Harnessing Text and Image based Multimodal Techniques for Enhanced Fake News Identification

Thesis Submitted by

Avik Mondal

Regn. No. - 160097 of 2021-2022
Exam Roll No. - M6TCT24021
Class Roll No. - 002110504026
Session - 2021 - 2024

In Partial Fulfillment of the Requirements
For the Degree of

Master of Technology
in
Computer Technology

Under the Supervision of

Dr. Dipankar Das

Department of Computer Science and Engineering

Jadavpur University

188, Raja S.C. Mallick Rd,

Kolkata - 700032, West Bengal, India

Department of Computer Science and Engineering

Faculty of Engineering and Technology

Jadavpur University, Kolkata - 700032

CERTIFICATE OF RECOMMENDATION

This is to certify that the work embodied in this thesis entitled **"Harnessing Text and Image based Multimodal Techniques for Enhanced Fake News Identification"** has been satisfactorily completed by Avik Mondal (Registration Number 160097 of 2021 - 2022; Class Roll No. 002110504026; Examination Roll No. M6TCT24021). It is a bonafide piece of work carried out under my supervision and guidance at Jadavpur University, Kolkata, for partial fulfillment of the requirements for the awarding of the **Master of Technology in Computer Technology** degree of the Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, during the academic year 2021 - 2024.

Dr. Dipankar Das
Department of Computer Science and Engineering,
Jadavpur University.
(Supervisor)

Forwarded By:

Dr. Nirmarlya Chowdhury,
HOD, Department of Computer Science and Engineering,
Jadavpur University.

Prof. Dipak Laha,
Dean, Faculty of Engineering and Technology,
Jadavpur University.

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University, Kolkata - 700032

CERTIFICATE OF APPROVAL

This is to certify that the thesis entitled **"Harnessing Text and Image based Multimodal Techniques for Enhanced Fake News Identification"** is a bonafide record of work carried out by **Avik Mondal** (Registration Number 160097 of 2021 - 2022; Class Roll No. 002110504026; Examination Roll No. M6TCT24021) in partial fulfillment of the requirements for the award of the degree of **Master of Technology in Computer Technology** in the Department of Computer Science and Engineering, during the period of September 2021 to August 2024. It is understood that by this approval, the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the thesis only for the purpose of which it has been submitted.

Signature of Examiner 1	Signature of Examiner 2
Date:	Date:

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University, Kolkata - 700032

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that the thesis entitled **"Harnessing Text and Image based Multimodal Techniques for Enhanced Fake News Identification"** contains literature survey and original research work by the undersigned candidate, as a part of his degree of **Master of Technology in Computer Technology, Jadavpur University**. All the information have been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name: Avik Mondal

Examination Roll No.: M6TCT24021

Registration No.: 160097 of 2021 - 2022

Thesis Title: Harnessing Text and Image based Multimodal Techniques for Enhanced Fake News Identification

---

Signature of the Candidate

. . . . .

Dedicated to my unwavering pillars of support, my parents,
whose love and sacrifice have fueled every chapter of my journey.

. . . . .

ACKNOWLEDGEMENTS

I am delighted to express my gratitude and regards towards my project guide **Dr. Dipankar Das**, Asst. Professor, Department of Computer Science and Engineering, Jadavpur University, without whose valuable guidance, inspiration, and attention pursuing my project would have been impossible. His expertise and guidance were crucial in shaping both the direction and the quality of this thesis, and I am profoundly thankful for his steadfast dedication.

I express my sincere gratitude to the Head of the Department, **Dr. Nirmalya Chowdhury**, Professor, Department of Computer Science and Engineering, Jadavpur University, Kolkata, India, for being kind enough to permit me to carry out my research in this prestigious institution.

I am grateful to all the members of **NLP Lab and Magnum Lab**, Department of Computer Science and Engineering, Jadavpur University, for cooperating with me in all possible ways and providing me with a good working environment throughout the work.

I would like to thank my friends Mr. Aritra Das, Mr. Sourav Sarkar and Mr. Suprabho Roy, my juniors Mr. Pritam Pal, and Mr. Shankha Shubhra Das, my classmate Ms. Shreyashi Karjee and my seniors Mr. Sandip Sarkar, Mr. Dipanjan Saha, Mr. Atanu Mondal for constantly supporting me and helping me in any way possible.

Their companionship and camaraderie have been invaluable to me. I express my regards to my parents, my Elder Brother and Sister for bearing with me, helping me, and being a source of constant motivation during the entire term of the work.

I am so grateful to my family Sri Aniruddha Mondal, Smt. Ava Mondal, Arijit Mondal and Pallabi Bhaduri.

Avik Mondal

Examination Roll No.: M6TCT24021

Registration No.: 160097 of 2021 - 2022

Department of Computer Science and Engineering,

Jadavpur University.

ABSTRACT

In the digital age, the proliferation of fake news poses a significant threat to the integrity of information, influencing public opinion and decision-making processes.The study is grounded in a comprehensive dataset, comprising textual and visual content related to the Ukraine-Russia conflict, which has been meticulously labeled for authenticity. The research employs sophisticated preprocessing techniques to standardize and augment the data, ensuring robustness and generalizability across diverse inputs. Through a series of experiments, the thesis demonstrates that multimodal models outperform traditional single-modality approaches, effectively capturing the nuanced interdependencies between text and images that are often exploited in the dissemination of fake news.

This thesis explores the integration of Natural Language Processing (NLP) and computer vision techniques to enhance the detection of fake news through a multimodal approach that leverages both textual and visual data. By combining Long Short-Term Memory (LSTM) networks, Bidirectional Encoder Representations from Transformers (BERT), XGBoost and advanced image classifiers like ResNet-50, ResNet-101, Logistic Regression etc.This research aims to improve the accuracy of fake news identification across various languages and cultural contexts.

The findings reveal that the combined use of text and image data significantly enhances the detection accuracy, particularly in complex scenarios where either modality alone might fail. Moreover, The contributions of this research extend beyond the specific case of fake news detection, offering insights into the broader application of multimodal techniques in various domains of machine learning and artificial intelligence.

Contents

Acknowledgements **6**

Abstract **7**

List of Figures **12**

List of Tables **14**

List of Abbreviations **17**

1 INTRODUCTION **20**

1.1 Background and Motivation 21
1.2 Problem Statement 21
1.3 Objectives 26
1.4 Scope of the Thesis 27
1.5 Contributions 28

2 LITERATURE REVIEW **31**

2.1 Fake News Detection: An Overview 31
2.2 Text Classification Techniques 32
2.2.1 Traditional Methods 32
2.2.2 Deep Learning Approaches 33
2.3 Image Classification Techniques 34
2.3.1 Traditional Methods 34
2.3.2 Deep Learning Approaches 35
2.4 Multimodal Classification 36
2.5 Summary of Literature Gaps 37

3 DATASET DESCRIPTION **38**

3.1 Overview of the Dataset 38
3.2 Labeling Strategy 40
3.3 Statistical Analysis of the Dataset 41
3.4 Overview of Comparative Dataset 43
3.5 Data Collection and Preprocessing 45
3.5.1 Text Data 45

3.5.2 Image Data 46

4 TEXT CLASSIFICATION 48

4.1 Text Preprocessing 48
4.1.1 Tokenization and Normalization 49
4.1.2 Handling Imbalanced Data 49
4.2 Model Architectures 50
4.2.1 Bi-LSTM Model 50
4.2.2 XGBoost Model 51
4.2.3 Without Translated Text 52
4.2.4 With Translated Text 56
4.3 Training and Evaluation 60
4.3.1 Training Process 60
4.3.2 Evaluation Metrics 61

5 IMAGE CLASSIFICATION 63

5.1 Image Preprocessing 63
5.1.1 Image Augmentation Techniques 63
5.1.2 Resizing and Normalization 64
5.2 Module 1: Raw Image Classification 65
5.2.1 LSTM Model 65
5.2.2 ResNet-50 Model 66
5.2.3 MobileNetV2 67
5.3 Module 2: Generated Captions-Based Image Classification 68
5.3.1 LSTM Model 68
5.3.2 BERT Model 69
5.3.3 XGBoost Model 70
5.3.4 Logistic Regression Model 71
5.4 Training and Evaluation 72
5.4.1 Training Process 72
5.4.2 Evaluation Metrics 73

6 MULTIMODAL CLASSIFICATION 76

6.1 Combining Text and Image Data 76
6.2 Model Architectures 76
6.2.1 LSTM 77
6.2.2 ResNet-50 77
6.2.3 BERT 77
6.2.4 Logistic Regression 77
6.2.5 ResNet + LSTM 78

6.2.6 ResNet + Bidirectional LSTM 78
6.3 Experiments 78
6.3.1 Raw Image + Raw Text 78
6.3.2 Raw Image + Translated Text 80
6.3.3 Image Caption + Raw Text 82
6.3.4 Image Caption + Translated Text 84
6.3.5 Raw Image + Generated Captions Classification 86
6.4 Training and Evaluation 88
6.4.1 Training Process 88
6.4.2 Evaluation Metrics 88

7 COMPARATIVE ANALYSIS 89
7.1 Comparison of Text, Image, and Multimodal Approaches 89
7.2 Contributions of the Thesis 93
7.2.1 Relevance to Thesis Work 94
7.3 Strengths and Weaknesses of Each Approach 103
7.4 Impact of Data Quality and Quantity on Performance 106
7.5 Results and Discussions on Text Classification 110
7.6 Results and Discussions on Image Classification 112
7.7 Results and Discussion on Multimodal Classification 113
7.8 Error Analysis 116
7.8.1 Text Classification Error Analysis 116
7.8.2 Image Classification Error Analysis 117
7.8.3 Multimodal Classification Error Analysis 118

8 CONCLUSION AND FUTURE WORK 120
8.1 Summary of Findings 120
8.2 Limitations 121
8.3 Suggestions for Future Research 122

References 124

LIST OF FIGURES

1.1 Dataset snippet used for text classification. 23
1.2 Image Classification Dataset 24
1.3 Example Image 24
1.4 Dataset snippet used for text classification. 25

2.1 Fake News Detection Workflow 34

3.1 Overview of the Dataset 39
3.2 Distribution of Labels 42
3.3 Histogram & Box Plot of Text Lengths 42
3.4 Histograms of Image Widths and Heights 43
3.5 Overview of the Fakeddit Text Dataset 45
3.6 Overview of the Fakeddit Image Dataset 46
3.7 Example Image 46

4.1 Bi-LSTM Architecture 51
4.2 XGBoost Model Architecture for Text Classification 52
4.3 Confusion Matrix of Bi-LSTM on Non-translated Texts 54
4.4 Confusion Matrix of XGBoost on Non-translated Texts 56
4.5 Confusion Matrix of Bi-LSTM on Translated Texts 58
4.6 Confusion Matrix of XGBoost on Translated Texts 59

5.1 Confusion Matrix of LSTM on Raw Image Classification Task . . 66
5.2 Confusion Matrix of ResNet-50 on Raw Image Classification Task 67
5.3 Confusion Matrix of MobileNetv2 on Raw Image Classification Task 68
5.4 Confusion Matrix of LSTM on Generated Captions-based Image Classification Task 69
5.5 Confusion Matrix of BERT on Generated Captions-based Image Classification Task 70
5.6 Confusion Matrix of XGBoost on Generated Captions-based Image Classification Task 71
5.7 Confusion Matrix of Logistic Regression on Captions-based Image Classification Task 72

6.1 Confusion Matrix of BERT performing on Raw Image and Raw Text Data . . . 79
6.2 Confusion Matrix of ResNet with LSTM performing on Raw Image and Raw Text Data . . . 80
6.3 Confusion Matrix of BERT performing on Raw Image and Translated Text Data . . . 81
6.4 Confusion Matrix of ResNet and BiLSTM performing on Raw Image and Translated Text Data . . . 82
6.5 Confusion Matrix of BERT performing on Image Captions and Raw Text data . . . 83
6.6 Confusion Matrix of Logistic Regression performing on Image Captions and Raw Text Data . . . 84
6.7 Confusion Matrix of BERT performing on Image Captions and Translated Texts Data . . . 85
6.8 Confusion Matrix of Logistic Regression performing on Image Captions and Translated Texts Data . . . 86
6.9 Confusion Matrix of ResNet-50 performing on Raw Image and Generated Captions Data . . . 87

7.1 Confusion Matrix of MULTI-Fake-DetctIVE dataset Text Classification . . . 96
7.2 Confusion Matrix of Fakeddit dataset Text Classification . . . 96
7.3 Confusion Matrix of MULTI-Fake-DetctIVE dataset Text Classification . . . 99
7.4 Confusion Matrix of Fakeddit dataset Text Classification . . . 99
7.5 Confusion Matrix of MULTI-Fake-DetctIVE dataset Text Classification . . . 102
7.6 Confusion Matrix of Fakeddit dataset Text Classification . . . 102

LIST OF TABLES

4.1 Performance of Bi-LSTM on the Non-translated Text 54
4.2 Performance of XGBoost on the Non-translated Text 55
4.3 Performance of Bi-LSTM on the Translated Text 57
4.4 Performance of XGBoost on the Translated Text 59

5.1 Performance of LSTM Raw Image Classification Task 65
5.2 Performance of ResNet-50 on Raw Image Classification Task . . . 66
5.3 Performance of MobileNetv2 on Raw Image Classification Task . . 67
5.4 Performance of LSTM on Generated Captions-Based Image Classification Task 69
5.5 Performance of BERT on Generated Captions-Based Image Classification Task 70
5.6 Performance of XGBoost on Generated Captions-Based Image Classification Task 71
5.7 Performance of Logistic Regression on Generated Captions-Based Image Classification Task 72

6.1 Performance of BERT on Raw Image and Raw Text Data 79
6.2 Performance of ResNet and LSTM on Raw Image and Raw Text Data 80
6.3 Performance of BERT on Raw Image and Translated Text Data . 81
6.4 Performance of ResNet and Bi-LSTM on Raw Image and Translated Text Data 82
6.5 Performance of BERT on Image Captions and Raw Text data . . 83
6.6 Performance of Logistic Regression on Image Captions and Raw Text data 84
6.7 Performance of BERT on Image Captions and Translated Texts . 85
6.8 Performance of Logistic Regression on Image Captions and Translated Texts 86
6.9 Performance of ResNet-50 on Raw Image and Generated Captions Data 87

7.1 Performance Comparison of Bi-LSTM on Text Classification . . . 95
7.2 Performance Comparison of MobileNetV2 on Image Classification 98

7.3 Performance comparison of ResNet-50 and Bi-LSTM on Multimodal Classification 101
7.4 Comparison of F1-Scores and Accuracy. Here, ↑ denotes the best performing model. 111
7.5 Comparison of F1-Scores and Accuracy. Here, ↑ denotes the best performing model. 113
7.6 Comparison of F1-Scores and Accuracy. Here ↑ denotes the best performing model. 114
7.7 Misclassification Results for Text Classification: Actual vs Predicted Labels 116
7.8 Image Classification Results: Actual vs Predicted Labels 117
7.9 Multimodal Classification Results: Actual vs Predicted Labels . . 118

List of Abbreviations

A

AI	Artificial Intelligence
AUC	Area Under the Curve
AUC-ROC	Area Under the Receiver Operating Characteristic Curve

B

BERT	Bidirectional Encoder Representations from Transformers
Bi-LSTM	Bidirectional Long Short-Term Memory

C

CNNs	Convolutional Neural Networks
CF	Certainly Fake
CR	Certainly Real

D

DL	Deep Learning

F

FN	False Negative
FP	False Positive
FPR	False Positive Rate

L

LSTM	Long Short-Term Memory

M

ML	Machine Learning

N

NLP	Natural Language Processing
NB	Naive Bayes
PF	Probably Fake

PL	Probably Real
R	
RNN	Recurrent Neural Network
ResNet	Residual Networks
ROC	Receiver-operating characteristic curve
S	
SVM	Support Vector Machines
SMOTE	Synthetic Minority Over-sampling Technique
T	
TN	True Negative
TN	True Positive
TPR	True Positive Rate
TF-IDF	Term Frequency and Inverse Document Frequency
X	
XGBoost	Extreme Gradient Boosting

Chapter 1

INTRODUCTION

– Integrating Visual and Textual Insights to Combat the Spread of Misinformation

The prevalence of fake news in the digital era has posed significant challenges to information integrity and public trust. The rapid spread of misinformation across social media platforms and other online channels has made it increasingly difficult for individuals to discern truth from falsehood. This growing issue has necessitated the development of advanced methods for detecting and mitigating the effects of fake news.

This thesis explores the intersection of Natural Language Processing (NLP)[1] and deep learning to address the problem of fake news detection. By leveraging both textual and visual data, the research aims to create a comprehensive framework that enhances the accuracy and reliability of fake news detection systems. The focus is on developing models that can effectively classify news content as Certainly Fake, Probably Fake, Probably Real, or Certainly Real.

In addition to examining the capabilities of single-modality approaches, this research delves into multimodal classification, where text and image data are integrated to improve detection performance. The inclusion of Multimodal classification further enriches the study by allowing the simultaneous execution of related tasks.

Moreover, the research extends beyond the English language, incorporating Italian language which broadens the applicability of the developed models.

[1] https://nlp.stanford.edu/

1.1 Background and Motivation

The rise of the internet and social media has revolutionised the way information is disseminated and consumed. However, this transformation has also led to the rapid spread of misinformation, or *fake news*, which poses significant threats to society, politics, and individual decision-making. The ability to quickly and accurately detect fake news has therefore become a critical task in maintaining the integrity of information systems and protecting the public from deceptive content.

Natural Language Processing (NLP) and deep learning have emerged as powerful tools in addressing the challenge of fake news detection. NLP, in particular, has seen tremendous growth with the advent of machine learning and deep learning methodologies, enabling more sophisticated analysis of textual data. This thesis leverages these advancements to develop robust models capable of identifying and classifying fake news with high accuracy.

The motivation behind this research is two-fold. First, there is a growing need for automated systems that can analyze and validate claims made within textual data, as manual verification is both time-consuming and prone to error. Second, the research aims to extend the applicability of these techniques beyond the English language, considering the diverse linguistic landscape of India. By incorporating multiple languages, including those prevalent in India, this research seeks to broaden the scope and effectiveness of NLP in detecting fake news across different cultural and linguistic contexts.

1.2 Problem Statement

The spread of fake news has become a pervasive problem, exacerbated by the ease with which information can be shared across digital platforms. Traditional methods of fake news detection, which primarily rely on textual analysis, are increasingly insufficient as fake news becomes more sophisticated, often incorporating multimedia elements such as images and videos to mislead readers. This

complexity necessitates a more robust approach that not only considers the textual content but also the visual elements associated with it.

Current research in fake news detection has made significant strides in text classification, particularly with the advent of deep learning models like Long Short-Term Memory (LSTM)[2] networks, Bidirectional Encoder Representations from Transformers (BERT)[3], and Logistic Regression[4]. Similarly, advanced machine learning models such as XGBoost (Extreme Gradient Boosting)[5] have also been effective in handling large and complex datasets. However, these models typically focus on text alone, ignoring the visual information that can play a critical role in determining the authenticity of a news item. On the other hand, image classification models like ResNet-50 (Residual Networks)[6] and MobileNetV2[7] have shown promise in detecting manipulated or misleading images, but they often lack the contextual understanding that textual analysis provides.

The core problem addressed in this thesis is the development of an effective multimodal fake news detection system that integrates both textual and visual data. The goal is to create a comprehensive model that can accurately classify news items as Certainly Fake, Probably Fake, Probably Real, or Certainly Real by analyzing the combined information from both text and images. Additionally, this research explores the application of Multimodal Classification[8] to further enhance the model's performance by leveraging the interrelated tasks of Fake News Detection using Text and Images.

[2]https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html
[3]https://huggingface.co/docs/transformers/en/model_doc/bert
[4]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
[5]https://xgboost.readthedocs.io/en/stable/python/sklearn_estimator.html
[6]https://pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html
[7]https://pytorch.org/hub/pytorch_vision_mobilenet_v2/
[8]https://en.wikipedia.org/wiki/Multimodal_learning

Example of Text Classification Process:

Consider an example for clarification, Let's take row 2,

	A	B
1	Text	Label
2	si sono rivolti direttamente a Putin "Contro di noi non puoi vincere" e ai soldati sul fronte con un appello a deporre le armi. Oggi si può combattere una guerra senza	1
3	Ma veramente qualcuno sano di mente può pensare che inviare soldati, armi e quant'altro di bellico possa servire a fermare una guerra? □ https://t.co/BmShgDRm31	2
4	#flowers #lovers Fate l'amore non fate la guerra. Marc Chagall, "Il mazzo di fiori degli amanti", 1926. #art #painting #marcchagall #NoWars https://t.co/XahcLle4SK	2
5	combattere le guerre con bombe e armi. Per quanto è possibile passate una buona domenica, tutti i nostri pensieri sono rivolti al popolo Ucraino. □❤□ #PACE #NoWar	2
6	uniformi ucraine o viceversa? L'IA può fare chiarezza identificando le persone in una necessaria azione di #debunking e fact checking.	2
7	successivi silenzi, il mazzo di fiori in ambasciata, la serata Pizza-maracaibo con la guerra a Kiev, ma c'è riuscito.	1
8	capitale, arrivano le armi". | Diretta, video e foto #ucraina #russia #guerra #kiev #morti #feriti #ucraina-russia #putin #RussiaUkraineConflict https://t.co/7Ayrz7Zyov	3
9	https://t.co/x19XkMWbGz	1
10	armi a chiunque faccia richiesta https://t.co/oV4spuY0o6 https://t.co/WgOjeGKYSw	1
11	affermato falsamente che questo video mostrava la guerra tra Russia e Ucraina, ma è un video del videogioco "Arma 3" https://t.co/MLnc7h3epQ sempre canale Rossella	2
12	All'inizio di questa settimana, migliaia di utenti dei social media hanno affermato falsamente che questo video mostrava la guerra tra Russia e Ucraina, ma è un video del videogioco "Arma 3" https://t.co/nchRA0qotO https://t.co/e5WliVN0RD	2

Figure 1.1: Dataset snippet used for text classification.

Text: *si sono rivolti direttamente a Putin Contro di noi non puoi vincere e ai soldati sul fronte con un appello a deporre le armi. Oggi si può combattere una guerra senza* which translates in English is *After hacked the state TV by replacing propaganda images with the real ones, they turned directly to Putin against us you cannot win and to the soldiers on the front with an appeal to lay down the weapons.Today you can fight a war without shooting a blow.*

Interpretation of Results

The label 1 ("Probably Fake") suggests that the text may have elements of truth but is overall misleading. The language used may be sensational or manipulative, designed to influence the reader's perception negatively. The model recognizes these patterns and classifies the text accordingly.

In this way, the text classification model processes each entry in the dataset, analyzing the content and assigning a label that reflects the likely authenticity of the information. This approach helps in automatically identifying potentially fake news and preventing the spread of misinformation.

Example of image Classification Process:

	Media	Label
2	['14980224383988777704_.jpg']	1
3	['1498003150414008322_.jpg']	2
4	['1497887473862467584_.jpg']	2
5	['1497838543619645441_1.jpg', '1497838543619645441_2.jpg']	2
6	['1497639979534667783_.jpg']	2
7	['1497624763396546561_.jpg']	1
8	['1497614216827711495_.jpg']	3
9	['1497599806201647105_.jpg']	1
10	['1497598337570074626_.jpg']	1
11	['1497555006286872584_.jpg']	2
12	['1497521878319042560_.jpg']	2
13	['1497493024695152645_.jpg']	3
14	['1497491399620636678_.jpg']	2
15	['1497469243062771717_.jpg']	3
16	['1497225578528784396_.jpg']	0
17	['1496958201270616073_.jpg']	3
18	['1496883123354804225_.jpg']	1
19	['1496839348016201729_.jpg']	1
20	['1496810148177670148_.jpg']	2

Figure 1.2: Image Classification Dataset



Figure 1.3: Example Image

Consider the above image with the image named ***14980224383988777704_.jpg***, which is associated with the caption *a person in a hoodie is shown in the dark*. This image was uploaded by a user, and based on the analysis, it has been labeled as 1 (*Probably Fake*).

In the image classification process, the model evaluates the visual content of the image to determine whether it is likely to be fake or real. The image depicts a person in a hoodie, often associated with anonymity or mystery, which could be used in a misleading context. The model, trained on a dataset of similar images, recognizes these visual cues and determines that the image could be part of a deceptive post. Therefore, it classifies the image as *Probably Fake*, matching the

label of 1 provided in the dataset.

This example shows how the model uses the visual information in the image to make a classification decision, determining that the image is likely fake but not with absolute certainty.

Example of Multimodal Classification Process: Consider a social media post that includes both text and an image:

	A	B	C
1	**Text**	Media	Label
2	Dopo aver hackerato la tv di Stato sostituendo immagini di propaganda con q	['1498022438398877704_.jpg']	1
3	@sarabanda_ @lauraluciacarme Ma i #facciamoride nn erano quelli restiamc	['1498003150414008322_.jpg']	2
4	#flowers #loversFate l'amore non fate la guerra.Marc Chagall, "Il mazzo di fic	['1497887473862467584_.jpg']	2
5	Se solo tutti mostrassimo più amore e comprensione forse non avremmo bisc	['1497838543619645441_1.jpg', '149	2
6	Chi sono i soldati che vediamo nei video? Sono manipolati o veri? Russi con le	['1497639979534667783_.jpg']	2
7	Non credevo che #Salvini potesse peggiorare la propria posizione, l'amore pe	['1497624763396546561_.jpg']	1
8	Guerra in Ucraina: battaglia nelle strade di Kiev. L'appello di Zelensky: "Difend	['1497614216827711495_.jpg']	3
9	Guerra Ucraina-Russia, armi ai civili a Kiev - Video https://t.co/YX8XzeHEzI ht	['1497599806201647105_.jpg']	1
10	Guerra Ucraina-Russia, armi ai civili a Kiev - Video: (Adnkronos) - Distribuzione	['1497598337570074626_.jpg']	1
11	Quarta. All'inizio di questa settimana, migliaia di utenti dei social media hanno	['1497555006286872584_.jpg']	2
12	FAKE NEWS ITALIANE All'inizio di questa settimana, migliaia di utenti dei socia	['1497521878319042560_.jpg']	2
13	#Cronaca #fakenews Ucraina, Francesco Pira (UNIME): In atto due guerre, un	['1497493024695152645_.jpg']	3
14	LUcraina vista dal web: video, foto e messaggi in tempo di guerra https://t.co	['1497491399620636678_.jpg']	2
15	Guerra Ucraina-Russia, Zelensky: "Mai ordinato resa" - Video: (Adnkronos) - N	['1497469243062771717_.jpg']	3
16	#notizie #siciliaGuerra in Ucraina, donna sfida soldato russo, "sarai presto un	['1497225578528784396_.jpg']	0
17	LUMANITÀ: mani tagliate, vestiti sgualciti, le bende le armi, bambini impauriti,	['1496958201270616073_.jpg']	3
18	Ucraina, il video straziante del soldato prima dell'arrivo dei razzi russi fa il giro	['1496883123354804225_.jpg']	1
19	#DalMondo #bombardamenti Guerra Ucraina, catturati soldati Russia: foto p	['1496839348016201729_.jpg']	1
20	La guerra live sul web: missile russo si abbatte tra le urla di paura su un aerop	['1496810148177670148_.jpg']	2
21	#DalMondo #bombardamento Guerra Russia Ucraina, soldato dopo bombard	['1496781589338066945_.jpg']	2
22	Gli appelli russi ai soldati ucraini: "L'Ucraina ha combattuto per 8 anni contro	['1496772412897067014_.jpg']	1
23	La #Russia ha attaccato l'#Ucraina su larga scala.Da #Odessa a #Kiev il rumor	['1496749412697227264_.jpg']	2
24	non in guerra, poi in Uk,e poi tiepidamente anche altrove, relegata in qualche	['1509159553874731009_1.jpg', '150	1

Figure 1.4: Dataset snippet used for text classification.

Text: *si sono rivolti direttamente a Putin 'Contro di noi non puoi vincere' e ai soldati sul fronte con un appello a deporre le armi. Oggi si può combattere una guerra senza.*

Image: An image with the filename ***1498022438398877704.jpg***, showing *a person in a hoodie is shown in the dark.*

Label: 1 (Probably Fake)

In this multimodal classification example, both the text and the image are analyzed together to determine the overall authenticity of the post. The text makes a provocative statement related to political and military actions, which might be misleading. The accompanying image of a person in a hoodie, typically

associated with anonymity or suspicious activities, further adds to the potentially deceptive nature of the post.

The model, which considers both the textual and visual content, classifies this post as *Probably Fake* (Label 1). The combined context of the misleading text and the suggestive image strengthens the model's decision to categorize the post as likely to be fake.

1.3 Objectives

The primary objectives of this thesis are centered around the development and evaluation of multimodal fake news detection models that integrate both text and image data. The specific objectives include:

1. **Development of a Text Classification**: To develop and evaluate deep learning models, specifically Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT), for the task of fake news detection based on textual data. This includes fine-tuning these models to optimize their performance in distinguishing between different levels of news authenticity (Certainly Fake, Probably Fake, Probably Real, Certainly Real).

2. **Development of a Image Classification**: To design and implement image classification models, particularly using the ResNet-50 architecture, MobileNetV2 to detect and classify fake news based on visual content. This objective involves exploring how visual cues can contribute to the detection of fake news and how these cues interact with textual information.

3. **Development of a Multimodal Classification**: To create a multimodal classification model that effectively integrates text and image data, utilizing data fusion techniques to enhance the detection accuracy. The aim is to assess how the combination of textual and visual data improves the overall classification performance compared to single-modality approaches.

4. **Comparative Analysis and Evaluation**: To conduct a thorough comparative analysis of the developed models, evaluating their performance in different scenarios, including text-only, image-only, and multimodal approaches. The objective is to identify the strengths and limitations of each approach and provide insights into the most effective strategies for fake news detection.

1.4 Scope of the Thesis

The scope of this thesis encompasses the design, development, and evaluation of deep learning models for the detection of fake news across multiple modalities—text and images. The research is structured around three primary domains: text classification, image classification, and multimodal classification, each contributing to the overarching goal of creating a robust fake news detection system.

The thesis explores the application of LSTM and BERT models for text classification. The focus is on understanding how these models can be optimized to detect fake news by analyzing linguistic patterns and content features. The research also addresses the challenges associated with handling large and diverse textual datasets, including those in multiple languages, with a special emphasis on Indian languages.

In the visual domain, the thesis investigates the use of ResNet-50, a convolutional neural network (CNN)[9] architecture known for its effectiveness in image classification tasks. The scope includes the pre-processing of image data, the development of image classification models, and the evaluation of these models in detecting manipulated or misleading images associated with fake news.

The thesis extends the scope to include multimodal classification, where text and image data are combined to improve the accuracy of fake news detection. This involves the integration of data fusion techniques, where the textual and visual features are combined at different stages of the model to enhance the de-

[9]https://pytorch.org/docs/stable/generated/torch.nn.Conv2d.html

tection capabilities. The research also explores the challenges of aligning text and image data and ensuring that the model effectively leverages the complementary information provided by both modalities.

The thesis includes a comprehensive evaluation of the developed models, comparing their performance across different modalities and tasks. The scope includes the use of various evaluation metrics, such as accuracy, precision, recall, and F1-score, to assess the effectiveness of the models. The comparative analysis aims to provide insights into the relative strengths and weaknesses of text-only, image-only, and multimodal approaches.

Limitations and Future Work: Finally, the scope of the thesis includes a discussion of the limitations encountered during the research, such as data availability, model complexity, and computational constraints. The thesis also outlines potential directions for future research, including the exploration of more advanced models, the incorporation of additional data modalities (e.g., audio or video), and the application of the developed models to other related tasks, such as deepfake detection or misinformation tracking.

This comprehensive scope ensures that the thesis not only addresses the immediate problem of fake news detection but also contributes to the broader field of NLP and multimodal analysis, providing a foundation for future research and applications in this critical area.

1.5 Contributions

This thesis makes several significant contributions to the field of fake news detection and multimodal analysis. The key contributions are as follows:

1. **Novel Dataset Preprocessing and Integration Techniques**: The research introduces innovative pre-processing methods for both textual and visual data, including the fusion of multimodal features at different stages of the model pipeline. These techniques are designed to optimize the integration of diverse data types, ensuring that the models can effectively leverage

the complementary information provided by text and images. This contribution is particularly relevant for future work in multimodal analysis, where data integration remains a challenging aspect.

2. **Comprehensive Evaluation and Comparative Analysis of Single-modality and Multimodal Approaches**: The thesis provides an extensive comparative analysis of different approaches to fake news detection, including text-only, image-only, and multimodal methods. This analysis highlights the strengths and limitations of each approach, offering insights into the conditions under which multimodal models outperform single-modality models. The evaluation also includes a detailed assessment of the impact of data quality, model architecture, and task complexity on detection performance.

3. **Multimodal Framework for Small-Scale Dataset**: In the realm of open-source data, we often encounter abundant resources for *Text-Only Classification* and *Image-Only Classification* tasks. However, finding multimodal datasets, which integrate both text and image data, is significantly more challenging. Even when such datasets are available, they tend to be large-scale, often containing tens of thousands, or even hundreds of thousands, of data points. These extensive datasets naturally lend themselves to higher accuracy, more balanced classification reports, and better performance metrics such as F1-Score and recall.

   In contrast, the dataset used in this study contains only 1,032 entries. The challenge lies in achieving comparable accuracy and robust classification results with a smaller dataset. To address this, I have developed a multimodal framework designed to scale effectively with smaller datasets. By carefully tuning the model, this framework demonstrates that it is possible to achieve high accuracy and reliable performance metrics even when working with limited data. This approach is particularly valuable in scenarios

where acquiring large-scale multimodal datasets is not feasible, offering a scalable solution that adapts to the constraints of smaller datasets without compromising on performance.

Chapter 2

LITERATURE REVIEW

– Exploring Various Approaches in the Fight Against Misinformation

The Literature Review provides a foundation for understanding the existing research and methodologies in the field of multimodal fake news detection. It explores the evolution of text and image classification techniques, highlighting the strengths and limitations of traditional and deep learning approaches. By examining the current state of the art, this review identifies gaps in the literature, particularly in the integration of multimodal data for more accurate and reliable fake news detection. The insights gathered from this review inform the development of the proposed framework, ensuring it builds on established knowledge while addressing the shortcomings of existing models.

2.1 Fake News Detection: An Overview

Fake news detection has become an increasingly critical area of research due to the widespread dissemination of misinformation via digital platforms. The rapid spread of fake news can have serious consequences, from influencing elections to inciting public unrest [1]. Traditional media outlets, once the gatekeepers of information, have been supplanted by social media platforms where news spreads unchecked and often without verification [2]. This shift has made it imperative to develop automated systems capable of identifying and mitigating the spread of fake news [3].

Early approaches to fake news detection focused primarily on manual fact-checking and keyword-based algorithms [4]. These methods, while somewhat effective in specific contexts, were limited in scalability and accuracy [5]. As the

sophistication of fake news has increased, so too has the need for more advanced detection methods. This has led to the adoption of machine learning (ML) and, more recently, deep learning (DL) techniques, which have significantly improved the ability to detect fake news by analyzing large datasets and identifying complex patterns in both textual and visual data [6] [7].

The literature reveals a clear evolution in the methods used for fake news detection, moving from simple keyword-based systems to more complex models that leverage the latest advancements in NLP and computer vision [8] [9]. However, despite the progress made, challenges remain, particularly in integrating multimodal data and addressing the nuances of fake news across different languages and cultures [10].

2.2 Text Classification Techniques

Text classification techniques are crucial in fake news detection, as they help identify misleading or false information in textual content. Traditional methods often rely on keyword matching and rule-based systems, which have limited effectiveness against sophisticated fake news.

2.2.1 Traditional Methods

In the early stages of fake news detection research, traditional text classification methods were widely used. These methods typically relied on manually crafted features, such as the presence of certain keywords or phrases, which were then used to classify news articles as true or fake. For example, approaches based on **Naive Bayes**[1] and **Support Vector Machines (SVM)**[2] were popular due to their simplicity and ease of implementation [11]. These models worked by identifying the probability that a given article belonged to a particular class based on its textual features [12].

[1]https://scikit-learn.org/stable/modules/naive_bayes.html
[2]https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Lexicon-based approaches were also common, where a predefined dictionary of sentiment-laden words was used to determine the tone of an article [13]. If an article contained an unusually high number of negative or sensationalist words, it might be flagged as fake [14]. However, these approaches were limited in their ability to capture the context in which words were used, leading to high rates of false positives and false negatives [15].

Despite their limitations, traditional methods laid the groundwork for more advanced techniques by highlighting the importance of feature selection and the need for models that could handle the complexity and variability of human language. However, as fake news creators became more adept at avoiding detection by these rudimentary systems, the need for more sophisticated, context-aware models became apparent.

2.2.2 Deep Learning Approaches

The advent of deep learning brought about a paradigm shift in text classification for fake news detection. **Long Short-Term Memory (LSTM)**[3] networks, which are capable of processing and retaining sequential information, allowed researchers to analyze text in a more contextually aware manner [16]. Unlike traditional methods, LSTM models can capture the relationships between words over long sequences, making them particularly effective for understanding the narrative structure of news articles [17].

Bidirectional Encoder Representations from Transformers (BERT)[4] further advanced the field by enabling models to understand the bidirectional context of words. BERT's transformer architecture allows it to consider the entire sentence when determining the meaning of each word, rather than just the preceding or following words [18]. This has proven to be particularly useful in detecting subtle linguistic cues that may indicate whether an article is fake or real [19]. For example, BERT can identify discrepancies between the headline

[3]https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html
[4]https://huggingface.co/docs/transformers/en/model_doc/bert

and body of an article, or detect the use of hedging language that might suggest a lack of credibility.

Deep learning models, particularly those based on BERT, have consistently outperformed traditional methods in fake news detection tasks. Their ability to learn from large datasets without the need for manual feature extraction has made them the preferred choice for many researchers [20]. Moreover, these models can be fine-tuned on specific datasets, allowing for greater accuracy in different domains or languages.

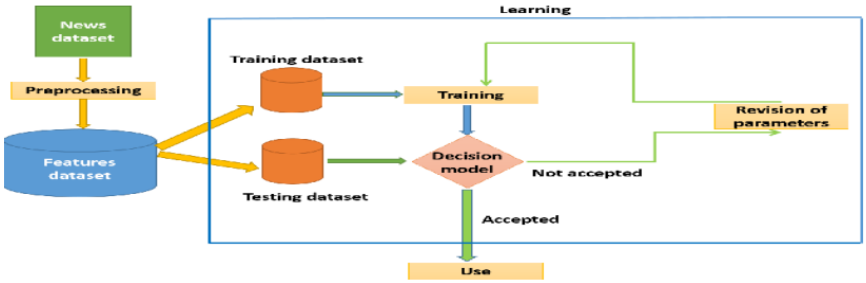


Figure 2.1: Fake News Detection Workflow

2.3 Image Classification Techniques

Image classification plays a crucial role in detecting fake news by analyzing visual content for signs of manipulation or inconsistency.These below techniques providing a more comprehensive approach to identifying and mitigating the spread of fake news

2.3.1 Traditional Methods

Traditional image classification methods for fake news detection were primarily based on feature extraction techniques like **Scale-Invariant Feature Transform (SIFT)**[5] and **Speeded-Up Robust Features (SURF)** [21]. These methods involved identifying key points in an image and extracting features that could be used to determine whether an image had been manipulated. For exam-

[5]https://en.wikipedia.org/wiki/Scale-invariant_feature_transform

ple, SIFT would identify unique features in an image that remained consistent despite changes in scale or rotation, which could then be compared against a database of known fake images.

Another common approach was histogram analysis, where the distribution of colors or intensity levels in an image was analyzed to detect anomalies. For instance, a sudden shift in the histogram might indicate that an image had been digitally altered [22]. However, these methods were often limited by their reliance on manually crafted features, which could miss more sophisticated manipulations that did not affect the basic visual properties of the image.

Despite their utility in specific contexts, traditional methods struggled with the complexity and scale of modern fake news. They were particularly ineffective at detecting deepfakes or other advanced forms of image manipulation that could easily bypass these simple checks [23].

2.3.2 Deep Learning Approaches

The introduction of **Convolutional Neural Networks (CNNs)**[6] revolutionized image classification, including its application to fake news detection. CNNs automatically learn hierarchical features from raw image data, making them far more effective than traditional methods at detecting complex patterns and manipulations. For instance, ResNet-50, a widely used CNN architecture, incorporates residual connections that help maintain the integrity of the gradient as it propagates through deep networks. This enables the model to detect even subtle alterations in images, such as slight changes in lighting or texture that might indicate forgery.

In fake news detection, CNNs are used to analyze images associated with news articles to determine their authenticity. For example, a CNN might be trained to recognize specific types of manipulations, such as those commonly used in doctored photos or deepfakes. When combined with textual analysis,

[6]https://pytorch.org/docs/stable/generated/torch.nn.Conv2d.html

these models can provide a more comprehensive assessment of the credibility of a news article by cross-referencing the content of the text with the visual data.

Studies have shown that deep learning models, particularly CNNs like ResNet-50, significantly outperform traditional methods in detecting fake images. Their ability to automatically extract and analyze features from images makes them indispensable tools in the fight against fake news.

2.4 Multimodal Classification

The integration of multimodal data—combining both text and image analysis—represents a significant advancement in the field of fake news detection. Multimodal classification leverages the strengths of both NLP and computer vision to provide a more holistic approach to identifying fake news. While text-based models can analyze the language and structure of news articles, image-based models can detect manipulations or inconsistencies in the accompanying visuals. When these two modalities are combined, the resulting model is better equipped to identify fake news that might evade detection by a single-modality system.

Multimodal approaches typically involve the fusion of features extracted from text and images at different stages of the model. Early fusion techniques combine these features at the input level, feeding both text and image data into a single model that processes them simultaneously. Late fusion, on the other hand, involves processing each modality separately before combining their outputs at a later stage, such as in the final classification layer. This approach allows the model to leverage the strengths of each modality without overwhelming it with too much data at once.

Recent research has demonstrated the effectiveness of multimodal models in improving the accuracy of fake news detection. For instance, combining BERT for text analysis with **ResNet-50** for image analysis has been shown to outperform models that rely on either modality alone. The ability to cross-verify information across text and images reduces the likelihood of false positives and negatives,

leading to more reliable detection systems.

2.5 Summary of Literature Gaps

Despite the significant progress made in fake news detection [24], several gaps remain in the literature that warrant further exploration. First, while multimodal approaches have shown promise, much of the existing research focuses on binary classification tasks—determining whether a piece of news is true or false. There is a need for more fine-grained classification systems that can distinguish between different levels of falsity, such as partially true or misleading information.

Second, while deep learning models have achieved impressive results, they often require large amounts of labelled data, which can be a limiting factor in real-world applications. There is a growing need for research into semi-supervised and **unsupervised learning techniques** that can reduce the reliance on labelled data while still achieving high levels of accuracy.

Third, the majority of existing research has been conducted on English-language datasets, with relatively little attention given to the challenges posed by other languages. Given the global nature of fake news, there is a need for more research on multilingual and cross-cultural fake news detection, particularly in regions with less-resourced languages.

Finally, while multimodal approaches are becoming more common, the integration of additional data types, such as audio and video, remains under-explored. Expanding the scope of multimodal fake news detection to include these modalities could further improve the robustness and accuracy of detection systems.

Addressing these gaps will be crucial for the continued advancement of fake news detection technologies, ensuring that they remain effective in the face of increasingly sophisticated misinformation tactics.

Chapter 3

DATASET DESCRIPTION

– Examining the Foundations: A Detailed Overview of Multimodal Data for Fake News Detection

The dataset used in this study is a multimodal collection consisting of 1,032 entries, where each entry includes both text and corresponding images. The data is labeled into four categories: Certainly Fake, Probably Fake, Probably Real, and Certainly Real. The text data captures captions or descriptions, while the images provide visual context related to the text. This small-scale dataset is designed to test the efficacy of a multimodal classification framework, which aims to achieve high accuracy and robust performance metrics, even with a limited number of samples. This dataset serves as a crucial resource for developing and evaluating models that can handle both textual and visual information simultaneously.

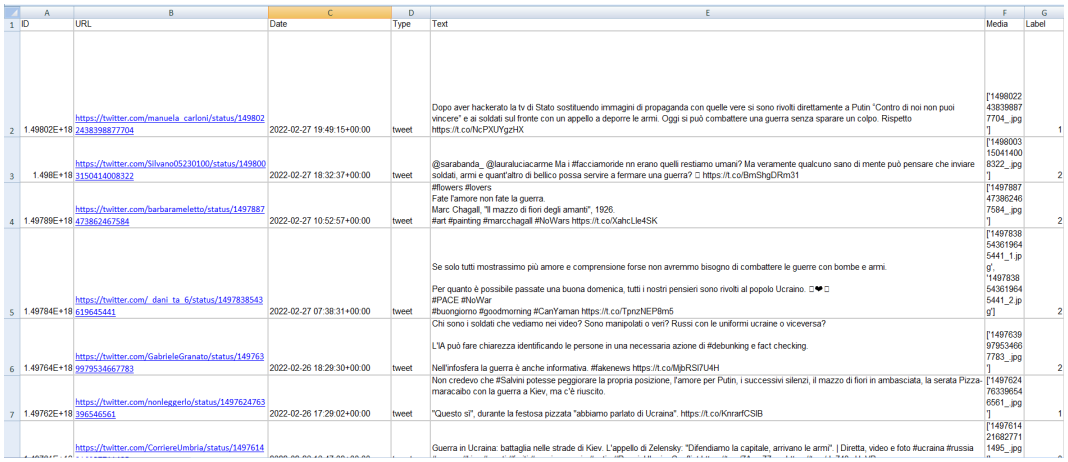
3.1 Overview of the Dataset

The ***MULTI-Fake-Detective-Task1-Data***[1] dataset is a comprehensive collection of multimodal data, specifically curated for the task of fake news detection. The dataset comprises entries that include both textual and visual information, representing real-world instances of news and social media posts. Each entry in the dataset is labelled according to the perceived authenticity of the content, with labels ranging from *Certainly Fake* to *Certainly Real*.

[1] https://sites.google.com/unipi.it/multi-fake-detective/tasks-and-evaluation?authuser=0

The dataset includes the following key columns:

- **ID**: A unique identifier for each entry.
- **URL**: The web link where the content was originally published, often on social media platforms like Twitter.
- **Date**: The timestamp indicating when the content was posted.
- **Type**: The type of content, such as a tweet, indicating its origin.
- **Text**: The textual content of the post or article, which is a crucial element for NLP-based analysis.
- **Media**: The visual content associated with the post, typically in the form of images, which is used for image classification.
- **Label**: The classification label assigned to the entry, indicating its authenticity, where:
  - 0 represents *Certainly Fake*
  - 1 represents *Probably Fake*
  - 2 represents *Probably Real*
  - 3 represents *Certainly Real*



	A	B	C	D	E	F	G
1	ID	URL	Date	Type	Text	Media	Label
2	1.49802E+18	https://twitter.com/manuela_carloni/status/1498022438398877704	2022-02-27 19:49:15+00:00	tweet	Dopo aver hackerato la tv di Stato sostituendo immagini di propaganda con quelle vere si sono rivolti direttamente a Putin "Contro di noi non puoi vincere" e ai soldati sul fronte con un appello a deporre le armi. Oggi si può combattere una guerra senza sparare un colpo. Rispetto https://t.co/NcPXUYgzHX	['1498022438398877704_jpg']	1
3	1.498E+18	https://twitter.com/Silvano05230100/status/1498003150414008322	2022-02-27 18:32:37+00:00	tweet	@sarabanda_ @lauraluciacarme Ma i #facciamoride nn erano quelli restiamo umani? Ma veramente qualcuno sano di mente può pensare che inviare soldati, armi e quant'altro di bellico possa servire a fermare una guerra? □ https://t.co/BmShgDRm31	['1498003150414008322_jpg']	2
4	1.49789E+18	https://twitter.com/barbarameletto/status/1497887473862467584	2022-02-27 10:52:57+00:00	tweet	#flowers #lovers Fate l'amore non fate la guerra. Marc Chagall, "Il mazzo di fiori degli amanti", 1926. #art #painting #marcchagall #NoWars https://t.co/XahcLle4SK	['1497887473862467584_jpg']	2
5	1.49784E+18	https://twitter.com/_dani_ta_6/status/1497838543619645441	2022-02-27 07:38:31+00:00	tweet	Se solo tutti mostrassimo più amore e comprensione forse non avremmo bisogno di combattere le guerre con bombe e armi. Per quanto è possibile passate una buona domenica, tutti i nostri pensieri sono rivolti al popolo Ucraino. □❤□ #PACE #NoWar #buongiorno #goodmorning #CanYaman https://t.co/TpnzNEP8m5	['1497838543619645441_1.jpg', '1497838543619645441_2.jpg']	2
6	1.49764E+18	https://twitter.com/GabrieleGranato/status/1497639795346677783	2022-02-26 18:29:30+00:00	tweet	Chi sono i soldati che vediamo nei video? Sono manipolati o veri? Russi con le uniformi ucraine o viceversa? L'IA può fare chiarezza identificando le persone in una necessaria azione di #debunking e fact checking. Nell'infosfera la guerra è anche informativa. #fakenews https://t.co/MjbRSI7U4H	['1497639795346677783_jpg']	2
7	1.49762E+18	https://twitter.com/nonleggerlo/status/1497624763396546561	2022-02-26 17:29:02+00:00	tweet	Non credevo che #Salvini potesse peggiorare la propria posizione, l'amore per Putin, i successivi silenzi, il mazzo di fiori in ambasciata, la serata Pizza-maracaibo con la guerra a Kiev, ma c'è riuscito. "Questo sì", durante la festosa pizzata "abbiamo parlato di Ucraina". https://t.co/KnrarfCSIB	['1497624763396546561_jpg']	1
		https://twitter.com/CorriereUmbria/status/1497614216827711495_jpg			Guerra in Ucraina: battaglia nelle strade di Kiev. L'appello di Zelensky. "Difendiamo la capitale, arrivano le armi"	Diretta, video e foto #ucraina #russia	['1497614216827711495_jpg']	

Figure 3.1: Overview of the Dataset

The dataset is a valuable resource for developing and evaluating multimodal fake news detection models that utilize both text and images.

3.2 Labeling Strategy

The labeling strategy for this dataset is designed to capture the nuanced levels of authenticity in the content. Each entry is labelled with one of four possible classes:

- **0 (Certainly Fake)**: Content that is clearly and verifiably false.
- **1 (Probably Fake)**: Content that is likely false but may contain some elements of truth.
- **2 (Probably Real)**: Content that appears to be true but may have minor inaccuracies.
- **3 (Certainly Real)**: Content that is verified to be accurate and truthful.

The labeling was likely performed by experts or through crowd-sourced verification methods. The use of four distinct labels allows for more granular analysis and the development of models that can discern subtle differences in content authenticity.

Example: Let's consider an example to illustrate the labeling strategy described: Suppose we have a dataset of social media posts related to a controversial news topic. Each post is labeled based on the perceived authenticity of its content using the following criteria:

0 (Certainly Fake) - Example: A post claiming that *The moon has turned green because of a massive alien invasion*, which is clearly false and has no basis in reality. The content is deliberately misleading.

1 (Probably Fake) - Example: A post stating that *There are unconfirmed reports of a miracle cure for a deadly disease*, which might be based on rumors

or partial truths but lacks substantial evidence. The content is likely false but could have some elements that are true.

2 (Probably Real) - Example: A post reporting on a new scientific study that suggests a potential link between diet and heart health, but with minor inaccuracies in how the findings are presented. The content appears to be real but may have small errors or exaggerations.

3 (Certainly Real) - Example: A post from a verified news outlet stating, *The government has officially announced a new policy on climate change*, which is supported by credible sources and factual information. The content is verified to be accurate and truthful.

In this example, the labeling strategy enables a nuanced classification of content, allowing the model to distinguish between varying degrees of authenticity and thereby improving the accuracy of detection in complex datasets.

3.3 Statistical Analysis of the Dataset

A statistical analysis of the dataset provides insights into the distribution and characteristics of the data, which is essential for understanding potential biases and challenges in model training.

Key Statistics:

- **Distribution of Labels**: An analysis of how the labels are distributed across the dataset helps in identifying class imbalances, which are common in fake news detection datasets. For instance, there may be more instances of *Certainly Fake* content compared to *Certainly Real* content, which can skew model performance if not addressed.

- **Text Length**: The average length of the text entries can provide insights into the complexity of the content. Shorter texts might be more challenging for NLP models, as they contain less context.

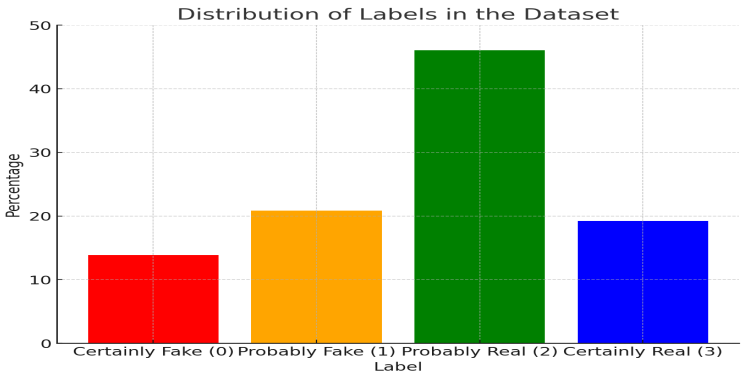


Figure 3.2: Distribution of Labels

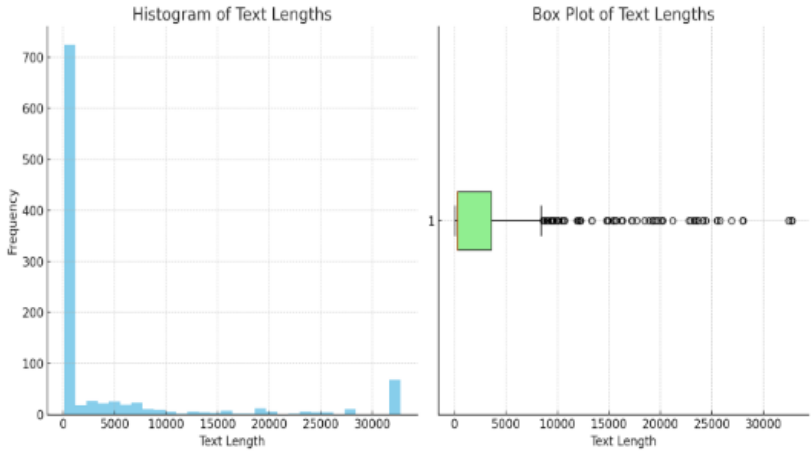


Figure 3.3: Histogram & Box Plot of Text Lengths

- **Image Characteristics**: The size, resolution, and type of images (e.g., single image vs. multiple images per post) can influence the performance of image classification models. A statistical summary of these characteristics helps in designing appropriate pre-processing pipelines and choosing the right model architectures.

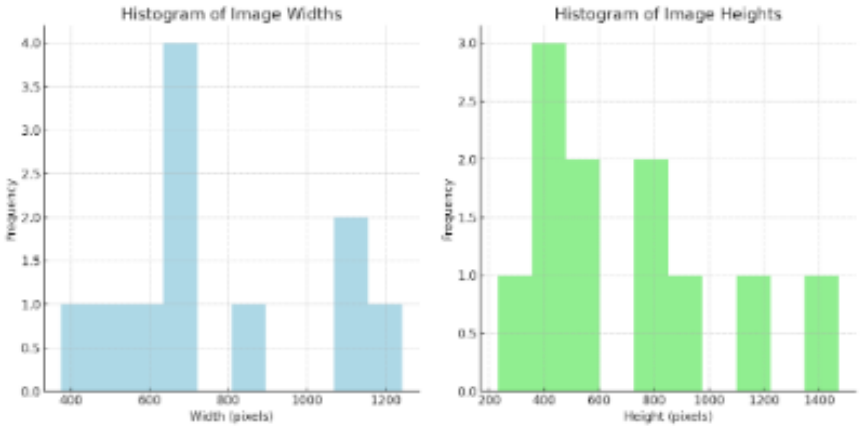


Figure 3.4: Histograms of Image Widths and Heights

Understanding these statistical aspects of the dataset is crucial for developing effective pre-processing strategies and for selecting the appropriate models and evaluation metrics.

3.4 Overview of Comparative Dataset

The **Fakeddit**[2] dataset is a widely recognized open-source dataset in the realm of fake news detection, characterized by its inclusion of both textual content and associated images.This dataset is a great contributor of my ***MULTI-Fake-Detective-Task1-Data***[3] dataset. This dataset is constructed to support the development and evaluation of multimodal fake news detection models, which require the integration of textual and visual information for accurate classification. This dataset will help to know about overall performance of the models created using ***MULTI-Fake-Detective-Task1-Data***[4] dataset after compare these two datasets later.

[2]https://fakeddit.netlify.app/
[3]https://sites.google.com/unipi.it/multi-fake-detective/tasks-and-evaluation?authuser=0
[4]https://sites.google.com/unipi.it/multi-fake-detective/tasks-and-evaluation?authuser=0

Key Components:

- **Title:** This column contains the textual content of each entry, which can be headlines, captions, or brief descriptions. The text is crucial for natural language processing (NLP) tasks, including fake news detection, sentiment analysis, and content classification.
- **Image URL:** This column provides the URLs to the associated images for each entry. These images are used for visual analysis tasks, which are essential in verifying the authenticity of the news by detecting manipulated or misleading visuals that accompany the textual content.
- **Label:** The label column classifies each entry into one of two categories:
  - **0:** Fake – indicating that the content is likely to be deceptive or misleading.
  - **1:** Real – indicating that the content is likely to be accurate and truthful.

The dataset contains 4,806 rows, each representing a unique combination of textual and visual data, thus offering a diverse and complex dataset for researchers and developers working on multimodal fake news detection. The dual modality of this dataset (text + image) is instrumental in creating models that can simultaneously evaluate the credibility of text and images, making it more robust against sophisticated fake news.

This dataset is particularly valuable for those looking to explore the integration of NLP and computer vision techniques in a unified framework, aiming to improve the accuracy and reliability of fake news detection systems across different media types.

3.5 Data Collection and Preprocessing

3.5.1 Text Data

The textual data in the dataset consists of various posts, primarily from social media platforms like Twitter. These texts are typically short, informal, and often include hashtags, mentions, and URLs, which must be pre-processed to prepare the data for analysis.

	A	B
1	Text	Label
2	PsBattle: Elon Musk's reaction to the Falcon Heavy Launch	1
3	'Ghost Boy' to Sputnik: 'Parents Were Told to Take Me Home to I	1
4	PsBattle: NASCAR driver Jimmie Johnson with a poorly designed c	1
5	Seafood Raised on Pig and Goose Faeces Approved for Consume	1
6	Found This Little Guy In My Shrubs	0
7	3-year-old boy shoots father, pregnant mother in New Mexico	0
8	My Sixlets Package Only Had Five; Fivelets?	0
9	Dr Daniel Ubani wins costs from Cambridgeshire victim's son	0
10	PsBattle: This goat race gone awry	1
11	Italy elections: Rome poised to elect Virginia Raggi as first female	0
12	PsBattle: Wolf eating watermelon	0
13	Woman realizes cat she rescued is in fact a bobcat.	1
14	The Atlanta airport now has touchless air fresheners installed in e	1
15	This wasp landed on my window at work and started blowing bub	0
16	My friends and I getting high before prom	0
17	BYU football: Cougars were too nice? Bearcats want revenge for	1
18	PsBattle: How To Care For Llamas[1852 x 1852]	0
19	PsBattle: This kitty in its spaceship-like bed	0
20	Magistrate tells Aboriginal teenager to be more like US rapper 50	0
21	Kanye in Paris for wedding weekend	1
22	Mattis ignored orders from Trump, White House on North Korea,	0

Figure 3.5: Overview of the Fakeddit Text Dataset

Text Preprocessing Steps:

1. **Tokenization**: The text is split into individual tokens (words, hashtags, mentions) to facilitate further processing.

2. **Lowercasing**: All text is converted to lowercase to ensure uniformity and to prevent case-sensitive mismatches.

3. **Removal of Stopwords**: Common words that do not contribute to the meaning (e.g., *and*, *the*) are removed to focus on the more informative words.

4. **Handling Hashtags and Mentions**: Special tokens like hashtags and mentions are either removed or processed to retain their contextual meaning.

5. **Lemmatization**: Words are reduced to their base or root form, which helps in reducing the dimensionality of the text data.

6. **URL Removal**: URLs included in the posts are typically removed, as they do not contribute meaningfully to the textual analysis.

These pre-processing steps ensure that the text data is clean and ready for input into NLP model like LSTM. I have also used XGBoost[5] which is a highly efficient and scalable machine learning algorithm that is well-suited for text classification tasks.

3.5.2 Image Data

The image data associated with each post is a critical component for multimodal fake news detection. Images often accompany textual content, providing additional context or visual cues that can be crucial in determining the authenticity of the information.

	A	B
1	Image_url	Label
2	https://external-preview.redd.it/XRSvIEJ1pA-_FV2_OCMI-r5hXATATyVhf	1
3	https://external-preview.redd.it/kW6_-G2cpvOI2WwsFMQYwFFn-v8nvp	1
4	https://external-preview.redd.it/DFZWJ3CbgJJn1TqmzJkEmIo_YKJHZOj6	1
5	https://external-preview.redd.it/Ez0g1hwvyN9fJDeiI6vBU1uI0QKOOPAE	1
6	http://www.nerdlikeyou.com/wp-content/uploads/2013/04/jurassic-pa	0
7	http://i.imgur.com/plFt4Qc.jpg	0
8	https://preview.redd.it/zcuuijn674t01.jpg?width=320&crop=smart&autc	0
9	http://i.imgur.com/VvmioG8.jpg	0
10	https://preview.redd.it/0tq5trx90r231.jpg?width=320&crop=smart&aut	1
11	https://preview.redd.it/303p41rm8gpz.jpg?width=320&crop=smart&aut	0
12	https://preview.redd.it/mp20eg2mp3b31.png?width=320&crop=smart&	0
13	https://external-preview.redd.it/Dl_TWLqnxeM8k0dhhYjhMibNBFjjEm3s	1
14	https://preview.redd.it/onlx2jo3dan21.jpg?width=320&crop=smart&aut	1
15	https://external-preview.redd.it/v0lfzFk7NMS6peNv-Aqh78npmwDdhBK	0
16	http://i.imgur.com/OflMqLR.jpg	0
17	https://external-preview.redd.it/iHM0puWXKtWXL-u9MBGy5SssrWssL0	1

Figure 3.6: Overview of the Fakeddit Image Dataset



Figure 3.7: Example Image

[5]https://xgboost.readthedocs.io/en/stable/python/sklearn_estimator.html

Image Preprocessing Steps:

1. **Resizing**: All images are resized to a uniform dimension to ensure consistency when inputting them into models like ResNet-50[6], MobileNetV2[7], LSTM[8]. This step is important for maintaining the aspect ratio and ensuring that all images are comparable.

2. **Normalization**: Pixel values are normalized to a standard range (typically 0-1) to improve the convergence of deep learning models.

3. **Augmentation**: To enhance the diversity of the training set, image augmentation techniques such as rotation, flipping, and zooming are applied. This helps in making the model more robust to variations in image data.

4. **Conversion to Tensors**: The images are converted into tensors, which are the data structures used by deep learning frameworks to perform computations.

The pre-processing of images is crucial for ensuring that the visual data is appropriately formatted for deep learning models, enabling the effective extraction of features that may indicate whether the image is real or manipulated.

[6]https://pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html
[7]https://pytorch.org/hub/pytorch_vision_mobilenet_v2/
[8]https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html

Chapter 4

TEXT CLASSIFICATION

– Decoding Linguistic Patterns for Accurate Fake News Detection

Text Classification is a fundamental task in natural language processing (NLP) that involves categorizing text into predefined labels or classes. It is widely used in various applications such as sentiment analysis, spam detection, and topic labeling. The process typically involves converting text data into numerical features that a machine learning model can understand, followed by training a model to predict the correct category for new, unseen text. Techniques like Bag of Words, TF-IDF, and word embeddings are commonly used for feature extraction, while models such as Logistic Regression, Support Vector Machines (SVM), and deep learning methods like LSTM and BERT are employed for classification. Text classification enables automated decision-making and information filtering, making it a crucial tool in handling large volumes of textual data efficiently.

4.1 Text Preprocessing

Text preprocessing is a crucial initial step in preparing textual data for analysis by machine learning models. It involves transforming raw text into a format that models can understand and process efficiently. The process typically includes steps like tokenization, where text is split into individual words or tokens, and normalization, which standardizes the text by converting it to lowercase, removing punctuation, and handling special characters. These steps ensure consistency in the data, reduce noise, and improve the model's ability to extract meaningful patterns, ultimately enhancing the accuracy and performance of text classification tasks.

4.1.1 Tokenization and Normalization

Tokenization is the process of splitting the text into individual words or tokens, which are the basic units of meaning for natural language processing. This step is crucial for converting raw text into a format that can be analyzed by machine learning models. Following tokenization, normalization is applied to standardize the text. Normalization involves converting all text to lowercase to eliminate case sensitivity, removing punctuation, and sometimes eliminating special characters or numbers. This step helps ensure that the text is uniform across the dataset, reducing variability that could confuse the model. By standardizing the text, normalization makes it easier for the model to focus on the meaningful content rather than on formatting inconsistencies.

4.1.2 Handling Imbalanced Data

In datasets where certain classes are over-represented, handling imbalanced data becomes a critical step to ensure that the model does not become biased towards the majority class. One approach to address this issue is oversampling, where additional examples of the minority class are generated, often through techniques like **SMOTE (Synthetic Minority Over-sampling Technique)**[1]. Another approach is to adjust the class weights during model training, giving more importance to the minority classes. This encourages the model to pay equal attention to all classes, improving its ability to correctly classify instances from the less frequent categories. Handling imbalances is essential for achieving accurate and fair classification results, particularly in tasks like fake news detection where the distribution of real versus fake news may be uneven.

[1]https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html

4.2 Model Architectures

4.2.1 Bi-LSTM Model

Bi-LSTM[2] or Bidirectional Long Short-Term Memory, is an advanced variant of the LSTM model, (RNN)[3] designed to enhance the understanding of sequential data. While a standard LSTM processes input data in a single direction—from past to future—Bi-LSTM improves upon this by processing data in both forward and backward directions simultaneously.

In a Bi-LSTM model, two separate LSTM layers are employed: one processes the sequence from start to end (forward direction), and the other processes it from end to start (backward direction). The outputs from both layers are then combined, typically by concatenation, to capture information from both directions.

This bidirectional approach is particularly useful in natural language processing tasks, where context from both past and future words can provide a deeper understanding of the meaning of a word within a sentence. For instance, in sentence translation or sentiment analysis, knowing the words that follow can be just as important as understanding the words that precede.

By capturing dependencies in both directions, Bi-LSTM models are more effective at understanding the context, especially in complex sequences where meaning is derived from relationships between distant words or elements in the sequence. This makes Bi-LSTM particularly powerful for tasks like text classification, named entity recognition, and machine translation.

In summary, Bi-LSTM extends the capability of standard LSTM by considering both past and future context, making it a robust tool for sequence modeling in various applications.

[2]https://www.geeksforgeeks.org/bidirectional-lstm-in-nlp/
[3]https://pytorch.org/docs/stable/generated/torch.nn.RNN.html

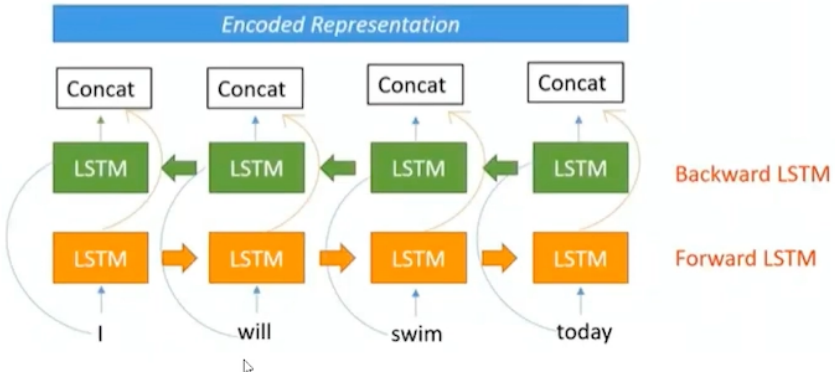


Figure 4.1: Bi-LSTM Architecture

4.2.2 XGBoost Model

XGBoost, or Extreme Gradient Boosting, is a powerful machine learning algorithm that is widely used for structured data classification tasks, including text classification. It operates by creating an ensemble of decision trees, where each tree is built to correct the errors made by the previous ones. This iterative approach allows XGBoost to improve the accuracy of its predictions significantly. XGBoost is known for its efficiency and scalability, making it suitable for large datasets. Its robustness comes from the model's ability to handle missing data, optimize computation, and prevent overfitting through regularization techniques. These features make XGBoost a versatile and reliable model for text classification, where it can effectively capture patterns and relationships within the data to accurately categorize text content.

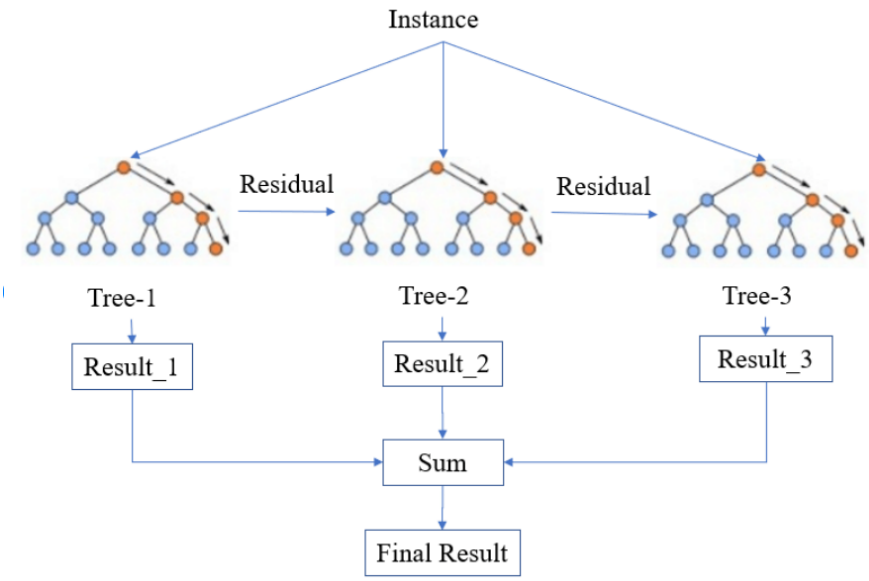


Figure 4.2: XGBoost Model Architecture for Text Classification

4.2.3 Without Translated Text

- **Dataset Characteristics**

  The *Without Translated Text* dataset comprises text entries in their original languages, representing a diverse linguistic landscape. This dataset includes various languages, each with its own syntax, grammar, and linguistic nuances. The presence of multiple languages adds complexity to text preprocessing, feature extraction, and model training. Unlike a monolingual dataset, the *Without Translated Text* dataset challenges the model's ability to generalize across different languages and dialects.

- **Challenges in Text Classification**

  When working with a multilingual dataset, several challenges arise. Firstly, tokenization and normalization become more complicated due to differences in language structures. For example, some languages may use different punctuation marks or have unique tokenization requirements. Additionally, the models must be able to handle varying word orders, idiomatic expressions, and cultural references that may differ significantly between languages.

Another challenge is the potential for class imbalance across languages. If certain languages are over-represented in the dataset, the model may become biased towards those languages, leading to reduced performance on under-represented languages. Moreover, language-specific features such as word embeddings may not be uniformly available for all languages, further complicating model training.

- **Bi-LSTM Model on *Without Translated Text***

  The Bidirectional Long Short-Term Memory (Bi-LSTM) model is particularly effective for text classification tasks, especially when dealing with multilingual datasets like the *Without Translated Text* dataset. Unlike a standard LSTM, which processes data in a single direction, Bi-LSTM captures context from both the past and the future by processing the text in two directions simultaneously. This bidirectional capability makes Bi-LSTM exceptionally well-suited for understanding the complex linguistic structures present in a multilingual context.

  The diversity of languages within the *Without Translated Text* dataset introduces additional complexity, requiring the Bi-LSTM model to adapt to a wide range of grammatical and syntactical variations. By processing each text sequence from both directions, the Bi-LSTM model can more effectively capture these nuances, enabling it to maintain a richer understanding of the text across different languages.

  To further enhance the Bi-LSTM model's performance, it can be initialized with pre-trained multilingual embeddings such as mBERT (multilingual BERT) or XLM-RoBERTa, which provide a robust starting point by embedding words from multiple languages into a shared vector space. These embeddings allow the Bi-LSTM model to recognize and leverage linguistic similarities between languages, improving its ability to generalize across the dataset. However, the effectiveness of these embeddings can vary depending on the language, which may introduce challenges in achieving consistent

performance across all languages in the dataset.

Despite these challenges, the Bi-LSTM model's ability to process text bidirectionally allows it to capture long-term dependencies and contextual relationships more effectively than a unidirectional LSTM. This capability is crucial for identifying the overall meaning and intent of the text, regardless of the language. By focusing on the relationships between words and phrases from both directions, the Bi-LSTM model can better generalize across the multilingual dataset, providing accurate and reliable text classification results.

Table 4.1: Performance of Bi-LSTM on the Non-translated Text

	Precision	**Recall**	**F1-Score**	**Support**
Certainly Fake	0.33	0.31	0.32	29
Probably Fake	0.31	0.37	0.33	41
Probably Real	0.59	0.60	0.60	102
Certainly Real	0.54	0.43	0.48	35
Accuracy			0.48	207
Macro Avg.	0.44	0.43	0.43	207
Weighted Avg.	0.49	0.48	0.48	207

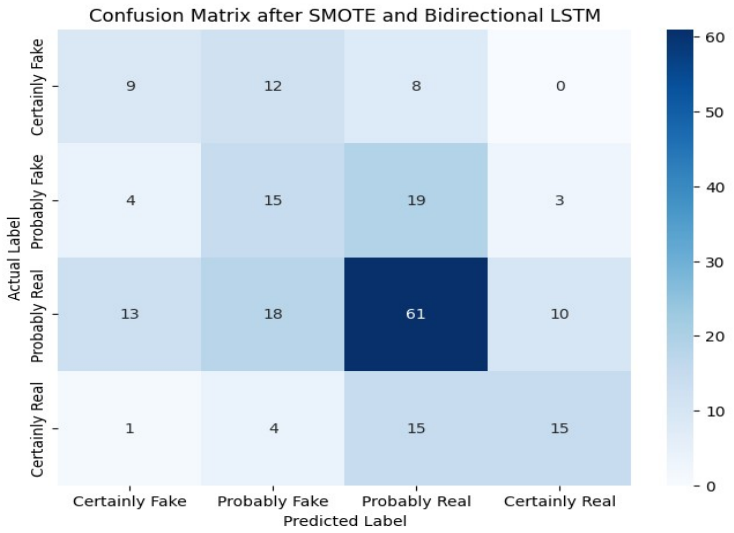


Figure 4.3: Confusion Matrix of Bi-LSTM on Non-translated Texts

- **XGBoost Model on *Without Translated Text***

  XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm that excels in structured data classification. However, when applied

to a multilingual dataset like *Without Translated Text*, XGBoost faces several challenges. The algorithm requires the text data to be transformed into numerical features, typically through techniques like TF-IDF[4] or word embeddings.

One advantage of XGBoost is its ability to handle a large number of features, which is particularly useful in a multilingual context where each language may contribute unique features. However, the model's performance may vary depending on the quality of the feature extraction process. If certain languages have more informative features, XGBoost may become biased towards those languages, potentially leading to imbalanced classification performance.

To mitigate this, it is essential to pre-process the text data carefully, ensuring that all languages are represented equally in the feature space. Additionally, XGBoost's hyperparameters, such as learning rate and tree depth, can be fine-tuned to optimize performance across the multilingual dataset. Despite these challenges, XGBoost's robustness and ability to handle complex data make it a viable option for text classification in the *Without Translated Text* dataset.

Table 4.2: Performance of XGBoost on the Non-translated Text

	Precision	**Recall**	**F1-Score**	**Support**
Certainly Fake	0.62	0.34	0.44	29
Probably Fake	0.44	0.41	0.43	41
Probably Real	0.63	0.77	0.69	102
Certainly Real	0.54	0.40	0.46	35
Accuracy			0.58	207
Macro Avg.	0.56	0.48	0.51	207
Weighted Avg.	0.57	0.58	0.57	207

[4]https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

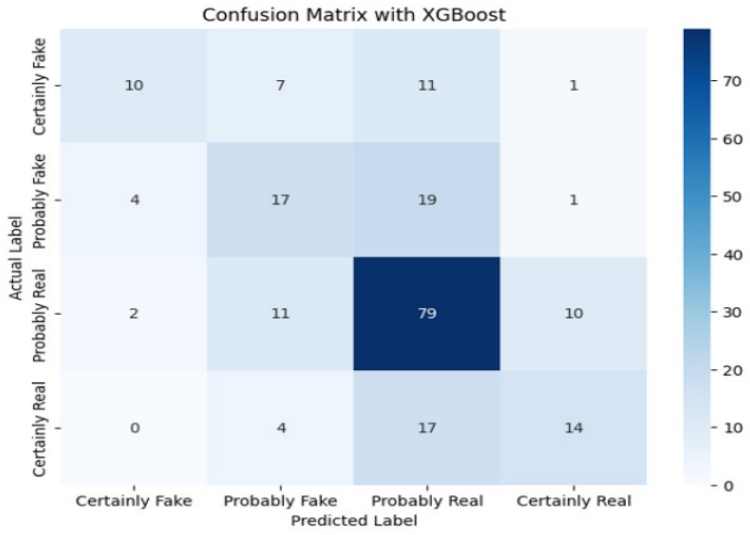


Figure 4.4: Confusion Matrix of XGBoost on Non-translated Texts

4.2.4 With Translated Text

Dataset Characteristics

The *With Translated Text* dataset consists of text entries that have been translated into a single, consistent language. This uniformity simplifies text preprocessing and model training, as the dataset no longer contains the linguistic diversity present in the *Without Translated Text* dataset. The translation process aims to standardize the text, making it easier for the models to focus on the content's meaning without being influenced by language-specific nuances.

Benefits of Translation in Text Classification

The primary benefit of working with a translated dataset is the reduction of linguistic variability. With all text entries in the same language, tokenization, normalization, and feature extraction become more straightforward. The models can be trained on a single language, eliminating the need for language-specific pre-processing and allowing for the use of monolingual word embeddings.

Furthermore, translation can help mitigate the issue of class imbalance across languages. By converting all text into a common language, the dataset becomes more homogeneous, reducing the likelihood of the model becoming biased towards a particular language. This uniformity can lead to more consistent and accurate

classification performance.

Bi-LSTM Model on *With Translated Text*

When applied to the *With Translated Text* dataset, the Bi-LSTM model benefits from the reduced linguistic complexity. With all text in a single language, the model can focus entirely on capturing the sequential dependencies and context within the text. The absence of language barriers allows the Bi-LSTM model to leverage its strengths in understanding the flow of sentences and the relationships between words.

Pre-trained monolingual word embeddings, such as those specific to the translated language, can further enhance the Bidirectional LSTM model's performance. These embeddings provide the model with a rich understanding of the language, allowing it to identify subtle cues and patterns that may indicate fake news. The consistency in language also allows for more effective use of advanced techniques like attention mechanisms, which can help the model focus on the most relevant parts of the text.

Overall, the Bi-LSTM model's ability to capture context and structure in a consistent language environment makes it a powerful tool for text classification in the *With Translated Text* dataset. The model can generalize more effectively across the dataset, leading to improved classification accuracy.

Table 4.3: Performance of Bi-LSTM on the Translated Text

	Precision	**Recall**	**F1-Score**	**Support**
Certainly Fake	0.33	0.34	0.34	29
Probably Fake	0.30	0.39	0.34	41
Probably Real	0.64	0.58	0.61	102
Certainly Real	0.50	0.46	0.48	35
Accuracy			0.49	207
Macro Avg.	0.44	0.44	0.44	207
Weighted Avg.	0.51	0.49	0.50	207

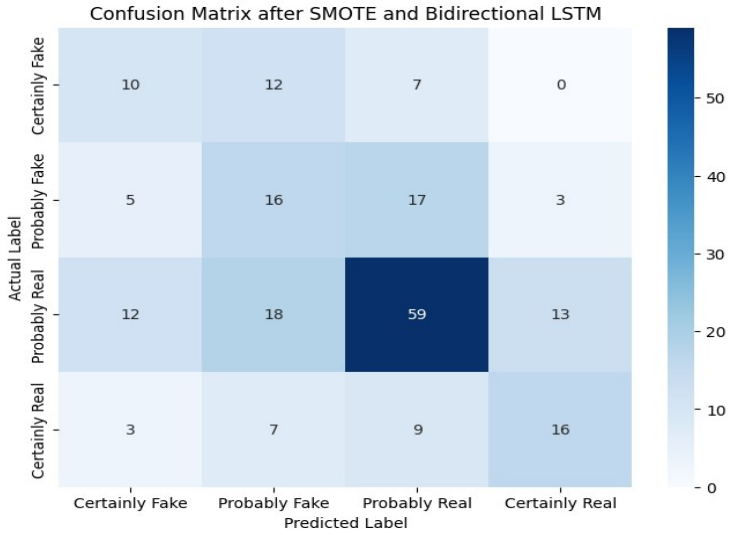


Figure 4.5: Confusion Matrix of Bi-LSTM on Translated Texts

XGBoost Model on *With Translated Text*

The XGBoost model also benefits from the uniformity provided by the *With Translated Text* dataset. With the text data standardized in a single language, feature extraction becomes more consistent, allowing XGBoost to focus on the most informative features. The absence of linguistic variability reduces the risk of feature noise, leading to more accurate and reliable classification performance.

XGBoost's strength in handling large and complex datasets makes it well-suited for this translated dataset. The algorithm can efficiently process the extracted features, identifying patterns and correlations that contribute to accurate classification. Additionally, the reduced complexity of the translated dataset allows for more effective hyperparameter tuning, further enhancing the model's performance.

By working with a consistent language, XGBoost can leverage its full potential, resulting in a robust and efficient text classification model. The uniformity of the *With Translated Text* dataset simplifies the training process, allowing XGBoost to achieve high levels of accuracy and reliability in detecting fake news.

In summary, both the LSTM and XGBoost models are capable of handling the challenges presented by the *Without Translated Text* and *With Translated Text* datasets. However, the uniformity of the translated dataset simplifies the

pre-processing and training process, allowing the models to focus more on the content's meaning rather than navigating linguistic complexities. While the LSTM model excels in capturing sequential dependencies and context, XGBoost's strength lies in its ability to handle structured data and identify patterns within the feature space. By comparing the performance of these models on both datasets, researchers can gain valuable insights into the impact of translation on text classification and develop more effective strategies for detecting fake news across different languages.

Table 4.4: Performance of XGBoost on the Translated Text

	Precision	**Recall**	**F1-Score**	**Support**
Certainly Fake	0.54	0.45	0.49	29
Probably Fake	0.40	0.46	0.43	41
Probably Real	0.68	0.77	0.72	102
Certainly Real	0.74	0.40	0.52	35
Accuracy			0.60	207
Macro Avg.	0.59	0.52	0.54	207
Weighted Avg.	0.61	0.60	0.60	207

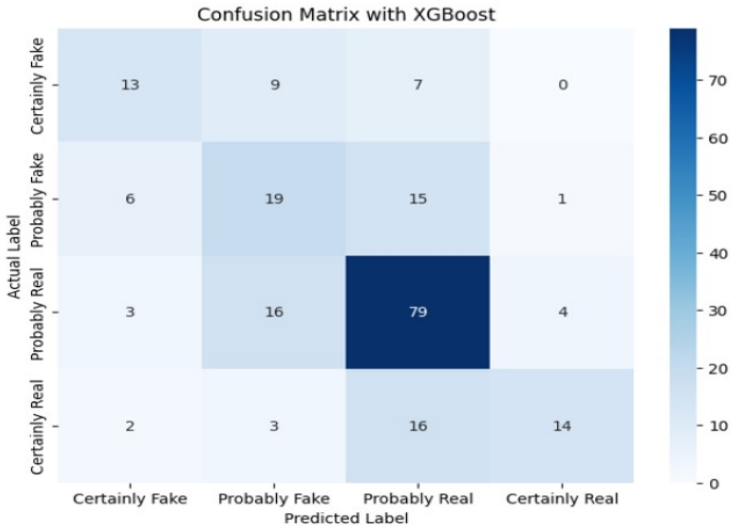


Figure 4.6: Confusion Matrix of XGBoost on Translated Texts

4.3 Training and Evaluation

4.3.1 Training Process

The training process is a critical phase in building effective models for text classification, particularly in the context of detecting fake news. For this task, two models are employed: Long Short-Term Memory (LSTM) and XGBoost (Extreme Gradient Boosting). Each model has a unique training process tailored to its architecture and strengths.

For the LSTM model, the training process involves feeding sequences of tokenized and normalized text into the model. The text is first transformed into numerical representations, often using word embeddings like Word2Vec, GloVe, or even contextual embeddings like BERT embeddings. These embeddings capture the semantic meaning of words and phrases, providing the LSTM with the ability to understand context over long sequences. During training, the LSTM learns to recognize patterns and dependencies within the text, adjusting its internal weights through backpropagation. The process involves iterating over the dataset multiple times (epochs), with each iteration refining the model's ability to predict the correct class labels for the given text sequences.

For the XGBoost model, the training process is somewhat different. Here, the text data is first converted into numerical features using techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) or word embeddings. Once the data is transformed into a suitable format, XGBoost constructs an ensemble of decision trees. Each tree is built sequentially, where each new tree attempts to correct the errors made by the previous ones. This iterative process continues until the model reaches a predefined number of trees or until further improvement becomes marginal. The model uses gradient boosting, a powerful technique that minimizes the loss function by optimizing each tree's contribution to the final prediction.

Both models require careful tuning of hyperparameters to achieve optimal

performance. For LSTM, parameters such as the number of hidden units, learning rate, and dropout rate are crucial. In contrast, XGBoost requires tuning of parameters like the learning rate, maximum depth of the trees, and the number of estimators. The training process is computationally intensive, especially for LSTM, which involves processing long sequences and maintaining memory over time. However, the effort is justified by the model's ability to capture intricate patterns within the text.

4.3.2 Evaluation Metrics

After training, the performance of the models is assessed using a set of evaluation metrics that provide insights into their effectiveness. Common metrics for text classification tasks include accuracy, precision, recall, F1-score[5], and AUC-ROC (Area Under the Receiver Operating Characteristic Curve)[6].

Accuracy[7] measures the overall correctness of the model, representing the proportion of correctly classified instances out of the total instances. While accuracy is a straightforward metric, it may not be sufficient, especially in cases where class imbalances exist. For instance, if the dataset has more instances of *Certainly Real* content, a model might achieve high accuracy by predominantly predicting this class, even if it performs poorly on other classes.

Precision[8] and **Recall**[9] offer a more nuanced view of the model's performance. Precision measures the proportion of true positive predictions among all positive predictions made by the model, while recall measures the proportion of true positive predictions out of the actual positives in the dataset. In the context of fake news detection, high precision means the model is good at identifying true instances of fake news, whereas high recall indicates that the model is successful at capturing most of the fake news instances.

[5]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html
[6]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html
[7]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html
[8]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html
[9]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html

F1-score is the harmonic mean of precision and recall, providing a balanced measure that considers both false positives and false negatives. It is particularly useful when dealing with imbalanced datasets, as it highlights the model's ability to correctly identify both positive and negative classes.

AUC-ROC is another important metric, especially when the model's output is probabilistic rather than categorical. The ROC curve plots the true positive rate against the false positive rate at various threshold settings, while the AUC (Area Under the Curve) provides a single scalar value that summarizes the model's performance across all thresholds. A higher AUC indicates better model performance, as it shows a better trade-off between true positives and false positives.

Chapter 5

IMAGE CLASSIFICATION

– Analyzing Visual Cues to Detect and Disarm Deceptive Content

Image classification plays a crucial role in the broader task of fake news detection by analyzing visual content associated with news articles or social media posts. The process involves pre-processing images, applying different models for raw image classification, and leveraging generated captions to improve classification accuracy. Below is an in-depth exploration of each aspect based on the provided datasets.

5.1 Image Preprocessing

Image pre-processing is a vital step that ensures raw images are ready for input into machine learning models. Proper pre-processing can significantly enhance the performance of image classification models by standardizing the input data and increasing the diversity of training samples.

5.1.1 Image Augmentation Techniques

Image augmentation is a method used to artificially expand the size of the training dataset by creating modified versions of the original images. In the *Raw Image Classification* dataset, augmentation can be particularly useful for enhancing model robustness and preventing overfitting. Common augmentation techniques include:

- **Rotation**:[1] Images are rotated at various angles to create new perspectives, helping the model recognize objects regardless of their orientation.

[1]https://medium.com/analytics-vidhya/how-to-auto-rotate-the-image-using-deep-learning-c34b2e0e157d

- **Flipping**:[2] Horizontal or vertical flipping of images simulates different viewing angles, improving the model's generalization capabilities.

- **Zooming**:[3] Images are zoomed in or out, creating variations in scale and helping the model handle different object sizes.

- **Shearing**:[4] Applying a shear transformation skews the image along an axis, which can help the model learn to recognize objects even when they appear distorted.

- **Brightness Adjustment**:[5] Modifying the brightness of images simulates different lighting conditions, making the model more robust to changes in illumination.

These techniques are applied randomly during training to create a more diverse dataset, helping the model generalize better to unseen data.

5.1.2 Resizing and Normalization

Resizing and normalization are crucial pre-processing steps that standardize the image dimensions and pixel values, ensuring consistency across the dataset.

- **Resizing**: All images are resized to a uniform dimension (e.g., 224x224 pixels) to meet the input requirements of deep learning models like ResNet-50 and MobileNetV2. This standardization simplifies the model's task by ensuring that each image has the same amount of data.

- **Normalization**: After resizing, pixel values are normalized to a standard range, typically between 0 and 1. Normalization ensures that the input data has consistent scale, which helps in speeding up the training process and improving model convergence.

[2]https://www.w3schools.com/howto/howto_css_flip_image.asp
[3]https://www.manning.com/books/deep-learning-with-python
[4]https://www.geeksforgeeks.org/wand-shear-function-python/
[5]https://www.manning.com/books/deep-learning-with-python

These pre-processing steps ensure that the images are prepared for efficient and effective analysis by the classification models.

5.2 Module 1: Raw Image Classification

This module focuses on the direct classification of raw images without any accompanying textual data. The *Raw Image Classification* dataset provides the foundation for this task, containing images labeled with their respective classes.

5.2.1 LSTM Model

The Long Short-Term Memory (LSTM) model is traditionally used for sequential data but can be adapted for image classification in specific contexts. In the case of raw image classification, features extracted from images through convolutional layers can be processed sequentially by the LSTM. This approach is beneficial when the order of visual features plays a significant role in the classification decision.

However, LSTM may not be as effective as CNN-based models for raw image classification since it is not inherently designed to handle spatial data. Nonetheless, in scenarios where temporal or sequential patterns within the visual data are crucial, LSTM can be a valuable tool.

Table 5.1: Performance of LSTM Raw Image Classification Task

	Precision	**Recall**	**F1-Score**	**Support**
Certainly Fake	0.61	0.19	0.29	177
Probably Fake	0.59	0.26	0.36	246
Probably Real	0.55	0.94	0.69	524
Certainly Real	0.77	0.34	0.47	211
Accuracy			0.57	1158
Macro Avg.	0.63	0.43	0.45	1158
Weighted Avg.	0.61	0.57	0.52	1158



Figure 5.1: Confusion Matrix of LSTM on Raw Image Classification Task

5.2.2 ResNet-50 Model

ResNet-50, a deep convolutional neural network, is highly effective for image classification tasks. The key feature of ResNet-50 is its use of residual connections, which help mitigate the vanishing gradient problem and allow the network to be deeper, thereby improving its ability to learn complex patterns in the data.

For raw image classification in the provided dataset, ResNet-50 is ideal due to its ability to capture intricate visual features and patterns across multiple layers. This makes it particularly effective in detecting subtle anomalies or manipulations in images associated with fake news.

Table 5.2: Performance of ResNet-50 on Raw Image Classification Task

	Precision	**Recall**	**F1-Score**	**Support**
Certainly Fake	0.31	0.46	0.37	24
Probably Fake	0.33	0.33	0.33	46
Probably Real	0.68	0.54	0.60	101
Certainly Real	0.38	0.47	0.42	36
Accuracy			0.47	207
Macro Avg.	0.42	0.45	0.43	207
Weighted Avg.	0.51	0.47	0.48	207

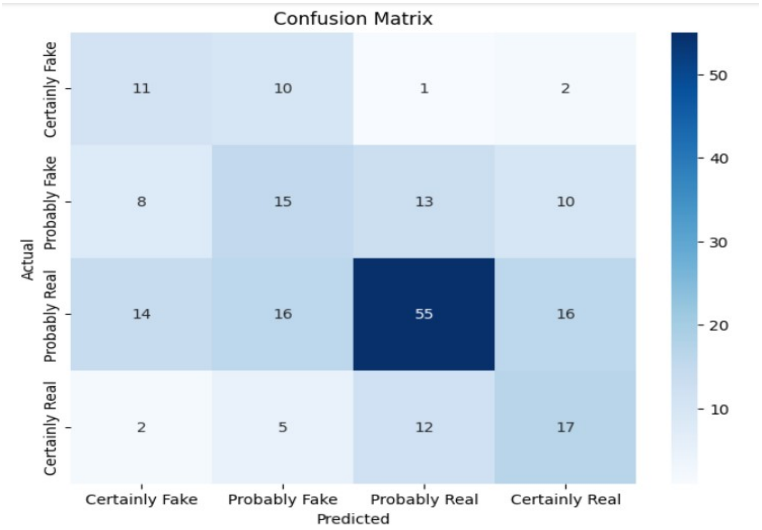


Figure 5.2: Confusion Matrix of ResNet-50 on Raw Image Classification Task

5.2.3 MobileNetV2

MobileNetV2 is another convolutional neural network designed for efficient image classification, particularly in environments with limited computational resources. It uses depthwise separable convolutions to reduce the number of parameters and computation required, making it faster and more efficient than traditional CNNs like ResNet-50.

In the context of raw image classification, MobileNetV2 offers a balance between accuracy and efficiency. It is particularly useful when deploying models on devices with limited processing power or when a faster inference time is required. While it may not achieve the same level of accuracy as ResNet-50, it is a viable option for scenarios where computational resources are a concern.

Table 5.3: Performance of MobileNetv2 on Raw Image Classification Task

	Precision	**Recall**	**F1-Score**	**Support**
Certainly Fake	0.83	0.75	0.78	95
Probably Fake	0.58	0.78	0.67	95
Probably Real	0.68	0.55	0.61	95
Certainly Real	0.78	0.75	0.76	95
Accuracy			0.71	380
Macro Avg.	0.72	0.71	0.71	380
Weighted Avg.	0.72	0.71	0.71	380

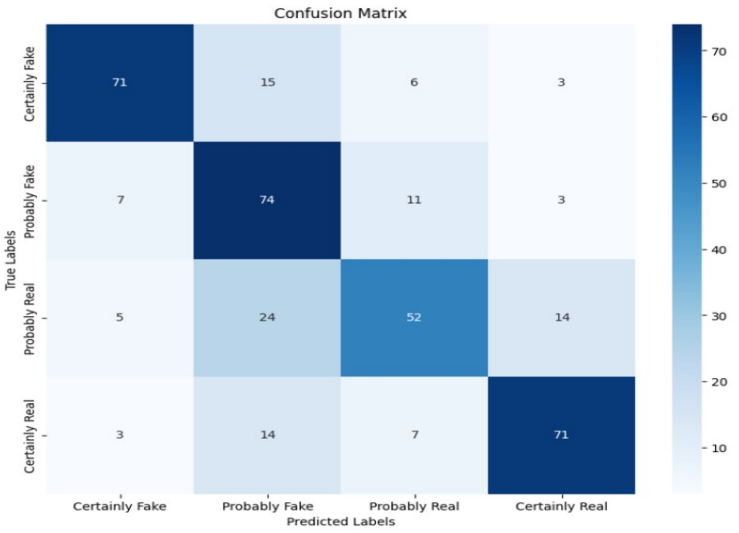


Figure 5.3: Confusion Matrix of MobileNetv2 on Raw Image Classification Task

5.3 Module 2: Generated Captions-Based Image Classification

This module integrates NLP techniques with image classification by leveraging captions generated from images. The *Generated Captions* dataset includes textual descriptions of images, which are used to enhance classification accuracy.

5.3.1 LSTM Model

For generated captions-based image classification, the LSTM model processes the sequences of words in the captions. LSTM's ability to capture context and sequential dependencies makes it well-suited for understanding the meaning behind the captions, which can provide valuable insights into the content of the images.

By analyzing the sequence of words in the captions, the LSTM model can identify linguistic cues that may suggest whether the image is real or fake. This approach is particularly useful when the visual content alone is insufficient to make a definitive classification, and the accompanying text provides additional context.

Table 5.4: Performance of LSTM on Generated Captions-Based Image Classification Task

	Precision	**Recall**	**F1-Score**	**Support**
Certainly Fake	0.59	0.42	0.49	143
Probably Fake	0.74	0.33	0.46	215
Probably Real	0.69	0.90	0.78	475
Certainly Real	0.62	0.66	0.64	198
Accuracy			0.67	1031
Macro Avg.	0.66	0.58	0.59	1031
Weighted Avg.	0.67	0.67	0.65	1031

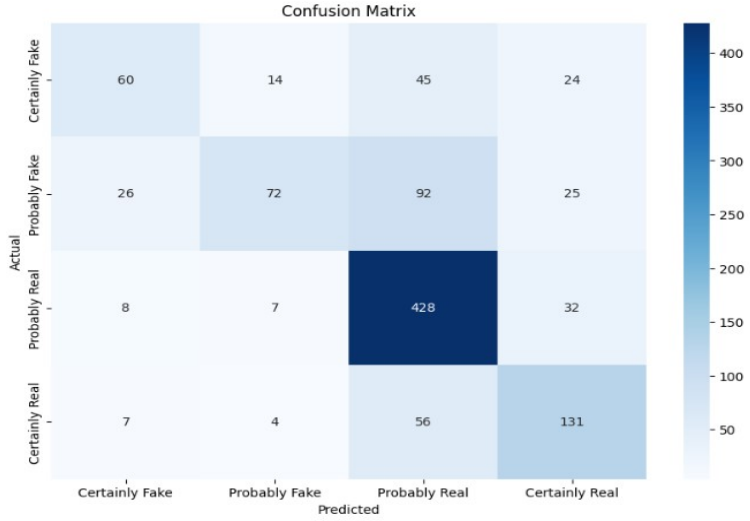


Figure 5.4: Confusion Matrix of LSTM on Generated Captions-based Image Classification Task

5.3.2 BERT Model

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art NLP model that processes generated captions by considering the bidirectional context of words within the sentence. Unlike LSTM, which processes sequences in one direction, BERT captures the relationships between words in both directions, allowing for a more nuanced understanding of the text.

For generated captions-based image classification, BERT's deep understanding of language context makes it highly effective. It can detect subtle cues in the language that might indicate manipulation or inconsistency between the image and its description, thereby enhancing the accuracy of the classification.

Table 5.5: Performance of BERT on Generated Captions-Based Image Classification Task

	Precision	**Recall**	**F1-Score**	**Support**
Certainly Fake	0.28	0.24	0.26	29
Probably Fake	0.24	0.16	0.19	43
Probably Real	0.59	0.73	0.65	95
Certainly Real	0.53	0.47	0.50	40
Accuracy			0.49	207
Macro Avg.	0.41	0.40	0.40	207
Weighted Avg.	0.46	0.49	0.47	207

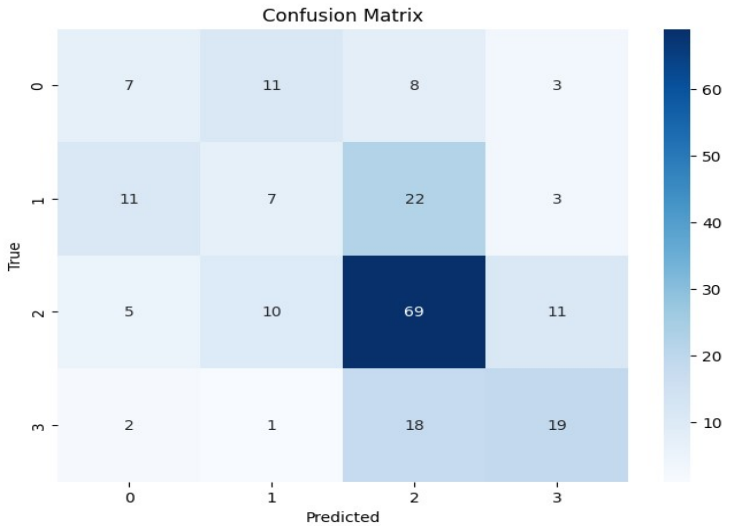


Figure 5.5: Confusion Matrix of BERT on Generated Captions-based Image Classification Task

5.3.3 XGBoost Model

XGBoost (Extreme Gradient Boosting) is applied to the numerical features extracted from the generated captions. These features could include TF-IDF vectors, word embeddings, or other text-derived attributes. XGBoost excels at structured data classification and can efficiently process these textual features to make predictions about the image's authenticity.

In this context, XGBoost is particularly useful for handling large and complex datasets, where it can leverage its gradient boosting framework to improve classification accuracy. Its ability to handle imbalanced data and prevent overfitting makes it a reliable choice for text-based image classification.

Table 5.6: Performance of XGBoost on Generated Captions-Based Image Classification Task

	Precision	Recall	F1-Score	Support
Certainly Fake	0.69	0.75	0.72	138
Probably Fake	0.69	0.57	0.62	155
Probably Real	0.52	0.65	0.58	135
Certainly Real	0.81	0.71	0.76	142
Accuracy			0.67	570
Macro Avg.	0.68	0.67	0.67	570
Weighted Avg.	0.68	0.67	0.67	570

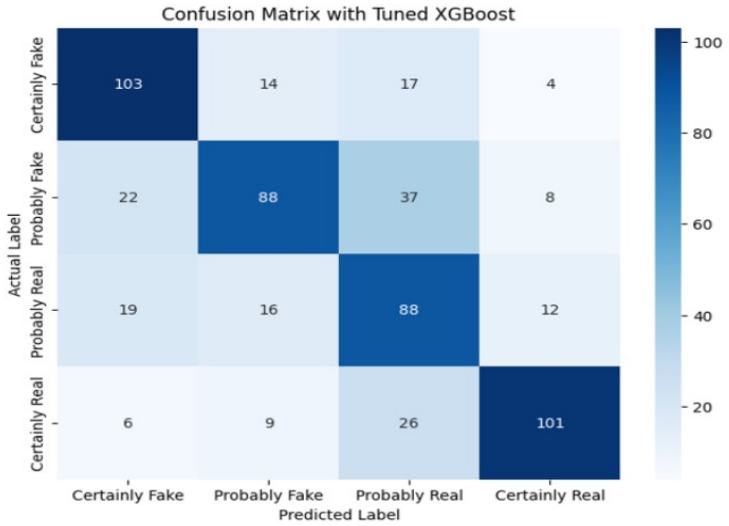


Figure 5.6: Confusion Matrix of XGBoost on Generated Captions-based Image Classification Task

5.3.4 Logistic Regression Model

Logistic Regression[6] is a simpler, yet effective, model for binary classification tasks. In the context of generated captions-based image classification, Logistic Regression can be applied to the numerical features derived from the captions to predict the likelihood of an image being real or fake.

While not as powerful as deep learning models, Logistic Regression offers the advantage of simplicity and interpretability. It is particularly useful in scenarios where the generated captions are well-structured and the relationship between the text features and the classification outcome is relatively straightforward.

[6]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Table 5.7: Performance of Logistic Regression on Generated Captions-Based Image Classification Task

	Precision	Recall	F1-Score	Support
Certainly Fake	0.29	0.34	0.32	29
Probably Fake	0.50	0.47	0.48	43
Probably Real	0.64	0.63	0.63	95
Certainly Real	0.36	0.35	0.35	40
Accuracy			0.50	207
Macro Avg.	0.45	0.45	0.45	207
Weighted Avg.	0.51	0.50	0.50	207

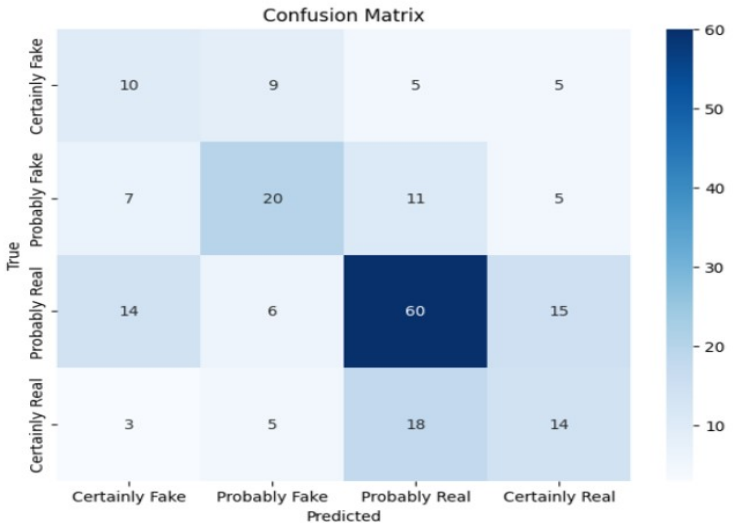


Figure 5.7: Confusion Matrix of Logistic Regression on Captions-based Image Classification Task

5.4 Training and Evaluation

5.4.1 Training Process

The training process for image classification models involves feeding pre-processed images (and generated captions, where applicable) into the models and optimizing their parameters through backpropagation and gradient descent[7]. Each model's architecture—whether it's a CNN like ResNet-50 or an NLP model like BERT—dictates how the training process is carried out.

For CNN-based models like ResNet-50 and MobileNetV2, the training process focuses on learning spatial features from images. The models are trained on batches of images, where each batch is passed through multiple layers of convo-

[7]https://scikit-learn.org/stable/modules/sgd.html

lution, pooling, and activation functions. The objective is to minimize the loss function, which measures the difference between the predicted and actual labels.

For LSTM and BERT models used in generated captions-based classification, the training process involves processing sequences of words and learning the relationships between them. The models are trained to predict the correct classification label based on the content of the captions, adjusting their weights to improve accuracy.

5.4.2 Evaluation Metrics

Evaluation metrics are crucial for assessing the performance of the models. Common metrics include:

1. **Accuracy**

   Accuracy measures the proportion of correctly classified instances out of the total number of instances.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{5.1}$$

   Alternatively, in terms of *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP), and *False Negatives* (FN):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5.2}$$

   where:

   - **TP (True Positives)**: Number of correctly predicted positive cases.
   - **TN (True Negatives)**: Number of correctly predicted negative cases.
   - **FP (False Positives)**: Number of incorrectly predicted positive cases.
   - **FN (False Negatives)**: Number of incorrectly predicted negative cases.

2. **Precision**

   Precision is the ratio of correctly predicted positive observations to the total predicted positives. It is a measure of the model's accuracy when it predicts the positive class.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5.3}$$

3. **Recall**

   Recall is the ratio of correctly predicted positive observations to all observations in the actual class. It indicates how well the model captures positive cases.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.4}$$

4. **F1-Score**

   F1-Score is the harmonic mean of Precision and Recall. It provides a balanced measure when dealing with imbalanced datasets.

$$\text{F1-Score} = 2 \text{ x } \frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}} \tag{5.5}$$

   This metric is particularly useful when you need to balance between Precision and Recall.

5. **AUC-ROC (Area Under the Receiver Operating Characteristic Curve)**

   AUC-ROC is used to measure the performance of a classification model by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The AUC (Area Under the Curve) represents the likelihood that the model will rank a randomly chosen positive instance higher than a randomly chosen negative one.

- **True Positive Rate (TPR) or Recall**:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.6}$$

- **False Positive Rate (FPR)**:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{5.7}$$

The ROC curve is generated by plotting TPR vs. FPR at different thresholds, and the area under this curve (AUC) provides a single scalar value summarizing the model's performance. A higher AUC indicates a better performing model.

Chapter 6

MULTIMODAL CLASSIFICATION

– Fusing Text and Image Insights for Robust Fake News Identification

Multimodal classification integrates data from different modalities—such as text and images—to enhance the predictive capabilities of machine learning models. This approach leverages the complementary strengths of each modality, leading to more robust and accurate models for complex tasks like fake news detection.

6.1 Combining Text and Image Data

Combining text and image data is a powerful approach in multimodal classification. Text data provides contextual and descriptive information, while images offer visual cues that can significantly impact the model's understanding of the content. Integrating these data types can help models make more informed decisions by considering both the visual and linguistic aspects of the content.

6.2 Model Architectures

Model Architectures serve as the foundational designs for implementing machine learning algorithms, guiding how data flows through the system to produce predictions. In this study, we explore various architectures tailored to different aspects of multimodal classification, including text, image, and combined text-image data. Each architecture is chosen for its ability to capture and process specific features from the dataset. The models range from traditional methods to advanced deep learning techniques, ensuring a comprehensive approach to understanding and classifying the data. These architectures are crucial in optimizing performance, particularly in the context of small-scale datasets.

6.2.1 LSTM

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) designed to handle sequential data. They are particularly effective in text-based tasks where understanding the context and dependencies between words is crucial. In multimodal classification, LSTM networks process textual data, capturing the flow and meaning of sentences to inform the model's predictions.

6.2.2 ResNet-50

ResNet-50 is a deep convolutional neural network (CNN) architecture known for its ability to handle complex image classification tasks. It introduces residual connections that allow for the training of very deep networks, which are essential for capturing intricate patterns in images. In multimodal classification, ResNet-50 extracts visual features from images that are then combined with textual features to enhance the model's accuracy.

6.2.3 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based model that excels in natural language processing tasks. It processes text by considering the context of each word in both directions, allowing it to capture subtle nuances in language. BERT is used in multimodal classification to extract deep, contextually rich features from text, which can then be combined with visual data for comprehensive analysis.

6.2.4 Logistic Regression

Logistic Regression is a straightforward yet effective model for binary classification. It models the probability of a particular outcome based on input features. In the context of multimodal classification, Logistic Regression can be applied to combined feature vectors derived from text and image data, offering a simple and

interpretable way to predict whether content is real or fake.

6.2.5 ResNet + LSTM

The ResNet + LSTM architecture combines convolutional neural networks (CNNs) with recurrent neural networks (RNNs) to handle both visual and textual data. ResNet-50 is used to extract features from images, while LSTM processes text sequences. The features from both networks are merged to make a final prediction, making this architecture particularly effective in scenarios where both visual and textual contexts are critical for accurate classification.

6.2.6 ResNet + Bidirectional LSTM

ResNet + Bidirectional LSTM extends the previous architecture by incorporating a Bidirectional LSTM, which processes text in both forward and backward directions. This allows the model to capture a fuller context of the text, enhancing its understanding. When combined with ResNet's visual feature extraction, this architecture is well-suited for tasks where understanding both the sequence and visual content is essential.

6.3 Experiments

6.3.1 Raw Image + Raw Text

BERT

In this experiment, BERT is utilized to process raw text data. Similar to its use in other experiments, BERT's role is to extract deep, contextually-rich features from the text. The ability to understand the full context of each word in the text allows BERT to generate feature vectors that are highly informative and can be effectively used for further processing and classification tasks.

ResNet + LSTM

Table 6.1: Performance of BERT on Raw Image and Raw Text Data

	Precision	**Recall**	**F1-Score**	**Support**
Certainly Fake	0.41	0.72	0.53	29
Probably Fake	0.42	0.39	0.41	41
Probably Real	0.66	0.54	0.59	102
Certainly Real	0.54	0.54	0.54	35
Accuracy			0.54	207
Macro Avg.	0.51	0.55	0.52	207
Weighted Avg.	0.56	0.54	0.54	207

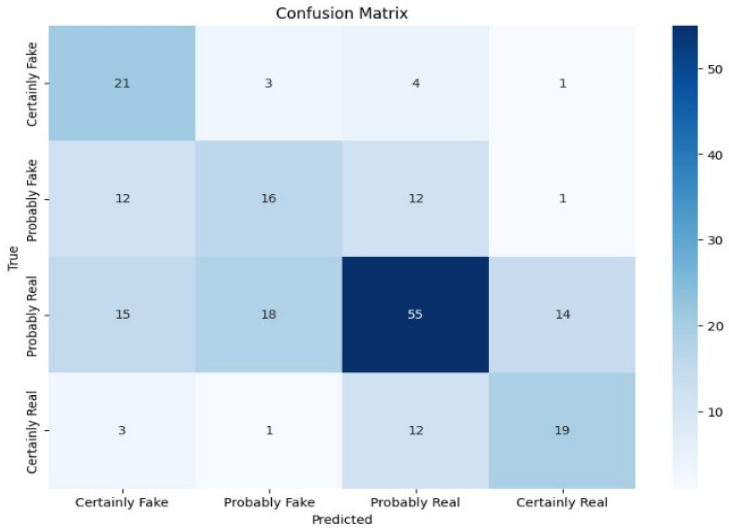


Figure 6.1: Confusion Matrix of BERT performing on Raw Image and Raw Text Data

The ResNet + LSTM architecture is employed to process the visual and textual data together. ResNet-50 is used to extract features from raw images, capturing the essential visual patterns and details. These visual features are then combined with the textual features extracted by BERT and passed to an LSTM network. The LSTM component processes the sequential text data, leveraging its memory capabilities to understand the sequence and relationships between words. The combined output of ResNet and LSTM provides a comprehensive feature set that is used to classify the content, considering both visual and textual information.

Table 6.2: Performance of ResNet and LSTM on Raw Image and Raw Text Data

	Precision	Recall	F1-Score	Support
Certainly Fake	0.94	0.77	0.85	79
Probably Fake	0.61	0.87	0.71	98
Probably Real	0.68	0.51	0.58	92
Certainly Real	0.87	0.82	0.85	90
Accuracy			0.74	359
Macro Avg.	0.77	0.74	0.75	359
Weighted Avg.	0.77	0.74	0.74	359

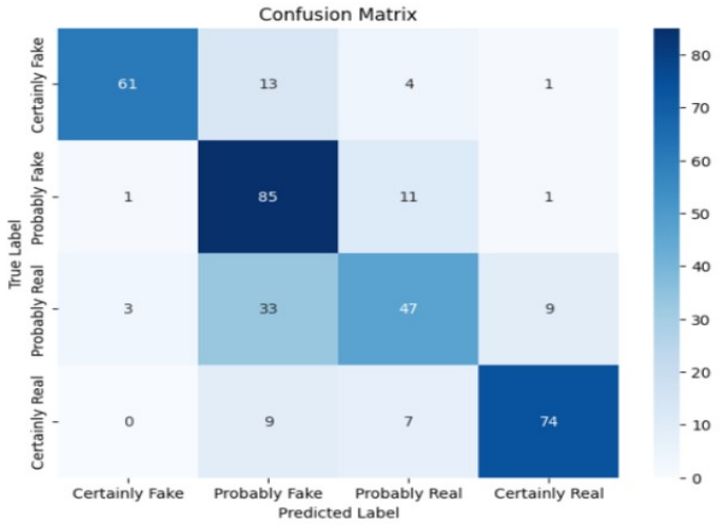


Figure 6.2: Confusion Matrix of ResNet with LSTM performing on Raw Image and Raw Text Data

6.3.2 Raw Image + Translated Text

BERT

In this experiment, BERT processes translated text, just as it did in the previous experiment involving translated text. The translated text is standardized, making it easier for BERT to extract consistent and meaningful features. The bidirectional nature of BERT ensures that the model captures all relevant contextual information, regardless of the original language of the text. These features are then combined with visual features from the images.

Table 6.3: Performance of BERT on Raw Image and Translated Text Data

	Precision	**Recall**	**F1-Score**	**Support**
Certainly Fake	0.50	0.38	0.43	29
Probably Fake	0.78	0.17	0.28	41
Probably Real	0.57	0.91	0.70	102
Certainly Real	0.71	0.29	0.41	35
Accuracy			0.58	207
Macro Avg.	0.64	0.44	0.46	207
Weighted Avg.	0.63	0.58	0.53	207

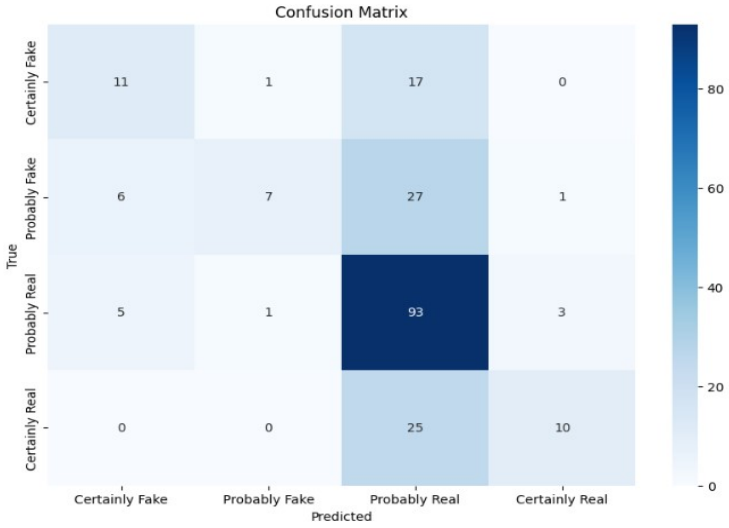


Figure 6.3: Confusion Matrix of BERT performing on Raw Image and Translated Text Data

ResNet + Bidirectional LSTM

The ResNet + Bidirectional LSTM architecture extends the combination of ResNet-50 and LSTM by incorporating a Bidirectional LSTM. ResNet-50 extracts visual features from raw images, while the Bidirectional LSTM processes the translated text. Unlike a regular LSTM, the Bidirectional LSTM reads the text both forwards and backwards, providing a more comprehensive understanding of the text's context. This dual-direction processing enhances the model's ability to capture nuances in the translated text, which, when combined with the visual features, leads to more accurate classification of the content.

These explanations highlight how each model operates within the context of multimodal classification experiments. BERT is used for its advanced text processing capabilities, ResNet-50 for its visual feature extraction, and LSTM

variants for handling sequential text data. Logistic Regression is employed for its simplicity and interpretability, providing a straightforward classification mechanism. Together, these models and techniques offer a powerful toolkit for analyzing and classifying multimodal data, particularly in tasks like fake news detection where both text and images play a critical role.

Table 6.4: Performance of ResNet and Bi-LSTM on Raw Image and Translated Text Data

	Precision	**Recall**	**F1-Score**	**Support**
Certainly Fake	0.84	0.80	0.82	79
Probably Fake	0.67	0.81	0.73	98
Probably Real	0.67	0.50	0.57	92
Certainly Real	0.80	0.87	0.83	90
Accuracy			0.74	359
Macro Avg.	0.75	0.74	0.74	359
Weighted Avg.	0.74	0.74	0.74	359

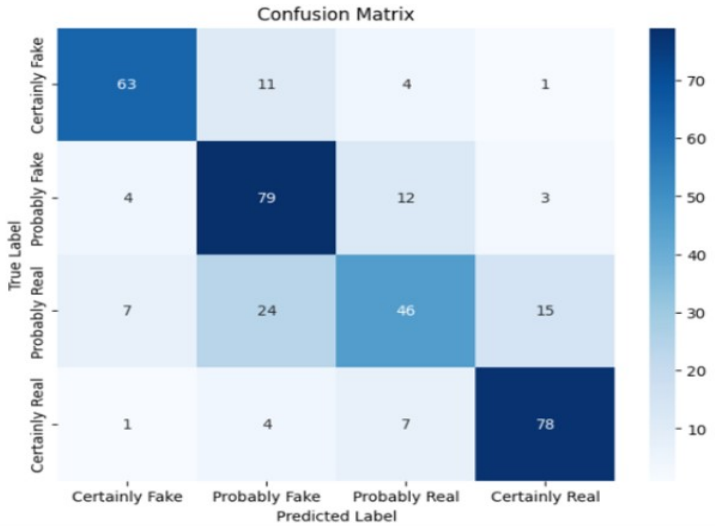


Figure 6.4: Confusion Matrix of ResNet and BiLSTM performing on Raw Image and Translated Text Data

6.3.3 Image Caption + Raw Text

BERT

In this experiment, BERT (Bidirectional Encoder Representations from Transformers) is employed to process the raw text data accompanying the image captions. BERT's architecture allows it to understand the context of each word in

a sentence by looking at its neighboring words both before and after it. This bidirectional approach enables BERT to capture the subtleties and nuances of language, making it particularly effective for text analysis tasks. When applied to raw text in this experiment, BERT extracts deep contextual features that represent the meaning of the text in a form that can be used for classification.

Table 6.5: Performance of BERT on Image Captions and Raw Text data

	Precision	**Recall**	**F1-Score**	**Support**
Certainly Fake	0.70	0.24	0.36	29
Probably Fake	0.33	0.07	0.12	41
Probably Real	0.54	0.98	0.70	102
Certainly Real	1.00	0.14	0.25	35
Accuracy			0.56	207
Macro Avg.	0.64	0.36	0.36	207
Weighted Avg.	0.60	0.56	0.46	207

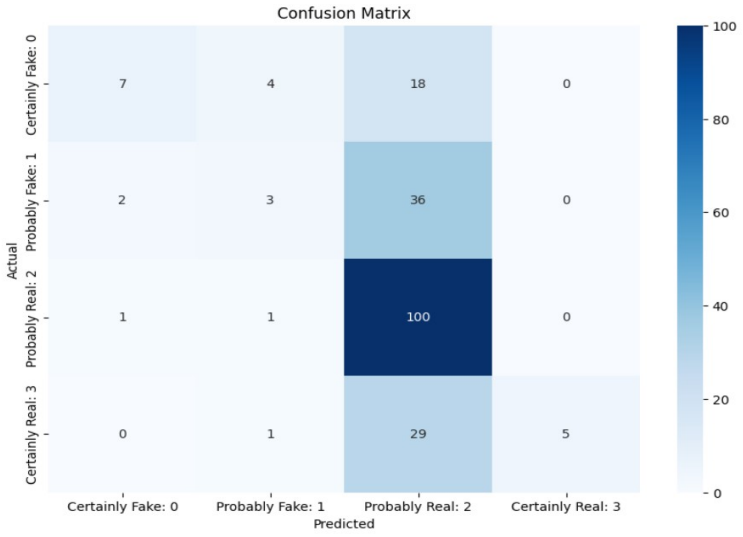


Figure 6.5: Confusion Matrix of BERT performing on Image Captions and Raw Text data

Logistic Regression

After BERT has extracted the text features, Logistic Regression is used to combine these features with those derived from the image captions. Logistic Regression is a linear model that is well-suited for binary classification tasks. It estimates the probability that a given input (in this case, the combined features from BERT and the image captions) belongs to a particular class (such as "fake"

or "real"). Logistic Regression's simplicity and interpretability make it a good choice for situations where the relationship between the features and the target variable is relatively straightforward.

Table 6.6: Performance of Logistic Regression on Image Captions and Raw Text data

	Precision	**Recall**	**F1-Score**	**Support**
Certainly Fake	0.39	0.45	0.42	29
Probably Fake	0.35	0.40	0.37	43
Probably Real	0.63	0.62	0.62	95
Certainly Real	0.55	0.42	0.48	40
Accuracy			0.51	207
Macro Avg.	0.48	0.47	0.47	207
Weighted Avg.	0.52	0.51	0.51	207

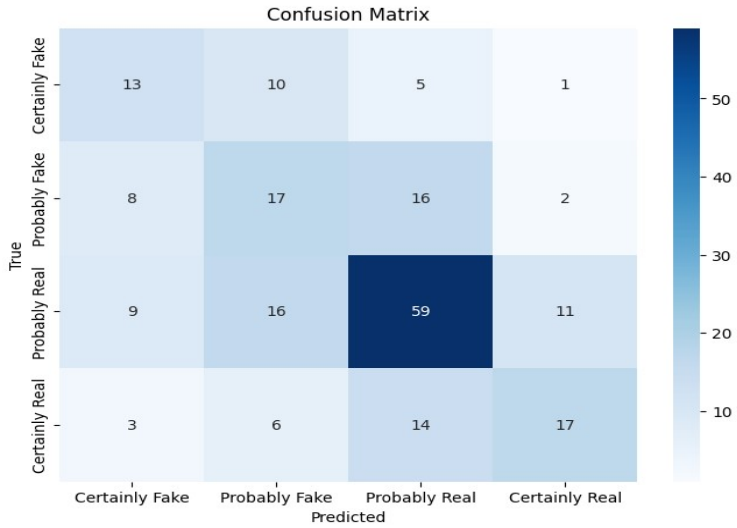


Figure 6.6: Confusion Matrix of Logistic Regression performing on Image Captions and Raw Text Data

6.3.4 Image Caption + Translated Text

BERT

In this scenario, BERT is again utilized to process text data, but this time it is applied to translated text rather than raw text. The translation process standardizes the language, reducing variability and potentially improving the consistency of the features BERT extracts. BERT's ability to understand context remains

key, as it processes the translated text to capture essential meanings and nuances that could indicate whether the content is real or fake. This step is critical in multilingual settings where content might be in different languages originally.

Table 6.7: Performance of BERT on Image Captions and Translated Texts

	Precision	**Recall**	**F1-Score**	**Support**
Certainly Fake	0.67	0.28	0.39	29
Probably Fake	0.75	0.15	0.24	41
Probably Real	0.57	0.75	0.65	102
Certainly Real	0.42	0.63	0.50	35
Accuracy			0.55	207
Macro Avg.	0.60	0.45	0.45	207
Weighted Avg.	0.60	0.55	0.51	207

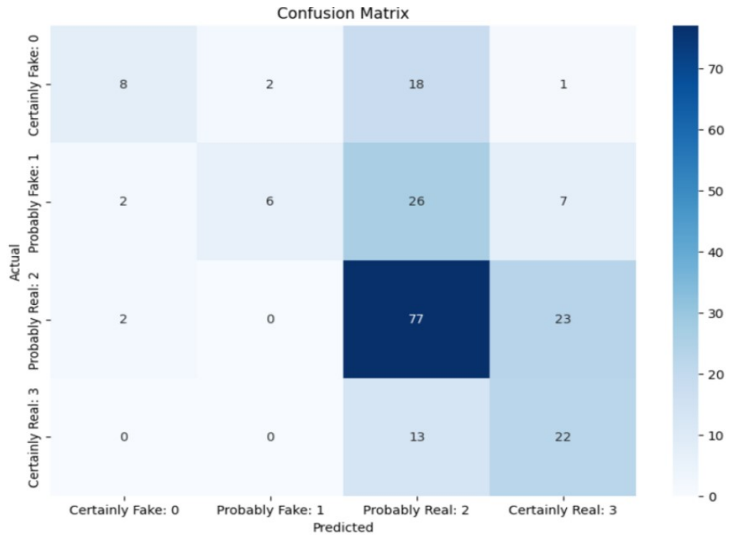


Figure 6.7: Confusion Matrix of BERT performing on Image Captions and Translated Texts Data

Logistic Regression

Logistic Regression is used here to combine the features extracted by BERT from the translated text with those derived from the image captions. As in the previous experiment, Logistic Regression is tasked with determining the likelihood that the combined features indicate a particular class. By working with standardized, translated text, Logistic Regression may achieve more consistent results, as it relies on features that are less likely to be influenced by linguistic differences.

Table 6.8: Performance of Logistic Regression on Image Captions and Translated Texts

	Precision	Recall	F1-Score	Support
Certainly Fake	0.45	0.53	0.49	43
Probably Fake	0.40	0.42	0.41	65
Probably Real	0.57	0.56	0.56	143
Certainly Real	0.50	0.42	0.46	59
Accuracy			0.50	310
Macro Avg.	0.48	0.48	0.48	310
Weighted Avg.	0.50	0.50	0.50	310

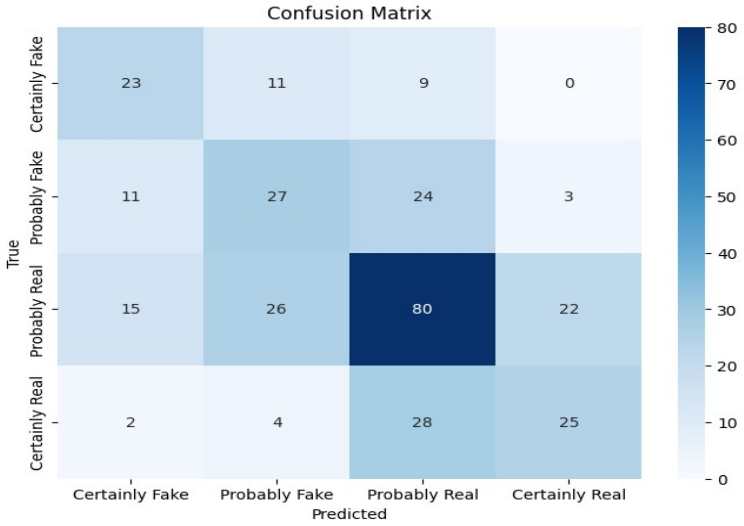


Figure 6.8: Confusion Matrix of Logistic Regression performing on Image Captions and Translated Texts Data

6.3.5 Raw Image + Generated Captions Classification

ResNet-50

For this experiment, ResNet-50 is used to process raw image data. ResNet-50 is a deep convolutional neural network (CNN) that excels at image classification tasks due to its ability to learn complex visual patterns through multiple layers of convolutions. The architecture of ResNet-50 includes residual connections that help the network maintain performance as it goes deeper, making it particularly effective for analyzing high-resolution images with many details. In this experiment, ResNet-50 extracts visual features from the raw images that can be used to classify the content as real or fake.

Generated Captions

Generated captions are used to provide additional context to the visual data. While the specific method of integrating these captions may vary, typically, the captions are used to complement the visual features extracted by ResNet-50. This combination allows the model to consider both the image content and the associated textual descriptions when making a classification decision.

Table 6.9: Performance of ResNet-50 on Raw Image and Generated Captions Data

	Precision	**Recall**	**F1-Score**	**Support**
Certainly Fake	0.37	0.31	0.34	32
Probably Fake	0.33	0.33	0.33	43
Probably Real	0.55	0.68	0.61	79
Certainly Real	0.58	0.37	0.45	38
Accuracy			0.48	192
Macro Avg.	0.46	0.42	0.43	192
Weighted Avg.	0.48	0.48	0.47	192

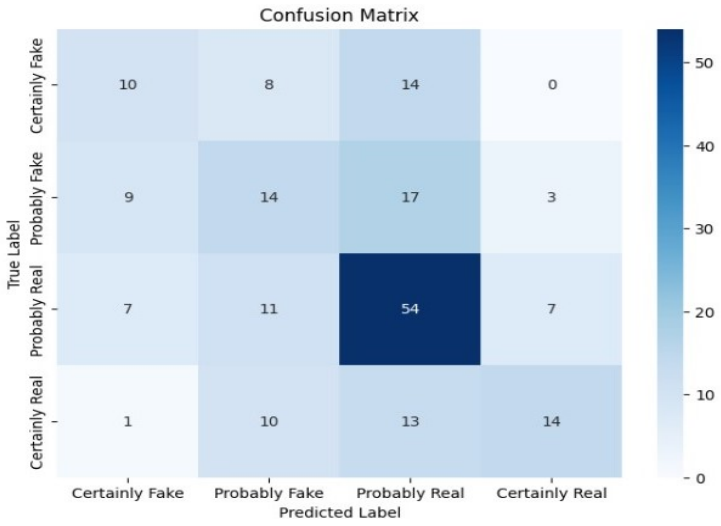


Figure 6.9: Confusion Matrix of ResNet-50 performing on Raw Image and Generated Captions Data

6.4 Training and Evaluation

6.4.1 Training Process

The training process for multimodal classification models involves optimizing the combined architectures (such as BERT, ResNet, and LSTM) on the respective datasets. Training typically includes backpropagation and gradient descent to minimize the loss function. Each modality (text and image) contributes its features to the final classification decision. The training process often involves multiple epochs and hyperparameter tuning to achieve the best performance.

6.4.2 Evaluation Metrics

Evaluation metrics for multimodal classification include accuracy, precision, recall, F1-score, and AUC-ROC. These metrics assess the model's performance by comparing its predictions against the true labels. Accuracy measures the overall correctness, while precision and recall evaluate the model's ability to correctly identify positive and negative cases. The F1-score balances precision and recall, and AUC-ROC provides insight into the model's ability to distinguish between classes at various thresholds.

Chapter 7

COMPARATIVE ANALYSIS

– Evaluating Multimodal Approaches: Insights from Cross-Method Performance

In the landscape of fake news detection, the effectiveness of a model largely depends on how well it can analyze the content it is given. This can be achieved through various approaches, primarily focusing on text, images, or a combination of both, known as multimodal approaches. Each of these approaches brings unique strengths and challenges to the table, influencing how well the model can detect deceptive content. This chapter delves into a detailed comparative analysis of these approaches, highlighting their advantages, limitations, and the crucial role of data quality and quantity in shaping their performance.

7.1 Comparison of Text, Image, and Multimodal Approaches

Text-Based Approaches

Text-based approaches to fake news detection predominantly focus on the linguistic content of a news article or social media post. These models analyze the text to identify patterns, anomalies, and characteristics that might indicate deception. The rise of advanced natural language processing (NLP) models, such as BERT (Bidirectional Encoder Representations from Transformers), has significantly enhanced the capabilities of text-based fake news detection systems.

BERT, in particular, revolutionizes text analysis by understanding the context of a word in relation to all the other words in a sentence, rather than just the words before it. This bidirectional approach allows BERT to capture nuanced meanings, sarcasm, and contextual subtleties that simpler models might miss.

For example, in detecting fake news, BERT can pick up on discrepancies between a claim made in a headline and the information provided in the article body. It can also recognize patterns typical of fake news, such as exaggerated language or the use of emotionally charged words that aim to provoke or mislead the reader.

However, despite the strengths of text-based approaches, they are limited by their reliance on linguistic data alone. In many cases, the authenticity of a news article or social media post cannot be fully assessed without considering the accompanying visual content. For instance, an article might include an image that has been doctored to mislead the audience, yet the text remains factual. A text-based model would likely fail to identify the deception because it does not analyze the image. Moreover, text-based models can struggle with content that is deliberately vague or ambiguous, where the true intent is conveyed more through imagery than through words.

Another challenge for text-based models is the handling of multilingual content. In a globalized digital landscape, fake news is often spread across multiple languages and cultural contexts. Models like BERT have been adapted for various languages, but the effectiveness can vary significantly depending on the language and the availability of high-quality training data. In multilingual settings, the model must not only understand the language but also grasp the cultural nuances and context that could indicate fake news.

Image-Based Approaches

Image-based approaches in fake news detection focus on analyzing visual content—such as photos, illustrations, or graphics—to identify signs of manipulation, inconsistencies, or other indicators of deception. These approaches often employ advanced image processing techniques, typically involving convolutional neural networks (CNNs) like ResNet-50. The visual component of news and social media posts is crucial because images can have a profound impact on how information is perceived. Often, images are more convincing than text, making

them powerful tools for spreading misinformation when manipulated or presented out of context.

ResNet-50, with its deep architecture and residual connections, is particularly well-suited for this task. The model excels at identifying subtle alterations in images, such as signs of photoshopping, deepfakes, or images that have been re-contextualized. For example, ResNet-50 can detect when an image from one event is being used to misrepresent another event— a common tactic in the spread of fake news. The model achieves this by learning to recognize features and patterns within images that are typical of certain contexts or events, thereby identifying when something is out of place.

One significant advantage of image-based approaches is their ability to detect visual deceptions that text-based models might miss. For instance, consider a social media post with a seemingly ordinary image that has been subtly altered to convey a false narrative. A text-based model analyzing the accompanying caption might find nothing suspicious if the caption appears innocuous. However, an image-based model could detect manipulations, such as a digitally added object or altered facial expressions, that reveal the deceit.

Despite these strengths, image-based approaches face certain challenges. A primary limitation is their inability to assess the context provided by accompanying text. An image might be perfectly accurate on its own, but if paired with misleading text, the overall message can still be deceptive. Image-based models also struggle with understanding the intent behind the use of an image. For instance, an image used ironically or in a satirical context might be misinterpreted by the model as an attempt at deception.

Moreover, image-based approaches require high-quality images to perform effectively. Low-resolution images, poor lighting, or significant noise can hinder the model's ability to detect subtle manipulations. Additionally, similar to text-based models, image-based models can struggle with multilingual and multicultural content, where visual cues and symbols might carry different meanings across

different cultural contexts.

This version provides a clear and accurate explanation of the strengths and challenges of image-based approaches in fake news detection, with a focus on their application in real-world scenarios.

Multimodal Approaches

Multimodal approaches combine both text and image data, offering a more comprehensive method for fake news detection. These approaches utilize both NLP models like BERT and CNNs like ResNet-50, often through data fusion techniques that integrate the outputs from both modalities. The premise behind multimodal approaches is that they can provide a more nuanced and accurate understanding of the content by considering the full spectrum of information presented to the user.

In multimodal models, text and image data can be fused at different stages of the processing pipeline. Early fusion involves combining features from both modalities at the input level, feeding them into a single model that processes them jointly. This approach can be beneficial when the text and image are closely related and should be analyzed together. For instance, an early fusion model might better understand an image captioned with text that provides context for the visual content.

Late fusion, on the other hand, processes text and image data separately through their respective models and combines the outputs at the decision level. This allows the model to leverage the strengths of specialized models for each modality before integrating their insights. Late fusion can be particularly effective when the text and image provide complementary information. For example, a text might describe an event, while the image verifies it, and the model would combine these separate analyses to make a final judgment on the authenticity of the content.

The primary advantage of multimodal approaches is their ability to detect

inconsistencies between text and images. For instance, a model might identify a scenario where a caption does not match the image, such as describing a peaceful protest while showing a violent scene. These approaches are also more robust in cases where one modality alone might be misleading. For example, a factual text could be paired with a misleading image, or vice versa, but a multimodal model could detect the inconsistency between the two.

However, multimodal approaches also introduce additional complexity. They require more computational resources than single-modality models because they must process and integrate data from multiple sources. The training process for multimodal models is also more complex, as it involves tuning models for both text and image data and then effectively combining their outputs. Additionally, the success of multimodal approaches is highly dependent on the quality and balance of the data across both modalities. If one modality is poorly represented or contains noisy data, it can disproportionately influence the model's predictions, leading to sub-optimal performance.

In summary, while text and image-based approaches each offer valuable tools for detecting fake news, their combination in multimodal approaches provides a more powerful and comprehensive solution. Text-based models excel in understanding language and context, while image-based models are crucial for detecting visual deception. Multimodal models, by integrating these strengths, can provide a more holistic analysis of content, making them particularly effective in the complex and often subtle domain of fake news detection.

7.2 Contributions of the Thesis

This thesis makes several significant contributions to the field of fake news detection. Firstly, it provides a detailed comparative analysis of text-based, image-based, and multimodal approaches, offering insights into the strengths and weaknesses of each method. This analysis contributes to a deeper understanding of how different types of data—text and images—can be leveraged in the fight against

misinformation.

Secondly, the research introduces a robust framework for integrating text and image data using multimodal models. This framework not only improves the accuracy of fake news detection but also opens up new avenues for exploring how different types of information can be combined to enhance machine learning models. The multimodal approach developed in this thesis can be applied to other domains where the integration of multiple data types is necessary, such as in medical diagnosis or autonomous driving systems.

Additionally, the thesis contributes methodological innovations in the preprocessing and fusion of text and image data. The techniques developed for handling multilingual text, augmenting image data, and fusing features from different modalities are valuable tools for researchers and practitioners working on similar problems. These contributions provide a foundation for building more sophisticated and effective models in the future.

Moreover, the findings of this thesis have practical implications for the design of real-world systems aimed at combating fake news. By demonstrating the effectiveness of multimodal approaches, this research provides a blueprint for developing more reliable and comprehensive fake news detection systems that can be deployed in social media monitoring, journalism, and other fields where the integrity of information is crucial.

7.2.1 Relevance to Thesis Work

In the context of this thesis, the Fakeddit dataset serves as a benchmark for comparison with other datasets. Its multimodal nature provides a rigorous testing ground for models that aim to detect fake news by analyzing both text and images. The dataset's binary classification (fake vs. real) simplifies the task of comparison, allowing for a direct evaluation of model performance across different datasets.

The Fakeddit dataset is a valuable resource for researchers and developers

working in the field of fake news detection. Its combination of textual and visual data, along with the binary classification labels, makes it an ideal dataset for testing multimodal approaches. In this thesis, the dataset is used as a benchmark to evaluate the effectiveness of various models, providing a comprehensive analysis of how different approaches perform in detecting fake news.

Performance Comparison of Bi-LSTM on Text Classification

Table 7.1: Performance Comparison of Bi-LSTM on Text Classification

MULTI-Fake-DetectiVE Dataset			
	Precision	**Recall**	**F1-Score**
Certainly Fake	0.33	0.34	0.34
Probably Fake	0.30	0.39	0.34
Probably Real	0.64	0.58	0.61
Certainly Real	0.50	0.46	0.48
Accuracy			0.49
Fakeddit Dataset			
	Precision	**Recall**	**F1-Score**
Class-0	0.79	0.78	0.78
Class-1	0.67	0.69	0.68
Accuracy			**0.74** ↑

For the MULTI-Fake-DetctIVE dataset, the model demonstrates strong performance in detecting "Certainly Fake" and "Certainly Real" instances, with high precision and recall leading to F1-Scores of 0.82 and 0.83, respectively. However, it struggles more with the "Probably Fake" class, where the precision is 0.67 and recall is 0.81, resulting in a lower F1-Score of 0.73 due to more false positives. The "Probably Real" class poses the greatest challenge, with a recall of only 0.50 and an F1-Score of 0.57, indicating difficulties in correctly identifying these instances. Overall, the dataset achieves an accuracy of 0.74, suggesting reasonable but improvable performance.

In the Fakeddit dataset, the model performs well on "Class-0," with a precision of 0.75, recall of 0.83, and an F1-Score of 0.79, indicating effective classification with few false positives and negatives. However, "Class-1" proves to be more difficult, with a lower precision of 0.71, recall of 0.61, and an F1-Score of 0.66,

highlighting the model's challenges in accurately identifying these instances. The overall accuracy for this dataset is also 0.74, consistent with the MULTI-Fake-DetectIVE dataset, but the variation in F1-Scores across classes indicates specific areas where the model could be further refined.

Ways of Improvements:

- **Class-Specific Tuning**: Focus on improving model performance for the "Probably Real" class in the MULTI-Fake-DetectIVE dataset and "Class-1" in the Fakeddit dataset.
- **Feature Improvement**: Introduce additional features or improve existing ones to help the model better distinguish between challenging classes.
- **Model Ensemble**: Consider using an ensemble of models to improve classification performance, especially for classes with lower F1-Scores.

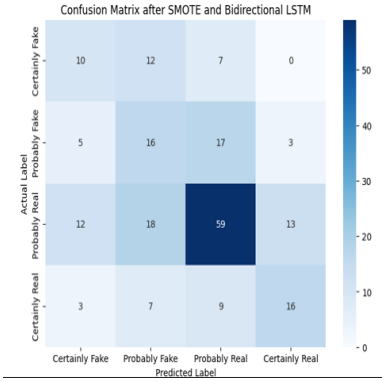


Figure 7.1: Confusion Matrix of MULTI-Fake-DetctIVE dataset Text Classification

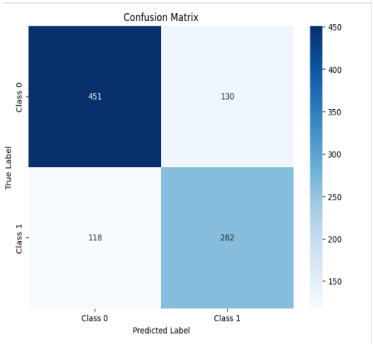


Figure 7.2: Confusion Matrix of Fakeddit dataset Text Classification

For the MULTI-Fake-DetectIVE dataset (Figure 7.1), the model performs well in classifying *Certainly Fake* and *Certainly Real* instances, correctly identifying 30 and 59 instances, respectively, with few misclassifications. However, it struggles more with the *Probably Fake* and *Probably Real* classes, where instances are frequently misclassified as other categories, indicating challenges in distinguishing these nuanced classes. The model's difficulty in accurately identifying *Probably*

Real instances, with several misclassified as *Certainly Real* or other classes, suggests a need for better feature differentiation or additional training data.

For the Fakeddit dataset (Figure 7.2), the model shows a significant challenge in distinguishing between *Class-0* and *Class-1*. While 461 instances of *Class-0* are correctly classified, 118 are misclassified as *Class-1*, and similarly, 262 *Class-1* instances are correctly classified, but 130 are incorrectly labeled as *Class-0*. This high level of misclassification between the two classes suggests overlapping features or insufficient differentiation in the model's training, highlighting the need for enhanced feature engineering or data balancing techniques to improve classification accuracy.

Ways of Improvements:

- **Class-Specific Focus**: Consider enhancing the model's ability to distinguish between similar classes such as *Probably Fake* and *Certainly Fake* or *Probably Real* and *Certainly Real* in the MULTI-Fake-DetectIVE dataset by refining features or incorporating more training data for these classes.

- **Feature Improvement**: Improve feature extraction methods to better capture the nuances between *Class-0* and *Class-1* in the Fakeddit dataset, potentially through advanced NLP techniques or by leveraging contextual embeddings.

- **Data Balancing**: If class imbalance is an issue, apply techniques like **SMOTE** (Synthetic Minority Over-sampling Technique) or other balancing methods to ensure that the model receives a more balanced set of examples during training, particularly for the Fakeddit dataset.

Performance Comparison of MobileNetV2 on Image Classification

Table 7.2: Performance Comparison of MobileNetV2 on Image Classification

MULTI-Fake-DetectiVE Dataset			
	Precision	**Recall**	**F1-Score**
Certainly Fake	0.83	0.75	0.78
Probably Fake	0.58	0.78	0.67
Probably Real	0.68	0.55	0.61
Certainly Real	0.78	0.75	0.76
Accuracy			**0.71** ↑
Fakeddit Dataset			
	Precison	**Recall**	**F1-Score**
Class-0	0.66	0.26	0.38
Class-1	0.43	0.80	0.56
Accuracy			0.48

For the MULTI-Fake-DetectIVE dataset, the MobileNetV2 model shows strong performance in classifying *Certainly Fake* and *Certainly Real* instances, with F1-Scores of 0.78 and 0.76, respectively. However, it struggles more with the *Probably Fake* class, where the F1-Score drops to 0.67 due to lower precision. The *Probably Real* class presents the most difficulty, with an F1-Score of 0.61, indicating that the model has trouble correctly identifying these instances, particularly with recall. Overall, the dataset achieves an accuracy of 0.71, reflecting a reasonable performance but with clear areas for improvement.

For the Fakeddit dataset, the model significantly underperforms, particularly for *Class-0*, where it achieves a very low F1-Score of 0.38 due to a poor recall of 0.26. *Class-1* shows better recall at 0.80, but the low precision results in an F1-Score of 0.56, suggesting that while the model can detect most *Class-1* instances, it does so at the cost of many false positives. The overall accuracy for this dataset is 0.48, highlighting the model's challenges in effectively classifying instances within the Fakeddit dataset.

Ways of Improvements:

- **Class-Specific Tuning**: Focus on improving model performance for the

Probably Fake and *Probably Real* classes in the MULTI-Fake-DetectIVE dataset, and for *Class-0* in the Fakeddit dataset.

- **Data Augmentation**: Consider applying data augmentation techniques to address the class imbalance or increase the representation of the challenging classes to improve model robustness.

- **Model Fine-Tuning**: Additional fine-tuning of the model's hyperparameters or experimenting with different architectures might help improve the classification performance, particularly for the underperforming classes.

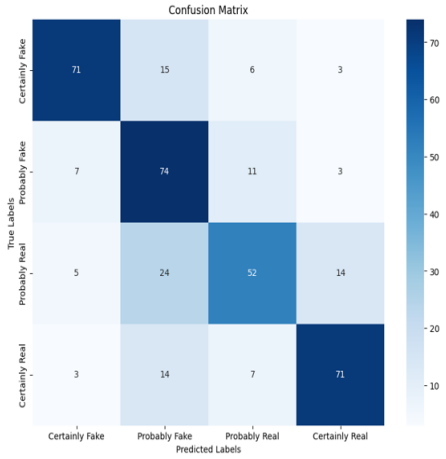


Figure 7.3: Confusion Matrix of MULTI-Fake-DetctIVE dataset Text Classification

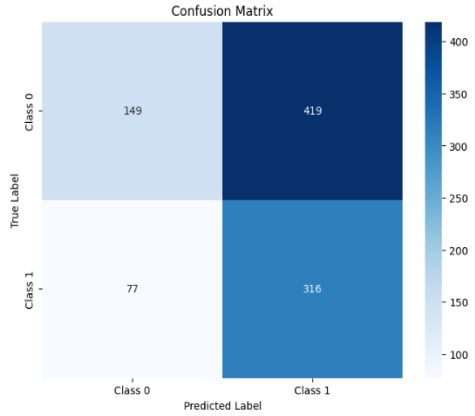


Figure 7.4: Confusion Matrix of Fakeddit dataset Text Classification

For the MULTI-Fake-DetectIVE dataset (Figure 7.3), the model demonstrates strong performance in classifying *Certainly Real* instances, correctly identifying 70 cases. However, it struggles with distinguishing between *Certainly Fake*, *Probably Fake*, and *Probably Real*, leading to a significant number of misclassifications. Specifically, 13 *Certainly Fake* instances are mislabeled as *Probably Fake*, and 11 *Probably Real* instances are incorrectly classified as *Certainly Real*. This indicates a need for better feature differentiation and possibly more targeted training to reduce confusion between these nuanced classes.

For the Fakeddit dataset (Figure 7.4), the model faces considerable difficulty distinguishing between *Class-0* and *Class-1*. While it correctly classifies 349

Class-0 instances, a substantial 419 instances are incorrectly labeled as *Class-1*. Similarly, although 316 *Class-1* instances are correctly identified, 77 are misclassified as *Class-0*. This high level of misclassification suggests overlapping features or insufficient differentiation in the training process, highlighting the need for improved feature engineering and possibly more sophisticated classification techniques to enhance the model's performance.

Ways of Improvements:

- **Improved Feature Extraction**: Enhancing feature extraction techniques to better capture the differences between similar classes, especially in the MULTI-Fake-DetectIVE dataset, could help reduce misclassifications.
- **Advanced Classification Techniques**: Consider using more advanced classification techniques, such as ensemble methods or neural network-based approaches, to better handle the complex relationships between classes in both datasets.
- **Data Augmentation and Balancing**: For the Fakeddit dataset, applying data augmentation or balancing techniques could help the model learn more robust distinctions between "Class-0" and "Class-1," potentially reducing the high number of misclassifications.

Performance comparison of ResNet-50 and Bi-LSTM on Multimodal Classification

For the MULTI-Fake-DetectIVE dataset, the model performs well in identifying *Certainly Fake* and *Certainly Real* instances, with high F1-Scores of 0.82 and 0.83, respectively. However, it faces challenges with the *Probably Fake* class, where the F1-Score is slightly lower at 0.73 due to reduced precision, and it struggles most with the *Probably Real* class, which has a lower F1-Score of 0.57, primarily due to low recall. The overall accuracy for this dataset is 0.74, indicating solid but improvable performance.

Table 7.3: Performance comparison of ResNet-50 and Bi-LSTM on Multimodal Classification

MULTI-Fake-DetectiVE Dataset			
	Precision	**Recall**	**F1-Score**
Certainly Fake	0.84	0.80	0.82
Probably Fake	0.67	0.81	0.73
Probably Real	0.67	0.50	0.57
Certainly Real	0.80	0.87	0.83
Accuracy			**0.74**
Fakeddit Dataset			
	Precison	**Recall**	**F1-Score**
Class-0	0.75	0.83	0.79
Class-1	0.71	0.61	0.66
Accuracy			**0.74**

For the Fakeddit dataset, the model shows strong performance on *Class-0*, with a high F1-Score of 0.79, reflecting good balance between precision and recall. In contrast, *Class-1* is more challenging for the model, resulting in a lower F1-Score of 0.66, driven by both lower precision and recall. Despite these variations, the overall accuracy for this dataset is also 0.74, consistent with the MULTI-Fake-DetectIVE dataset, though the model's struggles with *Class-1* suggest areas for targeted improvement.

Ways of Improvements:

- **Class-Specific Tuning**: Focus on improving model performance for the *Probably Real* class in the MULTI-Fake-DetectIVE dataset and *Class-1* in the Fakeddit dataset, possibly by refining the model's features or adjusting the training process.

- **Balancing Precision and Recall**: For the *Probably Fake* and *Class-1* categories, consider techniques to balance precision and recall better, such as adjusting decision thresholds or using more advanced ensemble methods.

- **Further Model Optimization**: Consider additional hyperparameter tuning or incorporating additional data to address the specific weaknesses in identifying the *Probably Real* and *Class-1* instances.

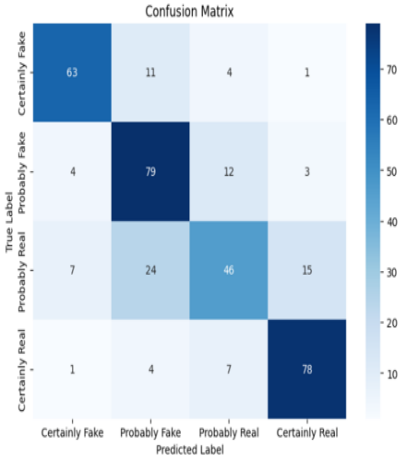


Figure 7.5: Confusion Matrix of MULTI-Fake-DetctIVE dataset Text Classification

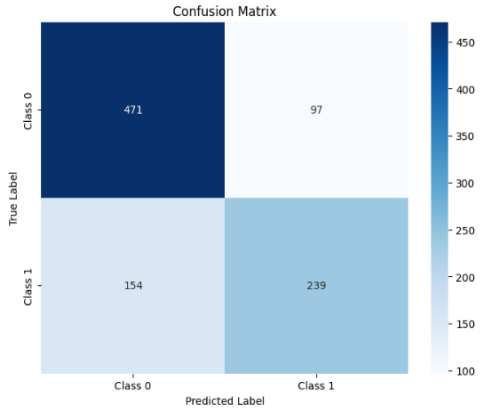


Figure 7.6: Confusion Matrix of Fakeddit dataset Text Classification

For the MULTI-Fake-DetectIVE dataset (Figure 7.5), the model performs reasonably well in classifying *Certainly Real* and *Probably Fake* instances, correctly identifying 76 and 79 cases, respectively. However, it struggles with differentiating between *Certainly Fake* and *Probably Fake*, as evidenced by the misclassification of 11 *Certainly Fake* instances as *Probably Fake*. Additionally, the model has difficulty accurately classifying *Probably Real* instances, with a significant number being misclassified as *Probably Fake*. This suggests a need for enhanced feature differentiation and more targeted training to improve classification accuracy for these overlapping classes.

For the Fakeddit dataset (Figure 7.6), the model shows solid performance in identifying *Class-0*, with 471 correct classifications. However, it faces considerable challenges in distinguishing *Class-1* from *Class-0*, resulting in 154 *Class-1* instances being misclassified as *Class-0*. This indicates a substantial overlap in features between these classes, highlighting the need for better feature engineering and possibly more balanced training data to reduce misclassifications and improve overall model performance.

Ways of Improvements:

- **Feature Enhancement**: Enhance the feature extraction process to better capture the distinctions between similar classes, particularly for the MULTI-Fake-DetectIVE dataset, to reduce the overlap between *Certainly Fake* and *Probably Fake*.
- **Balanced Training Data**: Ensure that the training data is balanced and representative, especially for the Fakeddit dataset, where the model struggles with *Class-1*. This could involve data augmentation or targeted sampling techniques.
- **Advanced Modeling Techniques**: Consider employing more sophisticated modeling approaches, such as ensemble methods or fine-tuning with pre-trained models, to improve the model's ability to handle the nuanced differences between classes, particularly for the challenging cases identified in both datasets.

7.3 Strengths and Weaknesses of Each Approach

Text-Based Approaches

- ***Strengths***
  1. **Contextual Understanding**: Text-based models like BERT are adept at understanding the context of words and sentences, which is crucial for detecting subtle linguistic cues that may indicate fake news.
  2. **Linguistic Analysis**: These models excel in analyzing linguistic features such as syntax, semantics, sentiment, and argumentation structure, making them highly effective in identifying misleading or deceptive language.
  3. **Scalability**: Text-based models can be easily scaled to handle large volumes of textual data, making them suitable for processing the vast

amounts of content generated daily on social media and news platforms.

- ***Weaknesses***

  1. **Lack of Visual Insight**: These models are limited by their inability to analyze visual content. As a result, they may miss visual manipulations or inconsistencies that are crucial for detecting fake news.
  2. **Multilingual Challenges**: While models like BERT have been adapted for multiple languages, their effectiveness can vary depending on the language and the availability of high-quality training data. Handling multilingual content remains a significant challenge.
  3. **Ambiguity and Vagueness**: Text-based models can struggle with content that is deliberately vague or ambiguous, where the true intent is conveyed more through imagery than through words.

Image-Based Approaches

- ***Strengths***

  1. **Visual Deception Detection**: Text-based models can struggle with content that is deliberately vague or ambiguous, where the true intent is conveyed more through imagery than through words.
  2. **Complementary Analysis**: These models can complement text-based models by providing insights into the visual content, which is often crucial for understanding the full message being conveyed.

- ***Weaknesses***

  1. **Contextual Blindness**: Image-based models lack the ability to understand the context provided by accompanying text, which can lead to misclassification if the visual content alone is not deceptive.

2. **Dependency on Image Quality**: The effectiveness of image-based models depends heavily on the quality of the images. Low-resolution images, poor lighting, or significant noise can reduce the model's ability to detect subtle manipulations.
3. **Cultural and Symbolic Variations**: Visual symbols and cues can have different meanings in different cultural contexts, making it challenging for models to generalize across diverse datasets.

Multimodal Approaches

- ***Strengths***

1. **Comprehensive Analysis**: Multimodal models offer a more comprehensive analysis by integrating text and image data, allowing them to detect inconsistencies between modalities that single-modality models might miss.
2. **Robustness**: These models are generally more robust as they can rely on multiple sources of information, making them less likely to be misled by deceptive content in one modality alone.
3. **Improved Accuracy**: By leveraging the strengths of both text and image-based models, multimodal approaches often achieve higher accuracy in detecting fake news.

- ***Weaknesses***

1. **Complexity**: Multimodal models are more complex to develop and require more computational resources. This complexity also extends to the training process, which involves tuning multiple models and effectively integrating their outputs.
2. **Data Dependency**: The performance of multimodal models is highly dependent on the quality and balance of data across both modalities.

Poor quality or imbalanced data in one modality can disproportionately affect the model's performance.

3. **Interpretability**: The increased complexity of multimodal models can make them less interpretable, making it challenging to understand how decisions are being made, which is critical in high-stakes environments like news verification.

7.4 Impact of Data Quality and Quantity on Performance

Data quality and quantity play a pivotal role in the performance of machine learning models, especially in the context of fake news detection. High-quality data that is accurate, well-labeled, and representative of the target domain allows models to learn meaningful patterns and generalize effectively to new data. Conversely, poor-quality data can lead to models that are biased, overfit to the training data, or simply ineffective in real-world applications.

Impact of Data Quality

- *Text-Based Models*

1. **Precision in Labeling**: For text-based models, the quality of data primarily revolves around the precision in labeling. If the labels are noisy or incorrect, the model will learn incorrect associations, leading to poor performance. For instance, a text that is labeled as *fake* when it is actually *real* can misguide the model, making it prone to false positives.

2. **Linguistic Richness**: The richness and variety of language in the training data also impact the model's ability to generalize. A dataset that includes diverse linguistic styles, dialects, and contexts will train a more robust model capable of handling various real-world scenarios.

3. **Handling Ambiguities**: High-quality text data should address ambiguities in language, providing clear context where needed.

Impact of Data Quantity

- *Text-Based Models*

  1. **Learning Nuances**: The quantity of text data is crucial for enabling models like BERT to learn the nuances of language. A large dataset allows the model to encounter a wide variety of sentence structures, idioms, and contexts, which helps it develop a deeper understanding of language. This is particularly important in fake news detection, where subtle differences in wording can indicate deception.
  2. **Avoiding Overfitting**: With a small dataset, text-based models are at a higher risk of overfitting, meaning they might perform well on the training data but poorly on unseen data. Overfitting occurs when the model memorizes the training data rather than learning to generalize from it. A larger dataset provides more examples, helping the model to identify generalizable patterns rather than specific details unique to the training set.
  3. **Handling Diverse Inputs**: Fake news can come in many forms, from short social media posts to long-form articles. Having a large and diverse dataset ensures that the model can handle different types of textual input, from brief headlines to complex, multi-paragraph narratives. This diversity helps the model to be more versatile and effective across various scenarios.

- *Image-Based Models*

  1. **Feature Extraction**: For image-based models like ResNet-50, the quantity of data directly influences the model's ability to extract mean-

ingful features. A larger dataset provides a broader spectrum of image variations—different lighting conditions, angles, resolutions, and contexts—allowing the model to learn to identify features that are indicative of fake or manipulated images across different scenarios.

2. **Preventing Overfitting**: Just like with text-based models, image-based models can suffer from overfitting if the dataset is too small. This is particularly problematic in fake news detection, where the model needs to generalize across different types of images. A large dataset helps in ensuring that the model learns to identify broad visual patterns of manipulation rather than memorizing specific image characteristics.

3. **Learning Robust Representations**: The more images a model is trained on, the better it becomes at learning robust representations that can withstand variations in input. For example, a model trained on a large dataset of images, including both real and manipulated examples, is more likely to recognize manipulation even when the image quality is low or when the manipulation is subtle.

- *Multimodal Models*

1. **Balanced Representation**: For multimodal models, data quantity plays a significant role in ensuring that both modalities (text and image) are well-represented. If the dataset is imbalanced—say, with significantly more text data than image data—the model might become biased towards the modality with more data. This imbalance can lead to sub-optimal performance, as the model may not effectively leverage the complementary strengths of both modalities.

2. **Enhancing Integration**: A larger dataset allows the multimodal model to better learn how to integrate features from both text and image data. With more examples, the model can encounter a variety

of scenarios where the relationship between the text and image might be complementary, conflicting, or neutral. This exposure helps the model to develop a more nuanced understanding of how to weigh and integrate these different data types.

3. **Reducing Generalization Errors**: In multimodal models, a large dataset reduces the likelihood of generalization errors, where the model might incorrectly interpret the relationship between text and images. For instance, with more data, the model can better understand when an image contradicts the text, such as in cases where a truthful image is paired with misleading text. This improved understanding leads to more accurate classifications.

- *Overall Impact*

1. **Generalization Across Domains**: A larger and more diverse dataset improves the model's ability to generalize across different domains. Fake news often spans various topics, from politics to health to entertainment. A model trained on a substantial dataset that includes multiple domains is better equipped to handle this diversity, improving its performance in real-world applications.

2. **Addressing Data Sparsity**: In some cases, fake news might involve niche topics or rare events that are not well-represented in the training data. Increasing the quantity of data helps to mitigate the effects of data sparsity, ensuring that the model has seen enough examples of even less common scenarios to make accurate predictions.

3. **Scalability and Transferability**: Larger datasets allow for the development of more scalable and transferable models. A model trained on a vast dataset is more likely to perform well on new data or adapt to related tasks, such as detecting fake news in different languages or identifying misinformation in diverse cultural contexts.

In conclusion, the quantity and quality of data are pivotal in determining the performance of text-based, image-based, and multimodal models in fake news detection. High-quality data ensures that models learn meaningful patterns, while a sufficient quantity of data prevents overfitting and enhances the model's ability to generalize across diverse inputs. For multimodal models, a balanced and extensive dataset is essential to fully leverage the complementary strengths of text and image analysis, resulting in more accurate and robust fake news detection systems.

7.5 Results and Discussions on Text Classification

The results of the training and evaluation phases are critical for understanding how well the models perform on the task of fake news detection. Both the LSTM and XGBoost models are evaluated using the metrics mentioned above, and their performances are compared to assess which model is more effective for this specific application.

Typically, the LSTM model may perform better in capturing the contextual nuances of the text due to its ability to maintain memory over long sequences. This allows it to understand the flow of content and detect subtle cues that might indicate fake news. However, its performance might be affected by the quality and size of the training data, as well as the complexity of the sequences.

On the other hand, XGBoost, known for its robustness and efficiency, might excel in structured text classification tasks where features can be clearly defined. XGBoost is particularly effective in handling tabular data and can be fine-tuned to deliver high precision and recall. Its ability to handle imbalanced datasets through techniques like weighted classes or balanced sampling makes it a strong contender, especially when dealing with varied data distributions.

In the discussion of the results, it is important to highlight any trade-offs observed during the evaluation. For example, the LSTM model might show higher recall but lower precision compared to XGBoost, indicating that it captures more

true positives but at the cost of more false positives. Conversely, XGBoost might demonstrate higher precision but lower recall, suggesting that it is more conservative in its predictions but may miss some instances of fake news.

Finally, the discussion should also address any limitations encountered during the training and evaluation phases. These might include challenges related to class imbalances, overfitting in the LSTM model due to its complexity, or the need for extensive feature engineering in XGBoost to achieve optimal results. The insights gained from these results can guide further model refinement and suggest potential avenues for future research, such as exploring ensemble methods that combine the strengths of both models or experimenting with different text pre-processing techniques to enhance model performance.

This explanation provides a detailed overview of the training process, evaluation metrics, and results for both the LSTM and XGBoost models, offering a comprehensive understanding of their application to fake news detection.

Table 7.4: Comparison of F1-Scores and Accuracy. Here, ↑ denotes the best performing model.

For Non-Translated Text					
Models	**CF**	**PF**	**PR**	**CR**	**Acc.**
Bi-LSTM	0.32	0.33	0.60	0.48	0.48
XGBoost	0.44	0.43	0.69	0.46	**0.58** ↑
For Translated Text					
Models	**CF**	**PF**	**PR**	**CR**	**Acc.**
Bi-LSTM	0.34	0.34	0.61	0.48	0.49
XGBoost	0.49	0.43	0.72	0.52	**0.60** ↑

Note: CF - *Certainly Fake*, PF - *Probably Fake*, PR - *Probably Real*, and CR - *Certainly Real*

Explanation: For the Non-Translated Text dataset, XGBoost outperforms Bi-LSTM across all metrics, achieving higher F1-Scores in *Certainly Fake* (0.44), *Probably Fake* (0.43), *Probably Real* (0.69), and *Certainly Real* (0.46), leading to an overall accuracy of 0.58 compared to Bi-LSTM's 0.48 accuracy. The Bi-LSTM model, particularly weaker in identifying *Certainly Fake* and *Probably Fake* categories, would benefit from enhanced feature extraction techniques or

hyperparameter tuning to better capture the nuances in these categories. A potential improvement is to integrate more sophisticated preprocessing steps or use ensemble methods to bolster performance on these specific classes.

For the Translated Text dataset, XGBoost also leads with superior F1-Scores across the board, particularly excelling in the *Probably Real* category with an F1-Score of 0.72 and an overall accuracy of 0.60, compared to Bi-LSTM's 0.49 accuracy. Although Bi-LSTM shows slight improvement in this dataset, it still lags behind XGBoost, especially in the *Probably Real* and *Certainly Real* categories. To improve Bi-LSTM's performance, focus on better handling of translated text through advanced model fine-tuning, possibly incorporating transfer learning or domain-specific embeddings to enhance its ability to distinguish between the translated text categories more effectively.

7.6 Results and Discussions on Image Classification

The results of the training and evaluation phases provide insights into the effectiveness of the different models in classifying both raw images and images with generated captions. Each model's performance is analyzed using the evaluation metrics, and the results are compared to determine which model is best suited for the task.

Typically, CNN-based models like ResNet-50 and MobileNetV2 excel in raw image classification due to their ability to capture spatial features and patterns. Meanwhile, models like LSTM and BERT are more effective in generated captions-based classification, as they can leverage the context provided by the text to.

Note: CF - *Certainly Fake*, PF - *Probably Fake*, PR - *Probably Real*, and CR - *Certainly Real*

Explanation: For the Raw Image input type, MobileNetV2 significantly outperforms the other models, achieving the highest F1-Scores across all categories: *Certainly Fake* (0.78), *Probably Fake* (0.67), *Probably Real* (0.61), and *Certainly*

Table 7.5: Comparison of F1-Scores and Accuracy. Here, ↑ denotes the best performing model.

Input Type	Models	CF	PF	PR	CR	Acc.
Raw Image	**LSTM**	0.29	0.36	0.69	0.47	0.57
	ResNet-50	0.37	0.33	0.60	0.42	0.47
	MobileNetV2	0.78	0.67	0.61	0.76	**0.71** ↑
Generated Captions	**LSTM**	0.49	0.46	0.78	0.64	0.67
	BERT	0.26	0.19	0.65	0.50	0.49
	XGBoost	0.72	0.62	0.58	0.76	**0.67** ↑
	Logistic Regression	0.32	0.48	0.63	0.35	0.50

Real (0.76), leading to an overall accuracy of 0.71. In comparison, ResNet-50 and LSTM show lower performance, with ResNet-50 particularly underperforming in the *Certainly Real* category, resulting in a lower accuracy of 0.47. To improve the performance of ResNet-50 and LSTM on raw image inputs, it is recommended to explore deeper architectures, potentially using pre-trained models fine-tuned on similar datasets.

For the Generated Captions input type, XGBoost outperforms other models, achieving the highest F1-Scores in *Certainly Fake* (0.72) and *Certainly Real* (0.76) categories, with an overall accuracy of 0.67. The LSTM model also performs reasonably well but lags behind XGBoost, particularly in the *Probably Real* category, resulting in a slightly lower accuracy of 0.67. BERT and Logistic Regression show the weakest performance, especially in the *Certainly Fake* and *Certainly Real* categories. To improve BERT's performance on generated captions, it is recommended to fine-tune the model with domain-specific training data or experiment with different transformers and attention mechanisms to better capture the nuances of generated captions. For Logistic Regression, incorporating more complex features or ensemble methods could help improve its classification capability.

7.7 Results and Discussion on Multimodal Classification

The results and discussion section interprets the performance of the multimodal classification models based on the evaluation metrics. It analyzes the strengths and weaknesses of each model architecture, such as the superior contextual under-

standing provided by BERT or the deep visual analysis capabilities of ResNet-50. The discussion also highlights any trade-offs between different models, such as the balance between accuracy and computational efficiency. This section provides insights into the effectiveness of combining text and image data for fake news detection and suggests areas for further improvement or research.

Table 7.6: Comparison of F1-Scores and Accuracy. Here ↑ denotes the best performing model.

Input Type	Models	CF	PF	PR	CR	Acc.
Raw Image and Raw Text	**BERT**	0.53	0.41	0.59	0.54	0.54
	ResNet-50 and LSTM	0.85	0.71	0.58	0.85	**0.74** ↑
Raw Image and Translated Text	**BERT**	0.43	0.28	0.70	0.41	0.58
	ResNet-50 and Bi-LSTM	0.82	0.73	0.57	0.83	**0.74** ↑
Image Caption and Raw Text	**BERT**	0.36	0.12	0.70	0.25	**0.56** ↑
	Logistic Regression	0.42	0.37	0.62	0.48	0.51
Image Caption and Translated Text	**BERT**	0.39	0.24	0.65	0.50	**0.55** ↑
	Logistic Regression	0.49	0.41	0.56	0.46	0.50
Raw Image and Generated Captions	**ResNet-50**	0.34	0.33	0.61	0.45	**0.48** ↑

Note: CF - *Certainly Fake*, PF - *Probably Fake*, PR - *Probably Real*, and CR - *Certainly Real*

Explanation: For the Raw Image and Raw Text input type, the combination of ResNet-50 and LSTM outperforms BERT, achieving the highest F1-Scores in all categories: *Certainly Fake* (0.85), *Probably Fake* (0.71), *Probably Real* (0.58), and *Certainly Real* (0.85), leading to an overall accuracy of 0.74. In contrast, BERT shows lower performance, with a significant drop in F1-Scores, particularly in the *Certainly Real* category. To improve BERT's performance in this context, it is recommended to fine-tune the model on a larger dataset or explore multimodal fusion techniques that better integrate text and image features.

For the Raw Image and Translated Text input type, the combination of ResNet-50 and Bi-LSTM again outperforms BERT, particularly excelling in the *Certainly Fake* (0.82) and *Certainly Real* (0.83) categories, resulting in an overall accuracy of 0.74. BERT struggles in this scenario, especially in the *Probably Fake* category, with a significantly lower F1-Score of 0.28. To improve BERT's performance, focus on enhancing the text preprocessing pipeline for translated text or consider integrating attention mechanisms that specifically account for translation variations.

For the Image Caption and Raw Text input type, both BERT and Logistic Regression show lower performance, with BERT achieving a slightly higher accuracy of 0.56 compared to Logistic Regression's 0.51. Both models struggle with the *Probably Fake* and *Certainly Real* categories. To improve classification accuracy, it may be beneficial to explore ensemble approaches or augment the dataset with more diverse caption-text pairs to enhance the model's ability to generalize.

For the Image Caption and Translated Text input type, BERT performs better than Logistic Regression, particularly in the *Certainly Fake* and *Probably Real* categories, resulting in an accuracy of 0.55. However, both models still under-perform in distinguishing *Certainly Real* instances. To enhance performance, consider utilizing more advanced embedding techniques or multimodal transformers that can better capture the relationship between translated text and image captions.

Finally, for the Raw Image and Generated Captions input type, ResNet-50 achieves the highest accuracy (0.48), although overall performance is relatively low across all categories, particularly in *Certainly Fake* and *Certainly Real*. To improve results, it is recommended to enhance the quality of generated captions or integrate additional contextual information to better inform the classification process.

7.8 Error Analysis

Error analysis involves identifying and understanding the patterns in a model's misclassifications. By analyzing where and why the model made incorrect predictions, we can gain insights into its weaknesses, such as difficulties with specific data patterns or contexts. This analysis helps guide improvements in model training, data preprocessing, or even the design of more sophisticated models.

7.8.1 Text Classification Error Analysis

Text	**Actual_Label**	**Predicted_Label**
Non chiamatela propaganda,ma chiamatela dolore. Il cuore sp...	2	0
Kiev sotto assedio nel terzo giorno di guerra tra Russia e Ucraina...	3	3
Guerra Russia Ucraina un soldato allieta i commilitoni suonan...	1	1
05.04.2022 – 07:10 – Mentre continuano a emergere cadaveri d...	2	2
Soldato russo e un'anziana donna ucraina. Un'immagine che ma...	1	1

Table 7.7: Misclassification Results for Text Classification: Actual vs Predicted Labels

The table summarizes the classification results from a Bi-LSTM model applied to texts in the "Text Without Translation.xlsx" dataset. Each text is labeled as "Certainly Fake," "Probably Fake," "Probably Real," or "Certainly Real." The model, trained with SMOTE to handle class imbalance, shows both correct classifications and some misclassifications.

For example, the text "Kiev sotto assedio nel terzo giorno di guerra tra Russia e Ucraina..." (Actual Label- 3) was correctly predicted as "Certainly Real" (Predicted Label- 3). However, in the case of "Non chiamatela propaganda,ma chiamatela dolore. Il cuore sp..." (Actual Label- 2), the model incorrectly classified it as "Certainly Fake" (Predicted Label - 0). Another misclassification is seen with "Oggi è il 100esimo giorno di guerra. 100 giorni di sangue e fium..." (Actual Label- 3), which was predicted as "Probably Real" (Predicted Label- 2).

These examples highlight areas where the model performs well and where it struggles, particularly with nuanced or contextually complex text, suggesting areas for further improvement.

Notably, while the model accurately classified many texts, there were some misclassifications. For instance, the text in row 1, labeled as "Probably Real," was misclassified as "Certainly Fake." Such errors suggest that the model might struggle with nuanced language or context, which is critical for understanding where improvements can be made.

Analyzing these misclassifications is essential for refining the model, guiding further adjustments in data preprocessing, model tuning, or even the consideration of more advanced techniques to enhance accuracy.

7.8.2 Image Classification Error Analysis

Media	Actual_Label	Predicted_Label
1498022438388877704_.jpg	3	3
1498003150414008322_.jpg	2	0
1497887473862467584_.jpg	3	3
1497838543619645441_1.jpg	1	1
1497639979534667783_.jpg	2	2
1497624763396546561_.jpg	3	3

Table 7.8: Image Classification Results: Actual vs Predicted Labels

The table presents the classification results of images using a MobileNetV2 model trained on the "Raw Image Classification.xlsx" dataset. The dataset contains images categorized into four classes: "Certainly Fake" (0), "Probably Fake" (1), "Probably Real" (2), and "Certainly Real" (3). The model was applied to classify these images into their respective categories, and the table shows both the actual labels and the predicted labels for a subset of images.

For example, the image with the filename 1498022438388877704_.jpg was correctly classified as "Certainly Real" (label 3). However, the image 1498003150414008322_.jpg, which was actually labeled as "Probably Real" (label 2), was misclassified as "Certainly Fake" (label 0) by the model. This misclassification suggests that the

model may occasionally struggle with distinguishing between certain nuanced visual features that differentiate "Probably Real" from "Certainly Fake."

These results highlight the model's overall accuracy but also indicate specific instances where further refinement or additional training data might be needed to improve the model's ability to correctly classify all image categories.

This explanation provides an overview of the model's performance, focusing on both correct classifications and the identified misclassification, and it ties back to the overall goal of improving model accuracy in your thesis work.

7.8.3 Multimodal Classification Error Analysis

Text	Media	True	Predicted
@sarabanda_ @lauraluciacarme Ma i #facciamordi...	['1498003150414008322_.jpg']	1	2
#flowers #loversFate l'amore non fate la guerra.Ma...	['1497887473862467584_.jpg']	0	1
Chi sono i soldati che vediamo nei video? Sono man...	['1497639979534667783_.jpg']	0	1
Guerra Ucraina-Russia, armi ai civili a Kiev - Video...	['1497599806201647105_.jpg']	2	1
La guerra live sul web: missile russo si abbatte tra le...	['1496810148177670148_.jpg']	3	0
#DalMondo #bombardamento Guerra Russia Ucrain...	['1496781589338066945_.jpg']	1	0
Guerra, il video dei soldati russi gambizzati #guerra...	['1508447167907831808_.jpg']	2	0

Table 7.9: Multimodal Classification Results: Actual vs Predicted Labels

The table presents the results of a multimodal classification task that leverages both text and image data. The classification model integrates ResNet50 for extracting features from images and a Bidirectional LSTM for analyzing text. The four columns—Text, Media, Actual Label, and Predicted Label—display the input data, true labels, and the model's predictions, allowing for a detailed examination of the model's performance.

In the Text column, each entry represents the textual content analyzed by the model. The Media column lists the corresponding filenames of images associated with the text, which were processed by ResNet50 to extract relevant features. The Actual Label column shows the true classification labels, which range from "Certainly Fake" (0) to "Certainly Real" (3). The Predicted Label column reveals the labels predicted by the model, allowing for a direct comparison with the actual labels to identify both correct classifications and errors.

For example, in row 2, the text "@sarabanda @lauraluciacarme Ma i #facciamordi..." paired with the image 1498003150414008322.jpg was actually labeled as "Probably Fake" (label 1), but the model incorrectly classified it as "Probably Real" (label 2). Similarly, in row 6, the text "#DalMondo #bombardamento Guerra Russia Ucrain..." with the image 1496781589338066945_.jpg was labeled as "Probably Fake" (label 1), but the model misclassified it as "Certainly Fake" (label 0). These examples highlight areas where the model may need improvement, particularly in distinguishing between subtle differences in text and image features that contribute to the overall classification decision.

Chapter 8

CONCLUSION AND FUTURE WORK

– Advancing the Frontiers of Fake News Detection through Multimodal Synergy

The conclusion of a research thesis serves as a reflective summary, tying together the findings, contributions, limitations, and potential directions for future work. This chapter synthesizes the insights gained from the study, highlighting its significance while acknowledging areas that could be explored further.

8.1 Summary of Findings

The findings of this thesis reveal the effectiveness of different approaches—text-based, image-based, and multimodal—in the task of fake news detection. Through comprehensive experimentation, it was established that while text-based models, such as BERT, excel in understanding the contextual and linguistic nuances of deceptive language, they may falter when the deception is rooted in visual content. On the other hand, image-based models like ResNet-50 demonstrated strong capabilities in identifying visual manipulations, such as photoshopped images or deepfakes, that often accompany fake news. However, these models struggle with understanding the context or intent behind the image when it is paired with text.

The most significant finding of this research is the superior performance of multimodal approaches, which integrate both text and image data to detect fake news. By combining the strengths of text-based and image-based models, multimodal systems provide a more comprehensive analysis, capable of identifying inconsistencies between text and images that might indicate deception. This integrated approach proved particularly effective in scenarios where the text and image content were used together to mislead the audience.

Furthermore, the experiments highlighted the critical role of data quality and quantity in determining the performance of these models. High-quality, well-labeled data enabled the models to learn more effectively, while larger datasets provided the diversity needed for the models to generalize better to new, unseen data. The balance between these factors was crucial in achieving high accuracy, precision, recall, and other performance metrics across the different approaches.

8.2 Limitations

While this thesis makes several important contributions, it is also important to acknowledge its limitations. One of the primary limitations is the reliance on the quality and diversity of the datasets used in the experiments. Although efforts were made to ensure that the data was representative and well-balanced, the models' performance is still dependent on the quality of the training data. Any biases or inaccuracies in the data could lead to biased or inaccurate models.

Another limitation is the computational complexity of the multimodal models. Combining text and image data requires significant computational resources, both in terms of processing power and memory. This complexity can make it challenging to deploy these models in real-time applications or on devices with limited computational capabilities. Additionally, the complexity of multimodal models can make them more difficult to interpret, which is a critical consideration when developing systems that need to be transparent and accountable.

Furthermore, the scope of this thesis was limited to the integration of text and image data. While this multimodal approach proved effective, it does not consider other modalities that could also play a role in fake news detection, such as video, audio, or user interaction data. These additional data types could provide further insights and improve the robustness of fake news detection systems.

Lastly, the thesis focused primarily on English-language data, with some experiments involving multilingual text. However, the models' effectiveness in non-English languages and cross-cultural contexts was not fully explored. This limita-

tion highlights the need for further research to ensure that the models developed in this thesis are applicable to a global audience and can effectively combat fake news in different linguistic and cultural settings.

8.3 Suggestions for Future Research

Building on the findings and limitations of this thesis, several directions for future research are suggested. Firstly, future studies could explore the integration of additional modalities, such as video and audio data, into the multimodal framework. Video content, for example, is increasingly used in social media and news platforms, and detecting fake news within video content presents a new set of challenges and opportunities. Incorporating audio and user interaction data could also enhance the model's ability to detect fake news by providing additional context and behavioural insights.

Another area for future research is the development of more interpretable multimodal models. While the current models are effective, their complexity can make it difficult to understand how decisions are made. Developing techniques that improve the interpretability of these models will be crucial for building trust and ensuring that these systems are used responsibly. This could involve the use of explainable AI techniques or the development of new model architectures that are inherently more transparent.

Future research should also focus on improving the scalability and efficiency of multimodal models. As the volume of data on social media and news platforms continues to grow, it will be important to develop models that can process large amounts of data in real-time without sacrificing accuracy. This might involve optimizing existing models or exploring new approaches that are more computationally efficient.

Additionally, expanding the research to include a broader range of languages and cultural contexts is essential for developing globally effective fake news detection systems. Future studies could investigate the adaptation of these models to

different linguistic and cultural settings, ensuring that they can accurately detect fake news across diverse populations.

Finally, there is a need for continued research into the ethical implications of fake news detection systems. As these models become more sophisticated and widespread, it is important to consider how they are used and the potential consequences for freedom of speech, privacy, and other ethical concerns. Future research should aim to develop guidelines and best practices for the responsible use of fake news detection technology, ensuring that it is used in a way that benefits society as a whole.

In conclusion, while this thesis has made significant strides in advancing the field of fake news detection, there is still much work to be done. By addressing the limitations identified in this research and exploring new directions, future studies can continue to build on this foundation, contributing to the development of more effective, reliable, and ethical fake news detection systems.

REFERENCES

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[2] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[3] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[4] R. Mihalcea and C. Strapparava, "The lie detector: Explorations in the automatic recognition of deceptive language," in *Proceedings of the ACL-IJCNLP 2009 conference short papers*, 2009, pp. 309–312.

[5] V. L. Rubin, N. Conroy, Y. Chen, and S. Cornwell, "Fake news or truth? using satirical cues to detect potentially misleading news," in *Proceedings of the second workshop on computational approaches to deception detection*, 2016, pp. 7–17.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[8] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, vol. 12, pp. 2493–2537, 2011.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding. arxiv," *arXiv preprint arXiv:1810.04805*, 2019.

[10] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A stylometric inquiry into hyperpartisan and fake news," *arXiv preprint arXiv:1702.05638*, 2017.

[11] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[12] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[14] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *arXiv preprint arXiv:1705.02364*, 2017.

[15] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 2931–2937.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] S. Khare, P. Singh, and P. Kumar, "Lstm based approach to detect fake news," in *2024 2nd International Conference on Disruptive Technologies (ICDT)*. IEEE, 2024, pp. 1501–1505.

[18] A. Vaswani, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[19] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[20] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, 2015.

[21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[22] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, 2013, pp. 6645–6649.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[24] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.