from various sources mentioned previously. These definitions comprised well-formed sentences and they comprised longer sequences of input terms. We repeated the experiments described previously after increasing the input maximum input sequence length to 256. This input length was decided based on the distribution of the number of tokens that were present in the term definitions following the augmentation step. We trained it until epoch 40, and found that its performance on the validation set was the best at 17^{th} epoch. We observed that this performance was significantly better than that of the models developed without data augmentation. This led us to conclude that the data augmentation steps we employed proved beneficial. We subsequently tested addition of data from various other sources as noted in Section 3.2.4.6. However, this did not yield any additional improvement in the performance of the model. This is probably because most of these terms were predominantly proper nouns and organization like entities.

We subsequently evaluated various alternative transformer-based models present in the Hugging Face [140] model repository. This included RoBERTa [125], FinBERT [69], FinEAS [141], and so on. We observed that FinBERT when fine-tuned using the expanded dataset yielded further improvements. Subsequently, we trained a new transformer-based model. Its objective was to predict two things simultaneously i) root node ii) hypernyms. This did not achieve promising results. We also attempted to fine-tune these models using a Masked Language Model based approach on the corpus of the prospectuses. Due to resource constraints, we were unable to train it beyond a few epochs. Its performance was not promising as well.

After extensively studying the failure cases and given the hierarchy of the labels we determined to evaluate a novel framework to generate negative instances and fine-tune it, using the sentence transformer [68] architecture. This has been elaborated in detail in Section 3.2.5. For creating the negative set mentioned in Algorithm: 1, we experimented with different sampling strategies and with various values of 'k'. The performance of the model improved when we employed the Sentence Transformer architecture with FinBERT as the backbone. It improved further on switching the underlying embedding model from FinBERT to FinISH. FinISH was developed and released by Yseop Labs[21] while participating in FinSim-3 [137]. We ran it for 45 epochs with a batch size of 30. It took approximately 1 hour 43 minutes to train.

Finally, we created an ensemble of the best performing models. We observed that an ensemble of the last two models which were trained using sentence transformers architecture with negative samples resulted in the best performance on the validation set. All the hyperparameters were empirically selected by monitoring model performance on the validation set.

3.2.7 Results and Discussions

In this section, we shall discuss the results presented in Table 3.7. We restrict our evaluation to just one dataset due to the unavailability of any alternative dataset suitable for financial hypernym detection.Models with serial numbers (SLN) 1 to 15 were developed during the FinSim-3 challenge while those with SLN 16 to 20 were developed later. After the event, the organizers declared the results for each submission of the participating teams. The number of submissions was restricted to 3. Thus, we present test set results for three of our models (SLN: 5, 6, 7). By comparing this with the test set results of other participants (SLN: 8 to 15), we observe that our previous model SFinBERT_neg (SLN: 7) [7] ranked third, and was marginally behind the model ranked second (SLN: 15) [137]. This model was developed by fine-tuning FinBERT [69] with negative samples using Sentence Transformer architecture. We attempted to contact the organisers to evaluate our new model (SLN: 20) on the test set as well. However, the test set has not yet been publicly released. Thus, we present our results on the held-out validation set.

It is interesting to observe that when using transformer-based pre-trained BERT embeddings (SLN: 3, 4), the model performs better than the baselines (SLN: 1, 2). This demonstrates the effectiveness of transformer-based embeddings such as BERT [1] compared to traditional embeddings like word2vec [121]. It likely occurred because transformer-based embeddings as they have been pre-trained on large datasets can capture greater linguistic

[21] https://yseop.com/ (accessed on 18th September, 2023)

complexity. Comparing the performance of models (SLN: 3 and 5) with those (hSLN: 4 and 6) we conclude that external data augmentation resulted in performance gains. We additionally observe that financial domain-specific embedding FinBERT [69] (SLN: 5, 6) resulted in improvement of the model performance compared to generic embedding like BERT [1] (SLN: 3, 4). Furthermore, it is particularly notable that fine-tuning FinBERT [69] using a classifier layer to top (SLN: 5 and 6) to predict hypernym did not perform as well as fine-tuning the FinBERT model using sentence transformer where negative samples were also included (reference: SFinBERT_neg with SLN: 7). This is because several hypernyms were interdependent as shown in Figure 3.3.

Models with SLN 8 to 15 have been developed by other participating teams. Since their models were not open-sourced, we cannot present the performance of their models on our held-out validation set. For team MXX (SLN: 13), we quote the performance as reported in their validation set [119].We mentioned the approaches followed by other teams in Table 3.2. In model SFinBERT_neg_th (SLN: 16) we changed 'k' (mentioned in Section 3.2.5) from 0.4 to 0.2. The remainder has been kept identical to the model SFinBERT (SLN: 7). Similarly, we experimented with changing the sampling strategy in the model SFinBERT_neg_ss (SLN: 17). Rather thna sampling across the entire set 'L' (as mentioned in Algorithm 1), we tested including all other hypernyms. Neither method yielded improvements.

Additionally, in model SFinBERT_neg (SLN: 7) we evaluated FinISH embeddings instead of FinBERT embeddings. We trained it for 45 epochs after increasing batch size to 30. This yielded improved model performance (Mean Rank: 1.072, and Accuracy: 0.952). We refer to this model as SFinHyp_neg (SLN: 18). As mentioned in Section 3.2.4.6, adding more data to this model led to performance degradation. This is due to the fact that this data is composed mainly of proper nouns. We denote to it as model SFinHyp_more_data (SLN: 19). Finally, ensembling models SFinHyp_neg (SLN: 18) with SFinBERT_neg (SLN: 7) resulted in the best performance (Mean Rank: 1.053, and Accuracy: 0.967). It performed even better than the old model we submitted at FinSim-3 (SLN: 7) and the existing state-of-the-art model MXX (SLN: 13)on the held-out validation set. We denote this ensemble model as Ensemble_7_18 (SLN: 20).

We further analyse the results for each label along with their root nodes. This is presented in Table 3.8. We observe that for all the labels having a root node 'CIV', 'SEC' and for the labels 'Forward', 'Option', 'Future', 'Credit Events' and 'Equity Index' the model performs the best. For labels 'Stock Corporation', 'Swap' the proposed model performs most poorly. For the remaining labels, model performance is mediocre.

Next, in Figure 3.6, we applied Principal Component Analysis (PCA) to visualize the embeddings of the hypernyms generated using the method SFinHyp_neg (SLN: 18) in two dimensions. Notably, that 'Option' and 'Future' despite lacking both the root node or the first child node in common, are proximal in the embedding space. This is likely because they are similar financial trading products, and thus share semantic similarity. This suggests that the model has captured the semantic aspect to some degree. We also observe that 'Regulatory Agency' and 'Central Securities Depository' that share the same root node 'FBC' are together. Similarly, hypernyms that share no common ancestors like 'Stock Corporation' and 'Debt pricing and yields' are separated from the remaining hypernyms. However, this pattern does not hold for most other hypernyms in the visualized space. This discrepancy arises because we lose out on significant information during projecting from 768 dimensions of the embeddings to 2-dimensionsal space. The PCA model retains only 28.3% of the variance in the original 768-dimensional space.

			Validation Set		Test Set	
SLN.	**Model**	**Data Aug.**	**MR**	**Acc.**	**MR**	**Acc.**
1	Base-1	No	2.158	0.498	1.941	0.564
2	Base-2	No	1.201	0.876	1.750	0.669
3	BERT	No	1.177	0.899	-	-
4	BERT	Yes	1.153	0.928	-	-
5	FinBERT	No	1.117	0.928	1.257	0.886
6	FinBERT	Yes	1.110	0.942	1.220	0.895
7	SFinBERT_neg (Our old model) [7]	Yes	1.086	0.947	1.156	0.917
8	dicoe_1 [136]	No	-	-	1.180	0.889
9	dicoe_2 [136]	Yes	-	-	1.162	0.904
10	MiniTrue_2 [124]	No	-	-	1.315	0.865
11	MiniTrue_1 [124]	No	-	-	1.346	0.855
12	MiniTrue_3 [124]	No	-	-	1.337	0.825
13	mxx [119]	Yes	*1.06*	*0.96*	**1.113**	**0.941**
14	yseop_1 [137]	Yes	-	-	1.236	0.883
15	yseop_2 [137]	Yes	-	-	1.141	0.917
16	SFinBERT_neg_th	Yes	1.110	0.938	-	-
17	SFinBERT_neg_ss	Yes	1.105	0.933	-	-
18	SFinHyp_neg	Yes	1.072	0.952	-	-
19	SFinHyp_more_data	Yes	1.306	0.813	-	-
20	**Ensemble_7_18 (Our new Model)**	Yes	**1.053**	**0.967**	-	-

TABLE 3.7: Results on validation and test set. Org. represents original and Ext. represents extended. Base refers to baseline. MR is Mean Rank.

Root	**Label**	**Mean Rank**	**Acc.**
BE	Stock Corporation	1.333	0.833
CIV	Funds	1.000	1.000
DER	Forward	1.000	1.000
DER	Option	1.000	1.000
DER	Swap	1.200	0.800
FBC	Future	1.000	1.000
FBC	Regulatory Agency	1.087	0.935
FBC	CSD	1.042	0.958
FBC	Credit Events	1.000	1.000
IND	Equity Index	1.000	1.000
IND	Credit Index	1.143	0.952
MD	Debt pricing and yields	1.059	0.941
SEC	Bonds	1.000	1.000
SEC	MMIs	1.000	1.000
SEC	Stocks	1.000	1.000
SEC	Parametric schedules	1.000	1.000
SEC	Securities restrictions	1.000	1.000

TABLE 3.8: Model performance for each label
CSD means Central Securities Depository.

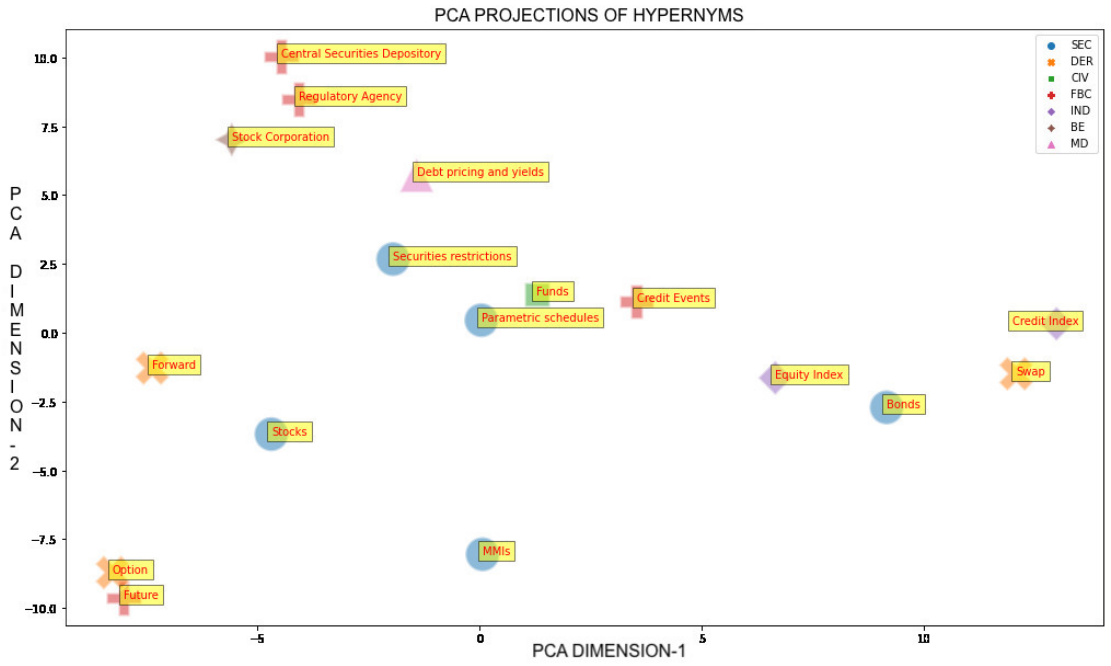

FIGURE 3.6: PCA projection of embeddings of Hypernyms in two dimensions. Same shape denotes same root nodes.

Model	Mean Rank	Acc.
Only FinBERT + cos. sim.	2.421	0.297
Only SFinHyp + cos. sim.	1.301	0.804
SFinBERT_neg	1.086	0.947
SFinHyp_neg	1.072	0.952
Ensemble (Our Model)	**1.053**	**0.967**

TABLE 3.9: Ablation Study on the validation set. cos. sim. means cosine similarity.

Ablation Study

To examine the significance of each component of the proposed model (See Figure 3.5) we conduct an ablation study. We present the results in Table 3.9. Analysing these results, we see that if one uses readily available FinBERT embeddings [69] or fine-tuned RoBERTa embeddings [137] to rank the hypernyms based on cosine similarity with the financial terms and their definitions, then the performance deteriorates drastically. This explains the necessity of the algorithm we developed for generating negative samples. The resulting ensemble model performs better than the constituent models.

3.2.8 Conclusion

In this chapter, we examine the approaches followed by participants of all three editions of the FinSim challenge. Furthermore, we present a novel method of fine-tuning FinBERT [69] and FinISH [137] embeddings leveraging hierarchies present in the FIBO structure. This approach enabled us to rank a set of hypernyms for any given financial term. We

conclude that pre-trained, transformer-based embeddings fine-tuned with domain-specific data yielded superior performance. We also observe that augmenting the existing dataset with external data enhanced the model performance. However, adding more data such as names of companies, mutual funds, and stocks yielded no performance gains.

When examining model stability of the model, we observed that during the training phase, we randomly sampled only in two places. During evaluation, we used two models to generate embeddings. These were subsequently used to calculate cosine similarities between any given set of financial terms and hypernyms. Final ranking is performed by computing "average of the two cosine similarities. Thus, the predictions generated from the ensemble model are stable, and reproducible.

Unlike the models developed by other participating teams ([119], [137], and so on), the proposed model is not a classification model. Thus, it is not necessary to retrain the model frequently when additional hypernyms are added. Moreover, the LSTM network which team MXX [119] trained cannot be easily parallelized and scaled. It will not be able to effectively handle with out-of-vocabulary words. It is simpler to compute the mean of two cosine similarity scores than employing two bi-directional LSTM networks to predict the hypernyms. This makes the proposed model simple, scalable, and easy to deploy compared to those of competing approaches.

3.2.9 Future Works

In the future, we plan to gather more data for training, and to explore leveraging Knowledge Graphs and Graph Neural Networks to improve these models. We aim to investigate on the interpretability of these models using various explainability plots and to participate in forthcoming challenges such as FinSim-4[22] [142]. Additionally, a promising direction for further research would be to create embeddings specifically for financial terms and their definitions. Currently, we have focused on exploring hierarchies and relation trees present within the FIBO structure. Although 'Future' and 'Options' represent similar trading products; they are present in different trees. We aim to account for this distinction when constructing the negative sample set. Employing Neural Network based ranking loss may result in the improved rank ordering for the hypernyms. Finally, we aim to evaluate statistical significance of predictions from these models over the baselines on a larger dataset.

[22] https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-2022/shared-task-finsim4-esg (accessed on 18th September, 2023)

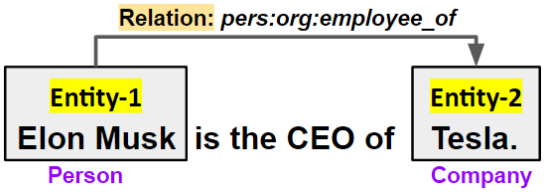

FIGURE 3.7: Relation between Financial Entities.

3.3 Extracting relationship between financial entities

3.3.1 Introduction

Automatically determining the relationships between financial entities helps in interpreting financial information about people, organizations, transactions, market sentiment which changes over time. This contextual information can help analysts, investors and others make better informed decisions. Moreover, it can provide valuable insights for a range of stakeholders, from financial institutions to regulators and companies. These insights can also help in risk management, complying with regulations, and so on. For example, as presented in Figure 3.7, *"Elon Musk is the CEO of Tesla."*, "Elon Musk" and "Tesla" are two entities of type person and company respectively. The relation between these two entities is *"employment"*. Although this appears to be simple, in real scenarios relations between entities are difficult to comprehend. For instance, a person who is a board member need not necessarily be an employee. Moreover, understanding relations between entities helps with the construction of Knowledge Graphs, which have a variety of use cases in the financial industry ranging from fraud detection to stock market prediction. Financial documents tend to be long, with lots of complex relations between entities. Reading and understanding these documents is a tedious task. Thus, an automated system for extracting insights relating to relations between entities present in these documents help.

In this chapter, we introduce the Mask One At a Time (MOAT) framework for automatically detecting the relationship between financial entities. Subsequently, we benchmark its performance with existing open source generative Large Language Models (LLMs). Our codebase can be accessed from here [23]

3.3.2 Related Works

Relation extraction is a well-studied field in natural language processing (NLP). It focuses on identifying the semantic relationships between entities in text. In the financial domain, relation extraction has been used to extract valuable insights from various financial documents, like news articles, financial reports, etc.

The task 8 of SemEval-2010 [143] was based on extracting 19 relations from 10,717 instances. With the advent of transformers [144], researchers have started fine-tuning BERT [1] like architectures for the task of relation extraction [145]. Sharma et al. [146] released the

[23] https://github.com/sohomghosh/REFinD

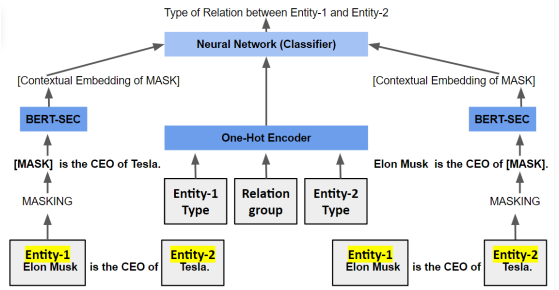

FIGURE 3.8: Architecture of MOAT.

FinRED dataset created from financial news and earning call transcripts. It comprises 29 relations and, 6767 instances.

Recently, Kaur et al. [147] released the REFinD dataset, consisting of more than 29 thousand instances involving 22 types of relations. They released a few baseline models, among which Luke Large [148] fine-tuned using the Matching the Blanks architecture [145] performed the best. With the onset of LLMs [149, 150], researchers have started exploring their applications for relation extraction.

3.3.3 Problem Statement

Given a financial entity pair $(e1, e2)$, our aim is to classify the relation between them into one of the categories mentioned in Table 3.10. For a given instance, $e1$ and $e2$ are subject and object, respectively.ORG, PER, UNI, GPE refer to organization, person, university and geopolitical entities respectively.
$e1 \epsilon \{ORG, PER\}$
$e2 \epsilon \{ORG, TITLE, UNI, GOV.AGENCY, DATE, GPE, MONEY\}$
$(e1, e2) \epsilon \{PER - TITLE, PER - ORG, PER - UNI, PER - GOV.AGENCY,$
$ORG - GPE, ORG - DATE, ORG - ORG, ORG - MONEY\}$

3.3.4 Dataset

We use the Relation Extraction Financial Dataset (REFinD) [147] for our analysis. It comprised 29,000 instances with 22 types of relations among 8 types of entity pairs. The detailed distribution of categories across the training, validation and test set is presented in Table 3.10. Subsequently, as the evaluation of generative LLMs was compute-intensive, we created a smaller test set comprising 500 instances randomly selected from the test set for benchmarking these LLMs. We use weighted average Precision (P), Recall (R) and F1-score (F1) for evaluation.

3.3.5 System Description

We present the architecture of MOAT in Figure 3.8. For a given text having two entities, we mask one entity at a time and extract the contextual SEC-BERT [151] based embeddings

categories	# train	# valid	# test
no_relation	9128	1965	1953
pers:title:title	3126	671	671
org:gpe:operations_in	2832	606	605
pers:org:employee_of	1733	372	374
org:org:agreement_with	653	141	141
org:date:formed_on	448	96	96
pers:org:member_of	441	94	95
org:org:subsidiary_of	386	82	83
org:org:shares_of	286	61	61
org:money:revenue_of	217	47	47
org:money:loss_of	141	30	31
org:gpe:headquartered_in	135	29	29
org:date:acquired_on	134	28	24
pers:org:founder_of	92	19	20
org:gpe:formed_in	81	17	17
org:org:acquired_by	55	11	12
pers:univ:employee_of	53	11	12
pers:gov_agy:member_of	40	8	8
pers:univ:attended	30	6	7
pers:univ:member_of	23	5	5
org:money:profit_of	20	4	5
org:money:cost_of	16	3	4
TOTAL	**20,070**	**4,306**	**4,300**

TABLE 3.10: Distribution of categories of relation.

of the [MASK] tokens separately. We use these embeddings along with the one-hot encoded vectors of the types of entities and their relation group as features. A relation group is the concatenation of entity types, i.e. if entity-1 is *PERSON* and entity-2 is *TITLE*, the relation group is *PERSON-TITLE*. Using these features, we train a neural network (NN) for 300 epochs with two hidden layers having 512 neurons and 128 neurons respectively. The NN classifies the relationship between the entities into one of the categories mentioned in Table 3.10.

3.3.6 Experiments and Results

We first masked the entities. We extracted the contextual embeddings of these masks, concatenated them, and fine-tuned a SEC-BERT [151] for the task of classification. The performance was suboptimal. We removed the masks, froze the underlying BERT model, and calculated the mean of the last hidden layer's embeddings of tokens constituting the entities. We concatenated these mean embeddings ($EMB_{E1,\,E2}$) and passed them through a feed forward neural network (NN). A few instances had more than 512 tokens. In those instances, we consider only 512 tokens around the entities, as the context length of SEC-BERT is 512 tokens only. This improved the performance steeply. Finally, on experimenting with the proposed MOAT architecture (described in §3.3.5) we obtained the best performance. We present the results on the test set in Table 3.11.

Model	P	R	F1
SEC-BERT	0.206	0.454	0.284
$\text{EMB}_{\text{E1, E2}}$+NN	0.731	0.701	0.709
MOAT	**0.748**	**0.743**	**0.736**

TABLE 3.11: Performance of discriminative LLMs.

Model	P	R	F1
MOAT	0.748	**0.743**	**0.736**
-relation group, entity types	0.736	0.725	0.720
$+\text{SBERT}_{\text{SDP}}$	0.694	0.687	0.679
+POS tags	0.747	0.738	0.737
MOAT (per relation group)	**0.839**	0.672	0.715
MOAT (LUKE)	0.467	0.545	0.497

TABLE 3.12: Ablation Study.

Type	LLM	P	R	F1
Zero Shot	Falcon	0.538	0.434	0.362
	Dolly	0.400	0.316	0.253
	MPT	0.295	0.380	0.255
	LLaMA-2	0.192	0.260	0.202
Few Shot	Falcon	0.246	0.258	0.242
	Dolly	0.348	0.234	0.245
	MPT	0.296	0.156	0.128
	LLaMA-2	**0.786**	0.352	0.314
Classifier	MOAT	0.726	**0.724**	**0.717**

TABLE 3.13: MOAT versus generative LLMs.

We conducted an ablation study in which we experimented by removing features constructed using the relation group (i.e. $(e1, e2)$) and entity types. We further added SEC-BERT [151] based sentence embeddings [68] of the shortest dependency path (SDP) provided in the dataset and one hot encoded vectors generated from POS tags of entities as features. Subsequently, we trained separate classifiers for each relation group and experimented by replacing SEC-BERT embeddings with that of Luke [148]. Training separate classifiers for each relation group improved the precision (P), but the recall (R) and f1-score (F1) fell drastically. The ablation study is presented in Table 3.12.

Recently, as generative-based LLMs have outperformed traditional methods of relation extraction [149], we benchmarked the performance of MOAT with three such open-source LLMs[24], namely Falcon, Dolly, MPT, and LLaMA-2 under zero shot and few shot setting. We used the smaller test set comprising 500 instances selected randomly for inference because evaluating these LLMs is computationally expensive. We observe that MOAT outperforms all these LLMs in both zero shot and few shot settings in terms of recall and F1 scores. However, LLaMA-2 gives the best precision. This is presented in Table 3.13.

[24]More details are in the Appendix

LLMs (Type)	**Prompt**
Falcon, Dolly, MPT, LLama-2 (Zero Shot)	Determine the relationship between entities [e1] of type [e1-type] and [e2] of type [e2-type] in the text given below. Choose any one from <list of relation categories> Input: [text] Response: Relation between [e1] of type [e1 type] and [e2] of type [e2 type] is
Falcon, Dolly, LLaMA-2 (Few Shot)	Determine the relationship between entities [e1] of type [e1-type] and [e2] of type [e2-type] in the text given below. Choose any one from <list of relation categories> Input: the capitalization table attached hereto as Exhibit H (the Cap Table) sets forth all of the outstanding capital of Microbot on a Fully - Diluted Basis as of the Effective Date , including the shares issued to Technion on account of Professor Shoham s involvement as a founder of Microbot ; . Response: Relation between 'Shoham' of type 'PERSON' and 'Professor' of type 'TITLE' is "pers:title:title" Input: volatility in exchange rates between the U.S. dollar and currencies of the countries in which SAN DIEGO GAS & ELECTRIC CO operate , as SAN DIEGO GAS & ELECTRIC CO discuss below . Response: Relation between 'SAN DIEGO GAS & ELECTRIC CO' of type 'ORG' and 'U.S.' of type 'GPE' is "org:gpe:operations in" Input: Dr. Fink joined Maxwell as President and Chief Executive Officer, and was appointed a director in May 2014, therefore his 2014 compensation in the table above reflects only a partial year . Response: Relation between 'Fink' of type 'PERSON' and 'Maxwell' of type 'ORG' is "pers:org:employee of" Input: [text] Response: Relation between [e1] of type [e1 type] and [e2] of type [e2 type] is
MPT (Few Shot)	Similar to Falcon and Dolly with <|endoftext|>added at the end of each response.

TABLE 3.14: Prompts for LLMs. The portion within box brackets, i.e. [content] is replaced by corresponding content from the dataset and is enclosed by backticks. <list of relation categories> refers to the categories mention in Table 3.10.

3.3.7 Conclusion

In this chapter, we discussed how language models can be used to determine the relation between a given pair of financial entities. We experimented with different LLMs and their variations. Finally, we propose MOAT, a novel architecture for determining relationship between financial entities.

Instruction fine-tuning generative LLMs, engineering better prompts, extracting entities and relations jointly, and benchmarking MOAT on other relevant datasets are a few directions for future work.

3.3.8 Prompts

Details relating to the prompts are mentioned in Table 3.14.

Chapter 4

Impactful Investing

Nowadays, in addition to risk and return, investors focus on making investments that have a lesser impact on the environment. In this chapter, we discuss how Natural Language Processing can be used to identify the ESG and sustainability-related aspects from financial texts.

4.1 Research Questions

- **RQ-4**: How to ensure that the investments are towards benefit of the Earth and have a positive impact on the environment?

  • **Relevant Contributions**: Detecting ESG and sustainability-related concepts [10], issues [11], impact [12], and impact duration [13].

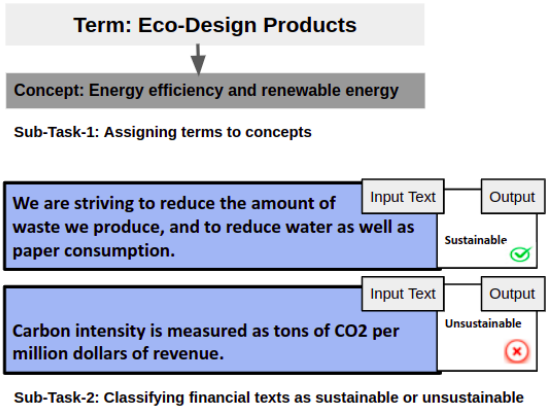

FIGURE 4.1: FinSim-4 ESG subtasks

4.2 Detecting ESG and sustainability related concepts from Financial Texts

4.2.1 Introduction

Nowadays, many investors have become socially responsible and environmentally conscious [1]. They tend to choose stocks and funds that do not harm the environment [2]. With this in mind, organisations are also putting in efforts to increase their ESG ratings. Organisations tend to publish reports mentioning the ESG aspect of their policies. However, reading through all such reports is time-consuming and inefficient. This highlights the need for an automated system for mapping terms to ESG concepts and automatically classifying financial texts as sustainable or not. FinNLP workshop of the IJCAI-2022 conference hosted a shared task with these problem statements. We present an example of this in Figure 4.1. Our team, LIPI participated in the shared task and ranked 4^{th} and 3^{rd}, in subtasks 1 and 2, respectively. In this chapter, we describe our approaches.

4.2.2 Related Works

The subtask of mapping terms with high-level concepts is similar to Hypernym Detection. For the Natural Language Processing (NLP) research community, Hypernym Detection has been an active area of research. Several SemEval tasks ([107], [108], [109], [110]) were organised on this topic. Subsequently, three editions of FinSim ([92], [93], [94]) shared tasks were held that adapted the task of hypernym detection for the financial domain. This year, while organising FinSim-4, the task was extended to ESG insights.

With rising popularity of green investing, understanding the sustainability aspects of financial texts has become increasingly important. Smeuninx et al. [154] studied the readability of annual reports of several organisations. They highlighted how formule-based

[1] https://news.gallup.com/poll/389780/investors-stand-esg-investing.aspx (accessed on 10 June 2022)

[2] https://bwdisrupt.businessworld.in/article/Sustainable-Investing-To-Surge-To-125-B-In-India-By-2026-Report/09-06-2022-432078/ (accessed on 10 June 2022)

readability scores classified these texts as complex documents. They also mentioned the need for NLP-based techniques to comprehend the readability of such documents. Luccioni et al. [155] fine-tuned RoBERTa-base [125] model to develop a question-answering-based tool, namely ClimateQA for extracting Sections related to climate from financial reports.

Guo et al. [156] proposed a framework, ESG2Risk for predicting stock market prices by analysing ESG-related events from financial news. They specifically utilised sentiments from these events.

Nugent et al. [157] pre-trained a BERT [1] model with financial news articles from Reuters News Archive for predicting ESG-related controversies. They subsequently used it to map financial news into one of the United Nations Sustainable Development Goals.

4.2.3 Problem Statements

The fourth edition of FinSim presented two subtasks. They are as follows:

4.2.3.1 subtask 1:

Given a set J consisting of n tuples of terms and their high-level concepts i.e. $J = \{(t_1, c_1), (t_2, c_2), ...(t_n, c_n)\}$ where c_i represents the high-level concept corresponding to the i^{th} term t_i and $c_i \in$ set of concepts mentioned in Table 4.1. For a given unknown term, the task was to develop a system to rank these concepts.

The evaluation metrics for this subtask were accuracy and mean rank. According to the evaluation script shared by the organisers, the rank of an instance was calculated by checking the presence of the true value in the first three elements of the predicted ranked list.

4.2.3.2 subtask 2:

Given a set F consisting of n tuples of financial texts and their sustainability labels i.e. $F = \{(f_1, l_1), (f_2, l_2), ...(f_n, l_n)\}$ where l_i represents the sustainability label corresponding to the i^{th} financial text f_i and $l_i \in$ {sustainable, unsustainable}. We need to develop a system to classify an unknown financial text as sustainable or not.

The evaluation metric for this subtask was accuracy.

4.2.4 Data

The datasets provided by the organisers consist of 190 financial documents in PDF format, 651 terms mapped to 24 concepts and 2,265 financial texts labelled as sustainable or unsustainable. We provide more details about the dataset in the following Sections.

Concept	**Count**
Energy efficiency and renewable energy	59
Sustainable Food & Agriculture	54
Product Responsibility	51
Circular economy	47
Sustainable Transport	46
Emissions	39
Shareholder rights	38
Board Make-Up	37
Injury frequency rate for subcontracted labour	35
Executive compensation	32
Biodiversity	29
Community	27
Employee engagement	23
Employee development	22
Water & waste-water management	21
Carbon factor	19
Future of work	18
Waste management	16
Recruiting and retaining employees	11
Human Rights	10
Audit Oversight	7
Injury frequency rate	2
Board Independence	2
Share Capital	2

TABLE 4.1: Distribution of concepts

4.2.4.1 Data Description

For subtask 1, the number of instances for each concept is shown in Table 4.1. For subtask 2, out of 2,265 financial texts 1,223 were sustainable whereas 1,042 were unsustainable. We maintained a training to validation split of 80% to 20% for both the subtasks.

4.2.4.2 Data Augmentation

Firstly, for subtask 1, we started by using 80% of 651 instances for training. To provide more context, we collected the definitions for each of the 24 concepts from various websites. For each term (t_i, concept c_i) pair, we obtained the corresponding concept definition d_i. Since each term t_i present here were mapped to a concept definition d_i, we had only positive instances, i.e., similarity score of 1.0 corresponding to the (t_i, d_i) pair. Subsequently, we considered adding negative samples in the training process. For each term, concept definition pair (t_i, d_i), we experimented by randomly pairing t_i with 1, 5, or 15 concept definitions. Later, we grouped the concepts manually. This is presented in Table 4.2. We were able to group 20 out of the 24 concepts. The remaining four concepts were singleton sets. For randomly selecting concept definitions for term t_i, we tried out the following sampling methods:

Group-1	**Group-2**	**Group-3**	**Group-4**
Carbon factor	Employee development	Injury frequency rate	Audit Oversight
Emissions	Recruiting and retaining employees	Injury frequency rate for subcontracted labour	Shareholder rights
Energy efficiency and renewable energy	Future of work	Human Rights	Executive compensation
	Employee engagement		Share Capital

Group-5	**Group-6**	**Group-7**
Waste management	Sustainable Transport	Board Independence
Waste Water management	Sustainable Food Agriculture	Board Make-Up

TABLE 4.2: Concepts divided into groups

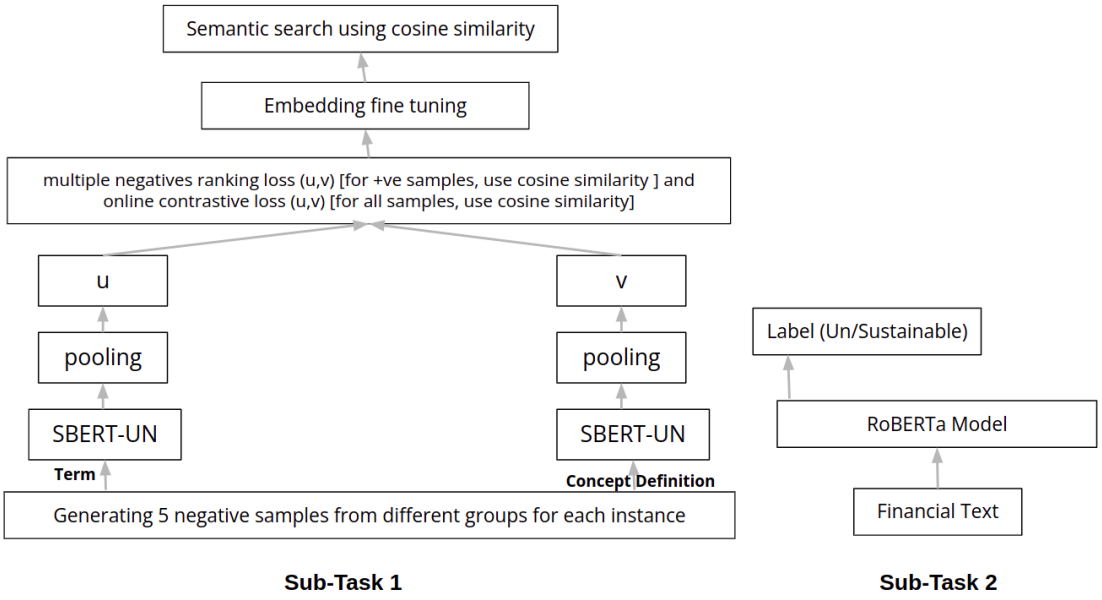

FIGURE 4.2: Methodology subtask 1 and 2

- Select any concept definition d_j such that concept $c_j \neq$ concept c_i, and assign a similarity score of 0.0 to the (t_i, d_j) pair.
- Select any concept definition d_j such that concept $c_j \notin$ the group where concept c_i is present, and then assign a similarity score of 0.0 to the (t_i, d_j) pair.
- Select any concept definition d_j, if concept $c_j \notin$ the group where concept c_i is present assign a similarity score of 0.0 to the (t_i, d_j) pair, else assign a similarity score of 0.5 to the (t_i, d_j) pair.

4.2.5 System Description

According to the rules, for every team, the number of submissions for each subtask was restricted to two. We describe each of our submissions here. We illustrate our methodology in Figure 4.2.

4.2.5.1 Subtask 1, System -1

We fine-tuned a Sentence Transformer [68] model [3] (SBERT-UN) which was pre-trained with United Nations (UN) sustainable development goals. For each of the terms in the training set, we randomly selected five concept definitions from different groups as mentioned in Section 4.2.4.2. Our objective was to minimise the Multiple Negatives Ranking Loss as well as the Online Contrastive Loss. This model was trained for 15 epochs with a batch size of 20. [4]. For subtask 1, among all our submissions, this performed the best in terms of both accuracy and mean rank. This is similar to the solution [7] presented at FinSim-3.

4.2.5.2 Subtask 1, System -2

This is a RoBERTa-base [125] based classifier. We fine-tune the pre-trained RoBERTa-base model so that its [CLS] token learns to classify terms into 24 pre-defined concepts or classes. Its hyperparameters are as follows: maximum length = 16, batch size = 256, epochs = 60, learning rate = 0.00002. We used the checkpoint created at the 57^{th} epoch as this was the best-performing one.

4.2.5.3 Subtask 2, System -1

This system consists of the pre-trained FinBERT [69] fine-tuned for classifying financial texts as sustainable or unsustainable. Its hyperparameters are as follows: maximum length = 128, batch size = 256, epochs = 60, learning rate = 0.00002. We used the checkpoint created at the 8^{th} epoch as this achieved the best validation performance.

4.2.5.4 Subtask 2, System -2

It consists of the pre-trained RoBERTa-base [125] fine-tuned for the task of classifying financial texts as sustainable or unsustainable. It's hyperparameters are as follows: maximum length = 128, batch size = 256, epochs = 60, learning rate = 0.00002. We used the checkpoint created at the 12^{th} epoch as this performed the best on the validation set. Among all our submissions, this performed the best on the test set.

4.2.6 Experiments and Results

We started by fine-tuning the all-mpnet-base-v2 model [158] using Sentence Transformer architecture. Our objective was to reduce the Multiple Negatives Ranking Loss as well as

[3] https://huggingface.co/Rodion/sbert_uno_sustainable_development_goals (accessed on 18^{th} September, 2023)

[4] The details are available at https://www.sbert.net/examples/training/quora_duplicate_questions/README.html (accessed on 18^{th} September, 2023)

Sl. No.	Base Model	Data Augmentation	Mean Rank	Accuracy
1	all-mpnet-base-v2	No (only positives)	1.4692	0.6923
2	all-mpnet-base-v2	Yes (1 negative per positive)	1.5769	0.7000
3	sbert_un	No (only positives)	1.5308	0.6769
4	sbert_un	Yes (1 negative per positive)	1.4769	0.7308
5	sbert_un	Yes (1 negative per positive) + concepts	1.4615	0.7154
6	sbert_un	Yes (1 negative per positive) - concept definitions + concepts	1.4846	0.7462
7	sbert_un	Yes (1 negative per positive) [out of group sampling]	1.4385	0.7462
8	sbert_un	Yes (5 negative per positive) [out of group sampling]	**1.4308**	**0.7615**
9	sbert_un	Yes (15 negative per positive) [out of group sampling]	1.5308	0.7000
10	sbert_un	Yes (5 negative per positive) [out of group sampling] {batch size = 40, epoch = 30}	1.4154	0.7462
11	sbert_un	Yes (5 negative per positive) [out of group sampling] {batch size = 40, epoch = 20}	1.4615	0.7462
12	roberta classifier	-	1.4846	0.7538
13	sbert_un	Yes (1 negative per positive) [same group & out of group sampling]	1.4615	0.7462
14	sbert_un	Yes (5 negative per positive) [same group & out of group sampling]	1.5000	0.7385
15	baseline-1	-	2.5308	0.3769
16	baseline-2	-	1.6846	0.7154

TABLE 4.3: Results of subtask 1 on the validation set.
NOTE: Where not mentioned, definitions of concepts were used with batch size of 20 for 15 epochs.

the Online Contrastive Loss for the task of Information Retrieval [5]. We also studied the effect of replacing this model with the SBERT-UN model, adding negative samples and concepts as is. We further experimented with different sampling methods as mentioned in Section 4.2.4.2. Furthermore, we fine-tuned a RoBERTa-base [125] based model to classify terms into 24 predefined concepts or classes.

Subsequently, we extracted texts from the documents provided in PDF format and fine-tuned a SBERT-UN model using Masked Language Modelling. However, this did not improve the performance. We also tested adding the definitions of 73 terms obtained from DBpedia [127]. However, this did not yield substantial improvement in the results. We present the results of subtask 1 in Table 4.3. The SBERT-UN model trained with negative samples (SL. No. 8) performed the best in the validation as well as in the test set.

For subtask-2, we fine-tuned four models for classifying financial texts into two classes: sustainable and non-sustainable. These models are: RoBERTa-base [125], FinBERT [69], SBERT-UN and SBERT-UN fine-tuned for subtask 1. We present the results in Table 4.4. FinBERT [69] performed the best in the validation set, while RoBERTa-base [125] performed the best in the test set. Each of these models were trained for a maximum of 128 input tokens with a batch size of 256, a learning rate of 0.00002 and for 60 epochs.

We present the test set results in Table 4.5.

[5] https://www.sbert.net/examples/training/quora_duplicate_questions/README.html (accessed on 18th September, 2023)

Sl. No.	Model	Accuracy
1	roberta-base	0.9338
2	finbert	**0.9426**
3	sbert_un	0.8653
4	subtask1 finetune	0.8543

TABLE 4.4: Results of subtask 2 on the validation set.

ST	Sub.	Accuracy	Mean Rank
1	1	**0.7103**	**1.5172**
1	2	0.7034	1.6689
2	1	0.9219	-
2	2	**0.9317**	-

TABLE 4.5: Test set results for subtasks (ST) 1 and 2. Sub.: Submission

4.2.7 Conclusion and Future Work

In this chapter, we elaborate on our team LIPI's approach to solving the FinSim-4-ESG subtasks. According to the official report, out of 28 registered teams, six and eight teams participated in subtask 1 and 2 respectively. For subtask 1, our team ranked 4^{th}, while for subtask 2, our team ranked 3^{rd}.

In future, we plan to collect more data and work to improve the models' performance. Developing a user-friendly tool for assigning terms to concepts and automatically evaluating the sustainable aspects of financial texts constitute other directions of future work.

4.3 Assessing ESG-related issues from Financial Texts

4.3.1 Introduction

Socially responsible and environmentally conscious investors prefer to evaluate investment opportunities based on parameters related to ESG. They focus on impact investing, i.e., they also look for ESG-related issues while managing risk and return. They rely on third-party agencies such as Bloomberg ESG Ratings, CDP Scores, FTSE Russell ESG Ratings, ISS (Institutional Shareholder Services) ESG Ratings and Rankings, MSCI (Morgan Stanley Capital International) ESG Ratings, etc. [6] to check the ESG performance of the organisations [159]. In 2020, the European Commission published the EU taxonomy, which is the classification of economic activities that are environmentally sustainable [160] to help companies formulate policies and publish ESG reports. To attract potential investment, the firms are following ESG reporting guidelines to disclose their environmental, social, and governance impact and getting their sustainability reports verified by third parties [161]. Studies show that ESG strategies positively affect stock returns in the utilities and energy sectors [162]. Zhou et al. [163] revealed how ESG performance improvement of a company leads to improvements in operating capacity and market value. However, for investors, reading through news articles manually and trying to align them to ESG-related issues is time-consuming and monotonous. We developed a SEC-BERT [151] based model and an **`ESG Issue Detector (EID)`** tool to automate this. Additionally, we seek answers to the following research questions.

Research Questions (RQ)

- RQ-1: Do domain-specific models help in improving the model performance?
- RQ-2: Does in-domain pre-fine-tuning help in improving the model performance?
- RQ-3: Does augmenting the training data with translated samples from the same domain (which are originally in a different language) help in improving performance?
- RQ-4: How well do existing open-source Large Language Models (LLMs) under a zero-shot setting perform the task of ESG issue classification?

Our contributions

- We conducted rigorous experiments to answer RQ-1, RQ-2, RQ-3 and RQ-4 with justification.
- We developed a SEC-BERT [151] based model to detect ESG-related issues from news titles and contents.
- We present a user-friendly ESG Issue Detector (EID) tool to help investors in identifying ESG-related issues easily. [7]

[6] `https://iriscarbon.com/a-beginners-guide-to-esg-rating-agencies-and-methodologies/` (accessed on 1st May 2023)

[7] The model artefacts and the EID tool can be accessed from `https://huggingface.co/spaces/sohomghosh/EID_ESG_Issue_Detector` (accessed on 18th September, 2023)

4.3.2 Related Works

Studying various aspects of ESG has been an active area of research. Kim et al. [164] performed multivariate regression analysis with variables such as credit rating, earnings before interest and taxes, and invested capital to study the effects of ESG factors on corporate finance. They found that social and governance had a positive impact while the environment had a negative effect on credit rating, and governance positively impacts corporate profitability. Koloski et al. [165] trained a classifier for classifying financial texts as sustainable or unsustainable. Kang et al. [142] introduced the FinSim4-ESG Shared Task. It comprised two subtasks - a) ranking ESG-related concepts for a given unknown term, and b) classifying a financial text as sustainable or unsustainable. Most of the teams ([166][10]) used transformer-based architecture such as Sentence BERT [68], BERT [1] and RoBERTa [125] to achieve state-of-the-art performance. Recently, Iazzolino et al. [167] investigated the effect of factors related to ESG on business efficiency. They performed gap analysis to estimate the impact of ESG factors on the efficiency of organisations across different sectors. Zhou et al. [163] demonstrated the positive relation between ESG performance and the market value of companies, while Chen et al. [168] stated the positive effect of ESG disclosures on the financial performance of organisations. Gillan et al. [169] and Bruna [170] studied the relationship between corporate finance and ESG. Guo et al. [156] analysed ESG-related events from news to predict stock prices. Nugent et al. [157] used BERT [1] model for detecting controversies relating to ESG from financial news. Although researchers have been studying the ESG field for quite some time, there is no tool to automatically align news articles to ESG-related issues.

4.3.3 Problem Statement

Given a news title, our objective is to develop a model for classifying it into one of the 33 ESG issue types. [8] We use accuracy, F1-score, precision, and recall for assessing the model performance.

4.3.4 Dataset

The dataset has been recently released in the FinSim-2023 Multilingual ESG Issue detection shared task [171]. It consists of financial news contents related to ESG that were obtained from the ESG Today news website. [9]Each instance has the URL of the news article, title of the news, and its content. Each instance has been annotated as one of the 33 ESG issues.[10] There are a total of 1,199 instances in English. Additionally, we have 1,077 annotated instances in French obtained from the Novethic website. [11] We use 80% of the English dataset for training the model and the remaining 20% for validation. Later, we translated the instances from French to English and added them to the training set to assess if it can help improve the performance of the model. The class-wise distribution for the English and the French dataset is given in Table 4.6.

[8] https://www.msci.com/our-solutions/esg-investing/esg-industry-materiality-map (accessed on 1st May 2023)

[9] https://www.esgtoday.com/ (accessed on 1st May 2023)

[10] https://www.msci.com/our-solutions/esg-investing/esg-industry-materiality-map (accessed on 1st May 2023)

[11] www.novethic.fr (accessed on 1st May, 2023)

ESG-Issue	count in English dataset	count in French dataset
Packaging Material & Waste	81	7
Board	74	22
Carbon Emissions	73	29
Financing Environmental Impact	71	66
Responsible Investment	70	45
Opportunities in Renewable Energy	68	105
Opportunities in Clean Tech	62	28
Human Capital Development	62	34
Product Carbon Footprint	55	44
Biodiversity & Land Use	53	22
Consumer Financial Protection	47	27
Opportunities in Green Building	44	16
Ownership & Control	42	27
Community Relations	40	30
Business Ethics	34	108
Climate Change Vulnerability	30	21
Raw Material Sourcing	29	18
Water Stress	27	16
Toxic Emissions & Waste	26	21
Pay	22	47
Opportunities in Nutrition & Health	22	26
Health & Demographic Risk	22	17
Access to Finance	20	25
Access to Health Care	19	21
Accounting	18	17
Chemical Safety	18	36
Privacy & Data Security	14	39
Access to Communications	13	4
Electronic Waste	12	18
Supply Chain Labor Standards	11	35
Product Safety & Quality	10	29
Labor Management	6	22
Controversial Sourcing	4	55
TOTAL	**1199**	**1077**

TABLE 4.6: Class-wise distribution of the dataset

4.3.5 Methodology

Firstly, we pre-fine-tuned SEC-BERT [151] on the news content present in the training set using masked language modelling. We randomly masked 15% of the tokens. Subsequently, we provided the title and content of the news articles as input to the pre-fine-tuned model. We fine-tuned this model further so that the [CLS] token learns to classify the input text into one of the 33 ESG issues. It was trained for seven epochs with training and validation batch size of 8 and 32 respectively. We used weight decay of 0.00001 and 500 steps to warm-up. The process is presented in Figure 4.3.

4.3.6 Experiments and Results

We carried out several experiments. We started by fine-tuning DistilBERT [172] for the task of classifying titles of news articles in English into one of the 33 ESG issue categories.

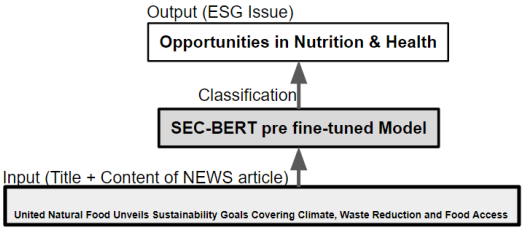

FIGURE 4.3: Model Architecture

On replacing DistilBERT [172] with finance and ESG domain specific embeddings like SEC-BERT [151] and finbert-esg[12] [173], the model performance improved. Subsequently, we noticed that adding the content of the news articles along with the title and pre-fine-tuning the existing models on the target dataset (ESG News content) using masked language modelling helped in enhancing the performance further. Lastly, we experimented by translating instances from French to English using Google Translate and adding them to the training set. However, this did not improve the performance of the model. This is probably because the process of machine translation is somewhat lossy. We present the results in Table 4.7. Each of the models was trained for 7 epochs with a training batch size of 8, evaluation batch size of 32 and 500 warm-up steps. We picked up the best performing model in each case. After analysing the results we conclude that pre-fine-tuned SEC-BERT [151] when trained only on the title and content of the English news articles, performed the best. We further tried ensembling results of the two models numbered as 6 and 7 in the Table 4.7. The ensembling was done by selecting the more confident model's output. However, this did not result in performance improvement. Recently, LLMs like ChatGPT[13] have revolutionized research in the domain of NLP. We experimented with FLAN-UL2[14] [74], [75] which is an open-source LLM released by Google and Dolly v2[15] released by Databricks. In a zero shot setting, we prompted FLAN-UL2 to classify the news titles (Sl. No. 13) and then news titles combined with their contents (Sl. No. 14) respectively. In both scenarios, the model performance deteriorated. Dolly v2 in zero shot setting when prompted with same inputs generated lengthy and ambiguous results. Thus, we exclude Dolly v2 from the comparison.

Since the actual labels of the hold out test set for the Multilingual ESG Issue identification shared task [171] are not released publicly, we are unable to benchmark our model with that of the other contestants. According to the publicly available leaderboard[16], team NCMU performed the best on the English dataset (Precision: 0.69, Recall: 0.70, F1-score: 0.69). Lastly, we conducted McNemar's test and concluded that the proposed model (Sl. No. 6) performs significantly better (p-value <0.05) than the baseline (Sl. No. 1).

Sl. No.	Model	Dataset	Accuracy	F1	Precision	Recall
1	distilbert	Eng (T)	0.637	0.624	0.662	0.637
2	sec-bert	Eng (T)	0.650	0.653	0.639	0.65
3	finbert-esg	Eng (T)	0.667	0.655	0.684	0.667
4	sec-bert	Eng (T+C)	0.717	0.708	0.732	0.717
5	finbert-esg	Eng (T+C)	0.688	0.678	0.697	0.688
6	sec-bert pre-ft	Eng (T+C)	**0.725**	**0.715**	**0.726**	**0.725**
7	finbert-esg pre-ft	Eng (T+C)	0.692	0.682	0.700	0.692
8	sec-bert pre-ft	Eng + Fr2Eng (T)	0.646	0.640	0.684	0.646
9	finbert-esg pre-ft	Eng + Fr2Eng (T)	0.629	0.621	0.645	0.629
10	sec-bert pre-ft	Eng + Fr2Eng (T+C)	0.692	0.687	0.703	0.692
11	finbert-esg pre-ft	Eng + Fr2Eng (T+C)	0.687	0.683	0.702	0.687
12	Ensemble 6 & 7	Eng (T + C)	0.696	0.688	0.722	0.696
13	FLAN-UL2	Eng (T)	0.367	0.339	0.627	0.367
14	FLAN-UL2	Eng (T + C)	0.417	0.417	0.596	0.696

TABLE 4.7: Results (Eng = English, Fr2Eng = French translated to English, C = content, T = Title, ft = fine tuned)

FIGURE 4.4: Error Analysis

Error Analysis

We use the transformer-interpret[17] library to study a few instances which were misclassified by the best performing model. We showcase five such instances in Figure 4.4. Tokens which contribute positively and negatively towards the predictions are highlighted with green and red respectively. Manual review of these misclassified instances reveal that in most cases the predicted and actual labels are similar. Additionally, in the majority of the cases (like Example 1,2,3 & 4 in Figure 4.4) the predicted label appears more

[12] https://huggingface.co/yiyanghkust/finbert-esg (accessed on 1st May, 2023)
[13] https://chat.openai.com/chat (accessed on 1st May, 2023)
[14] https://www.yitay.net/blog/flan-ul2-20b (accessed on 1st May, 2023)
[15] https://huggingface.co/databricks/dolly-v2-3b (accessed on 1st May 2023)
[16] https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-2023/shared-task-esg-issue (accessed on 30th) May, 2023
[17] https://github.com/cdpierse/transformers-interpret (accessed on 1st May, 2023)

FIGURE 4.5: User Interface of ESG Issue Detector

appropriate than the actual label. Another possible reason for misclassification seems to be risen in diversity of topics when corresponding news content are used as additional input for the classification of the news articles.

4.3.7 ESG Issue Detector (EID) Tool

For helping investors, we developed the ESG Issue Detector (EID) tool using Gradio [174]. This is presented in Figure 4.5. It takes title and content of news articles input from users. When the **`Submit`** button is clicked, it detects the ESG issue and presents it in the output box on the right-hand side. If the probability of the detected issue is below a threshold, it displays a message stating that it is not confident about the prediction. To collect feedback from users, we have kept a **`Flag`** button. Users can use this to flag instances where the output seems to be inappropriate. In the future, we would like to analyse and learn from these miss-classified instances.

4.3.8 Conclusion

In this chapter, we presented how we use a SEC-BERT [151] model for the task of identifying ESG-related issues. We also observed that domain specific embeddings help in improving the model performance (RQ-1). Moreover, when the embeddings were pre-fine-tuned on the target dataset, it improved the model performance further (RQ-2). However, translating instances from French to English and adding them to the training set degraded the performance of the model (RQ-3). We also found that LLMs like FLAN-UL2 and Dolly v2 in zero-shot setting does not perform as well as ones finetuned specifically for the task (RQ-4). Finally, we developed the ESG Issue Detector (EID) tool for aiding investors to understand ESG-related issues from news articles. In future, we would like to extract and feed that portion of the news content which is most relevant to the title of the news. Another direction of future work is to study how generative LLMs perform in few shot setting and after instruction fine-tuning.

Acknowledgements We thank the organisers of FinNLP-2023 multilingual ESG issue shared task[18] for sharing with us the multilingual ESG issue detection dataset.

[18] `https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-2023/shared-task-esg-issue` (accessed on 1st May 2023)

4.4 Assessing type of ESG impact from Financial Texts

4.4.1 Introduction

In recent times, the focus on Institutions' Environmental, Social, and Governance factors (ESG) has garnered considerable interest from the global investment and corporate governance communities. People have also grown to be socially responsible and environmentally conscious while investing. ESG serves as a third dimension beyond risk and return. Research also indicates that Institutions with better ESG performance directly correlate to better stock performance and risk management [175]. Keeping this in mind, many rating agencies quantify the nature and impact of ESG aspects of an institution and publish ratings [176]. Apart from ESG investing, Impact investing [177] has also gained traction where investors, instead of investing solely based on ESG benefits, would look for a combination of better returns as well as a positive influence in society. Hence, impact identification is crucial to determine whether statements are an opportunity or a risk for the Institution.

Most of the scoring processes involved in ESG and Impact assessments are extremely time-consuming and require expert involvement and manual annotations. To automate this, we propose a generalized pipeline capable of predicting the impact types of ESG-related news articles (as shown in Figure 4.6). This generalized pipeline can be scaled to other low-resource datasets as well.

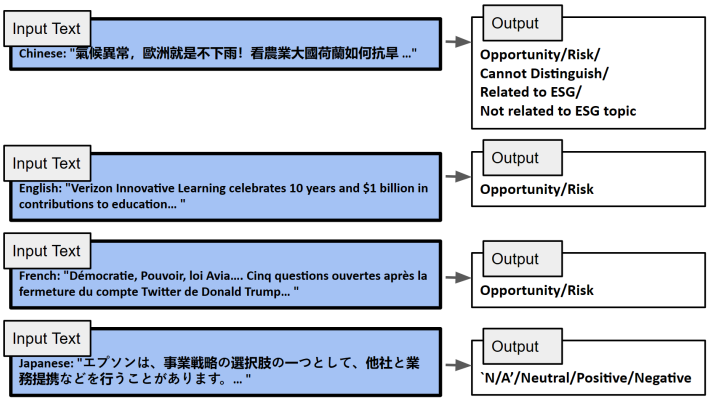

FIGURE 4.6: The Multilingual ESG Impact Assessment Task.

The labels primarily indicate if the given news is an opportunity or a risk from the ESG aspect. In this shared task, we participated in all four languages and were ranked 1^{st} in Chinese and Japanese subtasks, 4^{th} in French, and 7^{th} in English.

4.4.2 Related Work

With the advent of green investing, many approaches and models have been developed to automate processes in Financial and ESG-based NLP research, including the development of models like FinBERT [69], ESGBert [178], etc. While there has been much work on ESG-type classifications, including on multilingual datasets, more work needs to be done on impact-type classifications. FinNLP 2023 [171] focuses on a similar task where

participants were required to classify multilingual data into the ESG issue type, where the best results were obtained by using language-specific BERT models along with data augmentations using Large Language Models. Furthermore, it's important to note that extensive research has been conducted on sentiment analysis [179, 180], which can be considered a fundamental aspect of impact identification. Attempts have also been made for impact identification in Chinese [181] and Japanese [182].

4.4.3 Task Description

The task is primarily a classification task. We need to assess whether a given financial text poses a risk or an opportunity for the company. As shown in Figure 4.6, there are multiple languages with differences in classes.

4.4.4 Data

The dataset primarily contains news articles collected from four different languages, English (en), Chinese (zh), Japanese (ja), and French (fr), along with their impact types.

Language	Train	Test	C	W_c	W_h
English	808	218	2	412.48	76.83
Chinese	1400	156	5	-	33.68
Japanese	896	225	4	-	78.82
French	818	200	2	564.88	96.17

TABLE 4.8: Metrics across languages. C denotes the number of Classes, W_c denotes the average character length of content and W_h denotes the average character length of headline. Chinese and Japanese datasets do not have content columns.

Given that the dataset across languages is small, and the class-wise distribution is highly skewed. To overcome these challenges, we use a combination of translation and data paraphrasing on minority classes.

4.4.5 Approaches

We primarily used encoder-based models for this classification task. Given the limited sample size of the dataset, variations in languages, and disparity with class distribution among different languages, We tried to make a pipeline that accounted for such differences and performed consistently well across all languages. We tried a variety of approaches like Masked Language Modelling (MLM), paraphrasing for augmenting the minority classes, Translation, and Multilingual Models and used a combination to finalize our pipeline based on empirical experiments.

All the experiments have been run using a batch size of 32, a learning rate of $2e^{-5}$, weight decay of 0.01, and for ten epochs. The reported metrics are based on 80 : 20 train-test set splits with a constant random seed and not on the final validation sets used for the leaderboard. The code, data, and models used for inferences are available at the link.

4.4.5.1 Masked Language Modelling

We performed several experiments to decide the necessary models for classification. Also, we experimented with pre-training the models beforehand on the ESG corpus, which was the English dataset for the Multilingual ESG Issue Identification (ML-ESG) [171] and then using the fine-tuned models for classifications. We noticed that across all languages, the models pre-trained on the ESG corpus and then fine-tuned for classification outperformed those fine-tuned for classification.

Approach	Title	Content
Classification	74.89%	92.48%
MLM + Classification	85.48%	93.16%

TABLE 4.9: Comparison across Classification and MLM + Classification approaches along with news headlines and content using bert-base-cased [1] model. These reported numbers are the weighted $F1$ with the English dataset.

From Table 4.9, we also observe that using news content for training over title performs better. The French dataset exhibits similar trends, and hence, for all further analysis, we use the news content for English and French and the news title for Chinese and Japanese since they do not have news content available in the dataset.

4.4.5.2 Translation and Multilingual Models

We have also experimented with specific language models vs. translating and English-based models, primarily due to a larger number of specialized models pre-trained on ESG data being available in English. We used Google Translate to translate data from French, Chinese, and Japanese and leveraged this data as additional data while training for models. Also, by using English, we were able to use paraphrasing tools to augment and extend the minority classes of our dataset.

Approach	F1
Translated	68.92%
Chinese	68.45%

TABLE 4.10: Comparison of weighted F1 scores while using translated Chinese to train a bert-base-cased model vs. using Chinese data to train a bert-base-multilingual-cased model [1].

While the disparity between the translated text and the original language may not seem substantial, there exists a possibility that employing more specialized language models tailored to the Chinese language could have potentially delivered better results. However, this approach would have restricted our ability to employ paraphrasing-based techniques, as such tools are not as readily available in non-English languages. Furthermore, it would have limited our access to English models predominantly trained on ESG data. Accordingly, our primary strategy revolved around using translated text for classification.

4.4.5.3 Paraphrasing for Data Augmentation

Given that the dataset across languages is small, and the classwise distribution is highly skewed, one of the approaches we considered for improving the classification task is to augment the minority classes and extend the dataset. While rule-based paraphrasers are popular and widely used for such tasks, the variation within sentences is frequently minor and only offers a slight improvement during training. Hence, we considered a T5-based paraphraser [72], primarily fine-tuned on ChatGPT paraphrases. It offers a better range of sentence variations than any other approaches tried. We first translated the dataset from the respective languages to English and then generated paraphrased data on minority class data (For each minority instance, approximately 3–4 paraphrases were created, depending on the specific count of instances for that particular label. For the same reason we did not paraphrase for French language since the label distribution was already uniform.The paraphrased data can be accessed here.) and used this along with the original data for training the classification model.

Approach	F1 (en)	F1 (zh)
Paraphrased Data	98.91%	84.98%
Original Data	93.16%	68.45%

TABLE 4.11: Comparison of weighted F1 while using paraphrased text vs. original dataset for MLM + Classification on the English dataset and The original dataset for Chinese and the translated + Paraphrased version of the Chinese dataset. bert-base-cased model was used for English and bert-base-multilingual-cased for Chinese.

We observe that across languages, paraphrased data improved the F1 metrics of models to a great extent. This effect was more prominent in Chinese and Japanese datasets, where the number of classes was more prominent, and there was a wider class disparity. This supports our choice of using translated text rather than the original despite lacklustre results, while just translating and using that data for classification.

4.4.6 Final System Description

For the final system that was used, based on the empirical studies performed above, We used a pipeline that initially translated all of the given text into English using Google Translate. Then we use the T5-based paraphraser [72] to generate new minority class instances. We also use an ESG corpus to initially pre-train a model on this corpus and then fine-tune it for classification on the translated and augmented dataset. Figure 4.7 shows the exact process.

We also performed more experiments to decide which models best performed on the English dataset and chose bert-base-cased [1], Finbert [69], and finbert-tone [173]. We used the same models for the other languages as well. The model hyperparameters are the same as mentioned in the methodology.

We observe that despite using a generalized pipeline and models for all the languages, the results are good. Table 4.12 shows the performance of models for all of the languages and models used.

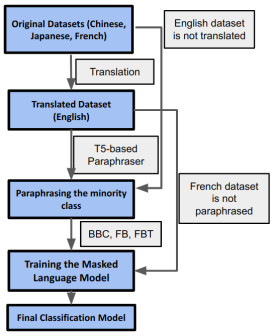

FIGURE 4.7: The final system pipeline. BBC = bert-base-cased, FB = FinBERT, FBT = FinBERT Tone

Models	F1(en)	F1 (zh)	F1 (ja)	F1 (fr)
BBC	98.91%	84.98%	89.64%	78.25%
FB	97.82%	85.31%	91.13%	78.55%
FBT	98.91%	82.26%	89.54%	70.79%

TABLE 4.12: Final weighted F1 metrics for the models used for submission. BBC = bert-base-cased, FB = FinBERT, FBT = FinBERT Tone

4.4.7 Conclusion

Comparing the performance of our models with that of other participants, we conclude that our models performed consistently well. We outperformed all other teams in the Chinese and Japanese subtasks. One unique feature is despite four different languages, we were able to use the same pipeline and same set of models and achieve consistently good results across languages, which leads us to believe that the pipeline is performant for low resource settings. All of the data generated and code used can be accessed here.

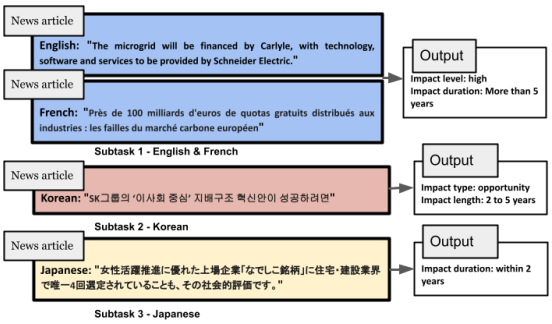

FIGURE 4.8: Overview of the ML-ESG3 task

4.5 Assessing duration of ESG impact from Financial Texts

4.5.1 Introduction

The Multi-Lingual ESG Impact Duration Inference (ML ESG-3) task being organised in conjunction with the FinNLP-KDF@LREC-COLING-2024 deals with predicting the impact of events on companies. Determining the duration of an impact, an event might have on a company in the context of Environmental Social and Governance (ESG) factors could be crucial for understanding and managing the risks or opportunities associated with that event.
Predicting the duration of an impact might involve fine-grained analysis of historical data, sentiment analysis, and other relevant information from news articles. In this chapter, we talk about our team LIPI's approach of solving the subtasks of ML ESG- 3. This can be the first step towards achieving the long-term goal of developing multilingual systems that can assess the potential short-term and long-term effects of specific events on a company's performance, reputation, or other ESG-related aspects. We present this in Figure 4.8.

Our contributions

Our contributions include developing a framework that fine-tunes pre-trained language models for classifying the impact and duration of an event associated with Multilingual news articles. We open-sourced the code[19] so that the research community can utilize them as baselines.

4.5.2 Problem Statement

The multilingual dataset of the shared task ML-ESG-3[20] consists of financial news articles in different languages such as English, French, Japanese, and Korean [183] [184]. The design of the task varies slightly across different languages. It is described as follows:

[19] https://github.com/Neel-132/ML-ESG3_LIPI
[20] https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-kdf-2024/shared-task-ml-esg-3 (accessed on 3rd Feb 2024)

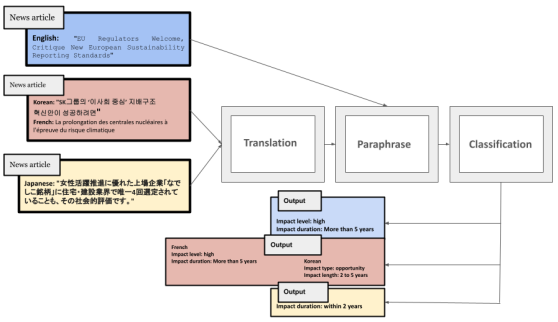

FIGURE 4.9: Proposed framework

- **English and French**: Given a financial news article in English or French, the objective is to determine its *impact level* and predict its *impact length*. The impact length can be "low", "medium" or "high". The impact duration can be "Less than 2 years", "2 to 5 years", or "More than 5 years".
- **Japanese**: Given a financial news article in Japanese, the objective is to predict its *impact duration*. The impact duration can be "Less than 2 years", "2 to 5 years", or "More than 5 years".
- **Korean**: Given a financial news article in Korean, the goal is to determine its *impact type* and predict its *impact length*. The impact type can be between "opportunity", "risk", or "cannot distinguish" and the impact length can be "less than 2 years", "2 to 5 years", or "more than 5 years".

4.5.3 System Descriptions

The pipeline for handling the tasks mentioned above comprises the following steps:

- Step 1: **Translation** - Although there are several powerful multilingual encoder models present, our experiments revealed that they were not very efficient in learning the intricate patterns in the dataset and thereby correctly predicting the impact type and duration of news articles. Thus, we primarily translated the non-English datasets into English before proceeding with modelling.
- Step 2: **Paraphrase** - We found that as the given dataset was small, the classification models were overfitting. To solve this, we paraphrased the translated dataset returned by the translation module as mentioned in Step 1 using a T5-based model [72].
- Step 3: **Classification** - After paraphrasing comes the final module of the pipeline. This is the classification module. Since the target variable differed slightly across different datasets, we designed two different classification modules for the three tasks given as follows:
  - Module 1 (for English, French & Korean): The English, French, and Korean dataset has two target variables. For English and French, they are *impact level* and *impact length*. For Korean, they are *impact type* and *impact length*. We used pre-trained encoder models like BERT [1], DistilBERT [172], etc. to learn the embeddings of the content as given by the paraphrase module, followed by a linear layer to predict the target which can be impact length, impact type, or

Dataset	Model	macro F-1	micro F-1
English	XGBoost	**0.35**	**0.31**
	SVM	0.29	0.26
	DNN	0.32	0.27
French	XGBoost	0.23	0.22
	SVM	0.21	0.21
	DNN	**0.33**	**0.34**
Japanese	XGBoost	**0.12**	**0.09**
	SVM	0.08	0.05
	DNN	0.11	0.10
Korean	XGBoost	0.34	0.34
	SVM	0.27	0.22
	DNN	**0.42**	**0.34**

TABLE 4.13: Result of the Baselines

impact level. The number of classes in each of these target variables is used as a hyperparameter to specify the output of the linear layer.

– Module 2 (For Japanese): The Japanese dataset has only one target variable, *impact duration*. The *impact type* was given for this dataset. So, we developed the second module to learn the pre-trained text embeddings using the same encoder models, but for two features which are news content and impact type, followed by a concatenation operation. Finally, we added a linear layer to predict the output.

We present this in Figure 4.9.

4.5.4 Experiments and Results

In this Section, we describe the experiments we performed, and the corresponding results.

4.5.4.1 Baseline

For the baseline, we chose BERT-base uncased (for English) and BERT-base multilingual [1] uncased (for other languages) to learn the pre-trained embeddings of news content and used them to train classical machine learning algorithms like XGBoost [71], Support Vector Machine [185], and deep learning based algorithms like Multi-layered Perceptron with just one hidden layer.
The results corresponding to it are presented in Table 4.13.

4.5.4.2 Experiment 1

The first experiment towards improving on the baseline had three stages, depending on the language of the dataset. For the non-English datasets like French, Korean, we firstly translated the news content into English using Google Translate. In the next step, we

Dataset	Model	macro F-1	micro F-1
English	BERT-base-uncased	**0.99**	**0.99**
	FinBERT	0.97	0.97
	DistillBERT-multiling	0.70	0.68
	DistillBERT-base	0.68	0.69
	NLI-Distilroberta-base	0.81	0.80
	Distilroberta Financial	0.75	0.78
	XLI Roberta base	0.83	0.81
	RoBERTa-base	0.98	0.97
Korean	BERT-base-uncased	**0.95**	**0.94**
	FinBERT	0.94	0.93
	DistillBERT-multiling	0.78	0.71
	DistillBERT-base	0.76	0.69
	NLI-Distilroberta-base	0.82	0.81
	Distilroberta Financial	0.67	0.64
	XLI Roberta base	0.75	0.71
	RoBERTa-base	0.96	0.93
French	BERT-base-uncased	0.93	0.93
	FinBERT	**0.94**	**0.93**
	DistillBERT-multiling	0.57	0.49
	DistillBERT-base	0.91	0.90
	NLI-Distilroberta-base	0.51	0.45
	Distilroberta Financial	0.47	0.46
	XLI Roberta base	0.63	0.67
	RoBERTa-base	0.91	0.92

TABLE 4.14: Results of Experiment-1

paraphrased each data point using a T5 based paraphraser [72] with a beam size of 5, temperature of 0.7, and repetition penalty of 10.
In the final step, we fine-tuned pre-trained encoder models like BERT [1], DistillBERT [172], RoBERTa [125], Fin-BERT [69], etc. for the task of classifying the news articles to their respective impact type/level. We used a learning rate of e^{-5}, and a weight decay of 0.005 and fine-tuned the models for 30 epochs. Our best-performing models were BERT-base-uncased [1] for English, RoBERTa [125] for Korean, and FinBERT [69] for French .
The results are presented in Table 4.14.

4.5.4.3 Experiment 2

Since the Japanese dataset had only one objective, i.e. to predict the *impact duration*, we used the *impact type* as another feature along with the news content. Like the first experiment mentioned above, we translated the data into English, followed by paraphrasing with the same model, and configurations as mentioned in Experiment 1. Finally, we fine-tuned pre-trained models mentioned in Experiment 1 for assessing impact duration of news articles in Japanese.
Furthermore, we concatenated the embeddings of news content and impact type followed by a linear layer before the final output layer. We used a learning rate of e^{-4} and a weight decay of 0.006 and trained the models for 30 epochs. Our top performing models were

Dataset	Model	macro F-1	micro F-1
Japanese	BERT-base-uncased	**0.67**	**0.69**
	FinBERT	0.55	0.52
	DistillBERT-multiling	0.52	0.48
	DistillBERT-base	0.36	0.32
	NLI-Distilroberta-base	0.51	0.48
	Distilroberta Financial	0.43	0.48
	XLI Roberta base	0.49	0.51
	RoBERTa-base	0.68	0.67

TABLE 4.15: Results of Experiment-2

Dataset	Model	macro F-1	micro F-1
English and French	BERT-base-uncased	**0.79**	**0.79**
	FinBERT	0.67	0.62
	DistillBERT-multiling	0.34	0.41
	DistillBERT-base	0.51	0.55
	NLI-Distilroberta-base	0.57	0.61
	Distilroberta Financial	0.47	0.45
	XLI Roberta base	0.51	0.50
	RoBERTa-base	0.76	0.76

TABLE 4.16: Results of Experiment-3

BERT-base-uncased [1], RoBERTA-base [125] and FinBERT [69].
The results are presented in Table 4.15.

4.5.4.4 Experiment 3

Since the English and French datasets had the same objective of predicting the impact level and impact length, we experimented with fine-tuning the pre-trained models (mentioned in both of the previous experiments) on the English dataset and testing them on the French dataset. The hyperparameters were the same as those of Experiment 2 and the results corresponding to it are mentioned in Table 4.16.

4.5.5 Conclusion

In this chapter, we share our approach for determining the duration of an event's impact on the company. We translated the non-English datasets into English and further paraphrased them before fine-tuning the encoder-based pre-trained language models on them. Our observations revealed the best performing models were BERT[1] for English and Japanese; RoBERTa [125] for Korean, and FinBERT[69] for French. We achieved a significant increase in performance with translation and paraphrasing. Finally, we proposed a unified framework for all the languages. Our team ranked 3^{rd} in both of the subtasks of the English dataset, 1^{st} in the first subtask (impact-length) and 8^{th} in the second subtask (impact-level) of the French dataset, 20^{th} in the first subtask (impact-length) and 13^{th} in the second subtask (impact-type) of the Korean dataset, and 11^{th} in the Japanese dataset.

Chapter 5

Informed Investing

With information overload constantly, it is essential to evaluate the authenticity of the information they are relying on. In this chapter, we describe how Natural Language Processing can be used to safeguard investors from misinformation.

5.1 Research Questions

- **RQ-5**: How to safeguard investors from misinformation?
    - **Relevant Contributions**: Detecting in-claim numerals in financial texts [14] [15] [16] [17], Detecting Exaggerated Numerals in Financial Texts[1], Estimating profitability and loss from financial social media posts [18], Deciding trustworthiness of social media posts by executives [19], Financial Argument Analysis [20].

[1] https://huggingface.co/spaces/sohomghosh/FENCE_Financial_Exaggerated_Numeral_ClassifiEr (accessed on 17th November, 2023)

5.2 Detecting In-claim Numerals in Financial Texts

5.2.1 Introduction

With the advent of the Internet and digitalisation, most financial services and investment platforms have moved online. Organizations publish their performance reports and brochure digitally. Earnings conference calls of executives get transcribed and preserved online. Most investors rely on this information to make investment decisions. Numbers present in such information may be claims or non-claims (i.e. facts). Facts are always true whereas claims may be true or false. It is expected that investors will rely only on facts and not be mislead by false claims. However, making such a distinction is not easy, especially for novice investors. Thus, we need to have an automated system that would be able to detect whether numbers in financial texts are claims (in-claims) or not (out-of-claims/facts). Figure 5.1 presents two instances. The number '23' present in the text "For the full year we continue to expect an adjusted effective tax rate of 23% to 24%" is a claim. The number '1.1' in the text "Free cash flow a really good start to the year at $1.1. billion." is not a claim.

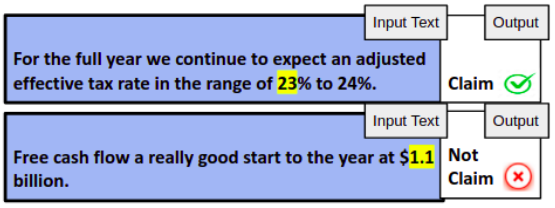

FIGURE 5.1: Claim detection in financial texts.

Our contributions

- We developed a system that can detect whether a numeral present in a given financial text is a claim or not. For this, we used the English version of the publicly available dataset FinNum-3 [186]. On the validation set, our system achieved macro F1-score is 0.8671.
- We studied how adding handcrafted features and information regarding the category of a target numeral affect the performance of the model.

This remaining chapter is structured as follows. In section 5.2.2 we discuss some of the existing works. We formally state the problem statement in section 5.2.3 and describe the dataset in section 5.2.4. In the subsequent sections 5.2.5, 5.2.6 and 5.2.7, we discuss the methodology, the experiments we performed and their results, respectively. Section 5.2.8 concludes and mentions some future work directions.

5.2.2 Related Works

Detecting claims from texts using Natural Language Processing (NLP) has been one of the trending areas of research. This has been applied in various domains like NEWS [187]

[188], Twitter [189], legal [190], etc. Hassan et al. [191] developed a system, ClaimBuster, to detect claims present in the 2016 US presidential primary debates. They evaluated ClaimBuster on statements selected for fact-checking by CNN and PolitiFact. They found that their system was able to detect several sentences with claims which were not selected for fact-checking by the above mentioned organizations. [188] created a new dataset by manually labelling the debates. They also proposed SVM and neural based systems to rank claims for prioritizing fact-checking. Subsequently, a similar application was presented by Konstantinovskiy et al. [192]. They used universal sentence representations for classification and outperformed existing claim ranking system [188] and ClaimBuster [191]. Furthermore, they proposed an annotation schema and a crowdsourcing methodology. This enabled them to create a dataset having 5,571 sentences with labels as claims or non-claims. Reddy et al. [187] released a new dataset NewsClaim which consisted of 529 manually annotated claims collected from 103 news articles mostly relating to COVID-19. They showed that zero-shot and prompt-based approaches perform well in detecting claims from news articles.

Aharoni et al. [193] developed a dataset for detecting claims in controversial topics. It consisted of 2,683 arguments which were collected from 33 controversial topics. Sundriyal et al. [194] proposed a novel framework called DESYR. It consisted of a gradient reversal layer and attentive orthogonal projection over Poincare embeddings. They evaluated it on informal datasets like online comments, web disclosures, Twitter, etc. Chakrabarty et al. [189] created a corpus from Reddit consisting of 5.5 million self-labelled claims which contain "IMO/IMHO (in my (humble) opinion)" tags. They fine-tuned ULMFiT [195] on this corpus. They further demonstrated how fine-tuning helped in argument detection tasks. Wright et al. [196] proposed a unified model called Positive Unlabelled Conversion. It consisted of a positive unlabelled classifier and a positive-negative classifier. They evaluated their model on three datasets namely Wikipedia citations, Twitter Rumours and Political Speeches.

Levy et al. [197] trained context-dependent classifiers for detecting claims on Wikipedia corpus. It primarily consisted of three components - Sentence Component, Boundaries Component and Ranking Component. Subsequently, Levy et al. [198] proposed an unsupervised framework to detect claims and evaluated its performance on the same corpus. Lippi et al. [199] used Partial Tree Kernels to generate features for detecting claims irrespective of the context. The inner nodes of these trees consisted of POS tags of the words in the leaf nodes. Furthermore, Lippi et al. [190] validated the effectiveness of this approach in the legal domain. To do this, they manually annotated claims from fifteen decisions of the European Court of Justice. Bar-Haim et al. [200] expanded the initial set of manually curated sentiment lexicons and added some contextual features (like headers, claim sentences, neighbouring sentences and neighbouring claims) to improve the existing claim stance classification systems. Botnevik et al. [201] proposed a browser-based extension BRENDA that helped users to verify facts within claims which are present in different websites.

Recently, with the increase in the availability of financial textual data, researchers have been focusing on detecting claims in financial texts as well [202] [203]. Chen et al. [203] presented a novel dataset NumClaim in Chinese which comprised financial texts, their categories and whether a target number within a text is in-claim or out-of-claim. They further proposed some neural architecture based baselines. Their best-performing model CapsNet resulted in a macro-F1-score 82.62% on the NumClaim Corpus. Recently, they released a similar dataset in English while organizing the FinNum-3 workshop [186].

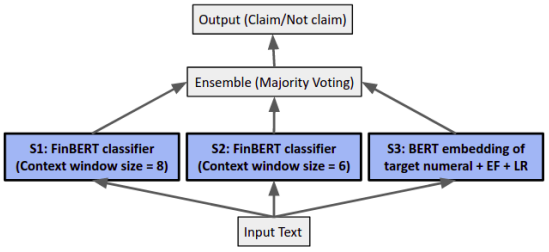

FIGURE 5.2: Methodology. EF = Engineered Features and LR = Logistic Regression

5.2.3 Problem Statement

Given a set F = {(t_1, n_1, s_1, e_1, c_1, m_1), (t_2, n_2, s_2, e_2, c_2, m_2) ... (t_k, n_k, s_k, e_k, c_k, m_k)} of k elements, the i_{th} element of F consists of a financial text 't_i', a number 'n_i' present within the text having starting and ending index positions 's_i' and 'e_i' respectively. Moreover, each element also contains c_i which denotes the category t_i belongs to and m_i which represents whether n_i is in-claim or out-of-claim. $m_i \epsilon \{0, 1\}$, 0 and 1 representing out-of-claim and in-claim, respectively. $c_i \epsilon$ {'date', 'other', 'money', 'relative', 'quantity absolute', 'absolute', 'product number', 'ranking', 'change', 'quantity relative', 'time'}. Our target is to develop a system for classifying an unknown numeral 'n'.

We evaluate the performance of our models using macro-averaged F1-score.

5.2.4 Dataset

Our experimental dataset comprises transcripts from earnings conference calls in English. They are formal financial documents. A similar dataset in Chinese consisting of reports written by analysts has been described in more detail [203]. Recently, a shared task, "NTCIR-16 FinNum-3: Investor's and Manager's Fine-grained Claim Detection"[2][186], is being held where participants are provided with this dataset. We registered in the shared task and obtained the training and validation data. The training data consists of 8,337 records whereas the validation data consists of 1,191 records. Of all these records the train and validation set has 1,039 and 114 in-claim instances respectively. There are 2,627 and 409 unique financial texts in the training set and validation set respectively. This indicates that most of the texts present in the training and validation sets have multiple numbers present in them. We present the category-wise distribution in Table 5.1.

5.2.5 Methodology

Our final system consists of an ensemble of 3 sub-systems. The first two sub-systems consist of fine-tuning pre-trained language model FinBERT [69] and are almost identical. The third one is a logistic regression based model built using contextual BERT embedding [1] of the numerals and other engineered features. BERT (Bidirectional Encoder Representations from Transformers) [1] is one of the state-of-the-art language models. It has

[2] https://sites.google.com/nlg.csie.ntu.edu.tw/finnum3/finnum-3 (accessed on 18th September, 2023)

Set	Category	Count
train	absolute	683
valid	absolute	85
train	change	398
valid	change	84
train	date	1616
valid	date	221
train	money	1496
valid	money	266
train	other	406
valid	other	106
train	product number	235
valid	product number	26
train	quantity absolute	1193
valid	quantity absolute	122
train	quantity relative	178
valid	quantity relative	45
train	ranking	35
train	relative	2089
valid	relative	236
train	time	8

TABLE 5.1: Category-wise distribution of the training and validation set

been pre-trained using masked language modelling (MLM) and next sentence prediction (NSP) objectives. We use the base and uncased version of it which consists of 768 hidden units, 12 attention heads and encoder blocks. It has a total of 110 million parameters and can be used to generate contextual embeddings of 768 dimensions. FinBERT [69] is a version of BERT which has been subsequently pre-trained on Financial text and fine-tuned for financial sentiment classification task. We fine-tune the FinBERT model even further for the text classification task to detect in-claim numerals. Since the given training set has multiple numbers that are present in the same text, we try to narrow down the context of the target numeral. For the first sub-system, we define context as 8 words before and after the numeral. For the second and third sub-system, we further narrow it down to 6 words around the numeral. The entire process is depicted in Figure 5.2.

5.2.5.1 Sub-system-1 (S1)

Firstly, we tokenize the financial texts and extract 8 words before and after the target numeral. We follow the standard method of fine-tuning a FinBERT model (768 dimensions) so that its [CLS] token learns to predict whether the target numeral is in-claim or out-of-claim. We run this model in batches of size 256 for 40 epochs with a learning rate of 0.00002. We consider a maximum of 64 tokens. Finally, we select the model which is tuned till 15^{th} epoch as it performs the best on the validation set (Macro F1-score = 0.8585).

5.2.5.2 Sub-system-2 (S2)

This sub-system is similar to the first one. The only differences are we narrow down the context around the target numeral from 8 to 6 and consider a maximum of 16 tokens. We do this to focus specifically on the target numeral. This model performs the best just after the 14th epoch (Macro F1-score = 0.8439).

5.2.5.3 Sub-system-3 (S3)

This sub-system is different from the previous two. In this sub-system, given a context window of 6 words, we first extract BERT base uncased embedding (768 dimensions) of the target numeral. Since we have been using subword tokenization, for many cases the target numerals resulted in more than one token. This is one of the drawbacks of transformer based models. It has also been mentioned by Wallace et al. in the paper [204]. To deal with such instances, we take the mean of the embeddings of all the constituent tokens. Moreover, being inspired by [205], [206], [207] and [208] we engineer several features from the target numerals. These features include -

- number of digits before the decimal
- number of digits after the decimal
- one-hot vectors of different categories extracted using Microsoft Recognizers for Text[3]
- one-hot vectors of parts of speech of the target numeral as well as the immediately preceding and succeeding words

Finally, we develop a logistic regression model which takes the embeddings and engineered features as input and predicts whether a given numeral is in-claim or out-of-claim. The hyperparameters[4] of the logistic regression model are: C=1.0, fit_intercept=True, intercept_scaling=1, max_iter=100, penalty=l2, solver=lbfgs, tolerance=0.0001. The macro F1-score of this model is 0.8318.

5.2.5.4 Final System

The final system is an ensemble model. It selects results from the three subsystems (S1, S2 and S3) using majority voting. The macro F1-score of this model is 0.8671.

5.2.6 Experiments

We performed the experiments in four phases as mentioned below.

[3] https://github.com/microsoft/Recognizers-Text (accessed on 18th September, 2023)
[4] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (accessed on 18th September, 2023)

5.2.6.1 Defining the context window

At first, while exploring the data we noticed that 1,867 and 285 financial texts from the training set and the validation set respectively had more than one target numerals. Thus, it was essential to define a context around the target numeral. We tried to extract the sentences in which the target numerals were present. This did not solve the problem as more than half of the data had multiple target numerals in a given sentence. We further tried to extract the portion of the text on which the target numeral was dependent using the dependency parser provided by spaCy[5]. However, the performance did not improve. Finally, we performed several experiments by varying the context window size from 2 to 10. The context window of size k means we consider k words before and after the target numeral. Context window of size 8 gave us the best results.

5.2.6.2 Exploring various embeddings and classification algorithms

We explored various ways to numerically represent texts starting from TF-IDF to sentence transformer [68] based embeddings generated using BERT [1], RoBERTa [125] and FinBERT [69]. We further trained several classifiers over it. These classifiers included Logistic Regression, Random Forest [66], XG-Boost [71], etc. The performances of these models were inadequate. Thus, we added several engineered features as mentioned in section 5.2.5. This improved the performance slightly but the improvement was not notably high.

5.2.6.3 Fine-tuning pre-trained transformer based models

We tried to fine-tune several variants of BERT [1] model for the task of classification. A FinBERT [69] model trained with batches of 256, for 15 epochs with a learning rate of 0.00002 gave the best performance. This model was trained on a context window of size 8.

5.2.6.4 Adding information regarding category and handcrafted features

We experimented by adding the categories to which the target numeral belonged as one hot vectors. We further engineered several features as mentioned in section 5.2.5.3. These actions improved the macro F1-score to 0.8315 and 0.8318 respectively.

5.2.6.5 Ensembling individual models

Finally, we tried to combine outputs of the individual models using majority voting. On combining the individual models which are described in section 5.2.5, the macro F1-score improved from 0.8585 to 0.8671.

[5] `https://spacy.io/` (accessed on 18th September, 2023)

5.2.6.6 Implementation Details

These experiments were performed in a node having Nvidia Tesla V100 GPU with 32 GB RAM. We used Python (3.7) for all the computations. The main libraries used here includes PyTorch[6], SentenceTransformers[7], pandas[8], NumPy[9], scikit-learn[10] and Microsoft recognizers-text-number.[11]

5.2.7 Results and Discussion

We present the overall results in Table 5.2. We observe that machine learning based classifiers built with TF-IDF (with ngrams ranging from 1 to 4 and ignoring terms with document frequency strictly lower than 0.0005) based features (Sl. No. 1 to 3) did not perform as well as those which were built with FinBERT embeddings as features (Sl. No. 4 to 6). We tried extracting the portion of the financial text on which the target numeral was dependent. We further fine-tuned a FinBERT model using only the words on which the target numeral was dependent. This did not yield any improvement in the model performance (Sl. No. 7, Macro F1-score = 0.7250). However, on adding handcrafted engineered features and using context words within a window of 6 for fine-tuning the FinBERT model, the Macro F1-score improved to 0.8244 (Sl. No. 8). On adding information relating to categories as one hot vectors the F1-score further improved to 0.8315 (Sl. No. 9). Details regarding models `S1, S2, S3` and their ensemble have already been mentioned in section 5.2.5. The ensemble model (Sl. No. 14) performed the best (Macro F1-score = 0.8671 on validation set, 0.8473 on the test set). This is a significant improvement over the existing baseline CapsNet [203] (Sl. No. 10, Macro F1-score = 0.5736 on the test set). The results on the test set have been provided by the organizers in the paper [186].

Next, we evaluate the performance of the ensemble model across different categories. We present this in Table 5.3. It is interesting to note that the model performs well for almost all categories except 'product number' and 'date'. This is because the training set did not have a single in-claim instance of the category 'date' and only 9 such instances of the category 'product number'.

5.2.7.1 Ablation Study

We conduct a detailed ablation study to assess the importance of each component present in the ensemble model. We present the results of this in Table 5.4. We observe that the ensemble model performs better than the constituent models. While testing the hypothesis that the ensembled model is better than `S1`, we obtained a p-value of 0.18. We modified `S3` by removing engineered features and considered only the largest sub-word token of the target numeral. It resulted in the reduction of macro F1-score. This proves the

[6] `https://pytorch.org/` (accessed on 18th September, 2023)
[7] `https://www.sbert.net/` (accessed on 18th September, 2023)
[8] `https://pandas.pydata.org/` (accessed on 18th September, 2023)
[9] `https://numpy.org/` (accessed on 18th September, 2023)
[10] `https://scikit-learn.org/stable/` (accessed on 18th September, 2023)
[11] `https://github.com/microsoft/Recognizers-Text` (accessed on 18th September, 2023)

Sl. No.	Model	F1-Macro
1	TF-IDF + LR	0.6345
2	TF-IDF + RF	0.6603
3	TF-IDF + XG-Boost	0.6646
4	FB + LR	0.7990
5	FB + RF	0.7763
6	FB + XG-Boost	0.7994
7	FB classifier (dependent text)	0.7250
8	FB classifier (CW=6) + EF	0.8244
9	FB classifier (CW=6) + category	0.8315
10	CapsNet (baseline) [203]	0.5736
11	`S1`	0.8585
12	`S2`	0.8439
13	`S3`	0.8318
14	**Ensemble `S1, S2, S3`**	**0.8671**

TABLE 5.2: Overall Results. LR = Logistic Regression, RF = Random Forest, FB = FinBERT, CW = Context Window Size and EF = Engineered Features

Category	Micro-F1	Macro-F1
product number	0.9615	0.4902
date	0.9864	0.4966
quantity absolute	0.9426	0.8002
money	0.8759	0.8214
relative	0.9788	0.8474
absolute	0.9059	0.8590
change	0.8810	0.8776
other	1	1
quantity relative	1	1

TABLE 5.3: Category wise performance of the ensemble model

effectiveness of every part of the final model. We further tried varying the context window size. We conclude that the context window of size 8 gives the best performance.

5.2.7.2 Qualitative Error Analysis

Subsequently, we performed a qualitative evaluation for instances where our model made wrong predictions. We present a sample of it in Table 5.5. We observe that more than 66% of the misclassified target numerals have a dollar ('$') symbol and 17% of them have a percentage ('%') symbol associated. Microsoft digit recognizer was able to effectively put these instances into categories 'currency' and 'percentage' respectively. Thus, creating a classifier to first predict the categories and then training separate classifiers for each category may have helped in achieving better performance.

Model	**Macro-F1**	**Micro-F1**
S1 (CW=8, only)	0.8585	0.6345
S2 (CW=6, only)	0.8439	0.6603
S3 (only)	0.8318	0.6646
S3 (-EF)	0.8238	0.7990
S3 (-EF, only largest token)	0.7934	0.7763
FinBERT classifier (CW=4)	0.8408	0.7994
FinBERT classifier (CW=5)	0.8318	0.7250
FinBERT classifier (CW=7)	0.8381	0.8244
FinBERT classifier (CW=9)	0.8247	0.8262
FinBERT classifier (CW=10)	0.8407	0.8585
Ensemble `S1, S2, S3`	**0.8671**	**0.9479**

TABLE 5.4: Ablation Study. CW = Context Window, EF = Engineered Features

paragraph	**numeral**	**target (claim)**	**predicted**
Fourth quarter free cash flow was $1.2 billion taking our full year cash generation to **$4.4** billion.	$4.4	0	1
On the synergies front Collins Aerospace team is on track to deliver approximately **$150** million of cost synergies this year and we still see at least $500 million of total cost synergy potential.	$150	0	1
On the synergies front Collins Aerospace team is on track to deliver approximately $150 million of cost synergies this year and we still see at least **$500** million of total cost synergy potential.	$500	0	1
Thanks Matt. Verizon's strategic priorities for 2019 are clear. I have outlined **5** priorities with my team that focus on our customers financial performance 5G leadership our valued employees and Verizon's role in creating benefits for our society.	5	1	0
Additionally United Technologies launched a hybrid-electric propulsion technology demonstrator program. This demonstrator program is expected to yield an average fuel savings for regional-sized aircraft of **30%** and we're targeting first flight within 3 years.	30%	1	0

TABLE 5.5: Qualitative Analysis of Misclassified Instances

5.2.8 Conclusion and Future Works

In this chapter, we introduced an ensemble based system to detect whether numerals in financial texts are in-claim or out-of-claim. This system consists of three sub-systems. Two of these sub-systems were created by fine-tuning FinBERT [69] on a context window of 8 and 6 words before and after the target numeral. The third sub-system is a logistic regression model. BERT based context embedding of target numeral and a few engineered features were used to train it. We conclude that adding hand-crafted features and information relating to category of the target numerals improves the performance slightly. However, training a model using only the portion of the text on which the target numeral is dependent, performs poorly. This is probably happening because the algorithm to extract dependent text is not yielding acceptable results. After conducting several experiments, we conclude the effectiveness of our model over the baseline CapsNet architecture [203].

In future, we would like to build a custom tokenizer that will tokenize other words into sub-tokens while keeping the target numeral as it is. We also want to experiment by changing the ensembling method from majority voting to a meta-classifier. Furthermore, a Convolutional Neural Network (CNN) or a Long Short Term Memory (LSTM) model trained using the context embeddings may yield better results.

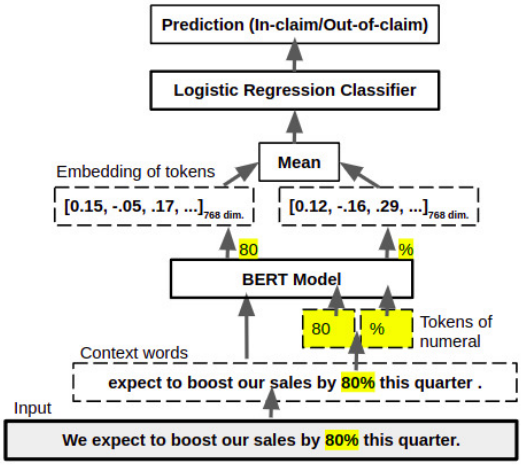

FIGURE 5.3: System Diagram of **FiNCAT**

5.2.9 FiNCAT: Financial Numeral Claim Analysis Tool

5.2.9.1 Introduction

Call transcripts, financial documents relating to stocks, funds and organizations enable investors to make data-driven investment decisions. However, to persuade the investors, narratives present in such documents may be just claims and not actual facts. Chen et. al released the NumClaim (Chinese) [203] and the NTCIR-16 FinNum-3 (English) [186] datasets in which the numerals present in the financial texts are annotated with in-claim and out-of-claim labels. We use the English dataset [186] to develop **FiNCAT** - a tool to analyse numerals present in financial texts.

Our contributions

- We developed a tool to automatically detect whether numerals present in financial texts are in-claim or out-of-claim. To the best of our knowledge, we are the first to develop such a tool.
- We have open-sourced[12] this tool as well as the embeddings and labels for further developments by the research community.

5.2.9.2 Experiments and Results

We initiated by exploring the "NTCIR-16 FinNum-3 (English): Investor's and Manager's Fine-grained Claim Detection" dataset [186]. The training and validation set had 8,337 and 1,191 records, respectively. Each of the target numerals in this dataset is labelled as in-claim or out-of-claim by experts. Most of these financial texts had more than one target numeral. We tried to define a context window around the target numeral by considering a

[12] https://github.com/sohomghosh/FiNCAT_Financial_Numeral_Claim_Analysis_Tool

Model	Training		Validation	
	F1-Micro	F1-Macro	F1-Micro	F1-Macro
BERT + LR	0.9698	0.9283	**0.9295**	**0.8223**
BERT + RF	0.9922	0.9826	0.9211	0.7869
BERT + GBM	**0.9996**	0.9992	0.9270	0.7738
BERT + LGBM	**0.9996**	0.9992	0.9286	0.8009
BERT + XGB	**0.9996**	0.9992	**0.9295**	0.8054
RoBERTa + LR	0.9478	0.8694	0.9261	0.8034
RoBERTa + RF	0.9681	0.9318	0.8992	0.7461
RoBERTa + GBM	**0.9996**	**0.9992**	0.9219	0.7248
RoBERTa + LGBM	**0.9996**	**0.9992**	0.9270	0.7699
RoBERTa + XGB	0.9993	0.9983	0.9244	0.7588

TABLE 5.6: Model Performance on Training and Validation sets (LR=Logistic Regression, RF=Random Forest, GBM=Gradient Boosting Machine, LGBM=LightGBM, XGB=XG-Boost)

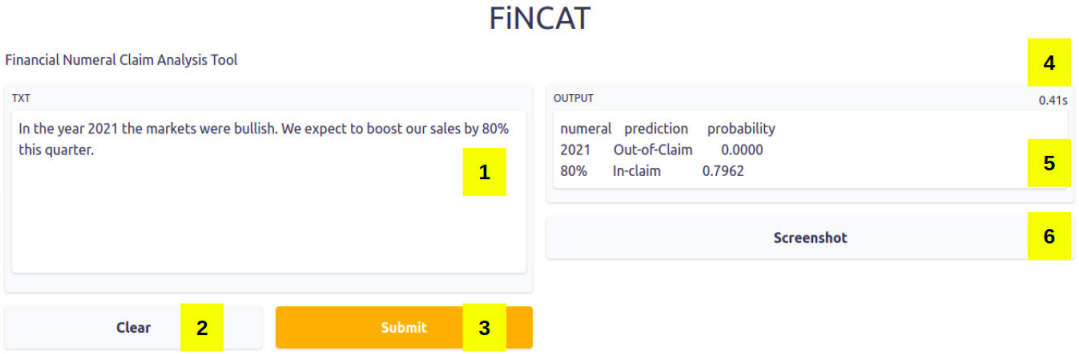

FIGURE 5.4: **FiNCAT**: Financial Numeral Claim Analysis Tool

certain number of words before and after it. We empirically decided to use 6 words before and after the target numeral as the context window.

We primarily experimented with two kinds of embeddings – BERT-base [1] and RoBERTa-large [125]. We extracted the mean of the embeddings of the constituent tokens of the target numeral given the words in the context window. We trained several machine learning models using the mean embeddings as features to detect whether the target numeral is in-claim or not. The models include Logistic Regression, Random Forest [66], Gradient Boosting Machine [67], LightGBM [70] and XG-Boost [71]. We kept the threshold at 0.5 and used F1-score for evaluation.

Analysing the results presented in Table 5.6, we finally decided to move ahead with the logistic regression based model trained using BERT [1] embeddings (768 dimensions). It performed the best and is more efficient, explainable than the others. We present the final architecture in Figure 5.3.

5.2.9.3 Tool Description

We deploy the tool using gradio[13] on Google Colab[14]. Figure 5.4 presents a screenshot of the tool. It comprises six parts: 1) **`input text box`**, 2) **`clear button`**, 3) **`submit button`**, 4) **`execution time`**, 5) **`output`**, and 6) **`screenshot button`**. The **`input text box`** takes any text as input. However, since this tool is specifically built for the financial domain, we recommend users provide texts related to finance like financial conversations, annual reports of organizations, etc. On pressing the **`submit button`** the tool looks for words in the input text which contains at least one digit. Each such word is evaluated using the model described in section 5.2.9.2. This consists of computing the mean of the contextual BERT [1] embeddings of the constituent tokens present in the target numeral. This mean (768 dimensions) of the contextual embeddings is used as features to score the Logistic Regression model. Finally, the tool presents the **`output`** in a tabular format consisting of three columns: i) numerals present in the input text, ii) prediction stating whether the numerals are in-claim or out-of-claim, and iii) probability predicted for each of them. The **`screenshot button`** and the **`clear button`** allows users to take screenshots and clear the entered text respectively.

We used Google Colab (free version CPU) to assess if it can detect in-claim numerals in real-time. We observed that the average time needed to generate predictions (**`execution time`**) for a given financial text consisting of 18 words and having 2 numerals is 0.25 seconds.

5.2.9.4 Conclusion

In this chapter, we present a tool, **`FiNCAT`**, which uses context-based embeddings and machine learning to detect in-claim numerals present in financial texts. Presently, it takes only text as input and checks for all the numerals present in the given text.

In the future, we want to take the target numeral as an input from the user. This is supposed to reduce the computational time. Further tuning of the hyperparameters of the tree-based models and threshold used for prediction may yield better results. Depending on the popularity we shall consider hosting it permanently using Hugging Face Spaces[15]. Another interesting direction for future research would be to explore different methods for generating the embeddings of the target numerals as a whole rather than taking the mean of embeddings of its constituent tokens.

[13] https://gradio.app/ (accessed on 18th September, 2023)
[14] https://colab.research.google.com/ (accessed on 18th September, 2023)
[15] https://huggingface.co/spaces (accessed on 18th September, 2023)

5.2.10 FiNCAT-2: An enhanced Financial Numeral Claim Analysis Tool

5.2.10.1 Introduction

Investors have moved online for performing financial activities of late. These activities include decision making based on financial content available over the internet. Organizations and executives try to persuade investors through claims which sound like facts. This brings in the need to develop a tool that is capable of automatically distinguishing 'in-claim' and 'out-of-claim' numerals present in financial texts. We developed a tool, **`FiNCAT`** [16] to address this. In this chapter, we present **`FiNCAT-2`** [209] which is an enhanced version of **`FiNCAT`**. The key enhancements include: i) adding more components to the user interface like examples and highlighting numerals along with the predicted types ii) hosting the tool permanently on HuggingFace Spaces iii) enabling an Application Programming Interface (API) for batch processing iv) minor bug fixes and, v) performance bench-marking by rigorous experimentation.

This tool has been trained on FinNum-3 (English) [186] dataset which consists of text transcripts from financial earnings conference calls. For a given numeral present in a financial text, this tool considers a context of six words before and after the numeral. Subsequently, it extracts 768 dimensional BERT embeddings [1] of the numeral given the context. For numerals consisting of multiple tokens, we consider the average of embeddings of the constituent tokens. We train a Logistic Regression model using these embeddings. For each numeral, it generates the predictions ('in-claim' or 'out-of-claim') along with their probabilities. The Macro F1-score of this model is 0.8223 on the validation set.

5.2.10.2 Related Works

Claim detection has been an interesting area of research across various domains like NEWS, Politics, Legal, etc. Some existing tools like ClaimBuster [210] and BRENDA [201] can be used to detect claims present in NEWS articles. Similarly, the system DESYR [194] works well for informal datasets like online comments, Twitter and so on. However, formal financial corpora have various nuances. To the best of our knowledge, there does not exist any such tool specifically for formal financial documents. This leads to the development of the tool **`FiNCAT-2`** exclusively for the financial domain.

5.2.10.3 Functionalities

As presented in Figure 5.5, the tool, **`FiNCAT-2`** consists of eight components. They are:
1) **`Input Textbox`**: It is used to take financial text as input from the user.
2) **`Clear Button`**: Using this button users can reset the entered text and output.
3) **`Submit Button`**: This button triggers the system to perform computation on the entered text.
4) **`Input Examples`**: We provide a few sample texts which can be used as inputs.
5) **`Execution Time`**: It displays the computing time.
6) **`Output 1`** (highlighted text): This is a text box consisting of 'out-of-claim' numerals in

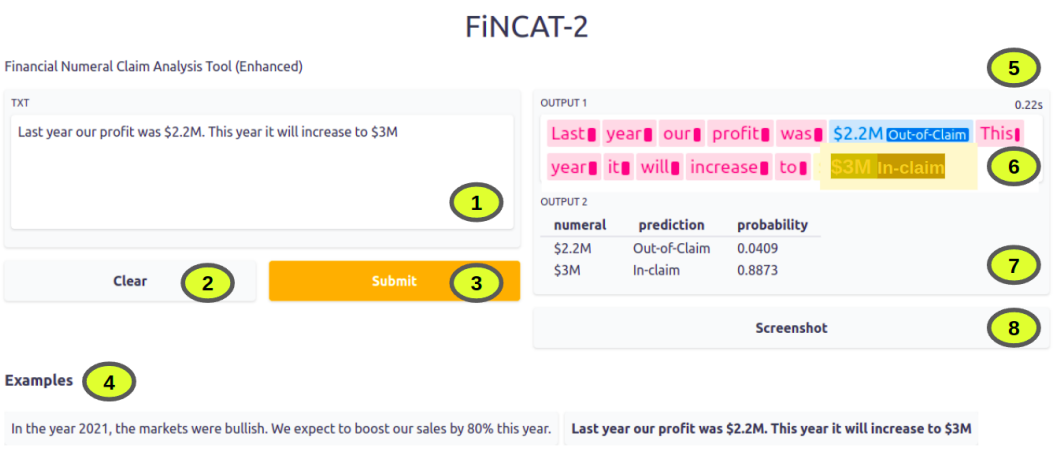

FIGURE 5.5: The User Interface of FiNCAT-2

blue, 'in-claim' numerals in yellow and other words in red.
7) **`Output 2`** (predictions in tabular format): It is a table having three columns: i) numeral ii) prediction denoting whether the target numeral is 'out-of-claim' or 'in-claim' numerals iii) probability for each prediction.
8) **`Screenshot button`**: This button enables user to take screenshots.
We collected 400 instances of financial texts having the number of numerals between 1 to 15 numerals and the number of tokens ranging from 7 to 230. We scored the model five times for each of these instances in Google Colab without using any hardware accelerator. The average time taken to score all the 400 instances was 114 seconds.

5.2.10.4 Impact overview

`FiNCAT-2` solves the problem of automatically detecting 'in-claim' numerals from financial texts. It has been hosted on HuggingFace Spaces[16] so it can be readily used by investors. To further increase its usability and reproducibility, we have also shared it as a Google Colab notebook[17]. We have open-sourced the tool on GitHub[18]. This tool has been very recently (Feb, 2022) launched. In less than one month, it has been cloned thrice on GitHub. We expect its popularity to grow over time among financial investors and researchers. Presently, it is suitable for non-commercial academic usage. We have made the training scripts available which can be effectively used to re-train the underlying model for commercial purposes. A list of scholarly publications enabled by the tool includes [16], [17], and [14].

[16] `https://huggingface.co/spaces/sohomghosh/FiNCAT_Financial_Numeral_Claim_Analysis_Tool` (accessed on 18th September, 2023)
[17] `https://colab.research.google.com/drive/1OEN48pPaEFAXiB972tYjC0Llfo0qrLcN?usp=sharing` (accessed on 18th September, 2023)
[18] `https://github.com/sohomghosh/FiNCAT_Financial_Numeral_Claim_Analysis_Tool/` (accessed on 18th September, 2023)

5.2.10.5 Limitations and Future Works

At present, the tool, **`FiNCAT-2`** checks whether each of the numbers present in a given financial text is 'in-claim' or 'out-of-claim'. This is computationally expensive. It would be nice to take the exact numeral which is to be checked as an input from the user. Another direction for extending this work further can be the creation of a web browser-based extension which will enable users to detect in-claim numerals present in financial web pages. Fine-tuning the underlying model and creating a Python package will be valuable contributions. A system capable of detecting claims by processing financial speeches in real-time will add a newer dimension to this research.

Sl. No.	Model	hyperparameters	F1 micro	F1 macro
1	LR	multi class='multinomial', solver='lbfgs', max iterations = 1000	0.543	0.509
2	NN	hidden_layer_sizes = [128], max iterations = 300	0.542	0.509
3	NN	hidden_layer_sizes = [512, 384, 256, 128, 64, 32, 16], max iterations = 300	0.558	0.527
4	NN	hidden_layer_sizes = [512, 256, 128, 64, 32, 16], max iterations = 300	0.569	**0.576**
5	**NN**	**hidden_layer_sizes = [512,128], max iterations = 300**	**0.581**	0.566

TABLE 5.7: Performance of different models. NN = Neural Networks, LR = Logistic Regression.

5.3 FENCE: Financial Exaggerated Numeral ClassifiEr

5.3.1 Introduction

Nowadays, investors tend to read financial contents which are available online while making investment decisions. These contents are generally blog posts, comments about the financial market, etc. Often, the numerals present in these contents are exaggerated and not correct. We present a tool to address these discrepancies. Firstly, we randomly sample five thousand instances each from market comments and blog titles present in the Numeracy-600K dataset [211]. We extract contextual embedding of the target numeral present in a financial text using SEC-BERT-NUM [151] and pass it through a neural network (NN) based classifier. Due to computational constraints, we train only the NN classifier after freezing the BERT [1] layers. We use 90% of the data for training and use the held-out 10% data as the test set. In Table 5.7, we present performance of various classification models on the test set. We select the NN having 512 and 128 neurons respectively in the hidden layers (Sl. No. 5 in Table 5.7) as the final model as it performs the best in terms of micro F1-score as it is less complex compared to others (SL. No. 3 and 4 in Table 5.7). Subsequently, we evaluated the performance of the selected model (Sl. No. 5) on real world datasets consisting of titles of 2,550 news articles, and we obtained a weighted average F1-score of 0.63. **`FENCE`** has been released on the code ocean platform [212] and can be accessed live from Hugging Face spaces[19]. A video demonstrating the functionalities of **`FENCE`** has been made available on YouTube.[20]

5.3.2 Related Works

Some of the tools that help users to apply Natural Language Processing in finance include [213], [23], and [15]. Liu et al. [213] developed an online tool, FIN10K for visualizing and analysing various 10K reports. We presented FiNCAT-2 [15] for detecting in-claim and out-of-claim numerals present in financial texts. We further released a comprehensive toolkit FLUEnT [23] which performs various natural language processing tasks on financial texts like readability assessment, hypernym extraction, etc. However, none of the existing tools solve the problem of exaggerated numeral detection in the finance domain. To the

[19] `https://huggingface.co/spaces/sohomghosh/FENCE_Financial_Exaggerated_Numeral_ClassifiEr` (accessed on 8th April 2023)

[20] `https://youtu.be/kXBflXu4EhQ` (accessed on 29th June 2023)

best of our knowledge, the **FENCE** tool is the first free, open-source and user-friendly tool to detect exaggerated numerals.

5.3.3 Tool Description

As with most other software tools, **FENCE** consists of a back-end and a front-end.

5.3.3.1 Back-end

As presented in Figure 5.6, in the back-end it replaces the target numeral present in financial texts with [NUM] token. Subsequently, for this numeral it extracts the contextual embedding of the [NUM] token using SEC-BERT-NUM [151] available on HuggingFace platform [140]. This embedding having 768 dimensions is passed through a neural network having two hidden layers with 512 and 128 neurons respectively. Finally, this network classifies the target numeral as Exaggerated or Non-exaggerated.

5.3.3.2 Front-end

We present the user interface of **FENCE** in Figure 5.7. The inputs are marked with numbers on yellow background, while the outputs are marked with numbers on orange background in Figure 5.7. Firstly, in the input text box `1`, users are supposed to enter financial texts and press the button marked as `2`. The tool highlights and shows the numerals present in the text along with their positions in the panel marked as `3`. Tab `4` helps users to choose whether the user wants to evaluate all or a few specific numerals present in the text. If they decide to evaluate only for specific numerals, they can get the numerals along with their positions in the text using button `5`. After this, they can use the multi-select option `6` to choose the specific numerals. On the other hand, for evaluating all the numerals, they can simply click on the "Predict for all numerals" button which appears on selecting the "All numerals" tab (Tab: `4`). For brevity, we do not show this button in Figure 5.7. Finally, on clicking this button or button `7`, the results get presented in panel `8`. For the convenience of the users, in the portion marked as `9`, the tool provides a few sample financial texts.

5.3.4 Impact Overview

The **FENCE** tool has been released very recently[21]. We expect that this tool will be readily used by investors and researchers to assess whether the numerals present in financial texts are exaggerated or not. For example, the numeral "30%" in the sentence "*Our sales will increase by 30% in the next year*" is not exaggerated. However, if "30%" is changed to "300%" it becomes exaggerated. Automatically detecting such exaggerated numerals helps financial analysts and investors in making decisions and staying away from false claims. Decisions made by expecting false claims to be true often lead to financial losses. **FENCE** uses SEC-BERT-NUM [151] embeddings at the back end. It has been developed using

[21]April 2023

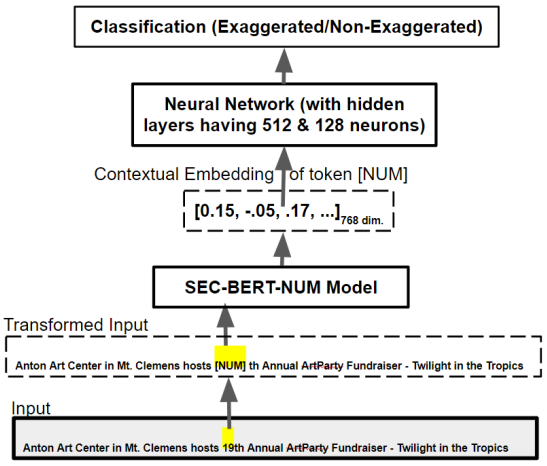

FIGURE 5.6: Architecture of **FENCE**

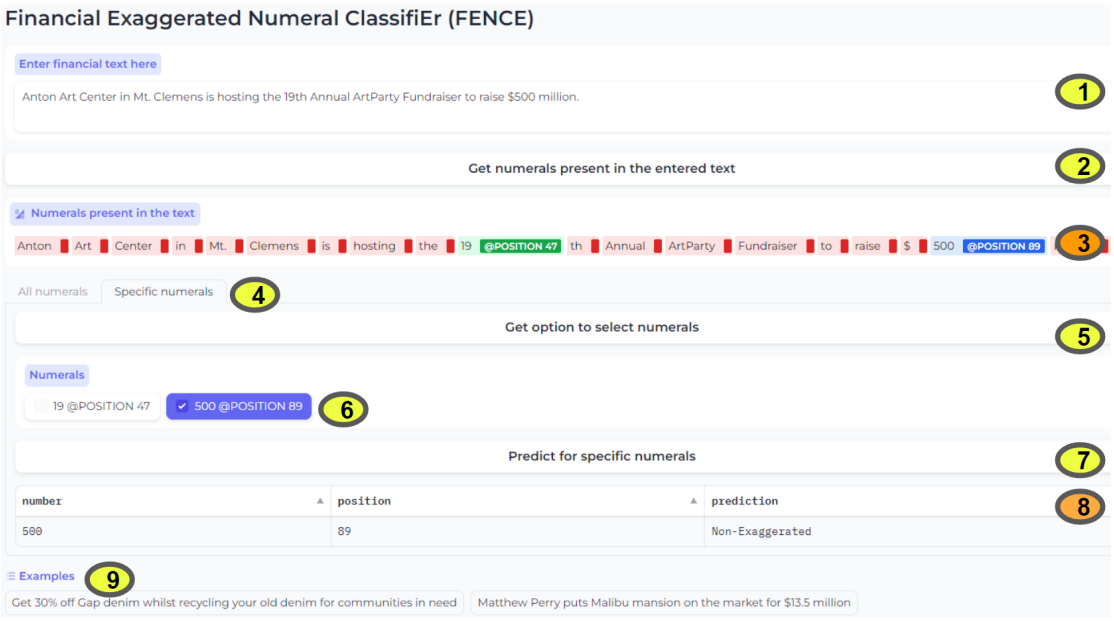

FIGURE 5.7: User Interface of **FENCE**

Gradio [174] and hosted on HuggingFace Spaces [140]. It has been open-sourced under the MIT licence. The **FENCE** tool does not require any installation. It can be used from any device with a Web browser and an Internet connection. It will help in spreading financial knowledge and thereby have a positive impact on improving financial literacy among users.

5.3.5 Conclusion

In this chapter, we introduced **FENCE**, a tool to detect exaggerated numerals from financial texts. As of now, the tool takes only one text as input, and that too in English. In future, we want to enhance its capability so that it can seamlessly process multiple texts in different other languages. Improving the model performance further by experimenting

with different architectures is another direction for future work. Based on the popularity of the tool, we shall consider releasing a browser based extension with additional capability to process audio files and scan Portable Document Format (PDF) files.

5.4 Estimating profitability and loss from financial social media posts

5.4.1 Introduction

Over the last few years, financial opinion mining has emerged to be an interesting area of research [214]. Several research studies (Mao et al. [215], Sprenger et al. [216], Lee et al. [217], Pagolu et al. [218], Asur and Huberman [219], Elliott et al. [220], Crowley et al. [221]) highlight the importance of social media posts for predicting stock markets. Although the wisdom of the crowd matters, it is still necessary to mine quality posts from the rest. Quantifying social media posts in terms of the expected profitability is an open area for research. Chen et al. [202] proposed two metrics: Maximum Possible Profit (**MPP**) and Maximum Loss (**ML**) for evaluating such posts. They recently hosted the FinNLP-2022 ERAI Task[22] (in conjunction with EMNLP-2022[23]). It comprises pairwise comparison (Task-1) and unsupervised ranking (Task-2) of financial social media posts with respect to **MPP** and **ML**. In this chapter, we describe our best-performing systems (Task-1 → **MPP**: 57.47% & **ML**: 59.77%; Task-2 → **MPP**:18.27% & **ML**: -3.90%).

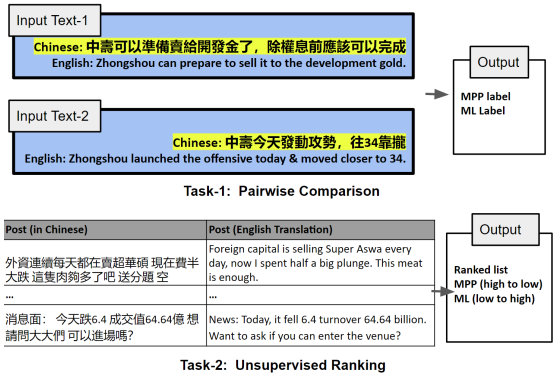

FIGURE 5.8: ERAI FinNLP-2022 Tasks

5.4.2 Problem Statement

For Task-1, given two posts, the task is to develop a system for evaluating which of them will lead to greater **MPP** and lower **ML**.
For Task-2, given a set of posts, the task is to develop a system for ranking these posts in terms of higher **MPP** and lower **ML** values.
Results of Task-1 were evaluated using accuracy. For Task-2, average **MPP** and **ML** values of top 10% posts were considered for evaluation.

[22] https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-2022-emnlp/erai-shared-task (accessed on 17th September 2022)
[23] https://2022.emnlp.org/ (accessed on 17th September 2022)

5.4.3 Datasets

The organizers initially provided the participants with two datasets. The first dataset (corresponding to Task-1) had 200 instances out of which 2 were null. We dropped the null instances from our experiments. Each instance consists of two posts (in Chinese as well as in English), their **MPP** and **ML** values, and labels corresponding to each post. In the dataset, the **ML** label is set to '1' for an instance (i.e., a pair of posts) when the **ML** value of the first post is less than that of the second post, otherwise the ML label is set to '0'. On the contrary, the **MPP** is set to '0' for an instance (i.e., a pair of posts) when the **ML** value of the first post is less than that of the second post, otherwise the **MPP** label is set to '1'. The posts in the dataset were collected from social media platforms like PTT[24] and Mobile01[25]. We refer to this as **D1**. For Task-2, a dataset consisting of 210 unlabelled posts (in Chinese as well as in English) were provided. This dataset is referred to as **D2**. **D2** serves as the test set for Task-2. Subsequently, the organizers released a test set consisting of 87 pairs of unlabelled posts (in Chinese and English) for pairwise comparison. We refer to this as **D3**.

Data Preparation

We created training and validation sets from **D1** maintaining a split ratio of 80:20. We extended **D1** in two ways.

Firstly, we treat each post from the pair individually, i.e., tuple (post-1, post-2, MPP-1, MPP-2, ML-1, ML-2) is converted into 2 tuples – (post-1, MPP-1, ML-1) and (post-2, MPP-2, ML-2). This gave us 320 instances for training and 80 for validation. We refer to this training set as **D4**. For sub-systems SB-1 (§5.4.3.2), SB-2 (§5.4.3.3) and SB-4 (§5.4.3.5), we used this set.

Secondly, we expanded **D4** by comparing each post to every other post after removing the null instances. It resulted in 97,032 instances of training. This is referred to as **D5**. The validation set was kept as it is. We use this in sub-systems SB-3 (§5.4.3.4) and SB-5 (§5.4.3.6).

Chen et al. [222] narrates the dataset and problem statement in more detail. The formulas for calculating **MPP** and **ML** are mentioned in [202]. In Figure 5.8, we present the problem statement and a sample dataset.

5.4.3.1 Sub-systems

Since our submitted systems are ensemble of multiple sub-systems, we explain each of the sub-systems here. More details regarding the hyperparameters of each sub-system are reported in the shared codebase.

[24] https://www.ptt.cc/index.html (accessed on 17th September 2022)
[25] https://www.mobile01.com/ (accessed on 17th September 2022)

5.4.3.2 Sub-System 1 (SB-1)

For all the Chinese posts in **D4**, we extracted the corresponding embeddings using sbert-chinese-qmc-finance.[26] We trained a linear regression model using the embedding as input to learn either **MPP** values or **ML** values based on requirements. We chose linear regression to start with as we did not have much data to train.

5.4.3.3 Sub-System 2 (SB-2)

This sub-system is similar to SB-1 (§5.4.3.2). The only difference is that we trained a neural network (multi-layer perceptron model) for 50 iterations instead of linear regression.

5.4.3.4 Sub-System 3 (SB-3)

For this sub-system, we used the **D5** dataset. For each pair of Chinese posts present in **D5**, we concatenated the embeddings for each of the posts obtained using sbert-chinese-qmc-finance[27]. We trained a linear regression model to learn the difference of either **MPP** values or **ML** values between each post present in a given pair.

5.4.3.5 Sub-System 4 (SB-4)

We customised the BERT model's architecture [1] for the task of regression such that its last layer learns to predict either the **MPP** values or the **ML** values. This was done by passing the representation of the [CLS] token through a fully connected linear layer having 128 neurons followed by a layer with *tanh* activation. We initialised it with the weights from the FinBERT model [69]. We used only the English posts present in **D4** for this.

5.4.3.6 Sub-System 5 (SB-5)

We extracted FinBERT [69] embeddings corresponding to all the English posts present in **D5**. We trained a multi-layer perceptron model for 500 iterations which takes this embedding as input and predicts the difference between either **MPP** values or **ML** values corresponding to each post present in a given pair.

5.4.4 best-performing Systems

In this section, we narrate the systems corresponding to our best-performing submissions.

[26] https://huggingface.co/DMetaSoul/sbert-chinese-qmc-finance-v1 (accessed on 17th September 2022)

[27] https://huggingface.co/DMetaSoul/sbert-chinese-qmc-finance-v1 (accessed on 17th September 2022)

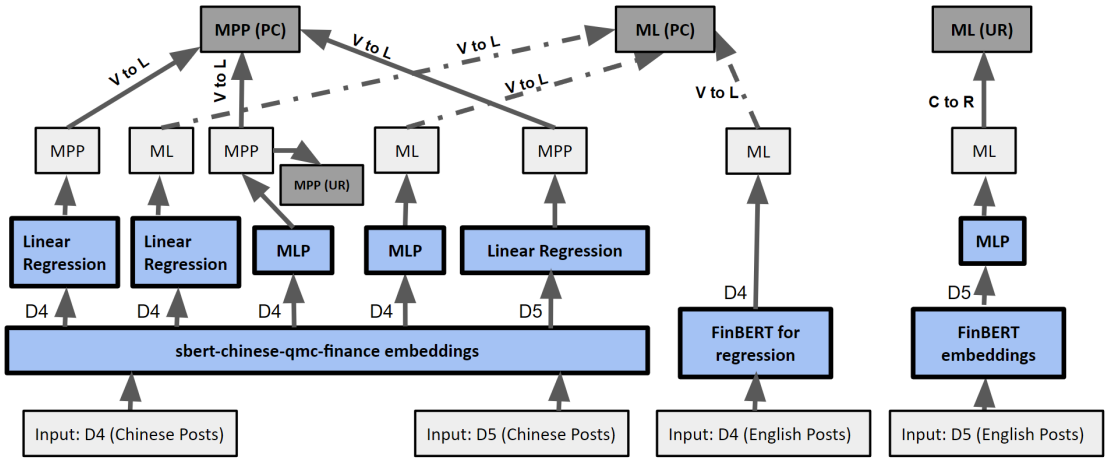

FIGURE 5.9: Ensemble Architecture. PC: Pairwise comparison, UR: Unsupervised Rankings, V to L: values to labels by comparison, C to R: comparison to rankings.

5.4.4.1 MPP calculation for Pairwise Comparison

This is an ensemble of three subsystems SB-1 (§5.4.3.2), SB-2 (§5.4.3.3) and SB-3 (§5.4.3.4). While SB-1 and SB-2 were trained with the objective of learning the **MPP** values, SB-3 was trained with the objective of learning the difference in **MPP** values for a given pair of posts. For SB-1 and SB-2, to obtain labels from raw **MPP** values, we computed and compared the **MPP** values of the posts constituting each pair in the test set. When **MPP** value of the first post was greater than **MPP** value of the second post, we assigned label '1', otherwise we assigned label '0'. For SB-3, we assigned label '1' when the predicted difference in **MPP** is greater than 0, otherwise we assigned label '0'. The final decision for the **D3** is made based on majority voting.

5.4.4.2 ML calculation for Pairwise Comparison

This system consists of selecting the final output from the predictions made by SB-1 (§5.4.3.2), SB-2 (§5.4.3.3) and SB-4 (§5.4.3.5) based on majority voting. Each of these constituent sub-systems were trained with the objective to learn the **ML** values. We scored each of these sub-systems on every post present in **D3**. Subsequently, we compared the raw **ML** values of posts constituting each pair in the test set. Label '1' was assigned when **ML** value of the first post was lesser than that of the second post, otherwise label '0' was assigned.

5.4.4.3 MPP calculation for Unsupervised Ranking

SB-2 (§5.4.3.3) was trained to predict the **MPP** value for a given post. We scored **D2** using SB-2 and ranked the posts in decreasing order of predicted **MPP** values.

		Train/Valid.		MPP (Pairwise Comparison)			MPP (Unsupervised Ranking)		
Sl.#	**Model**	**Data**	**Language**	**Train**	**Valid.**	**Test (D3)**	**Train**	**Valid.**	**Test (D2)**
1.1	SB-1	D4	Chinese	100.00%	70.00%	54.02%	8.04%	2.98%	11.83%
1.2	SB-2	D4	Chinese	62.18%	67.50%	48.28%	3.89%	2.45%	**18.27%**
1.3	SB-3	D5	Chinese	99.63%	60.00%	41.38%	-	-	17.46%
1.4	SB-4	D4	English	51.92%	47.50%	50.57%	2.11%	3.94%	4.17%
1.5	SB-5	D5	English	99.59%	45.00%	55.17%	-	-	16.63%
1.6	Ensemble (§5.4.4.1)	-	-	-	72.50%	**57.47%**	-	-	-

TABLE 5.8: MPP Results

5.4.4.4 ML calculation for Unsupervised Ranking

We trained SB-5 (§5.4.3.6) to learn the difference in `ML` values for a given pair of posts. We used this system to compare and sort the instances in `D2` in increasing order of predicted `ML` values.

Figure 5.9 gives a pictorial representation of all the ensemble models.

5.4.5 Experiments and Results

		Train/Valid.		ML (Pairwise Comparison)			ML (Unsupervised Ranking)		
Sl.#	**Model**	**Data**	**Language**	**Train**	**Valid.**	**Test (D3)**	**Train**	**Valid.**	**Test (D2)**
2.1	SB-1	D4	Chinese	97.44%	52.50%	50.57%	-10.26%	-2.16%	-7.81%
2.2	SB-2	D4	Chinese	57.69%	55.00%	50.57%	-5.55%	-8.01%	-5.56%
2.3	SB-3	D5	Chinese	99.65%	52.50%	47.12%	-	-	**-3.90%**
2.4	SB-4	D4	English	58.00%	50.00%	**59.77%**	-1.87%	-1.35%	-6.29%
2.5	SB-5	D5	English	91.24%	55.00%	44.83%	-	-	-4.11%
2.6	Ensemble (§5.4.4.2)	-	-	82.05%	57.50%	50.57%	-	-	-

TABLE 5.9: ML Results

This section states various experiments we performed and their results. We started with SB-1 which is a linear regression model trained over sentence embeddings. We tried financial sentence embeddings available for Chinese as well as the English language. Subsequently, we replaced the linear regression model with a multi-layer perceptron model. We further experimented by transforming the original training set `D1` to `D4` and `D5`. We also tried altering the last layer of the BERT [1] model for the task of regression. For the pairwise classification task, we used the regression models to get the `MPP/ML` values for each post in a pair. We then assigned a label to the pair by comparing these values as mentioned in §5.4.3. The results are presented in Tables 5.8 and 5.9. In this chapter, we focus on the best-performing systems among all our submissions due to page constraints. The other approaches we tried include classification of posts separated by *[SEP]* token using various variants of BERT [1]. Since the `D4` dataset consists of single posts, we use the same training and validation set for both the tasks. As the `D5` dataset comprises only of pairs of posts, we are unable to provide its performance in the unsupervised ranking task corresponding to the training and validation set. We ensembled models with varying lengths of the training set, therefore we do not report the performance of the model mentioned in §5.4.4.1 for the training set. Similarly, for the unsupervised ranking task, we do not report the performances of the models described in §5.4.4.1 and §5.4.4.2 as these models were suitable for pairwise comparison task only. The performance of the participating teams has been reported here[222].We used labelled instances from `D4` to assess the performance of the unsupervised ranking models as well. This helped us in choosing the best-performing models. As `D5` was suitable for pairwise comparison task

only, we could not use it to evaluate the models which were developed for the unsupervised ranking task. It is interesting to observe that our ensemble system's performance (Sl.# 1.6) is next only to that of team *Jetsons* in the pairwise comparison task using `MPP`. Moreover, in the same task using `ML` our subsystem SB-4 (Sl.# 2.4) performs as good as that of the best-performing team *DCU-ML* (accuracy: 59.77%). However, we did not submit this sub-system separately as it did not perform well on the validation set and submitted the results of the ensemble model (Sl.# 2.6) instead. In the unsupervised ranking using `MPP` task, only team *PromptShots*'s system performed better than that of ours (Sl.# 1.2). However, in the unsupervised ranking using `ML` task, the performance of the system developed by team *Yet* and the baseline solution were better than that of our systems (Sl.# 2.3 and 2.5). In this case as well we did not submit the result corresponding to SB-3 (Sl.# 2.3) where `ML` of top 10% post is -3.90% on the test set because the underlying system could not be evaluated on the validation set obtained from `D5`. We submitted results of SB-5 (Sl.# 2.5) instead.

5.4.6 Conclusion

Comparing the performance of our models with that of the other participants, we conclude that our models performed consistently well. We also observe that in most cases we achieve better performances using the Chinese texts than the translated version in English. This is because we are losing out on the nuances during translation. We further observe that ensembling helps in improving the overall performance.

Collecting more financial posts in a resource-rich language like English and incorporating prices of the stock whose `MPP` and `ML` are being discussed as input to the model are interesting directions for future work.

5.4.7 Limitations

The training dataset is very small in size and does not assure how the system will perform in real life. Fine-tuning large language models like BERT on `D5` is compute intensive. Moreover as the `MPP` and `ML` calculation differs for bullish and bearish market, it would be nice to take market conditions into consideration.

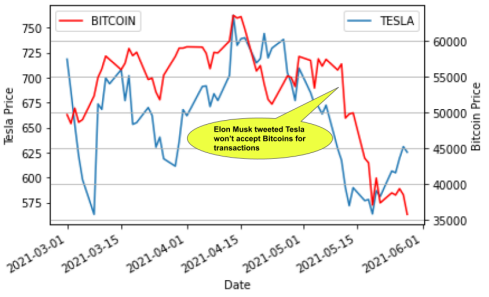

FIGURE 5.10: Elon Musk's tweets and its effect on stock prices

5.5 Deciding trustworthiness of social media posts by executives

5.5.1 Introduction

Real world outcomes are highly influenced by the opinion of people. Social media has become the top priority platform for people to share their opinion about products, services, movies, stocks etc. These opinions influence others' decisions and thought processes. Research [223] has shown that marketing and popularity of a product or stock is highly influenced by what the society and its people think and talk about it. This has given rise to 'meme stocks'. The world has witnessed how Elon Musk changing his Twitter bio to **'#bitcoin'** caused a hike in the price of bitcoin[28] as shown in Figure 5.10. In fact, Elon Musk's decision to buy $1.5 Billion of Bitcoin also caused the currency value to increase sharply [29]. This indicates to an underlying fact that the opinions of executives can bring changes in the real world. Motivated by this incident, we tried to find answers to the following three research questions.

- **RQ-1:** Does social media have any influence on close price movements of stocks over a longer period of time?
- **RQ-2:** Do opinions of executives on Twitter have greater influence on closing price of stocks than that of general people?
- **RQ-3:** How does Reddit fare compared to Twitter with respect to the task of close price prediction?

Initially, researchers believed that publicly available historical stock data is the only factor affecting the next day stock price. However, gradually people realized the power of social media and witnessed how opinion of executive people were affecting the stock market movements.

We designed our first experiment to validate the hypothesis that opinions expressed on social media have a deep influence on the close price of stocks and decentralised currencies. Performance improvements were observed on integrating sentiment mined from social

[28] https://www.blockchainresearchlab.org/2021/02/08/the-musk-effect-how-elon-musks-tweets-affect-the-cryptocurrency-market/ (accessed on 28th June, 2022)

[29] https://www.bbc.com/news/business-55939972 (accessed on: 28th June, 2022)

media data with historical stock data. The next set of experiments were performed to find out whether executive or general tweets have a deeper influence. Sentiment of general posts and sentiment of executive posts were separately integrated with historical stock data and different datasets were obtained on which we carried out the experiments. A better performance was witnessed on using sentiment of executive posts. Multiple experiments were performed on stocks (Tesla and Apple) and decentralised currencies (Bitcoin and Ethereum) to prove and validate the findings. Both Twitter and Reddit posts were considered for these experiments.

We made the following contributions in the chapter.

- We have validated that social media posts have an influence on close price movements. Subsequently, we have proposed how to use sentiments from social media to accurately predict the close prices.
- We have shown that opinions of the executives matter more than opinions of the crowd in predicting the close price movements.
- We showed that Reddit shows a similar trend like Twitter, however, Twitter is more effective in this task than Reddit.

Reproducibility: Our code has been open-sourced[30] so that researchers can leverage it for future research. To ensure reproducibility of our results we have released a dataset comprising of ids of social media posts which were used in this research. To comply with the terms and conditions of Twitter and Reddit, we could not share the text content of the social media posts.

To the best of our knowledge, we are the first to extensively study the effects of sentiments of tweets and Reddit posts on stock prices of listed companies (Apple, Tesla) as well as decentralised currencies (Bitcoin, Ethereum).

5.5.2 Related Works

Social media has become an integral part of our lives. We tend to express our thoughts and opinions by posting them on social media platforms like Twitter, Reddit, etc. Leskovec et al. [223] gathered information on how people converse regarding particular products and proved that it can be helpful in designing marketing and advertising strategies. Bollen et al. [224] did one of the pioneering work by aggregating mood signals from Twitter using and assessing how these mood signals correlated with one of the market index (Dow Jones Industrial Average) over time. They used OpinionFinder and Google-Profile of Mood States for the same. Traditionally, close price prediction of stocks were used to be done using methods like moving average, auto-regressive integrated moving average and so on. Presently, machine learning based algorithms have outperformed the traditional methods [225]. Vijh et al. [226] used Random Forest and Artificial Neural Networks for close price prediction. They collected historical data of five companies from Yahoo Finance for training these models. Using various performance metrics, they have proved that the Artificial Neural Network works better than Random Forest. However, they did not take into account the prevailing sentiments.

[30] https://github.com/datagodno/Evaluating-Impact-of-Social-Media-Posts-by-Executives-on-Stock-Price (accessed on 18th September, 2023)

Mao et al. [215] established that the number of daily tweets mentioning 'S&P 500' was correlated with its daily closing price and absolute change in price. They further proved this correlation is more for industries like Finance, Energy, Materials and Healthcare. Subsequently, they showed that the daily traded volume and absolute change in price of Apple Inc.'s stock was positively correlated with the number of daily tweets relating to Apple. Similarly, Sprenger et al. [216] showed how the positive sentiments and volumes of tweets were related to higher returns. These tweets were posted between 1st January and 30th June 2010.

Lee et al. [217] studied how the social media posts of corporates relating to product recall limited the harm on their firm's reputation. Pagolu et al. [218] have successfully established a correlation between stock data and Twitter data only for Microsoft stock. But, they have not explored the same for other stocks and platforms like Reddit. Asur and Huberman [219] used social media content to predict real-world outcomes like forecasting box-office revenues for movies.

Bartov et al. [227] focused on aggregated opinions of the people in general which is commonly referred to as wisdom of crowd. They proved that tweets of individuals could be used to forecast the earnings of an organization. On other side, Elliott et al. [220] and Chen et al. [228] emphasized on the importance of tweets by executives. Jung et al. [229] concluded that firms tweeted less regarding their financial when their earnings were poor. They further studied how tweets related to bad earnings tarnished organizations' images through media coverage. Jermann [230] used sentiment of executive tweets in predicting stock prices while ignoring that of the general people. Similarly, Elliott et al. [220] studied how tweets from CEOs after negative earnings helped in retaining investors. They concluded that CEOs who bonded with investors over Twitter gained were considered more trustworthy by the investors. Chen et al. [228] also presented similar findings. They studied how social media usage by executives of reputed firms impacted their stock prices and information environment. They further trained several machine learning models to classify executive tweets into three classes: company-related news announcement, work-related day-to-day activities and unrelated to-work (i.e. personal posts). However, they used a static list of negative words to assess the negativity of tweets. Seaton Kelton and Pennington [231] established that CEOs use social media platforms like Twitter to manipulate the investors. Crowley et al. [221] studied how the markets reacts to the tweets by executives and their firms during crucial business events.

Lately, Deshmukh et al. [232] used stock data of multiple companies and performed sentiment analysis of tweets using Vader [233] to predict close price. Chen et al. [234] presented an overview of various finance related opinion and argument mining techniques which are applicable on various sources like annual reports, earnings conference call, speeches, etc.

Xu and Cohen [235] proposed a neural-based model called StockNet for predicting rise or fall of various stock prices across nine industries. In addition to historical stock prices, this model used tweets for predicting the direction of movement of future stock prices.

Chen et al. [202] discussed how they used bi-directional Gated Recurrent Units [236] along with BERT [1] and Convolutional Neural Networks [237] for distinguishing social media posts of amateur from expert investment professionals. They further proposed two metrics, maximum possible profit (MPP) and maximum loss (ML) for quantitatively measuring the quality of these posts. Lastly, they released the Investor's ClaimRationale Dataset and proposed two tasks relating to rationale detection and claim-rationale inference.

Seroyizhko et al. [238] integrated sentiment of Bitcoin based on Reddit posts with Bitcoin stock data but did not achieve much improvements. They concluded that integration of social media information in the form of sentiment is still an open research.

Recently, Sawhney et al. [239] presented CryptoBubbles, a novel task of detecting market bubbles relating to crypto-currencies. They curated the dataset from Reddit and Twitter posts which related to crypto-currencies and meme stocks. Finally, they proposed a Multi Bubble Hyperbolic Network for solving this task.

5.5.3 Data

5.5.3.1 Data Collection

This section discusses how data was collected from three different sources. The procedures have been discussed below:

5.5.3.2 Twitter Data

Using snscrape[31], tweets about specific stocks were scraped using their stock tickers like **'TSLA'** (for Tesla), **'AAPL'** (for Apple), **'BTC'** (for Bitcoin), and **'ETH'** (for Ethereum). We refer to this scraped tweet dataset as dataset **T**. This dataset contains features like date, username, and tweet of both executive and general people from 1st January 2017 to 6th May 2022. A list of 122 executive Twitter handles was obtained from Forbes[32]. This list includes notable people like Elon Musk, Warren Buffett, etc. From dataset **T**, tweets of these executives were separated. Thus, two datasets were obtained, referred to as Dataset **E** and Dataset **G**, for executive and general tweets, respectively.

5.5.3.3 Reddit Data

Using pushshift.io[33] Reddit API, posts on particular subreddits were scraped. The dataset contains features like upvotes, date, posts and the subreddit. It is referred to as dataset **R**. According to Investopedia [34], there are subreddits that can influence the stock market. These subreddits include *'r/cryptocurrency'*, *'r/investing_discussion'*, *'r/robinhood'*, *'r/pennystocks'*, *'r/investing'*, and *'r/stock'*. Posts containing these executive subreddits were scraped to form a dataset, referred to as Dataset $\mathbf{E_r}$. Rest of the Tesla stock specific subreddits (*'tsla', 'TSLAtalk', 'teslainvestorsclub', 'TSLALounge', 'TSLAsexy', 'Tesla_Stock' and 'tslaq'*) were considered as general, and scraped to form a dataset referred to as Dataset $\mathbf{G_r}$.

[31] https://github.com/JustAnotherArchivist/snscrape, accessed on: 30th June, 2022
[32] https://www.forbes.com/sites/alapshah/2017/11/16/the-100-best-twitter-accounts-for-finance/?sh=783b0017ea0a, accessed on: 30th June, 2022
[33] https://github.com/pushshift/api, accessed on: 30th June, 2022
[34] https://www.investopedia.com/reddit-top-investing-and-trading-communities-5189322, accessed on: 30th June, 2022

Notation	Data Description
X (T)	Tweets relating to stock X
TSLA (R)	Reddit posts relating to TSLA
$\texttt{Y+T}_{vader}$	Closing prices & Vader based sentiment scores of all tweets
$\texttt{Y+T}_{finbert}$	Closing prices & FinBERT based sentiment scores of all tweets
Y+G	Closing prices & FinBERT based sentiment scores of general tweets
Y+E	Closing prices & FinBERT based sentiment scores of executive tweets
$\texttt{Y+G}_\texttt{r}$	Closing prices & FinBERT based sentiment scores of general reddit posts
$\texttt{Y+E}_\texttt{r}$	Closing prices & FinBERT based sentiment scores of executive reddit posts

TABLE 5.10: Notations and descriptions of the corresponding datasets

5.5.3.4 Historical Stock Data

Using Yahoo Finance[35], we obtained the historical stock data separately for each company stock or decentralised currency from 1st January 2017 to 6th May 2022. This dataset contains the features – 'open': the share price of a single stock at the start of the day, 'high': the highest price at which the stock was sold on that day, 'low': the lowest price the stock was sold on that day, 'volume' : total number of shares that were sold or bought on that day, and 'close': the closing price of a single stock on that day. Our objective was to build a model that can predict the 'close price shifted': the close price of the next day. This dataset is referred to as Dataset **Y**.

Notations

Table 5.10 presents a list of notations used in this chapter and their descriptions.

5.5.3.5 Exploratory Data Analysis

Tweets relating to Tesla, Apple, Bitcoin, and Ethereum were collected. The collected tweets were made by some executives and largely by general people. The total number of executive posts that were collected was 4,470 and the total number of non-executive tweets that were collected was 1,207,144. Since, tweets made by general people overshadow executive tweets, we perform under-sampling of the majority class. We limit the general posts to approximately 19,000 tweets for every stock. In case of Yahoo Finance, data were scraped from 1st January, 2017 to 5th May, 2022. However, no data was available in the weekends for different stocks and no data was available for Ethereum for the first 10 months of 2017 for Ethereum. This resulted in different counts of days with closing prices. We present the stock-wise statistics in Table 5.11. Figure 5.11, shows tweets made by executives and general users per day. The green line corresponds to executive posts and the blue line corresponds to general posts in Figure 5.11.

[35] https://finance.yahoo.com/, accessed on: 30th June, 2022

Stock	# Days with closing prices	Category	# Posts	Reduced # Posts	# Days with Posts	# Days with no Posts
TSLA (T)	1,346	Executive	2,617	NA	769	577
		General	4,82,375	19,164	1,346	0
AAPL (T)	1,346	Executive	260	NA	179	1,167
		General	21,383	19,057	1,324	22
BTC (T)	1,894	Executive	303	NA	242	1,652
		General	5,38,442	19,022	1,894	0
ETH (T)	1,543	Executive	51	NA	46	1,497
		General	1,53,362	19,091	1,535	8
TSLA (R)	952	Executive	1,239	NA	98	854
		General	11,582	NA	558	394

TABLE 5.11: Distribution of Social Media Posts. # represents number.

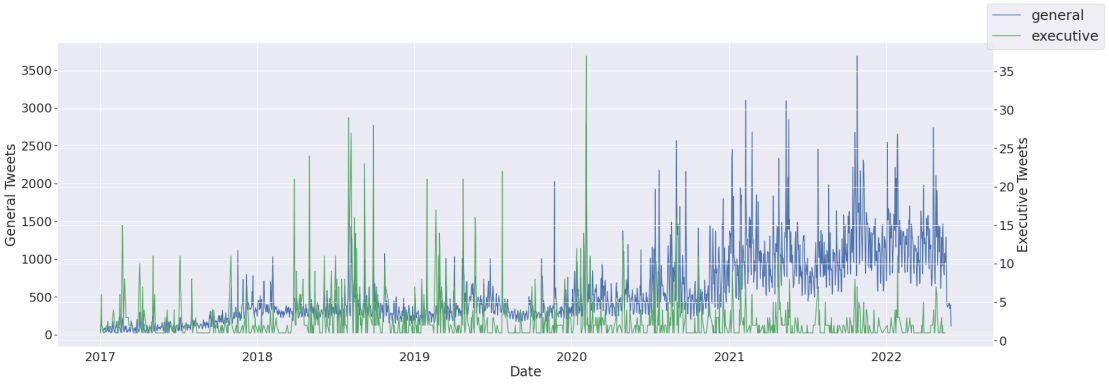

FIGURE 5.11: Number of executive and general tweets per day

5.5.4 Data Pre-processing

The tweets (i.e., datasets **E** and **G**) as well as Reddit posts were subjected to similar pre-processing steps. While extracting tweets, retweeted tweets were also considered as unique (i.e., separate) tweets. To avoid duplication, the duplicate tweets for every user were dropped. Every day thousands of people tweets relating to a given stock. To understand the sentiment associated with a given stock for a particular day, we extracted sentiments from all the tweets mentioning the stock in that day and averaged them. We used Vader [233] and a pre-trained language model FinBERT [69] for obtaining scores corresponding to three different types of sentiment – Positive, Negative, and Neutral. For FinBERT [69], these scores were normalized using the softmax function. For each day, we considered the average sentiment scores of all the tweets on that day. The day wise aggregated sentiment scores were aligned to the dataset **Y**. Dates for which no tweets were available, the sentiment scores of those dates were imputed using Cubic Spline Interpolation[36] technique for both executive and non-executive missing sentiment scores. A simple average method was not chosen since that approximation would be biased and much different from the actual value. Large language models like FinBERT [69] need high computational resources for training and scoring. Due to computational constraints, we considered a random sample of around 19,000 general tweets. Table 5.11, shows the number of days with no posts per stock for executives as well as general people. Unlike the decentralised currencies (Bitcoin and Ethereum), closing prices of listed companies (Apple and Tesla) are not available during the weekends. Subsequently, based on the availability of data we had to adjust the starting date for our analysis. Thus, the number of days with closing prices is different for different

[36] https://pythonnumericalmethods.berkeley.edu/notebooks/chapter17.03-Cubic-Spline-Interpolation.html, accessed on: 5th July, 2022

Stock	Category	Start	End
TSLA (T)	Train	4th, Jan 2017	14th, Apr 2021
	Test	15th Apr, 2021	5th May, 2022
AAPL (T)	Train	4th Jan, 2017	27th Apr, 2021
	Test	28th Apr, 2021	5rd May, 2022
BTC (T)	Train	2nd Mar, 2017	22rd Apr, 2021
	Test	23th Apr, 2021	6th May, 2022
ETH (T)	Train	16th Feb, 2018	2nd Jul, 2021
	Test	3rd Jul, 2021	6th May, 2022
TSLA (R)	Train	30th Jul, 2018	4th Aug, 2021
	Test	5th Aug, 2021	5th May, 2022

TABLE 5.12: Train and Test splits

stocks. Wherever we did not need to under-sample the number of posts, we mark it as NA (i.e. Not Applicable).

To predict close price more accurately, researchers [226] have introduced new variables which are derived from existing variables. In the present work we considered the exponentially weighted moving average (*ewma*) for the closing price and the sentiment scores for 3, 7, 14 and 30 days The intuition behind using four different ewma values is to cover all sudden and long-term changes to the price and sentiment of the stock. The derived dataset thus has more features: closing price, ewma closing prices, volume, open price, high, low, sentiment scores (corresponding to positive, negative and neutral classes) and ewma of sentiment scores. The goal of this work is to predict the close price of the next day, therefore a close price shift was added as the target variable for prediction. The dataset after being normalised using Standard Scaler was divided into training and test sets maintaining a ratio of 80% to 20%. The initial date range corresponding to the training data and the later date range corresponding to the test data are mentioned in Table 5.12. The starting dates are different for different stocks since missing posts right at the beginning of the date range could not be imputed. Imputation only works when data is missing in between the date range.

5.5.4.1 Experimental Setup

This section discusses the architecture and setup of the various models that we used in our experiments. Due to the sequential nature of the data, we primarily used sequence-based models such as Recurrent Neural Networks (**RNN**) [240], Gated Recurrent Unit (**GRU**) [241], Long Short Term Memory (**LSTM**) [120] and Auto Encoders (**AE**) [242]. We initiated by creating an **RNN** model. This was initialised by Glorot Normal [243] values. To generate the same random weights every time, the seed value was set to 42. It had three sequential RNN layers with a dropout [244] rate of 0.4 in each layer. The layers had 250, 200 and 150 neurons respectively. Subsequently, an output dense layer of a single neuron with a linear activation function was added. Adam optimiser [245] with a learning rate of 0.0001 was used. Mean Squared Error was used as the loss function. The model was run for 250 epochs with a batch size of 16 and a validation split of 0.1. Early stopping was performed with a patience value of 5 and the best weights were restored. Keeping everything else unaltered, we replaced the RNN layers by bi-directional RNN

(**Bi-RNN**) [246] layers in the above experiment. We further repeated the same experiment by replacing the RNN layers with **GRU** [241], bi-directional GRU (**bi-GRU**), **LSTM** [120] and bi-directional LSTM (**bi-LSTM**) layers. Lastly, we trained an Auto Encoder model (**AE**) consisting of a bi-directional LSTM layer with 250 neurons, tanh activation function, and a drop out rate of 0.4. This layer was followed by another LSTM layer with 200 neurons, a repeat vector layer, another two LSTM layers with 200 and 250 neurons each and drop out rate of 0.4 and 0.3 respectively. Finally, we added a flatten layer with a dropout rate of 0.4 and a dense layer with linear activation function.

Performance Metrics

We used MAE (Mean Absolute Error), RMSE (Root Mean Square Error), Adjusted R^2 (R^2_a) and MAPE (Mean Absolute Percentage Error) to evaluate the models. Among these metrics MAE, RMSE and MAPE are error metrics, i.e., the lower the value the better, while R^2_a is an accuracy metric.

5.5.5 Experiments and Results

To address the research questions, we carried out multiple experiments which are discussed in the following subsections. All of the experiments were performed in Google Colab with GPU.

5.5.5.1 Effect of Social Media Sentiment on Prediction of Close Price (Experiment 1)

This experiment was performed to answer RQ1, i.e., to investigate whether social media sentiment about a particular stock contributes to predicting its close price. For this study we chose the Tesla as the stock, twitter as the social media, and VADER [233] and FinBERT [69] tools as the sentiment analysis models. We used the historical stock data of Tesla from Yahoo (**Y**), and sentiment scores obtained from VADER and FinBERT on tweets (**T**) about Tesla from Twitter. Sentiment analysis was performed on dataset **T** using two sentiment analysis models, VADER – a rule based system, and FinBERT – a pre-trained model built by finetuning the BERT language model in the finance domain for performing sentiment analysis of financial text. After sentiment analysis, we obtained two different datasets, $\mathbf{T}_{vader}$ and $\mathbf{T}_{finbert}$. We obtained the scores corresponding to every type of sentiment ('positive', 'negative' and 'neutral'). These two datasets were merged according to dates with dataset **Y** giving rise to two datasets – $\mathbf{Y}$+$\mathbf{T}_{vader}$ and $\mathbf{Y}$+$\mathbf{T}_{finbert}$, having 1,346 instances of 24 features each. These 24 features consist of 5 features from the original dataset **Y** ('open', 'high', 'low', 'close' prices & 'volume' traded), 3 scores corresponding to sentiments (positive, negative & neutral) and 16 derived features (i.e. exponentially weighted moving averages for 3, 7, 14 & 30 days of 'close' price, positive, negative & neutral sentiment scores). The dependent variable (output) is the close price of the next day. Model trained on **Y** serves as our baseline model. The LSTM model was trained on these 3 datasets - **Y**, $\mathbf{Y}$+$\mathbf{T}_{vader}$ and $\mathbf{Y}$+$\mathbf{T}_{finbert}$, separately, and the results of these experiments are reported in Table 5.13. Results in Table 5.13 shows significant

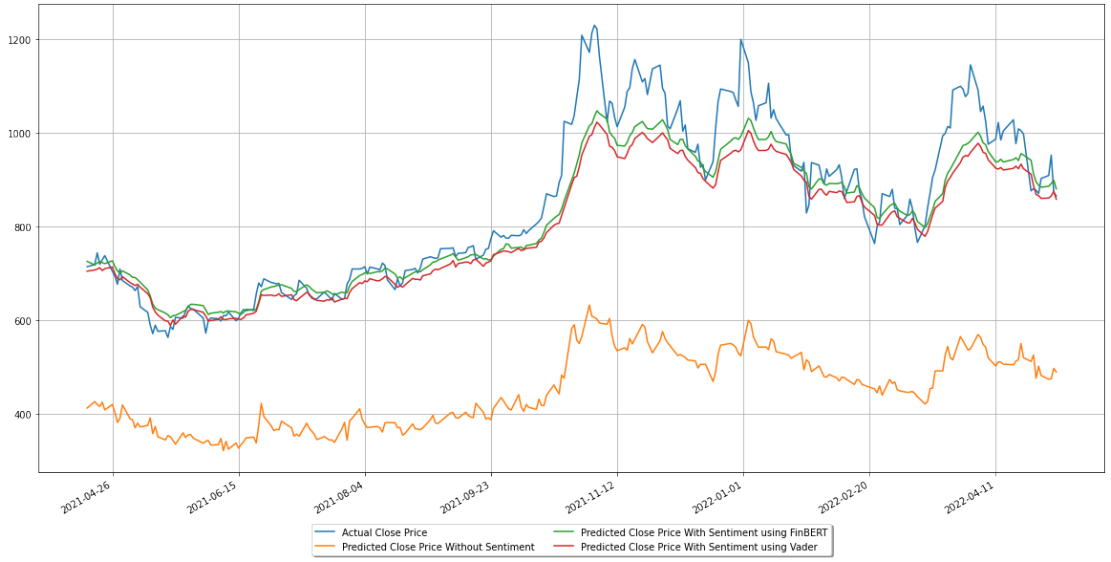

FIGURE 5.12: Close price prediction of Tesla with Y, Y+T$_{vader}$ and Y+T$_{finbert}$ datasets using LSTM

Dataset	MAE	RMSE	R_a^2	MAPE (%)
Y	393.195	405.845	-4.434	46.193
Y+T$_{vader}$	52.089	72.794	0.825	5.493
Y+T$_{finbert}$	**42.186**	**60.739**	**0.878**	**4.506**

TABLE 5.13: Results of Experiment 1

improvement in performance across all evaluation metrics for both **Y+T**$_{vader}$ and **Y+T**$_{finbert}$ over the baseline model **Y**. This proves the phenomenal fact that the sentiment of social media data has immense influence in predicting the close price and it comprehensively answers RQ1. The experimental results further suggest that FinBERT provides much better performance with respect to this extrinsic evaluation, hence FinBERT is used for sentiment analysis in all the experiments henceforward. Figure 5.12 pictorially presents the results of the prediction models. It shows that the curves obtained with the sentiment obtained using FinBERT (green) and the sentiment obtained using VADER (red) are much closer to the actual close price (blue) than the curve obtained without sentiment (orange).

5.5.5.2 Comparative Study of Models Predicting Close Price (Experiment 2)

Using the **Y+T**$_{finbert}$ dataset, we experimented with multiple models to find the best working model on this task. As the supremacy of neural networks over traditional methods like ARIMA [225] has been well-established for time series analysis, we tried various neural network-based architectures such as **RNN** [240], **GRU** [241], **LSTM** [120] and Auto-Encoder [242]. Table 5.14 presents the performance of these models on the **Y+T**$_{finbert}$ dataset. Among all these models, **GRU** provides the best working model across all evaluation metrics. Hence, in all further experiments **GRU** is used as the close price

Model	MAE	RMSE	R_a^2	MAPE (%)
RNN	184.194	233.086	-0.889	19.341
Bi-RNN	165.268	203.304	-0.437	18.141
GRU	**26.688**	**36.388**	**0.953**	**3.061**
Bi-GRU	30.509	42.166	0.938	3.478
LSTM	78.982	105.288	0.614	8.542
Bi-LSTM	87.046	107.885	0.595	9.399
AE	57.885	79.189	0.781	6.155

TABLE 5.14: Results of Experiment 2 on the $\texttt{Y+T}_{finbert}$ dataset

prediction model. This model follows the architecture of the **GRU** Model mentioned in section 5.5.4.1.

5.5.5.3 Influence of Executive Posts vs General Posts on Closing Prices (Experiment 3)

This set of experiments were carried out to answer RQ2, i.e., whether opinions of executives have greater influence on closing price than that of general people. Firstly, we use sentiment scores of tweets about Tesla from Twitter and historical stock data of Tesla from Yahoo. Two datasets were used, dataset **E** and dataset **G**, for executive and general posts, respectively. These datasets were subjected to sentiment analysis. Then they were merged according to dates with dataset **Y**. Thus, we had two new datasets, **Y+G** and **Y+E** having 1,346 instances of 24 features each. The output is a single feature, i.e., the next day close price. The **GRU** Model was trained on these datasets individually and next day close price was predicted. We refer to this as Experiment 3.1. We extended this experiment and replicated the same experiment by using tweets and stock prices of Apple instead of Tesla. We refer to this as Experiment 3.2.

We further extended our experiments to two unlisted decentralised currencies: Bitcoin and Ethereum. Unlike Tesla and Apple, these currencies do not depend on supply chain related factors like the availability of raw materials. Experiments 3.3 and 3.4 were carried out on Bitcoin and Ethereum datasets, respectively, keeping the same experimental framework, i.e., using the sentiment of posts about those currencies from Twitter and the corresponding historical stock data from Yahoo.

Finally, to answer RQ3, i.e., whether the above findings obtained using tweets also hold for Reddit, we repeated the same experiment using the sentiment of posts about Tesla on Reddit. This experiment is referred to as Experiment 3.5. This gives us a more comprehensive view of the bigger picture across two different social media platforms. This experiment has not been repeated with Apple, Bitcoin or Ethereum because of the tedious process of data collection which keeps failing multiple times due to payload.

Table 5.15 reports the results of Experiments 3.1–3.5. Results of Experiments 3.1–3.5 clearly suggest that opinions of executives matter much more than opinions of general people in the close price prediction task since the **Y+E** and $\texttt{Y+E}_\texttt{r}$ datasets provide much better performance than the corresponding **Y+G** and $\texttt{Y+G}_\texttt{r}$ datasets respectively across all the evaluation metrics. Since the trend holds true across both Twitter and Reddit and

Exp	Stock	Data	MAE	RMSE	R_a^2	MAPE (%)
Exp 3.1	**TSLA (T)**	**Y+G**	56.365	79.701	0.790	5.852
		Y+E	**34.362**	**48.261**	**0.923**	**3.817**
Exp 3.2	**AAPL (T)**	**Y+G**	4.700	5.922	0.859	2.932
		Y+E	**3.075**	**3.860**	**0.940**	**1.990**
Exp 3.3	**BTC (T)**	**Y+G**	4681.737	5584.437	0.572	9.675
		Y+E	**2842.190**	**3679.708**	**0.814**	**5.830**
Exp 3.4	**ETH (T)**	**Y+G**	315.075	407.849	0.641	8.663
		Y+E	**278.507**	**356.633**	**0.725**	**7.724**
Exp 3.5	**TSLA (R)**	**$Y+G_r$**	44.625	61.403	0.811	4.481
		$Y+E_r$	**42.283**	**58.994**	**0.826**	**4.247**

TABLE 5.15: Result of Experiments 3

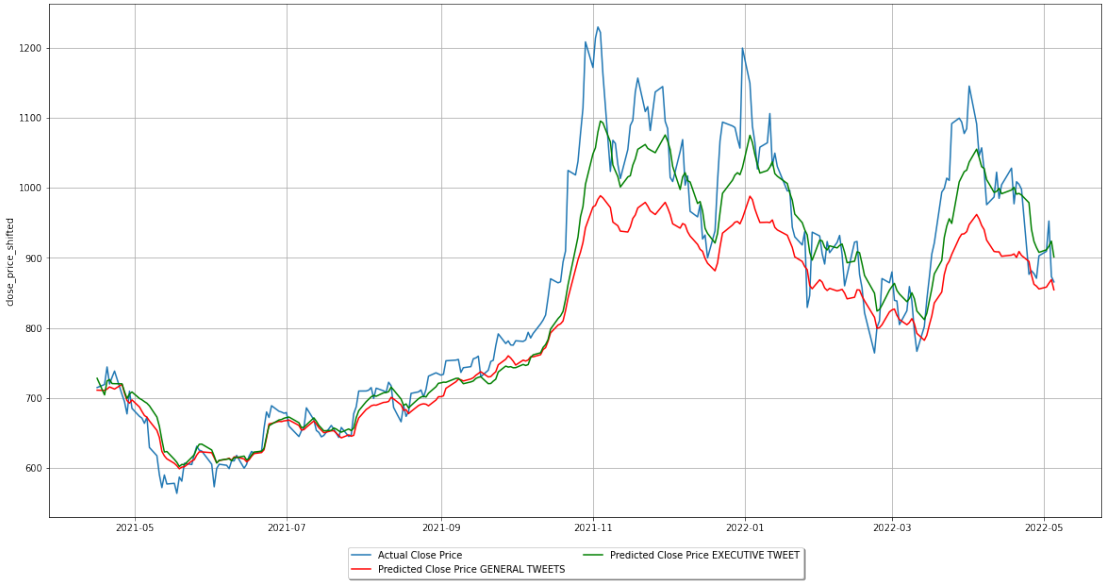

FIGURE 5.13: Close price prediction of Tesla with Y+G and Y+E Datasets

for all the stocks and decentralised currencies considered, it proves that the finding is widespread and effective in multiple domains. Hence, we can concretely conclude that the influence of executive posts on close price is much more than general posts not only for different stocks but also for different decentralised currencies. Figure 5.13, plots the actual data of the close price (blue) for Tesla, the predicted close prices obtained with the sentiment of executive posts (green), and general posts (red). It is evident from figure 5.13 that the green curve is much closer to the blue curve than the red curve, i.e., executive opinions are much more effective than general opinions with respect to the close prediction task.

Overall, sentiments expressed by executive in Twitter gives the best performance. It is equally interesting to note that unlike Twitter, the difference in performance between the general and executive datasets is not significant for Reddit.

Stock	Data	MAE	RMSE	R_a^2	MAPE (%)
TSLA (T)	**Y+G**	59.621	83.765	0.768	6.174
	Y+E	**34.362**	**48.261**	**0.923**	**3.817**
AAPL (T)	**Y+G**	3.899	5.004	0.899	2.459
	Y+E	**3.071**	**3.849**	**0.940**	**1.988**
BTC (T)	**Y+G**	3676.006	4631.177	0.706	7.492
	Y+E	**2842.190**	**3679.708**	**0.814**	**5.830**
ETH (T)	**Y+G**	309.507	392.286	0.668	8.557
	Y+E	**278.507**	**356.633**	**0.725**	**7.724**
TSLA (R)	**$Y+G_r$**	**46.953**	**68.659**	**0.844**	**4.954**
	$Y+E_r$	57.674	82.634	0.774	6.043

TABLE 5.16: Result of Experiment 4

5.5.5.4 Comparative study of close price prediction with and without imputation (Experiment 4)

This experiment was performed on all stocks and decentralised currencies using the datasets: **Y+G**, **Y+E**, **$Y+G_r$** and **$Y+E_r$**. The motivation behind this experiment is the observation that there is an abundance of tweets by general people and a scarcity of tweets by executives. To keep the research fair, we decided to equalise the number of tweets by general people and executives. We identified the dates on which there were no executive tweets. Tweets by general people were dropped for these dates. We performed Cubic Spline Interpolation for all the datasets. It resulted in an equal amount of data for the general and executive datasets. These datasets were used to train on the **GRU** model for predicting the close price of the next day. Table 5.16, shows the evaluation results for this experiment. We observe that for all the stocks, sentiments of tweets by executives are better predictors than that of the crowd. However, in case of Reddit, the sentiments of general posts prove to be more effective than that of the executives. Since there is not much difference in our findings from experiments 3 and 4, we conclude that the imputation methodology we followed and the non-availability of executive posts for all days did not have much effect on the results.

5.5.6 Conclusions

In this research, we studied how the sentiment of social media posts by executives and people in general affect stock prices of two popular companies Apple and Telsa. We primarily considered Twitter and Reddit posts for this research. We extended our study by predicting prices of two popular decentralised currencies Bitcoin and Ethereum. Our experiments successfully answer the research questions raised before.

RQ-1: Social media data from both Twitter and Reddit have a deep influence on close price movements. On integrating the sentiment of social media data, significant improvements were witnessed in close price prediction.

RQ-2: Sentiment of tweets by executives have a deeper influence on the prediction of close price. This is because the executives have more impact on the society and the mass tends to have more faith in executives and are easily influenced by the opinion of

executive people. However, the effect of tweets by general people should not be considered unimportant. This supports the claims made by Jermann [230], Elliott et al. [220] and Chen et al. [228].

RQ-3: Our findings using tweets also hold good for Reddit posts.

This work has a lot of directions where further research could be performed. Instead of just using the sentiment of the tweets, we would like to use the entire textual content for predicting close prices. A better way could be found to impute sentiment on days no tweets or posts are available. If these models are trained on more granular data, users can leverage them for choosing winning stocks by utilising the stock price prediction made every minute. While acknowledging that Twitter's (now X) transition to a restrictive API paywall model has substantially curtailed academic research access, the methodologies, and algorithms developed proposed remain methodologically sound and platform-agnostic. The analytical frameworks presented in this work are transferable to alternative social media platforms with accessible APIs, including Reddit, Bluesky, etc.

5.6 Financial Argument Analysis

5.6.1 Introduction

Earning call transcripts are an important source to know more about the financial performance of any organization. With the advent of social media, investors tend to discuss various investment strategies online. The FinArg-1 shared task [247] co-located with NTCIR-17 deals with mining arguments from financial texts. In this chapter, we discuss various approaches we followed for identifying argument units and relations in earning call transcripts and social media posts. This corresponds to Task-2 and Task-3 as mentioned in [247]. The dataset for Task-2 was in English while that for Task-3 was in Chinese. Furthermore, for the task of Argument Relation Identification, we also explored the applicability of Large Language Models under zero-shot and few-shot settings.

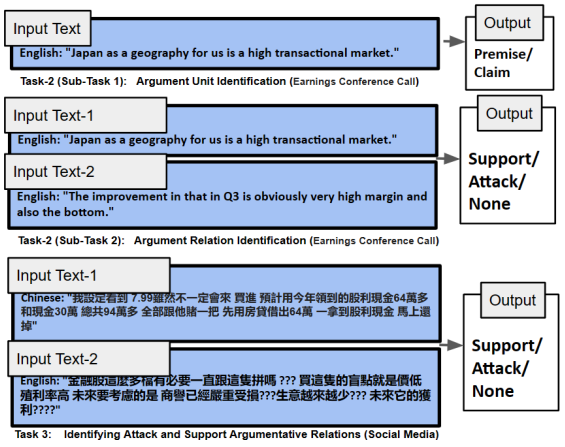

FIGURE 5.14: Argument Analysis in Financial Texts

5.6.2 Problem Statement

Task 2, Sub-Task 1: Given a financial argumentative text in English, we want to classify it as premise or claim.
Task 2, Sub-Task 2: Given two financial argumentative texts in English, our aim is to detect the relation between them. The relation can be 'Support', 'Attack', or None.
Task 3: Given two argumentative social media posts relating to finance in Chinese, the objective is to classify the relation between them. The relation can be 'Support', 'Attack', or None.
Chen et al. [247] described the tasks and datasets in more detail. We present it in Figure 5.14.

5.6.3 System Descriptions

In this section, we discuss our best-performing systems.

TASK	DATASET	LABEL	# ORIGINAL	# PARAPHRASED
2-2	Train	0	1600	3200
		1	3859	3859
		2	62	372
	Val	0	200	200
		1	482	482
		2	8	8
3	Train	0	684	4104
		1	3676	3676
		2	2158	3676
	Val	0	85	85
		1	460	460
		2	270	270

TABLE 5.17: Count (#) before & after paraphrasing. 2-2 refers to (Task-2, Sub-Task-2)

5.6.3.1 Task 2: Argument Identification

The Argument Identification task consists of two sub-tasks: Argument Unit Classification and Argument Relation Identification.

5.6.3.2 Sub Task 1: Argument Unit Classification

In this task, we had to identify and classify whether the given sentence was a claim or a premise. The training data had 7,753 sentences, and validation data had 969 sentences. In the training data given, the distribution was quite balanced, with 52.4% of the sentences labelled as claims and the remaining as premises. A similar distribution was seen with the validation data as well. After experimenting with various models, we found that a BERT-SEC [151] model trained for 5 epochs and a batch size of 32 performed the best (Micro-F1: 73. 89%, Macro-F1: 73. 86% in the test set).

5.6.3.3 Sub Task 2: Argument Relation Detection and Classification

In the given training dataset, we had 5,521 pairs of labelled sentences from which we had to identify and classify the relationship between them as support, attack, or none. In the validation data set, we had 690 pairs of labelled sentences. We identified the high class imbalance in the given dataset, so we had made an attempt to up-sample the minority class by paraphrasing the existing sentence pairs. The counts of each of the classes are given in table 5.17. We had used Contextual Word Embedding Augmenter and Synonym Augmenter from NLPAUG [248] library and FLANG-RoBERTa model [76] for paraphrasing sentences. However, paraphrasing was performed only on the training dataset, and hence the validation dataset remains the same.

We further fine-tuned the best-performing FinBERT model of Task-1 Sub-Task-2 for classification using the cross-encoder architecture [68]. This fine-tuning was done for 5 epochs with a batch size of 16 on the original dataset. This outperformed all other models we trained (Micro-F1: 79.42%, Macro-F1: 60.22% in the test set).

MODEL	MACRO-F1 (VALIDATION SET)	MICRO-F1 (VALIDATION SET)
RNN + Spacy Tokenizer	0.3571	0.5270
FastText + NN	0.7155	0.7173
GloVe Embeddings + CNN	0.6952	0.6957
BART-BASE-CASED + BERT TOKENIZER	0.7336	0.7337
BERT-SEC	**0.7426**	**0.7430**
FinBERT	0.7398	0.7401

TABLE 5.18: Results of Task 2, Sub-Task 1: Argument Unit Identification

5.6.3.4 Task 3: Identifying Attack and Support Argumentative Relations

In the training dataset, we had 6,518 pairs of labelled Chinese sentences. In the validation dataset, we had 815 pairs of labelled Chinese sentences. To increase the number of instances in the minority class, we paraphrased them. The distribution is presented in Table 5.17. Our aim was to infer from a sentence pair of social media posts if the argumentative posts were supportive, attacking, or neutral. Since the posts were in Chinese, we divide our work into 2 parts. Firstly, we translated Chinese texts into English using Google Translate. Secondly, we worked with the raw Chinese texts as it is.

By fine-tuning a BERT-SEC [151] model using cross encoder architecture on English texts obtained through translation, we obtained the best results on the test set (Micro F1: 64.79%, Macro F1: 69.45%). This fine-tuning was done with a batch size of 8, for 5 epochs on the original i.e. non-paraphrased dataset.

5.6.4 Experiments and Results

In this section, we mention the experiments we performed and their results.

5.6.4.1 Task 2: Argument Identification

5.6.4.2 Sub Task 1: Argument Unit Classification

For this task, we experimented with various classification techniques. We first used a simple Recurrent Neural Networks (RNNs) with a fully connected layer along with the Spacy [249] tokenizer to get our outputs, but this model did not perform well. The second experiment was using the FastText [250] model. This model had far fewer parameters than the previous model. It first calculated the word embedding for each word using the Embedding layer, then calculated the average of all the word embeddings and fed it to the linear layered Neural Network (NN). Next, we replaced the existing embeddings with Glove [251] and fed our embeddings into 3 convolutional layers and then finally to a fully connected layer to get the labels. We used a drop-out of 50%. Finally, we fine-tuned a few pre-trained language models like BERT [1], BERT-SEC [151], and FinBERT [69]. This led to significant improvement in performance. The results are mentioned in Table 5.18.

5.6.4.3 Sub Task 2: Argument Relation Identification

Firstly, we concatenated the texts in a given pair with separator ([SEP]) token in between them. We fine-tuned several encoder based pre-trained language models for classification. They are DistillBERT [172], FLANG-RoBERTa [76], and BERT-SEC [151]. Subsequently, we fine-tuned the cross encoder [68] architecture with BERT [1], BERT-SEC [151], and FinBERT [69] previously fine-tuned for Task-2 Sub-Task-1 embeddings. The scores of all the models are given in the result section, Table 5.19. We further tried to adapt the models to the given domain using Masked Language Modelling (MLM). However, this didn't improve the performance. Each model was trained with a batch size of 16 and 5 epochs. All the experiments were performed on the original as well as the paraphrased datasets.

Leveraging Small Language Models

Small Language Models (SLMs) have been re-defining the state-of-the-art in Natural Language Processing. We experimented like Dolly v2 [252] under zero shot and few shot settings.

Few shot learning is a method where we ask a language model to do a task and provide the model with a few examples of the task. Initially, we experimented with a static prompt where we choose one example from each classification category: 'Support', 'Attack', and 'None'. A static prompt is a prompt where the few shot examples are kept fixed with different query. But the performance was not satisfactory. This inspired us to come up with a novel dynamic prompt engineering algorithms where the few shot examples would not be fixed unlike static prompting. The motive of our algorithms is to dynamically choose such examples with each validation query which are similar to the query, hence giving the language model a better understanding of the classification task. Our algorithms have two steps and three steps, respectively. Our first proposed algorithm (Algorithm-1) has two steps, (1) Tweet Topic Classification and (2) Semantic Similarity. The algorithm initially finds the tweet topic of each instances present in train set and validation set. These topics were extracted using pre-trained model [87]. We append these tweet topics to the train set and validation set as columns. Now, we iterate through the validation set. For each validation instance, we choose a sample from the train set whose topic is equal to the topic of validation instance. Since, this task is like Natural Language Inference (NLI), and we have a pair of sentences whose relationship has to be determined, we merge the two sentences for simplicity. This gives us a train corpus whose embedding is found. Similarly, for the validation instance, we merge the sentences and find the embedding. Now, we find the cosine similarity between all the instances present in the train corpus and the validation instance. From this we choose top k sentences having maximum semantic similarity. These two steps ensures that the training examples provided with the validation instance belongs to the same topic as well as have the highest semantic similarity. However, in this algorithm, we are not ensuring whether each of the examples comes from different classification categories.

Our second proposed algorithm (Algorithm-2) has three steps, (1) Tweet Topic Classification using [87] (2) Semantic Similarity (3) Class Filter. This algorithm overcomes the limitation of the previous algorithm by making sure that the examples provided with the validation query comes from different classes ('Attack', 'Support', 'None') and has similar topic with high semantic similarity.

MODEL	DATA	VALIDATION	
		MICRO F1	MACRO F1
DistilBERT	Original	0.7913	0.5321
DistilBERT	Paraphrased	0.7942	0.4811
Flang-Roberta	Original	0.7971	0.5653
Flang-Roberta	Paraphrased	0.7971	0.5456
BERT-SEC	Original	0.813	0.5647
BERT-SEC	Paraphrased	0.7880	0.4900
Cross Encoder (BERT)	Original	0.7898	0.5383
Cross Encoder (BERT)	Paraphrased	0.7913	0.4956
Cross-Encoder (BERT-SEC)	Original	0.7695	0.476
Cross-Encoder (BERT-SEC)	Paraphrased	0.7681	0.4807
Cross Encoder (FinBERT Finetuned)	**Original**	**0.8275**	**0.5298**
Cross-Encoder (MLM-FinBERT)	Original	0.8000	0.5054
Cross-Encoder (MLM-FinBERT)	Paraphrased	0.7913	0.5482

Table 5.19: Results of Task 2, Sub-Task 2: Argument Relation Identification

LANGUAGE	MODEL	DATA	VALIDATION	
			MICRO F1	MACRO F1
English	BERT-base	Original	0.6453	0.6783
English	BERT-base	Paraphrased	0.6319	0.6568
English	FLANG-RoBERTa	Paraphrased	0.6392	0.6754
English	Cross Encoder (SBERT)	Original	0.6404	0.6796
English	Cross Encoder (SBERT)	Paraphrased	0.6500	0.6880
English	Cross Encoder (DistilROBERTA)	Original	0.7055	0.7472
English	Cross Encoder (DistilROBERTA)	Paraphrased	0.6920	0.7374
English	Cross Encoder (Flang-Roberta)	Original	0.6932	0.7342
English	Cross Encoder (Flang-Roberta)	Paraphrased	0.6858	0.7314
English	**Cross Encoder (BERT-SEC)**	**Original**	**0.6932**	**0.7342**
English	Cross Encoder (BERT-SEC)	Paraphrased	0.6800	0.7000
English	Cross Encoder (MLM on BERT-SEC)	Original	0.6846	0.7160
English	Cross Encoder (MLM on BERT-SEC)	Paraphrased	0.6871	0.7180
Chinese	SBERT-Chinese	Original	0.6321	0.6450
Chinese	Cross Encoder (SBERT-Chinese)	Original	0.6503	0.6432

Table 5.20: Result of Task 3: Identifying Argumentative Relation in Social Media Discussion

Sample prompts have been provided in the Appendix section. From the results, we can observe that prompts curated from Algorithm-1 are performing better than static prompts as well as prompts curated from Algorithm-2. From this observation, we can conclude that it is unnecessary to provide examples from different classes. This would simply add redundant information and noise to the language model, resulting in misclassification. Thus Algorithm-1 which curates prompts without the class filter works better than Algorithm-2. This is the first finding. Another observation is, language models tend to predict the first occurring classification category from the example for majority of the validation queries. We have performed three experiments with Algorithm 2, "v1" had "None" category as the first example, "v2" had "Attack" category as the first example, "v3" had "Support" category as the first example. In each of three experiments, the category of the first example became the majority category for prediction. This can be noticed from the confusion matrix given in the table. This is happening possibly because internally the language model is getting biased towards the first mentioned class. This is also a reason why Algorithm-2 does not perform as well as Algorithm-1. This is the other finding.

Task 3: Identifying Attack and Support Argumentative Relations

Firstly, we translated the Chinese texts to English so that we could comprehend them. To address the class imbalance, we paraphrased the English texts belonging to the minority classes. We fine-tuned several encoder based models like BERT-base-Uncased [1] and FLANG-RoBERTa [76] for classification after concatenating the texts in a given pair with a separator ([SEP]) token. We experimented with both the original and paraphrased data. Subsequently, we used cross-encoder architecture [68] with embeddings from Distil-RoBERTa [172], FLANG-Roberta [76], and BERT-SEC [151] for both the original and paraphrased datasets. We further used Masked Language Modelling (MLM) to adapt these models to the given domain.

To avoid the loss due to translation, we experimented with the original Chinese Texts as well. Firstly, we converted the raw Chinese text to simplified traditional Chinese texts using zhconv library. [37]. We trained a SBERT-Chinese[38] model for classification. We used the original dataset for training, as we couldn't find and validate a paraphraser suitable for Chinese texts. Subsequently, we replaced the embeddings in the cross-encoder architecture with SBERT-Chinese[39] embeddings and fine-tuned the model further. Each of the cross-encoder models were trained batch size of 8 and 5 epochs to train our dataset. The results are presented in Table 5.20.

Leveraging Small Language Models

Our experimentation focused on few-shot learning scenarios. Initially, we employed a static prompt strategy, selecting one example from each classification category ('Support', 'Attack', 'None'). However, this approach yielded unsatisfactory performance. This led us to innovate novel dynamic prompt engineering algorithms. The core idea behind these algorithms was to dynamically choose examples during validation that closely resembled the query, providing the LLM with a more profound understanding of the classification task.

As Algorithm-1 performed better than Algorithm-2 for Task 2-2, we experimented only with Algorithm-1 for Task-3. We translated the Chinese texts to English and evaluated Large Language Models like flan-t5-small [253], mpt-1b-redpajama-200b-dolly[40], and dolly-v2-3b [252] under various settings.

5.6.5 Conclusion

In this chapter, we shared our team, LIPI's approach for Argument Unit Classification, Argument Relation Detection, and Identifying Attack & Support Argumentative Relations in English and Chinese financial texts. We observed FinBERT[69] and BERT-SEC [151]

[37] https://pypi.org/project/zhconv/ (accessed on 15th August, 2023)

[38] https://huggingface.co/DMetaSoul/sbert-chinese-qmc-finance-v1 (accessed on 16th August, 2023)

[39] https://huggingface.co/DMetaSoul/sbert-chinese-qmc-finance-v1 (accessed on 16th August, 2023)

[40] https://huggingface.co/mosaicml/mpt-1b-redpajama-200b-dolly (accessed on 23rd August, 2023)

based models when fine-tuned using cross encoder architecture performed the best for relation identification. Paraphrasing and pre-fine-tuning using MLM did not help much in improving the performance of the model. LLMs under zero shot and few shot setting did not do as well. For Task-2, our team was ranked 13^{th} and 2^{nd} in sub-task-1 and sub-task-2 respectively. For task-3, we were ranked 4^{th}.

Regarding the limitations, it is necessary to mention that we have not considered semantic loss due to paraphrasing. In future, we would definitely try to improve it and we want to extend this solution to low resources Indian languages and create a user-friendly tool to help investors.

Chapter 6

Indic Investing

In India, the wealth disparity is huge. The richest 1% of Indians possess more than 40% of India's wealth.[1] A survey by National Center for Financial Education revealed that only 27% of Indians are financially literate.[2] Nations across the world spend billions of dollars in various financial inclusion schemes. But, for making these schemes successful, financial literacy is essential. The fear of losing money also deprives people of investing regularly. Due to lack of awareness, people tend to invest only in conventional avenues like gold, Fixed Deposits, etc.[3] Since return on investment from these traditional avenues has been constantly reducing, it is essential to enlighten the public with the knowledge of financial markets and products like equities, bonds, mutual funds, etc.[4] It is equally noteworthy that only a few active traders could earn more than the FD rates over the last 3 years.[5] Thus, apart from improving overall Financial Literacy, it is also essential to equip investors with data-driven tools for making trade-related decisions[6].

6.1 Research Questions

- **RQ-6**: How to keep investors informed in Indian Languages?
    - **Relevant Contributions**: Financial Argument Analysis in Bengali [21], Financial Natural Language Processing for Indian Languages [22], Predicting success of Indian IPOs [254], Predicting Ratings of Indian IPOs [255].

[1] https://www.livemint.com/news/india/richest-1-indians-own-more-than-40-of-country-s-wealth-report-11673830307891.html (accessed on 18th September, 2023)

[2] https://www.financialexpress.com/market/only-27-indians-are-financially-literate-sebis-garg/2134842/ (accessed on 18th September, 2023)

[3] https://indianexpress.com/article/business/market/less-than-1-of-rural-households-invest-in-stocks-sebi-survey-4601264/ (accessed on 18th September, 2023)

[4] https://economictimes.indiatimes.com/markets/stocks/news/financial-education-and-its-importance-in-making-investing-accessible-across-india/articleshow/91792025.cms (accessed on 18th September, 2023)

[5] https://economictimes.indiatimes.com/markets/stocks/news/no-easy-money-less-than-1-active-traders-beat-bank-fds/articleshow/88656009.cms (accessed on 18th September, 2023)

[6] https://www.analyticsinsight.net/leveraging-artificial-intelligence-to-simplify-financial-knowledge/ (accessed on 18th September, 2023)

6.2 Financial Argument Analysis in Bengali

6.2.1 Introduction

In natural language processing (NLP), argument mining has of late become one of the popular areas of research. The goal of argument mining is to automatically extract and identify argumentative structures from texts. Argumentative texts are present in various places like investor-generated text, social media posts, etc. While argument mining has been a topic of study for several years, financial argument analysis is still in the early stage. To the best of our knowledge, no work has been done to date on financial argument analysis in low-resource languages such as Bengali.

In this chapter, we focus on developing resources for mining arguments from financial texts in Bengali. First, we introduce the task of **Argument Unit Classification** to classify financial argumentative texts in Bengali into 'Premise' or 'Claim'. 'Premise', being more reliable than the 'Claim', helps investors to make data-driven decisions. Secondly, we present the task of **Argument Relation Identification** to understand if two financial argumentative texts in Bengali are supporting, attacking, or not related to each other. To quantify the effect of social media posts and understand an entire discussion, it is essential to examine the posts that support or attack each other.

Our contributions

- We created two datasets in Bengali for the task of Financial Argument Unit Classification and Argument Relation Identification. These datasets are released under CC BY-NC-SA 4.0 licence.[7]
- We fine-tuned two pre-trained language models to accomplish the above-mentioned tasks. We open-sourced these models so that the research community can use them as baselines.
- We developed a user-friendly tool (**Financial Argument Analysis in Bengali (FAAB)**) for demonstration [8] and hosted it in HuggingFace Spaces. [9]

6.2.2 Related Works

Argument analysis is one of the emerging research areas in Natural Language Processing. Lippi and Torroni [256] surveyed existing works relating to argument mining across various domains like economic sciences, policymaking, and information technology. Cabrio and Villata [257] narrated various machine learning and deep learning algorithms to predict relationships between texts. Schaefer and Stede [258] reviewed existing works on argument mining specifically for Twitter. They discussed the approaches used for modelling the structure of arguments in the context of tweets. Furthermore, they studied the current progress in detecting arguments, and their relations in tweets. Finally, they explored the overlap between stance detection and argument mining. Lawrence and Reed [259] also

[7] https://github.com/sohomghosh/FAAB_Financial_Argument_Ananlysis_Bengali (accessed on 11th Aug 2023)

[8] https://youtu.be/4JwVl4mbj6Q (accessed on 9th Aug 2023)

[9] https://huggingface.co/spaces/rima357/FinArgBengali (accessed on 9th Aug 2023)

surveyed recent advances in the domain of Argument Mining. Argument analysis has been widely adopted across various domains like legal [260] and finance [234]. Xu and Ashley [260] conceptualised argument mining as a word-level classification problem instead of a sentence-level classification problem. Chen et al. [234] proposed the structures between the opinions and those between the financial instruments. They discussed how opinions from various sources can be used to extract opinion components and detect the relation between the opinions. Chen et al. [261] explored the applicability of opinion mining in the financial domain. They analysed the investor's opinion. In [262], Zhai et al. proposed a dataset called AntCritic which consists of 10K free-form and visually-rich financial comments and supports both argument component detection and argument relation prediction task.

Although there have been significant advances in the field of Argument Mining on financial English texts, a lot is yet to be done for low-resource languages like Bengali. To address this knowledge gap, we release datasets, models, and, a tool for effectively analysing arguments in Bengali financial texts.

6.2.3 Problem Statement

We want to accomplish the following tasks.
Task-1: Given a financial argumentative text in Bengali, our aim is to classify it as 'Premise' or 'Claim'.
Task-2: Given two financial argumentative texts in Bengali, we want to classify the relationship between them. The relation can be 'support', 'attack' or 'none' (i.e. no relation).

6.2.4 Datasets

Due to a lack of resources to create ground truth for Bengali datasets from scratch, we leveraged the resources available in English. We translated the datasets released by Chen et al. [247] from English to Bengali using the translation system released by Ramesh et al. [263]. To understand the applicability of the translation system for argumentative financial text, we evaluated it using a two-step approach. Firstly, we manually translated 100 instances from English to Bengali and calculated the average BERTScore [262] and Cross-lingual Optimised Metric for Translation Evaluation (COMET) [264] scores for machine-translated sentences and human-translated gold standard reference translations in Bengali. The scores obtained were 0.964 and 0.859, respectively. Secondly, as manual translation was not scalable, we randomly picked 8000 instances from the overall dataset. We translated the English texts into Bengali using the MT system [263]. Then, we back-translated the Bengali texts to English using the same MT system (in the opposite language direction). We calculated the BERTScore between the original English text and the back translated English text. Due to the absence of actual ground truth for reference, we could not calculate the COMET score. We obtained an average BERTScore of 0.944. To further ensure the validity of the machine translation system, we calculated the LaBSe [265] based cosine similarity between sentence embeddings [266] of original texts in English and texts in Bengali obtained by machine translation. The average cosine similarity score was 0.845. As the scores obtained were high in all scenarios, we conclude that the system developed by Ramesh et al.[263] to translate to and from Indian languages is applicable for our dataset.

Text-1 (Bengali)	Text-2 (Bengali)	Label
তাই প্রথমবার আমরা ঘোষণা করেছিলাম যেআমাদের ২০ লক্ষেরও বেশি বিজ্ঞাপনদাতা আছেন যারা ফেসবুকে বিজ্ঞাপন কিনছেন।	৮২% মানুষ যারা আমাদের সাথে বিজ্ঞাপন শুরু করেন তারা আমাদের খুব সাধারণ বিজ্ঞাপন পণ্য দিয়ে শুরু করেন।	0 (No relation)
আমাদের আরও বেশি বিজ্ঞাপনদাতা রয়েছেন যারা সঠিক ব্যক্তিকে লক্ষ্য করে তাদের বিজ্ঞাপন দেওয়ার ক্ষমতা ব্যবহার করেন।	এবং আমি মনে করি আমরা এই সময় ক্ষেত্রে অভূতপূর্ব বৃদ্ধি দেখতে পাচ্ছি।	1 (Support)
আপনারা যদি চীনের মূল ভূখণ্ডের দিকে তাকান, যার ওপর আমি ব্যক্তিগতভাবে বিশেষ দৃষ্টি নিবদ্ধ করেছি, তা হলে চীনের মূল ভূখণ্ডে আমাদের সংখ্যা ১১ শতাংশ কমে গেছে।	আর তাই আমি যখন এর থেকেঘুরে দাঁড়াই এবং বৃহত্তর চিত্রের দিকে তাকাই, আমি মনে করি চীন ততটা দুর্বল নয় যতটা বলা হয়েছে।	2 (Attack)

TABLE 6.1: Some instances from Task-2 dataset

To further ensure the quality of our dataset, we segmented the dataset into several brackets based on the BERTScore and cosine similarity scores. For each of these brackets, we calculated the accuracy of translation by manually assessing how many of the 10 randomly picked instances from the corresponding bracket were translated properly. This is presented in Figure 6.1. Analysing this plot, we decided to retain only those translated instances having BERTScore and cosine similarity score above 0.925 and 0.800, respectively. We use the same training set, validation set, and labels as mentioned in the chapter [247]. After applying the filters mentioned, the label-wise distribution of instances for Task-1 and Task-2 are mentioned in Tables 6.2 and 6.3 respectively.

For Task-1, we have text in Bengali and the corresponding label (0=Premise, 1 = Claim). For Task-2, we have two texts in Bengali and their corresponding relation (0 = No relation, 1 = Support, 2 = Attack). Some samples of Task-1 and Task-2 are shown in Tables 6.4 and 6.1 respectively.

Dataset	Premise (Label-0)	Claim (Label-1)
training	2858	2667
validation	353	334

TABLE 6.2: Data distribution of Task-1

Dataset	No relation (Label-0)	Support (Label-1)	Attack (Label-2)
Training	1970	794	28
validation	230	108	5

TABLE 6.3: Data distribution of Task-2

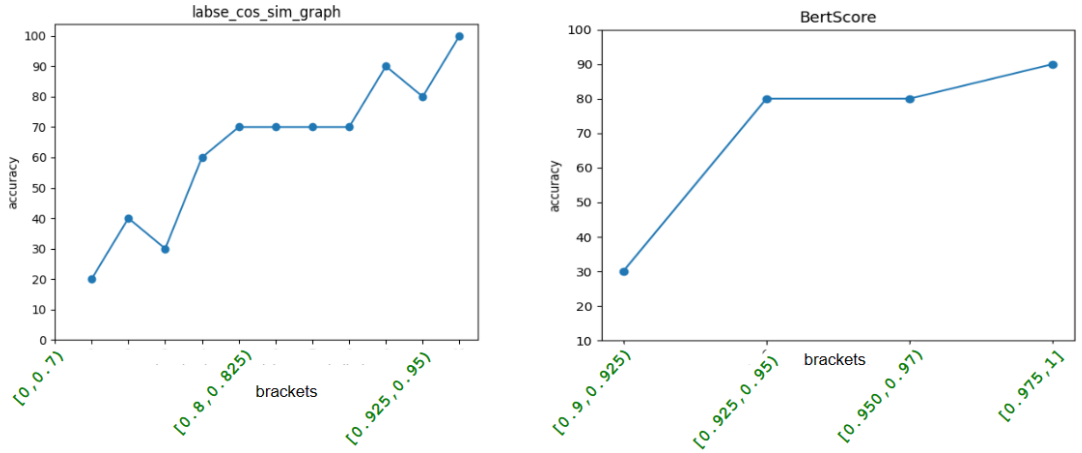

FIGURE 6.1: Accuracy of each bracket created using LaBSE based cosine-similarity and BERTScore

Text (Bengali)	Label
এবং এই প্রেক্ষাপটে, অবশ্যই, তারা পুরোনো কিছু কাজের বোঝা তুলে নিচ্ছে এবং স্থানান্তর করছে, কিন্তু তারা পুরো ব্যবসায়িক প্রক্রিয়ার আধুনিকীকরণ করছে।	0 (Premise)
হ্যাঁ, কোয়ার্টারের জন্য, তাই এটি একটি শক্তিশালী কোয়ার্টার ছিল।	1 (Claim)

TABLE 6.4: Some instances from Task-1 dataset

6.2.5 Experiments and Results

Firstly, for Task-1, we fine-tuned several variants of BERT [1] for classifying the given text into 'premise' and 'claim'. These variants are sagorsarker bangla bert base (SSB) [267], monsoon nlp bangla electra.[10] (MNB), aibharat indic bert (AIB) [268], distilbert base multilingual cased (DBMC) [172], and bert base multilingual cased (BBMC) [1]. The results of these models for Task-1 are presented in Table 6.5. We observe that **BBMC** when fine-tuned for 5 epochs, with batch size of 8, weight decay of 0.1, and 500 warm up steps, outperformed all other variants of the BERT model. We further fine-tuned a BERT-base-uncased [1] model on the English texts obtained by translating the Bengali texts. Although it performed slightly better than BBMC, the lift in performance was not statistically significant (p-value > 0.05) and deploying a machine translation system was an extra overhead.

[10] https://huggingface.co/monsoon-nlp/bangla-electra (accessed on 9th August, 2023)

Model Name	A	P	R	F1
SSB	0.700	0.690	0.694	0.692
MNB	0.714	0.691	0.745	0.717
AIB	0.711	0.700	0.712	0.706
DBMC	0.681	0.656	0.721	0.687
BBMC	**0.719**	**0.697**	**0.745**	**0.721**

TABLE 6.5: Results for Task-1. [A= Accuracy, P = Precision, R = Recall, F1 = F1 (binary)]

For Task-2, following [20], we used the cross-encoder architecture.[11] [68]. We trained it for classifying the relation between two texts in Bengali into one of 3 categories ('Support', 'Attack', or 'None'). We experimented with several variants of the BERT model, as done in Task-1. The results are presented in Table 6.6. We observe that similar to Task-1, **BBMC** when fine-tuned for 5 epochs, with batch size of 16, and 88 warm up steps, performed the best. Subsequently, we fine-tuned a BERT-base-uncased [1] model with the English texts obtained by translating the Bengali texts. Similar to Task-1, it performed slightly better than BBMC for Task-2 as well. However, the performance improvement was not statistically significant (p-value > 0.05) and deploying a Bengali to English machine translation system in production was an extra overhead.

Lastly, for both the tasks, we tried to adapt **BBMC** to the financial domain using Masked Language Modelling. However, this did not improve the performance.

Model Name	A	P mi	P ma	R mi	R ma	F1 mi	F1 ma
SSB	0.708	0.708	0.442	0.708	0.417	0.708	0.419
MNB	0.705	0.705	0.235	0.705	0.333	0.705	0.275
AIB	0.682	0.682	0.476	0.682	0.348	0.682	0.303
DBMC	0.699	0.699	0.442	0.699	0.392	0.699	0.386
BBMC	**0.755**	**0.755**	**0.488**	**0.755**	**0.46**	**0.755**	**0.466**

TABLE 6.6: Results for Task-2. [A= Accuracy, P = Precision, R = Recall, F1 = F1 score, mi = micro, ma = macro]

[11] https://www.sbert.net/examples/applications/cross-encoder/README.html (accessed on 9th August, 2023)

Prompts for LLM (TinyPixel/Llama-2-7B-bf16-sharded)
Zero-Shot:
Classify the following Input text into one of the following two categories: ['Premise', 'Claim'] Input : {text}
Few-Shot:
Classify the text given below into Premise or Claim based on the meaning of the text. Choose only one Class either: Premise or Claim for a text. Input : I mean, sometimes it is not that you came up with a bright strategy, it is like doing really good work continuously for a long time. Response : the class of the text is Premise. Input : See, first of all, I would like to say that the opportunity for our shareholders was never better when they thought of Microsoft. Response : The class of the text is Claim. Input : For example, we have never participated so much, I would call it a burden of all sophisticated work in the non-developed market, medium and small businesses. Response : The class of the text is Premise. Input : However, primarily, the feed for the video is going to focus on making money through advertisements. Response : The class of the text is Claim. Input : {text}

TABLE 6.7: The Zero-Shot and Few-Shot Prompts for Task-1

6.2.6 Large Language Models for Argument Analysis

Since Large Language Models (LLMs) have been re-defining the state of the art in NLP, we experimented with llama-2.[12] under zero-shot and few-shot setting. We further instruction-fine-tuned llama-2.[13] and evaluated the instruction-fine-tuned version under zero-shot setting. We performed these experiments with English texts obtained by translation due to the unavailability of LLMs for Bengali. The prompts which we have used are mentioned in Table 6.7. The results for Task-1 are presented in Table 6.8. Since the performance for Task-1 did not improve on using llama-2, we did not carry out experiments with LLMs for Task-2.

[12] https://ai.meta.com/llama/ (accessed on 18th Aug, 2023)

[13] https://huggingface.co/TinyPixel/Llama-2-7B-bf16-sharded (accessed on 18th Aug, 2023)

Model	Type	A	P	R	F1
llama-2	Zero-shot	0.505	0.484	0.278	0.354
llama-2	Few-shot	0.522	0.500	0.186	0.271
llama-2	IF + Zero-Shot	0.528	0.531	0.254	0.344

TABLE 6.8: Results for Task-1 using llama-2. [IF = Instruction-fine-tuned, A= Accuracy, P = Precision, R = Recall, F1 = F1 (binary)]

6.2.7 Tool Description

For helping investors, we developed a user-friendly tool, **Financial Argument Analysis in Bengali (FAAB)**. In the back-end, we use the best performing model, i.e., BERT-base-multilingual-cased [1]. This has been fine-tuned separately for Task-1 and Task-2. The front-end was developed using Gradio [174].

FIGURE 6.2: Screenshot of Tab-1 of the tool kit

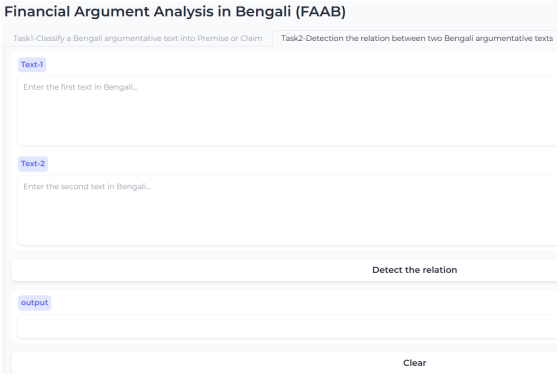

FIGURE 6.3: Screenshot of Tab-2 of the tool kit

As presented in Figures 6.2 and 6.3, the tool consists of two tabs. The first tab (Tab-1) is used to classify a Bengali argumentative text into premise or claim. The other tab (Tab-2) is used to detect the relation between two Bengali argumentative texts.

In Tab-1, there is a text box where users can write the text in Bengali. Below this there is the 'classify' button which on clicking will classify the text into 'Premise' or 'Claim'. The

output will be shown below. In Tab-2, there are two text boxes, one is for Text-1 and the other is for Text-2 where users can enter the texts in Bengali. By clicking on the 'Detect the relation' button, it detects the relation between these two Bengali texts and the output appears below. We have provided a list of examples for both the tasks at the bottom.

6.2.8 Conclusion

In this chapter, we proposed two datasets for mining argumentative financial texts in Bengali. Subsequently, we released the baseline models and open-sourced the **Financial Argument Analysis in Bengali (FAAB)** tool. The baseline models were obtained by fine-tuning BERT-base-multilingual-cased. Collecting more data (specifically for Task-2), determining profitability from these arguments, and experimenting with INDICXNLI datasets [269] are interesting directions for future work.

6.3 IndicFinNLP: Financial Natural Language Processing for Indian Languages

6.3.1 Introduction

In a nation, financial literacy leads to the overall well-being of citizens and economic prosperity. The financial literacy rate in India is only 27%.[14] The cultural diversity and existence of more than 100 major languages in India make it difficult to spread financial knowledge across the population. While most researchers working on FinNLP focused on creating datasets for high resource languages like English and Chinese, to the best of our knowledge such datasets do not exist for Indian languages, even though some of the Indian languages belong to most spoken languages worldwide. [15] To improve the overall financial literacy of the country, it is essential to educate the citizens in their own vernacular languages. As misinformation is a prevalent problem in today's society, it is extremely important to ensure that the financial knowledge being imparted is authentic. Many a time, numbers and figures are misrepresented to allure common people who are novice investors. To address this, we develop a system to detect exaggerated numerals in financial texts in Indian languages. With the ever-growing concern for climate change, investors are increasingly looking for avenues of green investing like sustainable and Environmental, Social, and Governance (ESG) aspect of funds. We propose frameworks to automatically assess the sustainability aspect and detect the ESG related themes present in financial texts written in Indian languages. Finally, we evaluate the necessity of having India specific FinNLP models. We summarize these tasks in Figure 6.4. Our datasets and models can be accessed from here[16]. We shall open-source the models and release the datasets after the acceptance of the manuscript.

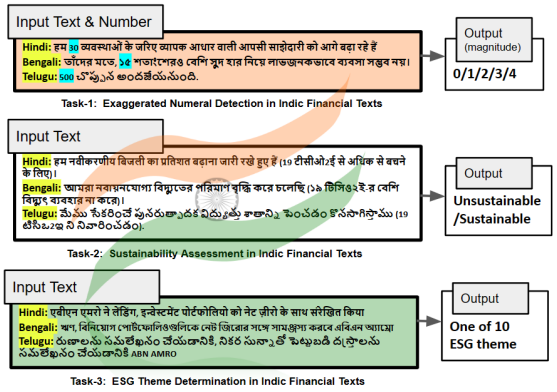

FIGURE 6.4: Financial Natural Language Processing for Indian Languages

Our Contributions

In this chapter, we present **IndicFinNLP** - a collection of 9 datasets corresponding to three FinNLP tasks in three most spoken [17] Indian Languages (Hindi, Bengali, and

[14] https://yourstory.com/2023/07/financial-literacy-is-key-to-unlocking-india-economy (accessed on 11th September, 2023)

[15] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

[16] https://github.com/annonymous-upload/resources/

[17] https://www.superprof.co.in/blog/indian-languages/ (accessed on 11th September, 2023)

Telugu). The tasks are: Exaggerated Numeral Detection, Sustainability Assessment, and ESG Theme Determination in financial texts. To the best of our knowledge, we are the first to create and open-source Indic FinNLP datasets.

6.3.2 Related Works

Over the last few years, research in the FinNLP space has evolved rapidly. Researchers have applied FinNLP on various languages, extending beyond English [270, 271] to encompass Japanese [182], Chinese [181] and multiple European languages like Danish, Spanish, Turkish, etc. [3]. Some of the shared tasks like Financial Narrative Summarization (FNS) have recently moved from monolingual corpus in English [272] to multilingual corpus in English, Greek and Spanish [273]. Similarly, the Financial Table of Content (FinTOC) [274] shared task has embraced multilingualism [275]; the corpus is available in various European languages (English, French, Spanish).

Financial language resources which are available in English are quite varied, ranging from ESG related aspects [142, 276] to stock market related sentiments [271]. Various FinNLP tasks have been proposed in English which use annual reports (EDGAR-CORPUS) [126], Earning Call Transcripts [277], Financial News [278, 279], Analyst reports [280], speeches [281], and Social Media [282]. In addition to this, several FinNLP related shared tasks are organized regularly, including FNS [273], FinTOC [274], FinCausal [283], FinNum [186, 284, 285], etc. Some of the FinNLP specific language models include FinBERT [69], FlangRoBERTA [76], SEC-BERT [151], FinGPT [286], BloombergGPT [2] etc.

Lately, some researchers have been working on India centric FinNLP [287, 288]. However, these works only focus on English texts, ignoring the linguistic diversity present in India. YubiBERT[18] is the only effort towards FinNLP in Indian languages. None of the existing works, including YubiBERT, have released datasets for addressing and bench-marking FinNLP related tasks in Indian languages.

6.3.3 Tasks

We focused on the following three FinNLP tasks in Indian languages.

Task-1: Given the position of an unknown numeral N in a financial text, the task is to determine its magnitude x such that $10^x <= N < 10^{x+1}$ where $x \in \{0,1,2,3,4\}$. Numerals with magnitude more than 4 are treated as 4. Detecting the magnitude of numerals helps in understanding if a certain number in a given context is exaggerated.
Task-2: Given a financial text, the task is to classify it into two classes: sustainable or unsustainable
Task-3: Given a financial text, the task is to determine the ESG theme related to it. The list of ESG themes are mentioned in Table 6.11.

[18] https://www.go-yubi.com/blog/yubibert-a-tiny-fintech-language-model/ (accessed on 11th September, 2023)

Language	0	1	2	3	4
Hindi	2435	3624	1444	2485	652
Bengali	1574	1886	931	1416	323
Telugu	1737	1800	983	1182	314

TABLE 6.9: Task-1 label-wise distribution. 0/1/2/3/4 are the magnitudes

6.3.4 Datasets

In this section, we describe the datasets. For each of the datasets, we used 80%, 10%, and 10% instances selected randomly for training, validation, and testing respectively.

6.3.4.1 Dataset for Task-1

For Task-1, we extracted texts from budget speeches delivered by Finance Ministries of different State Governments and the Central Government of India. We focused on Hindi, Bengali, and Telugu-speaking states– Punjab, Uttarakhand, Haryana, West Bengal, Telangana, and Andhra Pradesh. We considered the budget speeches starting from the year 2011 till 2023 since for most of the states we could get this data for the recent few years. Subsequently, we filtered sizeable volumes of texts in Hindi, Bengali, and Telugu which were related to finance from the Samanantar corpus [263]. For filtering, we extracted the topics using the topic classification model of Antypas et al. [87] and retained the ones belonging primarily to the 'Business & Entrepreneurship' topic. Subsequently, we added all the instances from Task-2 and Task-3. We kept only those texts which had atleast 6 words and one or more numerals in them. For preparing the final dataset, we created separate instances for each numeral in a given text. Statistics about the dataset is presented in Table 6.9.

6.3.4.2 Dataset for Task-2

Since we could not find resources relating to sustainability in the context of India, we translated the existing dataset proposed by Kang and El Maarouf [142] from English to Indian languages (Hindi, Bengali, and Telugu) using AI for Bharat Machine Translation System [263]. To assess the quality of translation, we back-translated the texts in Indian languages to English using the same system. We calculated BERTScore [262] between the original and back-translated sentences in English. Subsequently, we calculated LaBSe [265] based similarity score for the original texts in English and translated texts in Indian languages. We manually looked at the instances and empirically decided the thresholds for BERTScore and cosine similarity to ensure that we retain only high quality instances after translation. For details regarding the dataset, thresholds are presented in Table 6.10.

6.3.4.3 Dataset for Task-3

Although the ESG domain has increasingly become popular, we could not find any dataset related to the Indian domain. Thus, we translated existing resources [276] from English to Indic languages (Hindi, Bengali, and Telugu) using AI for Bharat Machine Translation

Language	BS(F1)	Sim.	Class	#
Hindi	>= 0.90	>=0.75	S	1212
			US	1026
Bengali	>=0.88	>=0.68	S	1203
			US	1025
Telugu	>=0.88	>=0.80	S	1119
			US	953

TABLE 6.10: Task-2 data distribution & thresholds. S=Sustainable, U=Unsustainable. BS=BERTScore, Sim.=Cosine Similarity

ESG Theme	#
climate change	92
corporate governance	91
environmental opportunities	72
product liability	68
natural capital	50
pollution waste	44
human capital	37
corporate behavior	30
social opportunities	27
stake holder opposition	21

TABLE 6.11: Task-3 label-wise distribution for Hindi, Bengali, and Telugu.

System (IndicTrans) [263]. We manually checked each of the instances and made corrections wherever necessary. As the number of instances per ESG label was very low,

we opted for coarse-grained classification. Thus, we mapped each ESG issue to the corresponding themes using the mappings provided by MSCI ESG Research LLC.[19] More details regarding the dataset are presented in Table 6.11.

6.3.5 Experiments and Results

We present our results in this section. The results for all the three tasks are presented in Table 6.12.

6.3.5.1 Task-1

For the task of detecting exaggerated numeral, we extracted multilingual BERT (M-BERT) [1] and IndicBERT [268] based embeddings of the numeral based on a context window of 512 tokens around it. We froze the underlying BERT models and trained LightGBM [265], XG-Boost [71], and Support Vector Machine (SVM) [185] models over it. We observed

[19] https://www.msci.com/documents/1296102/34424357/MSCI+ESG+Ratings+Methodology+-+Access+to+Health+Care+Key+Issue.pdf/683e8c43-7c81-ada7-307d-d9356ec84efb?t=1666182590869 (accessed on 26th September, 2023)

that, the SVM models perform the best in every case. For all the three languages, M-BERT outperforms IndicBERT.

6.3.5.2 Task-2

For the task of Sustainability Assessment from financial texts, we fine-tuned M-BERT and IndicBERT for classification. Unlike Bengali and Hindi, for Telugu, M-BERT outperformed IndicBERT. We pre-fine-tuned the best performing model in each case using Masked Language Modeling (MLM). We observed a notable improvement in the performance for the Bengali dataset only. Finally, we translated the Indic sentences to English using AI For Bharat MT system [263] and evaluated the RoBERTa based model (E-Ro) [10] which was trained using the original texts in English. The E-Ro model outperformed all other models. This reveals the fact that for sustainability assessment, it is better to translate Indic texts to English, and score the models trained on English texts rather than developing separate models for each language.

6.3.5.3 Task-3

For the task of the ESG theme determination, we fine-tuned IndicBERT, M-BERT (MB) and MLM with M-BERT (MLM-MB) with 426 instances. We also trained an M-BERT model with the original English texts (E-MB) and evaluated it using English texts obtained by translating Indic sentences to English using AI For Bharat. For all the three languages, the performance was less than 30% for each of the models. This is because the number of instances per ESG theme was very low (<100 for each label). To address this, we paraphrased the original English texts using a paraphraser [72] and expanded the training and validation set to 4774 and 539 instances respectively. We translated these instances to Indic languages using AI For Bharat. We re-trained all the models with the paraphrased data (referred to as *P in Table 6.12). Paraphrase based models provide significant improvements in performance over the baseline models for all the languages.

6.3.6 Conclusion

In this chapter, we narrated the datasets we created for solving three FinNLP tasks in three Indian Languages (Hindi, Bengali, and Telugu). These tasks are: exaggerated numeral detection, sustainability assessment, and ESG theme determination in financial texts. Subsequently, we released baselines for each of the tasks. We observed that for Task-2 and Task-3 instead of developing separate models for each language, we can simply translate these languages to English and score the existing model's trained English corpus. This improves the overall performance and saves the time and effort needed for training new models. Exploring other tasks in FinNLP and applying them on other low-resources languages are directions for future research. We would also want to work more on the datasets we created and further improve the baselines.

Ts	L	Model	Test			
			Pr	**Re**	**F1**	**Acc**
1	H	MB+LGB	0.63	0.64	0.63	0.64
1	H	IB+LGB	0.44	0.49	0.45	0.49
1	H	MB+XGB	0.63	0.64	0.63	0.64
1	H	IB+XGB	0.46	0.49	0.46	0.49
1	H	MB+SVM	**0.69**	**0.68**	**0.68**	**0.68**
1	B	MB+LGB	0.64	0.64	0.63	0.64
1	B	IB+LGB	0.51	0.51	0.50	0.51
1	B	MB+XGB	0.62	0.62	0.61	0.62
1	B	IB+XGB	0.51	0.50	0.48	0.50
1	B	MB+SVM	**0.66**	**0.65**	**0.65**	**0.65**
1	T	MB+LGB	0.59	0.61	0.59	0.61
1	T	IB+LGB	0.44	0.46	0.44	0.46
1	T	MB+XGB	0.59	0.60	0.59	0.60
1	T	IB+XGB	0.41	0.43	0.41	0.43
1	T	IB+XGB	**0.69**	**0.68**	**0.68**	**0.68**
2	H	IB	0.86	0.86	0.86	0.86
2	H	MB	0.77	0.77	0.77	0.77
2	H	MLM-IB	0.29	0.54	0.38	0.54
2	H	E-Ro	**0.95**	**0.95**	**0.95**	**0.95**
2	B	IB	0.80	0.80	0.80	0.80
2	B	MB	0.76	0.76	0.76	0.76
2	B	MLM-IB	0.81	0.81	0.81	0.81
2	B	E-Ro	**0.92**	**0.92**	**0.92**	**0.92**
2	T	IB	0.79	0.79	0.79	0.78
2	T	MB	0.90	0.89	0.89	0.89
2	T	MLM-IB	0.90	0.90	0.90	0.90
2	T	E-Ro	**0.92**	**0.92**	**0.92**	**0.92**
3	H	IB	0.03	0.17	0.05	0.17
3	H	MB	0.20	0.20	0.08	0.20
3	H	MLM-MB	0.20	0.20	0.11	0.20
3	H	E-MB	0.11	0.30	0.16	0.30
3	H	IB-P	0.12	0.26	0.16	0.26
3	H	MB-P	0.45	0.48	0.44	0.48
3	H	MLM-MB-P	0.43	0.46	0.44	0.46
3	H	E-MB-P	**0.56**	**0.63**	**0.59**	**0.63**
3	B	IB	0.03	0.17	0.05	0.17
3	B	MB	0.03	0.17	0.05	0.17
3	B	MLM-MB	0.11	0.20	0.10	0.20
3	B	E-MB	0.11	0.26	0.14	0.26
3	B	IB-P	0.20	0.30	0.23	0.30
3	B	MB-P	0.40	0.37	0.35	0.37
3	B	MLM-IB-P	0.32	0.37	0.33	0.37
3	B	E-MB-P	**0.55**	**0.59**	**0.55**	**0.59**
3	T	IB	0.03	0.17	0.05	0.17
3	T	MB	0.09	0.24	0.12	0.24
3	T	MLM-MB	0.07	0.22	0.11	0.22
3	T	E-MB	0.07	0.22	0.11	0.22
3	T	IB-P	0.27	0.31	0.22	0.31
3	T	MB-P	0.44	0.46	0.42	0.46
3	T	MLM-MB-P	0.36	0.41	0.37	0.41
3	T	E-MB-P	**0.56**	**0.63**	**0.58**	**0.63**

TABLE 6.12: Results for Tasks (Ts) 1, 2 & 3 for Languages (L) Hindi (H), Bengali (B), & Telugu (T). E=English, -P= -Paraphrased, IB=IndicBERT, MB=M-BERT, XGB=XGBoost, LGB=LightGBM, Pr=Precision, Re=Recall, Acc=Accuracy. **Bold** means the best.

6.4 Predicting success of Indian IPOs

6.4.1 Introduction

Recently, with the growth in the Indian economy, there is a huge surge in interest towards making investments in the stock market.[20] This is due to factors like easing the investment process, allowing low-ticket investments, providing liquidity, generating returns that help to hedge against inflation, etc. A private company transitions from private ownership to public trading through an Initial Public Offering (IPO). This allows the company to raise capital. Investors are lured towards subscribing for IPOs as it helps them to book profits quickly from the listing gains and provides them with access to early-stage companies.

A company needs to submit Draft Red Herring Prospectus (DRHP) to the Securities and Exchange Board of India (SEBI). It contains information regarding the company's fundamentals, business, operations, financial performance, prospects, and legal issues. The DRHP is circulated to potential investors for initial evaluation and feedback. Later, it is finalized and presented as the Red Herring Prospectus (RHP). SEBI oversees and regulates the IPOs. This makes the process of filing an IPO transparent and instills confidence among investors.

The price at which the company's shares are first offered to the public during an IPO is called the Issue Price. There are primarily two types of IPOs in India: Fixed Price Issue, and Book Building Issue. In a fixed price IPO, the share price is established in advance and communicated to investors prior to the opening of the issue. This price is determined by the company in collaboration with the merchant bankers, taking into account various factors such as the company's valuation, assets, liabilities, risks, and growth potential. In a book building IPO, the company provides a price range (from a minimum to a maximum price) rather than a fixed price. Investors have the flexibility to place bids at any price within this range. The final Issue Price is established based on the bids collected after the issue period concludes. Some other types of IPOs are: Rights Issue and Follow-on Public Offer (FPO). Rights Issue allows existing shareholders to purchase new shares. FPO enables companies that are already listed on stock exchanges to raise additional capital.

Based on size of the company, and issue sizes, IPOs can be categorized into Mainboard (MB) IPO and Small and Medium Enterprises (SME) IPO. Compared to MB IPOs, SME IPOs have more relaxed eligibility criteria, allowing smaller companies to access public funding more easily. SME IPOs are primarily vetted by the respective stock exchanges (BSE SME or NSE Emerge), while MB IPOs require scrutiny and approval from SEBI, which includes a more comprehensive review of the prospectus. The differences between these two categories relate to paid-up capital, the minimum number of allottees, underwriting requirements, minimum application size, and market-making practices[21].

[20] https://www.businesstoday.in/markets/top-story/story/demat-accounts-at-all-time-high-cdsl-nse-gain-market-share-425233-2024-04-12 (accessed on 19th August 2024)

[21] https://www.indiainfoline.com/knowledge-center/ipo/difference-between-mainboard-ipo-sme-i (accessed on 23rd August, 2024)

Overall, the Indian IPO landscape remains vibrant and dynamic.[22] Presently, India has been issuing the highest number of IPOs per year.[23] However, factors like market volatility, over-subscription, pricing, influence of investor sentiment, and social media chatter may have adverse effects on the expected return, and the market premium. SME companies have begun to exploit the more lenient regulatory framework. Recently, several instances of fraud have emerged, leading SEBI to issue enforcement orders against some of these firms[24]. SEBI Chairperson, expressed her concerns regarding potential manipulation within the Small and Medium Enterprises (SME) segment. She noted that the market regulator has detected indications of such manipulation, highlighting feedback from the market that suggests misuse of SME listing provisions.[25] These days, lots of retail investors are making speculative investments instead of relying on the fundamentals.[26]. Among the 10 largest first-day gainers in SME IPOs, nine have declined from their closing prices on day one. Additionally, 50% of SME stocks experience a drop after their initial listing day gains.[27] Over half (54%) of the IPO shares allocated to retail investors were sold within a week of the listing.[28] This indicates that a large chunk of investors seeks listing gains. Many investors blindly trust the Grey Market Premium (GMP)[29] for investing in an IPO. GMP refers to the difference between the Issue Price (the price at which shares are offered to the public) and the price at which the shares are traded in the unofficial and unregulated grey market. The Indian IPO market has witnessed significant growth in recent years, attracting speculative investors seeking to capitalize on the potential of emerging markets. These investors need to be educated. [30] Thus, we need a framework for understanding the success of Indian IPOs through thorough examination of the various factors influencing their performance.

Under-pricing in an IPO (Initial Public Offering) refers to the phenomenon where a company's shares are issued at a price lower than their actual market value, resulting in a significant increase in the share price on the first day of trading. The under-pricing percentage is calculated as: Under-pricing Cost = $[(P_m - P_0) / P_0] * 100$, where P_m is the closing price on the first day of trading and P_0 is the Issue Price. Most of the prior research work ([289],[290], [291], [292], [293], [294], [295]) relating to IPO studied under-pricing of MB IPOs.

[22] https://www.businesstoday.in/markets/ipo-corner/story/ipo-flood-rs-115-lakh-crore-worth-of-public-offers-likely-to-hit-markets-in-next-12-months-444559-2024-09-05 (accessed on 6th September, 2024)

[23] https://www.livemint.com/market/stock-market-news/matter-of-great-pride-madhabi-puri-buch-says-india-ipo-issuances-rank-no-1-in-global-league-tables-11722777200397.html (accessed on 19th August 2024)

[24] https://www.linkedin.com/pulse/jay-powell-says-let-party-continue-zerodha-dixtf/ (accessed on 27th August, 2024)

[25] https://finshots.in/archive/nse-cracks-down-on-shady-sme-ipos/ (accessed on 4th September 2024)

[26] https://www.moneycontrol.com/news/business/markets/raamdeo-agrawal-ola-electric-fundamenta
html (accessed on 31st August, 2024)

[27] https://www.financialexpress.com/market/50-sme-stocksnbsp-stumble-after-listing-day-gains-3596400/ (accessed on 31st August, 2024)

[28] https://www.moneycontrol.com/news/business/markets/sebi-study-retail-sold-ipo-12812542.
html (accessed on 3rd September, 2024)

[29] https://www.chittorgarh.com/book-chapter/ipo-grey-market-gmp/28/ (accessed on 19th August 2024)

[30] https://www.financialexpress.com/opinion/educate-retail-investors/3601919/ (accessed on 6th September, 2024)

In this section, we propose a Machine Learning (ML) and Natural Language Processing (NLP) based framework for determining if an IPO in the Indian market will be successful in the short term. We define success in terms of the difference between issue price and the opening-price, highest price, and closing price on the day of the IPO. We studied this separately for MB and SME IPOs. Furthermore, we investigate how much GMP of IPOs are trustable.

Our Contributions

- We present two multi-modal datasets, one for Mainboard IPOs, and the other for Small and Medium Enterprises (SME) IPOs. It consists of various features relating to the company going for IPOs, and other macroeconomic factors.
- We propose a Machine Learning and NLP based decision system for predicting if an IPO will be successful
- We study the impact of various macroeconomic factors, prevailing stock market performance, and financials on the success of an IPO
- Extracted important portions from documents like DRHP, RHP and used them as features for predicting the success of an IPO
- We investigate the relation between GMP and success of an IPO

The remainder of this chapter is organized as follows: related work is presented in Section 6.4.2. The problem has been described in Section 6.4.3. The data preparation steps are mentioned in section 6.4.4. Experiments are described in Section 6.4.5. Section 6.4.6 concludes the chapter.

6.4.2 Related work

The landscape of Initial Public Offerings (IPOs) in India has been a focal point for researchers aiming to understand the various factors influencing their success. This literature review synthesizes findings from multiple studies, highlighting key themes such as under-pricing, regulatory impacts, investor behaviour, and the overall performance of Indian IPOs. For the last few decades, several researchers have studied the IPO market of various countries like India ([289], [293], [294]), the USA ([291], [296], [297]), China ([298]), Turkey ([295], [299]), South Korea ([300]), and Malaysia ([301]). Most of these studies were related to under-pricing, its causes and effects ([289],[290], [291], [292], [293], [294], [295]).

[302] presents a structural review of IPOs in India, covering the period from pre-liberalization to the present. This study reveals significant insights into the evolution of the IPO market. The research emphasizes that IPO volume and valuation are heavily influenced by regulatory changes [303], economic growth [304] and global financial conditions. Notably, the fiscal year 2022-2023 witnessed a surge in IPO activity, driven by increased retail investor participation and a focus on technology and startups. This study underscores the importance of understanding India's unique socio-economic factors that shape the IPO landscape.

Most of the research works focus either on determining the short-run under-pricing ([305], [296], [291], [306], [307]) or the long-run underperformance ([308]). Traders are interested in short-run under-pricing, while investors are interested about long-run performance. Factors such as a company's age [309], pricing mechanism, retail subscription, market capitalization [289], size, return on assets, financial leverage, and the reputation of its underwriters [310] are crucial in determining the initial offering price. Additionally, political stability [311], market conditions — including general economic sentiment [**?**] and industry trends [313] also have a substantial impact on IPO performance and pricing. Other factors such as group affiliation [314], regulatory environment [303], and effective communication strategies with investors are recognized as critical in increasing the chances of IPO success.[31] A study [315] analysing the relationship between IPO offer price ranges and initial demand among investors indicates that lower-priced IPOs tend to attract less trading activity post-listing. This research highlights that institutional investors favour higher-priced offerings, while retail investors are more likely to subscribe to lower-priced IPOs. This behaviour contributes to the under-pricing phenomenon, which is positively correlated with over-subscription rates in the Qualified Institutional Buyers (QIB) category. Another empirical study [305] explores the impact of board composition and promoter ownership on IPO under-pricing. The findings suggest that reputable boards can mitigate information asymmetry, thereby reducing under-pricing. Conversely, high promoter ownership is associated with increased under-pricing, indicating potential conflicts of interest that may arise from insider control. Research [294] indicates that various factors influence the performance of Indian IPOs, including market conditions, pricing strategies, and the timing of offerings [308], [316]. A comprehensive analysis [294] of 290 IPOs from 2007 to 2017 reveals significant under-pricing, with an average raw return of 17.90% in the short term, which declines sharply after nine months. This suggests that while IPOs may perform well initially, long-term investments may not yield favourable outcomes. Additionally, the study [294] highlights that IPOs issued after 2013 generally performed better, with shorter listing delays correlating with higher returns. The impact of offer price and size on performance is also notable, with mid-range offerings typically yielding better results [293].

Research [308] reveals that initial under-pricing often leads to long-run under performance, with IPOs failing to maintain their initial momentum. While under-pricing offers short-term gains, the long-term performance of Indian IPOs presents a more nuanced picture. [317], [318], and [319] present a comparative analysis between the short term and long term performance.

The role of investor sentiment and macroeconomic conditions [320] in IPO performance are other critical areas of exploration. [291] indicates that positive media sentiment can significantly influence retail investor perceptions and demand, thereby affecting first-day returns. This suggests that external perceptions play a crucial role in shaping IPO success and investor behaviour. A study [321] focusing on pre-IPO earnings management reveals that firms utilising reputable investment banks are less likely to manipulate earnings, highlighting the importance of transparency in the IPO process. This research indicates that improved governance mechanisms can enhance investor confidence and potentially reduce under-pricing. Several studies ([308], [322]) suggest that investors perceive under-pricing as a signal of quality and future growth potential, leading to increased demand and subsequent price appreciation. However, the motivations for under-pricing also include

[31] https://fastercapital.com/topics/the-role-of-communication-in-a-successful-ipo.html (accessed on 1st September 2024

reducing risk for issuers and attracting investors, with potential consequences for long-term performance [317].

[323] presents a comparative analysis of IPO returns between India and China illustrates distinct return patterns influenced by differing political and economic systems. While Indian IPOs tend to exhibit positive initial returns, the sustainability of these returns diminishes over time, emphasizing the need for investors to consider risk factors associated with IPO investments in emerging markets.

While most of the research papers deal primarily with numeric features, there are a few which uses text based features as well ([291], [297]). [291] studies sentiment of news articles, [324] investigated effect of sentiments of tweets, while [297] analyses IPOs' prospectuses from the SEC database to estimate success of IPOs. Similarly, most of the research paper used regression models to predict the performance of the IPOs. Only a handful of research papers used advanced machine learning based algorithms like Support Vector Machines Random Forests [295], [290], and fuzzy techniques [292].

The literature on Indian IPOs presents a multifaceted view of the factors influencing their success. The prevalence of under-pricing, the impact of regulatory changes, and the role of investor behaviour are critical themes that emerge from the research. However, there are certain research gaps which can be addressed. Firstly, conducting sector-specific analyses of IPO performance may reveal unique challenges and success factors that are not captured in broader studies. Secondly, most of the prior works defined under-pricing in terms of closing price of the listing day. But, for benefiting short-term traders, it may be interesting to investigate how open price and highest price on listing day varied. Thirdly, the impact of grey market price and ratings of Indian IPOs by top analysts have not yet been thoroughly studied. Lastly, although Large Language Models (LLM) have shown remarkable performance in various financial tasks ([325], [326]), no one has ever used them to analyse DRHP and RHP of companies for predicting IPO related success. In this section, we would like to address these research gaps.

6.4.3 Problem statement

We define the success of a company's IPO by comparing issue price with the opening-price, highest price, and closing price on the listing day of the IPO.

- Opening-price
  · Predict if the opening-price on the listing day of the IPO will be greater than the issue price of the IPO. We refer to this as predicting the direction of opening-price movement.
  · Predict under-pricing with respect to opening-price on the listing day of the IPO, i.e. (opening-price - issue price)/(issue price)
- Highest price
  · Predict if the highest price on the listing day of the IPO will be greater than the issue price of the IPO
  · Predict under-pricing with respect to the highest price on the listing day of the IPO, i.e. (highest price - issue price)/(issue price)

- Closing price
  · Predict if the closing price on the listing day of the IPO will be greater than the issue price of the IPO
  · Predict under-pricing with respect to closing price on the listing day of the IPO, i.e. (closing price - issue price)/(issue price)

We conduct the separate experiments for SME and Mainboard IPOs with the same objective of predicting the success of IPOs.

6.4.4 Data preparation

Firstly, we prepared a list of companies which went for Mainboard IPO from 2009 to 2023. Similarly, we prepared another list of companies which went for SME IPO from 2017 to 2023. We collected this data from chittorgarh.com[32]. The date ranges were decided based on availability of the dataset. After removing all the instances where the IPO was withdrawn or no data was present, we were left with data for 418 Mainboard, and 681 SME IPOs.

For both SME and Mainboard IPOs, to train the models, we used the data till 2022. The data for the year 2023 was used to evaluate the performances of the trained models. We present the train, test split in Table 6.13. Figures 6.7 and 6.8 represent how the success rate of Mainboard and SME IPOs respectively varied over the years.

Subsequently, we collected historical values of Indian stock market indices, i.e., Nifty 50 and Nifty VIX at daily, weekly, and monthly granularities from investing.com[33]. We further extracted news articles related to the IPOs of the companies from Economic Times news portal. [34] We could get 215 and 33 news articles which presented information regarding companies participating in Mainboard and SME IPO respectively. In order to maintain data quality, we considered only those news articles which covered a single company. Furthermore, to get an understanding of the effect of various macroeconomic factors, we added them as features to our dataset. These features are: GDP per capita growth (annual), GDP growth (annual), GDP (current) Unemployment rate, stocks traded value, Personal remittances, net trade in goods and services, GNI per capita growth, Inflation, consumer prices, GNI (current), Foreign direct investment. For a given IPO, we obtained the values of these features from the World Bank's website.[35]. We obtained information regarding the sector and industry of the organization from stocksonfire.in[36]. We present the sector-wise and industry-wise distributions in Figures 6.5 and 6.6 respectively. We engineered additional features. They are the success rate of the IPOs launched in the previous quarter, and within the last 90 days from the launch of a given IPO. Information regarding the financials of the company, subscription rate till the penultimate day for subscription, ownership, ratings were obtained from chittorgarh.com[37], and DRHP, RHP

[32] https://www.chittorgarh.com/ipo/ipo_dashboard.asp (accessed on 23rd January, 2024)
[33] https://in.investing.com/indices/s-p-cnx-nifty-historical-data?end_date=1714933800&interval_sec=weekly&st_date=1136053800&interval_sec=daily
[34] https://economictimes.indiatimes.com/archive.cms
[35] https://data.worldbank.org/country/IN (accessed on 25th June, 2024)
[36] https://stocksonfire.in/trading-ideas/nse-stocks-sector-wise-sorting-excel-sheet/ (accessed on 25th June, 2024)
[37] https://www.chittorgarh.com/ipo/ipo_dashboard.asp (accessed on 25th June, 2024)

Type	Train	Test
Mainboard	361	57
SME	498	183

TABLE 6.13: Training-Test Split

reports present in the SEBI[38], National Stock Exchange (NSE)[39], and Bombay Stock Exchange (BSE)[40] websites. Subscription rate of a day is declared after the market closes on that day. We considered subscription rate till the penultimate day for subscription, as we want the investors to make a decision to opt for the IPO on the final day of subscription. In addition to this, we extracted texts, tables, and images from the prospectus (RHP, DRHP) of the companies. For these images, we performed Optical Character Recognition (OCR) using Tesseract[41] to retrieve texts. For a given company, we stored the extracted content in a JavaScript Object Notation (JSON) file, where the keys corresponded to the pages in the prospectus. We compiled a list of twenty-five questions which investors look for in the prospectus. These are presented in Table 6.17 of §6.4.7. These questions were formulated after interviewing several seasoned IPO investors and referring to eight reputed financial web-sites.[42][43] For each JSON file, we transformed the content of each page into embeddings using Nomic [327]. Nomic has an 8,192 context-length text encoder. Similarly, using Nomic we transformed each curated question to embeddings. From a given prospectus of a company, we retrieved pages relevant to the curated questions using cosine similarity and BM25 [328]. Cosine similarity was used for semantic matching, whereas BM25 was used for lexical matching. Subsequently, we passed the retrieved pages and the corresponding question to an LLM, Llama 3-8B [329] to generate the final answer. This approach is widely known as Retrieval-Augmented Generation (RAG). The RAG-based system treats each page of the PDF as a separate chunk.

Finally, for comparison, we obtained the GMP of the companies from investorgrain.com[44]. We could not use GMP as a feature because we could get only the GMP values of the IPOs which were launched in the year 2019 or later. The final list of features and their descriptions are presented in Table 6.18 of §6.4.8.

[38] `https://www.sebi.gov.in/` (accessed on 25th June, 2024)
[39] `https://www.nseindia.com/` (accessed on 25th June, 2024)
[40] `https://www.bseindia.com/` (accessed on 25th June, 2024)
[41] `https://github.com/tesseract-ocr/tesseract` (accessed on 25th June, 2024)
[42] https://www.motilaloswal.com/article-details/what-is-a-draft-red-herring-prospectus-and-why-is-it-important-for-investors/5259
https://www.fisdom.com/what-to-look-for-in-an-rhp-before-investing-in-ipo/
https://www.indiainfoline.com/knowledge-center/ipo/what-is-a-draft-red-herring-prospectus
https://www.nism.ac.in/2024/01/understanding-drhp-rhp-and-prospectus/
https://groww.in/blog/things-you-must-know-about-rhp https://www.chittorgarh.com/book-chapter/ipo-prospectus/18/
[43] https://www.5paisa.com/stock-market-guide/ipo/things-to-know-in-rhp
https://www.kotaksecurities.com/articles/6-things-to-look-for-in-a-draft-red-herring-prospectus/ (accessed on 7th September, 2024)
[44] `https://www.investorgain.com/report/live-ipo-gmp/331/` (accessed on 25th June, 2024)

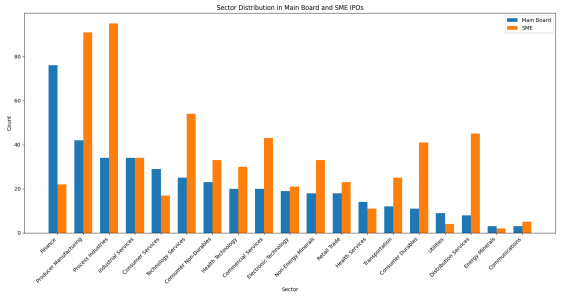

FIGURE 6.5: Sector-wise Distribution

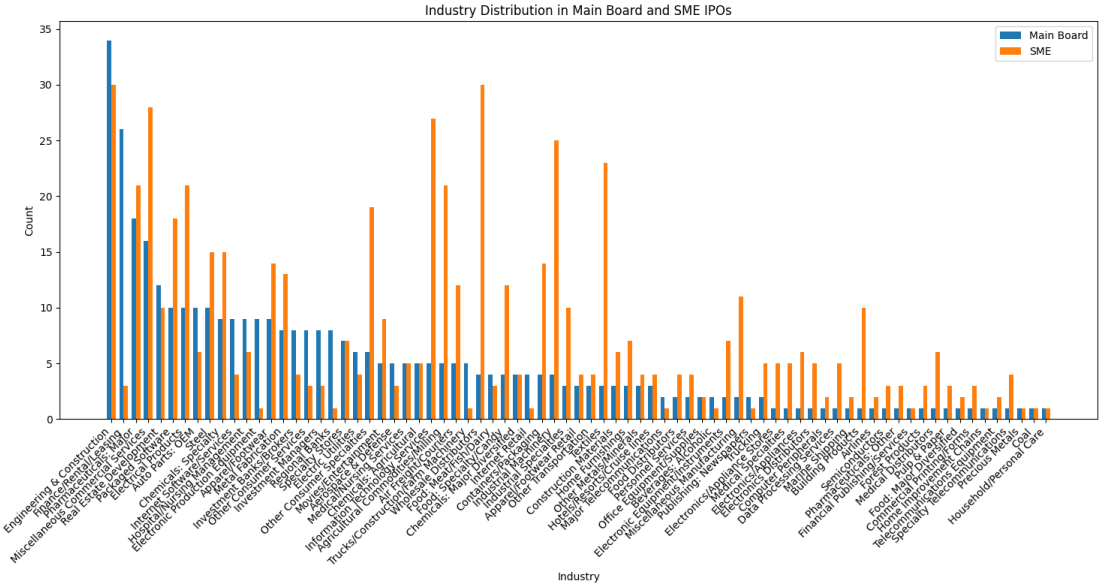

FIGURE 6.6: Industry-wise Distribution

6.4.5 Experiments and results

We present the overall experimental framework in Figure 6.9.

6.4.5.1 Predicting direction of opening, highest, and close prices of the listing day

Our objective is to train separate models for predicting the direction of Opening, Highest and Closing prices. We started with using only the numeric and categorical (N-C) features for prediction. These numeric features are: Issue Price, Lot Size, Market Variables (Nifty50, VIX), Macroeconomic variables (GDP, Stocks traded, Unemployment rate, etc.), Subscription rate per category (QIB, NII, Retail, etc.) up to the penultimate day for subscription, Success rate of the IPOs in the previous quarter and the last 90 days, recommendations by brokers and members, face value of a share, Shares and Amount allocated to per category (Retail, HNI, etc.), Assets, Revenue, Profit After Tax, Net Worth, Reserves and Surplus, Total Borrowing, and Total Income. A comprehensive list of the features in presented in Table 6.18 We used the AutoML open-source library

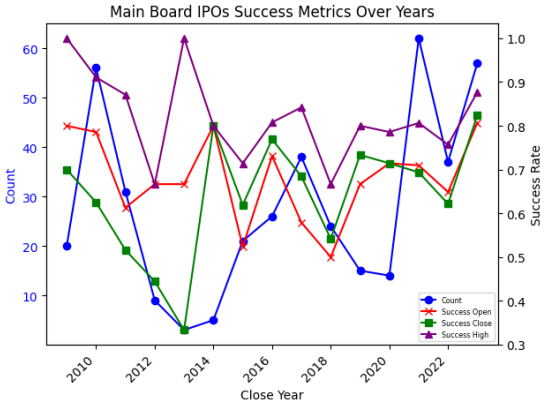

FIGURE 6.7: Success rates over the year for Mainboard IPOs

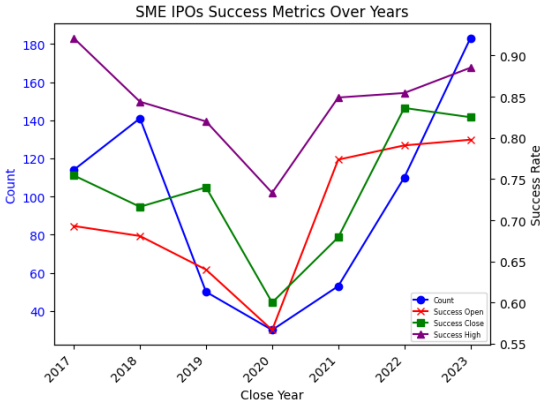

FIGURE 6.8: Success rates over the year for SME IPOs

developed by the H2O team[45] to train five kinds of models: Generalized Linear Model (GLM), Distributed Random Forest (DRF) [66], neural networks based Deep Learning (DL) models, XG-Boost (XGB) [71], and Gradient Boosting Machine (GBM) [67]. Subsequently, we ensembled (Ens) these models to get the final predictions.

Later on we used text content (T) related to the company, (i.e. columns 'full_text_content' and answer_of_question_n) in the modelling process. Moreover, we concatenated news (Nw) content related to the company (i.e. column: news_content) with the text content (T). We could not use news content as a separate feature because it was present in less than 50% and 10% instances for mainboard and SME respectively. To include the texts as features, we firstly extracted their embeddings separately using Nomic [327]. We used these embeddings only as input features to train five kinds of models (GLM, DRF, DL, XGB, and GBM) leveraging H2O AutoML library for classification. For each text feature, separate ML models were trained for predicting the direction of opening, highest and closing prices. We selected the best model in each case and appended the probabilities of the positive class as features to the list of existing numeric and categorical features. Direction equals to 1 (i.e. opening-price on the listing day of the IPO greater than the issue price of the IPO) is referred to as a positive class.

[45] `https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html` (accessed on 8th September, 2024)

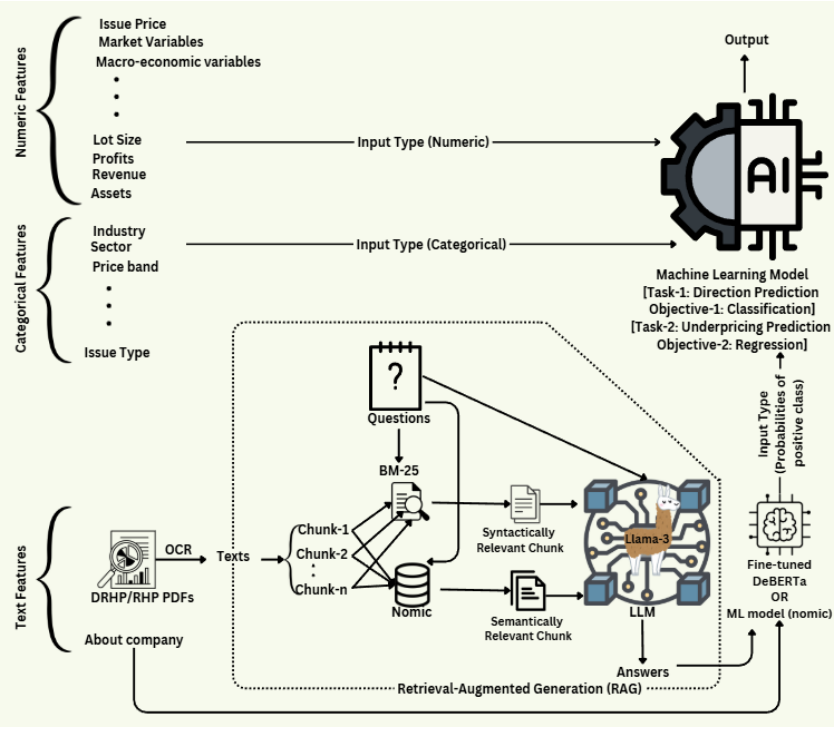

FIGURE 6.9: Methodology

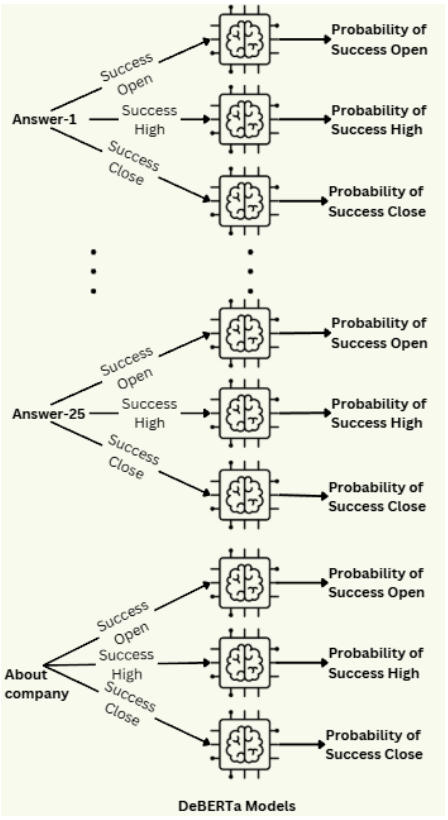

FIGURE 6.10: DeBERTa models

Furthermore, as depicted in Figure 6.10 we fine-tuned 26 DeBERTa [330] models for classification corresponding to each of the 26 text features (column: full_text_content,

answer_of_question_1 to 25). This process was repeated three times to obtain probability of the positive class i.e. direction of price movement with respect to opening, highest, and closing prices. We use these probabilities as inputs to the final machine learning based model. We replaced the previously mentioned Nomic based probabilities with the probabilities of the positive class obtained by fine-tuning separate DeBERTa-base [330] models. The fine-tuning was repeated three times separately, i.e. for predicting the direction of opening, highest, and closing prices. Each of the DeBERTa models was trained for three epochs with learning rate of 2e-5 and batch size of 8. At a time, depending on the objective, we use one out of these three models. This means for predicting success in terms of direction or under-pricing with respect to the opening-price, we use the corresponding DeBERTa model which was fine-tuned for classifying direction of opening-price.

The models for MB and SME IPOs were trained independently. For bench marking, we prompted a gemini-1.5-flash [331] model with all the necessary details for predicting the objectives. We repeated this with Llama-3.2 3b model. The details of the prompts are mentioned in section 6.4.9. We report Area Under the ROC curve (AUC), and F1 score for class 0 (i.e., F1(0)) and class 1 (i.e., F1 (1)) in Table 6.14.

Analysing the results, we observe that for predicting the direction of opening-prices, Deep Learning (DL) and Gradient Boosting Machines (GBM) trained with numeric and categorical feature only performs the best for MB and SME respectively. However, for predicting direction of highest prices, texts features do play a role. In this case, ML models trained with probability of positive class obtained by fine-tuning DeBERTa models as features dominate in terms of F1 (1) for MB. But, in case of SME, XGB model trained with numerical, categorical features, and probabilities of positive class (obtained from the best performing ML models trained using Nomic embeddings) outperforms all others in terms of F1 (1). Finally, for closing price, XGB models trained with numeric and categorical feature only performs the best in terms of F1 (1) for both MB and SME.

			Mainboard			**SME**		
Model	**Type**	**Input**	**AUC**	**F1 (0)**	**F1 (1)**	**AUC**	**F1 (0)**	**F1 (1)**
GLM	O	N-C	0.824	0.517	0.915	**0.698**	0.076	0.887
DRF	O	N-C	0.597	0.592	0.893	0.669	0.688	0.890
DL	O	N-C	**0.903**	**0.781**	**0.947**	0.679	0.003	0.887
XGB	O	N-C	0.773	0.233	0.893	0.670	0.077	0.887
GBM	O	N-C	0.712	0.597	0.893	0.606	**0.736**	**0.893**
Ens	O	N-C	0.824	0.278	0.902	**0.698**	0.027	0.887
GLM	O	N-C-Tn	0.466	0.345	0.893	0.534	0.427	0.887
DRF	O	N-C-Tn	0.372	0.268	0.893	0.517	0.356	0.887
DL	O	N-C-Tn	0.478	0.033	0.893	0.463	0.128	0.887
XGB	O	N-C-Tn	0.419	0.038	0.893	0.570	0.184	0.889
GBM	O	N-C-Tn	0.451	0.001	0.893	0.581	0.000	0.887
Ens	O	N-C-Tn	0.441	0.000	0.893	0.539	0.002	0.887
GLM	O	N-C-Tn-Nw	0.506	0.258	0.893	0.534	0.427	0.887
DRF	O	N-C-Tn-Nw	0.465	0.274	0.893	0.517	0.356	0.852
DL	O	N-C-Tn-Nw	0.502	0.005	0.893	0.463	0.128	0.887
XGB	O	N-C-Tn-Nw	0.544	0.229	0.911	0.570	0.184	0.889
GBM	O	N-C-Tn-Nw	0.577	0.000	0.893	0.580	0.000	0.887
Ens	O	N-C-Tn-Nw	0.545	0.001	0.893	0.539	0.002	0.887
GLM	O	N-C-Td	0.470	0.125	0.893	0.554	0.034	0.887

DRF	O	N-C-Td	0.370	0.387	0.893	0.483	0.231	0.888
DL	O	N-C-Td	0.520	0.000	0.893	0.561	0.000	0.888
XGB	O	N-C-Td	0.455	0.048	0.893	0.510	0.067	0.888
GBM	O	N-C-Td	0.472	0.054	0.893	0.524	0.001	0.888
Ens	O	N-C-Td	0.470	0.000	0.893	0.508	0.000	0.888
Gemini	O	N-C-T-Nw	0.524	0.000	0.880	0.479	0.000	0.855
Llama	O	N-C-T-Nw	0.533	0.167	0.891	0.530	0.128	0.863
GLM	H	N-C	0.679	0.667	0.934	0.669	0.519	0.939
DRF	H	N-C	0.387	0.741	0.934	0.617	0.719	0.939
DL	H	N-C	**0.799**	0.157	0.934	**0.692**	0.000	0.939
XGB	H	N-C	0.651	0.623	0.943	0.600	0.483	0.939
GBM	H	N-C	0.651	0.356	0.934	0.657	0.698	0.939
Ens	H	N-C	0.768	0.622	0.934	0.682	0.182	0.939
GLM	H	N-C-Tn	0.474	**0.833**	0.934	0.574	**0.912**	0.939
DRF	H	N-C-Tn	0.577	0.459	0.934	0.489	0.586	0.939
DL	H	N-C-Tn	0.465	0.012	0.934	0.527	0.846	0.939
XGB	H	N-C-Tn	0.516	0.212	0.934	0.621	0.087	0.939
GBM	H	N-C-Tn	0.598	0.000	0.934	0.559	0.001	0.939
Ens	H	N-C-Tn	0.537	0.006	0.934	0.486	0.742	0.939
GLM	H	N-C-Tn-Nw	0.446	0.788	0.934	0.586	**0.912**	0.939
DRF	H	N-C-Tn-Nw	0.511	0.521	0.934	0.507	0.586	0.939
DL	H	N-C-Tn-Nw	0.471	0.551	0.934	0.529	0.696	0.939
XGB	H	N-C-Tn-Nw	0.466	0.097	0.934	0.626	0.087	**0.942**
GBM	H	N-C-Tn-Nw	0.511	0.285	0.934	0.565	0.045	0.939
Ens	H	N-C-Tn-Nw	0.460	0.002	0.934	0.499	0.759	0.939
GLM	H	N-C-Td	0.466	0.557	**0.935**	0.654	0.019	0.939
DRF	H	N-C-Td	0.350	0.625	**0.935**	0.533	0.346	0.939
DL	H	N-C-Td	0.563	0.006	**0.935**	0.642	0.000	0.939
XGB	H	N-C-Td	0.411	0.509	**0.935**	0.638	0.112	0.939
GBM	H	N-C-Td	0.443	0.255	**0.935**	0.625	0.065	0.939
Ens	H	N-C-Td	0.477	0.497	**0.935**	0.621	0.006	0.939
Gemini	H	N-C-T-Nw	0.490	0.000	0.924	0.491	0.000	0.929
Llama	H	N-C-T-Nw	0.480	0.000	0.914	0.508	0.139	0.840
GLM	Cl	N-C	0.766	0.339	0.913	0.712	0.506	0.904
DRF	Cl	N-C	0.499	**0.599**	0.904	0.534	0.704	0.909
DL	Cl	N-C	**0.804**	0.001	0.911	**0.736**	0.064	0.904
XGB	Cl	N-C	0.728	0.181	**0.931**	0.697	0.312	**0.911**
GBM	Cl	N-C	0.614	0.443	0.904	0.636	0.869	0.907
Ens	Cl	N-C	0.766	0.084	**0.931**	0.724	0.489	0.906
GLM	Cl	N-C-Tn	0.600	0.304	0.904	0.527	0.784	0.904
DRF	Cl	N-C-Tn	0.531	0.207	0.904	0.509	0.483	0.907
DL	Cl	N-C-Tn	0.625	0.414	0.904	0.509	0.084	0.904
XGB	Cl	N-C-Tn	0.529	0.026	0.904	0.531	0.032	0.904
GBM	Cl	N-C-Tn	0.529	0.000	0.904	0.530	0.001	0.904
Ens	Cl	N-C-Tn	0.506	0.011	0.904	0.550	0.009	0.904
GLM	Cl	N-C-Tn-Nw	0.580	0.318	0.904	0.529	0.769	0.904
DRF	Cl	N-C-Tn-Nw	0.455	0.172	0.904	0.523	0.483	0.909
DL	Cl	N-C-Tn-Nw	0.604	0.404	0.904	0.461	0.018	0.904

XGB	Cl	N-C-Tn-Nw	0.479	0.128	0.904	0.518	0.115	0.904
GBM	Cl	N-C-Tn-Nw	0.445	0.000	0.904	0.549	**0.986**	0.905
Ens	Cl	N-C-Tn-Nw	0.449	0.005	0.904	0.523	0.910	0.905
GLM	Cl	N-C-Td	0.532	0.443	0.913	0.572	0.025	0.904
DRF	Cl	N-C-Td	0.526	0.481	0.913	0.467	0.143	0.904
DL	Cl	N-C-Td	0.538	0.010	0.913	0.571	0.000	0.904
XGB	Cl	N-C-Td	0.540	0.296	0.911	0.546	0.102	0.904
GBM	Cl	N-C-Td	0.545	0.403	0.904	0.475	0.000	0.904
Ens	Cl	N-C-Td	0.538	0.285	0.913	0.481	0.000	0.904
Gemini	Cl	N-C-T-Nw	0.500	0.000	0.904	0.486	0.000	0.876
Llama	Cl	N-C-T-Nw	0.485	0.125	0.845	0.502	0.139	0.840

TABLE 6.14: Results of predicting direction of opening, highest, and close prices. O=Open, H=Highest, Cl=Close, N Numeric, C= Categorical, T = Raw Texts, Nw = News, Tn = Text Embeddings Probability (Nomic), Td = Text Embeddings Probability (DeBERTa), Llama = Llama 3.2 3b, Ens = Ensemble. The best model (highest AUC, F1) of each type is highlighted in bold.

			Mainboard		**SME**	
Model	**Type**	**Input**	**MAE**	**MSE**	**MAE**	**MSE**
GLM	O	N-C	0.208	0.079	0.264	0.136
DRF	O	N-C	0.247	0.131	0.265	0.169
DL	O	N-C	0.184	0.066	0.248	0.152
XGB	O	N-C	0.231	0.084	0.282	0.157
GBM	O	N-C	0.223	0.100	0.291	0.214
Ens	O	N-C	0.171	0.057	0.248	**0.127**
GLM	O	N-C-Tn	0.202	0.083	0.264	0.133
DRF	O	N-C-Tn	0.266	0.142	0.289	0.206
DL	O	N-C-Tn	0.181	0.068	0.253	0.152
XGB	O	N-C-Tn	0.172	0.055	0.275	0.151
GBM	O	N-C-Tn	0.248	0.123	0.287	0.215
Ens	O	N-C-Tn	0.176	0.060	**0.239**	0.133
GLM	O	N-C-Tn-Nw	0.201	0.082	0.264	0.133
DRF	O	N-C-Tn-Nw	0.264	0.139	0.290	0.207
DL	O	N-C-Tn-Nw	0.177	0.057	0.250	0.155
XGB	O	N-C-Tn-Nw	0.196	0.067	0.296	0.168
GBM	O	N-C-Tn-Nw	0.248	0.123	0.286	0.215
Ens	O	N-C-Tn-Nw	0.171	**0.051**	0.247	0.139
GLM	O	N-C-Td	0.209	0.089	0.265	0.131
DRF	O	N-C-Td	0.256	0.133	0.278	0.192
DL	O	N-C-Td	**0.167**	0.058	0.258	0.150
XGB	O	N-C-Td	0.199	0.067	0.284	0.160
GBM	O	N-C-Td	0.241	0.124	0.287	0.206
Ens	O	N-C-Td	0.176	0.057	0.265	0.136
Gemini	O	N-C-T-Nw	0.259	0.135	0.331	0.264
GLM	H	N-C	0.232	0.106	0.276	0.150
DRF	H	N-C	0.299	0.185	0.289	0.206
DL	H	N-C	0.223	0.099	0.262	0.150

XGB	H	N-C	0.229	0.092	0.305	0.174
GBM	H	N-C	0.252	0.138	0.295	0.219
Ens	H	N-C	0.240	0.102	0.276	0.153
GLM	H	N-C-Tn	0.224	0.117	0.287	0.148
DRF	H	N-C-Tn	0.309	0.206	0.301	0.223
DL	H	N-C-Tn	0.206	0.092	**0.263**	0.165
XGB	H	N-C-Tn	0.235	0.094	0.304	0.181
GBM	H	N-C-Tn	0.279	0.169	0.301	0.237
Ens	H	N-C-Tn	0.206	0.087	0.269	0.276
GLM	H	N-C-Tn-Nw	0.223	0.117	0.287	0.148
DRF	H	N-C-Tn-Nw	0.311	0.207	0.301	0.224
DL	H	N-C-Tn-Nw	0.203	0.094	0.274	0.176
XGB	H	N-C-Tn-Nw	0.221	0.084	0.317	0.178
GBM	H	N-C-Tn-Nw	0.279	0.169	0.301	0.237
Ens	H	N-C-Tn-Nw	**0.193**	**0.080**	0.269	0.149
GLM	H	N-C-Td	0.220	0.112	0.276	0.147
DRF	H	N-C-Td	0.304	0.199	0.297	0.221
DL	H	N-C-Td	0.219	0.103	0.273	0.154
XGB	H	N-C-Td	0.205	0.065	0.313	0.184
GBM	H	N-C-Td	0.269	0.157	0.301	0.230
Ens	H	N-C-Td	0.211	0.087	0.269	**0.144**
Gemini	H	N-C-T-Nw	0.291	0.185	0.354	0.312
GLM	Cl	N-C	0.211	0.091	0.279	0.149
DRF	Cl	N-C	0.296	0.182	0.288	0.192
DL	Cl	N-C	0.243	0.122	0.259	0.168
XGB	Cl	N-C	0.265	0.115	0.297	0.181
GBM	Cl	N-C	0.239	0.119	0.304	0.229
Ens	Cl	N-C	0.237	0.097	0.268	0.148
GLM	Cl	N-C-Tn	0.206	0.088	0.278	0.146
DRF	Cl	N-C-Tn	0.289	0.179	0.307	0.238
DL	Cl	N-C-Tn	**0.194**	0.083	**0.256**	0.158
XGB	Cl	N-C-Tn	0.216	0.090	0.308	0.194
GBM	Cl	N-C-Tn	0.275	0.166	0.306	0.236
Ens	Cl	N-C-Tn	0.199	0.078	0.262	**0.144**
GLM	Cl	N-C-Tn-Nw	0.206	0.088	0.279	0.146
DRF	Cl	N-C-Tn-Nw	0.288	0.178	0.314	0.248
DL	Cl	N-C-Tn-Nw	0.201	0.089	0.264	0.161
XGB	Cl	N-C-Tn-Nw	0.222	0.089	0.294	0.169
GBM	Cl	N-C-Tn-Nw	0.281	0.166	0.301	0.232
Ens	Cl	N-C-Tn-Nw	0.201	0.077	0.262	**0.144**
GLM	Cl	N-C-Td	0.208	0.084	0.288	0.149
DRF	Cl	N-C-Td	0.275	0.165	0.305	0.223
DL	Cl	N-C-Td	0.200	0.079	0.273	0.160
XGB	Cl	N-C-Td	0.225	0.089	0.328	0.202
GBM	Cl	N-C-Td	0.256	0.145	0.311	0.239
Ens	Cl	N-C-Td	0.198	**0.068**	0.296	0.154
Gemini	Cl	N-C-T-Nw	0.267	0.146	0.348	0.295

TABLE 6.15: Results of predicting under-pricing with respect to opening, highest, and close prices.O=Open, H=Highest, Cl=Close, N Numeric, C= Categorical, T = Raw Texts, Nw = News, Tn = Text Embeddings Probability (Nomic), Td = Text Embeddings Probability (DeBERTa), Ens = Ensemble. Best model (lowest MAE, MSE) of each type is highlighted in bold.

6.4.5.2 Predicting under-pricing with respect to opening, highest, and close prices of the listing day

Our objective is to train separate models for predicting under-pricing with respect to Opening, Highest and Closing prices. Wherever we have these prices available in both BSE and NSE, we preferred to use the NSE prices. Similar to the previous section, we initiated the experiments with only the numeric features for prediction and added text features later. We trained the same five kinds of machine learning models described previously for regression. Everything else other than the objective was kept the same. Our evaluation metrics were: Mean Squared Error (MSE) and Mean Absolute Error (MAE). We present the results in Table 6.15. We also prompted Llama 3.2 3b model under zero-shot setting for predicting under-pricing of MB IPOs. It could not predict the under-pricing with respect to opening, highest, and closing prices for 36.84%, 17.54%, and 22.81% cases respectively. Thus, we did not repeat the experiments to predict under-pricing for SME IPOs.

We observed that for predicting under-pricing of SME IPOs with respect to opening-price, the Ensemble model trained using numerical inputs, categorical inputs, and probabilities of positive class as features performed the best in terms of MAE. These probabilities of positive class were obtained from best performing ML models trained using Nomic embeddings of text columns for predicting the direction of opening-prices. However, in case of predicting under-pricing of MB IPOs with respect to Opening-price, Deep Learning (DL) model trained using numerical, categorical inputs, and probabilities obtained by fine-tuning DeBERTa models performed the best in terms of MAE.

Similarly, for predicting under-pricing with respect to highest prices, Ensemble and DL models trained with probabilities of positive classes along with numeric and categorical features performed the best in terms of MAE for MB and SME respectively. As mentioned before, the probabilities of positive class were obtained from the best performing ML models trained using Nomic embeddings of text columns for predicting the direction of highest prices. We also observed that in case of MB, News content (Nw) played a role.

Finally, for predicting under-pricing with respect to closing prices, DL models trained with probabilities of positive classes along with numeric and categorical features performed the best in terms of MAE for both MB and SME. As discussed previously, the probabilities of positive class were obtained from the best performing ML models trained using Nomic embeddings of text columns for predicting the direction of closing prices.

6.4.5.3 Experiments related to Grey Market Premium

To understand the relation between GMP and success of the IPOs, we collected GMP of the 287 and 385 companies which went for IPOs in Mainboard and SME respectively. In

Table 6.16, we present how the listing price and issue price varied when GMP was negative, zero, and positive. We considered all the companies that went for Mainboard or SME IPO from 1st January 2019 to 12th July 2024. We eliminated those cases where GMP were not available. For the year, 2023 we present the results separately as it corresponds to the test set on which we are doing all our evaluation. It is interesting to note that at an overall level, GMP values aligned with the difference between Listing Prices and Issue Prices in 80.29% cases for the Mainboard IPOs and 21.29% cases for the SME IPOs.

Using GMP, we predicted the under-pricing with respect to opening-price on the listing day i.e. ((GMP + issue price) - issue price)/(issue price) = (GMP)/(issue price). We compared it with the actual under-pricing, i.e. (opening-price - issue Price)/(issue price). For the entire mainboard dataset, we obtained MAE, and MSE as 0.109, and 0.031 respectively. For the year 2023 only, the values of MAE, and MSE for mainboard IPOs are 0.091, and 0.019 respectively. Similarly, for the entire SME dataset, we obtained MAE, and MSE as 0.751, and 1.509 respectively. For the year 2023 only, the values of MAE, and MSE for SME IPOs are 0.531, and 0.771 respectively. Based on availability of data in investorgain.com, we present this analysis with respect to opening-price only.

We observe that GMP does a good job for predicting success of mainboard IPOs. However, for SME IPOs, GMP is not a good indicator.

		GMP (overall)			**GMP (2023)**		
		<0	=0	>0	<0	=0	>0
Mainboard	LP<IP	18	5	15	0	2	3
	LP=IP	1	0	3	0	0	1
	LP>IP	9	8	149	0	3	50
SME	LP<IP	16	42	232	14	9	98
	LP=IP	3	6	20	2	5	13
	LP>IP	1	5	60	1	3	34

TABLE 6.16: GMP analysis. IP = Issue Price, LP = Listing Price.

6.4.6 Conclusion

In this chapter, we thoroughly studied the Indian IPO landscape separately for mainboard and SME listed companies. We curated two new datasets. We mined relevant information from DRHP and RHP reports. We examined how different macroeconomic factors, the current performance of the stock market, and a company's financial health influence the success of its initial public offering (IPO). We also observed that, GMP can be used as a proxy for estimating success of Mainboard IPOs. However, for SME IPOs, GMP is not a good indicator.

We experimented with a multi-classifier decision system that fuses information from different modalities. We used texts, images, numerical data, and categorical features as inputs to predict the direction and under-pricing of stock prices at the opening, highest, and closing points on the IPO listing day. For predicting the direction of opening and closing prices, ML models like Deep Learning (DL), XG-Boost (XGB) and Gradient Boosting Machines (GBM) demonstrate superior performance when trained with numeric and categorical features. However, when it comes to predicting direction of highest prices with respect to the issue price of an IPO, incorporating text features enhances performance

of the predictors, particularly with prediction probabilities from DeBERTa models and ML models trained using Nomic embeddings. Interestingly, our approaches outperformed Gemini 1.5 flash and Llama 3.2 3b (popular large language models) under zero-shot setting. For under-pricing predictions with respect to opening-price, the Ensemble model excels for SME IPOs when leveraging a combination of numerical, categorical inputs, and probabilities derived from ML models. Conversely, the DL model is more effective for MB IPOs under similar conditions. This trend continues for under-pricing predictions with respect to highest and closing price, where both Ensemble and DL models trained with a blend of feature types consistently yield the best results. In summary, our analysis reveals the effectiveness of various machine learning models in predicting IPO price movements.

Overall, our findings underscore the importance of feature engineering in enhancing prediction accuracy in IPO pricing. This highlights the potential of advanced machine learning techniques to leverage both structured and unstructured data effectively.

This study has some limitations which can be the grounds for future works. Firstly, we have not considered the market premium which gets created due to discussions in social media. We could not consider the reviews written by expert analysts about the IPOs because this was only available from the year 2016 onwards. In future, we would like to extensively work in mining these reviews. Secondly, changing regulations can have effect on the success of an IPO. Although an IPO is regulated by SEBI, there can be instances of manipulating the Earnings Per Share (EPS) of a company before IPO. Impact of various local and international events like COVID-19, Russia Ukraine war, and General Elections of India in 2009, 2014, 2019 factors were not considered. We could gather information related to IPOs for 1099 instances in total. Expanding this dataset and capturing more features related to stock market dynamics are directions for further research. Furthermore, the models we proposed depend on historical performance data, which may not consistently reflect future results. Other future research directions include investigating how investor behaviour and psychological factors influence IPO pricing and performance, which could yield valuable insights into market dynamics. Employing qualitative methodologies such as interviews and case studies would enhance the understanding of the IPO process from the perspectives of both issuers and investors. Additionally, a comparative analysis with other emerging markets could provide broader context and insights. Further research exploring the impact of emerging technologies, and specific industry characteristics, will offer valuable insights for enhancing the efficiency and sustainability of the Indian IPO market.

Data and Code availability

The datasets can be downloaded from HuggingFace https://huggingface.co/datasets/sohomghosh/Indian_IPO_datasets.

For ensuring our work is reproducible, we have provided all the necessary details, including the hyper-parameters corresponding to the best performing models in GitHub https://github.com/sohomghosh/Indian_IPO.

6.4.7 Questions

A list of curated questions is presented in Table 6.17.

6.4.8 Variables

A list of variables which are used for predicting success of IPOs are presented in Table 6.18.

no	question
1	What are the background, qualifications, and experience of the promoters and key management team?
2	Are there any past or current criminal cases, police cases, or legal proceedings legal cases against the promoters and key management team or the company?
3	Are the promoters and key management team capable of managing the company and meet its objectives?
4	What are the company's business model, products/services?
5	What are the company's strengths, competitive advantages, and growth potential?
6	How is the company's position in the industry?
7	Is the company able to adapt to market changes?
8	What are the company's growth prospects?
9	What is the financial performance of the company in terms of revenue, profits, assets, and liabilities?
10	How does the company plan to use the funds raised through IPO?
11	Does the company plan to use IPO proceeds to repay debt?
12	Does the proposed use of funds raised from IPO aligns with the company's growth strategy?
13	What is the potential impact of the IPO on the company's future prospects?
14	What are the risks associated with investing in the company?
15	Is the company able to mitigate the risks and their potential impact?
16	What is the potential impact of market fluctuations on the company's performance?
17	Is the IPO price reasonable and offers potential for growth?
18	Does the IPO price reflect the company's intrinsic value and growth prospects?
19	Is the IPO price is reasonable and offers potential for growth?
20	Is the company's valuation right based on financial metrics (like P/E ratio, Enterprise value-to-EBITDA ratio) and industry comparisons?
21	How is the company's position in the market and among its competitors?
22	Has the company compiled with all relevant regulatory requirements?
23	Who are the lead, and co-lead managers/under-writers?
24	What is the company's corporate governance structure, including board composition, executive compensation, and shareholder rights?
25	What is the shareholding pattern, i.e.the ownership structure and potential conflicts of interest?

Table 6.17: List of Questions

TABLE 6.18: Description of Variables. P = Presence (B= Both, M = Mainboard, S = SME), T = Type of variable (I = Independent Variable i.e. Features, D = Dependent Variable i.e. Target)

P	T	Column Name	Description
B	I	mapping_key	Unique key for identifying each IPO
B	I	Company Name	Name of the company going for IPO in short
B	I	Issuer Company	Full Name of the company going for IPO
B	I	url	URL corresponding to the company's IPO in chittorgarh.com
B	I	subscription_link	URL to access the subscription information
M	I	Subscription Dates	Dates on which IPO can be subscribed
B	I	NSE_symbol	Ticker of the company in NSE
B	I	Total Issue Size	Total monetary value of all shares being offered to the public
B	I	Offer for Sale	Value of shares being offered for sale
B	I	Issue Type	Fixed Price Issue or Book Building Issue
B	I	Listing Date	Day on which IPO will get listed
B	I	Price Band	The range of prices within which investors can bid for shares
B	I	Industry	Industry of the company
B	I	Sector	Sector of the company
B	I	IPO Date	Duration for subscribing the IPO
B	I	Close Date	Last day for IPO subscription
B	I	Close Year	Year of the last day for IPO subscription
B	I	Close Year Previous	One year before the year of the last day for IPO subscription
M	I	Exchange	NSE or BSE
M	I	Issue Size (Rs Cr.)	Total shares that a company proposes to offer
B	I	Final_Issue_Price	Final Issue price of the company (on NSE for Mainboard) (on BSE/NSE for SME whichever available)
S	I	BSE_Final_Issue_Price	Issue price of the IPO on BSE
S	I	NSE_Final_Issue_Price	Final Issue price of the SME IPO on NSE
B	I	Fresh Issue	Value of shares being freshly issued
M	I	Lot Size	Minimum number of shares that an investor must bid for

Table 6.18 continued from previous page

P	T	Column Name	Description
M	I	Open Date	Date when IPO will be opened for subscription
B	I	<n>_assets	Asset of the company as on nth day
B	I	<n>_net worth	Net worth of the company as on nth day
B	I	<n>_profit after tax	Profit after tax of the company on the nth day
B	I	<n>_reserves and surplus	Reserves and surplus of the company on the nth day
B	I	<n>_revenue	Revenue of the company on the nth day
B	I	<n>_total borrowing	Total borrowing of the company on the nth day
B	I	<n>_total income	Total income of the company on the nth day
B	I	1 day before Close Day	Date before 1 day before the last day for IPO subscription
B	I	1 month before Close Day month number	Month number of the month which is 1 month before the last day of IPO subscription
B	I	1 month before Close Day year number	Year of the month which is 1 month before the last day of IPO subscription
B	I	1 week before Close Day week number	Week number of the week which is 1 week before the last day of IPO subscription
B	I	1 week before Close Day year number	Year of the week which is 1 week before the last day of IPO subscription
B	I	Basis of Allotment	Date when the final allocation of shares in an Initial Public Offering (IPO) is disclosed to investors
B	I	Brokers_Avoid	Number of top Brokers & Analysts who recommended to avoid the IPO
B	I	Brokers_Neutral	Number of top Brokers & Analysts with neutral recommended for the IPO
B	I	Brokers_Subscribe	Number of top Brokers & Analysts who recommended to subscribe for the IPO
B	I	Members_Avoid	Number of Members who recommended to avoid the IPO
B	I	Members_Neutral	Number of Members with neutral recommended for the IPO

Table 6.18 continued from previous page

P	T	Column Name	Description
B	I	Members_Subscribe	Number of Members who recommended to subscribe for the IPO
B	I	Face Value per share	Face value of a share
B	I	Credit of Shares to Demat	Date on which shares would be credited to the demat account
B	I	Cut-off time for UPI mandate confirmation	Time by which an investor must approve the UPI mandate request
B	I	day_<n>_date	Date of day n
B	I	day_<n>_bNII (bids above 10L)	Subscriptions by big Non-Institutional Investors on day n
B	I	day_<n>_emp	Subscriptions by employees on day n
B	I	day_<n>_nii	Subscriptions by Non-Institutional Investors (NII) on day n
B	I	day_<n>_nii*	Subscriptions by other type of Non-Institutional Investors (NII*) on day n
B	I	day_<n>_other	Subscriptions by others on day n
B	I	day_<n>_qib	Subscriptions by Qualified Institutional Buyers (QIB) on day n
B	I	day_<n>_retail	Subscriptions by Retail Investors on day n
B	I	day_<n>_total	Total subscriptions on day n
B	I	Retail (Max)_Amount	Maximum application amount for Retailers
B	I	Retail (Max)_Lots	Maximum number of lots that a Retailers must apply for
B	I	Retail (Max)_Shares	Maximum number of shares that a Retailers must apply for
B	I	Retail (Min)_Amount	Minimum application amount for Retailers
B	I	Retail (Min)_Lots	Minimum number of lots that a Retailers must apply for
B	I	Retail (Min)_Shares	Minimum number of shares that a Retailers must apply for
M	I	B-HNI (Min)_Amount	Minimum application amount for Big Highest Net-worth Individuals (B-HNIs)
M	I	B-HNI (Min)_Lots	Minimum number of lots that a B-HNI must apply for
M	I	B-HNI (Min)_Shares	Minimum number of shares that a B-HNI must apply for
M	I	S-HNI (Max)_Amount	Maximum application amount for Small High Net-worth Individuals (S-HNIs)

Table 6.18 continued from previous page

P	T	Column Name	Description
M	I	S-HNI (Max)_Lots	Maximum number of lots that a S-HNI must apply for
M	I	S-HNI (Max)_Shares	Maximum number of shares that a S-HNI must apply for
M	I	S-HNI (Min)_Amount	Minimum application amount for S-HNI
M	I	S-HNI (Min)_Lots	Minimum number of lots that a S-HNI must apply for
M	I	S-HNI (Min)_Shares	Minimum number of shares that a S-HNI must apply for
S	I	HNI (Min)_Amount	Minimum application amount for High Net-worth Individuals (HNIs)
S	I	HNI (Min)_Lots	Minimum number of lots that a HNI must apply for
S	I	HNI (Min)_Shares	Minimum number of shares that a HNI must apply for
B	I	Share Holding Post Issue	Distribution of ownership stakes in a company post IPO
B	I	Share Holding Pre Issue	Distribution of ownership stakes in a company pre IPO
B	I	Stocks traded, total value (% of GDP)	Value of stock traded as % of GDP in the year prior to IPO
B	I	dhrp_rhp_links	Links for DHRP, RHP, Anchor Investor files
B	I	dhrp_rhp_links_pdf	Links for DHRP, RHP, Anchor Investor pdf files
B	I	most_relevant_link	Link to download prospectus. Preference is given to RHP followed by DRHP and Anchor Investor files.
B	I	File_Rename_1st	Name of downloaded PDF file. Preference is given to RHP followed by DRHP and Anchor Investor files
B	I	Text_extracted_JSON	Page wise texts extracted from the PDF in JSON format
B	I	full_text_content	Text content related to the IPO obtained from chittorgarh.com
B	I	news_content	List of news relating to the company's IPO
B	I	news_headline	List of news headlines relating to the company's IPO
B	I	news_synopsis	List of news synopsis relating to the company's IPO
B	I	news_url	List of URLs corresponding to the news relating to the company's IPO
B	I	Chg%_nifty50_daily	Change in nifty 50 index during the day previous to the Close Date

Table 6.18 continued from previous page

P	T	Column Name	Description
B	I	Chg%_nifty50_monthly	Change in nifty 50 index during the month previous to the Close Date
B	I	Chg%_nifty50_weekly	Change in nifty 50 index during the week previous to the Close Date
B	I	Chg%_vix_daily	Change in vix index during the day previous to the Close Date
B	I	Chg%_vix_monthly	Change in vix index during the month previous to the Close Date
B	I	Chg%_vix_weekly	Change in vix index during the week previous to the Close Date
B	I	Open_nifty50_daily	Opening-price of nifty 50 index during the day previous to the Close Date
B	I	Open_nifty50_monthly	Opening-price of nifty 50 index during the month previous to the Close Date
B	I	Open_nifty50_weekly	Opening-price of nifty 50 index during the week previous to the Close Date
B	I	Open_vix_daily	Opening-price of vix during the day previous to the Close Date
B	I	Open_vix_monthly	Opening-price of vix during the month previous to the Close Date
B	I	Open_vix_weekly	Opening-price of vix during the week previous to the Close Date
B	I	High_nifty50_daily	Highest value of nifty 50 index during the day previous to the Close Date
B	I	High_nifty50_monthly	Highest value of nifty 50 index during the month previous to the Close Date
B	I	High_nifty50_weekly	Highest value of nifty 50 index during the week previous to the Close Date

Table 6.18 continued from previous page

P	T	Column Name	Description
B	I	High_vix_daily	Highest value of vix index during the day previous to the Close Date
B	I	High_vix_monthly	Highest value of vix index during the month previous to the Close Date
B	I	High_vix_weekly	Highest value of vix index during the week previous to the Close Date
B	I	Low_nifty50_daily	Lowest Price of nifty 50 index during the day previous to the Close Date
B	I	Low_nifty50_monthly	Lowest Price of nifty 50 index during the month previous to the Close Date
B	I	Low_nifty50_weekly	Lowest Price of nifty 50 index during the week previous to the Close Date
B	I	Low_vix_daily	Lowest Price of vix during the day previous to the Close Date
B	I	Low_vix_monthly	Lowest Price of vix during the month previous to the Close Date
B	I	Low_vix_weekly	Lowest Price of vix during the week previous to the Close Date
B	I	Price_nifty50_daily	Closing Price of nifty 50 index during the day previous to the Close Date
B	I	Price_nifty50_monthly	Closing Price of nifty 50 index during the month previous to the Close Date
B	I	Price_nifty50_weekly	Closing Price of nifty 50 index during the week previous to the Close Date
B	I	Price_vix_daily	Closing Price of vix during the day previous to the Close Date
B	I	Price_vix_monthly	Closing Price of vix during the month previous to the Close Date

Table 6.18 continued from previous page

P	T	Column Name	Description
B	I	Price_vix_weekly	Closing Price of vix during the week previous to the Close Date
B	I	Volume_nifty50_daily	Volume traded in nifty 50 during the day previous to the Close Date
B	I	Volume_nifty50_monthly	Volume traded in nifty 50 during the month previous to the Close Date
B	I	Volume_nifty50_weekly	Volume traded in nifty 50 during the week previous to the Close Date
B	I	dynamic_last_90Day_success_close	Average success rate calculated using close price in the last 90 days prior to IPO close day
B	I	dynamic_last_90Day_success_high	Average success rate calculated using high price in the last 90 days prior to IPO close day
B	I	dynamic_last_90Day_success_open	Average success rate calculated using open price in the last 90 days prior to IPO close day
B	I	previous_quarter	Quarter before the Close Date
B	I	previous_quarter_success_close	Average success rate calculated using close price in the previous calendar quarter
B	I	previous_quarter_success_high	Average success rate calculated using high price in the previous calendar quarter
B	I	previous_quarter_success_open	Average success rate calculated using open price in the previous calendar quarter
B	I	Foreign direct investment, net (BoP, current US$)	Net Foreign Direct Investment happened in the year prior to IPO
B	I	Foreign direct investment, net inflows (BoP, current US$)	Net inflow from Foreign Direct Investment happened in the year prior to IPO
B	I	GDP (current US$)	GDP of India in the year prior to IPO
B	I	GDP growth (annual %)	Annual growth % in GDP of India in the year prior to IPO

Table 6.18 continued from previous page

P	T	Column Name	Description
B	I	GDP per capita growth (annual %)	GDP per capita growth as annual % in the year prior to IPO
B	I	GNI (current US$)	Gross National Income in the year prior to IPO
B	I	GNI per capita growth (annual %)	Gross National Income per capita growth as annual % in the year prior to IPO
B	I	Inflation, consumer prices (annual %)	Inflation rate in the year prior to the IPO
B	I	Personal remittances, received (% of GDP)	Received personal remittances as % of GDP on the year prior to IPO
B	I	Initiation of Refunds	Date on which refunds are initiated to the investors who are not allocated any shares
B	I	Net trade in goods and services (BoP, current US$)	Net trade in goods and services in the year prior to the IPO
B	I	Unemployment, total (% of total labor force) (modelled ILO estimate)	Unemployment rate in the year prior to the IPO
B	I	answer_of_question_<n>	Answer generated using LLM for the n^{th} question. 1<=n<=25
S	D	BSE_High	Highest price of the stock on BSE on the Listing Day
S	D	BSE_Low	Lowest price of the stock on BSE on the Listing Day
S	D	BSE_Open	Opening-price of the stock on BSE on the Listing Day
B	D	BSE_Last_Trade	Closing price on BSE of the IPO on the Listing Day
B	D	NSE_High	Highest price of the stock on NSE on the Listing Day
B	D	NSE_Last_Trade	Closing price of the stock on NSE on the Listing Day
B	D	NSE_Low	Lowest price of the stock on NSE on the Listing Day
B	D	NSE_Open	Opening-price of the stock on NSE on the Listing Day
S	D	High	Overall highest price of the SME IPO on Listing day
S	D	Last_Trade	Overall closing price of the SME IPO on Listing Day
S	D	Low	Overall lowest price of the SME IPO on Listing Day
S	D	Open	Overall open price of the SME IPO on listing day

Table 6.18 continued from previous page

P	T	Column Name	Description
B	D	Success_Close	1 if closing price on listing day is more than issue price else 0
B	D	Success_High	1 if highest price on listing day is more than issue price else 0
B	D	Success_Open	1 if opening-price on listing day is more than issue price else 0
M	D	elasticity [Not used in this study]	((open price - issue price)/issue price)/(subscription rate - 1)
M	D	Total_subscriptions	Total subscriptions the IPO received

6.4.9 Prompts

6.4.9.1 Prompt for generating answers from prospectus

The prompt we used for generating answer using Llama-3 3b is as follows:
"*You are an expert financial analyst who have extensive experience of participating in Initial Public Offerings (IPOs) of Indian companies. Relevant content from Red Herring Prospectus (RHP) of an Indian company going for IPO is given to you. Your task is to analyse and answer the given question in less than 300 words as free text. Use just the content provided to you to answer the question and not anything else. If the content are not relevant, just return the word 'None'.*
CONTENT-1: {semantic-content}
CONTENT-2: {syntactic-content}
Question: {question}"

Here, {semantic-content} refers to the relevant information extracted using cosine similarity and {syntactic-content} refers to the relevant information retrieved using BM25 algorithm [328].

6.4.9.2 Prompt for estimating success of IPOs and under-pricing

We used the same prompt for Gemini 1.5 flash [331] and Llama 3.2 3b [329] to predict success of IPOs and under-pricing. The corresponding prompt is as follows:
For predicting Success
"*You are an expert financial analyst who have extensive experience of participating in Initial Public Offerings (IPOs) of Indian companies. You are given various facts of a company in JSON format where each key represents the type of content and value content itself. Your task is to analyse these content and predict if the (Open/Close/Highest) price of the IPO on the listing day will be more than the Issue price. Answer 1 if if the (Open/Close/Highest) price of the IPO on the listing day will be more than the Issue price, otherwise answer 0. If you are not confident answer -1. Your answer should be in -1, 0, 1 only.*
JSON CONTENT: {json_content}
Descriptions of keys of the JSON CONTENT are: {col_desc_dict}
Response:"

For predicting under-pricing "*You are an expert financial analyst who have extensive experience of participating in Initial Public Offerings (IPOs) of Indian companies. You are given various facts of a company in JSON format where each key represents the type of content and value content itself. Your task is to analyse these content and predict if the under-pricing with respect to (Open or Close or Highest) price of the IPO on the listing day i.e. (Open or Close or Highest - issue price)/(issue price). Answer should be a real number only. If you are not confident answer nan.*
JSON CONTENT: {json_content}
Descriptions of keys of the JSON CONTENT are: {col_desc_dict}
Response:"

6.5 Predicting Ratings of Indian IPOs from Red Herring Prospectus

6.5.1 Introduction

An Initial Public Offering (IPO) is the process by which a private company first offers its shares to the public, transitioning to public ownership. This event enables the company to raise capital by selling ownership stakes to individual and institutional investors.

In the Indian context, IPOs are categorized into Mainboard (MB) IPOs and Small and Medium Enterprises (SME) IPOs, each serving distinct market segments. Mainboard IPOs are intended for larger, established companies that meet stringent regulatory requirements and are listed on major stock exchanges such as the BSE and NSE. These offerings typically attract a broad investor base, involve larger issue sizes, and provide higher liquidity and market recognition.

Conversely, SME IPOs cater to small and medium enterprises, which often have more relaxed eligibility criteria and are listed on specialized platforms like BSE SME and NSE Emerge. SME IPOs generally involve smaller investment amounts and a limited number of allottees, making them accessible to retail investors but associated with higher risks and lower liquidity compared to Mainboard IPOs. The IPO prospectus is a critical legal document that provides potential investors with comprehensive information about the company and its offering. It serves as a transparency tool, enabling informed investment decisions. There are two types of IPO prospectuses: the Draft Red Herring Prospectus (DRHP) and the Red Herring Prospectus (RHP). The DRHP is the initial document filed with the Securities and Exchange Board of India (SEBI) prior to launching an IPO. It outlines the company's business model, financial statements, risk factors, and intended use of raised funds, subject to SEBI's review and approval. Upon receiving SEBI approval, the company issues the RHP, which includes updated information such as the final offer price and number of shares offered. The RHP is made available to potential investors during the offer period, providing essential details for investment decision-making. Both DRHP and RHP are vital components of the IPO process, ensuring that investors have access to accurate and comprehensive information about the company's financial health. Typically, DRHPs range from 100 to 300 pages in length, while RHPs are usually between 80 and 250 pages. Reading these lengthy documents can be time-consuming and overwhelming for novice investors. Reviews and ratings provided by experts can often be subjective and biased. For popular IPOs, investors are frequently inundated with expert reviews. However, for lesser-known IPOs, expert reviews are rarely available.

IPO grading is an evaluation process that assesses the fundamentals of a company's initial public offering (IPO) relative to its peers. This grading provides investors with an independent opinion on the quality and potential of the IPO, helping them make informed investment decisions. In India, IPO grading became mandatory in April 2007 for all new issues, as mandated by the Securities and Exchange Board of India (SEBI). This requirement aims to enhance transparency and encourage independent research in the equity market. However, The effectiveness of IPO grading was questioned, leading SEBI to make it optional starting February 4, 2014. [46]. Furthermore, IPO grading is inherently

[46] https://www.angelone.in/knowledge-center/ipo/ipo-grading (accessed on 19th January, 2025)

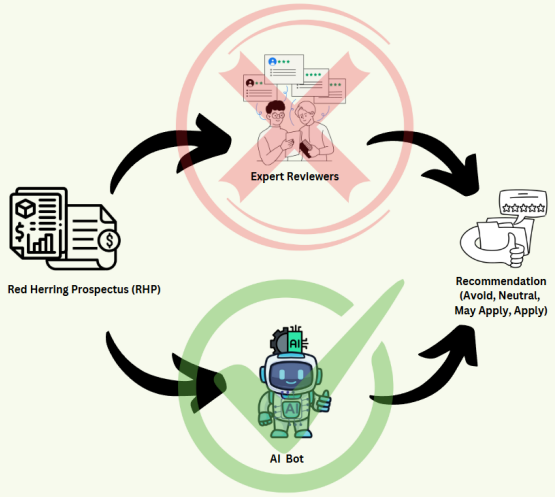

FIGURE 6.11: IPO Rating Prediction

subjective and can vary across different rating agencies. This variability raises questions about the consistency and reliability of the grades assigned. [47]

Therefore, an automated system capable of mining these prospectuses would facilitate the development of a decision-making tool to assist investors in determining whether to subscribe to an IPO. This is presented in Figure 6.11.

Our Contributions

Our contributions are as follows:

- We introduce two new India-specific datasets (one for Mainboard IPOs and another for SME IPOs) along with a task focused on predicting the ratings of these IPOs.
- We propose a novel method for mining prospectus of these IPOs which consists of a Retrieval Augmented Generation framework along with a fine-tuned small encoder based language model. This method outperforms state-of-the-art Large Language Models (LLMs) under zero-shot settings.

6.5.2 Related Works

The prediction of Initial Public Offering (IPO) performance has garnered significant attention in the financial literature, particularly due to its implications for investors and market efficiency. Various studies have investigated the determinants of IPO performance, emphasizing factors such as market conditions, investor behaviour, and corporate governance. A key aspect of this research is the phenomenon of IPO under-pricing, which is crucial for understanding overall IPO performance. Most prior studies have concentrated on short-run under-pricing [305] [291] or long-run underperformance [308].

[47] https://www.motilaloswal.com/blog-details/what-is-ipo-grading-process-in-india/21319 (accessed on 19th January, 2025)

Some researchers have explored the usefulness of IPO grading. The study [332] indicates that a substantial number of retail investors are familiar with the IPO grading process. However, perceptions of its effectiveness and influence on investment decisions vary. While IPO grading is considered a valuable tool for investors [333], its impact is not consistent across different segments of the investor population. As per [334], securities with higher IPO grades are observed to exhibit a lower degree of under-pricing. Additionally, higher IPO grades are associated with an increase in subscription rates across all kinds of investors. The influence of credit ratings on IPO under-pricing has been well-documented. Dhamija and Arora found that firms with credit ratings experience significantly less under-pricing than those without, indicating that improved corporate governance and transparency can lead to better IPO valuations [335]. Jacob and Agarwalla [336] explored the effects of mandatory IPO grading in India. They concluded that such certifications can enhance demand of institutional investors, but their impact on overall pricing efficiency is limited. All of these studies highlight the significance of IPO grading; however, none of them propose automated methods for grading IPOs. On the contrary, automated methods for predicting ratings from texts [337] have been well-studied in several domains like e-commerce [338], local service [339], etc. Consequently, we present the task of predicting ratings based on the prospectuses of Indian companies that are preparing for IPOs. This task is similar to automated grading of IPOs and it would pave the way for a valuable tool that empowers investors with data-driven insights to make more confident and informed decisions regarding IPO subscriptions. To the best of our knowledge, the proposed task represents a novel contribution to this field.

6.5.3 Problem Statement

Given a company's IPO prospectus, our objective is to comprehend its content and categorize it into one of four classifications: Apply, Neutral, May Apply, or Avoid, providing a concise and informed assessment of the investment opportunity. As this a classification problem with class imbalances, we will use Micro, Macro, and weighted F1 score for evaluation.

6.5.4 Dataset

We gathered data on MB and SME IPOs separately from the chittorgarh website.[48] The MB data is available from 2011, while SME data starts from 2012. Our collection of this data continued until November 7, 2024, and includes the following information: Review Title (this contains name of the company as well), Name of the Author / Organization who wrote the review, Year of the IPO, Link to access the review, Link to a webpage containing comprehensive details about the IPO, Key (Unique identifier of each row), Link to access the (D)RHP in PDF format, Name of the JSON file having text content extracted from (D)RHP, Text content of the review, Recommendation (Apply, Neutral, May Apply, or Avoid).

To ensure data quality, we excluded entries without reviews or recommendations. Notably, mainboard IPOs often have multiple reviews; in such cases, we retained only those reviews that aligned with the majority recommendation. For example, if a company has five

[48] https://chittorgarh.com/ (accessed on 19th January, 2025)

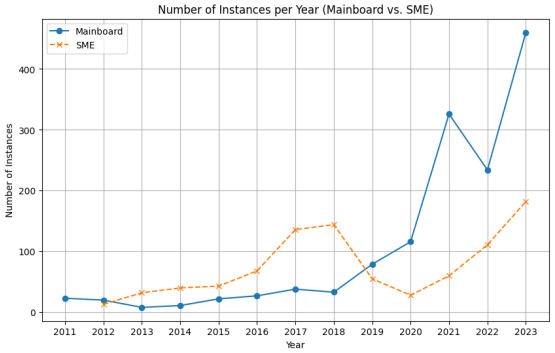

FIGURE 6.12: Data Distribution up to year 2023

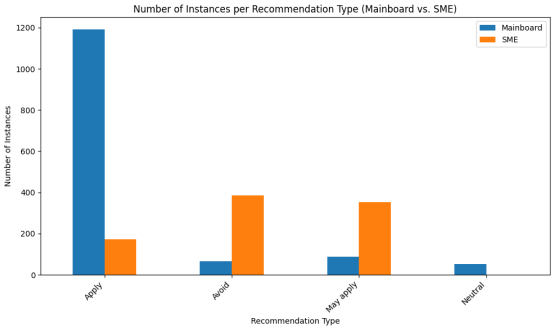

FIGURE 6.13: Distribution of Recommendations

reviews—three recommending “Apply" and two recommending “Avoid"—we would keep only the three “Apply” reviews. Conversely, 97% of SME IPOs have reviews authored by a single individual, leading us to discard the remaining 3% of data. For reviews provided in PDF format, we utilised PyPDF [49] to extract text. The Draft Red Herring Prospectuses (DRHP) and Red Herring Prospectuses (RHP), were available in PDF format. In instances where both DRHP and RHP were present, we prioritized the RHP. To further ensure the quality of our data, we compared IPO ratings with their actual opening-prices. For Mainboard IPOs, we found that in 82.17% of cases, an ’Apply’ recommendation corresponded to an opening-price higher than the issue price. For SME IPOs, this figure was 83.49%. In total, we collected 1,830 instances for mainboard IPOs and 1,131 for SME IPOs. Data up to 2023 was used for training purposes, while data from 2024 was reserved for testing. We present the data distribution with respect to year and recommendations in Figures 6.12 and 6.13 respectively.

6.5.5 Experiments and Results

In this section, we describe the experiments we conducted and discuss the corresponding results.

[49] https://pypi.org/project/pypdf/ (accessed on 19th January, 2025)

Due to budget limitations and computational constraints, we were unable to use the entire prospectus in PDF format into LLMs at once. Additionally, as noted in [340], larger context sizes can lead to a decrease in the performance and reasoning capabilities of LLMs. Therefore, it was essential for us to extract specific sections from the prospectus that were most relevant to determining the ratings of the IPOs. Thus, we conducted a randomized selection of 200 reviews for both MB and SME IPOs separately. The limitation to 200 reviews was necessitated by the rate limit of Groq [50] API's free tier. The selected reviews were processed using the Llama-3 8B model [329], from which we extracted questions utilising the prompt specified in Section 6.5.6.2. Subsequently, these questions were submitted to Perplexity.ai Pro [51] to compile a comprehensive list of distinct questions, which are presented in Section 6.5.6.1. The rationale for employing two different large language models (LLMs) stemmed from the superior capabilities of the Perplexity Pro model in handling complex tasks, albeit limited to two queries per day under the free tier. In contrast, the smaller Llama-3 8B [329] model allowed for multiple queries. We utilised the expert reviews solely for extracting the questions mentioned above and did not use them in any other steps of the process.

Following the methodology outlined in [254], we extracted text from the prospectus (RHP) which were present in PDF format. Optical character recognition (OCR) was performed using Tesseract to extract text from images within the documents. Each page was converted into embeddings utilising Nomic [327]. Employing a Retrieval-Augmented Generation (RAG) framework, for each of compiled questions mentioned in Section 6.5.6.1, we identified the two most pertinent pages based on two criteria: first, through cosine similarity for semantic matching, and second, via BM25 [328] for lexical similarity. The retrieved pages, along with their corresponding questions, were then passed into the Llama-3.2 3B [329] model to generate answers. Details relating to the prompt we used are mentioned in section 6.5.6.2. This process yielded a total of 16 answers for each instance, corresponding to the 16 questions posed.

We employed a zero-shot approach by prompting the Gemma-2 9B, Llama 3.1 70B, and Llama-3.2 3B models to classify the aggregate of 16 answers into one of four categories: Apply, Neutral, May Apply, or Avoid. Details of the prompts are provided in section 6.5.6.2. We then repeated these experiments by substituting the aggregate of answers with a single summary. These summaries were generated using Llama-3.2 3B [329]. Prompt details are presented in 6.5.6.2. We observed that this change led to improved model performance in most cases. Subsequently, we fine-tuned Llama-3.2 3B and Gemma-2 9B using supervised fine-tuning methods.

Finally, we trained three encoder based models (RoBERTa [125], LongFormer RoBERTa [52], and DeBERTa [330]) with the summaries for classification.

We observed that for MB IPOs, the LongFormer RoBERTa outperformed all other models in terms of micro, macro, and weighted F1 scores. In contrast, for SME IPOs, the Gemma-2 9B model excelled in micro F1 scores, while the Llama 3.1 70B model achieved the highest macro F1 scores. Additionally, the RoBERTa model demonstrated superior performance in terms of the macro FA score.

We present the overall flow in Figure 6.14 and results in Table 6.19.

[50] https://console.groq.com/docs/overview (accessed on 19th January, 2025)
[51] https://www.perplexity.ai/ (accessed on 19th January, 2025)
[52] https://huggingface.co/markussagen/xlm-roberta-longformer-base-4096 (accessed on 19th January, 2025)

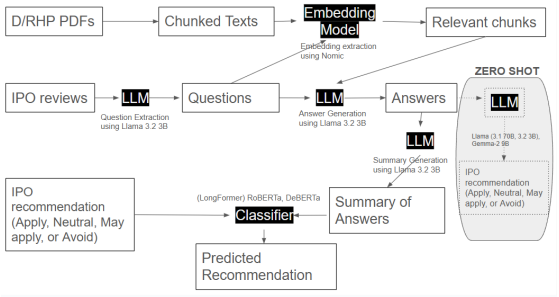

FIGURE 6.14: Detailed Flowchart narrating our methodology

		MB			SME		
Model	**Input**	**F1 (m)**	**F1 (M)**	**F1 (w)**	**F1 (m)**	**F1 (M)**	**F1 (w)**
Gemma-2 9B (Zero-Shot)	All Answers	0.009	0.007	0.005	0.411	0.189	0.368
Llama-3.1 70B (Zero-Shot)	All Answers	0.039	0.021	0.054	0.374	0.176	0.355
Llama-3.2 3B (Zero-Shot)	All Answers	0.484	0.184	0.348	0.076	0.038	0.114
Gemma-2 9B (Zero-Shot)	Summary	0.023	0.108	0.012	**0.516**	0.256	0.416
Llama-3.1 70B (Zero-Shot)	Summary	0.115	0.044	0.191	0.457	0.281	**0.423**
Llama-3.2 3B (Zero-Shot)	Summary	0.162	0.077	0.255	0.429	0.163	0.361
Llama 3.2 3b (SFT)	Summary	0.836	0.228	0.883	0.361	0.299	0.347
Gemma 2 9B (SFT)	Summary	0.716	0.233	0.814	0.402	0.298	0.349
RoBERTa	Summary	0.769	0.219	0.846	0.406	**0.335**	0.377
LongFormer RoBERTa	Summary	**0.968**	**0.246**	**0.952**	0.224	0.126	0.090
DeBERTa	Summary	0.912	0.239	0.925	0.457	0.319	0.383

TABLE 6.19: Model Performances. m = micro, M = Macro, w = weighted, SFT = Supervised Fine-tuning. Best performing models are highlighted in bold.

6.5.6 Conclusion

In this section, we introduce the task of mining the prospectuses of Indian companies preparing for IPOs to predict their ratings. To support this task, we propose two new datasets: one for SME IPOs and another for Mainboard IPOs. Additionally, we present a novel framework that utilizes Retrieval-Augmented Generation (RAG) to extract relevant sections from the prospectus, summarize them, and employ fine-tuned encoder-based small language models to predict the final ratings. Our approach demonstrates superior performance compared to existing state-of-the-art large language models, such as Llama 3.1 70B and Gemma-2 9B, when evaluated under zero-shot settings.

This research has a few limitations that highlight opportunities for future work. Firstly, the list of questions we curated to extract relevant portions from the prospectuses of Mainboard IPOs is not exhaustive and may not encompass all the critical details necessary for informed decision-making. In the future, we aim to make this list dynamic, adapting it based on various factors such as industry, profitability, and other relevant criteria.

Due to budget constraints, we were unable to process the entire text corpus (RHP) at once. While we are focused on rating predictions, the actual opening-price will ultimately determine the true performance. Analyzing the actual opening-price data will provide more meaningful insights. Additionally, we utilised APIs from various service providers, including Groq [53], Cerebras [54], and OpenRouter [55], to evaluate the performance of different LLMs,

[53] https://groq.com/ (accessed on 19th January, 2025)
[54] https://cerebras.ai/ (accessed on 19th January, 2025)
[55] https://openrouter.ai/ (accessed on 19th January, 2025)

such as Llama 3.1 70B [329] and Gemma-2 9B [341], under zero-shot settings. However, the same LLM may produce slightly varying results depending on the service used.

The datasets used in this paper can be obtained from https://huggingface.co/datasets/sohomghosh/indian_ipo_rating_prediction. Our codebase is available at https://huggingface.co/datasets/sohomghosh/indian_ipo_rating_prediction.

6.5.6.1 Questions

The list of 16 extracted questions is presented here.

- What is the price band and issue price of the IPO?
- What is the issue size and how many shares are being issued as part of the IPO?
- What is the implied market capitalization of the company after the IPO?
- How will the company utilize the funds raised through the IPO, and what is the purpose of the IPO?
- What is the company's revenue growth rate over recent financial years, and how has its financial performance been historically (including revenue, EBITDA, and net profit trends)?
- What are the key financial ratios, such as net profit margin, return on equity (RoE), return on capital employed (RoCE), and total debt?
- What is the shareholding pattern before and after the IPO, and who are the promoters?
- Are there any regulatory issues or conflicts of interest affecting the company?
- What are the company's plans for expansion and future growth, and how does it position itself in terms of competition within its industry?
- Who are the company's major customers, what is the revenue breakdown by sector, and is there a dependency on large institutional customers?
- What are the potential risks associated with increasing raw material costs, and what other risks does the company face?
- How does the company's valuation compare to its peers, and is the issue priced aggressively compared to industry standards?
- What is the competitive landscape of the industry in which the company operates?
- Has the company declared any dividends in the past, and what is its dividend policy?
- Who are the lead managers and registrar for the IPO, and what is their track record in terms of past IPO listings?
- Are there any concerns regarding transparency or missing details in the offer document?

6.5.6.2 Prompts

Question Extraction Prompt:
The prompt used for extracting questions is:
You are an expert financial analyst who have extensive experience of participating in Initial Public Offerings (IPOs) of Indian companies. You are given a review about an Indian company going for IPO. Extract a list of key questions which have been answered in the given review and which would help in determining whether to apply for the IPO. Return just a list of questions which can be answered from the review. Do not return anything other than the list of questions. Review: {review content}
Response:

Answer Generation Prompt:
This prompt was used for each of the 16 questions to generate the corresponding answer.
You are an expert financial analyst who have extensive experience of participating in Initial Public Offerings (IPOs) of Indian companies. Relevant content from Red Herring Prospectus (RHP) of an Indian company going for IPO is given to you. Your task is to analyse and answer the given question in less than 300 words as free text. Use just the content provided to you to answer the question and not anything else. If the content are not relevant, just return the word 'None'.
CONTENT-1: {semantically relevant content }
CONTENT-2: {syntactically relevant content}
Question: {question}
Response:

Summary Generation Prompt:
The prompt used for generating summary from answers is as follows:
You are an expert financial analyst who have extensive experience of participating in Initial Public Offerings (IPOs) of Indian companies. You are provided with various facts about a company going for IPO in the form of answers. Your task is to analyse these answers and generate a summary comprising of key points that investors need to know to decide if they should subscribe for the IPO or not. If you are not confident answer nan. Just return the summary in 300 words and nothing else. Facts about the company's IPOs are as follows: {answers of 16 questions}.
Response:

Rating Prediction Prompt:
The prompt used for zero-shot classification is:
"*You are an expert financial analyst who has extensive experience of participating in Initial Public Offerings (IPOs) of Indian companies. You are given various facts of a company. Your task is to analyse these facts and decide whether an investor should 'Avoid', 'May apply', 'Apply', or, be 'Neutral' for the IPO. Your answer should be in a JSON structure with two keys, 'prediction' and 'justification'. The value corresponding to 'prediction' key should be 0,1,2, or, 3 only where 0 represents 'Avoid', 1 represents 'Neutral', 2 represents 'May apply', and 3 represents 'Apply'. The value corresponding to 'justification' key should be the explanation behind the prediction. Facts: {answers of 16 questions concatenated side by side}.*
Response:"

Chapter 7

Tools for FinNLP

In this chapter, we describe some of the tools we developed for solving various FinNLP tasks.

7.1 Relevant Publications

- FLUEnT: Financial Language Understandability Enhancement Toolkit [23]
- Using Natural Language Processing to Enhance Understandability of Financial Texts [25]

7.2 Introduction

People who want to invest in stock markets often face various challenges due to the lack of financial knowledge. The financial domain is full of complicated concepts and jargon. Committing minor mistakes while investing can have an adverse effect on the returns. Professional investors also get perplexed by the information overload which inhibits them from making decisions in real-time. To address these challenges, we developed the Financial Language Understandability Enhancement Toolkit (**FLUEnT**) which consists of eight different tools to cater to the needs of the general people and the investors. Figure

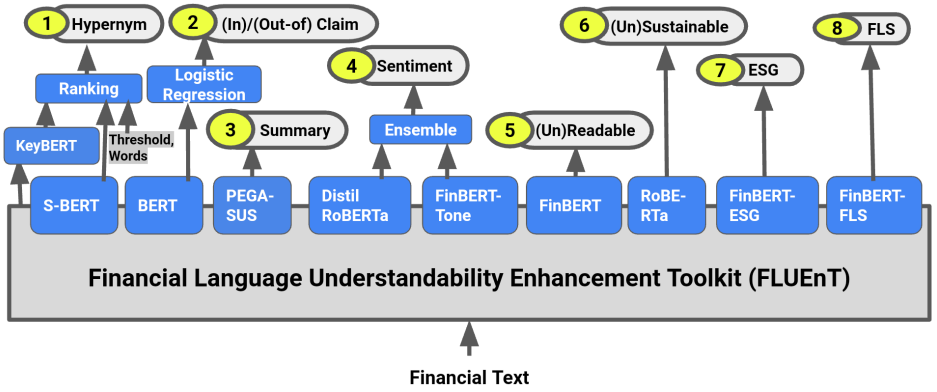

FIGURE 7.1: Overview of Financial Language Understandability Enhancement Toolkit

7.1 presents an overview of the toolkit and the functionalities of the constituent tools. We developed four of these tools. They are marked as **1, 2, 5** and **6** in Figure 7.1. For the remaining four tools (marked as **3, 4, 7** and **8** in Figure 7.1), we leverage existing open-source models and artefacts. The hypernym detection and readability assessment tools aim to enhance the financial literacy of the masses by providing them with suitable hypernyms (generic forms) of complex financial words (FW) and helping them to filter out the easy-to-understand (i.e., readable) content. The other tools of this toolkit help investors to summarise financial texts (FT) and understand the sentiment, sustainability, Forward-looking statements (FLS), Environmental, Social, and Governance (ESG) aspects of sentences present in FT. Furthermore, the claim detection tool (CD) looks to classify each numeral present in FT as in-claim or out-of-claim.

Our contributions

We have developed **FLUEnT** which can empower the investors in making data-driven decisions and aid in spreading financial literacy. Subsequently, we have deployed and open-sourced this toolkit[1] for non-commercial use. A live demonstration is available on YouTube[1]. The novelty of the system lies in the fact that, like a Swiss knife, it solves eight different use cases in real-time to empower seasoned as well as prospective investors. It intelligently picks up difficult words and numbers from financial texts and provides users with their hypernyms and 'claim' categories respectively. Moreover, it returns a summary of the entered FT in addition to sentence-wise sentiment, readability, sustainability, ESG and FLS classes.

[1] https://youtu.be/Bp8Ij5GQ59I (accessed on 18th September, 2023)

We expect that the popularity of **FLUEnT** will grow over time among professional investors and common people who want to invest in the stock markets. Governments, policy-makers and non-governmental organizations (NGOs) can use it readily for promoting financial literacy. Above all researchers working in this space can readily use the tools and libraries for their research.

7.3 Related Work

Table 7.1 presents a list of related tools and their functionalities alongside **FLUEnT**. As of July 2022, only nine out of twelve existing tools have a User Interface (UI) and only 6 of them are running live. Most of these tools deal with information extraction and present few analyses from financial reports. Unlike these tools, **FLUEnT** is relatively more comprehensive and it provides eight different functionalities. As we have two different variations of FinBERT, namely [69] and [342], we refer to them as FinBERT(a) and FinBERT(b) respectively. In addition to these tools, there are several proprietary tools and cloud services like SentiMine[2], Augmented Financial Analyst [3], etc. However, discussing them is beyond our scope as these tools require subscriptions.

Tools	UI	Live	Functionalities
Financial Term Visualisation [343]	No	No	Risk assessment, FT Identification & Visualisation from financial reports
FINCHAN [344]	Yes	No	Syntactic & semantic information extraction, & Summarisation, Text-to-speech conversion of financial instant messages
FIN10K [213]	Yes	Yes	Extracts relevant portions from 10-K reports & visualizes risk levels & sentiments of keywords
Financial Chatbot [345]	Yes	No	Document search, Topic extraction & Clustering
RegMiner [346]	Yes	Yes	Extraction & Visualization of restrictions present in regulatory documents
ClimateQA [155]	Yes	No	Extraction of climate related sections from financial reports using question answering
FinBERT(b) [342]	No	No	Sentiment Analysis, FLS Assessment & ESG Assessment
FedNLP [347]	Yes	Yes	Summarisation, Sentiment Analysis, Topic Models, Federal Funds Rate Movement Rate Prediction
EDGAR-CRAWLER [126]	No	No	Extraction of texts from financial reports
FinRead [5]	Yes	Yes	Readability Assessment
FiNCAT-2 [15]	Yes	Yes	Claim Detection
Financial_Analyst_AI [demo link]	Yes	Yes	Voice-to-Text, Summarisation, Sentiment Analysis, FLS Assessment, Company Names & Location Identification
FLUEnT [Demo] [Video] [Colab]	Yes	Yes	**Keyword & Hypernym Detection, Claim Detection, Summarisation, Sentiment Analysis, Readability Assessment, Sustainability Assessment, ESG Assessment & FLS assessment**

TABLE 7.1: Comparison of **FLUEnT** with existing non-proprietary tools

7.4 Constituent Tools

FLUEnT consists of eight different tools. Inputs, outputs, development process, and performance for each of these tools are summarised in Table 7.2. In this section, we present a detailed explanation for all of them. We chose the underlying models based on their performance and availability.

7.4.1 Hypernym Detection (HD)

Complex terms can be explained easily using their generic forms or hypernyms. For example, we can explain the FT "*alternative debentures*" by mentioning its hypernym

[2] https://www.lseg.com/about-lseg/labs/sentimine (accessed on 18th September, 2023)
[3] https://yseop.com/solutions/augmented-financial-analyst (accessed on 18th September, 2023)

Tool	Input	Output	Base Models	Developer	Development Dataset (Size)	Performance
HD	FW	Generic form of each terms	SBERT+FinBERT(a)	Ours (Ghosh et al.)	FinSim-3 (1,050 FW)	Accuracy: 0.9170
CD	FT	Each numeral in FT: in-claim or out-of-claim	BERT-base	Ours (Ghosh et al.)	FinNum-3 English (10,720 FT)	Macro-F1: 0.8238
SM	FT	Summary of the entire FT	PEGASUS	Passali et al.	Bloomberg articles (2,000 FT)	Rouge-L: 18.14
SA	FT	Each sentence present in FT: positive, negative or neutral	1) BERT-base 2) DistillRoBerta-base	1) Huang et al. (finbert.ai) 2) Romero M.	1) Analyst reports of S&P 500 firms (10,000 FT) 2) Financial PhraseBank (4,840 FT)	1) Accuracy: 0.882 2) Accuracy: 0.9823
RA	FT	Each sentences present in FT: readable or non-readable	FinBERT(a)	Ours (Ghosh et al.)	FinRAD (13,112 FW definitions)	AUROC: 0.9927
SN	FT	Each sentences present in FT: sustainable, non-sustainable or none	RoBERTa-base	Ours (Ghosh et al.)	FinSim-4-ESG (2,265 FT)	Accuracy: 0.9317
ESG	FT	Each sentences present in FT: Environmental, Social, Governance or none	FinBERT(b)	Huang et al. (finbert.ai)	Annual & ESG reports of firms (2,000 FT)	Accuracy: 0.895
FLS	FT	Each sentences present in FT: Specific-FLS, Non-specific FLS or Not-FLS	FinBERT(b)	Huang et al. (finbert.ai)	MD&A sections of annual reports of Russell 3000 firms (3,500 FT)	Accuracy: 0.853

TABLE 7.2: Different constituent tools and their characteristics. FT & FW mean financial texts & words respectively.

i.e. "*it is a kind of bond*". A tool to detect hypernyms is useful to learn financial jargon effortlessly. Given an FT, we extract the top three keywords from it using KeyBERT [348]. Users have an option to look for hypernyms of these keywords or other FW they manually enter. Chopra and Ghosh [7] fine-tuned a FinBERT(a) [69] model on the FinSim-3 dataset [94] using the sentence BERT architecture [68] for financial hypernym detection. For all the keywords or FW, we use the fine-tuned sentence BERT embeddings to calculate its cosine similarity with a set of seventeen predefined hypernyms. We provide users with the hypernyms corresponding to the entered financial words only when their similarity is more than the threshold set by the user using the slider present in the tool.

7.4.2 Claim Detection (CD)

Executives try to lure investors by making claims which may not always be true. The sentence, "*In the year 2021, the markets were bullish. We expect to boost our sales by 80% this quarter.*" has two numerals 2021 and 80%. Among these two, "*2021*" is 'out-of-claim' and "*80%*" is 'in-claim'. The CD tool can alert investors by detecting numerals in FT which are 'in-claim'. For each of the numbers present in an FT, we extract its BERT-base [1] embedding given a context window of six words before and after it. Subsequently, we use logistic regression to classify it as either in-claim or out-of-claim. The methodology of the CD tool is described in [17] and [14]. The CD model was trained on FinNum-3 (English) dataset [186]. We have further released two tools: FiNCAT [16] and FiNCAT-2 [15] to help investors in detecting claims present in numerals within FT.

7.4.3 Summarisation (SM)

In today's fast-paced world, time and money are almost equally valuable. With the advent of Big Data, investors are overloaded with information; they do not have the time to assimilate all the information. Thus, the SM tool aims to help them by removing irrelevant and fewer relevant facts and providing them with only the necessary information. We integrated the financial summariser built by Passali et al. [349] in our toolkit. The SM tool provides a summary of the entered FT using the PEGASUS [350] model.

7.4.4 Sentiment Analysis (SA)

Lately, financial opinion mining has gained significant interest. Some of the open-source models include FinBERT-tone [4] (a derivative of FinBERT(b) [342]) developed by fine-tuning BERT-base [1] on analyst reports of S&P 500 firms, and DistilRoBERTa-financial-sentiment[5] developed by fine-tuning DistillRoBERTa-base [172] on the financial phrase bank dataset [351]. For each sentence in an FT, we evaluate both models and produce the label with the greater probability. The output labels are: 'positive', 'negative' and 'neutral'.

7.4.5 Readability Assessment (RA)

To ensure that the non-investors who want to invest in the stock market do not get overwhelmed, it is essential to present them with information which is easy to understand (readable). Since the formula-based readability scores (like Automated Readability Index, Coleman-Liau index, etc.) are not suitable for the financial domain, we proposed a new financial readability assessment dataset, FinRAD [4], and a FinBERT(a) [69] based neural model to classify definition of financial terms. We use this model to assess whether each sentence in the entered FT is readable or not. Subsequently, we have developed a tool FinRead [5] to address this.

7.4.6 Sustainability Assessment (SN)

Socially conscious investors seek sustainable investment opportunities. We used the FinSim-4-ESG (Shared Task 2) [142] dataset to fine-tune a RoBERTa-base model [125] for classification of each sentence present in the FT into three classes 'sustainable', 'non-sustainable' or none (represented by '-') [10].

7.4.7 ESG Assessment (ESG)

Investors look for ESG ratings of companies they want to invest in. It is very tedious to read ESG reports of every organisation. For each sentence in an FT, this tool detects whether it is related to 'Environment', 'Social', 'Governance' or none. Huang et al.[6] developed the underlying model by fine-tuning the FinBERT(b) model [342].

7.4.8 FLS Assessment (FLS)

FLS help investors to understand the future conditions of the financial market. Huang et al. proposed FinBERT-FLS [7] for classifying financial texts as 'Specific-FLS', 'Non-specific

[4] `https://huggingface.co/yiyanghkust/finbert-tone` (accessed on 18th September, 2023)
[5] https://huggingface.co/mrm8488/DistilRoBERTa-finetuned-financial-news-sentiment-analysis (accessed on 18th September, 2023)
[6] `https://huggingface.co/yiyanghkust/finbert-esg` (accessed on 18th September, 2023)
[7] `https://huggingface.co/yiyanghkust/finbert-fls` (accessed on 18th September, 2023)

FLS' or 'Not-FLS'. It was developed by fine-tuning FinBERT(b) [342] on a set of 3,500 manually annotated financial sentences. We use the FinBERT-FLS model for classifying each sentence present in the entered FT into the above mentioned classes.

7.5 System Overview

In this section, we discuss the underlying technologies and elaborate the UI of **FLUEnT** in detail.

7.5.1 Implementation Details

These constituent models have been trained using PyTorch [352] using HuggingFace Transformers [140]. We carried out the experiments on Google Colab [8] (runtime: GPU). The UI has been created using Gradio [174] and hosted on Colab and HuggingFace Spaces https://www.youtube.com/watch?v=Bp8Ij5GQ59I.

7.5.2 Demonstration Interface

The Graphical User Interface (GUI) of **FLUEnT** primarily consists of two sections: the inputs (Ref: Figure 7.2) and the outputs (Ref: Figure 7.3). In the input section the user enters an FT in the text-box above (`TB-1`) and sets a confidence `threshold` using the slider. The GUI also provides a number of examples. When the user clicks the "*Get Keywords for Hypernym Detection*" button, the GUI shows the top three keywords extracted from the entered text based on KeyBERT [348]. The keywords are shown in the text-box (`TB-2`) below. The user can look for hypernyms corresponding to these keywords. The user can also alter the contents of this text-box (`TB-2`) by manually entering keywords of their choice.

The output section consists of eight different tools presented in four tabs. Each of these tools can be used independently. This saves time as well as computing resources. At any given time, users can select from the four tabs and press a `Get` button corresponding to the tool they want to use. The HD tool extracts generic forms of the FW entered in the `TB-2` (or, the extracted keywords). These hypernyms are presented only when their similarity score is above the threshold set by using the slider. The CD tool first extracts numerals from the FT entered in `TB-1`. Subsequently, it classifies each of the numerals as 'in-claim' or 'out-of-claim'. The 'in-claim' and 'out-of-claim' numerals are presented in red and green colours respectively. The remaining six tools use the FT entered in `TB-1` for making predictions. We highlight each sentence present in the FT and mention the predicted categories next to them. This enhances the usability of **FLUEnT**.

[8] https://research.google.com/colaboratory/(accessed on 18th September, 2023)

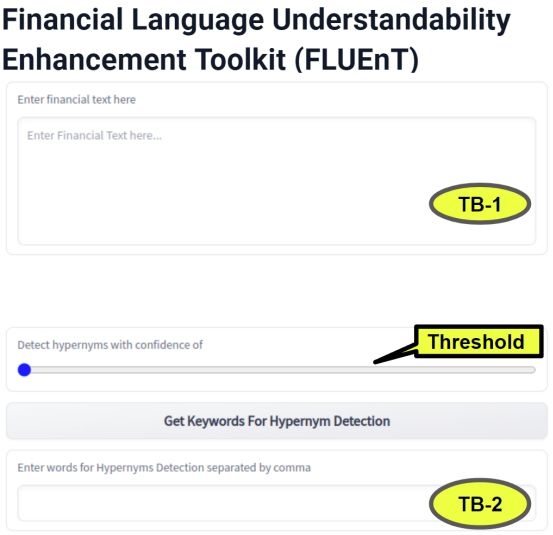

FIGURE 7.2: Inputs with text-boxes (`TB-1, TB-2`) and `threshold` field marked

FIGURE 7.3: Outputs (HD, CD, SM, and SA). Similar outputs are generated for RA, SN, ESG and FLS.

7.6 Conclusion

In this chapter, we presented **`FLUEnT`**, a toolkit that helps in improving the comprehensibility of complex FT. It performs several tasks on financial texts, such as HD, CD, SM, SA, RA, etc. In the future, we want to add various other features like uploading documents (PDFs) as input, extracting relevant portions from these documents which relate to finance and then perform various tasks on these portions. We also want to work on collecting feedback from the users and develop a browser-based extension that will scan content from financial web pages and help investors in understanding it. Another direction for future work is to develop a multi-task model which will reduce the overall size of the tool and improve its throughput.

Chapter 8

Conclusion

8.1 Conclusion

This thesis is aimed at improving financial well-being of investors and people in general. We focused on improving financial literacy by making the investment journey inclusive, improved, impactful, and informed. Financial literacy is essential for ensuring that the financial inclusion schemes are successful. Financial literacy leads to the overall prosperity of the nation and helps in reducing the wealth disparity. We developed and open-sourced several tools (like FLUEnT [23], FinLanSer, etc.) for analysing financial texts. In addition to this, we worked extensively on popular Indian languages like Hindi, Bengali and Telugu.

This thesis underscores the transformative potential of advanced methodologies and datasets in understanding and predicting crowd reactions, financial sentiment, and the intricate relationships between financial entities. By harnessing state-of-the-art language models and innovative approaches, this study contributes significantly to the fields of social media analytics, financial technology (FinTech), and NLP. At the core of this research is the introduction of the Crowd Reaction Estimation Dataset (CRED) and the Generator-Guided Estimation Approach (GGEA) [6]. CRED serves as a vital tool for analysing governmental social media managers' decision-making processes through comparative repost metrics. The GGEA not only showcases the capabilities of large language models (LLMs) but also emphasizes their role in enhancing traditional classification models. The experiments conducted reveal that integrating LLMs like Claude into predictive frameworks significantly improves accuracy in forecasting crowd reactions, thus demonstrating the importance of nuanced analysis in decision-making.

Furthermore, the exploration of financial entity relationships through various LLMs highlights how tailored embeddings can enhance model performance. The fine-tuning of models such as FinBERT [69] and FinISH [137] using domain-specific data illustrates a clear pathway for improving predictive analytics in finance. The findings indicate that augmenting datasets with external information can enhance performance.

The research also delves into multilingual applications, particularly focusing on low-resource languages. The consistent performance across different languages using a unified pipeline indicates a robust framework capable of adapting to diverse linguistic contexts. This

adaptability is crucial for expanding the applicability of financial tools globally, especially in regions where resources are limited. The development of tools like FiNCAT [16] and FENCE [26] demonstrates practical applications of these methodologies. These tools not only enhance the detection of in-claim numerals and exaggerated claims but also pave the way for future enhancements that could include multilingual support and improved user interface. The emphasis on user-friendly tools reflects a growing recognition of the need for accessible technology in finance.

Another significant aspect of this research is its focus on social media sentiment analysis, particularly concerning its impact on stock prices. The findings reveal that sentiments expressed by executives carries more weight than those from general users, underscoring the influence of authority figures in shaping public perception and market behaviour. This insight is critical for investors seeking to understand market dynamics influenced by social media discourse. The integration of sentiment analysis from platforms like Twitter and Reddit into price prediction models marks a significant advancement in financial forecasting. The results affirm that social media sentiment is a valuable predictor of stock movements, reinforcing the need for investors to consider these factors when making decisions.

This research encapsulates a multifaceted approach to understanding crowd reactions, financial sentiment, and entity relationships through advanced methodologies and innovative tools. The integration of LLMs with traditional models has proven effective in enhancing predictive capabilities across various domains within finance. As we move forward, it will be essential to continue refining these methodologies, expanding datasets thoughtfully, and developing accessible tools that empower investors with actionable insights. The intersection of technology and finance holds immense potential for future advancements, promising a more informed investment landscape driven by data-driven decision-making processes.

"Having systematically explored each dimension of financial text demystification, the answers to the research questions are as follows:

- **RQ-1:** How to quantify and improve readability of financial texts? The thesis addressed this through the creation of FinRAD, demonstrating empirically that standard readability formulae fail for financial definitions, and developing FinRead and FinLanSer tools that leverage domain-specific transformer models.
- **RQ-2:** How to ensure financial content reaches more people? The Generator-Guided Estimation Approach (GGEA) with the CRED dataset successfully predicted crowd reactions, demonstrating that LLM-augmented frameworks can optimize social media engagement for government communications.
- **RQ-3:** How to improve the investment process? Two methodological innovations: fine-tuned FinBERT embeddings with FIBO hierarchies for hypernym detection and the MOAT architecture for entity relationship extraction simplify information navigation for investors.
- **RQ-4**: How to ensure investments support environmental sustainability? Comprehensive ESG frameworks were established: SBERT-based sustainability classification, SEC-BERT ESG Issue Detector, and multilingual ESG impact identification across four languages, enabling investors to assess environmental impact systematically.

- **RQ-5**: How to safeguard investors from misinformation? Multiple detection systems were developed: FiNCAT/FiNCAT-2 for in-claim numerals, FENCE for exaggerated numerals, and empirical validation that executive social media sentiment predicts stock prices more accurately than general public sentiment.
- **RQ-6**: How to keep Indian investors informed? Pioneering contributions to Indic financial NLP include FAAB for Bengali argument analysis, IndicFinNLP datasets across Hindi, Bengali, and Telugu, and multi-modal IPO prediction frameworks outperforming state-of-the-art LLMs, establishing the first comprehensive Indian language financial NLP research programme.

8.1.1 Connecting Contributions to Research Pillars

The six research questions collectively address financial text demystification. Inclusive Investing (RQ-1, RQ-2) establishes the foundation by ensuring financial content is both comprehensible and accessible, creating the necessary conditions for widespread financial participation through FinRAD's readability quantification capabilities and GGEA's reach optimization. Improved Investing (RQ-3) builds upon this accessibility by providing navigational tools—hypernym detection and entity relationship extraction—that simplify the investor's information-seeking journey through complex financial ecosystems. Impactful Investing (RQ-4) extends the decision-making framework beyond traditional risk-return analysis to incorporate environmental sustainability through comprehensive ESG assessment capabilities, aligning investment choices with planetary well-being. Informed Investing (RQ-5) protects the integrity of this entire system by developing mechanisms to identify misinformation, from in-claim numerals to exaggerated claims, thereby safeguarding investors from manipulative content. Finally, Indic Investing (RQ-6) demonstrates the framework's adaptability and commitment to financial inclusion by extending all capabilities to underserved, low-resource Indian language contexts, ensuring emerging market participants have equal access to sophisticated financial analysis tools. This systematic coverage therefore demonstrates that demystifying financial texts requires simultaneous attention to readability, navigation, sustainability awareness, information integrity, and linguistic diversity.

8.1.1.1 Theoretical and Practical Contributions

This thesis advances computational linguistic theory through two principal architectural innovations that address domain-specific challenges in financial text processing. The MOAT (Mask One At a Time) framework introduces a novel approach to financial entity relationship extraction that outperforms general-purpose large language models in zero-shot settings by leveraging task-specific masking strategies tailored to financial discourse structures. The GGEA (Generator-Guided Estimation Approach) demonstrates a new paradigm for human-AI collaboration where large language models provide analytical reasoning that augments the performance of classification models, establishing a blueprint for combining interpretable LLM insights with discriminative model efficiency. Beyond theoretical contributions, the thesis delivers a comprehensive ecosystem of practical tools that democratize financial information access. FinRead enables content creators to assess definition readability in real time, FiNCAT-2 assists investors in distinguishing factual claims from speculative claims, FENCE identifies exaggerated numerical claims, and EID

classifies ESG issues across 33 predefined categories. These tools collectively transform financial text analysis from an expert-only activity to an accessible capability for retail investors, educators, and policymakers, particularly in resource-constrained settings where professional financial advisory services remain unaffordable. The open-source release of datasets (FinRAD, CRED, IndicFinNLP) and model implementations further amplify impact by enabling reproducible research and fostering community-driven innovation in financial NLP.

8.2 Limitations

In this thesis, we focused primarily on text data. However, with the rapid rise in digitalisation, investors consume content in various other forms like video, audio, etc. Due to the scarcity of experts with both financial domain knowledge and vernacular language expertise, we could not get large amounts of annotated data for low-resource languages. There is a clear opportunity to expand data collection efforts to further refine model performance across various tasks. In addition to this, we developed separate models for each task. Maintaining separate models is computationally expensive and operationally challenging. Separate models also lead to deployment challenges. This research began before large language models became prevalent, which limited their application to certain tasks. Additionally, owing to limited computational resources, we focused exclusively on small language models. For the majority of our evaluations, we employed standard metrics to assess model performance. Additionally, our limited resources and restricted access to a pool of investors prevented us from thoroughly studying the real impact of this research on their investment journeys.

We have not yet investigated Agentic AI frameworks and Reasoning Language Models. These emerging technologies offer substantial potential for advancing research, particularly in the analysis of financial documents and decision-making processes. By integrating Agentic AI with Reasoning Language Models, we could unlock new frontiers in financial analysis and strategic decision-making. Such integration could enable more sophisticated and autonomous systems capable of interpreting complex financial data and providing informed recommendations, thus representing a compelling direction for future work.

8.3 Future works

Potential directions of future work in general are: embracing more modalities (i.e., using audio, video, etc. along with texts), focusing on more languages (specifically the low-resource ones), and developing multitask models. The theme-wise possible extensions of our work are mentioned below:

Inclusive Investing:

• Annotating data with various degrees of readability instead of using coarse-grained classes to identify readable financial texts

• Creating a corpus having a simplified version of complex financial texts and using them to train a financial text simplification model

• Assessing reach of individual financial text rather than pairwise comparison

Improved Investing:

• The number of hypernyms our models can predict is limited to the taxonomy being used. Making this process loosely coupled to the overall system will help to deal with frequent changes in the taxonomy.
• The number of entities being considered in the MOAT [9] is limited. This can be extended to include more entities and relation types.
Impactful (Green) Investing:
• In addition to improving performance of the existing models, new tasks like automatically creating thematic green funds from existing equities and measuring their performance can be proposed.
Informed Investing:
• Assessing how much the content created by FinFluencers are in line with the guidelines prescribed by regulators like SEBI, FINRA, etc.
Indic Investing:
• Analysing the financial impact of policies implemented by different state and central governments
• Creating an India-specific knowledge graph exclusively for the financial domain
Tools:
• Integrating the open-source tools with popular trading platforms for helping the investors to enable data-driven decisions
• Assessing how much impact and value each proposed tool is creating. Accordingly, these tools can be refined further.

Bibliography

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[2] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance, 2023. URL https://arxiv.org/abs/2303.17564.

[3] Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. MultiFin: A dataset for multilingual financial NLP. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 894–909, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.66. URL https://aclanthology.org/2023.findings-eacl.66.

[4] Sohom Ghosh, Shovon Sengupta, Sudip Kumar Naskar, and Sunny Kumar Singh. Finrad: Financial readability assessment dataset - 13,000+ definitions of financial terms for measuring readability. In *Proceedings of the The 4th Financial Narrative Processing Workshop (FNP@LREC2022)*, pages 1–9, Marseille, France, 06 2022. European Language Resources Association. URL http://lrec-conf.org/proceedings/lrec2022/workshops/FNP/pdf/2022.fnp-1.1.pdf.

[5] Sohom Ghosh, Shovon Sengupta, Sudip Naskar, and Sunny Kumar Singh. FinRead: A transfer learning based tool to assess readability of definitions of financial terms. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 658–659, National Institute of Technology Silchar, Silchar, India, December 2021. NLP Association of India (NLPAI). URL https://aclanthology.org/2021.icon-main.81.

[6] Sohom Ghosh, Chung-Chi Chen, and Sudip Kumar Naskar. Generator-guided crowd reaction assessment. In *Companion Proceedings of the ACM Web Conference 2024*, 05 2024. URL https://doi.org/10.1145/3589335.3651512.

[7] Ankush Chopra and Sohom Ghosh. Term expansion and FinBERT fine-tuning for hypernym and synonym ranking of financial terms. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 46–51, Online, 9 August 2021. -. URL https://aclanthology.org/2021.finnlp-1.8.

[8] Sohom Ghosh, Ankush Chopra, and Sudip Kumar Naskar. Learning to rank hypernyms of financial terms using semantic textual similarity. *SN Computer Science*, 4 (5):610, 2023. URL https://doi.org/10.1007/s42979-023-02134-z.

[9] Sohom Ghosh, Sachin Umrao, Chung-Chi Chen, and Sudip Kumar Naskar. The mask one at a time framework for detecting the relationship between financial entities. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '23, page 40–43, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400716324. doi: 10.1145/3632754.3632756. URL https://doi.org/10.1145/3632754.3632756.

[10] Sohom Ghosh and Sudip Kumar Naskar. Ranking environment, social and governance related concepts and assessing sustainability aspect of financial texts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 243–249, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.finnlp-1.33.

[11] Priyank Soni, Sohom Ghosh, and Sudip Kumar Naskar. Detecting issues related to environmental, social, and corporate governance using sec-bert. In *Proceedings of International Conference on Data Science and Applications: ICDSA 2023*, volume 819, "Singapore", 2024. Springer, "Springer". URL https://doi.org/10.1007/978-981-99-7820-5_27.

[12] Harsha Vardhan, Sohom Ghosh, Ponnurangam Kumaraguru, and Sudip Naskar. A low resource framework for multi-lingual esg impact type identification. In Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, Hsin-Hsi Chen, Hiroki Sakaji, and Kiyoshi Izumi, editors, *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, pages 57–61, Bali, Indonesia, November 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.finnlp-2.8.

[13] Neelabha Banerjee, Anubhav Sarkar, Swagata Chakraborty, Sohom Ghosh, and Sudip Kumar Naskar. Fine-tuning language models for predicting the impact of events associated to financial news articles. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing (FinNLP), the 5th Knowledge Discovery from Unstructured Data in Financial Services (KDF), and The 4th Workshop on Economics and Natural Language Processing (ECONLP)*, pages 244—247, Torino, Italy, May 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.finnlp-1.25/.

[14] Sohom Ghosh and Sudip Kumar Naskar. Detecting context-based in-claim numerals in financial earnings conference calls. *International Journal of Information Technology*, 14:2559–2566, 2022. URL https://doi.org/10.1007/s41870-022-00952-7.

[15] Sohom Ghosh and Sudip Kumar Naskar. Fincat-2: An enhanced financial numeral claim analysis tool. *Software Impacts*, 12:100288, 2022. ISSN 2665-9638. doi: https://doi.org/10.1016/j.simpa.2022.100288. URL https://www.sciencedirect.com/science/article/pii/S2665963822000367.

[16] Sohom Ghosh and Sudip Kumar Naskar. Fincat: Financial numeral claim analysis tool. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion)*, New York, NY, USA, April 2022. Association for Computing Machinery. ISBN

978-1-4503-9130-6/22/04. doi: 10.1145/3487553.3524635. URL https://dl.acm.org/doi/10.1145/3487553.3524635.

[17] Sohom Ghosh and Sudip Kumar Naskar. Lipi at the ntcir-16 finnum-3 task: ensembling transformer based models to detect in-claim numerals in financial conversations. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, pages 92–94, Tokyo, Japan, June 2022. NII. URL https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings16/pdf/ntcir/02-NTCIR16-FINNUM-GhoshS.pdf.

[18] Sohom Ghosh and Sudip Kumar Naskar. LIPI at the FinNLP-2022 ERAI task: Ensembling sentence transformers for assessing maximum possible profit and loss from online financial posts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 111–115, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.finnlp-1.13. URL https://aclanthology.org/2022.finnlp-1.13.

[19] Anubhav Sarkar, Swagata Chakraborty, Sohom Ghosh, and Sudip Kumar Naskar. Evaluating impact of social media posts by executives on stock prices. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '22, page 74–82, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9798400700231. doi: 10.1145/3574318.3574339. URL https://doi.org/10.1145/3574318.3574339.

[20] Swagata Chakraborty, Anubhav Sarkar, Dhairya Suman, Sohom Ghosh, and Sudip Kumar Naskar. Lipi at the ntcir-17 finarg-1 task: Using pre-trained language models for comprehending financial arguments. In *Proceedings of the 17th NTCIR conference on evaluation of information access technologies.*, pages 29–36, Tokyo, Japan, 2023. NII. URL https://doi.org/10.20736/0002001281.

[21] Rima Roy, Sohom Ghosh, and Sudip Kumar Naskar. Financial argument analysis in bengali. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '23, page 88–92, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400716324. doi: 10.1145/3632754.3632763. URL https://doi.org/10.1145/3632754.3632763.

[22] Sohom Ghosh, Arnab Majhi, Aswartha Narayana, and Sudip Kumar Naskar. Indicfinnlp: Financial natural language processing for indian languages. In *LREC-COLING 2024*, pages 9010—9018, 05 2024. URL "https://aclanthology.org/2024.lrec-main.789/".

[23] Sohom Ghosh and Sudip Kumar Naskar. Fluent: Financial language understandability enhancement toolkit. In *6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD) (CODS-COMAD 2023)*, pages 258—262, New York, NY, USA, January 2023. Association for Computing Machinery. ISBN 978-1-4503-9797-1/23/01. doi: 10.1145/3570991.3571067. URL https://dl.acm.org/doi/abs/10.1145/3570991.3571067.

[24] Sohom Ghosh. Demystifying financial texts using natural language processing. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24),*, 10 2024. URL https://doi.org/10.1145/3627673.3680258.

[25] Sohom Ghosh and Sudip Kumar Naskar. Using natural language processing to enhance understandability of financial texts. In *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)*, CODS-COMAD '23, page 301–302, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450397971. doi: 10.1145/3570991.3571051. URL https://doi.org/10.1145/3570991.3571051.

[26] Sohom Ghosh and Sudip Kumar Naskar. Fence: Financial exaggerated numeral classifier, 2023. URL https://easychair.org/publications/preprint_download/s9Ds.

[27] Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3): 221, 1948. URL https://doi.org/10.1037/h0057532.

[28] E A Smith and R. Senter. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories*, pages 1–14, 1967. URL https://apps.dtic.mil/sti/pdfs/AD0667273.pdf.

[29] G Harry Mc Laughlin. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646, 1969. URL http://www.jstor.org/stable/40011226.

[30] Jeanne Sternlicht Chall and Edgar Dale. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, 1995. URL https://cir.nii.ac.jp/crid/1130282268845043712.

[31] R Timothy Rush. Assessing readability: Formulas and alternatives. *The Reading Teacher*, 39(3):274–283, 1985. URL https://www.jstor.org/stable/20199072.

[32] Bertram Bruce, Andee Rubin, and Kathleen Starr. Why readability formulas fail. *IEEE Transactions on Professional Communication*, PC-24(1):50–52, 1981. doi: 10.1109/TPC.1981.6447826. URL https://doi.org/10.1109/TPC.1981.6447826.

[33] Richard Chase Anderson and Alice Davison. Conceptual and empirical bases of readability formulas, 1986. URL https://psycnet.apa.org/record/1988-97085-002.

[34] Mostafa Zamanian and Pooneh Heydari. Readability of texts: State of the art. *Theory & Practice in Language Studies*, 2(1):43–53, 2012. URL https://doi.org/10.4304/tpls.2.1.43-53.

[35] Luo Si and Jamie Callan. A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, page 574–576, New York, NY, USA, 2001. Association for Computing Machinery. ISBN 1581134363. doi: 10.1145/502585.502695. URL https://doi.org/10.1145/502585.502695.

[36] Kevyn Collins-Thompson and James P. Callan. A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 193–200, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. URL https://aclanthology.org/N04-1025.

[37] Sarah E. Schwarm and Mari Ostendorf. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 523–530, USA,

2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219905. URL https://doi.org/10.3115/1219840.1219905.

[38] Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 460–467, Rochester, New York, April 2007. Association for Computational Linguistics. URL https://aclanthology.org/N07-1058.

[39] Jukka Ruohonen. Assessing the readability of policy documents on the digital single market of the european union, 2021. URL https://doi.org/10.1109/ICEDEG52154.2021.9530996.

[40] Tim Loughran and Bill McDonald. Measuring readability in financial disclosures. *the Journal of Finance*, 69(4):1643–1671, 2014. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12162.

[41] Anne-Marie Gosselin, Julien Le Maux, and Nadia Smaili. Readability of accounting disclosures: A comprehensive review and research agenda*. *Accounting Perspectives*, 20(4):543–581, 2021. doi: https://doi.org/10.1111/1911-3838.12275. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1911-3838.12275.

[42] Somya Arora and Yogesh Chauhan. Do earnings management practices define the readability of the financial reports in india? *Journal of Public Affairs*, n/a(n/a):e2692, 05 2021. doi: https://doi.org/10.1002/pa.2692. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pa.2692.

[43] Nicholas Schroeder and Charles Gibson. Readability of management's discussion and analysis. *Accounting Horizons*, 4(4):78–87, 1990. URL https://faculty.etsu.edu/POINTER/schroeder_n.pdf.

[44] James E. Smith and Nora P. Smith. Readability: A measure of the performance of the communication function of financial reporting. *The Accounting Review*, 46(3):552–561, 1971. ISSN 00014826. URL http://www.jstor.org/stable/244524.

[45] Kin Lo, Felipe Ramos, and Rafael Rogo. Earnings management and annual report readability. *Journal of Accounting and Economics*, 63(1):1–25, 2017. ISSN 0165-4101. doi: https://doi.org/10.1016/j.jacceco.2016.09.002. URL https://www.sciencedirect.com/science/article/pii/S0165410116300544.

[46] Samuel B Bonsall IV, Andrew J Leone, Brian P Miller, and Kristina Rennekamp. A plain english measure of financial reporting readability. *Journal of Accounting and Economics*, 63(2-3):329–357, 2017. URL https://doi.org/10.1016/j.jacceco.2017.03.002.

[47] Tim Loughran and Bill McDonald. Measuring readability in financial text, 2010. URL http://securitieseditor.com/wp/wp-content/uploads/2014/05/Plain_English_v5.pdf.

[48] Intan Waheedah Othman, Huda Hadi Hasan, Roszana Tapsir, Nurazzah Abdul Rahman, Indarawati Tarmuji, Suria Majdi, Seri Ayu Masuri, and Najma Omar. Text readability and fraud detection, 2012. URL https://doi.org/10.1109/ISBEIA.2012.6422890.

[49] Chansog Kim, Ke Wang, and Liandong Zhang. Readability of 10-k reports and stock price crash risk. *Contemporary accounting research*, 36(2):1184–1216, 2019. URL https://doi.org/10.1111/1911-3846.12452.

[50] Wei-Chih Chiang, Ted D Englebrecht, Thomas J Phillips Jr, and Ying Wang. Readability of financial accounting principles textbooks. *The Accounting Educators' Journal*, 18:47–80, 2008. URL https://www.aejournal.com/ojs/index.php/aej/article/view/74.

[51] Kenneth J. Plucinski and Mojtaba Seyedian. Readability of introductory finance textbooks. *Journal of Financial Education*, 39(1/2):43–52, 2013. ISSN 00933961, 2332421X. URL http://www.jstor.org/stable/41948697.

[52] Kenneth J Plucinski, John Olsavsky, and Linda Hall. Readability of introductory financial and managerial accounting textbooks. *Academy of Educational Leadership Journal*, 13(4):119, 2009. URL https://www.proquest.com/openview/c0e1c4cb0563d074e9c9cecb352b1b49/1?pq-origsite=gscholar&cbl=38741.

[53] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL https://aclanthology.org/D08-1020.

[54] Tim Loughran and Bill McDonald. Plain english, readability, and 10-k filings, 2009. URL https://www.researchgate.net/profile/Bill-Mcdonald/publication/228458241_Plain_English_Readability_and_10-K_Filings/links/5772a80d08aeeec3895410b0/Plain-English-Readability-and-10-K-Filings.pdf.

[55] Abhishek Ganguly, Arup Ganguly, Lin Ge, and Chad Zutter. Shareholder litigation and readability in financial disclosures: Evidence from a natural experiment, 2019. URL https://sbfc.sydney.edu.au/2019/papers/P100_Named.pdf.

[56] Arvid OI Hoffmann and Stefanie Kleimeier. Financial disclosure readability and innovative firms' cost of debt. *International Review of Finance*, 21(2):699–713, 2021. URL https://doi.org/10.1111/irfi.12292.

[57] R.A. Brealey, S.C. Myers, and F. Allen. *Principles of Corporate Finance*. Economia e discipline aziendali. McGraw-Hill Education, USA, 2019. ISBN 9781260565553. URL https://books.google.co.in/books?id=0280wAEACAAJ.

[58] Zvi Bodie and Alex Kane. Investments, 2020. URL http://elibrary.gci.edu.np/handle/123456789/707.

[59] Erik Banks. *The Palgrave Macmillan Dictionary of Finance, Investment and Banking*. Palgrave Macmillan, London, UK, 2010. URL https://doi.org/10.1057/9780230251212.

[60] John C Hull. *Options futures and other derivatives*. Pearson Education India, 2003. URL http://userhome.brooklyn.cuny.edu/bassell/myles_bassell/Myles_Bassell_704S.pdf.

[61] Frederic S Mishkin and Stanley G Eakins. *Financial markets and institutions*. Pearson Education India, 2006. URL https://www.solbridge.ac.kr/site/main/down/BBA/FIN407%20Financial%20Markets%20and%20Institutions.pdf.

[62] Paul Samuelson and V Nordhouse. Economics: a textbook, 2010. URL https://www.mheducation.com/highered/product/economics-samuelson-nordhaus/M9780073511290.html.

[63] Susmoy Chakraborty, Mir Tafseer Nayeem, and Wasi Uddin Ahmad. Simple or complex? learning to predict readability of bengali texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12621–12629, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17495.

[64] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975. URL https://api.semanticscholar.org/CorpusID:61131325.

[65] Meri Coleman and Ta Lin Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283, 1975. URL https://psycnet.apa.org/record/1975-22007-001.

[66] Tin Kam Ho. Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–282 vol.1, 1995. doi: 10.1109/ICDAR.1995.598994. URL https://ieeexplore.ieee.org/document/598994.

[67] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 00905364. URL http://www.jstor.org/stable/2699986.

[68] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

[69] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models, 2019. URL https://openreview.net/forum?id=HylznxrYDr.

[70] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, pages 3146–3154, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[71] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

[72] Maxim Kuznetsov Vladimir Vorobev. A paraphrasing model based on chatgpt paraphrases, 2023. URL https://huggingface.co/humarin/chatgpt_paraphraser_on_T5_base.

[73] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[74] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

[75] Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6ruVLB727MC.

[76] Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. When flue meets flang: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2022.

[77] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. *Proceedings of the International AAAI Conference on Web and Social Media*, 4(1):10–17, May 2010. doi: 10.1609/icwsm.v4i1.14033. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14033.

[78] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE Second International Conference on Social Computing*, pages 177–184, 2010. doi: 10.1109/SocialCom.2010.33.

[79] Chenhao Tan, Lillian Lee, and Bo Pang. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 175–185, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1017. URL https://aclanthology.org/P14-1017.

[80] Cheng Li, Jiaqi Ma, Xiaoxiao Guo, and Qiaozhu Mei. Deepcas: An end-to-end predictor of information cascades. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 577–586, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349130. doi: 10.1145/3038912.3052643. URL https://doi.org/10.1145/3038912.3052643.

[81] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM

'11, page 65–74, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450304931. doi: 10.1145/1935826.1935845. URL https://doi.org/10.1145/1935826.1935845.

[82] Michael S. Bernstein, Eytan Bakshy, Moira Burke, and Brian Karrer. Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 21–30, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318990. doi: 10.1145/2470654.2470658. URL https://doi.org/10.1145/2470654.2470658.

[83] Lauri Valkonen, Jouni Helske, and Juha Karvanen. Estimating the causal effect of timing on the reach of social media posts. *Statistical Methods & Applications*, pages 1–15, 2022. URL https://doi.org/10.1007/s10260-022-00664-z.

[84] Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3108–3127, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.222. URL https://aclanthology.org/2022.acl-long.222.

[85] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019. URL https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf.

[86] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[87] Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vitor Silva, Leonardo Neves, and Francesco Barbieri. Twitter topic classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3386–3400, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.299.

[88] Syeda Nadia Firdaus, Chen Ding, and Alireza Sadeghian. Retweet: A popular information diffusion mechanism – a survey paper. *Online Social Networks and Media*, 6:26–40, 2018. ISSN 2468-6964. doi: https://doi.org/10.1016/j.osnem.2018.04.001. URL https://www.sciencedirect.com/science/article/pii/S2468696417300952.

[89] Prithiviraj Damodaran. Parrot: Paraphrase generation for nlu., 2021. URL https://github.com/PrithivirajDamodaran/Parrot_Paraphraser".

[90] Silei Xu, Sina Semnani, Giovanni Campagna, and Monica Lam. AutoQA: From databases to QA semantic parsers with only synthetic training data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 422–434, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-main.31.

[91] Sohom Ghosh and Sudip Kumar Naskar. Lipi at fincausal 2022: Mining causes and effects from financial texts. In *Proceedings of the The 4th Financial Narrative Processing Workshop (FNP@LREC2022)*, pages 130–132, Marseille, France, June

2022. European Language Resources Association. URL https://aclanthology.org/2022.fnp-1.21.

[92] Ismail El Maarouf, Youness Mansar, Virginie Mouilleron, and Dialekti Valsamou-Stanislawski. The FinSim 2020 shared task: Learning semantic representations for the financial domain. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 81–86, Kyoto, Japan, 5 January 2020. URL https://www.aclweb.org/anthology/2020.finnlp-1.13.

[93] Youness Mansar, Juyeon Kang, and Ismail El Maarouf. *The FinSim-2 2021 Shared Task: Learning Semantic Similarities for the Financial Domain*, page 288–292. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450383134. URL https://doi.org/10.1145/3442442.3451381.

[94] Juyeon Kang, Ismail El Maarouf, Sandra Bellato, and Mei Gan. FinSim-3: The 3rd shared task on learning semantic similarities for the financial domain. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 31–35, Online, 19 August 2021. -. URL https://aclanthology.org/2021.finnlp-1.5.

[95] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*, pages –, -, 1992. -. URL https://aclanthology.org/C92-2082.

[96] Rion Snow, Daniel Jurafsky, and Andrew Y Ng. Learning syntactic patterns for automatic hypernym discovery. In *Advances in neural information processing systems*, pages 1297–1304, -, 2005. -. URL https://proceedings.neurips.cc/paper/2004/file/358aee4cc897452c00244351e4d91f69-Paper.pdf.

[97] George A Miller. *WordNet: An electronic lexical database*. MIT press, -, 1998. URL https://ieeexplore.ieee.org/servlet/opac?bknumber=6267389.

[98] Erik Tjong Kim Sang. Extracting hypernym pairs from the web. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 165–168, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/P07-2042.

[99] Erik Tjong Kim Sang and Katja Hofmann. Lexical patterns or dependency patterns: Which is better for hypernym extraction? In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 174–182, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL https://aclanthology.org/W09-1122.

[100] Alan Ritter, Stephen Soderland, and Oren Etzioni. What is this, anyway: Automatic hypernym discovery. In *AAAI Spring Symposium: Learning by Reading and Learning to Read*, pages 88–93, -, 2009. -. URL https://www.aaai.org/Papers/Symposia/Spring/2009/SS-09-07/SS09-07-015.pdf.

[101] Sharon A. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 120–126, College Park, Maryland, USA, June 1999. Association for Computational Linguistics. doi: 10.3115/1034678.1034705. URL https://aclanthology.org/P99-1016.

[102] Keiji Shinzato and Kentaro Torisawa. Acquiring hyponymy relations from web documents. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 73–80, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. URL https://aclanthology.org/N04-1010.

[103] Roberto Navigli and Paola Velardi. Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1318–1327, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/P10-1134.

[104] Stefano Faralli and Roberto Navigli. A Java framework for multilingual definition and hypernym extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 103–108, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-4018.

[105] Yixin Tan, Xiaomeng Wang, and Tao Jia. From syntactic structure to semantic relationship: Hypernym extraction from definitions by recurrent neural networks using the part of speech information. In Jeff Z. Pan, Valentina Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne, and Lalana Kagal, editors, *The Semantic Web – ISWC 2020*, pages 529–546, Cham, 2020. Springer International Publishing. ISBN 978-3-030-62419-4. URL https://doi.org/10.1007/978-3-030-62419-4_30.

[106] Jiaqing Liang, Yi Zhang, Yanghua Xiao, Haixun Wang, Wei Wang, and Pinpin Zhu. On the transitivity of hypernym-hyponym relations in data-driven lexical taxonomies. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 1185–1191, San Francisco, California, USA, 2017. AAAI Press. URL https://doi.org/10.1609/aaai.v31i1.10675.

[107] Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. SemEval-2015 task 17: Taxonomy extraction evaluation (TExEval). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 902–910, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-2151. URL https://aclanthology.org/S15-2151.

[108] Georgeta Bordea, Els Lefever, and Paul Buitelaar. SemEval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1081–1091, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1168. URL https://aclanthology.org/S16-1168.

[109] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2091. URL https://aclanthology.org/S17-2091.

[110] Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio

Saggion. SemEval-2018 task 9: Hypernym discovery. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 712–724, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1115. URL https://aclanthology.org/S18-1115.

[111] Gregory Grefenstette. INRIASAC: Simple hypernym extraction methods. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 911–914, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-2152. URL https://aclanthology.org/S15-2152.

[112] Alexander Panchenko, Stefano Faralli, Eugen Ruppert, Steffen Remus, Hubert Naets, Cédrick Fairon, Simone Paolo Ponzetto, and Chris Biemann. TAXI at SemEval-2016 task 13: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1320–1327, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1206. URL https://aclanthology.org/S16-1206.

[113] Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. MIT at SemEval-2017 task 10: Relation extraction with convolutional neural networks. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 978–984, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2171. URL https://aclanthology.org/S17-2171.

[114] Gabriel Bernier-Colborne and Caroline Barrière. CRIM at SemEval-2018 task 9: A hybrid approach to hypernym discovery. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 725–731, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1116. URL https://aclanthology.org/S18-1116.

[115] Sarthak Dash, Md. Faisal Mahbub Chowdhury, Alfio Gliozzo, Nandana Mihindukulasooriya, and Nicolas Rodolfo Fauceglia. Hypernym detection using strict partial order networks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7626–7633, New York, USA, 2020. AAAI Press. URL https://aaai.org/ojs/index.php/AAAI/article/view/6263.

[116] Yuhang Bai, Richong Zhang, Fanshuang Kong, Junfan Chen, and Yongyi Mao. Hypernym discovery via a recurrent mapping model. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2912–2921, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.257. URL https://aclanthology.org/2021.findings-acl.257.

[117] Vishal Keswani, Sakshi Singh, and Ashutosh Modi. IITK at the FinSim task: Hypernym detection in financial domain via context-free and contextualized word embeddings. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 87–92, Kyoto, Japan, 5 January 2020. URL https://www.aclweb.org/anthology/2020.finnlp-1.14.

[118] Emmanuele Chersoni and Chu-Ren Huang. *PolyU-CBS at the FinSim-2 Task: Combining Distributional, String-Based and Transformers-Based Features for Hypernymy Detection in the Financial Domain*, page 316–319. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450383134. URL https://doi.org/10.1145/3442442.3451387.

[119] Nadine Kroher, Aggelos Pikrakis, Simon White, and Joe Lyske. MXX@FinSim3 - an LSTM–based approach with custom word embeddings for hypernym detection in financial texts. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 36–39, Online, 19 August 2021. -. URL https://aclanthology.org/2021.finnlp-1.6.

[120] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

[121] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[122] Jan Portisch, Michael Hladik, and Heiko Paulheim. *FinMatcher at FinSim-2: Hypernym Detection in the Financial Services Domain Using Knowledge Graphs*, page 293–297. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450383134. URL https://doi.org/10.1145/3442442.3451382.

[123] Timen Stepišnik Perdih, Senja Pollak, and Blaž Škrlj. *JSI at the FinSim-2 Task: Ontology-Augmented Financial Concept Classification*, page 298–301. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450383134. URL https://doi.org/10.1145/3442442.3451383.

[124] Chao Feng and Shijie Wei. Exploiting network structures to improve semantic representation for the financial domain. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 58–62, Online, 19 August 2021. -. URL https://aclanthology.org/2021.finnlp-1.10.

[125] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://doi.org/10.48550/arXiv.1907.11692.

[126] Lefteris Loukas, Manos Fergadiotis, Ion Androutsopoulos, and Prodromos Malakasiotis. EDGAR-CORPUS: Billions of tokens make the world go round. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 13–18, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.econlp-1.2. URL https://aclanthology.org/2021.econlp-1.2.

[127] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data, 2007. URL https://doi.org/10.1007/978-3-540-76298-0_52.

[128] Yulong Pei and Qian Zhang. Goat at the finsim-2 task: Learning word representations of financial data with customized corpus. In *Companion Proceedings of the Web Conference 2021*, WWW '21, page 307–310, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383134. doi: 10.1145/3442442.3451385. URL https://doi.org/10.1145/3442442.3451385.

[129] Anuj Saini. Anuj at the FinSim task: Anuj@FINSIM¡VLearning semantic representation of financial domain with investopedia. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 93–97, Kyoto, Japan, 5 January 2020. URL https://www.aclweb.org/anthology/2020.finnlp-1.15.

[130] Gábor Berend, Norbert Kis-Szabó, and Zsolt Szántó. ProsperAMnet at the FinSim task: Detecting hypernyms of financial concepts via measuring the information stored in sparse word representations. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 98–103, Kyoto, Japan, 5 January 2020. URL https://www.aclweb.org/anthology/2020.finnlp-1.16.

[131] Vivek Anand, Yash Agrawal, Aarti Pol, and Vasudeva Varma. FINSIM20 at the FinSim task: Making sense of text in financial domain. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pages 104–107, Kyoto, Japan, 5 January 2020. URL https://www.aclweb.org/anthology/2020.finnlp-1.17.

[132] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2029. URL https://aclanthology.org/D18-2029.

[133] Ke Tian and Hua Chen. Aiai at the finsim-2 task: Finance domain terms automatic classification via word ontology and embedding. In *Companion Proceedings of the Web Conference 2021*, WWW '21, page 320–322, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383134. doi: 10.1145/3442442.3451388. URL https://doi.org/10.1145/3442442.3451388.

[134] Nhu Khoa Nguyen, Emanuela Boros, Gael Lejeune, Antoine Doucet, and Thierry Delahaut. *L3i LBPAM at the FinSim-2 Task: Learning Financial Semantic Similarities with Siamese Transformers*, page 302–306. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450383134. URL https://doi.org/10.1145/3442442.3451384.

[135] Tushar Goel, Vipul Chauhan, Ishan Verma, Tirthankar Dasgupta, and Lipika Dey. *TCS WITM 2021 @FinSim-2: Transformer Based Models for Automatic Classification of Financial Terms*, page 311–315. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450383134. URL https://doi.org/10.1145/3442442.3451386.

[136] Lefteris Loukas, Konstantinos Bougiatiotis, Manos Fergadiotis, Dimitris Mavroeidis, and Elias Zavitsanos. DICoE@FinSim-3: Financial hypernym detection using augmented terms and distance-based features. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 40–45, Online, 19 August 2021. -. URL https://aclanthology.org/2021.finnlp-1.7.

[137] Hanna Abi Akl, Dominique Mariko, and Hugues de Mazancourt. Yseop at FinSim-3 shared task 2021: Specializing financial domain learning with phrase representations. In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, pages 52–57, Online, 19 August 2021. -. URL https://aclanthology.org/2021.finnlp-1.9.

[138] Tomáš Kliegr. Linked hypernyms: Enriching dbpedia with targeted hypernym discovery. *Journal of Web Semantics*, 31:59–69, 2015. ISSN 1570-8268. doi: https://doi.org/10.1016/j.websem.2014.11.001. URL https://www.sciencedirect.com/science/article/pii/S1570826814001048.

[139] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply, 2017. URL https://doi.org/10.48550/arXiv.1705.00652.

[140] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[141] Asier Gutiérrez-Fandiño, Petter N Kolm, Miquel Noguer i Alonso, and Jordi Armengol-Estapé. Fineas: Financial embedding analysis of sentiment. *The Journal of Financial Data Science*, 4(3):45–53, 2022. doi: 10.3905/jfds.2022.1.095. URL https://www.pm-research.com/content/iijjfds/4/3/45.

[142] Juyeon Kang and Ismail El Maarouf. FinSim4-ESG shared task: Learning semantic similarities for the financial domain. extended edition to ESG insights. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 211–217, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.finnlp-1.28.

[143] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/S10-1006.

[144] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[145] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1279. URL https://aclanthology.org/P19-1279.

[146] Soumya Sharma, Tapas Nayak, Arusarka Bose, Ajay Kumar Meena, Koustuv Dasgupta, Niloy Ganguly, and Pawan Goyal. Finred: A dataset for relation extraction in financial domain. In *Companion Proceedings of the Web Conference*

2022, WWW '22, page 595–597, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391306. doi: 10.1145/3487553.3524637. URL https://doi.org/10.1145/3487553.3524637.

[147] Simerjot Kaur, Charese Smiley, Akshat Gupta, Joy Sain, Dongsheng Wang, Suchetha Siddagangappa, Toyin Aguda, and Sameena Shah. Refind: Relation extraction financial dataset. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 3054–3063, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3591911. URL https://doi.org/10.1145/3539618.3591911.

[148] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.523. URL https://aclanthology.org/2020.emnlp-main.523.

[149] Somin Wadhwa, Silvio Amir, and Byron Wallace. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.868. URL https://aclanthology.org/2023.acl-long.868.

[150] Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. A comprehensive survey on deep learning for relation extraction: Recent advances and new frontiers, 2023. URL https://doi.org/10.48550/arXiv.2306.02051.

[151] Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. FiNER: Financial numeric entity recognition for XBRL tagging. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4419–4431, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.303. URL https://aclanthology.org/2022.acl-long.303.

[152] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. Falcon-40B: an open large language model with state-of-the-art performance, 2023. URL https://huggingface.co/tiiuae/falcon-7b-instruct.

[153] Hugo Touvron, Louis Martin, and Kevin Stone et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/.

[154] Nils Smeuninx, Bernard De Clerck, and Walter Aerts. Measuring the readability of sustainability reports: A corpus-based analysis through standard formulae and nlp. *International Journal of Business Communication*, 57(1):52–85, 2020. doi: 10.1177/2329488416675456. URL https://doi.org/10.1177/2329488416675456.

[155] Alexandra Luccioni, Emily Baylor, and Nicolas Duchene. Analyzing sustainability reports using natural language processing, 2020. URL https://arxiv.org/abs/2011.08073.

[156] Tian Guo, Nicolas Jamet, Valentin Betrix, Louis-Alexandre Piquet, and Emmanuel Hauptmann. Esg2risk: A deep learning framework from esg news to stock volatility prediction, 2020. URL https://arxiv.org/abs/2005.02527.

[157] Tim Nugent, Nicole Stelea, and Jochen L. Leidner. Detecting esg topics using domain-specific language models and data augmentation approaches, 2020. URL https://arxiv.org/abs/2010.08319.

[158] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c3a690be93aa602ee2dc0ccab5b7b67e-Paper.pdf.

[159] Samuel Borms, Kris Boudt, Frederiek Van Holle, and Joeri Willems. Semi-supervised text mining for monitoring the news about the esg performance of companies. In *Data Science for Economics and Finance: Methodologies and Applications*, pages 217–239. Springer International Publishing Cham, 2021. URL https://doi.org/10.1007/978-3-030-66891-4_10.

[160] Caterina Lucarelli, Camilla Mazzoli, Michela Rancan, and Sabrina Severini. Classification of sustainable activities: Eu taxonomy and scientific literature. *Sustainability*, 12(16):6460, 2020. URL https://doi.org/10.3390/su12166460.

[161] Nicole Darnall, Hyunjung Ji, Kazuyuki Iwata, and Toshi H Arimura. Do esg reporting guidelines and verifications enhance firms' information disclosure? *Corporate Social Responsibility and Environmental Management*, 29(5):1214–1230, 2022. URL https://doi.org/10.1002/csr.2265.

[162] Mario La Torre, Fabiomassimo Mango, Arturo Cafaro, and Sabrina Leo. Does the esg index affect stock return? evidence from the eurostoxx50. *Sustainability*, 12(16): 6387, 2020. URL https://doi.org/10.3390/su12166387.

[163] Guangyou Zhou, Lian Liu, and Sumei Luo. Sustainable development, esg performance and company market value: Mediating effect of financial performance. *Business Strategy and the Environment*, 31(7):3371–3387, 2022. URL https://doi.org/10.1002/bse.3089.

[164] Sang Kim and Zhichuan (Frank) Li. Understanding the impact of esg practices in corporate finance. *Sustainability*, 13(7), 2021. ISSN 2071-1050. doi: 10.3390/su13073746. URL https://www.mdpi.com/2071-1050/13/7/3746.

[165] Boshko Koloski, Syrielle Montariol, Matthew Purver, and Senja Pollak. Knowledge informed sustainability detection from short financial texts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 228–234, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.finnlp-1.31.

[166] Elvys Linhares Pontes, Mohamed Ben Jannet, Jose G. Moreno, and Antoine Doucet. Using contextual sentence analysis models to recognize ESG concepts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 218–223, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.finnlp-1.29.

[167] Gianpaolo Iazzolino, Maria Elena Bruni, Stefania Veltri, Donato Morea, and Giovanni Baldissarro. The impact of esg factors on financial efficiency: An empirical analysis for the selection of sustainable firm portfolios. *Corporate Social Responsibility and Environmental Management*, 2023. URL https://doi.org/10.1002/csr.2463.

[168] Zhongfei Chen and Guanxia Xie. Esg disclosure and financial performance: Moderating role of esg investors. *International Review of Financial Analysis*, 83:102291, 2022. ISSN 1057-5219. doi: https://doi.org/10.1016/j.irfa.2022.102291. URL https://www.sciencedirect.com/science/article/pii/S1057521922002472.

[169] Stuart L. Gillan, Andrew Koch, and Laura T. Starks. Firms and social responsibility: A review of esg and csr research in corporate finance. *Journal of Corporate Finance*, 66:101889, 2021. ISSN 0929-1199. doi: https://doi.org/10.1016/j.jcorpfin.2021.101889. URL https://www.sciencedirect.com/science/article/pii/S0929119921000092.

[170] Maria Giuseppina Bruna, Salvatore Loprevite, Domenico Raucci, Bruno Ricca, and Daniela Rupo. Investigating the marginal impact of esg results on corporate financial performance. *Finance Research Letters*, 47:102828, 2022. ISSN 1544-6123. doi: https://doi.org/10.1016/j.frl.2022.102828. URL https://www.sciencedirect.com/science/article/pii/S1544612322001283.

[171] Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. Multi-lingual ESG issue identification. In Chung-Chi Chen, Hiroya Takamura, Puneet Mathur, Remit Sawhney, Hen-Hsen Huang, and Hsin-Hsi Chen, editors, *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 111–115, Macao, 20 August 2023. -. URL https://aclanthology.org/2023.finnlp-1.11.

[172] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019. URL https://doi.org/10.48550/arXiv.1910.01108.

[173] Allen H Huang, Hui Wang, and Yi Yang. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 2022. URL https://doi.org/10.1111/1911-3846.12832.

[174] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild, 2019. URL https://doi.org/10.48550/arXiv.1906.02569.

[175] Tensie Whelan and Ulrich Atz. Esg and financial performance : Uncovering the relationship by aggregating evidence from 1 , 000 plus studies published between 2015 – 2020, 2021. URL https://api.semanticscholar.org/CorpusID:232216565.

[176] George Serafeim and Aaron Yoon. Stock price reactions to esg news: the role of esg ratings and disagreement. *Review of Accounting Studies*, 28, 03 2022. doi: 10.1007/s11142-022-09675-3.

[177] Jonathan B. Berk and Jules H. van Binsbergen. The Impact of Impact Investing. Research Papers 3981, Stanford University, Graduate School of Business, October 2021. URL https://ideas.repec.org/p/ecl/stabus/3981.html.

[178] Srishti Mehra, Robert Louka, and Yixun Zhang. ESGBERT: Language model to help with classification tasks related to companies' environmental, social, and governance practices. In *Embedded Systems and Applications*. Academy and Industry Research Collaboration Center (AIRCC), mar 2022. doi: 10.5121/csit.2022.120616. URL https://doi.org/10.5121%2Fcsit.2022.120616.

[179] Stefan Pasch and Daniel Ehnes. Nlp for responsible finance: Fine-tuning transformer-based models for esg. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3532–3536, 12 2022. doi: 10.1109/BigData55660.2022.10020755.

[180] Tanja Aue, Adam Jatowt, and Michael Färber. Predicting companies' esg ratings from news articles using multivariate timeseries analysis, 2022.

[181] Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Dynamicesg: A dataset for dynamically unearthing esg ratings from news articles. In *Proceedings of the 32nd ACM International Conference on Information & Knowledge Management*, CIKM '23, New York, NY, USA, 2023. Association for Computing Machinery. URL http://nlg.csie.ntu.edu.tw/~cjchen/papers/DynamicESG.pdf.

[182] Naoki Kannan and Yohei Seki. Textual evidence extraction for esg scores. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, 2023.

[183] Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Hanwool Lee, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. Multi-lingual esg impact duration inference. In *Proceedings of Joint Workshop of the 7th Financial Technology and Natural Language Processing and the 5th Knowledge Discovery from Unstructured Data in Financial Services*, 2024.

[184] Naoki Kannan and Yohei Seki. Textual evidence extraction for ESG scores. In Chung-Chi Chen, Hiroya Takamura, Puneet Mathur, Remit Sawhney, Hen-Hsen Huang, and Hsin-Hsi Chen, editors, *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 45–54, Macao, 20 August 2023. -. URL https://aclanthology.org/2023.finnlp-1.4.

[185] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[186] Chung-Chi Chen, Hen-Hsen Huang, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. Overview of the ntcir-16 finnum-3 task: investor's and manager's fine-grained claim detection. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, pages 87–91, Tokyo, Japan, June 2022. NII. URL https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings16/pdf/ntcir/01-NTCIR16-OV-FINNUM-ChenC.pdf.

[187] Revanth Gangi Reddy, Sai Chinthakindi, Zhenhailong Wang, Yi R. Fung, Kathryn S. Conger, Ahmed S. Elsayed, Martha Palmer, and Heng Ji. Newsclaims: A new benchmark for claim detection from news with background knowledge, 2021. URL https://blender.cs.illinois.edu/paper/newsclaims2022.pdf.

[188] Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. A context-aware approach for detecting worth-checking claims in political debates. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 267–276, Varna, Bulgaria, September 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6_037. URL https://doi.org/10.26615/978-954-452-049-6_037.

[189] Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. IMHO fine-tuning improves claim detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 558–563, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1054. URL https://aclanthology.org/N19-1054.

[190] Marco Lippi, Francesca Lagioia, Giuseppe Contissa, Giovanni Sartor, and Paolo Torroni. Claim detection in judgments of the eu court of justice. In Ugo Pagallo, Monica Palmirani, Pompeu Casanovas, Giovanni Sartor, and Serena Villata, editors, *AI Approaches to the Complexity of Legal Systems*, pages 513–527, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00178-0. URL https://doi.org/10.1007/978-3-030-00178-0_35.

[191] Naeemul Hassan, Mark Tremayne, Fatma Arslan, and Chengkai Li. Comparing automated factual claim detection against judgments of journalism organizations. In *Computation+ Journalism Symposium*, pages 1–5, 2016. URL https://journalism.stanford.edu/cj2016/files/Comparing%20Automated%20Factual%20Claim%20Detection%20Against%20Judgments%20of%20Journalism%20Organizations.pdf.

[192] Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats: Research and Practice*, 2 (2), apr 2021. ISSN 2692-1626. doi: 10.1145/3412869. URL https://doi.org/10.1145/3412869.

[193] Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-2109. URL https://aclanthology.org/W14-2109.

[194] Megha Sundriyal, Parantak Singh, Md. Shad Akhtar, Shubhashis Sengupta, and Tanmoy Chakraborty. *DESYR: Definition and Syntactic Representation Based Claim Detection on the Web*, page 1764–1773. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450384469. URL https://doi.org/10.1145/3459637.3482423.

[195] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1031. URL https://aclanthology.org/P18-1031.

[196] Dustin Wright and Isabelle Augenstein. Claim check-worthiness detection as positive unlabelled learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 476–488, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.43. URL https://aclanthology.org/2020.findings-emnlp.43.

[197] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL https://aclanthology.org/C14-1141.

[198] Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, pages 79–84, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5110. URL https://aclanthology.org/W17-5110.

[199] Marco Lippi and Paolo Torroni. Context-independent claim detection for argument mining. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 185–191. AAAI Press, 2015. ISBN 9781577357384. URL https://www.ijcai.org/Proceedings/15/Papers/033.pdf.

[200] Roy Bar-Haim, Lilach Edelstein, Charles Jochim, and Noam Slonim. Improving claim stance classification with lexical knowledge expansion and context utilization. In *Proceedings of the 4th Workshop on Argument Mining*, pages 32–38, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5104. URL https://aclanthology.org/W17-5104.

[201] Bjarte Botnevik, Eirik Sakariassen, and Vinay Setty. *BRENDA: Browser Extension for Fake News Detection*, page 2117–2120. Association for Computing Machinery, New York, NY, USA, 2020. ISBN 9781450380164. URL https://doi.org/10.1145/3397271.3401396.

[202] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Evaluating the rationales of amateur investors. In *Proceedings of the Web Conference 2021*, WWW '21, page 3987–3998, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3449964. URL https://doi.org/10.1145/3442381.3449964.

[203] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Numclaim: Investor's fine-grained claim detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 1973–1976, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3412100. URL https://doi.org/10.1145/3340531.3412100.

[204] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do NLP models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages

5307–5315, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1534. URL https://aclanthology.org/D19-1534.

[205] Abderrahim Ait Azzi and Houda Bouamor. Fortia1@ the ntcir-14 finnum task: enriched sequence labeling for numeral classification. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*, pages 526–538, 2019. URL http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings14/pdf/ntcir/02-NTCIR14-FINNUM-AzziA.pdf.

[206] Alan Spark. Brnir at the ntcir-14 finnum task: Scalable feature extraction technique for number classification. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*, 2019. URL https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings14/pdf/ntcir/03-NTCIR14-FINNUM-SparkA.pdf.

[207] Chao-Chun Liang and Keh-Yih Su. Asnlu at the ntcir-14 finnum task: incorporating knowledge into dnn for financial numeral classification. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*, volume 192, 2019. URL https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings14/pdf/ntcir/04-NTCIR14-FINNUM-LiangC.pdf.

[208] Yu-Yu Chen and Chao-Lin Liu. Mig at the ntcir-15 finnum-2 task: use the transfer learning and feature engineering for numeral attachment task. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*, 2020. URL https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings15/pdf/ntcir/02-NTCIR15-FINNUM-ChenY.pdf.

[209] Sohom Ghosh and Sudip Kumar Naskar. Fincat-2 an enhanced financial numeral claim analysis tool. https://www.codeocean.com/, 3 2022.

[210] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1803–1812, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098131. URL https://doi.org/10.1145/3097983.3098131.

[211] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. Numeracy-600K: Learning numeracy for detecting exaggerated information in market comments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6307–6313, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1635. URL https://aclanthology.org/P19-1635.

[212] Sohom Ghosh and Sudip Kumar Naskar. Fence: Financial exaggerated numeral classifier. https://www.codeocean.com/, 4 2023.

[213] Yu-Wen Liu, Liang-Chih Liu, Chuan-Ju Wang, and Ming-Feng Tsai. Fin10k: A web-based information system for financial report analysis and visualization. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 2441–2444, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340731. doi: 10.1145/2983323.2983328. URL https://doi.org/10.1145/2983323.2983328.

[214] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Financial opinion mining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 7–10, Punta Cana, Dominican Republic & Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-tutorials.2. URL https://aclanthology.org/2021.emnlp-tutorials.2.

[215] Yuexin Mao, Wei Wei, Bing Wang, and Benyuan Liu. Correlating s&p 500 stocks with twitter data. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, HotSocial '12, page 69–72, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450315494. doi: 10.1145/2392622.2392634. URL https://doi.org/10.1145/2392622.2392634.

[216] Timm O Sprenger, Andranik Tumasjan, Philipp G Sandner, and Isabell M Welpe. Tweets and trades: The information content of stock microblogs. *European Financial Management*, 20(5):926–957, 2014. URL https://doi.org/10.1111/j.1468-036X.2013.12007.x.

[217] Lian Fen Lee, Amy P Hutton, and Susan Shu. The role of social media in the capital market: Evidence from consumer product recalls. *Journal of Accounting Research*, 53(2):367–404, 2015. URL https://doi.org/10.1111/1475-679X.12074.

[218] Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. Sentiment analysis of twitter data for predicting stock market movements. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pages 1345–1350, 2016. doi: 10.1109/SCOPES.2016.7955659. URL https://doi.org/10.1109/SCOPES.2016.7955659.

[219] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 492–499, 2010. doi: 10.1109/WI-IAT.2010.63. URL https://doi.org/10.1109/WI-IAT.2010.63.

[220] W Brooke Elliott, Stephanie M Grant, and Frank D Hodge. Negative news and investor trust: The role of $ firm and # ceo twitter use. *Journal of Accounting Research*, 56(5):1483–1519, 2018. URL https://doi.org/10.1111/1475-679X.12217.

[221] Richard M Crowley, Wenli Huang, and Hai Lu. Executive tweets. *Available at SSRN 3975995*, 2021. URL https://dx.doi.org/10.2139/ssrn.3975995.

[222] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. Overview of the FinNLP-2022 ERAI task: Evaluating the rationales of amateur investors. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 99–103, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.finnlp-1.11. URL https://aclanthology.org/2022.finnlp-1.11.

[223] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1):5–es, may 2007. ISSN 1559-1131. doi: 10.1145/1232722.1232727. URL https://doi.org/10.1145/1232722.1232727.

[224] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011. URL https://doi.org/10.1016/j.jocs.2010.12.007.

[225] Ayodele Ariyo Adebiyi, Aderemi Oluyinka Adewumi, and Charles Korede Ayo. Comparison of arima and artificial neural networks models for stock price prediction. *Journal of Applied Mathematics*, 2014, 2014. URL https://doi.org/10.1155/2014/614342.

[226] Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, and Arun Kumar. Stock closing price prediction using machine learning techniques. *Procedia Computer Science*, 167:599–606, 2020. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2020.03.326. URL https://www.sciencedirect.com/science/article/pii/S1877050920307924.

[227] Eli Bartov, Lucile Faurel, and Partha S. Mohanram. Can twitter help predict firm-level earnings and stock returns? *The Accounting Review*, 93(3):25–57, 07 2017. ISSN 0001-4826. doi: 10.2308/accr-51865. URL https://doi.org/10.2308/accr-51865.

[228] Hailiang Chen, Byoung-Hyoun Hwang, and Baixiao Liu. The emergence of social executives and its consequences for financial markets. *SSRN Electronic Journal*, 2013. doi: 10.2139/ssrn.2318094. URL https://app.dimensions.ai/details/publication/pub.1102387257.

[229] Michael J. Jung, James P. Naughton, Ahmed Tahoun, and Clare Wang. Do Firms Strategically Disseminate? Evidence from Corporate Use of Social Media. *The Accounting Review*, 93(4):225–252, 09 2017. ISSN 0001-4826. doi: 10.2308/accr-51906. URL https://doi.org/10.2308/accr-51906.

[230] Mike Jermann. Predicting stock movement through executive tweets, 2017. URL https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2743946.pdf.

[231] Andrea Seaton Kelton and Robin R Pennington. Do tweets from ceos matter to investors? *LSE Business Review*, 2019. URL https://eprints.lse.ac.uk/104106/.

[232] Rushali Deshmukh et al. Stock prediction by using nlp and deep learning approach. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(1S): 202–211, 2021. URL https://doi.org/10.17762/turcomat.v12i1S.1611.

[233] C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225, May 2014. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

[234] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. *From opinion mining to financial argument mining*. Springer Nature, 2021. URL https://library.oapen.org/handle/20.500.12657/49533.

[235] Yumo Xu and Shay B. Cohen. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1183. URL https://aclanthology.org/P18-1183.

[236] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using

RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179.

[237] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL https://aclanthology.org/D14-1181.

[238] Pavlo Seroyizhko, Zhanel Zhexenova, Muhammad Zohaib Shafiq, Fabio Merizzi, Andrea Galassi, and Federico Ruggeri. A sentiment and emotion annotated dataset for bitcoin price forecasting based on Reddit posts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 203–210, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.finnlp-1.27. URL https://aclanthology.org/2022.finnlp-1.27.

[239] Ramit Sawhney, Shivam Agarwal, Vivek Mittal, Paolo Rosso, Vikram Nanda, and Sudheer Chava. Cryptocurrency bubble detection: A new stock market dataset, financial task & hyperbolic models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5531–5545, Seattle, United States, July 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.naacl-main.405.

[240] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985. URL https://apps.dtic.mil/sti/citations/ADA164453.

[241] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014. URL https://arxiv.org/abs/1412.3555.

[242] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014. URL https://arxiv.org/abs/1409.0473.

[243] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/glorot10a.html.

[244] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

[245] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. URL http://arxiv.org/abs/1412.6980.

[246] M. Schuster and K.K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997. doi: 10.1109/78.650093. URL https://doi.org/10.1109/78.650093.

[247] Chung-Chi Chen, Chin-Yi Lin, Chr-Jr Chiu, Hen-Hsen Huang, Alaa Alhamzeh, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. Overview of the ntcir-17 finarg-1 task: Fine-grained argument understanding in financial analysis. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan*, pages 16–20, Tokyo, Japan, 2023. NII. URL https://doi.org/10.20736/0002001323.

[248] Edward Ma. Nlp augmentation. https://github.com/makcedward/nlpaug, 2019.

[249] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020. URL https://doi.org/10.5281/zenodo.1212303.

[250] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[251] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

[252] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

[253] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

[254] Sohom Ghosh, Arnab Maji, N Harsha Vardhan, and Sudip Kumar Naskar. Experimenting with multi-modal information to predict success of indian ipos, 2024. URL https://arxiv.org/abs/2412.16174.

[255] Sohom Ghosh and Sudip Kumar Naskar. Predicting ratings of indian ipos from red herring prospectus, 2025. URL https://easychair.org/publications/preprint/G1P2/open.

[256] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, 16(2), mar 2016. ISSN 1533-5399. doi: 10.1145/2850417. URL https://doi.org/10.1145/2850417.

[257] Elena Cabrio and Serena Villata. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on*

Artificial Intelligence, IJCAI-18, pages 5427–5433. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/766. URL https://doi.org/10.24963/ijcai.2018/766.

[258] Robin Schaefer and Manfred Stede. Argument mining on twitter: A survey. *it - Information Technology*, 63(1):45–58, 2021. URL https://doi.org/10.1515/itit-2020-0053.

[259] John Lawrence and Chris Reed. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818, 2020.

[260] Huihui Xu and Kevin Ashley. Multi-granularity argument mining in legal texts. *Frontiers in Artificial Intelligence and Applications*, 362:261–266, 2022.

[261] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. A research agenda for financial opinion mining. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):1059–1063, May 2021. doi: 10.1609/icwsm.v15i1.18130. URL https://ojs.aaai.org/index.php/ICWSM/article/view/18130.

[262] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

[263] Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162, 02 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00452. URL https://doi.org/10.1162/tacl_a_00452.

[264] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213.

[265] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.62. URL https://aclanthology.org/2022.acl-long.62.

[266] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.365. URL https://aclanthology.org/2020.emnlp-main.365.

[267] Sagor Sarker. Banglabert: Bengali mask language model for bengali language understanding, 2020. URL https://github.com/sagorbrur/bangla-bert.

[268] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.445. URL https://aclanthology.org/2020.findings-emnlp.445.

[269] Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. IndicXNLI: Evaluating multilingual inference for Indian languages. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10994–11006, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.755. URL https://aclanthology.org/2022.emnlp-main.755.

[270] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.300. URL https://aclanthology.org/2021.emnlp-main.300.

[271] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Issues and perspectives from 10,000 annotated financial social media data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6106–6110, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.749.

[272] Nadhem Zmandar, Mahmoud El-Haj, Paul Rayson, Ahmed Abura'Ed, Marina Litvak, Geroge Giannakopoulos, and Nikiforos Pittaras. The financial narrative summarisation shared task FNS 2021. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 120–125, Lancaster, United Kingdom, 15-16 September 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.fnp-1.22.

[273] Mahmoud El-Haj, Nadhem Zmandar, Paul Rayson, Ahmed AbuRa'ed, Marina Litvak, Nikiforos Pittaras, George Giannakopoulos, Aris Kosmopoulos, Blanca Carbajo-Coronado, and Antonio Moreno-Sandoval. The financial narrative summarisation shared task (FNS 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 43–52, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.fnp-1.6.

[274] Ismail El Maarouf, Juyeon Kang, Abderrahim Ait Azzi, Sandra Bellato, Mei Gan, and Mahmoud El-Haj. The financial document structure extraction shared task (FinTOC2021). In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 111–119, Lancaster, United Kingdom, 15-16 September 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.fnp-1.21.

[275] Juyeon Kang, Abderrahim Ait Azzi, Sandra Bellato, Blanca Carbajo Coronado, Mahmoud El-Haj, Ismail El Maarouf, Mei Gan, Ana Gisbert, and Antonio Moreno Sandoval. The financial document structure extraction shared task (FinTOC 2022). In

Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022, pages 83–88, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.fnp-1.12.

[276] Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, Hen-Hsen Huang, and Hsin-Hsi Chen. Overview of the FinNLP-2023 ML-ESG task: Multi-lingual esg issue identification. In *Proceedings of the Joint Workshop of the 5th Financial Technology and Natural Language Processing (FinNLP) and 2nd Multimodal AI For Financial Forecasting (Muffin)*, 2023.

[277] Ramit Sawhney, Mihir Goyal, Prakhar Goel, Puneet Mathur, and Rajiv Ratn Shah. Multimodal multi-speaker merger & acquisition financial modeling: A new task, dataset, and neural baselines. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6751–6762, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.526. URL https://aclanthology.org/2021.acl-long.526.

[278] Ching-Yun Chang, Yue Zhang, Zhiyang Teng, Zahn Bozanic, and Bin Ke. Measuring the information content of financial news. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3216–3225, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1303.

[279] Merve Alanyali, Helen Susannah Moat, and Tobias Preis. Quantifying the relationship between financial news and the stock market. *Scientific reports*, 3(1):3578, 2013.

[280] Katherine Keith and Amanda Stent. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 493–503, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1047. URL https://aclanthology.org/P19-1047.

[281] Agam Shah, Suvan Paturi, and Sudheer Chava. Trillion dollar words: A new financial dataset, task & market analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6664–6679, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.368. URL https://aclanthology.org/2023.acl-long.368.

[282] Takehiro Takayanagi, Chung-Chi Chen, and Kiyoshi Izumi. Personalized dynamic recommender system for investors. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2246–2250, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3592035. URL https://doi.org/10.1145/3539618.3592035.

[283] Dominique Mariko, Hanna Abi-Akl, Kim Trottier, and Mahmoud El-Haj. The financial causality extraction shared task (FinCausal 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 105–107, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.fnp-1.16.

[284] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. Overview of the ntcir-14 finnum task: Fine-grained numeral understanding in financial social media data. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*, pages 19–27, 2019. URL https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings14/pdf/ntcir/01-NTCIR14-OV-FINNUM-ChenC.pdf.

[285] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. Overview of the ntcir-15 finnum-2 task: Numeral attachment in financial tweets. *Development*, 850(194):1–044, 2020.

[286] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models, 2023.

[287] Ujwal Narayan, Pulkit Parikh, Kamalakar Karlapalem, and Natraj Raman. Detecting regulation violations for an indian regulatory body through multi label classification. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 610–614, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391306. doi: 10.1145/3487553.3524640. URL https://doi.org/10.1145/3487553.3524640.

[288] Sathvik Sanjeev Buggana, Deepti Saravanan, Shravya Kanchi, Ujwal Narayan, Shivam Mangale, Lini T. Thomas, Kamalakar Karlapalem, and Natraj Raman. Sebi regulation biography. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 598–603, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391306. doi: 10.1145/3487553.3524638. URL https://doi.org/10.1145/3487553.3524638.

[289] Rohit Bansal, Ashu Khanna, et al. Determinants of ipos initial return: Extreme analysis of indian market. *Journal of financial risk management*, 1(04):68, 2012.

[290] Eyup Bastı, Cemil Kuzey, and Dursun Delen. Analyzing initial public offerings' short-term performance using decision trees and svms. *Decision Support Systems*, 73: 15–27, 2015. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2015.02.011. URL https://www.sciencedirect.com/science/article/pii/S0167923615000317.

[291] Emanuele Bajo and Carlo Raimondo. Media sentiment and ipo underpricing. *Journal of Corporate Finance*, 46:139–153, 2017. ISSN 0929-1199. doi: https://doi.org/10.1016/j.jcorpfin.2017.06.003. URL https://www.sciencedirect.com/science/article/pii/S092911991730370X.

[292] David Quintana, Francisco Chavez, Rafael M Luque Baena, and Francisco Luna. Fuzzy techniques for ipo underpricing prediction. *Journal of Intelligent & Fuzzy Systems*, 35(1):367–381, 2018.

[293] B Ramesh and Akshay Sakharkar. Revisiting underpricing of initial public offerings (ipo's)-evidences from indian stock markets. *International Journal of Innovative Research and Advanced Studies (IJIRAS)*, 2019.

[294] Akshay Sakharkar and B Ramesh. Pricing and performance evaluation of initial public offerings (ipo's): Evidence from indian stock markets. *International Journal of Research and Analytical Reviews*, 2019.

[295] Boubekeur Baba and Güven Sevil. Predicting ipo initial returns using random forest. *Borsa Istanbul Review*, 20(1):13–23, 2020. ISSN 2214-8450. doi: https://doi.org/10.1016/j.bir.2019.08.001. URL https://www.sciencedirect.com/science/article/pii/S2214845019302686.

[296] Di Wang, Xiaolin Qian, Chai Quek, Ah-Hwee Tan, Chunyan Miao, Xiaofeng Zhang, Geok See Ng, and You Zhou. An interpretable neural fuzzy inference system for predictions of underpricing in initial public offerings. *Neurocomputing*, 319:102–117, 2018. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2018.07.036. URL https://www.sciencedirect.com/science/article/pii/S0925231218308713.

[297] Tuan Hao Ly and Khanh Nguyen. Do words matter: Predicting ipo performance from prospectus sentiment. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 307–310, 2020. doi: 10.1109/ICSC.2020.00061.

[298] Jing Chi and Carol Padgett. Short-run underpricing and its characteristics in chinese initial public offering (ipo) markets. *Research in International Business and Finance*, 19(1):71–93, 2005. ISSN 0275-5319. doi: https://doi.org/10.1016/j.ribaf.2004.10.004. URL https://www.sciencedirect.com/science/article/pii/S0275531904000807.

[299] Leila Bateni and Farshid Asghari. Study of factors affecting the initial public offering (ipo) price of the shares on the tehran stock exchange. *Research in World Economy*, 5(2):68, 2014.

[300] Yunhee Kim and Almas Heshmati. Analysis of korean it startups' initial public offering and their post-ipo performance. *Journal of Productivity Analysis*, 34(2):133–149, 10 2010. ISSN 1573-0441. doi: 10.1007/s11123-010-0176-0. URL https://doi.org/10.1007/s11123-010-0176-0.

[301] Edward Sek Wong, Ricky Wong WB, and Lee Sue Ting. Initial public offering (ipo) underpricing in malaysian settings. *Journal of Economic & Financial Studies*, 5(02):14–25, 2017.

[302] Jyothi Seepani and KVR Murthy. Initial public offerings in india–a structural review. *European Journal of Economic and Financial Research*, 7(4), 2023.

[303] Ajay Yadav and Sweta Goel. Research on underpricing concept of ipo (initial public offering) in indian stock market. *International Journal of Innovative Technology and Exploring Engineering*, 8(11):179–183, 2019.

[304] B Ramesh and Pournima Dhume. Performance analysis of initial public offering in indian context. *Splint International Journal of Professionals*, 2(9):47–64, 2015.

[305] Ramit Anand and Balwinder Singh. Effect of composition of board and promoter group retained ownership on underpricing of indian ipo firms: An empirical study. *Indian Journal of Corporate Governance*, 12(1):21–38, 2019. doi: 10.1177/0974686219836539. URL https://doi.org/10.1177/0974686219836539.

[306] K.R. Naveen Kumar Iqbal Thonse Hawaldar and T. Mallikarjunappa. Pricing and performance of ipos: Evidence from indian stock market. *Cogent Economics & Finance*, 6(1):1420350, 2018. doi: 10.1080/23322039.2017.1420350. URL https://doi.org/10.1080/23322039.2017.1420350.

[307] KS Manu and Chhavi Saini. Valuation analysis of initial public offer (ipo): the case of india. *Paradigm*, 24(1):7–21, 2020.

[308] Seshadev Sahoo and Prabina Rajib. After market pricing performance of initial public offerings (ipos): Indian ipo market 2002–2006. *Vikalpa*, 35(4):27–44, 2010. doi: 10.1177/0256090920100403. URL https://doi.org/10.1177/0256090920100403.

[309] Shikha Bhatia and Balwinder Singh. Examining the performance of ipos: an evidence from india. *Management and Labour Studies*, 37(3):219–251, 2012.

[310] Minawati Dewi, Ali Sadikin, and Fahmi Roy Dalimunthe. The factors that influence the level of underpricing of shares in non-financial companies that conduct ipo (initial public offering) on the indonesian stock exchange. *Open Access Indonesia Journal of Social Sciences*, 7(1):1332–1338, 2024.

[311] Waqas Mehmood, Rasidah Mohd-Rashid, Norliza Che-Yahya, and Chui Zi Ong. Determinants of heterogeneity in investors' opinions on ipo valuation: evidence from the pakistan stock market. *Review of Behavioral Finance*, 13(5):631–646, 2021. doi: 10.1108/RBF-04-2020-0078. URL https://doi.org/10.1108/RBF-04-2020-0078.

[312] Saurabh Ghosh. Underpricing of initial public offerings: The indian experience. *Emerging Markets Finance and Trade*, 41(6):45–57, 2005.

[313] Madhuri Malhotra and Manjusha Nair. Initial public offerings underpricing: A study on the short run price performance of bookbuilt ipos in india. *Paripex - Indian Journal of Research*, pages 63–77, 02 2015. URL http://old.gsu.by/biglib/GSU/%D0%A4%D0%B8%D0%B7%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9/files.joomla/Files_GP/Pinchuk/Paripex%20Feb%202015%20book%2005.pdf#page=67.

[314] Vijaya B Marisetty and Marti G Subrahmanyam. Group affiliation and the performance of initial public offerings in the indian stock market, 2006.

[315] Harsimran Sandhu and Kousik Guhathakurta. Effects of ipo offer price ranges on initial subscription, initial turnover and ownership structure — evidence from indian ipo market. *Journal of Risk and Financial Management*, 13(11):279, 2020.

[316] NAMRATA N Khatri. Factors influencing investors investment in initial public offering. *International Journal of Management and Applied Science*, 3(7):41–49, 2017.

[317] Disha Mehta and Akash Patel. Price performance of initial public offerings (ipos): Evidence from indian capital market from 2007-2014. *Apeejay Journal of Management and Technology*, 2016.

[318] Manas Mayur and Sanjiv Mittal. Relationship between underpricing and post ipo performance: Evidence from indian ipos. *Asia-Pacific Journal of Management Research and Innovation*, 10(2):129–136, 2014.

[319] Mohammed Arshad Khan, Khudsiya Zeeshan, Faiz Ahmad, Abdullah A Alakkas, and Rashid Farooqi. A study of stock performance of select ipos in india. *Academy of Accounting and Financial Studies Journal*, 25(6):1–11, 2021.

[320] Kedar M Phadke and Manoj S Kamat. Impacts of macroeconomic and ipo factors on under-pricing of initial public offerings on the national stock exchange (nse) in india. *International Journal of Management Studies*, 5(1):4, 2018.

[321] Ehsan Nikbakht, Sayan Sarkar, Garrett C. Smith, and Andrew C. Spieler. Pre-ipo earnings management: Evidence from india. *Journal of International Accounting, Auditing and Taxation*, 44:100400, 2021. ISSN 1061-9518. doi: https://doi.org/10.1016/j.intaccaudtax.2021.100400. URL https://www.sciencedirect.com/science/article/pii/S1061951821000252.

[322] K Srinivasa Reddy. The aftermarket pricing performance of initial public offers: Insights from india. *International Journal of Commerce and Management*, 25(1): 84–107, 2015.

[323] SM Locke and Kartick Gupta. The return to initial public offerings: a sino-indian comparison. *Venture Capital*, 11(3):255–277, 2009.

[324] Jim Kyung-Soo Liew and Garrett Zhengyuan Wang. Twitter sentiment and ipo performance: A cross-sectional examination. *Journal of Portfolio Management*, 42 (4):129, 2016.

[325] Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. The finben: An holistic financial benchmark for large language models. *arXiv preprint arXiv:2402.12659*, 2024.

[326] Qianqian Xie, Dong Li, Mengxi Xiao, Zihao Jiang, Ruoyu Xiang, Xiao Zhang, Zhengyu Chen, Yueru He, Weiguang Han, Yuzhe Yang, et al. Open-finllms: Open multimodal large language models for financial applications. *arXiv preprint arXiv:2408.11878*, 2024.

[327] Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic embed: Training a reproducible long context text embedder, 2024.

[328] Xing Han Lù. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring, 2024. URL https://arxiv.org/abs/2407.03618.

[329] AI@Meta. Llama 3 model card, 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[330] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[331] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[332] Vishal Sarin and Neeru Sidana. A study of perceptions of investors towards ipo grading in india. *International Journal of Economic Research*, 14(20):757–770, 2017.

[333] Saikat Sovan Deb and Vijaya B Marisetty. Information content of ipo grading. *Journal of banking & Finance*, 34(9):2294–2305, 2010.

[334] Sanjay Poudyal. *Grading Initial Public Offerings (IPOs) in India's Capital Markets: A Globally Unique Concept*. Indian Institute of Management, 2008.

[335] Sanjay Dhamija and Ravinder Kumar Arora. Impact of quality certification on ipo underpricing: Evidence from india. *Global Business Review*, 18(2):428–444, 2017. doi: 10.1177/0972150916668611. URL https://doi.org/10.1177/0972150916668611.

[336] Joshy Jacob and Sobhesh Kumar Agarwalla. Mandatory ipo grading: does it help pricing efficiency? *Vikalpa*, 40(2):132–144, 2015.

[337] Zahid Younas Khan, Zhendong Niu, Sulis Sandiwarno, and Rukundo Prince. Deep learning techniques for rating prediction: a survey of the state-of-the-art. *Artificial Intelligence Review*, 54:95–135, 2021.

[338] Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. The bag-of-opinions method for review rating prediction from sparse text patterns. In Chu-Ren Huang and Dan Jurafsky, editors, *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 913–921, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL https://aclanthology.org/C10-1103/.

[339] Xiaojiang Lei, Xueming Qian, and Guoshuai Zhao. Rating prediction based on social sentiment from textual reviews. *IEEE transactions on multimedia*, 18(9):1910–1921, 2016.

[340] Natanael Fraga. Challenging llms beyond information retrieval: Reasoning degradation with long context windows. *Preprints*, August 2024. doi: 10.20944/preprints202408.1527.v1. URL https://doi.org/10.20944/preprints202408.1527.v1.

[341] Gemma Team, Morgane Riviere, Shreya Pathak, and et al. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

[342] Allen Huang, Hui Wang, and Yi Yang. Finbert—a large language model approach to extracting information from financial text, July 2020. URL http://dx.doi.org/10.2139/ssrn.3910214.

[343] Ming-Feng Tsai and Chuan-Ju Wang. Visualization on financial terms via risk ranking from financial reports. In *Proceedings of COLING 2012: Demonstration Papers*, pages 447–452, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL https://aclanthology.org/C12-3056.

[344] Abejide Ade-Ibijola. Finchan: A grammar-based tool for automatic comprehension of financial instant messages. In *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists*, SAICSIT '16, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450348058. doi: 10.1145/2987491.2987518. URL https://doi.org/10.1145/2987491.2987518.

[345] Boris Galitsky and Dmitry Ilvovsky. On a chatbot conducting a virtual dialogue in financial domain. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 99–101, Macao, China, aug 2019. URL https://aclanthology.org/W19-5517.

[346] Karolin Winter, Manuel Gall, and Stefanie Rinderle-Ma. Regminer: Taming the complexity of regulatory documents for digitalized compliance management. In *Proceedings of the Best Dissertation Award, Doctoral Consortium, and Demonstration & Resources Track at BPM 2020*, pages 112–116, 2020. URL http://ceur-ws.org/Vol-2673/paperDR10.pdf.

[347] Jean Lee, Hoyoul Luis Youn, Nicholas Stevens, Josiah Poon, and Soyeon Caren Han. Fednlp: An interpretable nlp system to decode federal reserve communications. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2560–2564, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462785. URL https://doi.org/10.1145/3404835.3462785.

[348] Maarten Grootendorst. Keybert: Minimal keyword extraction with bert., 2020. URL https://doi.org/10.5281/zenodo.4461265.

[349] Tatiana Passali, Alexios Gidiotis, Efstathios Chatzikyriakidis, and Grigorios Tsoumakas. Towards human-centered summarization: A case study on financial news. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 21–27, Online, April 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.hcinlp-1.4.

[350] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/zhang20ae.html.

[351] Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. Financialphrasebank-v1.0, 07 2013. URL https://www.researchgate.net/publication/251231364_FinancialPhraseBank-v10.

[352] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Sudip Kumar Naskar

Soham Ghosh