Extracting and Visualizing Ontologies of Characters, Topics and Relations from Texts

Thesis Submitted by

Apurba Paul

Doctor of Philosophy (Engineering)

Department of Computer Science & Engineering

Faculty Council of Engineering & Technology

Jadavpur University

Kolkata, India

2024

JADAVPUR UNIVERSITY
KOLKATA-700032, INDIA

INDEX NO. 215/17/E

1. **Title of the Thesis: Extracting and Visualizing Ontologies of Characters, Topics and Relations from Texts**

2. **Name, Designation & Institution of the Supervisor: Dr. Dipankar Das, Assistant Professor, Department of Computer Sc. & Engg., Jadavpur University, Kolkata – 700 032, India**

3. **List of Publications:**

   (A) **Journal:**

   i. Paul, A., Seal, S. & Das, D. Transformer-based Pouranic topic classification in Indian mythology. *Sādhanā* 49, 263 (2024). https://doi.org/10.1007/s12046-024-02598-6

   (B) **Conferences:**

   i. Apurba Paul and Dipankar Das. 2017. A Deep Dive into Identification of Characters from Mahabharata. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 447–455, Kolkata, India. NLP Association of India.

   ii. Apurba Paul and Dipankar Das. 2017. Identification of Character Adjectives from Mahabharata. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 569–576, Varna, Bulgaria. INCOMA Ltd.

   iii. Apurba Paul, Anupam Mondal, Sainik Kumar Mahata, Srijan Seal, Prasun Sarkar, and Dipankar Das. 2023. Mytho-Annotator: An Annotation tool

for Indian Hindu Mythology. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 574–578, Goa University, Goa, India. NLP Association of India (NLPAI).

4. **List of Patents:** N/A

5. **List of Presentations in International Conferences :**

   (A) Apurba Paul and Dipankar Das. 2017. A Deep Dive into Identification of Characters from Mahabharata. In Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017), pages 447–455, Kolkata, India. NLP Association of India.

   (B) Apurba Paul, Anupam Mondal, Sainik Kumar Mahata, Srijan Seal, Prasun Sarkar, and Dipankar Das. 2023. Mytho-Annotator: An Annotation tool for Indian Hindu Mythology. In Proceedings of the 20th International Conference on Natural Language Processing (ICON), pages 574–578, Goa University, Goa, India. NLP Association of India (NLPAI).

Statement of Originality

I, Apurba Paul, registered on 16/10/2017, do hereby declare that this thesis entitled `Extracting and Visualizing Ontologies of Characters, Topics and Relations from Texts` contains literature survey and original research work done by the undersigned candidate as part of Doctoral studies.

All information in this thesis have been obtained and presented in accordance with existing academic rules and ethical conduct. I declare that, as required by these rules and conduct, I have fully cited and referred all materials and results that are not original to this work.

I also declare that I have checked this thesis as per the "**Policy on Anti Plagiarism, Jadavpur University, 2019**", and the level of similarity as checked by **iThenticate software is 5%**.

Signature:

Apurba Paul

Certified by the supervisor (signature and seal)

Dipankar Das. 05/11/24

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

Certificate from the Supervisor

This is to certify that the thesis entitled `Extracting and Visualizing Ontologies of Characters, Topics and Relations from Texts` submitted by **Shri Apurba Paul**, who got his name registered on **16/10/2017** for the award of **Ph.D. (Engg.) degree of Jadavpur University** is absolutely based upon his own work under the supervision of **Dr. Dipankar Das** and that neither his thesis nor any part of the thesis has been submitted for any degree/diploma or any other academic award anywhere before.

Signature of the Supervisor

Dipankar Das. 05/11/2021

Dr. Dipankar Das
Assistant Professor,
Department of Computer Sc. & Engg., Jadavpur University,
Kolkata – 700 032, India

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

Dedicated to my Guru,

Shri Satyaban Banerjee

Acknowledgements

Acknowledging the invaluable support and guidance that has shaped this thesis, I would first like to express my deepest gratitude to **the Almighty**, whose blessings and grace have been my strength throughout this academic journey.

This thesis embodies not just my efforts but also the unwavering support from many individuals. I thank my parents, family and friends for their boundless motivation and belief in me. They have been a constant source of inspiration to me besides my research guide.

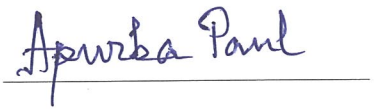
A special thanks to my research mentor and guide, **Dr. Dipankar Das**, whose mentorship brought me into the light of the fascinating world of Natural Language Processing. His constructive feedback, advice and guidance have been pivotal in refining my research and shaping my academic growth. He has been a constant support and a propelling force in this journey of my research.

My Guru, **Shri Satyaban Banerjee**, has been a powerful source of encouragement, fueling my determination to pursue excellence even in the most challenging times.

I also extend my sincere appreciation to the members of the Research Group in Natural Language Processing at *Jadavpur University*, especially **Dr. Anupam Mondal**, **Monalisa Dey**, and **Dr. Sainik Kumar Mahata**, for their invaluable insights that enhanced this thesis.

I would like to thank everyone who has contributed to this journey. This thesis stands as a testament to the collective support I have received along the way.

Signature:



Apurba Paul

Abstract

This thesis proposes a robust and innovative framework for the extraction, classification, and visualization of ontologies related to characters, topics, and their interrelationships within Indian mythological texts, particularly focusing on the *Ramayana*, the *Mahabharata*, the *Devi Bhagavata*, the *Harivamsha Purana*, the *Srimad-Bhagavatam*, the *Caitanya Caritamrita*, and *Krsna, the Supreme Personality of Godhead*. Using advanced Natural Language Processing (NLP) techniques, this work addresses the complex challenges posed by these intricate and expansive narratives. The research integrates state-of-the-art transformer models and methodologies to enhance character identification, thematic classification, and ontology development.

In the first segment, the thesis explores **Character Identification** within these ancient texts, utilizing linguistic feature extraction and machine learning models like Neural Networks and KNN classifiers. By developing a dedicated dataset for mythological characters, the study successfully disambiguates character names and their associated attributes, contributing to a more precise recognition system.

The second section presents **Pouranic Topic Classification**, introducing transformer-based models such as BERT, RoBERTa, and DistilBERT for classifying topics across mythological texts. The development of a novel annotated dataset, *PouranicTopic*, allows for accurate classification of cantos, and topics, tackling the complexity of overlapping topics and enhancing the understanding of these mythological texts.

Next, **Ontology and Character-Topic Relationship** is explored through the creation of SBC-Ontology and SBT-Ontology, offering a structured framework that maps characters to their associated topics. By employing advanced transformer models, the study uncovers deeper insights into the relationships between characters and the topics they represent in texts like the *Srimad-Bhagavatam*.

The fourth contribution is **MythoBERT 1.0**, a custom-built language model designed for Indian hindu mythology. MythoBERT surpasses general models in tasks like Named Entity Recognition (NER), text classification, and topic modeling, offering a specialized

tool for understanding the rich linguistic and cultural nuances of mythological texts.

The thesis also introduces **Mytho-Annotator**, a web-based annotation tool tailored to Hindu mythological texts. This tool provides an efficient framework for labeling entities, relationships, and events, accelerating the annotation process and enabling the creation of high-quality datasets for further research.

Finally, the study delves into the **Visualization of Character-Centric Summaries**, employing models like T5, BART, and PEGASUS to generate summaries focused on individual characters. KeyBERT is used to extract key phrases, linking them to specific characters and enhancing the interactive exploration of these texts through visual summaries.

Overall, this thesis significantly advances the application of NLP in Indian mythology, offering new models and insights for character identification, topic classification, ontology development, domain specific language model, annotation tool and finally the visualization of mythological characters of complex narratives. These contributions have broad implications for digital humanities, cultural preservation, and the creation of educational tools, enabling deeper engagement with India's hindu mythological heritage.

Extracting and Visualizing Ontologies of Characters, Topics and Relations from Texts

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Exploring Indian Mythology with NLP

1.1 Background

Indian mythology, particularly epics like the Mahabharata and Ramayana, is not just a treasure trove of stories but also a complex web of characters, philosophies, and cultural lessons. These texts have captured the imagination of readers for millennia, offering rich, layered narratives filled with nuanced characters. However, extracting meaningful information from these vast mythological corpora—especially identifying characters, understanding their roles,relations between characters and topics and development of ontologies—are the challenging tasks for readers and scholars alike.

With the rise of Natural Language Processing (NLP), we now have the tools to analyze such complex texts systematically. Modern transformer-based models like BERT and RoBERTa have revolutionized the way we approach language understanding, offering unprecedented accuracy in tasks such as text classification and entity extraction. However, their application to Indian mythological texts is still in its early stages, with challenges posed by the diversity of the texts and their linguistic intricacies.

Research in Indian mythology through Natural Language Processing (NLP) holds significant importance due to the richness of its narratives, cultural relevance, and the potential for various applications. Here's a detailed exploration of its importance, benefits and applications:

Importance of Indian Mythology Research in NLP

- Cultural Heritage:
  - Preservation of Tradition: Indian mythology is an integral part of the country's cultural heritage, encapsulating moral values, historical events, and social

norms. NLP can help to digitize, preserve, and analyze these texts, making them accessible for future generations.

  - Diversity of Narratives: India has a vast array of mythological texts, including the Ramayana, Mahabharata, and Puranas, each with unique stories and teachings. Analyzing these texts using NLP can reveal insights into different cultural perspectives.

- Understanding Societal Values:

  - Moral and Ethical Insights: The narratives often encapsulate ethical dilemmas and moral teachings. Researching these texts can help to understand the evolution of societal values and norms in Indian culture.
  - Gender Roles and Social Structure: Mythological stories frequently reflect the roles and perceptions of gender and society. NLP analysis can highlight these aspects and facilitate discussions on societal changes.

- Interdisciplinary Approach:

  - Collaboration Across Fields: Research in Indian mythology using NLP fosters collaboration among scholars from fields like linguistics, anthropology, religious studies, and computer science, leading to richer insights and interdisciplinary methodologies.

Benefits of Indian Mythology Research in NLP

- Enhanced Text Analysis:

  - Automated Information Extraction: NLP techniques enable automated extraction of characters, topics, and relationships from extensive mythological texts, saving time and effort compared to manual methods.
  - Improved Text Classification: Utilizing machine learning algorithms can enhance the categorization of mythological texts based on topics, narratives, or cultural contexts.

- Accessibility and Engagement:

  - Digital Humanities Initiatives: Digitizing mythological texts and making them available online increases public access and engagement, allowing more people to explore and appreciate these narratives.

  - Interactive Learning: NLP can be used to create interactive educational tools, such as chatbots and educational apps, that help users learn about Indian mythology in an engaging manner.

- Data-Driven Insights:
  - Quantitative Analysis: NLP facilitates the quantitative analysis of mythological texts, enabling researchers to identify trends, patterns, and anomalies that may not be evident through qualitative analysis alone.
  - Sentiment Analysis: Understanding the sentiments expressed in mythological narratives can offer insights into historical and cultural attitudes toward various characters and events.

Applications of Indian Mythology Research in NLP

- Character Identification and Mapping:
  - Developing Character Databases: NLP techniques can be employed to identify and classify characters from various texts, enabling the development of comprehensive character databases that track their roles, relationships, and attributes.
- Event Extraction and Summarization:
  - Automated Summarization: Summarizing lengthy mythological texts to create concise versions that capture the essence of the narratives, useful for educational purposes and quick reference.
  - Event Timeline Generation: Extracting significant events from mythological narratives to create timelines that visualize character interactions and major plot developments.
- Interactive Storytelling:
  - Gaming and Multimedia Applications: Using NLP to develop interactive storytelling experiences in games and multimedia applications, where users can explore mythological narratives and make choices that influence the storyline.
- Cross-Cultural Studies:
  - Comparative Mythology Research: Applying NLP to compare Indian mythology with other global mythologies, highlighting similarities, differences, and common themes across cultures.

- Educational Tools:
  - Chatbots and Virtual Assistants: Developing educational chatbots that provide information about characters, events, and themes in Indian mythology, enhancing learning experiences for students.
- Semantic Web Applications:
  - Knowledge Graphs: Developing semantic networks that map out relationships between characters, events, and places in Indian mythology, enabling better information retrieval and exploration.

Moreover, the vast range of topics embedded in mythological literature—ranging from philosophical discourses to detailed genealogies—demands sophisticated topic modeling techniques. In this regard, Pouranic (mythological) topic modeling aims to unravel the deep, thematic structure of these texts, allowing us to categorize and understand the wealth of content in Indian mythology.

In this context, my research seeks to bridge the gap between classical Indian mythology and modern NLP techniques, advancing how we can both extract and interpret characters, topics and the relations from these ancient texts using cutting-edge methods.

1.2 Motivation

Research on Mythology has been a central theme in literature for centuries, contributing to cultural, religious, and moral frameworks across civilizations. Mythological narratives, with their archetypal characters and symbolic motifs, provide valuable insights into human nature, societal values, and the collective unconscious. While mythology has long been a topic of research in literature, its intersection with computer science is still evolving.

1.2.1 Indian Mythology

Indian mythology offers a vast and complex narrative landscape that has inspired scholars across disciplines, including literature, philosophy, history, and, more recently, Natural Language Processing (NLP) and Artificial Intelligence (AI).One of the few computational analyses of Indian mythology, particularly centering on the *Mahabharata*, is provided by (Das et al., 2016b). It applies Natural Language Processing, sentiment analysis, and social network analysis to uncover character dynamics and emotional arcs in the epic. The work

offers fresh insights into the structure and leadership roles within the story, providing a new dimension to traditional literary analysis.

(Buddhi et al., 2022) conducts a fundamental NLP-based experiment on the Mahabharata by training a word2vec model on the corpus to identify the most similar words within the text. The choice of word2vec is motivated by its ability to effectively manage high-dimensional word vectors. Our analysis encompasses essential aspects such as unigram vocabulary, sentence distribution, 100-dimensional vector representation, and the calculation of word similarities.

In exploring the intricate layers of Indian mythology, (R. and Aithal, 2023)'s study on Mahabharata characters provides a fresh perspective by connecting the Big Five personality traits—captured in the OCEAN model (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism)—with the nuanced personalities within this epic narrative. Through analysis, the researchers illuminate how these mythological characters exhibit universal personality traits that align with modern psychology's foundational models. This work reveals that, beyond the typical binary of good versus evil, Mahabharata's characters embody complex, multifaceted behaviors that evolve across diverse contexts, reflecting both unique individuality and shared human tendencies. Such insights not only deepen our understanding of ancient Indian narratives but also offer a practical framework for psychology and management education, highlighting mythology's enduring relevance in interpreting human behavior.

(Gadesha et al., 2023)'s study introduces a computational approach to analyzing the Mahabharata using Natural Language Processing (NLP) techniques. This research leverages modern language models to uncover statistical insights and semantic relationships within the epic, offering a data-driven exploration beyond traditional sentiment-based interpretations. By identifying frequently co-occurring words, sentence structures, and popular lemmas, this approach opens new avenues for understanding the Mahabharata across various domains. This work highlights how computational methods can reveal complex linguistic patterns within Indian mythology, enriching cross-disciplinary studies in literature, AI, and cultural studies.

(Varadarajan et al., 2022) introduces GENOME, the first dedicated methodology for iterative ontological modeling of Indian epics. It addresses shortcomings in previous ad-hoc approaches by integrating best practices and allowing for the reuse of existing ontological models. The methodology is validated through its application to the Indian epic Mahabharata, demonstrating both thoroughness and performance efficacy.

(Gultepe and Mathangi, 2023) applies social network analysis to the *Mahabharata*, preprocessing the text into verses and utilizing TF-IDF and Latent Semantic Analysis (LSA) to derive word vectors. A novel locally weighted K-Nearest Neighbors (KNN) algorithm transforms these vectors into a social network, achieving an F-score of 0.812 in predicting key character relationships and effectively distinguishing between the Pandavas and Kauravas, showcasing its significance in computational analysis of Indian mythology.

1.2.2 Global Mythology

There are numerous studies on Greek mythology, whereas research on Celtic, Nordic, Slavic, and Norse mythologies is relatively limited. Some of the recent works are follows:

(Syamili and Rekha, 2017) focuses on building an ontology for the heroes of ancient Greek mythology, addressing the lack of such ontologies in the domain. The authors employ a combined methodology, integrating elements from Gruninger and Fox's motivating scenario, METHONTOLOGY's development cycle, and YAMO's Analytico-synthetic approach, to create a comprehensive framework for this ontology.

In (Pastor-Sánchez et al., 2021), the authors address the absence of an ontology for Greek mythology in the Semantic Web by creating the Ontology of Greek Mythology (OGM) from a selection of 5377 Wikidata items. The OGM incorporates properties that define classes for mythological figures and accommodates contradictions inherent in classical narratives. Additionally, a retrieval tool utilizing SPARQL queries was developed, allowing users to query and download results in various formats. The study also explores a semantic enrichment workflow to extract additional data from classical sources, highlighting the need for scalable automation to handle the complexities of mythological knowledge.

In (Kostkan et al., 2023), authors present OdyCy, a versatile NLP pipeline crafted specifically for Ancient Greek, which attains SOTA performance on the Perseus UD Treebank. It excels in tasks including POS tagging, dependency parsing and morphological analysis. The pipeline aims to be a reproducible, open-source tool that can process Ancient Greek texts of varying quality. The authors also evaluate its performance against comparable tools and analyze lemmatization errors, showcasing its effectiveness in the field of computational linguistics for ancient languages.

The study by (Besnier, 2020) leverages social network analysis to explore the character networks, offering a framework for comparing the importance and relationships of mythological figures over time. Specifically, the Siegfried-Sigurd mythology, which is believed

to originate from events in the Merovingian dynasty of the fifth and sixth centuries, is dissected through the Nibelungenlied, the Völsunga saga, and the History of the Franks. By tracking the interactions and centrality of characters in these texts, the study highlights how certain figures maintain prominence or diminish across retellings, allowing for a nuanced view of mythological transformation. This methodology provides a basis for further comparative mythology research by allowing consistent annotations and analyses across disparate sources.

(Kenna and MacCarron, 2017) introduced a novel method for analyzing ancient narratives using network science, which quantifies the interactions between characters in texts like The Iliad, Beowulf, and Icelandic sagas. This approach, rooted in statistical physics, reveals the social structures within these stories, enabling comparisons across mythologies and with modern social networks. By examining the interrelationships within epics, this method provides fresh insights into comparative mythology and character dynamics across different cultural narratives.

(Fumanal-Idocin et al., 2021) introduced an advanced social network analysis that incorporates "semantic affinity" to capture the meaning associated with each character in mythological networks. This method allows for nuanced comparisons of gods and heroes in Greek, Celtic, and Nordic myths, revealing connections like the close ties between Celtic and Nordic deities and the Greek emphasis on heroism over divinity. By adding semantic value to traditional network structures, this approach uncovers deeper, non-structural relationships in comparative mythology.

(Fumanal-Idocin et al., 2023) expands social network analysis by integrating external information, introducing a **semantic value** centrality measure and **semantic affinity** to capture deeper relational meanings in mythological networks. Applied to Greek, Celtic, and Nordic myths, this method uncovers nuanced connections, such as cross-cultural affinities among gods and heroes, offering richer insights than traditional metrics. This approach also proves effective in other domains, like news and social media networks, demonstrating its broader applicability for capturing complex, context-specific relationships.

(Sarkanych et al., 2022) applied network science to analyze the **Kyiv bylyny**, a set of *East Slavic epic narratives* from Ukraine, alongside prominent European epics. By examining community structures and key character rankings, they reveal unique and shared network features within the **bylyny**, validating hypotheses about figures like **Prince Volodymyr** and generating fresh insights. This study positions the **Kyiv bylyny** within

the broader tradition of heroic narratives, offering a quantitative framework to explore Ukraine's epic heritage and its central heroes.

The existing research on Indian mythology, particularly in the context of Natural Language Processing (NLP) and ontological modeling, reveals significant advancements, yet notable gaps persist. While works such as (Das et al., 2016b), (R. and Aithal, 2023), (Gadesha et al., 2023), (Varadarajan et al., 2022), and (Gultepe and Mathangi, 2023) provide valuable insights into character dynamics and social structures in the *Mahabharata*, there remains a lack of comprehensive methodologies that systematically integrate diverse computational techniques for a broader analysis of Indian epics. Furthermore, the potential for developing robust ontologies that encompass the entirety of Indian mythology, while leveraging existing models for scalability and adaptability, has yet to be fully realized. This creates a substantial gap in both the linguistic and cultural aspects of NLP, as the narratives, character dynamics, topics and relation between character and topics inherent in Indian mythology have yet to be effectively modeled or analyzed. Consequently, there is a pressing need for developing specialized NLP tools and techniques that can cater to the complexities of Indian mythological literature. We believe that the models developed in the context of Indian mythology can be utilized for visualizing stories and news, thereby enhancing public interest and providing insights into public opinion.

1.3 Problem Statement

Understanding complex narratives, especially in Indian mythology, presents multiple challenges for automatic systems. These challenges can be broadly classified into two categories: technical challenges and non-technical challenges.

1.3.1 Technical challenges

The study of Indian mythology, particularly through texts such as the *Mahabharata*, *Ramayana*, *Srimad-Bhagavatam*, *Puranas* etc., presents a unique set of technical challenges in the realms of Information extraction and NLP. The core problems addressed in this research include:

- **Lack of Annotated Datasets:** The absence of structured and annotated datasets tailored for Indian mythological texts poses significant barriers for developing effective automated systems in NLP. This lack of resources limits the ability of researchers and practitioners to apply advanced Natural Language Processing techniques.

- **Character Identification complexities:** The presence of a vast array of characters—both protagonists and antagonists—along with their varied representations through proper names and noun phrases complicates the task of automatic character identification. Existing methods often fall short in accurately capturing the multifaceted nature of these characters, hindering comprehensive narrative analysis.

- **Topic Classification Challenges:** Indian mythology encompasses diverse themes and narratives that overlap, making the topic classification of related texts complex. The intricacies involved in categorizing texts based on topics demand sophisticated methodologies to capture the nuanced relationships within the narratives.

- **Domain Ontology Development:** Developing domain-specific ontologies, such as character and topic ontologies, is essential for effectively organizing and managing knowledge within Indian mythology. The lack of established ontologies hinders comprehensive understanding and analysis of mythological texts.

- **Character-Topic Relationship Challenges:** Understanding the relationships between characters and the topics they are involved in presents a significant challenge. This complexity necessitates advanced models to capture these relationships accurately, which is crucial for deeper narrative analysis.

- **Mythology Text-Based Embeddings:** The development of specialized embeddings, such as MythoBERT 1.0, tailored for mythology texts is crucial. Existing embeddings may not adequately capture the unique linguistic and thematic elements found in mythological narratives, limiting their effectiveness in NLP tasks.

- **Inefficient Annotation Processes:** Existing annotation tools often prove inadequate for the specific needs of mythological texts, resulting in inefficient and cumbersome annotation processes. There is a pressing need for tools that can enhance readability while providing effective annotation capabilities, thereby facilitating better understanding and analysis of complex narratives.

- **Character-centric Summarization and Visualization:** Creating effective character-centric summaries and visualizations of mythological texts remains a challenge. Developing methodologies to achieve this can significantly enhance understanding and engagement with the narratives, providing valuable insights for researchers and enthusiasts alike.

1.3.2 Non-technical challenges

The study of mythology extends beyond its narratives and characters, playing a crucial role in shaping cultural, literary, and scientific understandings. Mythology serve as foundational stories that inform societal values, artistic expressions, and human experiences. Mythology often reflect early human attempts to explain natural phenomena, and an understanding of these narratives can enrich scientific perspectives. However, individuals who do not engage with these narratives may encounter various challenges in different contexts. Below are some of the significant issues that arise from a lack of familiarity with mythology:

- **Navigating the Voluminous Nature of Indian Mythology:** One significant challenge in working with Indian mythology is the sheer voluminous size of these texts. These texts contain a vast array of characters, narratives, and philosophical discussions, making them difficult to navigate and comprehend in their entirety. For readers, this can lead to mental fatigue and a sense of overwhelming complexity, often resulting in a nagging frustration as they attempt to follow intricate storylines, character relationships, and moral teachings. This challenge makes it crucial to develop methods that assist readers in engaging with the texts in a more accessible and organized manner.

- **Social Context Challenges:** Individuals who do not study mythology may struggle to understand cultural references and shared narratives that shape societal values and beliefs. This lack of understanding can lead to misinterpretations of social norms and practices that are often deeply rooted in mythological stories, resulting in a disconnection from cultural heritage.

- **Literary Context Challenges:** Without an understanding of mythology, readers may miss significant topics, motifs, and archetypes in literature. Many literary works draw upon mythological references to convey deeper meanings, and failure to recognize these connections can hinder literary analysis and appreciation, reducing the richness of the reading experience.

- **Understanding Historical Contexts:** Many scientific concepts have roots in mythological explanations. A lack of knowledge in mythology may hinder one's ability to appreciate the historical evolution of scientific thought, including how ancient beliefs laid the groundwork for modern scientific theories.

- **Interpreting Cultural Influences on Science:** Mythology influence cultural approaches to science, technology, and ethics. Without recognizing these connections, individuals may struggle to understand how cultural narratives shape scientific inquiry and its applications, leading to potential biases in scientific reasoning.

- **Problem-Solving and Critical Thinking:** Mythology often present complex problem-solving scenarios that mirror scientific challenges. Engaging with these narratives can enhance critical thinking and creativity. A lack of exposure to mythology may limit one's ability to think abstractly or consider diverse solutions to scientific problems.

- **Inspiration for Scientific Inquiry:** Mythological themes can inspire scientific exploration and innovation. For instance, concepts of transformation and duality found in mythology can motivate research in fields such as biology and physics. Individuals unfamiliar with these narratives may miss opportunities for inspiration and interdisciplinary collaboration.

By addressing these intertwined issues, this research aims to pave the way for advancements in the overall analysis of Indian mythological texts.

1.4 Objectives

The primary aim of this research is to develop robust NLP models and datasets that facilitate character identification, topic classification, and character-topics relationship in the domain of Indian mythology. The key objectives of this thesis are as follows:

- **Preparation of Dataset for Character Identification Models:** To curate a specialized dataset for identifying characters from Indian mythological texts. This involves annotating characters, disambiguating names with multiple references, and preparing the data for fine-tuning models like BERT and domain specific embeddings for accurate character extraction.

- **Preparation of Dataset for Topic Classification Models:** To build a dataset for training topic classification models that can accurately categorize different sections of mythological texts. This involves labeling texts according to predefined topics, facilitating the development of transformer-based models for Pouranic and contemporary topic classification.

- **Preparation of Dataset for Domain Ontologies Development:** To develop a dataset supporting the creation of domain-specific ontologies for Indian mythology. The dataset will include entities, relationships, and properties that define mythological characters, events, and places, forming the foundation for structured representations and knowledge graphs in this domain.

- **Preparation of Dataset for Character-Topic Relationship Models:** To create a dataset that maps the relationships between characters and topics within mythological narratives. This dataset will serve to train models that not only identify characters but also link them to relevant topics, thus improving understanding of character dynamics within narratives.

- **Preparation of Dataset for MythoBERT 1.0:** To design and compile a large corpus for training MythoBERT 1.0, a domain-specific transformer model for Indian mythology. This corpus will be drawn from mythological texts like the Ramayana, Mahabharata, and others, with appropriate preprocessing, tokenization, and annotation to support various NLP tasks such as Named Entity Recognition (NER) and text classification.

- **Preparation of Dataset for Character-Centric Summarization and Visualization:** To prepare a dataset that enables character-centric summarization and visualization of keyphrases extracted from mythological texts. This involves linking summaries and key phrases to specific characters, allowing for the creation of interactive visual interfaces where characters are mapped to both their narrative summaries and key topics they are associated with.

By achieving these objectives, the thesis aims to advance the application of NLP in Indian mythology and other narrative domains, improving the capabilities of automated systems to analyze, summarize, and visualize complex, character-driven texts.

1.5 Contributions

This thesis contributes to the field of Natural Language Processing (NLP) through the development and preparation of various datasets and models aimed at enhancing the understanding of Indian mythology and related narrative structures. The key contributions are outlined as follows:

- **Identification of Character and Character Adjectives Models:** Developed a robust model for identifying characters and their associated adjectives within Indian mythological texts. This model enhances the understanding of character traits and narratives, which is crucial for in-depth literary analysis.

  Curated a comprehensive dataset specifically designed for this task, which not only facilitates accurate character recognition but also serves as a benchmark for future studies in character identification and description.

- **Pouranic Topic Classification Models:** Proposed a novel model for Pouranic topic classification, enabling the categorization of themes and topics present in mythological narratives. This classification is vital for researchers to explore and analyze the thematic structures of various texts.

  Established a dedicated dataset to train and evaluate the model, contributing valuable insights into the thematic dynamics of Indian mythology and providing a rich resource for further academic inquiry.

- **Domain Ontology and Character-Topic Relationship Models:** Created domain ontology frameworks that map relationships between characters and topics, offering a structured approach to understanding narrative interconnections. This ontology is essential for applications in semantic analysis and information retrieval.

  Developed models to analyze and quantify these relationships, along with respective datasets that provide foundational resources for character-topic analysis, significantly advancing research in the field.

- **MythoBERT 1.0 and Mytho-Annotator Models: MythoBERT 1.0** is a specialized transformer-based language model designed for Indian mythological texts. Trained on a vast corpus of texts like the Mahabharata, Ramayana, and Puranas, it excels in tasks such as character identification, named entity recognition (NER), and topic classification. By using a custom vocabulary, MythoBERT 1.0 captures unique linguistic patterns and entities in mythology, offering improved accuracy over general models in mythology-specific NLP tasks. This model is crucial for advancing NLP applications in mythology, enabling more effective character recognition and context understanding.

  The **Mytho-Annotator** is a tool developed to facilitate the annotation of characters, events, and themes in Indian mythological texts. It automates processes like

named entity recognition (NER), topic labeling, and relationship extraction, reducing manual effort. With a user-friendly interface and visualization capabilities, the Mytho-Annotator helps researchers easily explore complex mythological narratives and generate richly annotated datasets. This tool significantly aids in the efficient analysis of vast mythological corpora, making it invaluable for research in digital humanities and cultural studies.

- **Visualization of Character-Centric Summary Model:** Designed a model for visualizing character-centric summaries, providing an interactive interface to present summarizations alongside relevant visual elements. This model enhances user engagement and understanding of mythological narratives.

  Curated a dataset to support this model, ensuring that the visualizations are informative and relevant to the character narratives, which is important for educational and research purposes in the field of mythology.

In summary, this thesis provides substantial advancements through the development of datasets and models that address critical challenges in character identification, topic classification, ontology development, MythoBERT 1.0, Mytho-Annotator and visualization of character-centric summary in the context of Indian mythology. These contributions not only enhance the field of NLP but also pave the way for future research and applications in related domains.

1.6 Thesis Overview

This thesis is structured into five chapters, each representing a distinct research paper that contributes to the fields of character identification, annotation tools, and topic modeling within Indian mythology.

1.6.1 Identification of Characters in Hindu Mythology (Chapter 3)

The first chapter focuses on identifying prominent characters and their associated adjectives in the English texts of the Indian mythological epics. Unlike traditional named entity recognition methods, the current approach extracts hidden attributes linked to each character. We discovered specific phrase-level linguistic patterns that reveal the presence of characters within various parts of the text. To extract the characters, six specific patterns were applied. Furthermore, a distinctive set of new features—such as multi-word expressions, parse tree nodes and paths, and immediate ancestors—was used. We also

assessed the correlation of these features to evaluate their significance. Finally, we used several machine learning algorithms—including Naive Bayes, KNN, Logistic Regression, Decision Tree, and Random Forest—alongside deep learning methods to classify patterns as either characters or non-characters, achieving commendable accuracy. The evaluation confirms that the phrase-level linguistic patterns and selected features successfully identify characters and their descriptive adjectives.

Moreover, it was observed that these characters can be identified at both word and phrase levels within sentences, following distinct patterns. To extract characters from the text, we considered two sets of features at both levels. Using a semi-supervised learning approach, we prepared the training datasets. We then employed the Chi-squared statistic to identify significant features, followed by an analysis of the associations among those selected features. Subsequently, we developed training models using Neural Networks and KNN classifiers for both word and phrase levels and conducted tests on these models. Our findings indicate that Neural Networks outperformed KNN, achieving accuracies of 88% and 76% at the word and phrase levels, respectively. Lastly, we analyzed various error measures and visualized the co-occurrence of frequently appearing story characters.

1.6.2 Transformer-based Pouranic Topic Classification in Indian Mythology (Chapter 4)

This chapter offers an in-depth study on Pouranic topic classification, employing transformer-based models for the automated classification of topics within Indian mythological texts. This research aims to address the existing gap in structured and annotated datasets, providing a novel method for interpreting the complex narratives and themes found within the mythological corpus.

Topic classification poses significant challenges in understanding the subject matter or themes of Indian mythology. However, it can enhance the performance of NLP-based systems, such as recommendation engines and semantic search tools, when processing mythological texts. This study focuses on developing transformer-based models for the automated classification of Indian mythological documents, tackling the difficulties associated with organizing and analyzing this rich and diverse body of work. We introduce `PouranicTopic`, a newly annotated dataset comprising over 200,000 verses from seven major Hindu texts, complete with canto, topic, and sentence labels. Additionally, we create two supplementary datasets: `Similarity-based` and `Log-likelihood-based`, utilizing sentence clustering techniques. The performance of the BERT, RoBERTa, and

DistilBERT models is evaluated for canto and topic classification across these datasets. While clustering significantly enhances results on the `Similarity-based` dataset, the `Log-likelihood-based` dataset continues to present challenges.

1.6.3 Ontology and Character-Topic Relationship (Chapter 5)

This chapter proposes the development of a comprehensive domain ontology and a character-topic relationship analysis using models like BERT, RoBERTa, and DistilBERT. The research focuses on addressing gaps in datasets and computational techniques, aiming to uncover deeper insights into character dynamics and thematic relationships within the text. Contributions include specialized datasets, advanced models, and an ensemble approach for improved analysis of the text.

`The Srimad-Bhagavatam` is a foundational text in Hindu mythology, characterized by intricate narratives and profound philosophical teachings. This paper addresses the analytical challenges of the text by developing tailored ontologies: `SBC-Ontology` for characters and `SBT-Ontology` for topics, which together create a structured framework for deeper understanding. Advanced transformer models, including BERT, DistilBERT, and RoBERTa, were utilized to analyse character-topic relationships, with an ensemble approach(F1-score 0.79) demonstrating superior performance in capturing nuanced relationship. Despite the text's inherent complexities and limited data availability, this research significantly contributes to digital humanities, providing innovative tools for exploring and interpreting the rich narratives of `the Srimad-Bhagavatam`.

1.6.4 Application Development in Hindu Mythology (Chapter 6)

This chapter presents **MythoBERT 1.0**, a specialized language model developed to overcome the limitations of general-purpose models like BERT when processing Hindu mythological texts along with Mytho-Annotator, an annotation tool for hindu mythology.

MythoBERT is an innovative pretrained language model specifically tailored for the analysis of Hindu mythology. By introducing a custom vocabulary known as Mytho-Vocab and incorporating domain-specific embeddings, MythoBERT significantly surpasses general-purpose models like BERT-Base in key NLP tasks such as text classification, named entity recognition, and topic modeling. Fine-tuned on a comprehensive corpus of Hindu mythological literature, **MythoBERT** achieves improved accuracy, F1-scores, precision, and recall, demonstrating its exceptional capacity to capture the rich cultural and symbolic nuances of these texts. Furthermore, **MythoBERT** exhibits strong gen-

eralization capabilities in downstream tasks related to moral stories and news articles, highlighting its versatility. This model not only addresses a crucial gap in the processing of culturally rich, domain-specific texts but also paves the way for new applications of NLP in the realm of mythology and beyond. Future research will aim to expand the training data and investigate additional NLP tasks, further enhancing the capabilities of specialized language models in complex narrative contexts.

Mytho-Annotator- Recognizing the complexity of mythology, which encompasses a collection of myths, particularly those belonging to specific religious or cultural traditions, we identify the need for an annotation tool to effectively extract important and intricate information from mythological texts or corpora. Additionally, obtaining high-quality annotated datasets for complex information extraction, including labeled text segments, can be both costly and time-consuming. To address this, we introduce **Mytho-Annotator**, an annotation tool designed specifically for Hindu mythology. This intuitive web-based text annotation tool utilizes Natural Language Processing (NLP) technology, focusing on labeling three main categories: named entities, relationships, and event entities. **Mytho-Annotator** provides a comprehensive and adaptable annotation framework.

1.6.5 Visualization of Character-centric Summary (Appendix A)

This appendix explores character-centric summarization and keyphrase extraction in Indian mythology, focusing on the Mahabharata, a complex narrative with numerous characters and intricate relationships. The study leverages transformer-based models—T5, BART, and PEGASUS—to generate character-specific summaries by fine-tuning these models on a custom dataset ($DS_{character}$). The dataset consists of character-centric sentences that highlight the roles, actions, and significance of individual characters within the text. The models were evaluated using ROUGE and BLEU scores, with PEGASUS exhibiting the best overall performance in capturing the coherence of character-specific summaries. Additionally, the paper introduces a novel keyphrase extraction framework using KeyBERT, achieving strong precision and recall in identifying key themes associated with characters. Visualization of summaries and keyphrases further enhances understanding of each character's importance within the narrative. This research provides a deeper insight into mythology processing and suggests future directions for improving model performance in character-centric tasks.

Overall, this thesis aims to contribute significantly to the understanding of character dynamics, annotation processes, and topic classification in Indian mythology, thereby

advancing research in the intersection of natural language processing and literary studies.

Chapter 2

Literature Survey

Earlier related works

The study of Natural Language Processing (NLP) in the domain of mythology, particularly Indian epics, has seen a surge of interest in recent years. This growing interest is driven by the unique linguistic challenges and cultural richness present in these ancient texts, offering opportunities to explore narrative structure, character analysis, and thematic categorization through computational techniques.

This chapter presents a comprehensive survey of the existing literature relevant to the scope of this research. Each section provides an overview of prior work, organized according to the specific tasks undertaken in this thesis. By examining the methodologies, datasets, and models used in related studies, we aim to contextualize our contributions within the broader field of computational mythology.

2.1 Identification of Characters

Several studies have addressed Character Identification in texts. (Mamede and Chaleira, 2004) developed a system (DID) which was applied to children stories starts by classifying the utterances. The utterances belong to the narrator (indirect discourse) as well as belong to the characters taking part in the story (direct discourse). Afterwards, this DID system tries to associate each direct discourse utterance with the character(s) in the story. (Goyal et al., 2010) proposed a system that exploits a variety of existing resources to identify affect states and applies "projection rules" to map the affect states onto the characters in a story.

(Calix et al., 2013) developed a methodology to detect sentient actors in the spoken stories. (Valls-Vargas et al., 2013) proposed a method for automatically assigning narrative roles to characters in stories. (Valls-Vargas et al., 2014) proposed a case-based approach to character identification in natural language text in the context of their Voz system. (Valls-Vargas et al., 2015) also proposed a feedback-loop-based approach to identify

the characters and their narrative roles where the output of later modules of the pipeline is fed back to earlier ones. In the context of keyword extraction, statistical approaches are often built for extracting general terms (Vandeck et al., 2010); the most basic measure is frequency. C/NC-value (Frantzi et al., 2000), another statistical method is well known in the literature and combines statistical and linguistic information for the extraction of multi-word and nested terms. Lastly, in the realm of keyword extraction, statistical methods are often employed for extracting general terms (Vandeck et al., 2010).

2.2 Pouranic Topic Classification

In this research, we are motivated to classify the important information, such as cantos and topics, from the mythological corpora using various state-of-the-art transformer models. This information assists in understanding the overall meaning in the form of semantics and linguistics for the native user. Furthermore, we have observed that over the last few decades, linguists and researchers have contributed various topic classification models in different public domains like legal, political, and news articles, etc. In spite of that, we observed that there is still a huge scope for improvement in the domain of Indian mythological texts for topic classification, especially Hindu mythology. In order to recognize the research gap and bridge the topic classification models, we provide references to previous studies across a range of domains. At first, we will explore the research work done in previous years in the Indian Hindu mythology domain.

(Das et al., 2016a) presented a computational analysis of the Indian epic Mahabharata, applying natural language processing, sentiment/emotion analysis, and social network analysis methods. (Srijeyarankesh et al., 2022) demonstrated how applying NLP techniques like semantic analysis, sentiment analysis, and Named Entity Recognition (NER) to the Mahabharata can yield insights into events, character actions, and emotional shifts across its eighteen parvas. (Varadarajan et al., 2022) introduced GENOME, the first methodology for iterative ontological modeling of epics, addressing reliance on ad-hoc methods and neglect of existing models.

(Gadesha et al., 2023) and (Buddhi et al., 2022) proposed an NLP pipeline to extract statistical and computational insights, as well as implement a relevant word-searching method, from the epic 'Mahabharata.' This approach addresses human biases and memory limitations, providing valuable data for diverse domains referencing the epic.

Since the aforementioned works are not directly relevant to our research, we endeavored to explore topic classification tasks in different domains below.

(Noguti et al., 2020) developed three different classification methods, such as linear models, boosted trees, and neural networks, to predict the area of law of petitions to the Public Prosecutor's Office automatically. (Chalkidis et al., 2021) introduced a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer to classify the topics from their prepared multi-lingual dataset. (Papaloukas et al., 2021) worked on the task of multi-class Greek legal topic classification to identify the relevant thematic topic. (Braun and Matthes, 2022) examined seven approaches for the automatic classification of clause topics in standard legal contract forms. (Xenouleas et al., 2022) used the MULTI-EURLEX dataset to investigate zero-shot cross-lingual transfer in legal topic classification and demonstrated that translation-based methods outperform multilingually pre-trained models by a wide margin.

Additionally, (Karan et al., 2016) addressed the challenge of supervised topic classification in Croatian political texts by developing a new dataset comprising 7,300 titles, which were manually coded according to the Comparative Agendas Project codebook. (Glavas et al., 2017) introduces a method for cross-lingual topic coding of sentences from political parties' electoral manifestos in various languages that outperform those models trained solely on one language. (Wei and Santos, 2020) have developed a sequence classifier to predict the origin of texts in history book excerpts and newspaper articles about the Israeli-Palestinian conflict. (Osnabrügge et al., 2021) studied and assessed the use of supervised-learning approaches to cross-domain topic classification of political texts.

(Akdemir and Hurriyetoglu, 2022) explored the method of framing the classification task as an entailment problem and utilizing zero-shot ranking for socio-political texts. (Wang, 2023) developed supervised topic classification with pre-trained language models as an alternative to cross-domain topic classification to overcome the data scarcity problem. (Ness et al., 2023) proposed a model that incorporates articles from three distinct sources, where BERT outperformed the other classification models.

The mentioned background study motivates us to build an automated topic classification model in Hindu mythology. Additionally, we have realized the importance of preparing a benchmark dataset to build the models.

2.3 Ontology and Character-Topic Relationship

Research on domain ontology and character-topic relationships in religious texts, such as `the Srimad-Bhagavatam`, has seen significant scholarly attention, primarily focusing on enhancing understanding and interpretation through structured frameworks. While there

is a wealth of literature on ontology development and application in various domains, few studies specifically address the ontological representation of characters and topics and their relationships within religious texts. In order to recognize the research gap and bridge the character and topic ontology models, we provide references to previous studies across a range of domains.

The paper presented by (Varadarajan et al., 2022) introduces GENOME, a novel methodology addressing shortcomings in ontological modelling of epics, offering a unified solution for iterative development and potential reuse of existing models.

(Syamili and Rekha, 2017) presented the development of an ontology for ancient Greek mythology, addressing a significant gap in existing ontological research. The Ontology of Greek Mythology (OGM) presented by (Pastor-Sánchez et al., 2021) addresses the absence of a semantic representation of Greek mythology on the Semantic Web by utilizing Wikidata items and properties to construct a comprehensive ontology.

(Watrobski, 2020) proposes a comprehensive approach to ontology learning methods, aiming to gather knowledge from various sources and highlight their differences, drawing on ontology engineering best practices. (Espinoza-Arias et al., 2018) explores existing smart city ontologies, highlighting their diverse ontological commitments and lack of interoperability.

(Noy and McGuinness, 2001) trace the evolution of ontologies from AI research labs to widespread adoption on the World Wide Web. They highlight initiatives like RDF and DAML designed to encode knowledge for electronic agents. The paper discusses the development of standardized ontologies across diverse fields, including medicine (e.g., SNOMED) and general-purpose domains like UNSPSC for product and service terminology. (Mohan and Arumugam, 2010) introduces a methodology leveraging Protege for developing and inferring the Indian medicinal plant ontology, crucial for understanding and utilizing medicinal plants properties and classifications, thus advancing knowledge dissemination and application in healthcare. The study proposed by (Sanabila and Manurung, 2014) outlines an automated approach for constructing and populating an ontology of wayang mythology using relation extraction and clustering techniques, evaluated against a reference ontology to assess its accuracy and completeness.

(Tuamsuk et al., 2018) conducted research to merge digital humanities concepts with the study of folktales in the Greater Mekong Subregion (GMS). They collaborated with experts in folktales, literary studies, and ontology development to create domain-specific ontologies comprising 74 concepts.

(Sattar et al., 2020) reviewed ontology development methodologies in the Interlocking Institutional Worlds (IWs) domain from 2015 to 2020. Their goal was to identify strengths and weaknesses in different approaches and provide guidelines for selecting methodologies that effectively manage the knowledge in IWs. (Pannach and Blaschke, 2023) present a technique to organize background details in various myth variants, using twelve versions of the Orpheus and Eurydice myth as a focal point. They employ shallow ontologies to simplify comparison among different versions, assisting scholars in identifying and understanding variations more effectively.

(Afolabi et al., 2014) created a domain ontology for Nigerian history to overcome the issue of imprecise and anecdotal documentation, employing a semi-automated method. (Gersberg and Ebecken, 2014) propose a method for swiftly constructing domain ontologies starting from scratch, by identifying contextual concepts and leveraging text mining, link analysis, and graph analysis methods. Additionally, they develop a web prototype tool for visualization and content retrieval to complement this approach.

(Paul et al., 2024) prepared and annotated dataset with over 200k verses from 7 major Hindu texts, labeled for canto, topic, and sentence. It also includes Similarity-based and Log-likelihood-based datasets created via sentence clustering. The BERT, RoBERTa, and DistilBERT models are evaluated for classification tasks, with clustering improving results on the Similarity-based dataset, while the Log-likelihood-based dataset remains difficult.

2.4 Application development in Hindu Mythology

In this thesis we consider MythoBERT 1.0 and Mytho-Annotator as developed applications within Hindu mythology.

2.4.1 MythoBERT 1.0

The evolution of language models has seen significant advancements in recent years, greatly impacting the field of Natural Language Processing (NLP). This section reviews key contributions that are relevant to our work, focusing on BERT and its variants, as well as other specialized models.

BERT and Its Variants: The seminal work by Devlin et al. (2019) introduced BERT (Bidirectional Encoder Representations from Transformers), which revolutionized NLP by leveraging deep bidirectional transformers for pre-training on a large corpus and fine-tuning on specific tasks. BERT's ability to capture context from both directions has set a new standard in language understanding and has been widely adopted across various

applications.

Following BERT's success, several domain-specific variants have emerged. BioBERT, developed by Lee et al. (2019), adapts BERT for biomedical text mining by pre-training on large-scale biomedical corpora. This model has shown significant improvements in tasks related to biomedical text processing, demonstrating the value of domain-specific pre-training. Similarly, SciBERT, introduced by Beltagy et al. (2019), extends BERT for scientific literature, improving performance on tasks within the scientific domain.

Specialized Models and Applications: LegalBERT, presented by Chalkidis et al. (2020), applies BERT to legal texts, enhancing its performance in legal document classification and similarity measurement. ClinicalBERT, developed by Huang et al. (2019), focuses on clinical notes and has demonstrated improvements in predicting hospital readmission, showcasing the adaptability of BERT to healthcare applications.

In addition to domain-specific adaptations, there have been efforts to improve BERT's embeddings for specialized tasks. Phrase-BERT, introduced by Wang et al. (2021), enhances phrase embeddings for better corpus exploration, while ArgueBERT by Behrendt and Harmeling (2021) improves argument similarity measurements. ColBERT, developed by Khattab and Zaharia (2020), presents an efficient approach to passage search by employing late interaction over BERT.

Feng et al. (2020) introduce Language-agnostic BERT Sentence Embedding, which enhances BERT's capabilities by developing sentence embeddings that work effectively across multiple languages. This advancement broadens the applicability of BERT-based models in multilingual contexts, enabling better performance and integration in diverse linguistic settings.

Xu et al. (2021) present a comparative study on Contextualized Embeddings for Neural Machine Translation. This work investigates the performance of various BERT-based models, including mBERT and BiBERT, in the context of neural machine translation. Their findings contribute to refining the understanding of how different contextualized embeddings can be optimized for translation tasks, offering insights into improving model performance for multilingual translation applications.

Embedding Relations and Knowledge Graphs: RelBERT, proposed by Ushio et al. (2023), extends BERT for embedding relations within textual data, providing insights into relational patterns that can be leveraged for knowledge graph construction. This model emphasizes the importance of relational understanding in enhancing language models' capabilities.

Fundamental Techniques and Architectures: The foundational principles of transformer models are rooted in the work by Vaswani et al. (2023), which introduced the Transformer architecture and the attention mechanism that underpins BERT and its derivatives. Additionally, Wu et al. (2016) highlighted the role of neural machine translation systems in bridging gaps between human and machine translation, further influencing subsequent developments in language modeling.

Optimization and Fine-tuning: Finally, the Adam optimization algorithm, described by Kingma and Ba (2014), has become a standard in training deep learning models, including language models like BERT, due to its effectiveness in handling large-scale optimization problems. The Universal Language Model Fine-tuning (ULMFiT) approach by Howard and Ruder (2018) also plays a crucial role in text classification, demonstrating the benefits of fine-tuning pre-trained models on specific tasks.

2.4.2 Mytho-Annotator

In the realm of Natural Language Processing (NLP) annotation tools, (Bikaun et al., 2022) present **Quickgraph**, an annotation tool for knowledge graph extraction from technical text, while (Frei et al., 2022) introduce **DrNote**, an open medical annotation service. (Mondal et al., 2022) utilized a medical annotation system to prepare a structured medical corpus. (Dutta et al., 2020) proposed an annotation system to annotate healthcare information from tweets. (Bojars et al., 2018) developed a Semantic Annotation Tool for Cultural Heritage Content, whereas (Tyers et al., 2017) present **UD Annotatrix**, an annotation tool for universal dependencies. (Kiesel et al., 2017) introduced **WAT-SL**, a web annotation tool for segment labeling, and (de Castilho et al., 2014) propose **WebAnno**, a flexible, web-based annotation tool for CLARIN. (Bollmann et al., 2014) introduced **CorA**, a web-based annotation tool for historical and non-standard language data, while (Stenetorp et al., 2012) present **BRAT**, a web-based tool for NLP-assisted text annotation. (Tesconi et al., 2010) developed **KAFnotator**, a multilingual semantic text annotation tool, and (Kenter and Maynard, 2005) utilized **GATE** as an annotation tool for various applications. (Maeda and Strassel, 2004) discussed annotation tools for large-scale corpus development, and (Morton and LaCivita, 2003) introduced **WordFreak**, an open tool for linguistic annotation. (Cunningham, 2002) presents **GATE**, a general architecture for text engineering, providing a comprehensive framework for text processing and annotation tasks.

2.5 Visualization of Character-centric Summary

Text Summarization has seen great progress with models like T5, BART, and Pegasus, which leverage deep learning for summarizing text effectively. T5 (Raffel et al., 2020) treats all NLP tasks as text-to-text problems, excelling in generating coherent summaries by converting tasks into text generation. BART (Lewis et al., 2020) uses a denoising autoencoder to reconstruct corrupted input, creating fluent and ac-curate summaries by understanding text structure and context.

Pegasus (Zhang and et al., 2020) is designed for abstractive summarization, using gap-sentence generation during pre-training, making it highly effective for generating rich, context-aware summaries. Further research includes BERTSUM (Lin and Liu, 2019), which enhances extractive summarization with BERT embeddings, and Hierarchical Transformer (Liu and Lapata, 2019), which handles complex, long documents. SummaRuNNer (Gao et al., 2020) leverages attention mechanisms to generate abstractive summaries.

Other contributions include optimizing encoder-decoder models for coherence (Zhou et al., 2021) and using contrastive learning to enhance key sentence extraction (Gong and Liu, 2021). Despite advancements, summarizing mythological texts remains underexplored. Applying these models can simplify narratives, support comparative studies, aid digital projects, and create accessible educational resources, enhancing understanding and relevance of mythology.

Keyphrase extraction identifies meaningful phrases in a text, supporting indexing, summarization, and retrieval. In mythology, where texts are rich and symbolic, keyphrase extraction helps capture essential characters and narratives.

KeyBERT (Grootendorst, 2020) is an unsupervised model that uses BERT embeddings to extract keyphrases without labeled data, excelling in document classification and summarization tasks. EmbedRank (Bennani-Smires et al., 2018) combines embeddings with graph-based ranking to find keyphrases, ranking them based on semantic similarity to the text. It works well in unsupervised keyphrase extraction tasks.

PositionRank (Florescu and Caragea, 2017) (Florescu & Caragea, 2017) enhances keyphrase extraction by consid-ering word position and co-occurrence. It is ideal for long-form texts like mythological narratives, improving document analysis. CopyRNN (Meng et al., 2017) (Meng et al., 2017) uses a supervised approach to generate keyphrases, even those not in the text, making it effective for summarizing research papers and complex domains like mythology.

Multi-Head Attention Networks (Vaswani et al., 2023) leverage attention mechanisms to focus on important areas of the text, aiding in document summarization and keyphrase extraction. Yake! (Campos et al., 2020) is a lightweight, unsupervised model that relies on word frequency and position to extract keyphrases, performing well across different domains, including long texts. Attention-based Neural Keyphrase Extraction (Sun et al., 2019) uses neural networks with attention to focus on document structure and content, improving context-dependent keyphrase extraction.

In mythology, keyphrase extraction faces challenges like capturing symbolic meanings and multifaceted roles of characters. Existing models need adaptation to meet these complexities, but they hold potential to enhance comparative analysis, summarization, and content retrieval in mythological texts.

2.6 Comparative Analysis and Research Gap

Most existing approaches to character identification are built for structured, modern narratives and fail to accommodate the complexity and symbolic depth of mythological characters. Mythological texts often involve multifaceted characters with multiple identities, roles, and symbolic representations, which existing systems struggle to handle. Techniques like those presented above may not fully capture the cultural and contextual nuances required to identify characters in such texts.

While significant progress has been made in character identification in general narratives, there remains a noticeable gap in applying these techniques to mythological texts, where the complexity and richness of characters demand specialized approaches.

The existing works on topic classification across legal, political, and historical domains provide a foundation for our research, but they also highlight key differences when applied to mythological texts. For instance, (Noguti et al., 2020) and (Braun and Matthes, 2022) effectively used linear models and neural networks for legal document classification, where the structure is formalized and predictable. In contrast, mythological texts are rich in symbolic meaning and narrative complexity, requiring models that can grasp both surface-level topics and deeper, implicit themes. Political text classification, such as the work by (Karan et al., 2016) and (Glavas et al., 2017), introduced challenges in multilingual and cross-lingual environments, which are relevant but still distinct from the unique linguistic and cultural diversity in Hindu mythology. Moreover, these studies typically rely on labeled datasets and fixed taxonomies, which may not fully capture the nuanced, multi-faceted characters and events in mythological stories. Our approach contrasts with these

by focusing specifically on the adaptation of state-of-the-art transformers to classify topics in mythology, filling a gap in the research where narrative and cultural richness have been underrepresented in prior classification models.

Despite extensive research in domain ontology across various fields, significant contrasts emerge when comparing approaches in mythological and religious texts. Studies like (Varadarajan et al., 2022), which introduces GENOME for epic ontology modeling, focus on creating reusable and adaptable models, filling a gap in structured frameworks for complex narratives. In contrast, ontologies in other domains, such as those proposed by (Syamili and Rekha, 2017) and (Pastor-Sánchez et al., 2021) for Greek mythology, emphasize representing mythological figures and relationships, but often lack the flexibility and scalability needed for iterative development. While (Mohan and Arumugam, 2010) showcases the use of ontologies in medicinal plants and healthcare, addressing practical applications, the ontological frameworks for religious texts like `the Srimad-Bhagavatam` remain under-explored. Approaches like (Sattar et al., 2020) in institutional knowledge and (Watrobski, 2020) in ontology learning offer comprehensive insights into capturing domain-specific knowledge but struggle to incorporate the symbolic complexity and narrative depth found in religious or mythological texts. In particular, the work by (Paul et al., 2024) on annotating Hindu scriptures highlights the challenge of contextualizing vast and intricate datasets, something current ontology frameworks from other domains may find difficult to manage without further adaptations. This contrast underscores the need for tailored, domain-specific ontologies that bridge the gap between character-topic relationships and cultural narratives in religious texts.

MythoBERT 1.0- The primary research gap in MythoBERT 1.0 lies in its limited capacity to handle the complex symbolic language, multi-layered narratives, and culturally specific references inherent in mythological texts, revealing a need for further model refinement to enhance its interpretative accuracy in this unique domain. While BERT and its domain-specific variants such as BioBERT (Lee et al., 2019), SciBERT (Beltagy et al., 2019), and LegalBERT (Chalkidis et al., 2020) have demonstrated remarkable success in improving task performance within specialized domains, these models are designed for highly structured text like scientific, legal, and biomedical literature. They excel in capturing specific domain knowledge but are less effective when applied to complex, culturally-rich narratives like Hindu mythology, where character interactions and symbolic meaning play a crucial role. Efforts like Phrase-BERT (Wang et al., 2021) and ColBERT (Khattab and Zaharia, 2020) have focused on improving embeddings for spe-

cific tasks such as phrase similarity or passage retrieval, but they do not address the unique challenges posed by mythological texts, which require deeper contextual and narrative understanding. While models like RelBERT (Ushio et al., 2023) have extended BERT for relation embedding, these approaches still fall short in handling the symbolic and historical nuances present in mythological narratives. Therefore, a significant gap remains for models specifically designed for mythological texts, which led to the development of more culturally and contextually tuned models for this under-explored domain.

Mytho-Annotator- While numerous annotation tools have been developed across domains such as medicine, technical text, and cultural heritage, a gap remains in the context of mythological texts. Tools like Quickgraph (Bikaun et al., 2022) and DrNote (Frei et al., 2022) are tailored for structured domains like knowledge graphs and medical data, where the content is often factual and well-defined. In contrast, mythological narratives are rich with symbolic meanings, complex character interactions, and evolving themes, which require more nuanced annotation capabilities. Tools such as WebAnno (de Castilho et al., 2014) and BRAT (Stenetorp et al., 2012) focus on flexible, web-based annotation but may lack the depth needed for capturing the cultural and narrative complexities of mythology. While these existing tools excel in their respective domains, they fall short of addressing the intricacies of mythological texts, highlighting the need for a specialized annotation system that can handle the unique challenges of mythological character identification, narrative role classification, and symbolic analysis.

The research gap in this study lies in the underutilization of advanced text summarization and keyphrase extraction models—such as T5, BART, Pegasus, KeyBERT, and EmbedRank—specifically in the context of mythological texts. While these models have demonstrated effectiveness in generating coherent and contextually relevant summaries for general text, their potential to capture the unique characteristics of mythological narratives remains largely unexplored.

First, mythological narratives often contain intricate character relationships, layered meanings, and rich symbolism, which traditional summarization models have not been explicitly designed to handle. This complexity necessitates a tailored approach to ensure that summarization accurately reflects the depth and nuance of these texts.

Second, although models like T5 and BART possess frameworks that could theoretically be beneficial for summarizing mythology, their application to these texts has not been thoroughly investigated. For instance, while Pegasus excels in abstractive summarization, it may lack the contextual awareness needed to interpret the multifaceted meanings

embedded in mythological stories.

Third, current keyphrase extraction models, such as KeyBERT and PositionRank, focus on identifying significant phrases but may fail to capture the symbolic nuances and thematic intricacies that characterize mythological literature. The relationships between characters, their roles, and overarching themes may be overlooked, leading to insufficient representation of the source material.

Furthermore, the existing methodologies for summarization and keyphrase extraction require adaptation to effectively address the specific challenges posed by mythological texts. This includes the development of models that can consider the narrative's character-centric aspects and the interplay of symbolic meanings to facilitate more accurate summaries and keyphrase identification.

Addressing this research gap could pave the way for enhanced comparative analyses of mythological texts across different cultures and traditions. Additionally, it would provide educational resources that facilitate a deeper understanding of these narratives, enriching both scholarly research and public engagement with mythology.

Finally, while significant advancements have been made in text summarization and keyphrase extraction, their direct application to mythological texts presents unique challenges that require focused research and model adaptation to unlock their full potential in this area.

This literature survey has explored key contributions to the field of character-centric analysis in Indian mythology, encompassing various dimensions such as Character Identification, Topic Classification, and the development of tools like Mytho-Annotator and MythoBERT. Additionally, the exploration of Domain Ontology and Character Topic Relationships has provided valuable insights into the complex interactions within mythological narratives. The emphasis on Character-Centric Summary and Visualization highlights the potential for enhancing user engagement and understanding of these rich texts. Collectively, these studies illuminate the evolving landscape of research in this domain and set a solid foundation for future investigations that can further bridge the gap between mythology and advanced computational techniques.

Chapter 3

Identification of Characters in Hindu Mythology

Mythological Characters

3.1 Introduction

The identification of characters in narrative texts has garnered significant attention in recent research. A character, often referred to as a fictional figure, represents an entity within a narrative work of art. In literature, characters lead readers through the story, aiding in the understanding of plots and themes (Freeman, 2016). According to (Iosif and Mishra, 2014), characters can be either human or non-human, including animals and inanimate objects, often depicted with human-like qualities.

The interactions among characters can be categorized as either human-to-human or human-to-nonhuman. Notably, a character does not always need to be a speaker within the narrative; there are instances where a character exists without any associated dialogue or monologue, indicating that they may not serve as an active speaker. Characters play crucial roles in comprehending the narrative context, thus enhancing readers' understanding of the story. Generally, characters can be classified into two types: protagonists and antagonists. A protagonist is the central character who influences the actions of other characters and drives the story forward (Duncan, 2006), while an antagonist similarly affects the narrative trajectory.

In the context of analyzing characters in ancient Indian epics, it is crucial to consider a range of texts. In addition to the `Mahabharata`, which centers on the conflict between the Pandavas and the Kauravas, other texts offer rich narratives involving multiple characters. These include the `Ramayana`, the story of Rama's quest to rescue his wife, Sita, from the demon king Ravana; the `Srimad-Bhagavatam`, which presents the life and

teachings of Lord Krsna; the `Devi Bhagavata`, which focuses on the Goddess Devi and her divine acts; the `Caitanya Caritamrita or Chaitanya Charitamrita`, documenting the life and teachings of Chaitanya Mahaprabhu; the `Harivamsha Purana`, which serves as a supplement to the Mahabharata and delves into the lineage of Krsna; and `Krsna, the Supreme Personality of Godhead`, a detailed account of Krishna's life, emphasizing his divine and human attributes. In this research work, we will refer to the texts by the following abbreviations: **RAMA** for `the Ramayana`, **MBH** for `the Mahabharata`, **DEVI** for `the Devi Bhagavata`, **HVM** for `the Harivamsha Purana`, **SB** for *the Srimad-Bhagavatam*, **CC** for *the Caitanya Caritamrita or Chaitanya Charitamrita*, and **KRSNA** for `Krsna, the Supreme Personality of Godhead`. These abbreviations will be used consistently throughout the document for clarity and brevity.

Characters may be represented at the word level (NNP) or phrase level ($NP \ll NNP$). For instance, `Janamejaya` (NNP) represents a character at the word level, while `Janamejaya, the son of Parikshit` ($NP \ll NNP$) denotes a character at the phrase level. However, relying solely on NNP or ($NP \ll NNP$) is insufficient for accurate character identification. Thus, the identification of characters at both word and phrase levels is the primary research issue addressed in this chapter.

The presence of adjectives, and sometimes adverbs, in these noun phrases serves as critical attributes that reveal character traits and qualities. For instance in **MBH**, in the following phrase **Yudhisthira**, the full characterization as **The Kuru King Yudhisthira**, highlights his role as the ruler of the Kuru dynasty, showcasing the importance of the descriptive element **The Kuru King**.

Example 1: **The Kuru King**$_{adj}$ **Yudhisthira**$_{character}$

Similarly, in another example, **Krishna** can be referred to as **the highly intelligent and high-souled Krishna**, emphasizing his attributes within the Mahabharata.

Example 2: **the highly intelligent**$_{adj}$ **and**$_{cc}$ **high-souled**$_{adj}$ **Krishna**$_{character}$

The identification process followed specific policies to determine whether a phrase qualifies as a character or character adjective. Some important policies are outlined below:

- Every name of a person followed by a verb is a character.
  - e.g., `Yudhisthira`
- Each name of a person accompanied by its qualities, often mentioned before or after the name, is a character.

  - e.g., `the wonderful warrior Drona`, `Arjuna the foremost`

- Living, non-living, and celestial beings that perform actions such as speaking, walking, or feeling, and actively participate in the narrative, are considered characters.

  - e.g., `the celestial Sakti`, `the celestial Ganges`

- Any word related to a person's profession (e.g., sage, brahmana) is regarded as a character.

  - e.g., `The Asura architect`

- An animal that plays an active role in the text is also classified as a character.

  - e.g., `the celestial steed Uchchaihsrava`

- Any special weapon recognized for its destructive power is termed a character due to its specific identity, such as:

  - e.g., `the Sudarshana Chakra (the celestial disc)`, `the terrible weapon Narayana`

This research tackles the challenges of character identification in narrative texts using two complementary approaches. The first approach, **Task 1**, captures descriptive noun phrases, including key adjectives and modifiers, to identify characters. The second approach focuses on, **Task 2**, the identification of characters at both word and phrase levels.

This research work is structured as follows: we discuss about the mythological texts followed by the preparation of datasets , Classifier Models, experiments and result analysis for the tasks, error analysis and finally summary.

3.2 Mythological Texts

In this research, we considered the entire text of MBH and a partial set of sentences from all others mythological texts. The table 3.1 presents statistics of several Hindu mythological texts, highlighting key aspects such as the total number of sentences, total uni-grams, average sentence length, and the percentage of sentences considered for analysis. The texts analyzed include Mahabharata (MBH), Ramayana (RAMA), Srimad-Bhagavatam (SB), Devi Bhagavata Purana (DEVI), Caitanya Caritamrita or Chaitanya Charitamrita (CC), Harivamsha Purana (HVM), and Krsna, the Supreme Personality of Godhead (KRSNA).

Mahabharata (MBH) has the highest number of sentences (117.3k) and total uni-grams (1,711.9k), with an average sentence length of 16.8 words, and 100% of its sentences are considered in the analysis. For all other texts, only 20% of the total sentences are considered, with the Ramayana (RAMA) containing 20.5k sentences, Srimad-Bhagavatam (SB) having 18.2k, and Devi Bhagavata Purana (DEVI) having 34.2k. The average sentence lengths vary between 14 and 16 words across these texts.

Table 3.1: Statistics of Mythological Texts

Mytho Text	**Total Sentences**	**Total Uni-grams**	**Average Sentence Length (words)**	**% of Sentences Considered**
MBH	117.3k	1,711.9k	16.8	100%
RAMA	20.5k	290.3k	14.2	20%
SB	18.2k	273.0k	15.0	
DEVI	34.2k	522.0k	15.3	
CC	20.1k	311.6k	15.5	
HVM	14.7k	206.0k	14.0	
KRSNA	14.6k	233.6k	16.0	

In these texts it is essential to note that not only proper names identify characters; noun phrases can also refer to characters, necessitating the extraction, identification, and analysis of such phrases to capture significant attributes associated with each character.

3.3 Task 1: Identification of Characters

3.3.1 Preparation of datasets

In **Task 1** we employ two distinct approaches to identify characters within the text. We have established a set of `97 features` at the word level and a set of `51 features` at the phrase level. Utilizing these features, we developed our datasets and devised two training models based on a semi-supervised approach. Subsequently, we identified the significant features and their interrelations. We then evaluated our model, assessing precision, recall, F-measure, kappa, and error rates.

We observe that numerous characters play significant roles within these texts. Initially, we manually annotated these characters and compiled a list. To analyze the positions and occurrences of each character, we examined each sentence in the texts using the Stanford

Table 3.2: Confusion matrix of Not_a_Character by Annotator 1 and 2

Character Identified in dataset (D = 159487)		Annotator 1	
		Yes	No
Annotator 2	Yes	156300	200
	No	400	2587
Not_a_Character Identified in dataset (D = 98211)		Annotator 1	
		Yes	No
Annotator 2	Yes	96000	200
	No	200	1811

CoreNLP suite[1]. Each sentence was tokenized, annotated with part-of-speech (POS) tags, and a syntactic parse tree was generated by the suite. After a thorough examination of each sentence in the texts, we developed the notion that characters can be identified at both the word and phrase levels. We also noted that, in most cases, a word is considered a character at the word level when its POS tag is NNP. Similarly, at the phrase level, a phrase is deemed a character when the root of the phrase is NP and one of its descendants is NNP. Examples are provided below.

At word Level: `(NNP Narayana)` = $\texttt{[Narayana]}_{\text{Character}}$

At phrase Level:

`(NP (DT the) (JJ holy) (NNP Rishi) (NNP Vyasa))` = $\texttt{[The holy Rishi Vyasa]}_{\text{Character}}$

The above observation helps us to extract different features at word level and phrase level. The total number of instances we observed is 257698.

Inter Annotator Agreement:

In the analysis of inter-annotator agreement, two confusion matrices in Table 3.2 were evaluated to determine the consistency between Annotator 1 and Annotator 2 in categorizing instances as either *Character Identified* (D_{char} = 159487) or *Not_a_Character Identified* ($D_{notchar}$ = 98211). The kappa score for *Character Identified* was calculated to be approximately $\kappa \approx 0.894$, based on the confusion matrix values where Annotator 1 identified 154000 characters as **Yes** and 2900 as **No**, while Annotator 2 identified 156300 characters as **Yes** and 200 as **No**. This indicates a very high level of agreement, suggesting that the annotators were largely in sync regarding their assessments.

Similarly, the kappa score for *Not_a_Character Identified* yielded an even higher value

[1] https://stanfordnlp.github.io/CoreNLP/

of approximately $\kappa \approx 0.898$. In this case, the confusion matrix revealed that Annotator 1 classified 96000 instances as **Yes** and 200 as **No**, while Annotator 2 classified 75000 instances as **Yes** and 400 as **No**. This further reinforces the findings of a robust consensus between the two annotators in their classification efforts. The close agreement not only reflects the effectiveness of the annotators in distinguishing between characters and non-characters but also highlights the reliability of the dataset itself.

3.3.2 Feature Engineering

Word Level Features : For each NNP present in a sentence at word level we have considered 97 different features.

Table 3.3: List of Word Level Features (WL_F)

Word Level Features(WLF)								
Basic Features			Immediate Pre and Post . . . Features of Cw			Immediate Pre and Post . . . Distance from Cw		
Sl(W)	Name	Freq.	Sl(W)	Name	Freq.	Sl(W)	Name	Freq.
1	Extracted NNP word(Cw)	4152	14-17	verb word and tag	2645;3158	70,71	verb distance	2645;3158
2	NNP-tag	4152	18-21	adverb word and tag	1218;1381	72,73	adverb distance	1218;1381
3	Length of Cw	4152	22-25	preposition word and tag	2590;3042	74,75	preposition distance	2590;3042
4	Starting Index of Cw	4152	26-29	noun word and tag	2741;3412	76,77	noun distance	2741;3412
5	Ending Index of Cw	4152	30-33	NNP word and tag	2199;2229	78,79	NNP distance	2199;2229
6	Previous word of Cw	3583	34-37	adjective word and tag	1445;2089	80,81	adjective distance	1445;2089
7	Previous word tag of Cw	2584	38-41	C. Conjunc. word and tag	1052;1611	82,83	C Conjunction distance	1052;1611
8	Next word of Cw	4152	42-45	determiner word and tag	2608;2554	84,85	determiner distance	2608;2554
9	Next word tag of Cw	4152	46-49	existential word and tag	92;43	86,87	existential distance	92;43
10	Porter Stemmer word of Cw	4152	50-53	interjection word and tag	1;1	88,89	interjection distance	1;1
11	Is porter Stemmed word same with Cw?	4152	54-57	TO word and tag	541;959	90,91	TO distance	541;959
12	Snowball Stemmer word of Cw	4152	58-61	Cardinal Number and tag	255;263	92,93	Cardinal Number distance	255;263
13	Is snowball stemmed word same with Cw?	4152	62-65	pronoun word and tag	1162;1773	94,95	pronoun distance	1162;1773
			66-69	Wh word and tag	625;771	96,97	Wh distance	625;771

In Table 3.3, we categorize the set of features into three distinct subcategories. The features labeled from W1 to W13 are classified as basic features of a context word (Cw). Features W14 to W69 are classified as immediate pre- and post-words and tags of Cw. Features W70 to W97 pertain to the verb situated to the left of Cw, referred to as immediate pre-verb distance, which has a frequency of 2645. Similarly, W71 calculates the word distance of the verb situated to the right of Cw, known as immediate post-verb distance, with a frequency of 3158.

To count the word distance from the context word (Cw) as immediate pre- and post-word distances, we consider their frequencies. Here, frequency indicates the number of occurrences of a distinct feature in our sample space. For instance, the feature set W14-17 (verb word and tag) consists of four different types of features. W14 represents the immediate pre-verb word situated to the left of Cw, while W15 denotes its POS tag, with

a frequency of 2645. Next, W16 refers to the immediate post-verb word situated to the right of Cw in a sentence, and W17 identifies its POS tag, with a frequency of 3158.

As another example, consider W70 and W71 (verb distance). Here, W70 calculates the word distance. Consider a sentence:

S_1 = `Having bowed down unto Narayana, and to Nara, the foremost of men, as also to the goddess Sarasvati, should the word Jaya be uttered.`

In the above sentence, our context word (Cw) is $\texttt{Narayana}_{\text{Character}}$. Some of the features extracted from the sentence S_1 with respect to $\texttt{Narayana}_{\text{Character}}$ are illustrated in Figure 3.1.

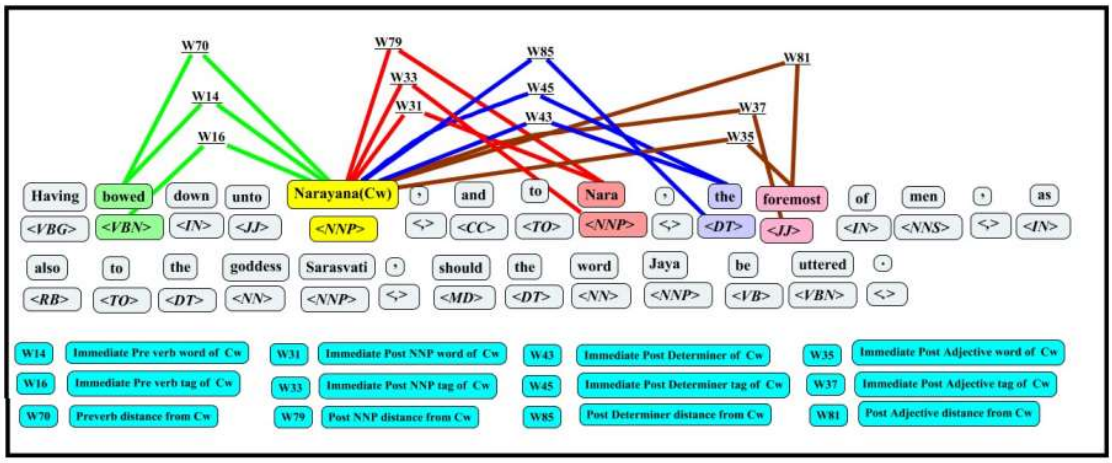


Figure 3.1: Example of Word Level Features

Phrase Level Features: At phrase level, we have considered 51 different features displayed in Table 3.4 for each $NP \ll NNP$ pattern present in the sentences

In Table 3.4, it can be observed that there are primarily two different subcategories of features. All the features from P1 to P27 are related to the phrase (Cw), which is assumed to be a Character, while the remaining features pertain to the two levels up ancestor (the parent of a parent of the current head node, AnCh), along with their frequencies.

In this context, frequency denotes the number of occurrences of a particular feature in the sample space. For example, P1 represents the current head node of the phrase (Ch) with a frequency of 2991, and P36 identifies the presence of any NP as a sibling of the ancestor node of Ch (AnCh), with a frequency of 1239. Consider the sentence S_2:

S_2 = **The king, in honour of Hari and naming him repeatedly, fed the Island-born Vyasa, and Narada, and Markandeya possessed of wealth of penances, and Yajnavalkya of Bharadwaja's race, with many delicious viands.**

Table 3.4: List of Phrase Level Features (PL_F)

Phrase Level Features (PL_F)					
Sl(P)	**Name**	**Freq.**	**Sl(P)**	**Name**	**Freq.**
1	Current head Node of the phrase (Ch)	2991	27	has STOP as siblings of Ch?	313
2	The pre-terminal yield Nodes of Ch	2991	28	Ancestor Node of Ch (AnCh)	2991
3	Leaves of the Ch (Cw)	2991	29	has ADJP as siblings of AnCh?	8
4	Path from Ch to ancestor Node	2991	30	has ADJP as siblings of AnCh?	88
5	has ADJP as siblings of Ch?	18	31	has CONJP as siblings of AnCh?	28
6	has ADJP as siblings of Ch?	128	32	has FRAG as siblings of AnCh?	1
7	has CONJP as siblings of Ch?	6	33	has INTJ as siblings of AnCh?	1
8	has FRAG as siblings of Ch?	1	34	has LST as siblings of AnCh?	1
9	has INTJ as siblings of Ch?	1	35	has NAC as siblings of AnCh?	1
10	has LST as siblings of Ch?	1	36	has NP as siblings of AnCh?	1239
11	has NAC as siblings of Ch?	1	37	has NX as siblings of AnCh?	1
12	has NP as siblings of Ch?	790	38	has PP as siblings of AnCh?	210
13	has NX as siblings of Ch?	1	39	has PRN as siblings of AnCh?	16
14	has PP as siblings of Ch?	271	40	has PRT as siblings of AnCh?	18
15	has PRN as siblings of Ch?	16	41	has QP as siblings of AnCh?	1
16	has PRT as siblings of Ch?	1	42	has RRC as siblings of AnCh?	1
17	has QP as siblings of Ch?	1	43	has UCP as siblings of AnCh?	1
18	has RRC as siblings of Ch?	3	44	has VP as siblings of AnCh?	470
19	has UCP as siblings of Ch?	3	45	has WHADJP as siblings of AnCh?	1
20	has VP as siblings of Ch?	619	46	has WHAVP as siblings of AnCh?	1
21	has WHADJP as siblings of Ch?	1	47	has WHNP as siblings of AnCh?	25
22	has WHAVP as siblings of Ch?	1	48	has WHPP as siblings of AnCh?	1
23	has WHNP as siblings of Ch?	1	49	has X as siblings of AnCh?	1
24	has WHPP as siblings of Ch?	1	50	has COMMA as siblings of AnCh?	706
25	has X as siblings of Ch?	4	51	has STOP as siblings of AnCh?	446
26	has COMMA as siblings of Ch?	788			

The important part of the parse tree of the above sentence S_2 is:

$$S_{2parsed} = \text{(VP (VBN fed) (NP (NP (DT the) (JJ Island-born) (NNP Vyasa)) (, ,) (CC and) (NP (NNP Narada))))}$$

In the above sentence, our target phrase is the `Island-born Vyasa`. Figure 3.2 provides a detailed explanation of the features P_1, P_2, P_3, and P_{28}.

Features Associativity Analysis: It is observed from the data sets, D_{wt} and D_{pt}, that some feature or set of features coexists with other feature or set of features. This type of relations can be found from the texts very frequently in our sample space. To address this issue we have applied **FP-Growth algorithm** in word and phrase level.

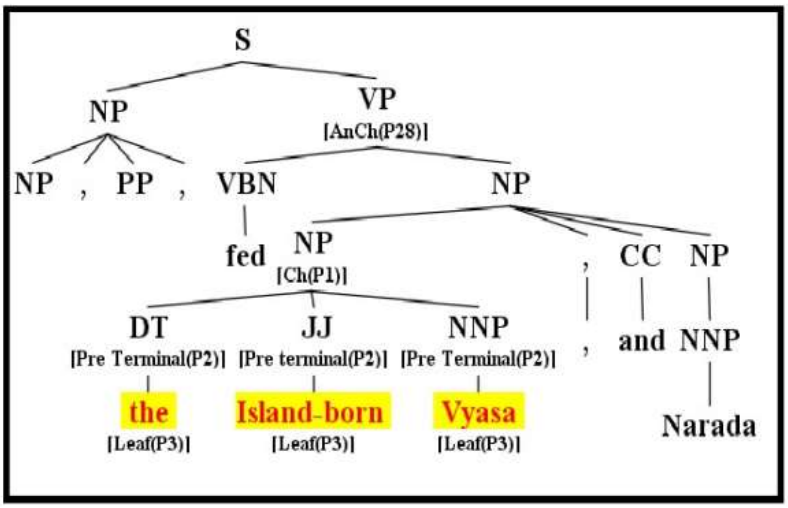


Figure 3.2: Example of Phrase Level Features P1,P2,P3,P28

This algorithm calculates all frequent feature/feature set from the data set by building a **FP-Tree data structure** on the data sets D_{wt} and D_{pt}. Some frequent relations of word and phrase level are given in table 3.5 and 3.6 .

From Table 3.5 we can understand that for a distinct context word, Cw,when we identify a value for the feature,W20 in a sentence in the sample space, simultaneously we can find a value for the feature W14 also with a confidence value 0.258. Similarly in the Table 3.6, when we can observe a value for the feature P3 then P26 is also observed for context word Cw in a sentence of our sample space with confidence value 0.261 . Where $X \Rightarrow Y$ implies that if X occurred then Y also occurred; X means antecedent and Y means consequent.

3.4 Task 2: Identification of Character Adjectives

In this study, **Task 2**, involves developing a set of rules to extract words, phrases, or groups of words from parsed sentences that are likely to represent characters. We then assess the efficacy of these rules using a range of linguistic and statistical features to identify the properties of the extracted textual units. Each unit is manually annotated as `Character` or `Not_a_Character` to create a comprehensive tagged dataset. Subsequently, we apply various classifiers to this dataset to evaluate *precision, recall, and F-measure*, followed by a discussion of results, error analysis, and observations. The primary aim of this approach is to not only identify characters in the text but also to uncover their attributive qualities, which we refer to as **character adjectives**.

Table 3.5: Features Associativity at Word Level

Word Level		
Antecedent	**Consequent**	**Confidence**
W20	W14	0.258
W20	W83,W35	0.408
W23	W83,W31,W27	0.352
W20,W35	W14	0.381
W20,W83	W35,W31	0.381
Antecedent ->Consequent		
W14=Immediate pre verb word of Cw		
W20=Immediate post adverb word of Cw		
W23=Immediate pre position tag of Cw		
W27=Immediate pre noun tag of Cw		
W31=Immediate pre NNP tag of Cw		
W35=Immediate pre adjective tag of Cw		
W83=Post C. Conjunction distance from Cw		

3.4.1 Preparation of datasets

The primary challenges in character identification involve managing sentences of varying lengths and recognizing multiple characters appearing in different text segments. We utilized the Stanford CoreNLP[2] suite for sentence tokenization and to annotate with Part-of-Speech (POS) tags and syntactic parse trees. Initial analysis of the parsed sentences revealed that noun phrases (NP) with a verb phrase (VP) as a right sibling are more likely to represent characters. Instances were also found where a proper noun (NNP) immediately follows an NP. Based on these observations, we developed a set of rules to extract subtrees from parsed sentences where the NP exhibits these characteristics, classifying them as entities. For instance:

$$\text{R1: } \{\text{NP}<\text{NNP } \${+}{+} \text{ VP, NP}<<-\text{NNP}\}$$

In this notation, X<Y means X immediately dominates Y in parse tree, X \$++ Y means X is a left sister of Y in parse tree, and X<< −Y means Y is the rightmost descendent of A in parse tree of a sentence.

[2] https://stanfordnlp.github.io/CoreNLP

Table 3.6: Features Associativity at Phrase Level

Phrase Level		
Antecedent	**Consequent**	**Confidence**
P3	P26	0.261
P26	P3	0.412
P3	P12	0.418
P12	P3	0.743
P26	P12	0.659
P12	P26	0.795
Antecedent ->Consequent		
P3 = Leaves of the Ch(Cw)		
P12= hasNP as siblings of Ch?		
P26= hasCOMMA as siblings of Ch?		

```
Entity (e_1) = (NP (DT the) (ADJP (RB highly) (JJ intelligent)
(CC and) (JJ high-souled)) (NNP Krishna))
= [the highly intelligent and high-souled Krishna]

Entity (e_2) = (NP (NNP Krishna))
= [Krishna]
```

The average Support (S_{Avg}) and Confidence (C_{Avg}) of each rule from all the texts has been given in Table 3.7.

Table 3.7: Average Support and Confidence of rules

Rule #	Rules	S_{Avg}	C_{Avg}
R1	NP<NNP	55.45	64.34
R2	NP<NNP $++ (VP<VBD)	7.67	89.35
R3	NP<NNP $++ (VP<VBG)	1.3	83.11
R4	NP<NNP $++ (VP<VBN)	1.16	79.2
R5	NP<NNP $++ (VP<VBP)	0.48	64.1
R6	NP<NNP $++ (VP<VBZ)	0.99	75.21

The table 3.7 presents the average support (S_{Avg}) and confidence (C_{Avg}) of six syntactic rules extracted from natural language text. Rule 1 (R1) shows the highest average support

(55.45%) but relatively lower confidence (64.34%), meaning it occurs more frequently but with moderate confidence. Rule 2 (R2) has high confidence (89.35%) but much lower support (7.67%). Other rules (R3 to R6) show decreasing support, but varying confidence levels, indicating how certain syntactic patterns, though less common, are strong indicators of specific grammatical structures.

Next,the effectiveness of a rule R can be evaluated based on its coverage and accuracy. Given a tuple X from a class-labeled dataset D, let N_{covers} represent the total number of tuples covered by R, N_{correct} denote the number of tuples correctly identified by R, and $|D|$ indicate the total number of tuples in D. The Coverage and Accuracy of R can be defined as follows:

$$\text{Coverage}(R) = \frac{N_{\text{covers}}}{|D|}, \quad \text{Accuracy}(R) = \frac{N_{\text{correct}}}{N_{\text{covers}}} \tag{3.1}$$

Table 3.8: Coverage and Accuracy of Rules

Rules #	N_{covers}	$N_{correct}$	Coverage %	Accuracy %
R1	201522	128053	85.45	63.54
R2	18622	16467	7.89	88.42
R3	3971	3280	1.68	82.59
R4	4417	3978	1.87	88.97
R5	3224	2677	1.36	85.82
R6	4054	3472	1.71	85.64

Coverage measures the percentage of tuples covered by the rule, while accuracy assesses the proportion of correctly identified tuples among those covered. The coverage and accuracy results for each rule are detailed in Table 3.8, where $|D| = 235810$.

The table 3.8 summarizes the coverage and accuracy of various rules used in a classification or extraction task, with each row representing a specific rule (R1 to R6) and its associated performance metrics. The N_{covers} column indicates the total number of instances each rule covers, with Rule 1 (R1) covering the most instances at 201,522, while Rule 2 (R2) covers 18,622. The $N_{correct}$ column shows the number of correctly identified instances, with R1 identifying 128,053 correctly.

The Coverage % reflects the portion of the total dataset addressed by each rule, with R1 covering 85.45%, significantly higher than the other rules. Conversely, the Accuracy % measures the correctness of the classifications, where R2 exhibits the highest accuracy at 88.42%, despite its lower coverage. This table illustrates the trade-off between coverage

and accuracy, emphasizing that while R1 covers more instances, it has lower accuracy compared to R2, which, although it covers fewer instances, demonstrates greater precision in its classifications. This information is valuable for guiding decisions on rule optimization in classification tasks.

3.4.2 Quality Measures of Rules

Sometimes, accuracy alone is not a reliable metric for assessing the quality of a rule. A rule might cover many tuples for a given class, but if most of these tuples belong to different classes, its effectiveness is compromised. Therefore, it is essential to utilize additional metrics for evaluating quality that integrate aspects of accuracy and coverage (Han et al., 2012). In this study, we examine three measures: Entropy (R), FOIL_Gain, and Likelihood Ratio statistics. The quality measures for each rule are presented in Table 3.9.

Table 3.9: Quality Measures of Each Rule

Rule #	Entropy	FOIL_GAIN	Likelihood
R1	0.30	-2540.24	380.35
R2	-3.53	2740.40	3196.96
R3	-48.77	301.00	249.28
R4	-39.35	300.00	196.06
R5	-71.41	-28.44	2.79
R6	-46.98	156.76	69.72

Table 3.9 presents the quality measures for each rule applied in the analysis, highlighting their performance in terms of entropy, FOIL gain, and likelihood. The entropy values indicate the level of uncertainty associated with each rule, where R1 exhibits an entropy of 0.30, suggesting a relatively low level of uncertainty, while the other rules display negative entropy values, reflecting varying degrees of uncertainty in their predictions. The FOIL gain measures the increase in accuracy due to applying the rules, with R2 achieving the highest gain of 2740.40, demonstrating its effectiveness in improving model performance. In terms of likelihood, R2 also stands out with a likelihood score of 3196.96, indicating a strong predictive capability, while R5 has the lowest likelihood score of 2.79, suggesting less reliability in its predictions. Overall, the table emphasizes the varying effectiveness of the rules, with R2 leading in both FOIL gain and likelihood, which may indicate its critical role in the modeling process.

In addition to these critical rules, we have sought to extract more features for use in

a machine learning framework. In this task, we present two types of features: linguistic features and statistical features for each of the entities. To the best of our knowledge, these features have not yet been explored in the literature for character identification.

3.4.3 Feature Engineering

Linguistic Features: To extract the linguistic features for each of the rules R_{e}, we have identified the following set of attributes:

- Current head node of the extracted entity (C_{h})
- The preterminal node list of C_{h} (P_{l})
- The desired character adjective entity (C_{adj})
- List of siblings of C_{h} (S_{l})
- List of preterminal yields of all siblings of C_{h} (SP_{l})
- Path from C_{h} to two levels up the parent node ($\mathrm{Path}_{\mathrm{2up}}$)
- Immediate ancestor node of C_{h} (AN_{n})
- List of head nodes of the siblings of the immediate ancestor node from C_{h} (AN_{l})
- List of preterminal yield nodes of the siblings of the immediate ancestor node from C_{h} (ANP_{l})

Consider the rule R_{e} defined as $\mathrm{NP} \ll \mathrm{NNP}\,\$ + + (\mathrm{VP} \ll \mathrm{VBD})$, and let the sentence S_1 be:

$\mathrm{S}_1 =$ `O ye ascetics, the great Vyasa hath composed one hundred and eighty-six sections in this Parva.`

The corresponding parsed tree of the sentence S_1 is given by:

$\mathrm{S}_{1\mathrm{p}}$ = (ROOT (S (NP-TMP (NP (NN O)) (NP (PRP ye) (NNS ascetics))) (, ,) (NP (DT the) (JJ great) (NNP Vyasa)) (VP (VBP hath) (VP (VBN composed) (NP (NP (CD one) (CD hundred) (CC and) (CD eighty-six) (NNS sections)) (PP (IN in) (NP (DT this) (NN Parva)))))) (. .)))

The linguistic features ($C_{\mathrm{h}}, P_{\mathrm{l}}, C_{\mathrm{adj}}, S_{\mathrm{l}}, SP_{\mathrm{l}}, \mathrm{Path}_{\mathrm{2up}}, AN_{\mathrm{n}}$) extracted from S_{1p} are illustrated in Figure 3.3.

Now, consider another sentence S_2 and its corresponding parsed tree S_{2p}:

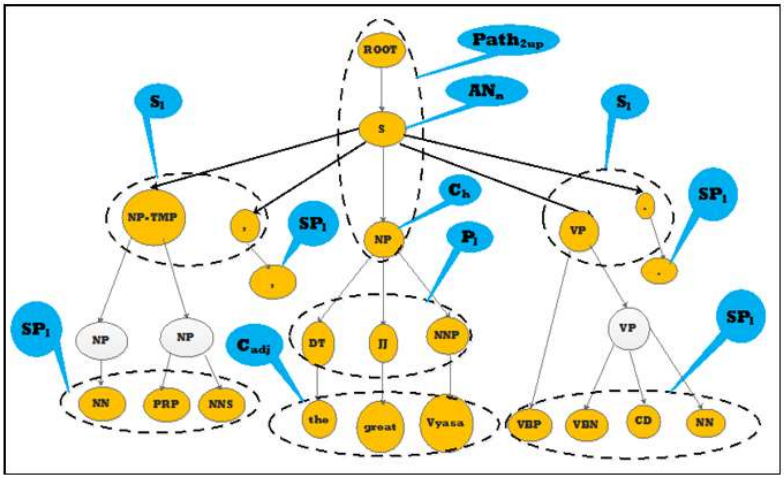


Figure 3.3: Example of Linguistic Features

S_2 = `The mighty Jayatsena, the son of Jarasandha, the prince of the Magadhas, O king, hath been slain in battle by the high-souled son of Subhadra.`

The parsed tree S_{2p} is given by:

S_{2p} = (ROOT (S (NP (NP (NP (DT The) (JJ mighty) (NNP Jayatsena)) (NP (NP (DT the) (NN son)) (PP (IN of) (NP (NNP Jarasandha))))) (, ,) (NP (NP (DT the) (NN prince)) (PP (IN of) (NP (NP (DT the) (NNPS Magadhas)) (, ,) (NP (NNP O) (NN king))))) (, ,)) (VP (VBP hath) (VP (VBN been) (VP (VBN slain) (PP (IN in) (NP (NN battle))) (PP (IN by) (NP (NP (DT the) (JJ high-souled) (NN son)) (PP (IN of) (NP (NNP Subhadra)))))))) (. .)))

The linguistic features AN_l and ANP_l extracted from S_{2p} are displayed in Figure 3.4.

Next, we applied the Stanford Universal Dependency parser[3] to each character adjective entity C_{adj}. The parser provides the relation, governor, and dependency of the words present in each NP entity as an additional set of features. These include the relation name of C_{adj} (STR_{n}), the governor value of C_{adj} (STG_{v}), the governor tag of C_{adj} (STG_{t}), the dependent value of C_{adj} (STD_{v}), and the dependent tag of C_{adj} (STD_{t}).

In S_{2p}, we have

$$C_{\text{adj}} = \texttt{The mighty Jayatsena.}$$

Its dependency relations and governor are detailed as follows:

det(mighty/JJ, The/DT) , appos(mighty/JJ, Jayatsena/NNP)

[3]https://stanfordnlp.github.io/CoreNLP

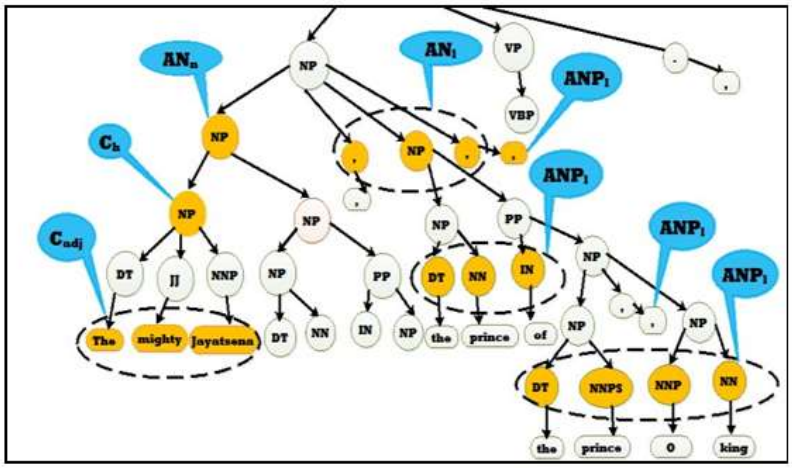


Figure 3.4: Example of Linguistic Features

The two relation names found for the desired character adjective entity, STR_{n}, are `det` and `appos`. The governor value of the desired character adjective entity, STG_{v}, is `mighty`. The governor tag of the desired character adjective entity, STG_{t}, is `JJ`. The dependent values of the desired character adjective entity, STD_{v}, are `The` and `Jayatsena`. The dependent tags of the desired character adjective entity, STD_{t}, are `DT` and `NNP`.

Statistical Features: To extract the statistical features, we computed the term frequency (TF) and term frequency-inverse document frequency (TF-IDF) for each character adjective entity (C_{adj}) identified in the various mythological texts, treating each text as a separate document. The analysis of the Term Frequency-Inverse Document Frequency (TF-IDF) reveals a variance of 0.02500 and a standard deviation of 0.15000. Additionally, we calculated the C-value and NC-value for each character adjective entity (C_{adj}). Among the extracted entities, we observed both single-word entities and multi-word term entities. The degree to which a linguistic unit relates to domain-specific concepts is referred to as Termhood (Frantzi et al., 2000). To evaluate the Termhood of each entity, we applied a modified C-value function to all of them. The C-value is a domain-independent approach that aims to enhance the extraction of nested terms, assigning a Termhood score to an entity and ranking it in the output list of each candidate character adjective (C_{adj}).

$$C\text{-}value(a) = \begin{cases} \log_2 |a| \cdot f(a), & \text{if } a \text{ is not nested,} \\ \log_2 |a| \left(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \right), & \text{otherwise.} \end{cases} \tag{3.2}$$

Where,

a = candidate character adjectives entity (C_{adj}),

b = longer entity,

$\|a\|$ = length of the entity (number of words),

f(a) = frequency of a in the corpus,

T_a = set of extracted candidate terms that contain a,

P(T_a) = number of candidate terms in T_a,

f(b) = frequency of longer entity b in the corpus.

Subsequently, we employed the NC-value method, which incorporates contextual information into the C-value method. This method re-ranks the C-value list for each candidate character adjective entity (C_{adj}).

The NC-value measure is formally described as:

$$NC\_value = 0.8C\_value(a) + 0.2 \sum_{b \in C_a} f_a(b)w(b) \tag{3.3}$$

where a = candidate character adjectives entity (C_{adj}), C_a = the set of distinct context words of a, $f_a(b)$ = the frequency of b as a term context word of a, and $w(b)$ = weight of b as a term context word. For example, if we consider

C_{adj} = `the Kuru king Yudhishthira`,

then its C-value is 36.00 and NC-value is 46.56. The details of average C and NC values are given in 3.10

Table 3.10: Average C-value and NC-value for Mythological Texts

Mythological Text	**Average C-value**	**Average NC-value**
MBH	36.00	46.56
RAMA	28.75	40.10
SB	25.00	35.50
DEVI	30.50	42.30
CC	20.80	30.70
HVM	22.25	33.40
KRSNA	24.00	36.00

The table 3.10 presents the average C-value and NC-value for various mythological texts, which are critical metrics for evaluating the significance of terms within these texts. The **C-value** quantifies the intrinsic importance of a term based on its frequency and context, while the **NC-value** incorporates additional contextual factors, enhancing its relevance. For instance, the **Mahabharata (MBH)** demonstrates a **C-value** of **36.00**

Table 3.11: Confusion matrix of Characters/Character Adjectives by Annotator 1 and 2

Character Identified in dataset (D = 157927)		Annotator 1	
		Yes	No
Annotator 2	Yes	154000	400
	No	500	3027
Not_a_Character Identified in dataset (D = 77883)		Annotator 1	
		Yes	No
Annotator 2	Yes	75000	200
	No	400	2283

and an **NC-value** of **46.56**, indicating its prominent role in the dataset. Other texts, such as **RAMA**, **SB**, **DEVI**, **CC**, **HVM**, and **KRSNA**, exhibit varying C-values and NC-values, reflecting their respective term significance and contextual relevance within the realm of Hindu mythology. This analysis underscores the diversity of term importance across different mythological narratives.

Thus, we collected all linguistic and statistical features and organized them in a set denoted as $Attr_{Total}$ for all candidate character adjectives (C_{adj}):

$Attr_{Total}$ = { { Re }, { Ch, Pl, Cadj, Sl, SPl, Path2up, ANn, ANl, ANPl },
{ STRn, STGv, STGt, STDv, STDt }, { TF, TF-IDF }, { C-value, NC-value } }

From these data preprocessing steps, we manually tagged all entities as Character and $Not\_a\_Character$. A total of 235810 objects extracted by the algorithm were manually annotated, and two independent domain experts were tasked with identifying characters and non-characters in the dataset D according to their logic and perception. Additionally, they identified characters along with their attributive qualities, termed as character adjectives.

Inter Annotator Agreement: Every object in dataset D possessing anthropomorphic traits is classified as a character. The confusion matrices for the identification of Character and $Not\_a\_Character$ provided by annotator 1 and annotator 2 are presented in Table 3.11.

Table 3.11 summarizes the agreement between two annotators in identifying Characters ($D_{\text{char}} = 157927$) and $Not\_a\_character$ adjectives within the dataset of size ($D_{\text{notchar}} = 77883$). Annotator 1 and Annotator 2 both identified characters with strong consistency, as shown by the high number of agreements: 154,000 instances where both annotators marked **Yes** and 3,027 where both marked **No**. There were relatively few disagreements,

with 400 instances where Annotator 2 said **No** while Annotator 1 said **Yes** and 500 instances where the opposite occurred. The kappa score for this part of the dataset was calculated to be approximately 0.8678, indicating a strong level of agreement beyond chance.

In case of *Not_a_Character*, containing 77,883 entries,both the annotators again showed a high level of agreement. There were 75,000 instances where both annotators agreed on marking **Yes** and 2,283 where both marked **No**. Disagreements were minimal, with 200 cases where Annotator 2 said **No** and Annotator 1 said **Yes**, and 400 cases where the reverse happened. The kappa score for this portion was calculated to be approximately 0.8814, again reflecting a very strong agreement between the annotators, with minimal disagreements in classification. Both kappa scores suggest almost perfect agreement, indicating that the annotators worked in close alignment with each other.

We conducted feature engineering on the dataset D, which comprises 15 linguistic features and 4 statistical features, totaling 235,810 entities. The dataset represents a two-class problem, with each entity manually labeled as *Character* or *Not_a_Character*. To meet the requirements of various classifiers, data preprocessing was performed to convert textual information into numerical values.

Our feature ablation studies were conducted in two stages: the first at the individual feature level using different attribute selection measures, and the second at the subset level utilizing *Forward Selection* and *Backward Elimination* schemes. Finally, we applied several classification algorithms available in the **RapidMiner Studio tool**[4] to our dataset along with the key attributes.

For the **Gain Ratio (Gr)**, the attribute with the maximum value was chosen as the splitting attribute, with the top three attributes identified as {NC-value, TF-IDF, C-value}.

Conversely, the **Gini Index (Gi)** serves as a measure of impurity in a dataset, where a higher weight for an attribute indicates greater relevance. The top three relevant attributes from our dataset according to this measure are $\{C_{adj}, P_l, STD_v\}$.

Finally, the **Chi-Squared Statistic** is a nonparametric statistical technique used to assess whether the distribution of observed frequencies differs from the expected theoretical frequencies. Attributes with higher Chi-Square values are deemed more relevant. The top three attributes selected based on the Chi-Squared Statistic from our dataset are $\{ANP_l, C_{adj}, SP_l\}$.

[4]RapidMiner Studio: https://rapidminer.com/products/studio/

Feature Subset Selection: We utilized two different schemes, namely Forward Selection and Backward Elimination, available in RapidMiner to identify various groups of relevant attributes or features. By employing the Forward Selection scheme, we obtained a new set of attributes denoted as FS_F, which is a subset of $Attr_{Total}$.

$FS_F = \{ R_e, P_l, C_{adj}, STG_t, STD_v, STD_t, \text{TF, TF-IDF} \}$ where $FS_F \subset Attr_{Total}$

Table 3.12: Attribute Selection measures

Attribute	Ig	Gr	Gi	Chi
R_e	0.028	0.038	0.016	7804.304
C_h	0	0	0	0
P_l	0.623	0.126	0.339	161885.5
C_{adj}	0.924	0.082	0.461	303530.6
S_l	0.261	0.045	0.152	73003.12
SP_l	0.575	0.058	0.297	269168.3
$Path_{2up}$	0.332	0.039	0.183	114295.3
AN_n	0.063	0.035	0.04	18843.35
AN_l	0.18	0.03	0.107	51561.98
ANP_l	0.435	0.05	0.226	337090.9
STR_n	0.425	0.142	0.25	119360
STG_v	0.503	0.071	0.278	209395.5
STG_t	0.208	0.116	0.128	61067.99
STD_v	0.541	0.097	0.304	145828.3
STD_t	0.407	0.146	0.24	114517.4
TF	0.121	0.287	0.072	33953.84
TF-IDF	0.189	0.335	0.115	40422.39
C-value	0.149	0.312	0.082	34288
NC-value	0.198	0.348	0.119	34571.75

The table 3.12 demonstrates the evaluation of different attributes using four attribute selection measures: Information Gain (Ig), Gain Ratio (Gr), Gini Index (Gi), and Chi-Square (Chi). Attributes like C_{adj} and ANP_l exhibit significantly high Chi-Square values, suggesting strong contributions to classification. Conversely, C_h holds zero values across all metrics, indicating irrelevance. Other attributes such as P_l, SP_l, and STD_v show

balanced importance across multiple metrics, reflecting their relevance. This comparative analysis helps in identifying the most influential features for the classification task.

Next, using the Backward Elimination scheme, we have acquired a new set of attributes BE_F which is also a subset of $Attr_{Total}$.

$$BE_F = \{\ R_e,\ C_h,\ P_l,\ C_{adj},\ S_l,\ SP_l,\ Path_{2up},\ AN_n,\ AN_l,\ ANP_l,\ STG_t,\ STD_v,\ STD_t,\ TF,\ TF\text{-}IDF\ \} \text{ where } BE_F \subset Attr_{Total}$$

In both the schemes, we received a list of attributes as an end product. Then, we have prepared two different data sets D_{FS} and D_{BE} with the relevant attributes.

3.5 Models

For the Task 1, we trained **MLP Classifier**, **KNN Classifier** on our datasets in the **RapidMiner tool**.

Similarly, for the Task 2, we employed MLP Classifier, KNN Classifier, Logistic Regression Classifier, Naive Bayes Classifier, Decision Tree, and Random Forest Classifier to train the models in the **RapidMiner tool**.

The **MLP Classifier** is a neural network-based classifier consisting of multiple layers of perceptrons that capture non-linear relationships. The **KNN Classifier** is a non-parametric model that classifies data points based on the majority vote of their k nearest neighbors. The **Logistic Regression Classifier** is a linear model that predicts binary outcomes by modeling the probability using the logistic function. The **Naive Bayes Classifier** is a probabilistic classifier that relies on Bayes' Theorem and operates under the assumption of strong independence among features. The **Decision Tree Classifier** is a tree-structured model that learns decision rules by recursively splitting the data based on feature values. Finally, the **Random Forest Classifier** is an ensemble of decision trees that improves predictive accuracy by aggregating the outputs of multiple trees.

3.6 Experiments & Results

3.6.1 Task 1: Experiments & Results

To prepare the training sets for both word level and phrase level, we consider a semi-supervised learning approach. Initially, we extracted all the features of each NNP present in mythological texts like **MBH**, **RAMA**, **SB** and **DEVI** at the word level and compared each NNP with a manually tagged list of Characters. The NNPs found in the list are

annotated as *Character*, while those not found are labeled as $Not\_a\_Character$. In this manner, we prepared a dataset, denoted as $WD_{training}$.

Next, we extracted all the features of each NNP present in the mythological texts like **CC**, **HVM** and **KRSNA** and prepared a dataset called WD_{test}. Subsequently, we developed a learning model trained on the $WD_{training}$ dataset using a **KNN-classifier** and tested the model on the WD_{test}. We then calculated the accuracy, precision, recall, and F-measure for WD_{test}.

Afterward, all the NNPs in the WD_{test} dataset were correctly annotated and incorporated into the $WD_{training}$ dataset by appending the newly annotated WD_{test} dataset. This process was repeated for all other chapters in our sample space. Ultimately, we obtained an updated dataset, $WD_{training}$, which is considered the training set containing word-level features for our system. The results are discussed in Table 3.13.

Table 3.13: Precision, Recall, F-measure at Word Level

Word Level ($WD_{Training}$)			
Mytho-Text	**P**	**R**	**F**
CC	0.43	0.44	0.43
HVM	0.62	0.64	0.6
KRSNA	0.63	0.63	0.63
P=Precision; R=Recall; F=F-measure			

Similarly, at the phrase level, we extracted all the features of each $NP \ll NNP$ present in the **MBH**, **RAMA**, **SB** and **DEVI** and prepared a dataset named $PD_{training}$. We trained a model using the **KNN Classifier**. Next, we extracted all the features of each $NP \ll NNP$ present in the **CC**, **HVM** and **KRSNA** and prepared a dataset called PD_{test}, which was applied to the trained model, following a process similar to that of the word level. Here, we calculate precision, recall, and F-measure for PD_{test}. This process was repeated for other chapters, and finally, we obtained an updated $PD_{training}$ dataset as the training set for the phrase level. The results are shown in Table 3.14.

Next, we prepared the datasets for both word and phrase levels, WT_{Test} and PT_{Test}, respectively.

To determine the important features in the $WD_{Training}$ and $PD_{Training}$ datasets, we calculated the relevance of the features by computing the **Chi-squared statistic** with respect to the class level feature using the **RapidMiner tool**[5]. The higher the weight

[5]https://rapidminer.com/products/studio/

Table 3.14: Precision, Recall, F-measure on Phrase Level

Phrase Level ($PD_{Training}$)			
Mytho-Text	**P**	**R**	**F**
CC	0.52	0.56	0.5
HVM	0.67	0.65	0.63
KRSNA	0.6	0.62	0.6
P=Precision; R=Recall; F=F-measure			

of a feature, the more relevant it is considered. The value of the Chi-squared statistic is given by

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{3.4}$$

where, χ^2 is the **chi-square statistic**, O_i is the observed frequency, and E_i is the expected frequency. Using this measure, we identified 45 important features at the word level(D_w) and 9 at the phrase level (D_p). The list of relevant features are as follows:

List of Relevant Features at Word Level (D_w) = {W1, W4, W5, W6, W7, W8, W9, W10, W11, W12, W13, W14, W15, W16, W20, W22, W23, W26, W27, W28, W29, W30, W31, W32, W35, W38, W39, W64, W68, W70, W71, W73, W74, W75, W76, W77, W78, W80, W82, W83, W84, W85, W95, W96, W97}

List of Relevant Features at Phrase Level (D_p) = {P1, P3, P4, P12, P20, P26, P27, P28, P36}

With the help of D_w and D_p, we prepared our new training sets as D_{wt} and D_{pt}, and new test sets, using these important features, as D_{wtest} and D_{ptest} with 7:3 ratio.

Now, we developed a training model using MLP and KNN classifiers with the newly prepared datasets, D_{wt} and D_{pt}. Subsequently, we tested these models using our newly created test sets, D_{wtest} and D_{ptest}. At the word level, the MLP classifier demonstrated better precision, recall, and F-measure than the KNN classifier. At the phrase level, the MLP classifier showed superior precision and F-measure compared to the KNN classifier. However, KNN achieved a higher recall than the MLP classifier at the phrase level. The precision, recall, and F-measure of the two classifiers are detailed in Table 3.15.

For the word level, MLP demonstrates a high precision of 91.84 and recall of 84.91, resulting in an F-measure of 88.24, indicating effective identification of relevant words. In

Table 3.15: Precision, Recall, F-measure on D_{wtest} and D_{ptest}

	Word Level			Phrase Level		
Classifiers	P	R	F	P	R	F
MLP	91.84	84.91	88.24	79.07	69.39	73.91
KNN	90.7	73.58	81.25	61.67	75.51	67.89
P=Precision; R=Recall; F=F-measure						

contrast, KNN achieves a precision of 90.70, a lower recall of 73.58, and an F-measure of 81.25, showing that while it maintains competitive precision, it struggles more with recall compared to MLP. At the phrase level, MLP's performance drops slightly, with a precision of 79.07, recall of 69.39, and an F-measure of 73.91, indicating some challenges in phrase-level classification. KNN, on the other hand, exhibits a significant drop in performance, achieving a precision of 61.67, a recall of 75.51, and an F-measure of 67.89, suggesting it is less effective at the phrase level compared to word level. Overall, MLP consistently outperforms KNN across both classification levels, particularly in word-level tasks.

Both classifiers performed well in classifying the test datasets, D_{wtest} and D_{ptest}. The **MLP classifier** exhibited better accuracy than the **KNN classifier** at both the word level and phrase level.

3.6.2 Task 2: Experiments & Results

Here, we transformed our datasets (D, D_{FS} and D_{BE}) into formats compatible with the classifiers in the **RapidMiner tool**. The datasets were then split into a 7:3 ratio for training and testing purposes. Afterward, these classifiers were applied to the datasets to assess their accuracy, along with other relevant statistical metrics.

A detailed observation of Precision, Recall, and F-Measure for each classifier applied to the datasets D, D_{FS} and D_{BE} is presented in Table 3.16 to 3.18.

Table 3.16 provides a detailed overview of the performance of various machine learning classifiers—namely MLP, KNN, Logistic Regression, Naive Bayes, Decision Tree, and Random Forest—on the D dataset, which comprises character instances from notable mythological texts, including the Mahabharata (MBH) and Ramayana (RAMA). Key performance metrics include Precision, Recall, and F-measure, which assess the classifiers' effectiveness in identifying characters accurately. The KNN classifier emerges as the top performer, achieving the highest Precision, Recall, and F-measure scores for both the MBH and RAMA datasets, with values reaching 0.87 for Precision, 0.90 for Recall, and an

Table 3.16: Precision, Recall and F-measure of classifiers on D dataset

Mytho Text	MLP			KNN			Logistic			NaiveBayes			DecisionTree			RandomForest		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
MBH	0.82	0.85	0.83	**0.87**	**0.90**	**0.89**	0.75	0.76	0.75	0.84	0.86	0.85	0.72	0.73	0.73	0.78	0.80	0.79
RAMA	0.78	0.79	0.78	**0.80**	**0.82**	**0.81**	0.72	0.74	0.73	0.73	0.75	0.74	0.69	0.70	0.69	0.71	0.72	0.71
SB	0.80	0.82	0.81	0.79	0.80	0.79	0.70	0.72	0.71	0.78	0.79	0.78	0.67	0.68	0.68	0.75	0.76	0.75
DEVI	0.77	0.78	0.77	0.75	0.76	0.75	0.68	0.69	0.68	0.71	0.73	0.72	0.66	0.67	0.66	0.69	0.70	0.69
CC	0.79	0.80	0.80	0.77	0.78	0.77	0.71	0.72	0.71	0.76	0.78	0.77	0.68	0.69	0.68	0.74	0.75	0.74
HVM	0.82	0.83	0.82	0.80	0.81	0.80	0.73	0.74	0.73	0.79	0.80	0.79	0.71	0.72	0.71	0.77	0.78	0.77
KRSNA	0.84	0.85	0.85	0.81	0.82	0.81	0.74	0.75	0.74	0.80	0.81	0.80	0.70	0.71	0.70	0.76	0.77	0.76

F-measure of 0.89 for MBH. These results indicate that KNN is particularly effective in accurately identifying character instances while also capturing a significant proportion of all relevant instances. In contrast, the Random Forest classifier records the lowest Recall and F-measure, suggesting it fails to capture a substantial number of relevant instances, which raises concerns about its applicability for tasks requiring high sensitivity and accuracy in complex narrative texts. This disparity underscores the importance of selecting the appropriate model for specific tasks; the KNN classifier's performance indicates its suitability for character identification in mythological narratives, while the limitations of Random Forest highlight the challenges associated with its more complex ensemble method.

Similarly, the Precision, Recall, and F-Measure for the datasets D_{FS} and D_{BE} are presented in Table 3.17 and 3.18. In these datasets, the KNN classifier consistently outperformed the other classifiers. Conversely, Naive Bayes exhibited the poorest results for both datasets. The confusion matrices for each classifier applied to the datasets D, D_{FS} and D_{BE} are also provided. Here, N and C denote *Not_a_Character* and *Character*, respectively.

Notably in table 3.17, the MBH and RAMA texts exhibit superior performance, highlighted by their robust F-measure values, indicating their effectiveness in the classification task. For the Mahabharata (MBH), KNN delivers the highest performance, with Precision: 0.90, Recall: 0.92, and F-measure: 0.91. In comparison, MLP achieves Precision: 0.85, Recall: 0.87, and F-measure: 0.86. Similarly, for the Ramayana (RAMA), KNN again leads with Precision: 0.85, Recall: 0.87, and F-measure: 0.86, whereas MLP has Precision: 0.81, Recall: 0.82, and F-measure: 0.81. In the Srimad Bhagavatam (SB), Naive Bayes slightly outperforms the others with Precision: 0.79, Recall: 0.80, and F-

Table 3.17: Precision, Recall and F-measure on D_{FS} dataset

Mytho Text	MLP			KNN			Logistic			NaiveBayes			DecisionTree			RandomForest		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
MBH	0.85	0.87	0.86	**0.90**	**0.92**	**0.91**	0.76	0.77	0.76	0.82	0.84	0.83	0.73	0.74	0.73	0.78	0.80	0.79
RAMA	0.81	0.82	0.81	**0.85**	**0.87**	**0.86**	0.74	0.75	0.74	0.75	0.77	0.76	0.70	0.71	0.70	0.72	0.73	0.72
SB	0.78	0.80	0.79	0.76	0.77	0.76	0.72	0.73	0.72	0.79	0.80	0.79	0.69	0.70	0.69	0.76	0.78	0.77
DEVI	0.76	0.77	0.76	0.74	0.75	0.74	0.69	0.70	0.69	0.70	0.72	0.71	0.67	0.68	0.67	0.70	0.71	0.70
CC	0.80	0.81	0.80	0.78	0.79	0.78	0.73	0.74	0.73	0.77	0.79	0.78	0.68	0.69	0.68	0.75	0.76	0.75
HVM	0.84	0.85	0.84	0.79	0.80	0.79	0.71	0.72	0.71	0.76	0.77	0.76	0.69	0.70	0.69	0.74	0.75	0.74
KRSNA	0.83	0.84	0.84	0.80	0.81	0.80	0.72	0.73	0.72	0.77	0.78	0.77	0.71	0.72	0.71	0.76	0.77	0.76

measure: 0.79, while KNN records Precision: 0.76, Recall: 0.77, and F-measure: 0.76. For the other texts, KNN consistently performs well across all datasets, with Devi Bhagavata (DEVI) scoring Precision: 0.74, Recall: 0.75, and F-measure: 0.74, and Caitanya Caritamrita (CC) with Precision: 0.78, Recall: 0.79, and F-measure: 0.78. IKNN shows the best overall performance, especially for MBH and RAMA, followed closely by MLP and Naive Bayes.

Table 3.18: Precision, Recall and F-measure on D_{BE} dataset

Mytho Text	MLP			KNN			Logistic			NaiveBayes			DecisionTree			RandomForest		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
MBH	0.80	0.81	0.80	**0.91**	**0.93**	**0.92**	0.75	0.76	0.75	0.78	0.80	0.79	0.71	0.72	0.71	0.77	0.79	0.78
RAMA	0.83	0.84	0.83	**0.88**	**0.90**	**0.89**	0.76	0.77	0.76	0.76	0.78	0.77	0.72	0.73	0.72	0.70	0.71	0.71
SB	0.79	0.80	0.79	**0.85**	**0.87**	**0.86**	0.71	0.72	0.71	0.77	0.78	0.77	0.68	0.69	0.68	0.74	0.75	0.74
DEVI	0.75	0.76	0.75	0.72	0.73	0.72	0.67	0.68	0.67	0.71	0.73	0.72	0.65	0.66	0.65	0.68	0.69	0.68
CC	0.78	0.79	0.78	0.76	0.77	0.76	0.72	0.73	0.72	0.74	0.75	0.74	0.67	0.68	0.67	0.72	0.73	0.72
HVM	0.82	0.83	0.82	0.78	0.79	0.78	0.70	0.71	0.70	0.75	0.76	0.75	0.69	0.70	0.69	0.73	0.74	0.73
KRSNA	0.80	0.81	0.81	0.77	0.78	0.77	0.72	0.73	0.72	0.76	0.77	0.76	0.70	0.71	0.70	0.75	0.76	0.75

The table 3.18 displays the precision, recall, and F-measure for various classifiers evaluated on the dataset D_{FS}. For the Mahabharata (MBH) text, KNN performs best with Precision: 0.91, Recall: 0.93, and F-measure: 0.92. MLP follows with Precision: 0.80, Recall: 0.81, and F-measure: 0.80. Logistic Regression, Naive Bayes, and the other models show lower performance metrics in comparison. In the Ramayana (RAMA) text, KNN

again leads with Precision: 0.88, Recall: 0.90, and F-measure: 0.89, while MLP shows decent performance with Precision: 0.83, Recall: 0.84, and F-measure: 0.83. For the Srimad Bhagavatam (SB), KNN achieves the best performance, with Precision: 0.85, Recall: 0.87, and F-measure: 0.86. MLP has lower values, while Naive Bayes shows Precision: 0.77, Recall: 0.78, and F-measure: 0.77. The results for the other texts (DEVI, CC, HVM, and KRSNA) show that KNN generally outperforms other models, while MLP and Naive Bayes perform moderately well. Across all texts, Decision Tree and Random Forest consistently show lower performance compared to KNN and MLP. Finally, KNN exhibits the highest Precision, Recall, and F-measure across most of the texts, making it the best-performing model on the D_{BE} dataset.

3.7 Error Analysis

An in-depth error analysis of Task 1 and Task 2 is conducted here, highlighting key areas of misclassification and offering insights to enhance performance.

3.7.1 Task 1: Error analysis

It can be observed that the **MLP classifier** has the lowest classification error and the highest kappa value at both the word level and phrase level. The classification error rate and kappa measure are presented in Table 3.19.

Table 3.19: Error and Kappa of D_{wtest} and D_{ptest}

Classifiers	Word Level		Phrase Level	
	CE	K	CE	K
MLP	12.00%	0.76	24.00%	0.519
KNN	18.00%	0.643	35.00%	0.303
CE=Classification Error rate; K=Kappa measure				

The average absolute deviation of the predictions from the actual values, known as Absolute Error, of the **MLP classifier** at the word level is lower than that of the **KNN classifier**. Similarly, the average absolute deviation of the predictions from the actual values, divided by the actual values (referred to as *Relative Error*), as well as the *Root Mean Squared Error* of the **MLP classifier** at the word level, are significantly lower than those of the **KNN classifier**. The details are provided in Table 3.20.

Table 3.20: Error Analysis of D_{wtest} and D_{ptest}

	Word Level		Phrase Level	
Measures	MLP	KNN	MLP	KNN
AE	0.16+/-0.28	0.31+/-0.19	0.36+/-0.24	0.41+/-0.25
RE	16.83+/-28.09	31.83+/-18.97	36.93+/-23.97	41.89+/-25.84
RMSE	0.32+/-0.00	0.37+/-0.00	0.44+/-0.00	0.492+/-0.00
AE=Absolute Error; RE=Relative Error(%)				
RMSE= Root mean Squared Error				

At the phrase level, the **MLP classifier** also exhibits lower *Absolute Error*, *Relative Error*, and *Root Mean Squared Error* compared to the **KNN classifier** in table 3.20.

3.7.2 Task 2: Error analysis

Table 3.21 shows that the KNN classifier achieves the lowest average error rate across all datasets, indicating that KNN outperforms the other five classifiers. Conversely, we observe that the Random Forest classifier exhibits the worst performance on dataset D, while Naive Bayes has the highest average error rate on datasets D_{FS} and D_{BE}.

Table 3.21: Error Rate of D, D_{FS} and D_{BE} datasets

	Classification error(%)		
	D dataset	D_{FS} dataset	D_{BE} dataset
MLP	10.29	12.63	12.86
KNN	2.67	4.34	7.81
Logistic Regression	3.83	15.47	15.46
NaiveBayes'	5.35	25.58	23.89
DecisionTree	25.94	9.3	9.31
RandomForest	27.08	10.7	11.8

3.8 Summary

In this research, we have explored two critical aspects of character identification from the Indian mythological texts: Character Adjective Identification and Character Identification. Through these contributions, we have laid the groundwork for a more structured and

computationally effective analysis of characters within complex narrative texts.

The first major contribution of this work is the identification of character adjectives from un-annotated mythological texts. By employing specific phrase-level rules, we were able to successfully extract both characters and the adjectives that describe them. This nuanced approach helped capture the rich descriptive nature of mythological narratives, where characters are often introduced and elaborated upon through a combination of proper nouns and descriptive phrases. Using various machine learning classifiers such as KNN, Logistic Regression, Naive Bayes, and MLP models, we demonstrated that the KNN classifier consistently outperformed other models in accurately classifying these entities. The incorporation of feature subset selection further enhanced the effectiveness of our classification models, proving the importance of the right set of features in character adjective identification.

The second key contribution focuses on character identification. By manually annotating a subset of the mythological texts, we constructed a semi-supervised system capable of identifying characters at both word and phrase levels. The use of the *Chi-Square Statistic* for feature selection, combined with association analysis via the *FP-Growth algorithm*, revealed key feature dependencies and co-occurrences that bolstered the accuracy of character extraction. Our models, trained using both *KNN* and *MLP classifiers*, were rigorously tested, showing robust performance across metrics such as accuracy, precision, recall, and F-measure. This method lays the foundation for more generalized character identification techniques applicable to larger corpora of mythological and narrative texts.

Chapter 4

Pouranic Topic Classification

Classifying Mythological Topics

4.1 Introduction

Indian Hindu mythology such as ***the Mahabharata***, ***the Srimad-Bhagavatam***, etc., characterized by a rich tapestry of stories, characters, and themes, has captivated scholars and enthusiasts for generations. Researchers have observed that a significant portion of the mythological texts, containing around 400,000 verses from the 18 *Maha Puranas* and 18 *Upa Puranas*, have been translated into English(Dominic, 2021).

Text classification, a foundational task in natural language processing, entails assigning text to predefined categories or classes. It efficiently organizes and analyzes vast volumes of textual data. One prominent application of text classification is topic labeling, wherein text is categorized into distinct topics or themes (Lang, 1995; Wang and Manning, 2012). Researchers have observed that topic classification poses challenges due to the potentially large and overlapping number of classes, intensifying the complexity of the task (Gentzkow et al., 2017; Liu, 2015). Given the broad and intricate nature of Indian mythology, topic classification aids in simplifying the comprehension and navigation of textual content, encompassing diverse narratives, characters, and themes (Paul and Das, 2017b). The absence of ***structured and annotated datasets*** in Indian mythological texts poses challenges for automated topic classification systems. Addressing this, the research aims to introduce automated models for topic classification and assess their performance against sentence clustering-based approaches. This endeavor responds to challenges like ***labeling datasets***, ***a lack of automated models***, and ***the need for comparative performance studies***. In our research, a collection of words or phrases is presented as a topic that carries the important findings and observations from large clusters of texts.

To support the above studies, we prepared datasets to build the topic classification

model for the mythological texts due to their unavailability. Initially, we collected seven Hindu mythological texts,[1][2] which contain 200k ***Pouranic*** (connected to Purana texts) verses, to prepare the dataset, represented as the `PouranicTopic` dataset in this work. Next, it has been observed that every topic in mythological texts has several sentences that are semantically related to the topic name. In this regard, we developed two distinct datasets: one comprising sentences clustered by `semantic textual similarity` for each topic, and the other comprising sentences clustered by `log likelihood value` for each topic.

Log-likelihood ratio-based sentence clustering is a technique that is used to group similar sentences together (Manning et al., 2008). Thereafter, we have pre-processed and structured the dataset into three different hierarchical labels, such as cantos, topics, and subject matter. Apart from the development of topic classification models, we investigated and compared the performances of these datasets and reported it accordingly. A canto refers to a section of the mythological text where the topic represents the title of the chapter, which belongs under a canto. On the other hand, subject matter indicates individual sentences about a topic.

In the following example, we describe the hierarchical organization of our `PouranicTopic` dataset. **Mythological text:** *"the Harivamsha Puranas"*, **Canto:** *"Harivamsa Parva"*, **Topic:** *"The Origin of Men: the Birth of Daksha"*, and **Subject matter:** *"Vaishampayana said:—When the work of his creation of progeny was complete the Patriarch Vashishtha obtained Shatarupa, not born of a woman, as his spouse."*.

Transformers are pivotal for precise topic classification in Indian mythology, adeptly handling vast text data and intricate semantic relationships. With their attention mechanisms and deep learning architectures, transformers excel at capturing nuanced patterns within the rich content of Indian mythology. This study employs BERT (Bidirectional Encoder Representations from Transformers), RoBERTa (Robustly Optimized BERT), and DistillBERT (distilled version of BERT), along with an ensemble approach, to enhance topic classification performance. The proposed model framework is described in Figure 4.1. In this research, we propose a transformer-based model framework for automated topic classification of Indian mythological texts, which provides a structured and annotated dataset for this domain. The rest of the chapter is structured as follows: the Preparation of datasets is in the Preparation of Datasets section, the System Framework has been described in the System Framework section. Thereafter, Result Analysis, Error

[1] https://www.wisdomlib.org/hinduism
[2] https://vedabase.io/en/library

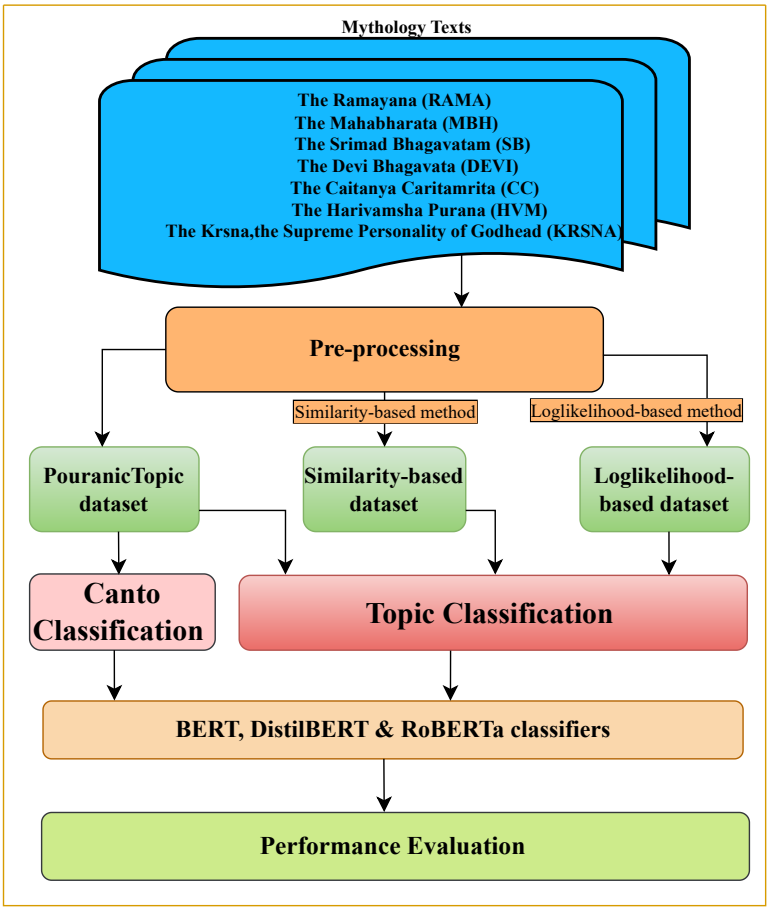


Figure 4.1: Proposed Model Framework

Analysis, and Discussion & Observations describe the experimental results, errors, and observations of the models. Finally, Summary contains the closing remarks and future scopes of this research work.

4.2 Preparation of Datasets

Primarily, we have observed that a huge amount of unstructured corpora are available on the web related to Hindu religions. In order to build a topic classification model for Hindu mythology, a structured dataset is essential. Hence, in this research, we are motivated to prepare a structured corpus that assists in building an automated topic classification system.

4.2.1 PouranicTopic Datasets

Firstly, we have crawled various Hindu mythological texts in English from different web sources[3][4] [5].

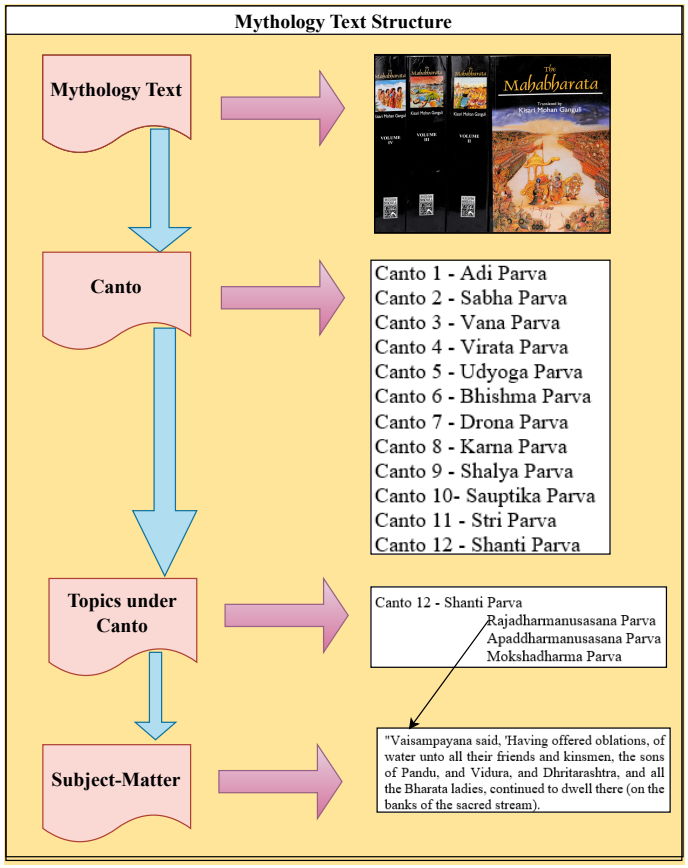


Figure 4.2: Hierarchical structure of the mythological texts

Afterward, we combined all the crawled data and prepared our experimental datasets, which include ***the Ramayana***, ***the Mahabharata***, ***the Srimad-Bhagavatam***, ***the Devi Bhagavata***, ***the Caitanya Caritamrita***, ***the Harivamsha Purana***, and ***Krsna, the Supreme Personality of Godhead***. The dataset contains a number of cantos; each canto has a number of topics, and each topic has a number of sentences referred to as subject matters. The following example illustrates the detailed relationship between the mentioned three levels of a mythological text which is shown in Figure 4.2.

In order to prepare our `PouranicTopic` dataset, we have employed the spaCy[6] tool along with the nltk[7] package, which assists in mapping each subject matter to its corres-

[3]https://www.wisdomlib.org/hinduism
[4]https://vedabase.io/en/library
[5]https://hindi.webdunia.com/religion/religion/hindu/ramcharitmanas
[6]https://spacy.io/
[7]https://www.nltk.org/

Table 4.1: Basic details of all the datasets

Mytho-Text	PouranicTopic dataset				Similarity based dataset				Loglikelihood based dataset			
	C	T	SM	Qp+	C	T	SM	Qs+	C	T	SM	Ql+
RAMA	7	648	20.5k	10.81	7	530	2.2k	98.67	7	638	18.04k	12.35
MBH	18	74	117.3k	7.61	18	66	10.9k	68.55	18	68	47.4k	15.23
SB	12	335	18.2k	23.22	12	349	9.7k	71.68	12	331	17.3k	46.49
DEVI	12	318	34.2k	28.48	12	314	10.05k	84.94	12	315	18.9k	38.20
CC	3	62	20.1k	14.16	3	62	2.7k	99.67	3	62	17.4k	15.42
HVM	3	235	14.7k	6.29	3	177	0.97k	94.68	3	214	11.2k	8.26
KRSNA	1	90	14.6k	9.87	1	90	1.8k	80.26	1	90	7.3k	19.74
C: Total Cantos; T: Total Topics; SM: Total Subject Matters(number of sentences)												
Qp+ : Quality Index of PouranicTopic Dataset; Qs+ : Quality Index of Similarity-based Dataset; Ql+ : Quality Index of Loglikelihood-based Dataset;												

ponding topic. The topic refers to the name of the chapter, which belongs to a specific canto. In this section, we conduct a thorough quantitative analysis of the final datasets, which consist of seven *Pouranic texts*. From this stage forward, in this paper, ***the Ramayana*** will be referred to as **RAMA**, ***the Mahabharata*** as **MBH**, ***the Devi Bhagavata*** as **DEVI**, ***the Harivamsha Purana*** as **HVM**, ***the Srimad-Bhagavatam*** as **SB**, ***the Caitanya Caritamrita*** as **CC**, and lastly, the ***Krsna, the Supreme Personality of Godhead***, as **KRSNA**. In reference to the total number of cantos, topics, and the subject matters or sentences in each text, as observed in Table 4.1.

The data presented in Table 4.1 highlights that ***the Mahabharata*** comprises the highest number of sentences (117,370) and cantos (18), whereas ***the Ramayana*** encompasses the largest number of topics (648). Conversely, ***the Caitanya Caritamrita*** exhibits the fewest topics in the dataset, while ***Krsna, the Supreme Personality of Godhead***, features only one canto with a total of 14,664 sentences.

Table 4.2 presents the statistical analysis of our `PouranicTopic` dataset. Notably, ***the Mahabharata*** exhibits the highest word and stopwords counts, featuring the longest sentence in our dataset. Conversely, ***the Ramayana***, ***the Devi Bhagavata***, and ***the Srimad-Bhagavatam*** demonstrate fewer words, stopwords, and punctuation compared to ***the Mahabharata***. Furthermore, our observations indicate that sentences in ***the Ramayana*** tend to be longer on average, while ***Krsna, the Supreme Personality of Godhead***, contains fewer words and stopwords, with an average sentence length of 21.56 words.

Table 4.2: Statistical Analysis of the PouranicTopic dataset

Mytho Text	Words	Unique words	Stop words	Avg Length (Als)	Sent ≥ Als	Canto Name	Topic with max sentences
RAMA	518k	10.7k	124	25.02	7.7k	Yuddha Kanda (Canto of War)	Kumbhakarna's Exploits- He is slain by Rama
MBH	1170k	35.2k	128	24.52	15.3k	Anusasana Parva (Canto of Instructions)	Anusasanika Parva
SB	390k	17.4k	129	24.22	7.9k	The Summum Bonum (Canto of the ultimate good)	Talks Between Krsna and Rukmini
DEVI	608k	24.2k	132	17.75	14.6k	Canto-9	On the description of Prakrti
CC	326k	19.6k	127	16.21	8.6k	Madhya Lila (Middle Pastimes)	Lord Sri Caitanya Mahaprabhu's Travels to the Holy Places
HVM	278k	16.1k	125	19.10	6.5k	Vishnu Parva (Canto of Vishnu)	Citralekha Unites Aniruddha with Usha
KRSNA	316k	13.8k	125	21.56	6.2k	Krsna	Prayers by the Personified Vedas

Table 4.2 displays various cantos and their corresponding topics, along with the number of sentences associated with each topic. Notably, in ***the Ramayana***, the canto *Yuddha kanda* contains 131 different topics, with the topic *Kumbhakarna's Exploits-He is slain by Rama* having the highest count of 187 sentences. In ***the Mahabharata***, the *Anusasana Canto* features the topic *Anusasanika Parva* with 13,800 sentences, while ***the Devi Bhagavata***'s *Canto 9* includes the topic *On the description of Prakrti* with 362 sentences. Similarly, in ***the Harivamsha Purana***, the topic *Citralekha Unites Aniruddha with Usha-Fight Aniruddha's with Vana's Soldiers* under *Vishnu-Parva* comprises 263 sentences. In ***the Srimad-Bhagavatam***, *the Summum Bonum* canto hosts 333 sentences within the topic *Talks Between Krsna and Rukmini*. Lastly, in ***the Caitanya Caritamrita***, the topic *Lord Sri Caitanya Mahaprabhu's Travels to the Holy Places* under the *Adi Lila* canto has the highest sentence count of 626, while ***Krsna, the Supreme Personality of Godhead***, includes the topic *Prayers by the Personified Vedas* with 1,363 sentences.

We noted that each topic in mythological texts is elaborated upon by a set of sentences, which significantly contribute to understanding the topic thoroughly. Considering this, we constructed two distinct datasets, one based on *semantic similarity* and the other on *log likelihood*, to advance our research efforts.

4.2.2 `Similarity`-based dataset

Semantic textual similarity is the task of evaluating how similar two texts are in terms of meaning. Sentence similarity models convert input texts into vectors or embeddings that capture semantic information and calculate how close (similar) they are between them.

For our study, we performed the semantic similarity between each sentence and the name of topic using Sentence-BERT (Reimers and Gurevych, 2019).Once we have the embeddings for all sentences and the topics, the similarity between each sentence embedding and the topic embedding is calculated using a cosine similarity metric . We finally selected only those sentences that had a high similarity score ($\geq$ 0.65) and prepared `Similarity based` dataset. The semantic similarity score, **S_score**, of a sample sentence from *Harivamsa-Parva* of **HVM** is:

Sentence-1: ***"Hear, I shall describe the fourth Manvantara."*** = **0.7376**

Here, the Sentence-1 belongs to the topic named ***An Account of Manvantaras***.

Table 4.1 describes this dataset in detail. ***The Ramayana*** has the most number of cantos (7) and topics (530), but ***the Mahabharata*** contains the highest number of sentences (10984). ***The Srimad Bhagavatam*** has 349 topics, second highest after ***the Ramayana***. ***The Devi Bhagavata*** contains 314 topics and 10053 sentences. ***The Harivamsha Purana*** has 177 topics and 977 sentences. ***The Caitanya Charitamrita*** has 62 topics and 2700 sentences. ***Krsna, the Supreme Personality of Godhead*** text has 90 topics and 1803 sentences, despite having only one canto. This dataset contains semantically clustered sentences for each topic, selected using sentence similarity analysis.

4.2.3 `Log-likelihood`-based dataset

Log-likelihood ratio based sentence clustering is based on the idea that sentences that are semantically similar will have a higher probability of being generated by the corpus (Manning et al., 2008). For each word in the sentence, the log-likelihood ratio is computed using the observed and expected counts of the word. Once the log-likelihood ratios are calculated for each word in the sentence, they are aggregated to obtain a single score representing the overall likelihood of the sentence. This process helps identify sentences that contain words that are significantly more or less frequent in the given context compared to the entire corpus. The formula for the log-likelihood ratio (LLR) is presented in equation 4.1.

$$LLR = -2\left(OCT \log\left(\frac{OCT}{ECT}\right) - (OCT + CCT)\log\left(\frac{OCT + CCT}{SS + CS}\right)\right) \tag{4.1}$$

where: **OCT** is the frequency of the word in the sentence, **ECT** is the expected frequency of the word based on its frequency in the corpus, **CCT** is the frequency of the word in the entire corpus, **SS** represents the number of words in the sentence, while **CS** denotes the total number of words in the corpus. The LLR score of a sample sentence from *Harivamsa-Parva* of **HVM** is as follows:

Sentence-1: ***"He is always beyond the path of righteousness."*** = **-839.83**

In this regard, we performed the similarity between each sentence and the name of the topic using this technique and selected only those sentences with a higher score($\geq$ *median_value*) and prepared the `Log-likelihood-based` dataset. The statistical details of this dataset are described in Table 4.1. ***The Ramayana*** has the most cantos (7), but ***the Mahabharata*** has the highest number of topics (68) and sentences (47400). ***The Srimad Bhagavatam*** contains 331 topics and 17355 sentences. ***The Devi Bhagavata*** has 315 topics and 18926 sentences. ***The Harivamsha Purana*** contains 214 topics and 11284 sentences. ***The Caitanya Charitamrita*** has 62 topics and 17469 sentences. ***Krsna, the Supreme Personality of Godhead*** has 90 topics but only 7329 sentences, despite having one canto. This dataset contains sentences clustered using log-likelihood analysis to select semantically relevant sentences for each topic. Overall, Table 4.1 outlines the structure of the log-likelihood-based sentence clustered corpus of mythological texts.

4.2.4 Quality Index of datasets

The Q_{p+}, Q_{s+} and Q_{l+} in Table 4.1 refer to the quality index of each dataset. These are used to calculate the percentage of common sentences in the `PouranicTopic`, `Similarity-`based, and `Loglikelihood-` based datasets, respectively, when compared to a combined set of common sentences from all three datasets. Q_{p+} calculates the proportion of sentences in the `PouranicTopic` dataset that are also found in the combined common set of sentences from all datasets. This value represents the percentage of sentences in `PouranicTopic` dataset that are shared across all three datasets.

Q_{s+} determines the percentage of sentences in the `Similarity-` based dataset that are common to the combined set of sentences from all the datasets. This indicates how much of the `Similarity-` based dataset overlaps with the common set of sentences.

Q_{l+} measures the proportion of sentences in the `Loglikelihood-` based dataset that appear in the combined common set of sentences from all the datasets. This percentage

shows the extent of overlap between the `Loglikelihood-` based dataset and the common set of sentences. The generic equation for the above three indices can be represented as :

$$Q_N = \left(\frac{\text{total common sentences from all the datasets}}{\text{total sentences in } \texttt{PT/SM/LL} \text{ dataset}} \right) \times 100 \tag{4.2}$$

where, N can be either p+ or s+ or l+ and PT= PouranicTopic, SM= Similarity-based and LL= Loglikelihood-based dataset.

These indices collectively aid in evaluating the extent of commonality or overlap among the three datasets, offering insights into the similarity or diversity of the datasets with regard to their shared content.

Table 4.3: Inter-Annotator Agreement on all the datasets based on Cohen's Kappa coefficient

PouranicTopic dataset		Annotator2		
		Relevant	Not-Relevant	PT_{Kappa}
Annotator1	Relevant	120070	8190	0.85
	Not-Relevant	10340	117921	

Similarity based dataset		Annotator2		
		Relevant	Not-Relevant	SM_{Kappa}
Annotator1	Relevant	20034	1034	0.89
	Not-Relevant	1250	19750	

Loglikelihood based dataset		Annotator2		
		Relevant	Not-Relevant	LL_{Kappa}
Annotator1	Relevant	69300	5098	0.76
	Not-Relevant	12099	62300	

4.2.5 Inter-Annotator Agreements

It is a crucial metric when assessing the reliability and consistency of annotations performed by multiple annotators on the same dataset. We designated two separate annotators for our datasets, namely `PouranicTopic`, `Similarity`, and `Loglikelihood`-based datasets. To assess the agreement between these annotators on our datasets, we employed *Cohen's Kappa coefficient* (McHugh, 2012) as a measure. Annotations were conducted following specific guidelines to ensure uniformity and clarity in the assessment process. The guidelines are as follows:

- The objective is to evaluate the correlation between sentences and topics within the Cantos of the `PouranicTopic`, `Similarity`, and `Loglikelihood`-based datasets.

- Employ a binary scale to signify the relevance of each sentence to its corresponding topic. Award a score of 1 if the sentence aligns with the topic. Conversely, assign a score of 0 if the sentence lacks relevance to the topic.
- Concentrate on gauging the semantic alignment of sentences with topics, taking into account the context and thematic coherence within the Cantos.
- Exercise discretion in assessing whether each sentence contributes substantively to the topic under consideration.

Within the `PouranicTopic` dataset, there are a total of 256,521 sentences spread across 1,762 topics. The Similarity-based dataset comprises 42,068 sentences allocated across 1,588 topics. On the other hand, in the Loglikelihood-based datasets, we have 148,797 sentences in 1,718 topics. We are interested in how similar both annotators are in finding the relevancy of the sentences under each topic in both datasets. The `PouranicTopic` method shows the highest level of agreement between annotators, followed by the `Loglikelihood`-based method, and then the `Similarity`-based method.

The `PouranicTopic` method demonstrates consistent agreement on both relevant and non-relevant instances, indicating the effectiveness of this method for annotation. The `Loglikelihood`-based method shows relatively high agreement on relevant instances but lower agreement on non-relevant instances compared to the `PouranicTopic` method.

The `Similarity` based method exhibits the lowest level of agreement among the three methods, suggesting potential challenges in annotating instances based on similarity. Here in Table 4.3 shows the IAA using *Cohen's Kappa coefficient*. Our analysis revealed a substantial level of agreement between annotators for all the `PouranicTopic` (PT_{kappa} = 0.85), `Similarity` based (SM_{kappa} = 0.89) and `Loglikelihood` based (LL_{kappa} = 0.76) datasets, indicating consistency in their assessments. These findings underscore the reliability of the annotations within our datasets and validate their utility for further analysis and interpretation.

Splitting of datasets - Table 4.4 depicts the training, testing, and validation sets, which are divided in the ratio of 80-10-10. We observe that in the `PouraniTopic` dataset, ***the Mahabharata*** has the most number of sentences with a training dataset count of 93,877, test dataset count of 11,744, and validation dataset count of 11,750, whereas ***Krsna, the Supreme Personality of Godhead*** has the least data, with a training dataset count of 11,729, test dataset count of 1,467, and validation dataset count of 1,467.

In the `Similarity-based` dataset in Table 4.4, ***the Mahabharata*** has the highest number of sentences, with a training dataset count of 8,787, test dataset count of 1,098,

Table 4.4: Train-Test-Val ratio of all the datasets

Mytho-Text	PouranicTopic dataset			Similarity based dataset			Loglikelihood based dataset		
	#Train	#Test	#Val	#Train	#Test	#Val	#Train	#Test	#Val
RAMA	16.4k	2.06k	2.06k	1.8k	0.22k	0.22k	14.4k	1.8k	1.8k
MBH	93.8k	11.7k	11.7k	8.7k	1.09k	1.09k	37.9k	4.7k	4.7k
SB	27.8k	3.4k	3.4k	7.7k	0.97k	0.97k	13.8k	1.7k	1.7k
DEVI	27.3k	3.4k	3.4k	8.04k	1.0k	1.0k	15.1k	1.8k	1.8k
CC	16.1k	2.01k	2.01k	2.1k	0.27k	0.27k	13.9k	1.7k	1.7k
HVM	11.7k	1.4k	1.4k	0.7k	0.97k	0.97k	9.02k	1.1k	1.1k
KRSNA	11.7k	1.4k	1.4k	1.4k	0.18k	0.18k	5.8k	0.73k	0.73k
Train-Test-Val Ratio 80-10-10									

and validation dataset count of 1,099. On the other hand, ***the Harivamsha Purana*** has the least data, with a training dataset count of 781, test dataset count of 97, and validation dataset count of 97.

In the `Loglikelihood-based` dataset in Table 4.4, ***the Mahabharata*** has the highest number of sentences, with a training dataset count of 37,920, test dataset count of 4,740, and validation dataset count of 4,740. Conversely, ***Krsna, the Supreme Personality of Godhead*** has the least data, with a training dataset count of 5,863, test dataset count of 732, and validation dataset count of 732.

4.3 System Framework

Topic classification is a supervised technique where a set of documents is trained on a list of predefined topics so that the trained model can automatically predict the topic of an unseen document. The systems that we developed for the classification tasks are discussed here:

- Our initial system involves assigning sentences from a mythological text to their respective cantos, known as Canto classification.
- Our second system is to classify sentences into topics within the Cantos. This process, known as Topic classification, involves developing various models using `PouranicTopic`, `Similarity-based` and `Loglikelihood-based` datasets.
- Lastly, we conducted comparisons and analyses of the model performances using the aforementioned datasets.

We trained our models on different state-of-the-art transformer models, applying callbacks based on training accuracy. Using the Adam optimizer, we compiled all models and employed categorical cross-entropy as the primary loss function. Evaluation of the results is based on weighted average *precision (P)*, *recall(R)*, and *f1-score (F1)* metrics on the test set.

Before going into the task, we will discuss different types of BERT-based deep learning models and their parameters that are used in this research work.

4.3.1 Models

In this research, we utilized the BERT, DistilBERT, and RoBERTa models. ***BERT (Bidirectional Encoder Representations of Transformers)*** in Figure 4.3, proposed by (Devlin et al., 2019), is intended for pre-training deep bidirectional representations of unlabeled text data in each layer, taking into account both left and right context. In the Figure 4.3 The input sentence is broken down into tokens Tok1, Tok2, ... , TokN. At the beginning of the training tasks for each input sentence [CLS] is used. The embeddings for each token in the input sentence are E_1, E_2, ... , E_n. The embedding for the [CLS] token is represented by $E_{[CLS]}$. C is the classification token ([CLS]). In BERT, every input sequence starts with this special token. When BERT is trained on a classification task, the final hidden state corresponding to this token is used by the classification layer to determine the class of the input sequence. T_1, T_2, ... , T_n are the tokens that make up the input sequence. Each token Ti represents a word or subword in the sentence. BERT processes these tokens bidirectionally, meaning it considers the context from both the left and the right of each token to understand its meaning within the sentence. Finally, Class Label, the output of the model, is the predicted class label for the input sentence.

As proposed by (Sanh et al., 2019), ***DistilBERT (Distilled Bidirectional Encoder Representations of Transformers)*** is a smaller and speedier variant of BERT that was developed following extensive training using a distilled BERT basis. Having 40% fewer parameters than the original BERT, it performs about 60% quicker while still retaining about 95% of the performance on the General Language Understanding Evaluation (GLUE) (Wang et al., 2018) language comprehension benchmark.

RoBERTa(Robustly Optimized BERT approach): is a variation of the BERT model. (Liu et al., 2019), assessed several design choices of BERT models thoroughly before pretraining. They found that training the model longer, with larger batches over more data, eliminating the next sentence prediction objective, training on longer sequences, and

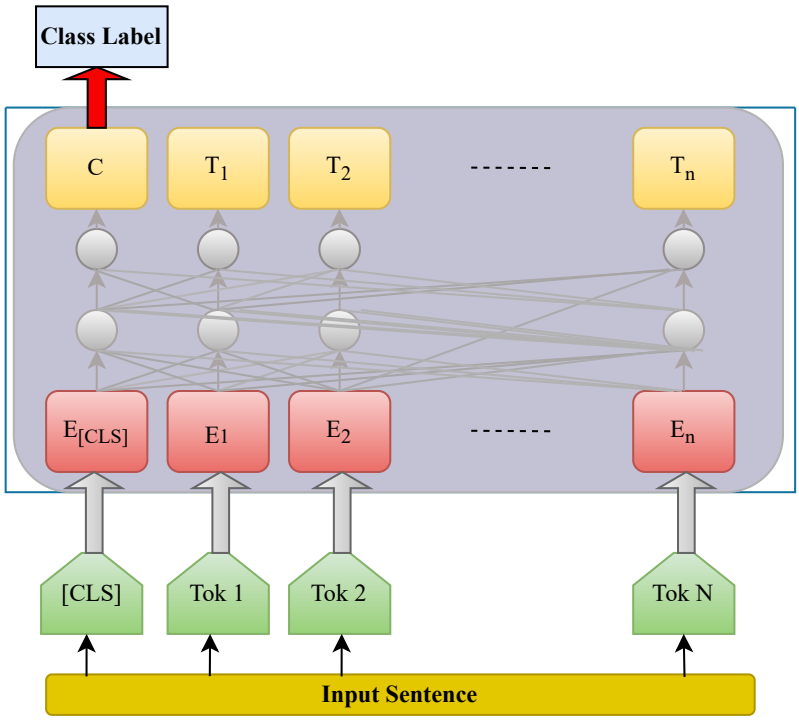


Figure 4.3: BERT architecture

dynamically varying the masking pattern applied to the training data can all significantly improve performance. RoBERTa achieves state-of-the-art performance on GLUE, ReAding Comprehension Dataset From Examinations (RACE) (Lai et al., 2017), and Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016).

4.3.2 Canto classification task

Our dataset comprises a total of 56 cantos. Initially, for canto classification, we processed all cantos of the texts simultaneously. We first trained an SVM model, followed by BERT, DistilBERT, and RoBERTa models, on the `PouranicTopic` datasets depicted in Figure 4.4. Due to the unsatisfactory performance of the SVM model, we discontinued its use for subsequent classification tasks.

4.3.3 Topic Classification task

To develop a topic classification model, we used a canto from a mythological text and applied BERT, DistilBERT, and RoBERTa to evaluate the weighted average precision, recall, and f1-score. Figure 4.5 displays these metrics for the ***Aswamedha Parva*** from ***the Mahabharata*** of `PouranicTopic` dataset. This canto includes two topics: ***Aswamedhika Parva*** and ***Anugita Parva***. The highest f1-score and weighted average f1-score were achieved by both BERT and DistilBERT, with scores of 97% and 94%, respectively,

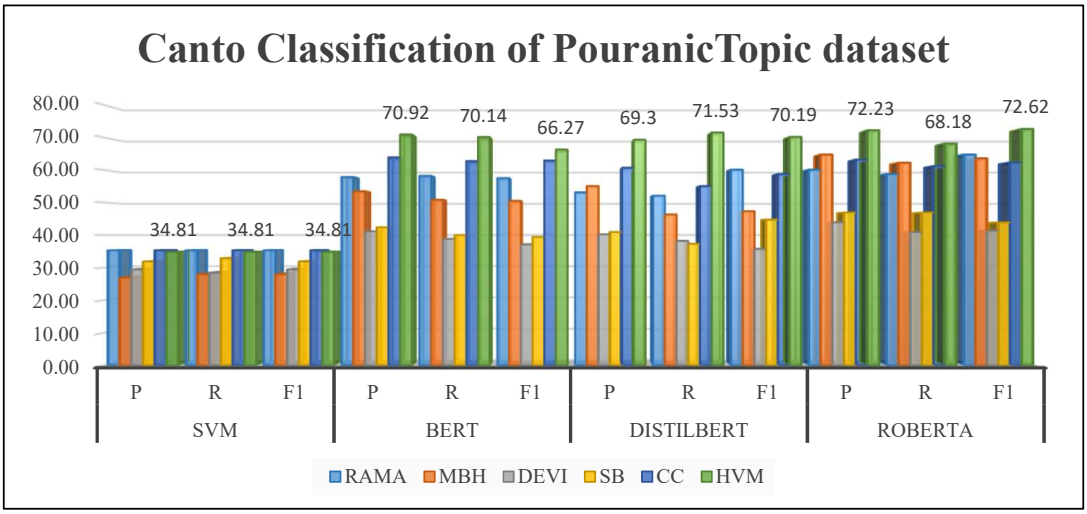


Figure 4.4: Precision(P), Recall(R), F1-Score(F1) at Canto classification

for the ***Anugita Parva*** topic. For every canto of each mythological text, we received the average weighted precision, recall, and f1-score from all the models and reported the mean of these values as final data.

In addition to the `PouranicTopic` dataset, we also developed three topic classification models for both the `Similarity based` and `Loglikelihood based` datasets. The results are presented in Table 4.5 and analyzed in the Result Analysis section .

4.4 Result Analysis

This section provides a comprehensive analysis of canto and topic classification, focusing on weighted average precision, recall, and f1-scores. Model performances across all mythological texts in Canto classification are presented in Figure 4.4, and the performances in Topic classification are described in Table 4.5.

4.4.1 Results of Canto classification

BERT emerges as the top performer in ***the Harivamsha Purana*** for canto classification in Figure 4.4, excelling across all metrics. It is followed by ***the Caitanya Caritamrita***, ***the Ramayana***, and ***the Mahabharata***. However, we excluded ***Krsna, the Supreme Personality of Godhead*** from this comparison across all models due to having only one canto.

In canto classification, as depicted in Figure 4.4, DistillBERT emerges as the top performer in ***the Harivamsha Purana***, followed by ***the Caitanya Caritamrita*** and

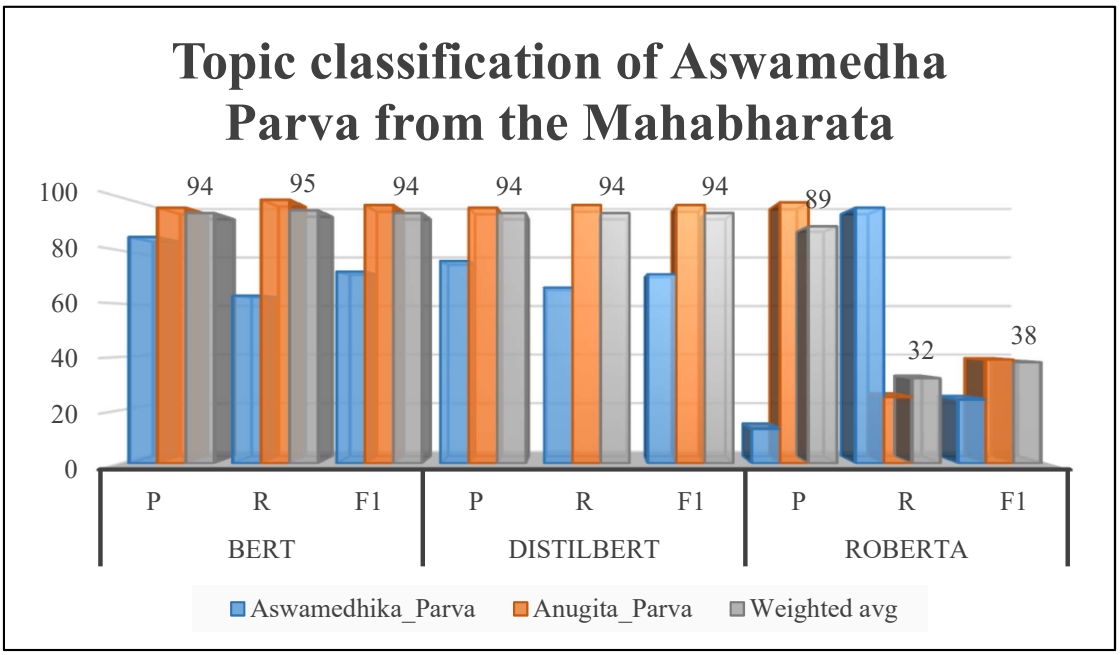


Figure 4.5: Results of Topic classification of Aswamedha Parva from the Mahabharata

the Mahabharata. However, DistillBERT exhibits poor performance in ***the Devi Bhagavata***.

In Figure 4.4, RoBERTa demonstrates strong performance in ***the Harivamsha Purana***, followed by ***the Caitanya Caritamrita*** and ***the Mahabharata***. However, RoBERTa exhibits poor performance in ***the Devi Bhagavata*** in this context.

4.4.2 Results of Topic classification

Results of BERT-based model: In topic classification, as observed in Table 4.5, within the `PouranicTopic` dataset, BERT achieves the highest precision of 72.10%, recall of 68.11%, and f1-score of 67.79% in ***the Mahabharata***. BERT demonstrates strong performance in both the `Similarity-based` and `Log-likelihood-based` datasets, while displaying moderate performance in the `PouranicTopic` dataset. Specifically, in ***Krsna, the Supreme Personality of Godhead***, BERT achieves the highest precision of 82.50%, recall of 81.44%, and f1-score of 81.26% in the `Similarity-based` dataset. Similarly, in the `Log-likelihood-based` dataset, BERT attains the highest precision of 91.41%, recall of 91.34%, and f1-score of 91.20% in ***Krsna, the Supreme Personality of Godhead***.

Results of DistillBERT-based model: In Table 4.5, within the ***the Mahabharata*** of the `PouranicTopic` dataset, DistilBERT achieves the highest precision at 69.83%, along with a recall and f1-score of 67.53%. Across both the `Similarity-based` and `Log-likelihood-based` datasets, DistilBERT has performed exceptionally well. In

Table 4.5: Comparison of Precision(P), Recall(R) and F1-Score(F1) on all datasets-using BERT, DistilBERT, RoBERTa and Ensemble approach

Mytho-Text	BERT			DistilBERT			RoBERTA			Ensembled		
PouranicTopic dataset												
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
RAMA	18.64	14.90	17.63	17.37	16.94	17.65	21.78	21.84	20.28	21.78	21.84	20.28
MBH	**72.10**	**68.11**	**67.79**	**69.83**	**67.53**	**67.53**	**86.00**	**73.99**	**73.99**	**86.00**	**73.99**	**73.99**
SB	47.30	43.62	42.79	39.03	38.65	38.58	47.45	47.31	44.43	47.45	47.31	44.43
DEVI	51.85	44.50	43.05	43.31	43.31	43.05	50.35	49.02	47.60	51.85	49.02	47.60
CC	35.59	35.59	36.07	35.64	35.64	35.64	38.19	38.19	36.19	38.19	38.19	36.19
HVM	30.75	27.96	26.99	30.32	30.75	29.08	33.54	35.57	32.90	33.54	35.57	32.90
KRSNA	35.34	31.17	31.63	27.44	26.44	26.44	36.21	37.87	34.33	36.21	37.87	34.33
Similarity based dataset												
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
RAMA	68.19	64.79	64.09	67.42	64.31	63.67	70.55	65.73	67.85	70.55	65.73	67.85
MBH	80.27	80.21	79.59	**84.03**	**84.93**	**83.35**	76.57	75.09	72.05	**84.03**	**84.93**	**83.35**
SB	78.98	76.81	76.92	73.55	72.17	72.13	78.35	77.20	77.04	78.98	77.20	77.04
DEVI	77.72	76.64	76.61	68.96	68.25	67.30	73.00	71.64	71.67	77.72	76.64	76.61
CC	79.80	79.23	79.84	76.38	76.70	76.11	55.25	55.50	55.09	79.80	79.23	79.84
HVM	63.11	63.14	63.60	67.22	66.97	66.36	67.39	66.10	65.53	67.39	66.97	66.36
KRSNA	**82.50**	**81.44**	**81.26**	82.07	81.10	80.82	**83.02**	**81.91**	**81.55**	83.02	81.91	81.55
Loglikelihood based dataset												
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
RAMA	64.98	64.74	64.44	60.09	59.92	59.70	0.02	0.98	0.04	64.98	64.74	64.44
MBH	73.23	71.01	69.66	88.70	88.25	88.10	68.20	72.89	68.89	88.70	88.25	88.10
SB	66.79	65.40	64.98	71.77	70.05	69.93	38.04	40.91	37.52	71.77	70.05	69.93
DEVI	56.17	54.89	54.37	64.24	62.27	62.22	4.63	6.95	3.34	64.24	62.27	62.22
CC	68.43	64.69	66.62	72.59	73.00	72.52	0.29	5.05	0.54	72.59	73.00	72.52
HVM	57.46	57.11	56.93	59.22	59.37	59.07	18.18	18.57	17.39	59.22	59.37	59.07
KRSNA	**91.41**	**91.34**	**91.20**	**92.29**	**92.25**	**92.22**	0.01	0.95	0.02	**92.29**	**92.25**	**92.22**

the Mahabharata of the `Similarity based` dataset, it attained the highest precision of 84.03%, recall of 84.93%, and f1-score of 83.35%. Conversely, in ***Krsna, the Supreme Personality of Godhead*** of the `Log-likelihood-based` dataset, DistilBERT achieved the highest precision of 92.29%, recall of 92.25%, and f1-score of 92.22%. The lowest performance of DistilBERT is reported in ***the Ramayana*** with precision 17.37%, recall as 16.94%, and f1-score as 21.78%, respectively.

Results of RoBERTa-based model: In the `PouranicTopic` dataset, as depicted in Table 4.5, RoBERTa excels in ***the Mahabharata***, achieving the highest precision at 86%, with recall and f1-score both reaching 73.99%.In the `Similarity-based` dataset, RoBERTa has surpassed the performance of the `Log-likelihood-based` dataset. Specifically, in ***Krsna, the Supreme Personality of Godhead***, RoBERTa achieved a precision

of 83.02%, recall of 81.91%, and f1-score of 81.55%. In general, RoBERTa performed very well in the `Similarity-based` dataset compared to the `PouranicTopic` dataset. Overall, it can be observed that RoBERTa did not perform well in the `Log-likelihood-based` dataset in comparison to the other two models. In particular, in ***Krsna, the Supreme Personality of Godhead*** of the `Log-likelihood-based` dataset, the performance is very low.

Ensemble-based model: We utilized ensemble-based models incorporating BERT, RoBERTa, and DistilBERT in our study, leveraging their benefits such as enhanced performance, minimized biases, and broader representation. Employing a majority voting approach, we applied this ensemble method to classify topics on all the datasets depicted in Table 4.5.

In the context of Ensemble-based topic classification, the weighted average precision, recall, and f1-score of each model (BERT, DistilBERT, and RoBERTa) are treated as **votes**. The final metrics are determined by selecting the class label with the highest number of votes from the individual models. This ensemble model was applied to all the datasets. As depicted in Table 4.5, our ensemble-based model outperformed the other models across all texts. In particular, within the ***Mahabharata*** of the `PouranicTopic` dataset, our model achieved a precision of 86.20%, a recall of 75.79%, and an F1 score of 74.16%. In contrast, the ***Devi Bhagavata*** had a precision of 55.26%, a recall of 49.23%, and an F1 score of 48.38%. In the `Similarity-based` dataset, the performance of ***the Mahabharata*** is very good in comparison to other texts, with precision of 84.03%, recall of 84.93%, and f1-score of 83.35%.

4.5 Error Analysis

It has been observed that in the Topic classification task, all the models have misclassified a set of sentences in all the datasets. Table 4.6 has demonstrated the total percentage of error done by each model for each dataset.

In the case of the `Similarity-based` dataset, in ***the Ramayana***, DistilBERT has the highest percentage of misclassified sentences (23.30%), followed by RobERTa and BERT. In ***the Mahabharata***, ***the Srimad Bhagavatam***, and ***the Devi Bhagavata***, RoBERTa has the highest percentage (34.11%, 30.08%, and 37.57%, respectively) of misclassified sentences. In ***the Caitanya Caritamrita***, BERT has the highest percentage of misclassified sentences. After that, in ***the Harivamsha Purana*** , RoBERTa (30.56%) and in ***Krsna, the Supreme Personality of Godhead***, DistilBERT (21.44%) have the

Table 4.6: Percentage of misclassified sentences of the models

Mytho-texts	Similarity-based			Log-likelihood-based		
	BERT	DistilBERT	RoBERTa	BERT	DistilBERT	RoBERTa
RAMA	21.03	**23.30**	19.88	35.67	35.97	**98.35**
MBH	25.39	28.86	**34.11**	15.70	10.83	**30.17**
SB	23.49	24.07	**30.08**	30.32	38.50	**64.25**
DEVI	30.21	33.87	**37.57**	34.23	34.25	**62.82**
CC	**28.06**	13.59	12.67	**53.62**	28.50	39.95
HVM	16.63	18.89	**30.56**	56.00	37.34	**76.94**
KRSNA	11.58	**21.44**	9.28	8.92	9.11	**98.58**

highest percentage of misclassified sentences.

Now, in the case of the `Log-likelihood-based` dataset, BERT (53.62%) has the highest percentage of misclassified sentences in ***the Caitanya Caritamrita***. In all other mythological texts, RoBERTa has the highest percentage of misclassified sentences.

To understand the errors while predicting the topic names in depth, we adopted two approaches to investigate the results of both datasets. In the first approach, we tried to find all those sentences for which all the models had made false predictions. Next, in the second approach, we tried to find those topics that were mostly misclassified by the models.

First approach- In ***the Mahabharata***, the following sentence is wrongly predicted by all the three models.

S=*Dhritarashtra said, I think, Destiny is supreme.*

Topic_*trueLabel* = ***Jayadratha-VadhaParva***

BERT_*predLabel* = *Ghatotkacha-badhaParva*

DistilBERT_*predLabel* = *Ghatotkacha-badhaParva*

RoBERTa_*predLabel* = *Abhimanyu-badhaParva*

After that, in ***the Srimad Bhagavatam***, under *the Summum Bonum* canto the following sentence is falsely predicted by all the models.

S=*After all he said, this Krsna is outside the system of Vedic social and spiritual orders and the society of respectable families.*

$\text{Topic}_{trueLabel}$ = ***The Deliverance of Sisupala at the Rajasuya Sacrifice***

$\text{BERT}_{predLabel}$ = *Lord Krsna Returns to the City of Hastinapura*

$\text{DistilBERT}_{predLabel}$ = *Lord Krsna Shows the Universal Form Within His Mouth*

$\text{RoBERTa}_{predLabel}$ = *Draupadi Meets the Queens of Krsna*

From the above cases, it can be understood that to be predicted correctly by the

models, each sentence must have sufficient information related to the assigned label. Less information will tend to give the wrong results.

Second approach- In the case of mostly misclassified topics from our datasets, the observations are given in the following tables 4.7 and 4.8. In these tables, we described the topics with the maximum number of misclassifications by the models.

Table 4.7: Maximum misclassified topics on Similarity based dataset

Mytho texts	True topic	Sample Predicted topics	Models
RAMA	The Departure of Hanuman	Hanuman calms Sita's Fears, Hanuman's takes leave of Sita, Sita gives Hanuman her Jewel,	B,D,R
MBH	Sambhava Parva	Paushya Parva, Khandava-daha Parva, Arjuna-vanavasa Parva,	B,D
SB	The Advent of Lord Krsna-Introduction	Lord Siva Saved from Vrkasura, Draupadi Meets the Queens of Krsna, Lord Krsna's Daily Activities,	B,D,R
DEVI	On the anecdote of Manasa	On the glory of Tulasi, On the anecdote of Sasthi Devi, On the anecdote of Daksina,	B,D,R
CC	Lord Sri Caitanya Mahaprabhu's Travels to the Holy Places	The Lord Travels to Vrndavana, The Lord's Return to Jagannatha Puri, The Lord's Attempt to Go to Vrndavana	B,D,R
HVM	Brahma's Creation	The Creation of Gandharvas Etc, Bali Promises to Give Lands to the Dwarf, An Account of Pirthu and the Churning of the Ocean,	D,R
KRSNA	Delivery of the Message of Krsna to the Gopis	Lord Krsna's Daily Activities, Mother Yasoda Binds Lord Krsna, Krsna Returns to the Gopis,	B,D,R
B=BERT; D=DistilBERT; R=RoBERTa ;			

In the `Similarity-based` dataset (table 4.7), in ***the Ramayana***, in 10 instances, the sentences that are tagged with the topic named *the Departure of Hanuman* are misclassified by all the models. Some sample predicted topics are *Hanuman calms Sita's Fears*, *Hanuman's takes leave of Sita*, and *Sita gives Hanuman her jewel*. In ***the Mahabharata***, 11 times, the ***Sambhava Parva*** topic was not correctly identified by the models BERT and DistilBERT. A few samples of the predicted topics are *Paushya Parva*, *Khandava-daha Parva*, and *Arjuna-vanavasa Parva*. Similarly, in ***the Srimad Bhagavatam***, 14 times the topic named *the Advent of Lord Krsna-Introduction* is falsely identified by all the models.

Examples of some of the predicted topics include *Lord Siva Saved from Vrkasura*, *Draupadi Meets the Queens of Krsna*, and *Lord Krsna's Daily Activities*.

The topic named *On the anecdote of Manasa* of ***the Devi Bhagavata*** was incorrectly identified by all the models 23 times. Among the predicted topics *On the glory of Tulasi*, *On the anecdote of Sasthi Devi*, and *On the anecdote of Daksina* are the few of them. After that, the topic named *Lord Sri Caitanya Mahaprabhu's Travels to the Holy Places* from ***the Caitanya Caritamrita*** was not identified correctly 23 times by all the models.

Some of the predicted topics are exemplified by *The Lord Travels to Vrndavana*, *The Lord's Return to Jagannatha Puri*, and *The Lord's Attempt to Go to Vrndavana.*

Similarly in ***the Harivamsha Purana***, the topic named *Brahma's Creation* was not identified correctly by DistilBERT and RoBERTa five times. Examples of some of the predicted topics include *The Creation of Gandharvas Etc*, *Bali Promises to Give Lands to the Dwarf*, and *An Account of Pirthu and the Churning of the Ocean.* Finally, the topic *Delivery of the Message of Krsna to the Gopis* from ***Krsna, the Supreme Personality of Godhead*** was not identified correctly 16 times by all the models. A few samples from the predicted topics are *Lord Krsna's Daily Activities*, *Mother Yasoda Binds Lord Krsna*, and *Krsna Returns to the Gopis.*

Table 4.8: Maximum misclassified topics in Loglikelihood based dataset

Mytho texts	True topic	Sample Predicted topics	Models
RAMA	Puru takes the place of his Father cursed by Shukra	The Story of the three Sons of Sukesha, The Story of Madhu, Lakshmana takes Sita away,	B,D,R
MBH	Sambhava Parva	Rajya-labha Parva, Paushya Parva, Khandava-daha Parva	B,D,R
SB	Lord Krsna's Daily Activities	Five Queens Married by Krsna, The Story of King Nrga, The Rasa Dance	B,D,R
DEVI	On the anecdote of Manasa	On the glory of Tulasi, On the anecdote of Ganga, On Manasa's story	B,D,R
CC	Lord Sri Caitanya Mahaprabhu Instructs Sanatana Gosvami in the Science of the Absolute Truth	The Activities of Saksi-gopala, The Sixty-One Explanations of the Atmarama Verse, The Process of Devotional Service	B,D,R
HVM	Citralekha Unites Aniruddha with Usha	Shalya Meets Kalayavana, The Defeat of the Asura Naraka, Trial of Arms	B,D,R
KRSNA	Prayers by the Personified Vedas	Akrura's Arrival in Vrndavana, The Deliverance of Lord Siva, Description of Autumn	B,D,R
B=BERT; D=DistilBERT; R=RoBERTa			

In table 4.8, on the other hand, the topic named *Puru takes the place of his Father cursed by Shukra* of ***The Ramayana*** was not correctly identfied by all the models 32 times. Examples of some of the predicted topics includes *The Story of the three Sons of Sukesha*, *The Story of Madhu*, *Lakshmana takes Sita away.* Again, the *Sambhava Parva* of ***the Mahabharata*** was falsely identified 17 times by all the models. A few samples from the predicted topics are *Rajya-labha Parva*, *Paushya Parva* and *Khandava-daha Parva.* Lord *Krsna's Daily Activities* of ***the Srimad Bhagavatam*** was misidentified 31 times. Among the predicted topics the few samples are like *Lord Siva Saved from Vrkasura*, *Draupadi Meets the Queens of Krsna* and *Lord Krsna's Daily Activities.*

The topic *On the Anecdote of Manasa* from ***the Devi Bhagavata*** was incorrectly identified 27 times by all the models. Some of the predicted topics are exemplified by *On the glory of Tulasi*, *On the anecdote of Sasthi Devi* and *On the anecdote of Daksina*. Hereafter, the topic *Lord Sri Caitanya Mahaprabhu Instructs Sanatana Gosvami in the Science of the Absolute Truth* from ***the Caitanya Caritamrita*** was not correctly identified 24 times by the models. A few samples from the predicted topics are *The Lord Travels to Vrndavana*, *The Lord's Return to Jagannatha Puri* and *The Lord's Attempt to Go to Vrndavana*.

The topic *Citralekha Unites Aniruddha with Usha-Aniruddha's Fight with Vana's Soldiers* from ***the Harivamsha Purana*** was incorrectly identified 31 times. Among the predicted topics the samples are like *The Creation of Gandharvas Etc*, *Bali Promises to Give Lands to the Dwarf* and *An Account of Pirthu and the Churning of the Ocean*. Finally, the topic named *Prayers by the Personified Vedas* from ***Krsna, the Supreme Personality of Godhead*** was not correctly identified 44 times by the models. The sample predicted topics are *Akrura's Arrival in Vrndavana*, *The Deliverance of Lord Siva* and *Description of Autumn* etc.

4.6 Discussion and Observations

The comparative performance analysis of BERT, DistilBERT, RoBERTa, and an ensembled model across the ***PouranicTopic***, ***Similarity-based***, and ***Log-likelihood-based*** datasets reveals several key insights into the effectiveness of transformer-based models for topic classification in Hindu mythological texts.

Overall performance- RoBERTa consistently outperforms BERT and DistilBERT in F1-scores across all datasets, especially for the `PouranicTopic` dataset, suggesting its superior architecture for classifying complex mythological texts. The ensemble model, combining BERT, DistilBERT, and RoBERTa, performs comparably to RoBERTa alone, indicating that while ensembling is beneficial, RoBERTa's standalone performance is already strong.

Dataset-specific observations- For the `PouranicTopic` dataset, the performance is generally lower compared to the `Similarity-based` and `Log-likelihood-based` datasets. This is likely due to the inherent complexity and diversity of topics within the `Pouranic` texts, which may challenge the models' ability to generalize.

In the `Similarity-based` dataset, all models exhibit improved performance, with f1-scores notably higher than in the `PouranicTopic` dataset. The use of cosine similarity

scores appears to facilitate better topic classification, likely by providing additional context for semantic alignment.

In the `Log-likelihood-based` dataset, the results show high variability, with some models performing exceptionally well for certain texts, e.g., ***Krsna, the Supreme Personality of Godhead*** with an f1-score of 92.22, but poorly for others, e.g., ***the Ramayana*** with an f1-score of 0.04 using RoBERTa. This suggests that while log-likelihood values can enhance classification for certain topics, they may introduce instability or overfitting issues in others.

Model specific Insights- BERT demonstrates moderate performance, indicating a competent but not outstanding ability in classification tasks. DistilBERT underperforms compared to BERT and RoBERTa, highlighting the trade-off between model size and performance. RoBERTa demonstrates the best overall performance, with consistently high precision, recall, and F1-scores across datasets, including texts like ***the Mahabharata*** and ***Krsna, the Supreme Personality of Godhead***, underscoring its robustness and effectiveness for complex text classification.

The ensemble model's similar performance to RoBERTa suggests that while ensembling can help, the choice of a strong base model (like RoBERTa) is crucial for achieving high performance. The ensembled approach yields high performance, particularly in the `Log-likelihood-based` dataset, demonstrating the potential of model ensembling to enhance robustness and accuracy.

Impact of dataset characteristics- The substantial performance improvement in the `Similarity` and `Log-likelihood-based` datasets compared to the `PouranicTopic` dataset suggests that features like cosine similarity and log-likelihood values provide valuable context. However, the variability in the `Log-likelihood based` indicates that this approach can also introduce noise or complexity, affecting models differently. **Challenges and Limitations**- The low f1-scores for certain texts, e.g., ***the Ramayana*** in the `Log-likelihood-based` dataset with RoBERTa highlight potential limitations in model generalizability and the need for further refinement or additional data preprocessing steps.

4.7 Summary

This study investigates Indian mythological texts using a comprehensive analysis of topic classification. The study begins with an in-depth exploration of cantos and then delves into topic classification. Various BERT-based models and datasets are employed, and precision, recall, and F1 scores are computed to evaluate their performance. While some

models yield satisfactory results, others fall short of expectations. Overall, this research contributes significantly to the field in the following ways:

- **PouranicTopic dataset**-A newly annotated dataset named `PouranicTopic` is developed, compiled from seven major Hindu texts. This dataset serves as a valuable resource for NLP research within the mythological domain.
- **Sentence clustering techniques**- An analysis of sentence clustering techniques is conducted, leveraging semantic similarity and log-likelihood. Two supplementary datasets—`Similarity-based` and `Log-likelihood-based` are crafted.
- **Classification models**- Classification models are devised to identify cantos and topics within mythological documents. Ensemble approaches that combine transformer models show promising results, effectively managing the diversity observed across various texts.
- **Model evaluation**- The study evaluates model performance, covering both canto classification and topic classification across multiple datasets.

The future direction of this research includes expanding the mythological text corpus with multilingual texts, enhancing sentence annotations for better entity relationship mapping, refining clustering techniques for mythological language, and exploring advanced transformer models like T5 and ALBERT. It also aims to create ontologies and knowledge graphs for deeper domain understanding, investigate hybrid models with new features and domain-specific training, and analyze misclassified cases to improve model accuracy. The overarching goal is to continuously innovate in methodologies for progress in mythological text analysis.

Chapter 5

Ontology and Character-Topic Relationship

Ontology and Relations of Character and topics

5.1 Introduction

Analyzing complex ancient texts such as `The Srimad-Bhagavatam` poses significant challenges due to their intricate narratives and philosophical depth. This text, a cornerstone of Hindu philosophy, spans twelve cantos with extensive narratives, hymns, and discourses on Lord Vishnu and his avatars. The complexity of its teachings—encompassing concepts like cosmogony, karma, and the path to liberation—creates difficulties for scholars in interpretation and analysis.

To address these challenges, this research explores two key methods: *domain ontology development* and *character-topic relationship* analysis. **Domain Ontology**, which specifies concepts within a domain, identifies entities, reveal familial relationships, thematic links and constraints (Gruber, 1993). These ontologies provide valuable insights (Jiang, 2012), helping humans and intelligent software agents understand character and topic in the text. On the other hand, **Character-Topic relationship** analysis utilizes computational techniques to uncover thematic connections between characters and topics, enhancing scholarly understanding and accessibility.

The research gaps in domain ontology and character-topic relationship analysis for the Srimad-Bhagavatam are significant and multifaceted. Firstly, there is a marked **deficiency in comprehensive domain ontology datasets** specifically tailored to Hindu mythology. This gap restricts the development of detailed ontologies that accurately capture the intricate relationships and thematic elements within `Srimad-Bhagavatam`. The existing datasets are insufficient to represent the complex network of characters and their

interactions in the text. Secondly, the lack of **robust character-topic relationship datasets** hampers the ability to analyze how characters are associated with various topics throughout `Srimad-Bhagavatam`, affecting the depth and accuracy of such analyses.

Furthermore, challenges in **representing complex character roles, addressing polysemy, and integrating cultural nuances** add to the difficulty of developing effective domain ontologies and character-topic analyses. The rich cultural and historical context of `Srimad-Bhagavatam` makes it challenging to align computational models with traditional interpretations.

To overcome these challenges, there is a need for improved methods to **validate findings with established scholarly perspectives** and to bridge the gaps between computational results and traditional interpretations. Addressing these issues is crucial for advancing both domain ontology and character-topic relationship analysis, thereby enhancing the overall understanding of `the Srimad-Bhagavatam`. The primary objectives of this research are twofold:

- **Domain Ontology Development**: Develop a comprehensive domain ontology for the Srimad-Bhagavatam that identifies characters, their relationships, and related topics. This ontology will offer a structured framework to capture intricate connections and provide a clearer understanding of the text's narrative and philosophical dimensions.
- **Character-Topic Relationship Analysis**: Conduct a detailed analysis of character-topic relationships using advanced computational techniques. This includes:
  - Developing models based on BERT, RoBERTa, and DistilBERT to capture nuanced connections between characters and topics.
  - Exploring an ensemble approach to improve accuracy and effectiveness of the analysis.

  This multi-model strategy aims to uncover deeper insights into character dynamics and thematic relationships.

The contributions of this research include:

- Development of specialized datasets for Character Ontology and Topic Ontology, forming the foundation for a comprehensive domain ontology tailored to the Srimad-Bhagavatam.

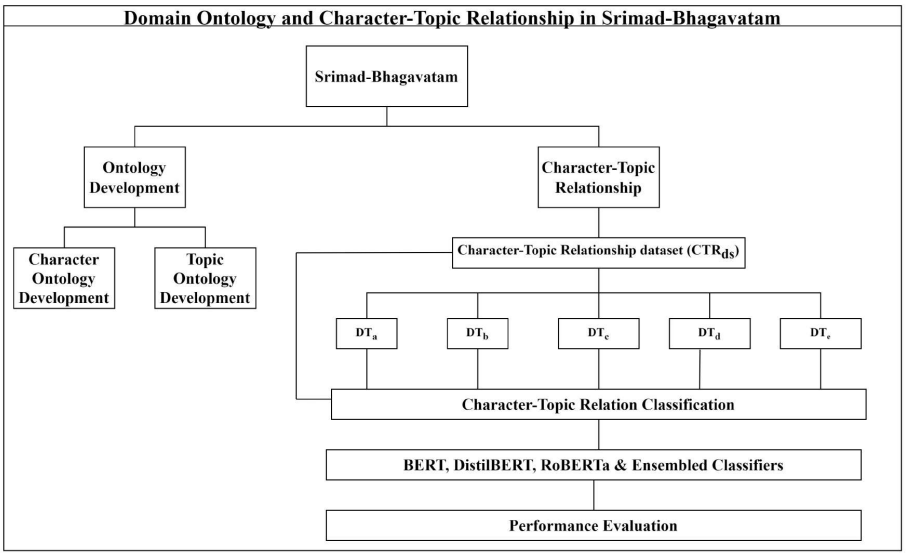


Figure 5.1: Proposed Research Framework

- Creation of a Character-Topic Relationship dataset to facilitate detailed analysis of character-topic connections within the text.
- Development of a domain ontology revealing intricate familial relationships and thematic links.
- Implementation of models using BERT, RoBERTa, and DistilBERT, as well as an ensemble approach, to effectively analyze character-topic relationships.
- Application of these analyses to uncover hidden patterns and gain deeper insights into the Srimad-Bhagavatam's complex narratives and philosophical dimensions.

The proposed research framework is illustrated in Figure 5.1. The figure highlights two main components of this research. The first component focuses on the development of domain ontologies, including Character Ontology and Topic Ontology. The second component involves utilizing the CTR_{ds} dataset to create five distinct sub-datasets (DT_a, DT_b, DT_c, DT_d, DT_e). These sub-datasets along with CTR_{ds} are then used to evaluate the performance of various models, such as BERT, DistilBERT, RoBERTa, and an ensemble approach.

The overall structure of the research work is organized as follows: The *Preparation of Datasets* section details the dataset preparation process. The *System Framework* section outlines the proposed approach. This is followed by the *Result Analysis*, *Error Analysis*, and *Discussion & Observations* sections, which present the results, discuss errors, and offer insights from the tasks. The paper concludes with the *Summary* section, summarizing the

findings and suggesting directions for future research.

5.2 Preparation of Datasets

Preparing datasets for studying domain ontology and character-topic relationships in Srimad-Bhagavatam involves a methodical approach to capture and analyze the extensive thematic and narrative complexities of this ancient scripture. (Paul et al., 2024) created an extensive data set for the Pouranic topic classification challenge based on Indian mythology. We have selected `the Srimad-Bhagavatam` from that dataset for the domain ontology development and character-topic relationship tasks in this study. The procedures for creating the datasets for the character-topic relationship and domain ontology tasks are covered in the sections below.

5.2.1 Datasets for domain ontology development

We separated our work into two distinct subtasks for the domain ontology: character and topic ontology development, and we then prepared the dataset accordingly.

Characters and their categories- The text contains 1714 individuals, or characters of which some are male, some female, and others of a different gender. There are 17 various categories of characters we found in the text, viz., *God*, *Demigod*, *Devarsi*, *Gandharva*, *Celestial-beings*, *Yaksaraja*, *Raksasa*, *Raksasi*, *Danava*, *Daitya*, *Demon*, *She-demon*, *Witch*, *Human*, *Animal*, *River*, *Ocean*.

Under *God* category, there are 4 different sub-categories present such as, *Godhead*, *Goddess*, *Lord's-spiritual-energy* and *Bhagavan*. Again,under *Human* category, there are 15 sub-categories found such as- *King*, *Queen*, *Maharaja*, *Prince*, *Sage*, *Muni*, *Prajapati*, *Priest*, *Devotee*, *Rsi*, *Saint*, *society-girl*, *the-commander-in-chief*, *Manu*, *Vasu*.

Additionally, under the *Animal* category, the sub-categories identified are: *Elephant*, *Horse*, *Snake*, and *Serpent*. Examples to illustrate the categories of characters are provided below:

Ex-1: `Durga` ***rdf:type*** `Goddess`

Ex-2: `Narada` ***rdf:type*** `Devarsi`

Ex-3: `Bharata` ***rdf:type*** `Maharaja`

Here `Durga` is classified as a `Goddess`, `Narada` as a `Devarsi`, and `Bharata` as a `Maharaja`.

Family Relations between Characters- There are 24 different types of family relations found in the text, including: *brother-in-law of*, *brother of*, *carrier of*, *daughter*

of, *descendant of*, *devotee of*, *disciple of*, *enemy of*, *father-in-law of*, *father of*, *friend of*, *grandson of*, *great-grandson of*, *husband of*, *killer of*, *known as*, *master of*, *mother of*, *servant of*, *sister of*, *son of*, *spiritual master of*, and *wife of*.

An example to illustrate family relations in the text:

Ex-4: *My dear King Pariksit, your father, Abhimanyu, was born from the womb of Subhadra as the son of Arjuna.*

In the above sentence we can find that the character ***Subhadra*** is the mother of ***Abhimanyu*** and ***Arjuna*** is his father. Whereas, the character ***Pariksit*** is a King, who is the son of ***Abhimanyu***.

Topics in the Srimad-Bhagavatam- The `Srimad-Bhagavatam` is divided into 12 cantos. Each canto contains a varying number of chapters, some with fewer chapters and others with more extensive ones. These topics encompass stories, dialogues, teachings, and philosophical discussions. For example, the Second canto, titled ***The Age of Deterioration***, contains 10 topics: *The First Step in God Realization*, *The Lord in the Heart*, *Pure Devotional Service: The Change in Heart*, *The Process of Creation*, *The Cause of All Causes*, *Purusa-sukta Confirmed*, *Scheduled Incarnations with Specific Functions*, *Questions by King Pariksit*, *Answers by Citing the Lord's Version*, and *Bhagavatam Is the Answer to All Questions*.

The detailed statistics of cantos and topics in the `Srimad-Bhagavatam` are discussed in the Table 5.1.

Table 5.1: Canto and Topic Statistics in Srimad-Bhagavatam

Text	Canto Name	#Topics
Srimad-Bhagavatam	Creation	19
	The Cosmic Manifestation	10
	The Status Quo	33
	The Creation of the Fourth Order	31
	The Creative Impetus	26
	Prescribed Duties for Mankind	19
	The Science of God	15
	Withdrawal of the Cosmic Creations	24
	Liberation	24
	The Summum Bonum	90
	General History	31
	The Age of Deterioration	13

Each topic discussed about different characters or individuals in the text. In the 9th canto named as `Liberation`, the character ***Ramacandra*** has appeared in the topics such as, *The Pastimes of the Supreme Lord, Ramacandra*; *Lord Ramacandra Rules the World* and *The Dynasty of Kusa, the Son of Lord Ramacandra*.

5.2.2 Datasets for Character-Topic relationship

Initially, we extracted every sentence from `the Srimad-Bhagavatam` for each character. We then computed sentiment-related attributes for each sentence, including *polarity*, *subjectivity*, *positivity*, and *negativity*, using the *NaiveBayesAnalyzer* of `TextBlob`[1]. The formulas for these parameters are described below.

$$\text{Polarity} = \frac{\text{Pos sentiment} - \text{Neg sentiment}}{\text{Total sentiment}} \quad (5.1)$$

$$\text{Subjectivity} = \frac{\text{Subjective words or phrases}}{\text{Total words or phrases}} \quad (5.2)$$

$$\text{Positivity} = \frac{\text{Positive words or phrases}}{\text{Total words or phrases}} \quad (5.3)$$

$$\text{Negativity} = \frac{\text{Negative words or phrases}}{\text{Total words or phrases}} \quad (5.4)$$

We then calculated these metrics for each character based on the topics they appear in and aggregated them to determine the relationships and patterns.

Positivity and negativity scores are calculated as the ratio of positive and negative words to the total words in each sentence. Relevant events are extracted as verbs or verb particles. Using the **PouranicTopic** dataset, sentences with specific characters are tagged with corresponding cantos and topics, creating the CTR_{ds} dataset. This dataset, which includes 45,941 observations, illustrates how characters appear across various sentences, topics, and cantos. For an example, refer to Table 5.2.

The table 5.2 analyzes a sentence related to the character Ramacandra from the Liberation section. It assesses the sentence's polarity as positive (1), and subjectivity as objective (0). The positivity score is 0.62, and the negativity score is 0.37, resulting in an overall positive sentiment (pos). The verbs associated with the sentence are *brahmanas* and *engaged*, and the topic is ***Lord Ramacandra Rules the World***.

We created five distinct datasets (DT_a, DT_b, DT_c, DT_d, DT_e) from the CTR_{ds} dataset to evaluate the impact of various features on character-topic relationships. Each dataset comprises different feature combinations: DT_a includes *Character*, *Canto*, *Sentence*,

[1]https://textblob.readthedocs.io/en/dev/

and *VP*; DT_b contains *Character*, *Canto*, *Sentence*, *Pol*, *Senti*, and *VP*; DT_c features *Character*, *Sentence*, and *VP*; DT_d has *Character*, *Canto*, and *VP*; and DT_e comprises *Character*, *Canto*, *Sentence*, *Pos*, *Senti*, and *VP*. We tested these datasets using advanced transformer classifiers including BERT, DistilBERT, RoBERTa, and an ensemble model that integrates all three classifiers.

Table 5.2: Sample observation of CTR_{ds}

Character	Canto	Sentence	Pol	Subj	Pos	Neg	Senti	VP	Topic
Ramacandra	Liberation	All the brahmanas who were	1	0	0.62	0.37	pos	'brahmanas', 'engaged'	Lord Ramacandra Rules the World
Pol=Polarity; Subj=Subjectivity; Pos=Positivity; Neg=Negativity; Senti=Sentiment; VP=Verb or a particle attached to a verb									

Table 5.3: The statistical analysis of the CTR_{ds} dataset

Pol Range	Subj Range	Pos Range	Neg Range	Total Pos	Total Neg	Frequent Top 5 Topics
-1 to +1	0 to 1	0.002 to 1	0.002 to 0.99	38296	7645	Gajendras Prayers of Surrender The Killing of the Demon Trnavarta The Descendants of Ajamidha The Gopis Songs of Separation Five Queens Married by Krsna The Liberation of King Jarasandha The Sages Teachings at Kuruksetra
Pol Range: Polarity score range; Subj Range: Subjectivity score range; Pos Range: Positivity range; Neg Range: Negativity score range; Total Pos/Neg: Total Positively/Negatively classified sentences						

Table 5.3 provides a statistical summary of the CTR_{ds} dataset. It details the range of Polarity scores (-1 to +1), Subjectivity (0 to 1), Positivity (0.002 to 1), and Negativity (0.002 to 0.99). The dataset includes 38,296 positive and 7,645 negative sentences. Key topics include *Gajendras Prayers of Surrender*, *The Killing of the Demon Trnavarta*, *The Descendants of Ajamidha*, *The Gopis Songs of Separation*, and *Five Queens Married by Krsna*.

5.2.3 Inter-Annotator Agreements

The assessment of Inter-Annotator Agreement (IAA) is a crucial step in ensuring the reliability and consistency of annotations within the dataset used for this research. In this study, we conducted IAA,*Cohen's Kappa coefficient* (McHugh, 2012), for the Character Topic Relationship Dataset (CTR_{DS}) to evaluate the alignment between two annotators

in determining whether a character belongs to a specific topic. The guidelines are as follows:

- Clearly define the goal of identifying whether a character is associated with a particular topic. Emphasize the importance of contextual relevance and thematic coherence in making these determinations.
- Use a binary scale to assess the relationship between characters and topics. Assign a score of 1 if a character is deemed relevant to the topic and a score of 0 if there is no relevance. This straightforward system facilitates consistency in decision-making.
- Concentrate on gauging the semantic alignment of characters with topics, taking into account the context.
- Exercise discretion in assessing whether each character contributes substantively to the topic under consideration.

Table 5.4: Inter-Annotator Agreement for Topic-Centric Dataset Based on Cohen's Kappa Coefficient

Topic-Centric Dataset	Annotator 2			$\kappa = 0.77$
		Belongs to Topic	Does Not Belong to Topic	
Annotator 1	Belongs to Topic	38000	300	
	Does Not Belong to Topic	1500	5140	

By employing Cohen's Kappa coefficient, we aimed to quantify this agreement, yielding a Kappa score of 0.77 in table 5.4. This score reflects a substantial level of agreement, indicating that the annotations are both reliable and valid for the subsequent analyses performed in this research.

5.2.4 Splitting of datasets

To train the models, we divided the dataset into training, validation, and test sets using an 80-10-10 split. Specifically, the training set contains 36,752 samples, while the validation and test sets contain 4,594 samples each. The dataset includes a total of 1,635 unique characters and 335 unique topics. The training set consists of 1,568 unique characters and

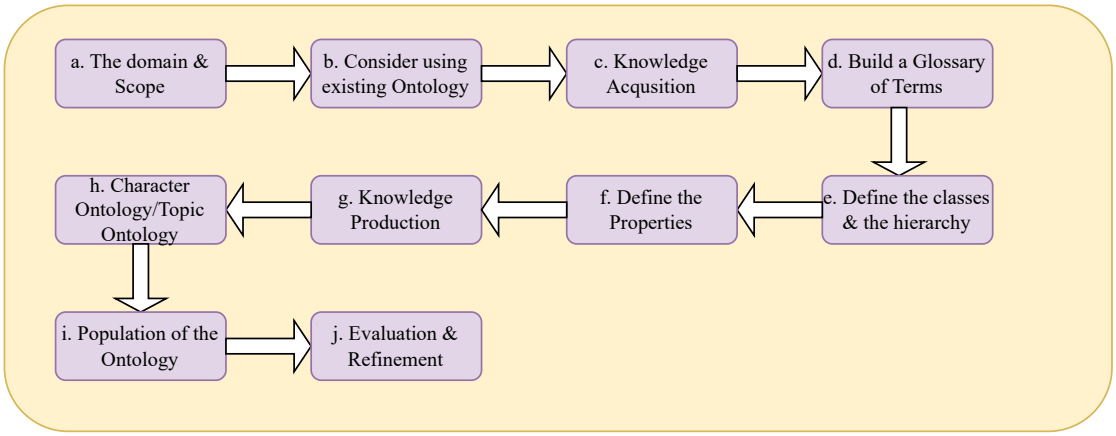


Figure 5.2: Proposed Model Framework for Ontology Development

335 unique topics. The validation set has 761 unique characters and 333 unique topics, while the test set contains 758 unique characters and 331 unique topics. Since we created five distinct datasets (DT_a, DT_b, DT_c, DT_d, DT_e) from the CTR_{DS} dataset, the number of unique characters and unique topics remain the same across all datasets. It can be observed in Table 5.5

Table 5.5: Train, Test and Val splitting of CTR_{DS}

Set	Samples	Unique Characters	Unique Topics
Train	36,752	1,568	335
Validation	4,594	761	333
Test	4,594	758	331

5.3 Domain Ontology Development System

The Character and Topic ontologies for `the Srimad-Bhagavatam` were developed following Stanford University's guidelines Noy and McGuinness (2001). The Character ontology is referred to as `SBC-Ontology`, and the Topic ontology as `SBT-Ontology`. The framework is illustrated in Figure 5.2.

To clarify the motivation, competency questions in natural language are used to define the scope and coverage of both the ontologies. These questions help ensure that the ontologies, once developed, can address all relevant queries. Below are examples of informal queries for both `SBC-Ontology` and `SBT-Ontology`:

- SBC-Q1. Who is Abhimanyu?

- SBC-Q2. Who is the Son of Arjuna and Subhadra?
- SBC-Q3. Who is Ramacandra?
- SBT-Q4. What is Srimad-Bhagavatam?
- SBT-Q5. What are the names of the cantos present in Srimad-Bhagavatam?
- SBT-Q6. What are the names of the topics available in each canto in Srimad-Bhagavtama?

Existing Ontology: A comprehensive literature search revealed that there are very few ontologies specifically for mythology. Notably, the Mahabharata ontology [2], created by a research team at IIT Madras, is a unique contribution and the only one of its kind in this area.

Knowledge acquisition and glossary of terms : Based on research by (Paul and Das, 2017b) and (Paul and Das, 2017a), we extracted and categorized characters and topics from `the Srimad-Bhagavatam`. The `SBC-Ontology` includes characters, their genders, and genealogical relations, while `SBT-Ontology` covers cantos, topics, hierarchies, and sentences.

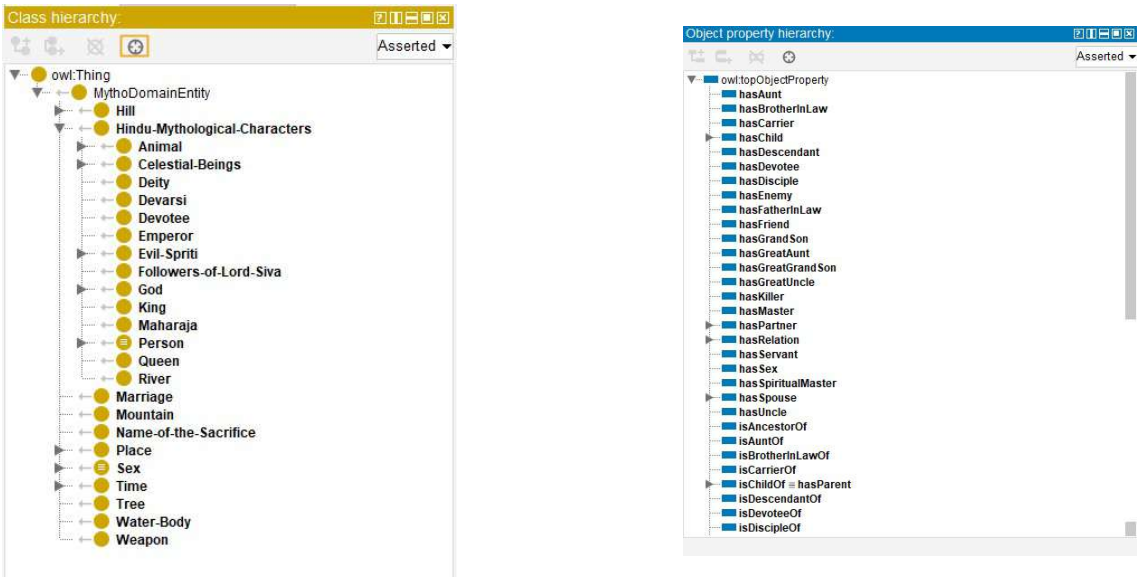
Define the classes and the hierarchy: Terms in `SBC-Ontology` are grouped into categories like Animal, Celestial-Beings, Deity, Devotee, Evil-Spirit, God, and Person, as shown in Figure 5.3a. `SBT-Ontology` features classes such as Canto, Topic, Person, and Sentence, depicted in Figure 5.4a.

Define the Properties: In `SBC-Ontology`, key properties include *hasFather*, *isFatherOf*, *hasChild*, *isChildOf*, *hasWife*, and *isWifeOf* (Figure 5.3b). `SBT-Ontology` includes properties such as *hasSentence*, *hasTopic*, *isBelongsToCanto*, and *isUnderCanto* (Figure 5.4b).

Knowledge production: In this phase, fragmented data is integrated to enhance the ontologies' semantic depth. In `SBC-Ontology`, *MythoDomainEntity* is the top-level class, with other classes defined using the *subClassOf* property.

Character Ontology: `SBC-Ontology` for `the Srimad-Bhagavatam` uses annotation properties for metadata (e.g., ***isGrandsonOf***, ***isServentOf***) and object properties for complex relationships, like lineage (***hasAncestor***) and parent-child connections (***hasChild***). It also includes inverse familial properties like ***hasAunt*** and ***hasBrotherInLaw***. The `SBC-Ontology` comprises 3,530 axioms (2,551 logical and 972 declaration), defines 86 classes, and 85 object properties. It includes 797 individuals and 4 annotation properties, reflecting its extensive and detailed semantic structure.

[2] https://sites.google.com/site/ontoworks/about-us



(a) Class-Hierarchy **(b)** Object-Properties

Figure 5.3: Class-Hierarchy and Object Properties of `SBC-Ontology`

(a) Class-Hierarchy **(b)** Object-Properties

Figure 5.4: Class-Hierarchy and Object Properties of `SBT-Ontology`

Topic Ontology: `SBT-Ontology` organizes concepts into classes like Mythology, Canto, Topic, Sentence, and Person. It defines properties linking these classes, such as *hasTopic* and *isUnderCanto*. The `SBT-Ontology` includes 1,457 axioms, with 1,258 being logical. It defines 8 classes, 9 object properties, and 451 individuals. The ontology features 4 subclass relationships, 1 equivalent class, and 1 disjoint class. Object properties include 1 inverse and 1 functional property, along with 9 domains and ranges. Individual axioms comprise 178 class assertions, 1,053 property assertions, and 1 same individual. Additionally, there are 182 annotation assertions.

Obtain expert feedback and Population of the ontology: We consulted domain experts to validate the ontology structure and content. Protégé[3] was used for formalization, as it is a preferred tool among experts.

Evaluation and refinement: Ontology evaluation focuses on determining two crucial aspects of ontologies: their quality and correctness (Hlomani and Stacey, 2014). Here we have done manual evaluation by the experts. Domain experts reviewed the ontologies to ensure its correctness and completeness.

Comparative Study: Performance and Insights of the Proposed Ontologies We compared our ontology with two prominent models: the Mahabharata Ontology and the Ontology of Greek Mythology. The metrics used for comparison are as follows:

- **Coverage:** Our ontology includes 86 classes and 85 object properties, surpassing the Mahabharata Ontology (40 classes, 50 properties) and the Greek Mythology Ontology (60 classes, 70 properties).
- **Semantic Depth:** Our ontology comprises 3,530 axioms, including 2,551 logical axioms, providing deeper semantic insights compared to the existing models.
- **Interoperability:** Designed for cross-domain use, our ontology can be linked with global knowledge graphs like DBpedia[4].

5.4 Character-Topic Relationship System

Understanding the relationship between characters and topics in `the Srimad-Bhagavatam` involves analyzing how different characters engage with and contribute to various topics throughout the text.

[3] https://protege.stanford.edu/
[4] https://www.dbpedia.org/

Initially, we analyzed the CTR_{ds} dataset by performing several key tasks such as: **Frequent characters and topics**, **Correlation Analysis**, **Aggregate sentiment analysis**, **Topic distribution**, **Verbal phrases analysis**, **Character comparisons** and finally **Model development**.

Frequent characters and topics - We first pinpointed the top 5 characters and topics by frequency to highlight their importance in `the Srimad-Bhagavatam`. Table 5.6 shows that ***Krsna*** is the most frequently mentioned character, while ***Krsna, the Supreme Personality*** is the most common topic.

Table 5.6: Top 5 Frequent Characters and Topics

Characters	**Freq.**	**Topics**	**Freq.**
Krsna	6445	Krsna, the Supreme Personality	792
Supreme Personality of Godhead	1574	Descendants of Ajamidha	641
Balarama	1017	Advent of Lord Krsna	531
Vasu	956	Diti Vows to Kill Indra	509
Brahma	951	Pastimes of Lord Ramacandra	421

Character-Topics Associativity The table 5.7 presents the most frequent topics associated with various characters in the CTR_{ds} dataset, along with the count of occurrences of each topic. For instance, the character **Krsna** is most frequently mentioned in the context of *"Krsna, the Supreme Personality of Godhead"*, with a total of 106 occurrences. Other characters like **Balarama** and **Vasu** are also strongly associated with this topic, appearing 21 and 20 times, respectively. Additionally, characters such as **Kamsa**, **Brahma**, and **Vasudeva** are primarily linked with *"The Advent of Lord Krsna-Introduction"*, highlighting the frequent connections between these characters and topics related to Krsna's life.

The table 5.8 presents the most frequent characters associated with various topics from a dataset related to Hindu mythology. Each row corresponds to a topic and lists the characters linked to it, along with the frequency of their occurrences.

For instance, in the topic *Genealogical Table of the Daughters of Manu*, Krsna appears 46 times, followed by Vasu with 11 occurrences and Brahma with 9. The topic *The Advent of Lord Krsna-Introduction* highlights Krsna's prominence with a total of 85 mentions, alongside Kamsa and the Supreme Personality of Godhead, each appearing 20 times. This demonstrates the significant roles these characters play in the narratives associated with each topic.

Table 5.7: Most Frequent Topics for Characters

Character	**Most_Frequent_Topic**	**Count**
Krsna	Krsna, the Supreme Personality of Godhead	106
the Supreme Personality of Godhead	The Descendants of Ajamidha	22
Balarama	Krsna, the Supreme Personality of Godhead	21
Vasu	Krsna, the Supreme Personality of Godhead	20
Kamsa	The Advent of Lord Krsna-Introduction	20
Brahma	The Advent of Lord Krsna-Introduction	14
Bali	Bali Maharaja Conquers the Heavenly Planets	14
Indra	Diti Vows to Kill King Indra	13
Vasudeva	The Advent of Lord Krsna-Introduction	12
Sukadeva Gosvami	The Advent of Lord Krsna-Introduction	12
Arjuna	Krsna, the Supreme Personality of Godhead	12
Siva	Lord Krsna Fights with Banasura	11

Correlation Analysis - Next, we performed a correlation analysis on sentiment attributes—polarity, subjectivity, positivity, and negativity—to reveal their relationships with characters and topics. We compared **Pearson**[5], **Spearman**[6], and **Kendall**[7] correlations using the four numerical features in the CTR_{ds} dataset, with results presented in Table 5.9.

The correlation matrices for *polarity*, *subjectivity*, *positivity*, and *negativity* show consistent patterns across **Pearson**, **Spearman**, and **Kendall** methods. All three methods indicate a positive correlation between `polarity` and `subjectivity`, with **Pearson** showing the strongest correlation (0.53) and **Kendall** the weakest (0.48). `Positivity` and `negativity` are perfectly inversely correlated (-1.00) in all matrices. The correlations between `polarity` and `positivity`, and between `polarity` and `negativity`, are moderate and inverse, showing similar values across methods. `Subjectivity` has a moderate positive correlation with `positivity` and a moderate negative correlation with `negativity`, with **Spearman** indicating slightly stronger correlations than **Pearson** and **Kendall**. Overall, the matrices reveal a robust relationship between sentiment attributes, with `positivity` and `negativity` showing the most significant inverse correlation.

[5] *https://en.wikipedia.org/wiki/Pearson_correlation_coefficient*
[6] *https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient*
[7] *https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient*

Table 5.8: Top Frequent Characters Associated with Various Topics

Topic	Top_Characters
Genealogical Table of the Daughters of Manu	Krsna (46), Vasu (11), Brahma (9), the Supreme Personality of Godhead (9), Balarama (7)
King Sudyumna Becomes a Woman	Krsna (37), Siva (9), Manu (7), Vasu (7), Pariksit (6)
Questions by Vidura	Krsna (26), the Supreme Personality of Godhead (5), Brahma (4), Indra (4), Nara (4)
Questions by the Sages	Krsna (8), Balarama (4), the Supreme Personality of Godhead (4), Siva (3), Vasu (3)
The Advent of Lord Krsna-Introduction	Krsna (85), Kamsa (20), the Supreme Personality of Godhead (20), Balarama (15), Brahma (14)
The Curse upon the Yadu Dynasty	Krsna (20), Brahma (5), Nara (5), Pariksit (4), Vasu (4)
The First Step in God Realization	Krsna (4), Asvini-kumaras (2), Balarama (2), Bharata (2), Nara (2)
The History of the Life of Ajamila	Krsna (12), Balarama (5), Nara (5), Indra (4), Visnu (4)
The Supreme Lord Is Equal to Everyone	Krsna (40), Vasu (11), Brahma (6), Sukadeva Gosvami (6), the Supreme Personality of Godhead (6)
Ajamila Delivered by the Visnudutas	Krsna (15), the Supreme Personality of Godhead (8), Ajamila (5), Visnu (5), Nara (4)
Daksa Curses Lord Siva	Krsna (9), Siva (5), Bali (4), Daksa (4), Brahma (2)

Aggregate sentiment analysis - In this case, we carried out aggregate sentiment analysis to assess the overall emotional tone across different characters, cantos, and topics. This provided insights into the narrative's emotional landscape.

The table 5.10 lists the top 5 characters with the highest positive and negative sentiment scores. ***Krsna*** leads with the highest positive score (4693), while ***Abhramu*** has the highest negative score (152).

Topic distribution - This task aims to identify the most frequent topics linked to specific characters. Understanding how topics are distributed across the text allows us to see which topics are most relevant to particular characters. The character ***Krsna*** appears most frequently in the following top 5 topics: *Krsna, the Supreme Personality*

Table 5.9: Comparative Study of different correlation metrics on the dataset

	Polarity			Subjectivity			Positivity			Negetivity		
	PS	SP	KD	PS	SP	KD	PS	SP	KD	PS	SP	KD
Polarity	1.00	1.00	1.00	**0.53**	0.51	0.48	0.23	0.22	0.17	-0.22	-0.22	-0.17
Subjectivity	**0.53**	0.51	0.48	1.00	1.00	1.00	0.29	0.33	0.26	-0.29	-0.33	-0.26
Positivity	0.22	0.22	0.17	0.29	**0.34**	0.26	1.00	1.00	1.00	-1.00	-1.00	-1.00
Negetivity	-0.22	-0.22	**-0.17**	-0.29	-0.34	-0.26	-1.00	-1.00	-1.00	1.00	1.00	1.00
PS: Pearson correlation; SP: Spearman correlation; KD: Kendall correlation												

Table 5.10: Top 5 characters with highest positive/negative sentiments

Character	Pos senti score	Character	Neg senti score
Krsna	4693	Abhramu	152
The Supreme Personality of Godhead	936	Airavana	144
Brahma	936	Arimardana	132
Balarama	873	Asana	116
Vasu	824	Asanga	107

of Godhead (106 times), *The Descendants of Ajamidha* (90 times), *The Advent of Lord Krsna-Introduction* (85 times), *Five Queens Married by Krsna* (74 times), and *The Story of the Syamantaka Jewel* (66 times).

Verbal phrases analysis - This task involves analyzing the most common verbal phrases (VP) associated with different topics. Identifying these phrases helps reveal how characters express themselves and how certain topics are discussed, adding depth to our understanding of the text. In the topic named as *Krsna, the Supreme Personality of Godhead*, the most common verbal phrases include *named*, *appeared*, *killed*, *married*, and *worshiped* among others.

Character comparisons - This task compares characters based on their sentiment profiles, or textual patterns. By juxtaposing different characters, we can draw comparisons that highlight similarities, differences, and the unique contributions each character makes to the overall narrative. In the following tabel 5.11 we described a list of the top 3 terms (words) associated with each character, based on their TF-IDF scores. For instance, Krsna is linked with ***lord*** (0.63), ***sri*** (0.37), and ***supreme*** (0.26). Each character has a unique set of words that reflect their prominence in the text, with their corresponding TF-IDF scores.

Table 5.11: List of top 3 terms (words) associated with each character, based on their TF-IDF scores

Character	words (tf-idf score)
Krsna	lord (0.63), sri (0.37), supreme (0.26)
the Supreme Personality of Godhead	supreme (0.62), personality 0.61),godhead (0.24)
Abhimanyu	son (0.32) ,pariksit (0.25), womb (0.23)
Balarama	krsna(0.71), lord(0.30), cowherd (0.12)
Vasudeva	lord(0.64), devaki(0.11), kamsa(0.11)

Finally, we developed and employed various classification models using advanced transformer architectures, such as BERT, DistilBERT, and RoBERTa, to analyze and refine character-topic associations.

Model Training- The Character-Topic Relationship Model is designed to analyze and predict the relationships between characters and topics in mythological texts using advanced natural language processing techniques. This model leverages various transformer-based architectures to understand how different characters interact with specific topics across texts, identify recurring patterns or associations, and observe the evolution of these relationships over time.

For this analysis, we utilized the datasets specifically curated for this study, comprising extensive mythological texts with detailed annotations on character-topic relationships. These datasets provide a robust foundation for training and evaluating our models.

To achieve our objectives, we employed several cutting-edge transformer models, including BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), and RoBERTa (Liu et al., 2019). BERT offers comprehensive contextual understanding, making it ideal for capturing nuanced relationships. DistilBERT, a lighter and faster variant, facilitates quicker training and inference, while RoBERTa enhances BERT's capabilities with improved pretraining techniques. Each of these models was chosen for their unique strengths, which collectively contribute to a more nuanced analysis of character-topic relationships.

Each model was trained using specific parameters tailored to optimize performance for our task. For BERT, we used the base version featuring 12 layers, 768 hidden units, and 110 million parameters, with a learning rate of 2e-5, a batch size of 32, and a maximum sequence length of 128 tokens. DistilBERT was configured with 6 layers, 768 hidden units, and 66 million parameters, employing a learning rate of 3e-5, a batch size of 32, and the same maximum sequence length of 128 tokens. RoBERTa was set up with 12 layers, 768

hidden units, and 125 million parameters, utilizing a learning rate of 1e-5, a smaller batch size of 16, and a longer maximum sequence length of 256 tokens. During training, we applied callbacks based on accuracy to monitor and optimize model performance. To evaluate the models, we used categorical cross-entropy as the loss function and compiled them with the Adam optimizer.

To evaluate our models, we utilized weighted average precision, recall, and F1-score metrics. Precision measures the accuracy of predicted relationships, ensuring the model minimizes false positives. Recall assesses how well the model identifies all actual relationships, capturing the completeness of its predictions. The F1-score balances precision and recall, providing a comprehensive view of model performance, especially useful in imbalanced datasets. Together, these metrics offer a detailed assessment of the model's effectiveness in analyzing and predicting character-topic relationships within the mythological texts.

We implemented various classification models, including an ensembled approach, to maximize performance.

Ensembled Classifier- This research introduces an ensembled classifier combining BERT, DistilBERT, and RoBERTa to enhance performance in Character-Topic relationship tasks. The ensemble trains each model independently and merges their outputs via majority voting, leveraging each model's strengths to improve Precision, Recall, and F1-score.

5.5 Result Analysis

The table 5.12 provides a detailed comparison of the performance metrics across different models on various datasets. The models evaluated include BERT, DistilBERT, RoBERTa, and an Ensemble approach. Each row in the table represents a different dataset, denoted as DT_a, DT_b, DT_c, DT_d, DT_e, and CTR_{ds}.

The BERT classifier in table 5.12 demonstrates consistent performance across datasets. For DT_a, it achieves a Precision of 0.70, Recall of 0.65, and F1-Score of 0.67. In DT_b, it has Precision of 0.68, Recall of 0.64, and F1-Score of 0.66. For DT_c, BERT scores 0.69 in Precision, 0.67 in Recall, and 0.68 in F1-Score. Performance is slightly lower in DT_d with Precision at 0.66, Recall at 0.63, and F1-Score at 0.64. In DT_e, it records Precision of 0.67, Recall of 0.64, and F1-Score of 0.65. For the CTR_{ds} dataset, BERT achieves Precision of 0.65, Recall of 0.63, and F1-Score of 0.64.

DistilBERT in table 5.12 performs robustly across datasets. For DT_a, it achieves a

Table 5.12: Comparison of Precision(P), Recall(R) and F1-Score(F1) on all datasets using BERT, DistilBERT, RoBERTa and Ensemble approach

	BERT			DistilBERT			RoBERTa			Ensemble		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
DT_a	0.70	0.65	0.67	0.75	0.70	0.72	0.72	0.68	0.70	**0.80**	**0.78**	**0.79**
DT_b	0.68	0.64	0.66	0.73	0.69	0.71	0.70	0.66	0.68	**0.78**	**0.76**	**0.77**
DT_c	0.69	0.67	0.68	0.74	0.71	0.73	0.71	0.69	0.70	**0.79**	**0.77**	**0.78**
DT_d	0.66	0.63	0.64	0.71	0.68	0.69	0.68	0.65	0.66	**0.76**	**0.74**	**0.75**
DT_e	0.67	0.64	0.65	0.72	0.69	0.70	0.69	0.66	0.67	**0.77**	**0.75**	**0.76**
CTR_{ds}	0.65	0.63	0.64	0.70	0.68	0.69	0.67	0.65	0.66	0.75	0.73	0.74

$DT_a = \{Character, Canto, Sentence, VP\}$

$DT_b = \{Character, Canto, Sentence, Pol, Senti, VP\}$

$DT_c = \{Character, Sentence, VP\}$

$DT_d = \{Character, Canto, VP\}$

$DT_e = \{Character, Canto, Sentence, Pos, Senti, VP\}$

$CTR_{ds} = \{Character, Canto, Sentence, Pol, Subj, Pos, Neg, Senti, VP\}$

Precision of 0.75, Recall of 0.70, and F1-Score of 0.72. In DT_b, it records Precision of 0.73, Recall of 0.69, and F1-Score of 0.71. For DT_c, DistilBERT shows Precision of 0.74, Recall of 0.71, and F1-Score of 0.73. In DT_d, it scores 0.71 in Precision, 0.68 in Recall, and 0.69 in F1-Score. For DT_e, DistilBERT's metrics are 0.72 for Precision, 0.69 for Recall, and 0.70 for F1-Score. In the CTR_{ds} dataset, it achieves a Precision of 0.70, Recall of 0.68, and F1-Score of 0.69.

RoBERTa in table 5.12 performs well across datasets. For DT_a, it achieves Precision of 0.72, Recall of 0.68, and F1-Score of 0.70. In DT_b, it scores 0.70 in Precision, 0.66 in Recall, and 0.68 in F1-Score. For DT_c, RoBERTa has Precision of 0.71, Recall of 0.69, and F1-Score of 0.70. In DT_d, it records 0.68 for Precision, 0.65 for Recall, and 0.66 for F1-Score. For DT_e, RoBERTa's metrics are 0.69 in Precision, 0.66 in Recall, and 0.67 in F1-Score. In CTR_{ds}, it achieves Precision of 0.67, Recall of 0.65, and F1-Score of 0.66.

The Ensemble approach in table 5.12 achieves the highest performance across all datasets. For DT_a, it scores 0.80 in Precision, 0.78 in Recall, and 0.79 in F1-Score. In DT_b, the metrics are 0.78 for Precision, 0.76 for Recall, and 0.77 for F1-Score. For DT_c, it shows 0.79 in Precision, 0.77 in Recall, and 0.78 in F1-Score. In DT_d, the model achieves 0.76 for Precision, 0.74 for Recall, and 0.75 for F1-Score. For DT_e, it records 0.77 in

Precision, 0.75 in Recall, and 0.76 in F1-Score. The Ensemble approach excels in CTR_{ds} with Precision of 0.75, Recall of 0.73, and F1-Score of 0.74.

Based on the comparison of precision (P), recall (R), and F1-score (F1) across all datasets using BERT, DistilBERT, RoBERTa, and the Ensemble approach, the Ensemble model consistently outperforms the individual models (BERT, DistilBERT, and RoBERTa) across all datasets. The highest performance is observed with DT_a, which includes the features Character, Canto, Sentence, and VP, achieving the best F1-scores across the board, particularly when using the Ensemble approach. DT_a emerges as the most effective dataset, while the Ensemble model is identified as the best performing model, providing superior accuracy and balanced precision and recall for all evaluated datasets.

5.6 Error Analysis

Now, in table 5.13 we demonstrated the percentage of misclassified characters to predict the respective topics of the models.

Table 5.13: Percentage of misclassified topics of the models

	BERT	DistilBERT	RoBERTa	Ensemble
DT_a	35%	30%	32%	22%
DT_b	36%	29%	32%	23%
DT_c	33%	27%	30%	22%
DT_d	37%	32%	35%	25%
DT_e	36%	31%	33%	24%
CTR_{ds}	37%	31%	34%	26%

BERT consistently misclassifies characters across all datasets, with the highest misclassification rate being 37% in the DT_d and CTR_{ds} datasets. This indicates BERT's difficulty in capturing complex relationships between characters and attributes in diverse datasets.

DistilBERT shows better performance than BERT, with misclassification rates between 29% and 32% across the datasets. Its lower misclassification rate highlights its efficiency and better generalization, although it still struggles with complex datasets like CTR_{ds}.

RoBERTa improves over BERT and DistilBERT, with misclassification rates ranging

from 30% to 35%. Despite its robust training and extensive pre-training dataset, RoBERTa still has challenges with complex datasets, as evidenced by a 35% misclassification rate in DT_d.

The Ensemble model significantly reduces the misclassification rate, achieving the lowest percentages from 22% to 26% across all datasets. This approach leverages the strengths of BERT, DistilBERT, and RoBERTa, excelling particularly in complex datasets like CTR_{ds}, reducing the misclassification rate to 26%.

This analysis in table 5.13 indicates that model performance is impacted by dataset complexity. Datasets with more attributes and nuanced relationships, such as CTR_{ds}, tend to have higher misclassification rates, highlighting the need for more sophisticated models or training strategies to handle such complexity.

Now we are discussing about the maximum number of misclassified topics by all the models accross all the datasets. Table 5.14 presents the top five topics with the highest number of misclassifications, highlighting significant discrepancies in the predicted topics. The topic **The Descendants of Ajamidha** is the most misclassified, with a total of 56 occurrences where it was incorrectly predicted as **Krsna, the Supreme Personality of Godhead**, **Lord Parasurama Destroys the Worlds Ruling Class**, or **Parasurama, the Lords Warrior Incarnation**. Following closely is **Diti Vows to Kill King Indra**, which was misclassified 44 times, often being predicted as **The Pastimes of the Supreme Lord, Ramacandra**, **Krsna, the Supreme Personality of Godhead**, or **The Advent of Lord Krsna-Introduction**. The third most misclassified topic is **Krsna, the Supreme Personality of Godhead**, with 42 misclassifications where it was predicted as **Diti Vows to Kill King Indra**, **The Dynasties of the Sons of Yayati**, or **The Descendants of Ajamidha**. The topic **Lord Krsna Shows the Universal Form Within His Mouth** was misclassified 40 times, with predictions including **Diti Vows to Kill King Indra**, **Krsna, the Supreme Personality of Godhead**, or **The Advent of Lord Krsna-Introduction**. Lastly, **The Killing of the Demon Putana** experienced 38 misclassifications, frequently being predicted as **Krsna, the Supreme Personality of Godhead**, **Diti Vows to Kill King Indra**, or **The Descendants of Ajamidha**.

5.7 Discussion and Observations

The paper presents a comprehensive framework for understanding the complex narratives and character-topic relationships in the Srimad-Bhagavatam, a central text in Hindu

Table 5.14: Top 5 maximum misclassified topics

Topic	Sample Predicted Topics	Count
The Descendants of Ajamidha	Krsna, the Supreme Personality of Godhead, Lord Parasurama Destroys the Worlds Ruling Class, Parasurama, the Lords Warrior Incarnation	56
Diti Vows to Kill King Indra	The Pastimes of the Supreme Lord, Ramacandra, Krsna, the Supreme Personality of Godhead, The Advent of Lord Krsna-Introduction	44
Krsna, the Supreme Personality of Godhead	Diti Vows to Kill King Indra, The Dynasties of the Sons of Yayati, The Descendants of Ajamidha	42
Lord Krsna Shows the Universal Form Within His Mouth	Diti Vows to Kill King Indra, Krsna, the Supreme Personality of Godhead, The Advent of Lord Krsna-Introduction	40
The Killing of the Demon Putana	Krsna, the Supreme Personality of Godhead, Diti Vows to Kill King Indra, The Descendants of Ajamidha	38

mythology. The study introduces two core ontologies—`SBC-Ontology` for characters and `SBT-Ontology` for topics—which together aim to provide a structured way to analyze this intricate scripture.

The `SBC-Ontology` categorizes the 1,714 characters in the Srimad-Bhagavatam into 17 groups, including Gods, Demigods, Celestial Beings, Humans, and Animals, with further subdivisions, such as Godhead, Goddess, Kings, Queens, and Sages. It also identifies 24 family relations like father, son, disciple, and master, capturing intricate genealogies. The `SBT-Ontology` organizes the text into 12 Cantos, each addressing specific stories and philosophical teachings. It maps character interactions within these topics, such as Ramacandra featuring in topics like The Pastimes of the Supreme Lord and Lord Ramacandra Rules the World.

The study uses a **Character-Topic Relationship dataset** (CTR_{ds}), which is extracted from sentences in the Srimad-Bhagavatam and includes 45,941 observations. This dataset analyzes the presence of characters across topics, linking them to sentiment scores such as polarity, subjectivity, positivity, and negativity. This sentiment-based analysis provides insights into the emotional tone and dynamics within the text. For example, **Krsna** is linked with high positivity scores, reflecting his association with topics like ***Krsna, the Supreme Personality of Godhead***.

To analyze the **character-topic relationships**, the paper applies various transformer

models such as **BERT**, **DistilBERT**, and **RoBERTa**, with an **ensemble approach** outperforming the individual models. The ensemble model achieved the highest F1-score of 0.79, demonstrating its ability to capture nuanced connections between characters and topics more effectively. Additionally, five distinct datasets were created to test different feature combinations such as character, topic, sentiment, and verb phrases, which were used to evaluate model performance across multiple metrics.

The analysis further revealed **Krsna** as the most frequently mentioned character, and ***Krsna, the Supreme Personality of Godhead*** as the most common topic. A correlation analysis of sentiment attributes (polarity, subjectivity, positivity, and negativity) showed robust relationships, with positivity and negativity being perfectly inversely correlated. This suggests that characters associated with highly positive topics tend to have minimal negative interactions, and vice versa.

5.8 Summary

The development of character and topic ontologies for the Srimad-Bhagavatam presents a structured approach to understanding the complex narratives and philosophical themes within the text. Through the creation of **SBC-Ontology** (Character Ontology) and **SBT-Ontology** (Topic Ontology), the study provides a detailed framework for categorizing the vast array of characters, their relationships, and the topics they engage with across the scripture's 12 Cantos. The integration of advanced computational techniques, such as transformer-based models like **BERT**, **DistilBERT**, and **RoBERTa**, has enabled a more nuanced analysis of the character-topic relationships. The ensemble model, in particular, demonstrated superior performance with an F1-score of 0.79, effectively capturing the intricate links between characters and topics. This work represents a significant contribution to the field of digital humanities, offering a comprehensive dataset and ontology tailored specifically for Hindu mythology, a domain that has historically lacked computational resources.

The current research opens numerous avenues for further exploration. One key area of improvement is the expansion of the dataset to include other significant texts from Hindu mythology, such as ***the Mahabharata*** and ***the Ramayana***, allowing for the development of a more extensive ontology that spans multiple mythological sources. Additionally, integrating the ontologies with global knowledge graphs like **DBpedia** or **Wikidata** could improve the interoperability of the models, making them applicable beyond the domain of Hindu mythology. Another potential direction is enhancing the sentiment analysis

by incorporating more sophisticated sentiment scoring techniques, which could capture deeper emotional and philosophical nuances. The use of fuzzy logic to handle ambiguous or polysemous terms in mythology could also be explored to refine the representation of character relationships. Lastly, the framework could be extended to support interactive digital tools for scholars and educators, facilitating easier exploration and analysis of these texts through a user-friendly interface. This research paves the way for further advancements in ontology development and natural language processing in mythological studies.

Chapter 6

Application Development in Hindu Mythology

MythoBERT 1.0 and Mytho-Annotator

6.1 Introduction

Recent advancements in Natural Language Processing (NLP) have been largely driven by pretrained language models such as `BERT` (Bidirectional Encoder Representations from Transformers) Devlin et al. (2019). These models have revolutionized a variety of NLP tasks, including text classification, topic modeling, translation, and question answering. However, while general-purpose models like `BERT` have proven highly effective in a broad range of applications, their performance tends to falter in domain-specific contexts, particularly when applied to culturally and historically rich areas like mythology. Texts in these domains often feature unique terminology, intricate references, and profound symbolism, which are difficult for general models to interpret accurately. Hindu mythology, with its complex narratives, diverse characters, and philosophical depth, exemplifies this challenge, necessitating the development of more specialized models.

While the trend toward domain-specific NLP models is gaining momentum—with examples like `BioBERT` Lee et al. (2019), `SciBERT` Beltagy et al. (2019) and `LEGAL-BERT` Chalkidis et al. (2020) demonstrating significant improvements in their respective fields—there remains a gap when it comes to applying such models to religious and mythological texts. Current models lack the specialized knowledge to accurately interpret mythological references and cultural nuances, leading to misinterpretation or oversimplification of the content. Hindu mythology, in particular, with its complex symbolism, poses significant challenges for general-purpose models. To date, limited research has been conducted on developing NLP models tailored to such mythological texts, leav-

ing a critical gap in our ability to process and analyze these culturally significant works. To address this gap, we introduce `MythoBERT`, a specialized model tailored specifically for Hindu mythological texts.

Developing an NLP model for Hindu mythological texts presents unique challenges. First, *the richness of the content*, *including diverse characters*, *places*, and *philosophical concepts*, creates a demand for models that can understand and interpret intricate cultural and historical references. The vast and varied nature of these texts, along with the complexity of symbolic and metaphysical ideas, often makes it difficult for general models to capture their full depth. Furthermore, the language and terminology in these texts are often deeply contextual and require a vocabulary that general models do not possess. Without a specialized approach, the meaning of many mythological concepts is either lost or misrepresented.

The primary objective of the `MythoBERT` model is to develop a specialized variant of `BERT` tailored specifically to Hindu mythological texts. This involves creating a custom vocabulary, `MythoVocab`, and embedding model, `Mytho-Embedding`, designed to better capture the terminology and semantic richness of these texts. `MythoBERT` is pretrained on a comprehensive corpus of Hindu mythological works, enabling it to outperform general-purpose models on NLP tasks within this domain. Key objectives include:

- Developing `MythoBERT`, a model fine-tuned for Hindu mythology.
- Evaluating `MythoBERT's` performance on tasks such as topic modeling, text classification, and named entity recognition (NER) specific to mythological content.
- Introducing downstream tasks, including identifying characters in Indian moral stories and English news articles, to evaluate `MythoBERT's` generalization across different narrative domains.
- Comparing `MythoBERT's` performance with that of general-purpose models like `BERT` to demonstrate its superiority in handling culturally nuanced texts and its ability to adapt to diverse domains.

This research makes several novel contributions to the field of domain-specific NLP models:

- `MythoVocab`: A custom vocabulary designed specifically to capture the unique terminology and cultural references found in Hindu mythology. This vocabulary enables the model to better understand and process specialized content.

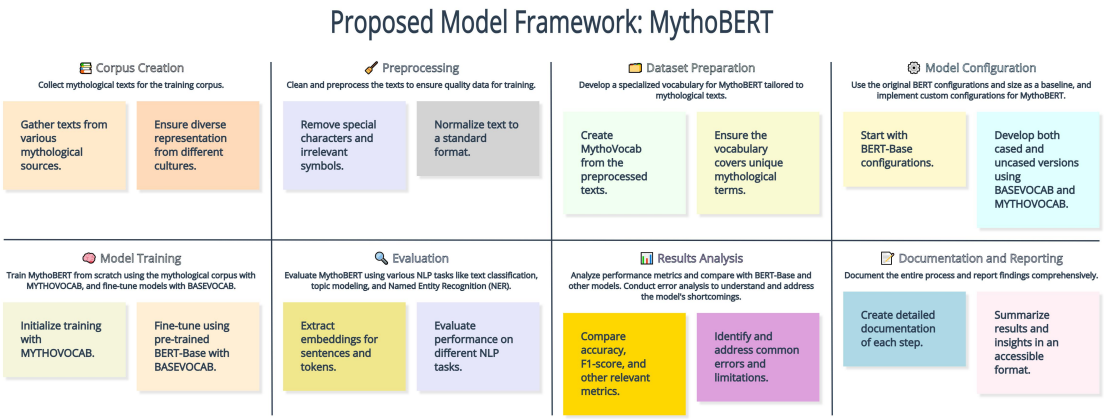


Figure 6.1: MythoBERT development process

- `Mytho-Embedding`: A set of specialized embeddings derived from `MythoVocab`, offering a more accurate representation of the semantic relationships within Hindu mythological texts.

- `MythoBERT`: A pretrained language model built on the foundation of MythoVocab and `Mytho-Embedding`, demonstrating superior performance in tasks such as topic modeling, text classification, and named entity recognition when compared to general-purpose models.

By addressing the limitations of general-purpose models, `MythoBERT` sets a new precedent for NLP research in religious and mythological texts. It not only enhances our ability to process and understand these texts but also opens up new possibilities for applying NLP to culturally significant domains. This research contributes to a growing trend of developing domain-specific language models, providing valuable insights into the application of NLP in the field of mythology. The figure 6.1 illustrates the development process of `MythoBERT`.

The proposed model framework for `MythoBERT` involves several key steps. First, in Corpus Creation, texts are gathered from the `PouranicTopic` dataset Paul et al. (2024), focusing specifically on Hindu mythology. Then, Preprocessing involves cleaning the texts by removing special characters and normalizing them to a standard format. Dataset Preparation follows, where a specialized vocabulary, `MythoVocab`, is created to cover unique mythological terms not typically found in general language models.

The Model Configuration begins by setting up the architecture based on `BERT`-Base, developing both cased and uncased versions with the custom vocabulary. Two types of training tasks were conducted:

- **Fine-tuning BERT:** Both cased and uncased versions of BERT were fine-tuned using `BaseVocab` specifically for mythology. This process adapted the general model to the intricacies of mythological texts.
- **Training MythoBERT from Scratch:** Both cased and uncased versions of `MythoBERT` were trained from scratch using `MythoVocab`. This approach allowed the model to develop a deeper understanding of mythological content and vocabulary from the outset.

Evaluation focuses on extracting embeddings and assessing performance on tasks like named entity recognition (NER), topic modeling, text classification, and the downstream tasks of identifying characters in Indian moral stories and English news articles. This ensures that the model meets the requirements of these specific NLP applications. Afterward, Results Analysis compares metrics such as accuracy and F1-score, identifying errors and limitations. Lastly, Documentation and Reporting involves detailing every step and summarizing results for accessible reporting.

Beyond MythoBERT, the accurate annotation of the mythological texts is crucial for further linguistic and computational analysis. An annotation tool is an application or platform that assists in the process of labeling textual material with various linguistic and semantic properties in the context of Natural Language Processing (NLP) (Mondal et al., 2022). The most widely used annotation tools are designed to be multipurpose. Examples include the web-based BRAT (Stenetorp et al., 2012) and WebAnno (de Castilho et al., 2014) tools, as well as the integrated editors of GATE (Cunningham, 2002). Their extensive nature results in a more complex interface, often making the texts to be annotated less readable. Hindu mythology comprises a multitude of intricate and interwoven narratives, often spanning multiple texts and traditions. To this end, we introduce **Mytho-Annotator**, a specialized tool designed to facilitate the annotation of characters, adjectives, and other narrative elements in Hindu mythology. This tool offers a user-friendly interface that enhances text readability and supports collaboration among annotators. By streamlining the annotation process, **Mytho-Annotator** enables researchers to focus on the intricate details of character interactions and attributes, leading to deeper insights into the text.

The motivation behind **Mytho-Annotator** is to enhance accessibility, accuracy, and collaboration in the annotation and analysis of Hindu mythological texts. All features needed to effectively manage and operate text labeling projects are included in **Mytho-Annotator**. Remote annotation procedures will find it particularly useful due to its

self-descriptive annotation interface, which only requires a web browser. Additionally, it ensures that texts requiring annotation stay clear and readable throughout the entire annotation process. This process operates on a server and may be interrupted at any time. Since all pertinent interactions between the annotators are recorded in a straightforward plain text format with a key-value basis, the annotator's progress may be continuously monitored.

The structure of this research work is organized to provide a comprehensive overview of the development and evaluation of `MythoBERT` and `Mytho-Annotator`. Section 2: Preparation of datasets for `MythoBERT`, Section 3: System Framework covers the development of `MythoVocab`, development of `Mytho-Embedding`, and task specific details of `MythoBERT`. Section 4: Experimental Setup details the tasks, pretrained BERT variants, finetuning procedures, datasets and preprocessing, evaluation metrics, and task-specific details. Section 5: Results Analysis presents visualization characters using `Mytho-Embedding` and findings for the specified tasks, including an overall performance comparison and an analysis of `MythoVocab` and `Mytho-Embedding`. Section 6: Error Analysis explain the errors. Section 7: System Architecture of `Mytho-Annotator`followed by User Interface and Experience. Then Section 9: Discussion and Observations of `MythoBERT` provides an in-depth analysis of the results and their implications. Finally, Section 10: summarizes the study's contributions and suggests directions for future research.

6.2 MythoBERT: Preparation of Dataset

The dataset used to train `MythoBERT` was carefully compiled from a variety of Hindu mythological texts, including the `Ramayana`, `Mahabharata`, `Srimad-Bhagavatam`, `Devi Bhagavata`, `Caitanya Caritamrita`, `Harivamsha Purana`, and `Krsna, the Supreme Personality of Godhead`. For clarity, in the remainder of this work, ***Ramayana*** will be abbreviated as **RAMA**, ***Mahabharata*** as **MBH**, ***Devi Bhagavata*** as **DEVI**, ***Harivamsha Purana*** as **HVM**, ***Srimad-Bhagavatam*** as **SB**, ***Caitanya Caritamrita*** as **CC**, and ***Krsna, the Supreme Personality of Godhead*** as **KRSNA**.

These texts were selected for their cultural importance and their representation of a broad spectrum of mythological themes. On average, each text contains around 70,000 sentences, resulting in a corpus of approximately 2.17 million tokens. The dataset is specifically structured to support the three core NLP tasks: Topic Modeling (TM), Text Classification (TC), and Named Entity Recognition (NER).

6.2.1 Datasets

Text Classification (TC): For text classification (TC), the dataset contains sentences labeled with their corresponding mythological text, canto, and chapter, enabling `MythoBERT` to classify sentences into predefined categories based on content.

Named Entity Recognition (NER): The NER dataset contains sentences annotated with entity tags for characters, locations, and mythological concepts, evaluating `MythoBERT's` ability to recognize and classify named entities in mythological texts.

Topic Modeling (TM): The TM dataset comprises sentences extracted from Hindu mythological texts. As an unsupervised technique, it lacks predefined thematic labels, allowing the model to uncover hidden structures and insights into the thematic elements within the narratives.

Downstream task: The datasets for the downstream tasks include text from the *Panchatantra*, a collection of ancient Indian fables, annotated with character names to enable `MythoBERT` to identify and classify characters within these moral stories. Additionally, the *20 Newsgroups* dataset comprises approximately 20,000 documents across 20 distinct topics, such as politics, sports, and technology. This dataset allows the model to recognize and classify named entities like people, organizations, and locations within diverse discussions. Together, these datasets aim to enhance the model's ability to understand and identify entities in both narrative and contemporary contexts.

6.2.2 Preprocessing

To ensure the datasets are suitable for training and evaluation, we applied the following preprocessing steps:

Tokenization: All text data was tokenized using the WordPiece tokenizer[1] associated with the respective BERT variant (e.g., BASEVOCAB for BERT-Base, MythoVocab for MythoBERT). This step ensures consistency in how the text is broken down into tokens, aligning with the model's vocabulary.

Lowercasing (for uncased models): For uncased versions of BERT and MythoBERT, all text was converted to lowercase. This helps in reducing variability due to case sensitivity.

Sentence Truncation and Padding: To accommodate the varying lengths of sentences, we truncated sentences longer than 128 tokens initially and later extended this to 512 tokens for further training. Shorter sentences were padded with special tokens to

[1] https://huggingface.co/learn/nlp-course/en/chapter6/6

ensure consistent input size across the models.

Entity Annotation for NER: For the NER task, manual annotation was performed to mark characters, locations, and mythological concepts with appropriate tags. We ensured high-quality annotations by conducting multiple rounds of validation with domain experts, ensuring that the entities were accurately labeled.

Preprocessing for the downstream tasks involves cleaning the text data by removing any irrelevant content, such as headers or footers, and normalizing the text format. Additionally, tokenization is applied to prepare the data for named entity recognition, ensuring consistency across different document types.

6.3 MythoBERT: System Framework

6.3.1 Development of MythoVocab

`MythoVocab` is a custom vocabulary designed to handle the unique terminology and references found in Hindu mythological texts.

`MythoVocab` was set to a size of 30,000 tokens, aligning with the size of `BERT's` original vocabulary (`BASEVOCAB`). This size balances the need to capture the full richness of mythological texts while maintaining computational efficiency during training. The structure of `MythoVocab` is tailored to the specific needs of Hindu mythology, with approximately 60% of tokens being unique to the domain, ensuring effective processing of domain-specific content.

An intentional overlap of approximately 40% between `MythoVocab` and `BASEVOCAB` allows the model to retain general-domain knowledge while focusing on the specialized vocabulary required for mythological texts. This overlap ensures that `MythoBERT` can effectively process general language constructs while being attuned to specialized content.

6.3.2 Development of Mytho-Embedding

`Mytho-Embedding` consists of specialized 768-dimensional vectors generated using the 30,000-token `MythoVocab` to capture the semantic complexity of Hindu mythological texts. The steps in its creation are:

Embedding Initialization: Each token in `MythoVocab` is initially mapped to a random 768-dimensional vector, matching the BERT-Base architecture.

Training of Mytho-Embedding: During `MythoBERT` pretraining, these vectors are adjusted based on the context of words and phrases within a corpus of Hindu mythological

texts. This pretraining used a corpus size of approximately 950K sentences, allowing the embeddings to reflect not only the direct meanings but also the symbolic and cultural context of the texts.

As shown in Algorithm 1, the process of generating Mytho-Embeddings involves tokenizing the input corpus using MythoVocab, followed by training with Masked Language Modeling (MLM), and fine-tuning embeddings using the Adam optimizer.

6.3.3 Task-Specific Details

At first we visualized the prominent characters of Mahabharata using t-SNE based on `Mytho-Embedding`. Then we experiment on the following core NLP tasks:

- Masked Language Modeling (MLM)
- Text Classification (TC)
- Named Entity Recognition (NER)
- Topic Modelling (TM)

Using `MythoBERT` for Topic Modeling, Text Classification and Named Entity Recognition enables precise identification and categorization of themes, characters, and relationships within Hindu mythological texts, enhancing the analysis and understanding of these rich narratives.

Masked Language Modeling (MLM): The MLM task involves randomly masking a percentage of tokens in the input sentences and training `MythoBERT` to predict these masked tokens. This self-supervised learning approach enhances the model's understanding of context and word relationships, improving its performance on downstream tasks. Challenges such as ensuring diverse and meaningful masking strategies are addressed by carefully tuning the masking rate and leveraging context from surrounding tokens to maintain semantic coherence.

Text Classification (TC): For Text Classification, the final [CLS] token representation from BERT is fed into a linear classification layer to predict class probabilities. We use the cross-entropy loss function to manage multi-class classification, minimizing the discrepancy between predicted and actual labels. Challenges such as class imbalance are addressed through weighted loss functions, data augmentation, and careful tuning of learning rates and batch sizes to improve model performance across all classes.

Named Entity Recognition (NER): For Named Entity Recognition, `MythoBERT` is used with a token classification head that includes a dense layer and softmax activation

Algorithm 1 Mytho-Embedding Generation for MythoBERT

Input: Tokenized sentences from mythological corpus C (950K sentences), Mytho-Vocab V_m (30,000 tokens), Random initialization of embeddings E_m for each token in V_m, Training epochs T, batch size B, learning rate η

Output: Fine-tuned Mytho-Embeddings E_m^{final} (768-dimensional vectors)

Tokenization and Initialization

Tokenize input corpus C using the WordPiece tokenizer with custom MythoVocab V_m.

$C_{\text{tokenized}} \leftarrow \text{Tokenize}(C, V_m)$

Initialize random embeddings for each token in V_m.

$E_m \leftarrow \text{RandomInitialize}(V_m, \text{dim} = 768)$

Pretraining the Model

for each epoch $t \in [1, T]$ **do**

    for each batch $b \in B$ **do**

        Select a batch of tokenized sentences S_b from $C_{\text{tokenized}}$.

        for each sentence $s \in S_b$ **do**

            Pass s through the BERT architecture with randomly initialized embeddings E_m.

            Perform Masked Language Modeling (MLM): Mask 15% of the tokens in the input sentences.

            Compute the loss function L (MLM)

            $L \leftarrow \text{Loss}(s, E_m)$

            Update embeddings E_m using Adam optimizer with learning rate η:

            $E_m \leftarrow \text{AdamOptimizer}(E_m, L, \eta)$

Optimization of Embeddings

Fine-tune the embeddings E_m over multiple epochs T, adjusting the vectors based on token context within sentences.

Output of Fine-Tuned Embeddings

After completion of training, output the fine-tuned Mytho-Embeddings E_m^{final}, now optimized for capturing the semantic and cultural richness of Hindu mythological texts.

to predict entity labels for each token. The cross-entropy loss function is employed at the token level to ensure accurate alignment between predicted and actual entity labels. Challenges in recognizing complex and varied mythological names are addressed by training on a diverse dataset and evaluating performance with entity-level precision, recall, and F1-score.

Topic Modeling (TM): For Topic Modeling, a topic extraction layer is added to `MythoBERT`, featuring a dense layer and softmax activation to predict topic distributions. The cross-entropy loss function is used to compare predicted topic probabilities with actual distributions. Challenges like capturing nuanced themes in mythological texts are addressed by training on diverse sentences and using coherence scores to refine topic interpretability.

Evaluation Metrics: To evaluate `MythoBERT's` performance across Topic Modeling (TM), Text Classification (TC), Named Entity Recognition (NER), and Masked Language Modeling (MLM), we employ task-specific metrics. For TC and NER, we use Accuracy, Precision, Recall, and F1-Score to address imbalanced datasets. For MLM, we assess Accuracy, F1-Score, Cross-Entropy Loss, and Perplexity. TM is evaluated using Coherence Score and Perplexity to measure interpretability and generalization. These metrics ensure a robust evaluation of `MythoBERT's` capabilities across tasks.

6.4 MythoBERT: Experimental Setup

This section outlines the key elements of the setup, including hardware specifications, training duration, and details of the pretrained BERT variants used.

Hardware and Training Duration: Training `MythoBERT` required high-performance resources, utilizing a cluster of 8 NVIDIA V100 GPUs. Training from scratch on the `MythoVocab` corpus took about 3 weeks, while fine-tuning with **BASEVOCAB** took around 3 days, highlighting the hardware's efficiency in managing large-scale computations.

Pretrained BERT Variants: Two versions of `MythoBERT` were developed: one with `BASEVOCAB` and another with the custom `MythoVocab`, each having cased and uncased configurations. The BASEVOCAB models were fine-tuned from BERT-Base, while MythoVocab models were trained from scratch to capture domain-specific terminology.

Fine-Tuning and Hyperparameters: Fine-tuning occurred over 2 to 5 epochs with a batch size of 32 and learning rates between 5e-6 and 5e-5, typically achieving best results at 2 or 4 epochs with a learning rate of 2e-5. The models were optimized using the Adam

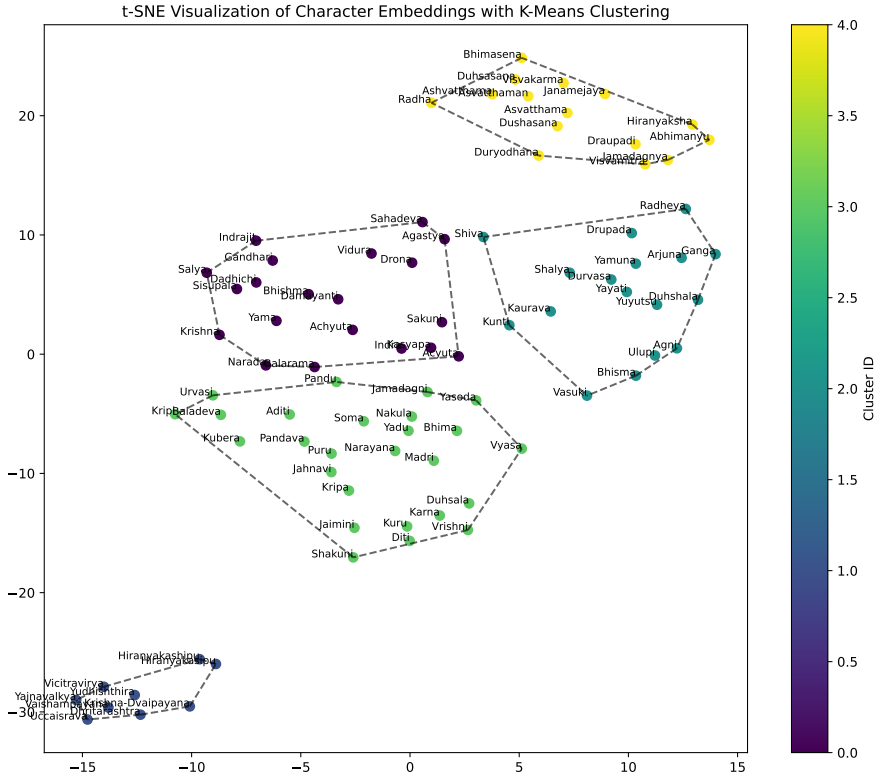


Figure 6.2: t-SNE Visualization of Character Embeddings with K-Means Clustering

optimizer, ensuring robust performance across tasks.

This combination of custom vocabulary and fine-tuning allowed `MythoBERT` to excel in domain-specific tasks, emphasizing the need for tailored models in specific linguistic and cultural contexts.

6.5 MythoBERT: Results Analysis

6.5.1 Visualizing Mythological Characters

Visualizing characters using `Mytho-Embeddings` with **t-SNE** and **KMeans** clustering provides insightful patterns in the relationships between mythological figures. It helps uncover thematic groupings and latent connections that are not immediately apparent, aiding in a deeper understanding of character dynamics. The combination of **t-SNE's** dimensionality reduction and **KMeans** clustering enables clear identification of clusters, revealing how characters are related based on semantic proximity. This method enhances exploration of large mythological texts by offering an intuitive visual summary of complex

relationships.

The figure 6.2 presents a t-SNE visualization of character embeddings from the Mahabharata, plotted in 2D space with K-Means clustering. Different colors represent clusters, revealing thematic and relational connections among characters.

Cluster 0 includes central figures like *Krishna*, *Bhishma*, and *Drona*, while Cluster 1 features significant roles such as *Dhritarashtra* and *Vyasa*. Cluster 2 mixes divine entities and heroes like *Agni* and *Arjuna*, reflecting the intersection of divine and heroic narratives. Cluster 3 encompasses a range of influential characters, including gods and key warriors, and Cluster 4 focuses on crucial figures from both the *Kaurava* and *Pandava* sides, such as *Abhimanyu* and *Duryodhana*. Overall, these clusters provide insight into the relationships and themes in the **Mahabharata**.

6.5.2 Masked Language Modeling (MLM)

Model Performance: The performance of `MythoBERT` on the Masked Language Modeling task was evaluated using various metrics to assess its predictive capabilities. The results indicate the model's effectiveness in understanding contextual relationships and token dynamics within mythological texts.

Table 6.1: MLM Performance Metrics for MythoBERT

Model	Accuracy	F1-Score	Cross-Entropy Loss	Perplexity
BERT-Base (Cased)	75.0%	0.60	0.70	3.50
BERT-Base (Uncased)	74.5%	0.58	0.72	3.60
MythoBERT (Cased)	83.0%	0.76	0.48	2.20
MythoBERT (Uncased)	82.5%	0.75	0.50	2.30

Analysis: The table 6.1 summarizes the performance metrics for Masked Language Modeling (MLM) across various models. BERT-Base models (both cased and uncased) show lower accuracy and F1-scores compared to `MythoBERT`, indicating that `MythoBERT` outperforms the general-purpose BERT models in predicting masked tokens. The cross-entropy loss and perplexity values also reflect `MythoBERT's` superior predictive capabilities, showcasing its effectiveness in handling domain-specific tasks.

6.5.3 Text Classification (TC)

Model Performance: For Text Classification, we evaluated the performance of MythoBERT and compared it with the baseline BERT-Base models. The results are summar-

ized in Table 6.2.

Table 6.2: Text Classification Performance

Model	Accuracy	Precision	Recall	F1-Score
BERT-Base (Cased)	85.3%	84.7%	86.1%	85.4%
BERT-Base (Uncased)	84.8%	84.2%	85.6%	84.9%
MythoBERT (Cased)	89.2%	88.7%	89.8%	89.2%
MythoBERT (Uncased)	88.5%	88.0%	89.0%	88.5%

Analysis: `MythoBERT` models significantly outperform the BERT-Base models in all evaluation metrics. The cased version of `MythoBERT` achieved the highest accuracy of 89.2%, which indicates a better classification performance for sentences from Hindu mythological texts. The improvement in F1-Score and Precision demonstrates `MythoBERT's` superior capability to handle nuanced class distinctions in this specialized domain.

6.5.4 Named Entity Recognition (NER)

Model Performance: The Named Entity Recognition results are presented in Table 6.3.

Table 6.3: Named Entity Recognition Performance

Model	Precision	Recall	F1-Score
BERT-Base (Cased)	78.5%	76.2%	77.3%
BERT-Base (Uncased)	77.9%	75.8%	76.8%
MythoBERT (Cased)	83.7%	82.5%	83.1%
MythoBERT (Uncased)	82.3%	81.0%	81.7%

Analysis: `MythoBERT` outperforms the BERT-Base models in identifying and classifying named entities within mythological texts. The cased version of `MythoBERT` achieved an F1-Score of 83.1%, indicating a substantial improvement in both Precision and Recall. This suggests that `MythoBERT` is better at recognizing and categorizing complex mythological entities.

6.5.5 Topic Modeling (TM)

Model Performance: For Topic Modeling, we evaluated the coherence and perplexity of topics extracted using `MythoBERT` and BERT-Base models. The results are shown in Table 6.4.

Table 6.4: Topic Modeling Performance

Model	Coherence Score	Perplexity
BERT-Base (Cased)	0.62	210.4
BERT-Base (Uncased)	0.60	215.3
MythoBERT (Cased)	0.72	175.6
MythoBERT (Uncased)	0.70	180.2

Analysis: `MythoBERT` models demonstrate superior performance in Topic Modeling compared to BERT-Base models. The cased version of `MythoBERT` achieves the highest Coherence Score of 0.72, reflecting better interpretability of the extracted topics. The lower Perplexity indicates a better generalization of topics and a more effective representation of thematic elements within the mythological texts.

6.5.6 Downstream Tasks

Model Performance: The performance of MythoBERT on the downstream tasks of Story Named Entity Recognition (NER) and Newspaper NER is detailed in the following tables 6.5 and 6.6.

Table 6.5: Story Named Entity Recognition Performance

Model	Precision	Recall	F1-Score
BERT-Base (Cased)	68.5%	71.2%	69.8%
BERT-Base (Uncased)	68.0%	70.5%	69.2%
MythoBERT (Cased)	81.0%	83.8%	82.4%
MythoBERT (Uncased)	80.5%	82.0%	81.2%

Table 6.6: Newspaper Named Entity Recognition Performance

Model	Precision	Recall	F1-Score
BERT-Base (Cased)	70.0%	73.5%	71.7%
BERT-Base (Uncased)	69.8%	72.0%	70.9%
MythoBERT (Cased)	84.5%	85.8%	85.1%
MythoBERT (Uncased)	84.0%	85.5%	84.8%

Analysis: The tables 6.5 and 6.6 demonstrates that `MythoBERT` significantly outperforms BERT-Base models in both Story NER and Newspaper NER tasks. The enhanced

precision and F1-scores indicate `MythoBERT's` effectiveness in recognizing named entities in varied contexts, underscoring the benefits of domain-specific training for these applications.

6.6 MythoBERT: Error Analysis

Error analysis is crucial for understanding the limitations of `MythoBERT` and guiding future improvements. In this section, we present detailed observations on the errors encountered during the experiments, providing quantitative insights into their impact.

Text Classification Errors- The Table 6.7 explains the errors rates of different models.BERT-Base (Cased) has an error rate of 14.7%, with notable misclassifications including *Sukadeva Gosvami's final instructions* being predicted as *a glorification of Lord Siva and Uma*. Similarly, *Dronabhisheka Parva* was misclassified as *Jayadratha-Vadha Parva*. The `BERT-Base (Uncased)` model shows a slightly higher error rate of 15.2%, misidentifying *the activities of Maharaja Agnidhra* as those of*Jada Bharata* and misclassifying *Markandeya-Samasya Parva* as *Tirtha-yatra Parva*. In contrast, `MythoBERT (Cased)` achieved a lower error rate of 10.8%, with misclassifications such as *The Movements of the Sun* being predicted as *The Orbits of the Planets*. Finally, `MythoBERT (Uncased)` has an error rate of 11.5%, misidentifying *The Glories of Lord Ananta* as *The Activities of Maharaja Priyavrata*. Overall, `MythoBERT` models demonstrate improved performance in accurately classifying labels compared to the `BERT-Base` models.

Named Entity Recognition (NER) Errors- The table 6.8 presents a summary of named entities misclassified by each model across various mythological texts. The `BERT-Base (Cased)` model incorrectly identified *Maharaja Pariksit* as *King Janaka* and *Sita* as *Urmila*. The Uncased version confused *King Dhruva* with *King Harishchandra* and misclassified *Rama* as *Lakshmana*. In `MythoBERT (Cased)`, *Maharaja Agnidhra* was misidentified as *Maharaja Yudhishthira*, and *Hanuman* as *Sugriva*. The Uncased variant made similar errors, misclassifying *King Dhruva* as King *Priyavrata* and *Sita* again as *Urmila*. This illustrates the challenges each model faces in accurately recognizing named entities in mythological texts.

Topic Modeling Errors- The **Topic Coherence Error Rate (TCER)** is a metric used to measure how well a model's predicted topics align with actual meaningful themes in a corpus. It's particularly useful in tasks like Topic Modeling, where the goal is to extract coherent and distinct topics from a large collection of texts. This metric, as explained in Equation 6.1, is important for ensuring that the topics generated by the model aren't

Table 6.7: Percentage of misclassified labels by the models

Model	Error Rate	True label	Predicted label
BERT-Base (Cased)	14.7%	Sukadeva Gosvami's Final Instructions to Maharaja Pariksit Dronabhisheka Parva Shri Rama is given the celestial weapons	Lord Siva and Uma Glorify Markandeya Rsi Jayadratha-Vadha Parva The story of the king of Himalayas' younger daughter Uma
BERT-Base (Uncased)	15.2%	The Activities of Maharaja Agnidhra Markandeya-Samasya Parva Ravana's Threats	Jada Bharata Instructs King Rahugana Tirtha-yatra Parva Hanuman's astonishment on beholding Ravana
MythoBERT (Cased)	10.8%	The Movements of the Sun Tirtha-yatra Parva Hanuman allows himself to be taken captive by the Titans	The Orbits of the Planets Arjunabhigamana Parva Hanuman's Dilemma
MythoBERT (Uncased)	11.5%	The Glories of Lord Ananta Kichaka-badha Parva Hanuman's Reflections on seeing Sita	The Activities of Maharaja Priyavrata Pandava-Pravesa Parva Sita sees Hanuman

just collections of unrelated terms but represent consistent themes or narratives—crucial in applications like analyzing mythological texts

$$\text{TCER} = 1 - \frac{1}{N}\sum_{i=1}^{N}\text{Coherence}(T_i) \tag{6.1}$$

Where:

- N is the number of topics.
- T_i represents the i-th topic.
- Coherence(T_i) is a measure of the coherence of the i-th topic.

The table 6.9 provides the **Topic Coherence Error Rate (TCER)** for different models, showing the percentage of error each model encounters in achieving the optimal topic coherence. Models with higher coherence scores, such as `MythoBERT` (Cased) and `MythoBERT` (Uncased), have lower TCERs, indicating better topic modeling performance compared to the BERT-Base models. The TCER is calculated as the deviation from

Table 6.8: Named Entities Misclassified by Each Model

Model	True Entity	Predicted Entity
BERT-Base (Cased)	Maharaja Pariksit	King Janaka
	Dronacharya	Krishna
	Sita	Urmila
BERT-Base (Uncased)	King Dhruva	King Harishchandra
	Bhishma	Drona
	Rama	Lakshmana
MythoBERT (Cased)	Maharaja Agnidhra	Maharaja Yudhishthira
	Arjuna	Bhima
	Hanuman	Sugriva
MythoBERT (Uncased)	King Dhruva	King Priyavrata
	Duryodhana	Yudhishthira
	Sita	Urmila

Table 6.9: Topic Coherence Error Rate (TCER) for Different Models

Model	**Coherence Score**	**TCER (%)**
BERT-Base (Cased)	0.62	38%
BERT-Base (Uncased)	0.60	40%
MythoBERT (Cased)	0.72	28%
MythoBERT (Uncased)	0.70	30%

an ideal coherence score of 1, highlighting `MythoBERT's` superior capability in capturing coherent topics from mythological texts.

To analyze coherence scores, we examined the top 5 frequent topics from the **SB** dataset, as summarized in Table 6.10. Topic coherence scores for topics A, B, C, D, and E across four models reveal that Topic C consistently has the highest scores, indicating strong definition. `MythoBERT` models outperform BERT-Base, with Topic A scoring 0.60 for BERT-Base (Cased) and 0.63 for `MythoBERT`(Uncased). Topic E achieves a score of 0.68 with `MythoBERT` (Cased), further highlighting its coherence. Overall, `MythoBERT` demonstrates superior coherence, affirming its effectiveness in topic modeling for Hindu mythological texts.

The downstream task of Named Entity Recognition (NER) based on two datasets—mythological stories and newspapers—demonstrates that `MythoBERT` outperforms

Table 6.10: Topic Coherence Score Comparison for SB

Topic	BERT-Base (Cased)	BERT-Base (Uncased)	MythoBERT (Cased)	MythoBERT (Uncased)
Topic A	0.60	0.58	0.65	0.63
Topic B	0.55	0.52	0.60	0.59
Topic C	0.62	0.61	0.67	0.64
Topic D	0.57	0.55	0.62	0.60
Topic E	0.64	0.63	0.68	0.65

BERT-Base models in both domains. `MythoBERT` outperforms BERT-Base in both the Story and Newspaper NER tasks. In the story dataset, `MythoBERT` (Cased) achieves an F1-score of 82.4%, compared to 69.8% for BERT-Base (Cased). Similarly, in the newspaper dataset, `MythoBERT` (Cased) scores 85.1% vs. BERT-Base's 71.7%. This highlights `MythoBERT's` superior precision and recall in both mythological and general text contexts.

6.7 Mytho-Annotator: System Description & User Interface

Mytho-Annotator is a collaborative tool developed using the modern full-stack framework MERN Stack[2]. Our tool consists of three components: i) client side, ii) MongoDB database, and iii) server side.

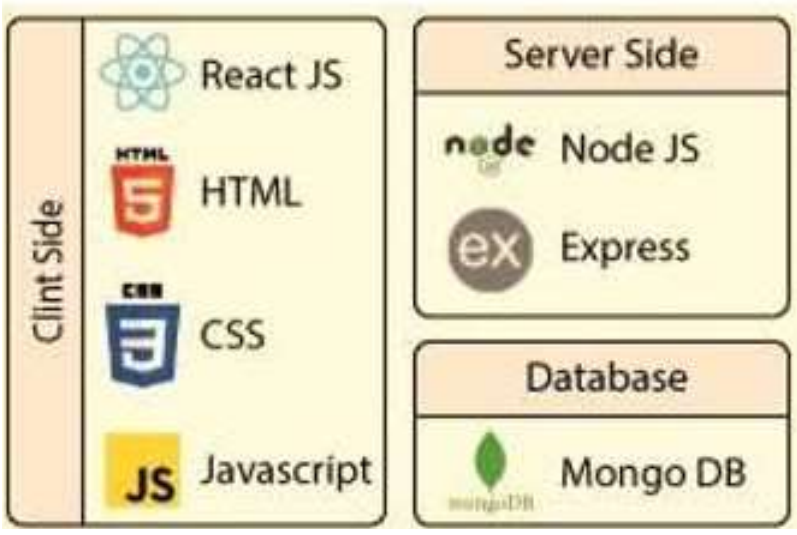


Figure 6.3: Mytho-Annotator's system architecture and technology stack

[2]https://www.mongodb.com/mern-stack

The MongoDB database of **Mytho-Annotator** consists of two collections: **Projects** and **Users**. The **Projects** collection contains information pertinent to various projects, including fields such as `userId`, `projectName`, `namedEntities`, `eventEntities`, `relations`, `namedEntityTags`, `eventEntityTags`, `namedEntityAppearances`, `eventEntityAppearances`.

The **Users** collection stores information about the users, including `username`, `hashed and salted password`, and `email`.

6.7.1 Text Document Handling

Mytho-Annotator's prowess in text document handling goes beyond mere compatibility. It excels in seamlessly processing a wide array of textual data formats, accommodating the diverse sources of text that researchers and annotators encounter. Whether it's standard text documents, PDFs, web content, or more specialized formats, **Mytho-Annotator** ensures flexibility and adaptability which significantly simplifies the intricate process of importing and managing documents for annotation tasks.

6.7.2 Annotation Sections

Mytho-Annotator's annotation framework is thoughtfully structured into three dedicated sections, each tailored to serve a specific annotation purpose.

Named Entity Annotation- Within the realm of named entity recognition, **Mytho-Annotator** stands out by offering annotators a comprehensive set of predefined categories. These categories encompass a wide spectrum of common named entities, including individuals, organizations, geographic locations, and much more. Annotators benefit from a user-friendly interface that streamlines the selection and annotation of named entities throughout the text.

Relationship Annotation- Mytho-Annotator's Relationship Annotation section is a robust tool designed for capturing intricate connections and associations between entities present in the text. This section features two essential components that work in harmony to facilitate thorough relationship annotation. First, annotators can seamlessly select relevant phrases within the text, pinpointing specific sections that hold significance for relationship identification. This phrase selection capability serves as the foundation for building meaningful relationships. Second, annotators can assign and define relationships between the identified entities, fostering a comprehensive understanding of how different elements within the text interact.

Event Entity Recognition- Mytho-Annotator extends its annotation capabilities to encompass event entity recognition, a crucial aspect of text understanding. Within the Event Entity Recognition section, annotators have the power to assign events to phrases while considering essential attributes such as modality, frequency, and time. This feature empowers users to capture nuanced information related to events described in the text. Whether it's identifying the likelihood of an event, its occurrence frequency, or temporal context, **Mytho-Annotator** provides a versatile framework for annotating these critical details.

6.7.3 User Interface and Experience

In our annotation tool, we have thoughtfully structured it into three distinct sections, each aimed at providing users with a comprehensive and user-friendly annotation experience. We understand the importance of efficiency and precision in annotating mythological texts, and our tool has been designed with these considerations in mind. The `Named Entity Annotation`, `Relationship Annotation`, and `Event Entity Annotation` sections are the cornerstones of our tool's functionality.

Within the `Named Entity Annotation` section, we have included a sub-bar that offers two crucial functions: **Assign Tag** and **Assign Gender**. With **Assign Tag**, users have the flexibility to apply a diverse range of tags to specific words or phrases within the text. This feature ensures that annotations are not only accurate but also enriched with contextual information, making it a valuable resource for mythological analysis.

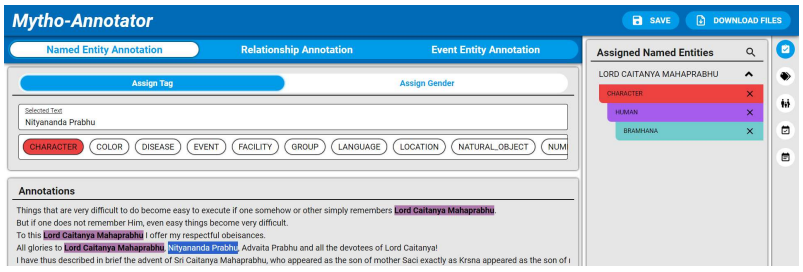


Figure 6.4: Interface portraying Assign Tag section

In the visual representation, Figure 6.4 of `Named Entity Annotation` depicts the **Assign Tags** feature, where a diverse set of tags, including *Character*, *Color*, *Disease*, *Event*, *Facility*, and others, are observed. Specifically, ***Lord Caitanya Mahaprabhu*** is identified and annotated as a character. The illustration further features a sidebar, introducing a hierarchical structure that visually organizes the annotated entities. Notably, under the category *Character*, a hierarchical structure is established, with a broader classifica-

tion denoted as *Human*. Within this hierarchy, a more specific subclass is identified as *Brahmana*, and this annotation will be reflected in the whole text.

Furthermore, the **Assign Gender** function under this section is specifically designed to address the nuanced gender dynamics often present in mythological narratives.

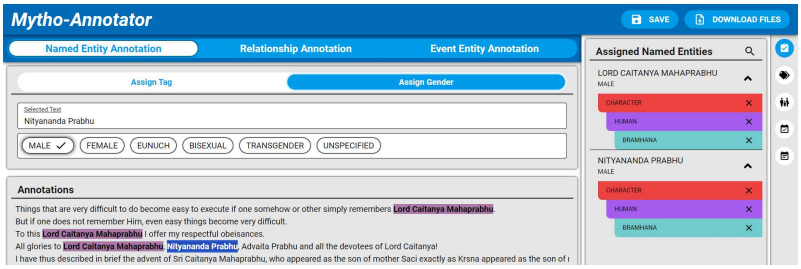


Figure 6.5: Interface portraying Assign Gender section

In Figure 6.5, the `Named Entity Annotation` module includes a distinctive segment titled **Assign Gender**. Within this section, a range of gender categories is presented, encompassing *Male*, *Female*, *Eunuch*, *Bisexual*, *Transgender*, and others.

In the sidebar, we can observe that both ***Lord Caitanya Mahaprabhu*** and ***Nityananda Prabhu*** selected from the text have been designated with the gender label *Male*, shown in the right sidebar. This module facilitates the assignment of specific gender attributes to identified characters within the analyzed text. Such a feature proves valuable in the context of character analysis, allowing for the nuanced classification of individuals based on their gender identity.

Now, in the second section, which is the `Relationship Annotation` section, we have incorporated two separate yet seamlessly integrated components. The first component allows users to select phrases within the text, providing a foundation for identifying key elements for relationship annotation. This phrase selection process ensures that users can pinpoint specific sections of text that are relevant to the relationships they intend to annotate.

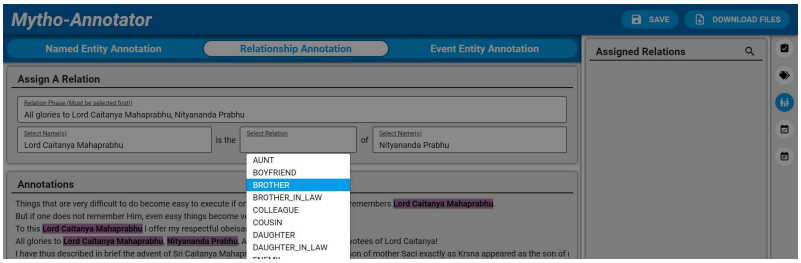


Figure 6.6: Interface portraying Relationship Annotation

In Figure 6.6, the illustration showcases `Relationship Annotation`. The accompa-

nying sentence provides contextual information, revealing that ***Lord Caitanya Mahaprabhu*** is associated with the relationship label *brother* in relation to ***Nityananda Prabhu***. This feature elucidates the capability of the system to discern and annotate intricate relationships between characters, enriching the semantic understanding of their affiliations within the given textual context.

Within the `Event Entity Annotation` section, our tool offers a versatile platform for annotating events and associating them with specific text phrases. This feature allows users to capture information related to events, including their modality, frequency, and temporal attributes. Whether it's identifying the certainty of an event (Modality), specifying how often it occurs (Frequency), or noting the time it takes place (Time), our annotation tool empowers users to comprehensively capture event-related information within the text.

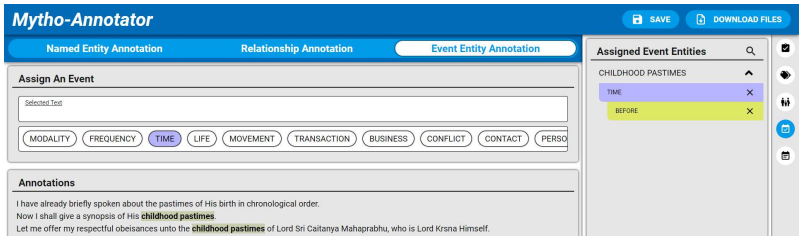


Figure 6.7: Interface portraying Event Entity Annotation

In Figure 6.7, the depiction centers on `Event Entity Annotation`, featuring various entities such as modality, frequency, time, and others. Notably, when the specific text segment ***childhood pastimes*** is selected, the associated entity assigned is *time*. Furthermore, a contextual insight is provided through the accompanying right-hand bar, where the event is characterized as representing the past, and the temporal relationship is specified as *before*. This nuanced annotation of temporal attributes within the Event Entity adds a layer of sophistication, allowing for a more detailed analysis of the chronological aspects associated with events in the annotated text.

To enhance user experience and provide a visual overview of the annotation process, we have incorporated a dedicated section where we can see the results as well as the hierarchical structure.

Thus, this visual representation, Figure 6.8, serves as a dynamic dashboard, representing the hierarchical relationships of Named Entities and that of Event Entities to validate their annotations and refine their work as needed. Overall, our annotation tool is designed to cater to the unique demands of annotating mythological texts. Whether users are identifying named entities, capturing intricate relationships, or recognizing signific-

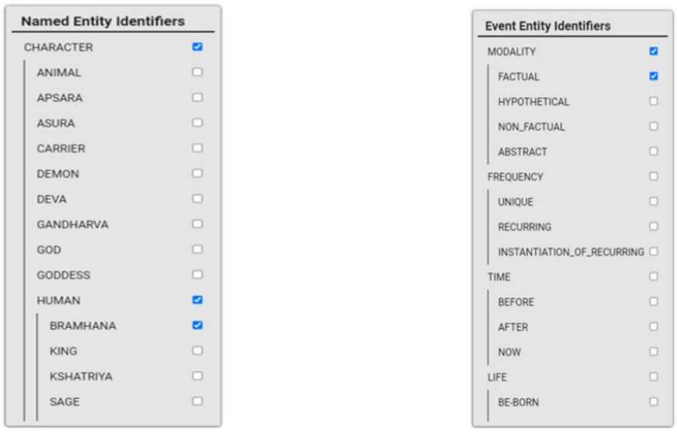


Figure 6.8: Name and Event Entities

ant event entities, our tool serves as a reliable and user-centric companion, facilitating a seamless and productive annotation journey.

6.8 MythoBERT: Discussion and Observations

The `MythoBERT` model demonstrated clear advantages over general-purpose models such as BERT-Base across several core NLP tasks, specifically *text classification*, *named entity recognition (NER)*, and *topic modeling*. These improvements can be attributed to `MythoBERT's` specialized vocabulary (*MythoVocab*) and embeddings (*Mytho-Embedding*) tailored to the complex terminology and cultural nuances of Hindu mythology.

In **text classification**, `MythoBERT` achieved superior performance, with the cased version reaching an accuracy of 89.2%, significantly outperforming BERT-Base (85.3%). This demonstrates the model's enhanced ability to classify sentences into mythological categories based on content.

For **NER**, `MythoBERT` showed an F1-score of 83.1% (cased version), surpassing BERT-Base's performance (77.3%), reflecting the model's strength in recognizing mythological entities, which are often complex and varied.

Topic modeling further highlighted `MythoBERT's` proficiency, as it achieved a higher *coherence score* (0.72 for the cased version) and lower *perplexity* compared to BERT-Base, indicating better topic extraction and interpretability within the mythological texts.

In the **downstream tasks** of named entity recognition, `MythoBERT` continued to excel. In the *story NER* task, `MythoBERT` achieved an F1-score of 82.4%, outperforming BERT-Base (69.8%). This shows its ability to generalize well to traditional Indian stories beyond the specific mythological domain it was trained on. In the *newspaper NER*

task, `MythoBERT` scored 85.1%, again outperforming BERT-Base (71.7%), demonstrating robustness in handling contemporary text data despite its mythological specialization.

Error analysis revealed fewer misclassifications in `MythoBERT`compared to BERT-Base, demonstrating its refined capacity for understanding domain-specific content. While misclassifications were present, they were less frequent and severe, indicating MythoBERT's deeper understanding of mythological characters and topics.

Overall, `MythoBERT's` specialized training allowed it to outperform general-purpose models in all tasks, reflecting the importance of domain-specific adaptations in NLP models, especially for texts with rich cultural and symbolic content.

6.9 Summary

In summary, `MythoBERT` has proven to be an effective domain-specific language model, outperforming general-purpose models like BERT-Base in tasks such as text classification, named entity recognition (NER), and topic modeling, particularly in the context of Hindu mythology. The use of a custom vocabulary and specialized embeddings allowed `MythoBERT` to capture the cultural and symbolic richness of the texts, demonstrating the value of domain-specific adaptations in NLP.

For future work, expanding the training data to include a broader range of mythological and cultural texts could enhance `MythoBERT's` generalization capabilities. Further exploration of additional NLP tasks such as sentiment analysis and relation extraction within these texts is also a promising direction. Additionally, cross-lingual extensions and integration with knowledge graphs could broaden its applicability, making `MythoBERT` a versatile tool for both academic research and real-world applications.

Finally, the introduction of **Mytho-Annotator** represents a significant advance in the annotation of mythological texts. As a specialized annotation tool, it was developed to address the specific challenges posed by the unique narrative structures of Indian mythology. **Mytho-Annotator** is equipped with a user-friendly interface and supports remote collaboration, making it a powerful resource for annotators and researchers alike. Looking forward, the tool's future development promises exciting possibilities, such as expanding entity recognition capabilities to support more complex relationships between characters, improving interoperability with other annotation platforms, and integrating machine learning models for semi-automated annotations.

Chapter 7

Conclusion

Final remarks

This thesis has made significant strides in advancing the computational analysis of Hindu mythological texts through a multi-dimensional approach that includes character identification, topic classification, ontology development, and character-topic relationship analysis. Each of these aspects plays a critical role in deepening our understanding of these ancient narratives and paves the way for future research in the field of natural language processing (NLP) applied to mythological and cultural studies.

The first notable contribution is the development of methodologies for **Character Identification** and **Character Adjective Classification**. Characters in mythological texts like the Mahabharata are often introduced and described through a combination of proper nouns and vivid descriptive adjectives. We addressed this challenge by employing machine learning classifiers—such as KNN, Logistic Regression, and MLP models—to identify both characters and their associated adjectives. By incorporating feature subset selection and leveraging phrase-level rules, we improved the accuracy and robustness of the classification models. The success of these methods highlights the importance of carefully engineered features in processing culturally rich content, and the results can be extended to larger mythological corpora for more structured textual analysis.

Following this, the **PouranicTopic dataset** was introduced to tackle the task of **topic classification** within Hindu mythological texts. The dataset, compiled from seven major Hindu texts, served as a key resource for training and evaluating BERT-based and ensemble models. Through the use of various sentence clustering techniques and log-likelihood-based analysis, we were able to capture thematic structures and classify cantos effectively. While ensemble models showed promising results, we also identified areas where topic classification models could be improved, particularly in handling the nuances of ancient narrative structures. Future research in this area could focus on expanding the dataset to multilingual texts, improving sentence annotations, and incorporating more

sophisticated transformers like T5 and ALBERT to enhance model performance.

One of the thesis's major accomplishments is the development of **ontologies** for mythological texts. The construction of **SBC-Ontology** (Character Ontology) and **SBT-Ontology** (Topic Ontology) for the Srimad-Bhagavatam marked a significant step in structuring the complex relationships and themes within Hindu mythology. These ontologies provide a formal framework for organizing and categorizing characters and topics, allowing for a more detailed computational analysis of these ancient texts. The use of transformer models like BERT and DistilBERT facilitated the extraction of relationships between various entities and topics, and our ensemble models demonstrated a high degree of accuracy in identifying and structuring these relationships. The ontologies built in this thesis can be further expanded to include other significant texts like the Mahabharata and Ramayana, and integrated with global knowledge graphs such as **DBpedia** and **Wikidata**. This will enhance interoperability and allow for cross-domain research in the digital humanities.

In addition to ontology development, a separate focus was placed on analyzing **Character-Topic Relationships** within Hindu mythological texts. By examining how characters engage with different themes and topics across the Srimad-Bhagavatam's 12 Cantos, we developed a framework that reveals the intricate interplay between characters and the philosophical themes they embody. This task required a careful balance between symbolic and thematic analysis, facilitated by transformer models like RoBERTa and BERT. The high F1-scores achieved in capturing these relationships underscore the value of using computational models to unravel complex narrative structures. Future research in this area could incorporate fuzzy logic to handle ambiguous or polysemous terms in mythology, as well as develop sentiment analysis models that delve deeper into the emotional and philosophical nuances present in these texts.

The development of **MythoBERT**, a domain-specific language model, represents another pivotal contribution of this thesis. Hindu mythology, with its unique cultural and linguistic context, poses challenges for general-purpose models like BERT. MythoBERT addresses these challenges by incorporating a custom vocabulary and specialized embeddings designed specifically for mythological texts. The model consistently outperformed BERT-base in tasks like **text classification**, **named entity recognition (NER)**, and **topic modeling**, demonstrating the advantages of domain-specific adaptations in NLP. Future work with MythoBERT includes expanding its training data to encompass a broader array of cultural texts and exploring additional NLP tasks such as **sentiment analysis** and

relation extraction. Integrating MythoBERT with knowledge graphs could further enhance its applicability, making it a versatile tool for both academic research and real-world applications in digital preservation and interactive storytelling.

Additionally, the creation of **Mytho-Annotator** represents an important step forward in addressing the specific challenges posed by the unique narrative structures in Indian mythology. This annotation tool offers a user-friendly interface designed for collaborative annotation, supporting remote work by multiple researchers. The ability to annotate characters, relationships, and themes within these texts enables more detailed and structured analyses of the narratives. Future developments of Mytho-Annotator could involve enhancing entity recognition capabilities to support more complex character relationships and integrating machine learning models to assist in **semi-automated annotation** tasks, making it an even more powerful resource for digital humanities research.

In conclusion, this thesis has laid a strong foundation for the computational analysis of Hindu mythology through character identification, topic classification, ontology development, and summarization. The research has not only contributed to the growing field of NLP applied to cultural texts but also demonstrated the value of domain-specific tools like MythoBERT and Mytho-Annotator in tackling the unique challenges posed by mythological narratives. As we look to the future, expanding the datasets, refining models, and developing interactive digital tools will further enhance the field of computational mythology, offering richer insights into the timeless stories that have shaped human culture. The methods and models developed in this work have the potential to unlock new avenues for research and digital preservation, ensuring that these ancient narratives continue to inform and inspire future generations.

Appendix A

Visualization of Character-centric Summary

An application in Hindu Mythology

A.1 Introduction

Indian mythology, with its rich and intricate narratives, presents unique challenges for natural language processing (NLP) due to the complexity of its stories and the vast number of characters involved. The Mahabharata, one of the two major epics, offers an especially complex case with its deep and interconnected character relationships. This complexity makes it a prime text for exploring character-centric summarization, where the focus is placed on individual characters' roles, actions, and interactions. Recent advancements in NLP, particularly transformer-based models like T5, BART, and PEGASUS, have provided powerful tools for text summarization. However, character-centric approaches to summarization remain relatively under explored. This paper proposes fine-tuned versions of these models to tackle the challenge of character-specific summaries within the Mahabharata.

Transformer-based models, such as T5, BART, and PEGASUS, have significantly advanced the field of text summarization. These models generate coherent and con-textually appropriate summaries by leveraging their deep understanding of language structures. While traditional summarization focuses on summarizing entire documents, there is a growing need to tailor these models for more specialized tasks, such as character-centric summarization. For mythology, especially in narratives like the Mahabharata, character interactions are central to the story's meaning, and a focused approach is needed to capture each character's significance.

The motivation for this work arises from the desire to enhance the quality and relevance

of summaries for mythological texts by emphasizing the roles and actions of individual characters. Traditional summarization approaches often overlook these critical aspects, leading to less informative summaries when applied to character-driven narratives. By fine-tuning T5, BART, and PEGASUS specifically for character-centric summarization, this paper aims to generate summaries that are more reflective of each character's contributions to the Mahabharata, providing richer, more context-specific insights into the narrative.

Several challenges arise in the context of character-centric summarization for the Mahabharata. First, the complex structure of the narrative—with its multiple inter-woven storylines—makes isolating and summarizing character-specific actions diffi-cult. Additionally, the vast array of characters, from major figures like Arjuna and Krishna to more minor players, demands a delicate balance to ensure that key ac-tions are appropriately represented. Creating a dataset that captures the essential character-based sentences from the text is also labor-intensive, as there are few exist-ing resources that meet this need. Finally, fine-tuning models like T5, BART, and PEGASUS to effectively generate character-focused summaries while maintaining narrative coherence presents its own technical difficulties.

The objectives of this work are threefold: (1) to fine-tune T5, BART, and PEGASUS models for character-centric summarization, focusing on capturing the unique roles and actions of individual characters in the Mahabharata; (2) to create a custom dataset ($DS_{character}$) that consists of character-specific sentences, which are used to fine-tune these models for better summarization outcomes; and (3) to per-form keyphrase extraction and visualization, highlighting the key themes and actions associated with each character to provide a clearer understanding of their importance in the narrative.

This research work makes several important contributions. First, it introduces fine-tuned versions of T5, BART, and PEGASUS for character-centric summarization, specifically tailored for complex mythological texts like the Mahabharata. Second, it presents a custom dataset $DS_{character}$) that focuses on character-specific content, which serves as a critical resource for training and evaluation. Third, it offers a novel keyphrase extraction and visualization framework, providing an intuitive way to explore and understand each character's role and their relationships within the story.

The rest of this research work is organized as follows: Section 2 details the preparation of the $DS_{character}$ dataset, describing how character-based sentences were extracted for fine-tuning. Section 3 presents the system framework, outlining how T5, BART, and PEGASUS were fine-tuned and implemented for character-centric summarization and key-

phrase extraction. Section 4 analyzes the results, evaluating the performance of each model on the character-specific dataset. Section 5 offers a discussion and observation section, exploring the implications of the findings and how they contribute to the understanding of mythological texts. Finally, Section 6 concludes the research work and suggests directions for future research.

A.2 Preparation of Dataset

The preparation of the dataset is a crucial step in ensuring the effectiveness of the character-centric summarization and keyphrase extraction processes. For this study, the $DS_{character}$ dataset was utilized, which is tailored for the task of summarizing and extracting key information about characters in Indian mythology. The preparation involved several key steps:

The $DS_{character}$ dataset was sourced from PouranicTopic dataset developed by (Paul et al., 2024). This dataset contains a comprehensive collection of texts related to Indian mythology, specifically focused on character descriptions and narratives. To ensure the quality and usability of the dataset, the following preprocessing steps were carried out. The text was tokenized into sentences and words using SpaCy [1]. Named entities, especially character names, were identified and tagged using SpaCy's pre-trained models. Irrelevant information, such as meta-data or unrelated text, was removed to focus solely on character-centric content.

A.3 System Framework

The system framework for this research is designed to provide a comprehensive approach to character-centric summarization, keyphrase extraction, and visualization. The framework consists of three primary components: Character-Centric Summarization, Keyphrase Extraction, and Visualization. Each component plays a crucial role in processing and presenting the data effectively presented in figure A.1.

Once the character names were identified through Named Entity Recognition (NER), sentences related to each character were extracted from the text. This extraction involved segmenting the text into individual sentences using tools such as SpaCy and then filtering these sentences to include only those that referenced the identified characters. For each character, a comprehensive list of related sentences was compiled. Each character has a

[1] https://spacy.io/

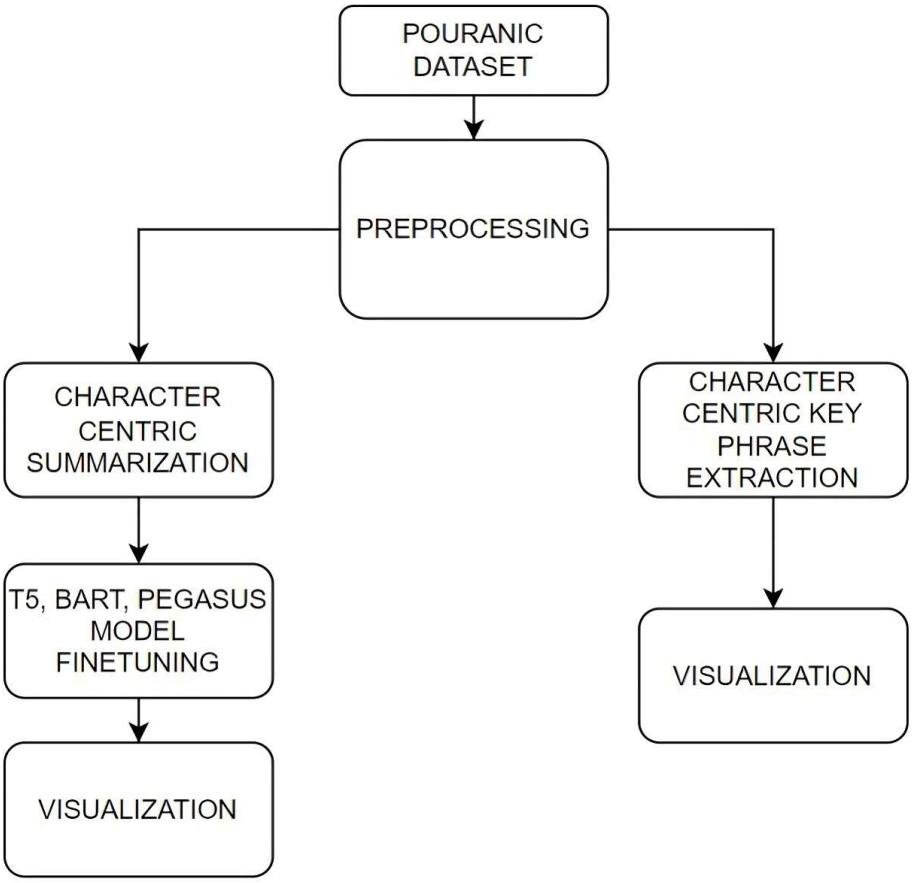


Figure A.1: Proposed System Framework

mapped image that will be used for visualization, allowing users to associate the summaries with the correct character visually.

To facilitate fine-tuning, summaries were manually created for each list of sentences. These summaries provided a condensed representation of the content related to each character, ensuring that the model training would be informed by accurately crafted and contextually relevant summaries. A quality assurance process was then applied, involving both manual review of the summaries to verify their accuracy and contextual relevance, as well as automated validation to ensure consistency and completeness across the dataset. To illustrate the preparation of the dataset, consider the following example:

Original: *Abhimanyu was born to Subhadra, the sister of Vasudeva, and Arjuna, making him the grandson of the illustrious Pandu. Known in his previous life as the mighty Varchas, son of Soma, he was reborn as Abhimanyu, a warrior of extraordinary deeds and accomplishments. Abhimanyu fathered a son, Parikshit, who was beloved by Vasudeva himself. Among all the warriors, Abhimanyu stood as the perpetuator of his family line, earning renown for his strength and valor.*

Summary: *Abhimanyu, son of Subhadra and Arjuna, was a renowned warrior known for his extraordinary deeds and valor. A reincarnation of Varchas, he continued his legacy through his son Parikshit, earning admiration as the perpetuator of his family line.*

This preparation process ensured that the dataset was well-suited for character-centric summarization and keyphrase extraction tasks, providing a solid foundation for the subsequent tasks.

A.3.1 Character-Centric Summarization

The first component of the framework is the Character-Centric Summarization. This process involves generating concise summaries that focus on individual characters within the dataset. The steps involved include:

- **Model Fine-Tuning:** Pre-trained language models, such as T5, BART, and PEGASUS, are fine-tuned on the character-specific sentences. The fine-tuning process adjusts the models to generate summaries focused on the narrative and attributes of each character. T5 was fine-tuned to optimize for reduced loss between generated and reference summaries, evaluated using ROUGE scores. BART followed a sequence-to-sequence approach, training the model to reconstruct summaries with similar evaluation criteria. PEGASUS, known for abstractive summarization, was fine-tuned for concise summaries, also evaluated using ROUGE and qualitative assessments. Each model was tailored to produce high-quality summaries that capture the essence of characters' narratives in Indian mythology.

- **Summary Generation:** The fine-tuned models produce summaries that highlight key aspects of each character's story, attributes, and significance within the context of Indian mythology.

A.3.2 Key-Phrase Extraction using KeyBERT

The second component involves Keyphrase Extraction using KeyBERT, a specialized model for extracting meaningful phrases from text. The steps include:

- **Keyphrase Identification:** KeyBERT, which leverages BERT embeddings to capture contextual meaning, is used to identify key phrases within the character-related sentences. The model helps in pinpointing the most relevant and meaningful phrases associated with each character.

- **Keyphrase Extraction:** KeyBERT extracts key phrases that highlight important aspects of the character's story and attributes. This extraction focuses on terms and phrases that are crucial for understanding the character's significance.

A.3.3 Visualization

The final component, Visualization, presents the summarized data and key phrases in an engaging and interactive manner. This component involves:

- **Summary Display:** Character-centric summaries are displayed on a web interface. To enhance visual appeal and recognition, each character's name in the summaries is replaced with their respective image. This approach makes it easier for users to associate the summary with the correct character.

- **Keyphrase Visualization:** The key phrases extracted by KeyBERT are visualized using methods such as word clouds. This helps users understand the prominence and relationships of key phrases associated with each character.

A.4 Result Analysis

A.4.1 Character-Centric Summarization

The performance of the T5, BART, and PEGASUS models in generating character-centric summaries was evaluated based on ROUGE and BLEU score. To evaluate the summarization quality of the T5, BART, and PEGASUS models, the ROUGE metrics—ROUGE-1, ROUGE-2, and ROUGE-L—were applied. The results are described in Table A.1.

Table A.1: ROUGE scores of the models for Character-centric summary

MODEL	ROGUE-1	ROGUE-2	ROGUE-L	BLEU
T5	0.65	0.70	0.68	0.66
BART	0.76	0.66	0.72	0.77
PEGASUS	0.71	0.74	0.75	0.70

Table A.1 presents a comparison of the T5, BART, and PEGASUS models across four evaluation metrics: ROGUE-1, ROGUE-2, ROGUE-L, and BLEU. These metrics are typically used to assess the quality of text generation models, particularly in tasks like summarization or machine translation. Each model exhibits varying performance across the different metrics, revealing their strengths and weaknesses.

In terms of ROGUE-1, which measures the overlap of unigrams between the generated and reference texts, BART performs the best with a score of 0.76. This indicates that BART captures more relevant individual words compared to the other models. PEGASUS follows closely with a score of 0.71, while T5 lags behind at 0.65, suggesting it is less effective at capturing relevant unigrams.

For ROGUE-2, which measures bigram overlap, PEGASUS takes the lead with a score of 0.74, demonstrating its strength in capturing context beyond single words. T5 is not far behind, scoring 0.70, while BART falls to third place with a score of 0.66, indicating its relative weakness in bigram matching. When it comes to ROGUE-L, which focuses on the longest common subsequence between generated and reference texts, PEGASUS continues to outperform the other models with a score of 0.75. BART follows with 0.72, while T5 scores 0.68, again showing that PEGASUS is more effective in capturing more extended sequences of relevant information.

Finally, for the BLEU metric, which measures n-gram precision between the generated and reference texts, BART excels with a score of 0.77, indicating its superiority in this evaluation. PEGASUS scores 0.70, while T5 trails behind slightly at 0.66, high-lighting that BART is particularly strong in generating text that closely matches the reference in terms of precision across n-grams. Overall, PEGASUS consistently performs well across the ROGUE metrics, indicating its effectiveness in generating coherent and contextually relevant sequences, while BART shows strong precision in terms of n-gram matching, especially with BLEU. T5, although generally competitive, appears to lag slightly behind the other two models across all metrics.

A.4.2 Key-Phrase Extraction Evaluation

The KeyBERT model was evaluated on its ability to extract significant key phrases from character-related sentences. The evaluation focused on the Precision, Recall and F1-score of the model shown in Table A.2. The performance of the KeyBERT model for keyphrase

Table A.2: Performance metrics of keyphrase extraction model

Precision	Recall	F1-score
0.73	0.78	0.75

extraction was evaluated using precision, recall, and F1-score metrics, yielding average values of 0.73, 0.78, and 0.75, respectively. The precision score of 0.73 indicates that 73% of the key phrases identified by the model were relevant, demonstrating its effectiveness in

selecting pertinent phrases without many irrelevant ones. The recall score of 0.78 reflects that the model successfully captured 78% of all relevant key phrases present in the dataset, showcasing its ability to identify a substantial portion of important phrases. The F1-score of 0.75, which balances precision and recall, further highlights the model's overall performance by showing a strong trade-off between accurately identifying key phrases and ensuring comprehensive coverage. These metrics collectively illustrate that KeyBERT performed effectively in extracting key phrases, providing valuable insights into the character-centric analysis.

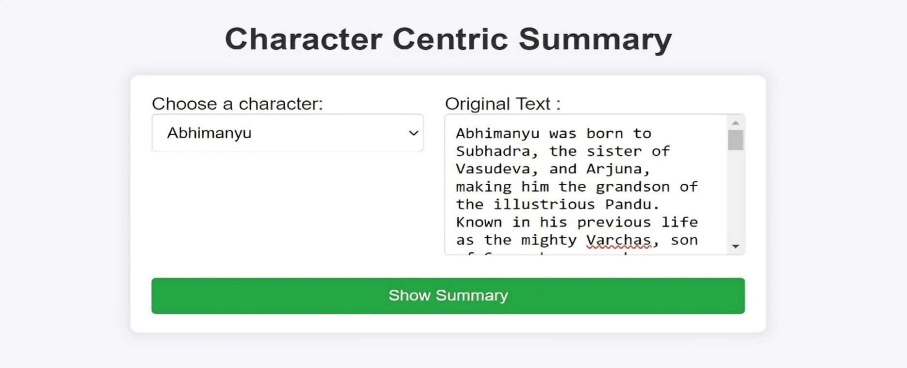


Figure A.2: Character-centric Summary Interface

A.4.3 Visualization

The visualization component of the system enhances the presentation and interaction with character-centric summaries and keyphrases through an intuitive web inter-face. This component includes several key features:

Summary Display: Each character-centric summary is prominently displayed on the web interface. To illustrate how the summaries and keyphrases are visualized, here we present several images of the web interface. These images showcase the key features and functionalities of the visualization system. Figure A.2 displays the web interface of Character-centric summary. Upon choosing a character, the lines from the text related to the character is displayed under Original text.

The generated summary is displayed under Original Summary and the character names in the summary are replaced by images in the next section, displayed in figure A.3.

From figure A.4, we can see that the keyphrases of the character **Abhimanyu** - *King Pariksit*, *Abhimanyu*, *Subhadra* ,*son of Arjuna*, *great hero Abhimanyu* in the list are visualized as word cloud.

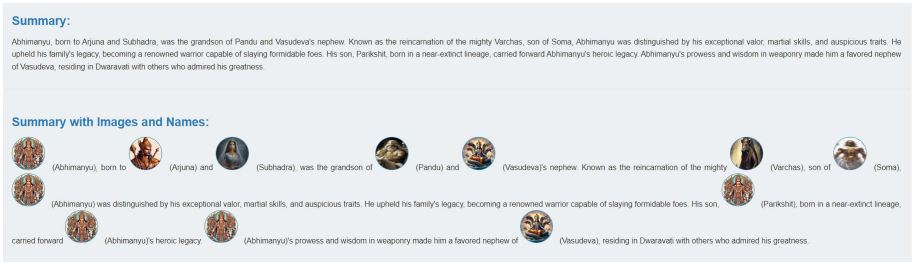


Figure A.3: Character-centric Summary output

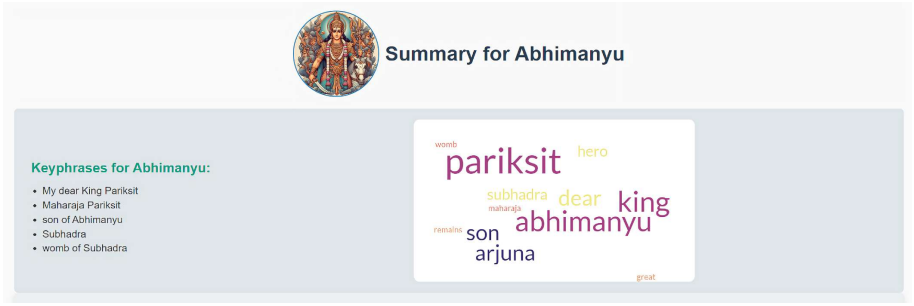


Figure A.4: Character-centric keyphrases

A.5 Discussion and Observations

The data highlights the performance of T5, BART, and PEGASUS across both the ROGUE and BLEU metrics, revealing distinct patterns in their efficiency and effectiveness. T5 shows moderate effectiveness, maintaining relatively balanced scores across all ROGUE metrics and a BLEU score of 0.66, indicating a consistent but not outstanding performance in both precision and recall tasks. BART, with a strong BLEU score of 0.77, further confirms its high efficiency in language translation and summarization, though its dip in ROGUE-2 suggests challenges in capturing complex relationships. PEGASUS, while slightly behind BART in BLEU with a score of 0.70, exhibits superior effectiveness across the ROGUE metrics, particularly excelling in ROGUE-2 and ROGUE-L. This suggests that PEGASUS strikes a strong balance be-tween precision and the ability to capture broader textual relationships, making it the most effective model overall across both performance metrics.

KeyBERT demonstrated varying performance in key-phrase extraction across different summaries. In particular, Summary achieved the highest Precision (0.73) and Recall (0.78), reflecting effective extraction of key phrases. However, there was noticeable variability in performance across other summaries, high-lighting areas where the method could be improved to achieve more consistent results.

A.6 Summary

This study evaluates T5, BART, and PEGASUS models trained on the PouranicTopic Dataset for summarization and key-phrase extraction tasks. PEGASUS excelled in ROUGE metrics, particularly ROUGE-L, indicating its strength in capturing coherent summaries. BART achieved the highest BLEU score, showing precision but lower bigram performance. T5 performed well in ROUGE-2 but could improve in overall coherence. KeyBERT was effective in extracting key phrases, though its consistency varied across summaries. Future research could focus on combining model strengths, expanding the dataset for better generalization, and exploring hybrid approaches for improved summarization and key-phrase extraction.

Bibliography

Afolabi, I. T., Daramola, O. J., and Adio, T. A. (2014). Developing domain ontology for nigerian history. *ArXiv*. 23

Akdemir, K. and Hurriyetoglu, A. (2022). Zero-shot ranking socio-political texts with transformer language models to reduce close reading time. In *CASE*. 21

Behrendt, M. and Harmeling, S. (2021). Arguebert: How to improve bert embeddings for measuring the similarity of arguments. In *Conference on Natural Language Processing*. 24

Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert: A pretrained language model for scientific text. In *Conference on Empirical Methods in Natural Language Processing*. 24, 28, 109

Bennani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., and Jaggi, M. (2018). Simple unsupervised keyphrase extraction using sentence embeddings. *arXiv preprint arXiv:1801.04470*. 26

Besnier, C. (2020). History to myths: Social network analysis for comparison of stories over time. In *LATECHCLFL*. 6

Bikaun, T., Stewart, M., and Liu, W. (2022). Quickgraph: A rapid annotation tool for knowledge graph extraction from technical text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278. 25, 29

Bojars, U., Rašmane, A., Žogla, A., Balina, S., and Salna, E. (2018). Semantic annotation tool for cultural heritage content. *Baltic Journal of Modern Computing*, 6:10.22364/bjmc.2018.6.4.09. 25

Bollmann, M., Petran, F., Dipper, S., and Krasselt, J. (2014). Cora: A web-based annotation tool for historical and other non-standard language data. In *Proceedings of*

the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 86–90. 25

Braun, D. and Matthes, F. (2022). Clause topic classification in german and english standard form contracts. In *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*. 21, 27

Buddhi, D., Joshi, A., and Negi, P. (2022). Language model based related word prediction from an indian epic-mahabharata. In *International Interdisciplinary Humanitarian Conference for Sustainability (IIHC)*. 5, 20

Calix, R. A., Javadpour, L., Khazaeli, M., and Knapp, G. M. (2013). Automatic detection of nominal entities in speech for enriched content search. In *The Twenty-Sixth International FLAIRS Conference*, pages 190–195. 19

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., and Jatowt, A. (2020). Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289. 27

Chalkidis, I., Fergadiotis, M., and Androutsopoulos, I. (2021). Multieurlex - a multilingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *CoRR*. 21

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). Legal-bert: "preparing the muppets for court'". *ArXiv*, abs/2010.02559. 24, 28, 109

Cunningham, H. (2002). Gate, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254. 25, 112

Das, D., Das, B., and Kavi, M. (2016a). A computational analysis of mahabharata. In *13th International Conference on Natural Language Processing, NLP Association of India*. 20

Das, D., Das, B., and Mahesh, K. (2016b). A computational analysis of mahabharata. In *Proceedings of the 13th International Conference on Natural Language Processing*, pages 219–228. 4, 8

de Castilho, R. E., Biemann, C., Gurevych, I., and Yimam, S. M. (2014). Webanno: a flexible, web-based annotation tool for clarin. In *Proceedings of the CLARIN Annual Conference (CAC)*, pages 4372–4380. 25, 29, 112

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186. 23, 72, 101, 109

Dominic, K. (2021). Puranas - the literary way of indian philosophy. *International Research Journal of Management Sociology & Humanity (IRJMSH)*. 61

Duncan, S. V. (2006). *A Guide to Screenwriting Success: Writing for Film and Television*. Rowman & Littlefield. 31

Dutta, N., Mondal, A., and Paul, P. (2020). An annotation system to annotate healthcare information from tweets. In *Emerging Technology in Modelling and Graphics*, page 30. Springer, Singapore. 25

Espinoza-Arias, P., Poveda-Villalón, M., García-Castro, R., and Corcho. (2018). Ontological representation of smart city data: From devices to cities. *Applied Sciences*. 22

Feng, F., Yang, Y., Cer, D. M., Arivazhagan, N., and Wang, W. (2020). Language-agnostic bert sentence embedding. In *Annual Meeting of the Association for Computational Linguistics*. 24

Florescu, C. and Caragea, C. (2017). Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. 26

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms: The c-value/nc-value method. *International Journal of Digital Libraries*, 3(2):117–132. 20, 46

Freeman, M. (2016). *Historicising Transmedia Storytelling: Early Twentieth-Century Transmedia Story Worlds*. Routledge. 31

Frei, J., Soto-Rey, I., and Kramer, F. (2022). Drnote: An open medical annotation service. *PLOS Digital Health*, 1(8):e0000086. 25, 29

Fumanal-Idocin, J., Cordón, O., Dimuro, G. P., de Hierro, A.-F. R. L., and Bustince, H. (2023). Quantifying external information in social network analysis: An application to comparative mythology. *IEEE Transactions on Cybernetics*, 54:3417–3430. 7

Fumanal-Idocin, J., Cordón, O., Dimuro, G. P., Min'arov'a, M., and Bustince, H. (2021). The concept of semantic value in social network analysis: an application to comparative mythology. *ArXiv*, abs/2109.08023. 7

Gadesha, V., Joshi, K. D., and Naik, S. T. (2023). Estimating related words computationally using language model from the mahabharata - an indian epic. *ArXiv*. 5, 8, 20

Gao, Y., He, J., and Li, Z. (2020). Summarunner: A recurrent neural network based sequence model for extractive summarization. *arXiv preprint arXiv:2004.08056*. 26

Gentzkow, M., Kelly, B. T., and Taddy, M. (2017). Text as data. *Capital Markets: Market Efficiency eJournal*. 61

Gersberg, I. W. and Ebecken, N. F. F. (2014). A simple strategy to start domain ontology from scratch. In *International Journal of Advanced Computer Science and Applications*. 23

Glavas, G., Nanni, F., and Ponzetto, S. P. (2017). Cross-lingual classification of topics in political texts. In *NLP+CSS@ACL*. 21, 27

Gong, Y. and Liu, X. (2021). Extractive summarization with contrastive learning. *arXiv preprint arXiv:2105.00119*. 26

Goyal, A., Riloff, E., and Hal, D. I. (2010). Automatically producing plot unit representations for narrative text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 77–86. Association for Computational Linguistics. 19

Grootendorst, M. (2020). Keybert: Simple and effective keyphrase extraction with bert. *arXiv preprint arXiv:2010.11972*. 26

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220. 85

Gultepe, E. and Mathangi, V. (2023). A quantitative social network analysis of the character relationships in the mahabharata. *Heritage*, 6(11):7009–7030. 6, 8

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques*. Elsevier Inc. 43

Hlomani, H. and Stacey, D. A. (2014). Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey. *ArXiv*. 96

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Annual Meeting of the Association for Computational Linguistics*. 25

Huang, K., Altosaar, J., and Ranganath, R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*. 24

Iosif, E. and Mishra, T. (2014). From speaker identification to affective analysis: A multi-step system for analyzing children stories. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL) @EACL 2014*, pages 40–49. 31

Jiang, L. P. (2012). Designing family ontology with the protégé owl plugin. In *Materials Science and Information Technology II*, pages 836–840. 85

Karan, M., Šnajder, J., Širinić, D., and Glavas, G. (2016). Analysis of policy agendas: Lessons learned from automatic topic classification of croatian political texts. In *LaTeCH@ACL*. 21, 27

Kenna, R. and MacCarron, P. (2017). A networks approach to mythological epics. 7

Kenter, T. and Maynard, D. (2005). Using gate as an annotation tool. In *University of Sheffield, Natural Language Processing Group*. 25

Khattab, O. and Zaharia, M. A. (2020). Colbert: Efficient and effective passage search via contextualized late interaction over bert. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 24, 28

Kiesel, J., Wachsmuth, H., Al Khatib, K., and Stein, B. (2017). Wat-sl: a customizable web annotation tool for segment labeling. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 13–16. 25

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980. 25

Kostkan, J., Kardos, M., Mortensen, J. P. B., and Nielbo, K. L. (2023). Odycy – a general-purpose nlp pipeline for ancient greek. *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. 6

Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. (2017). Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 73

Lang, K. (1995). Newsweeder: learning to filter netnews. In *Proc. of the International Conference on Machine Learning (ICML)*. 61

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. 24, 28, 109

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., and Gouws, S. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*. 26

Lin, C. Y. and Liu, X. (2019). Bertsum: Bert-based extractive summarization. *arXiv preprint arXiv:1904.02340*. 26

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press. 61

Liu, Y. and Lapata, M. (2019). Hierarchical transformer for extractive summarization. *arXiv preprint arXiv:1907.12428*. 26

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., and Chen, D. (2019). Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692. 72, 101

Maeda, K. and Strassel, S. M. (2004). Annotation tools for large-scale corpus development: Using agtk at the linguistic data consortium. In *LREC*. 25

Mamede, N. and Chaleira, P. (2004). Character identification in children stories. In *Advances in Natural Language Processing, Springer Berlin Heidelberg*, pages 82–90. 19

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. 62, 67

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3):276–282. 69, 91

Meng, R., Huang, X., Yuan, J., and Wang, D. (2017). Copyrnn: A sequence-to-sequence model with copy mechanism for keyphrase generation. *arXiv preprint arXiv:1704.06377*. 26

Mohan, A. R. and Arumugam, G. (2010). Developing indian medicinal plant ontology using owl and swrl. In *International Conference on Data Engineering and Management*. 22, 28

Mondal, A., Cambria, E., and Dey, M. (2022). An annotation system of a medical corpus using sentiment-based models for applications. In *Computational Intelligence Applications for Text and Sentiment Data Analysis*, pages 163–178. Academic Press. 25, 112

Morton, T. and LaCivita, J. (2003). Wordfreak: An open tool for linguistic annotation. In *HLT/NAACL 2003: demonstrations*. 25

Ness, E., Fatima, A., and Oghaz, M. M. (2023). Data driven model to investigate political bias in mainstream media. *IEEE Access*, pages 41880–41893. 21

Noguti, M. Y., Vellasques, E., and Oliveira, L. (2020). Legal document classification: An application to law area prediction of petitions to public prosecution service. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. 21, 27

Noy, N. F. and McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. Technical Report KSL-01-05, Stanford Knowledge Systems Laboratory. 22, 93

Osnabrügge, M., Ash, E., and Morelli, M. (2021). Cross-domain topic classification for political texts. *Political Analysis*, pages 59–80. 21

Pannach, F. and Blaschke, T. (2023). Modeling and comparison of narrative domains with shallow ontologies. In *International Conference on Language, Data, and Knowledge*. 23

Papaloukas, C., Chalkidis, I., Athinaios, K., Pantazi, D.-A., and Koubarakis, M. (2021). Multi-granular legal topic classification on greek legislation. In *ArXiv*. 21

Pastor-Sánchez, J.-A., Kontopoulos, E., Saorín, T., Bebis, T., and Darányi, S. (2021). Greek mythology as a knowledge graph: From chaos to zeus and beyond. *ArXiv*. 6, 22, 28

Paul, A. and Das, D. (2017a). A deep dive into identification of characters from mahabharata. In *ICON*. 94

Paul, A. and Das, D. (2017b). Identification of character adjectives from mahabharata. In *Recent Advances in Natural Language Processing*. 61, 94

Paul, A., Seal, S., and Das, D. (2024). Transformer-based pouranic topic classification in indian mythology. *Sādhanā*. 23, 28, 88, 111, 139

R., S. B. and Aithal, P. S. (2023). Five factor taxonomy of personality traits & ocean model from mahabharata characters. *International Journal of Case Studies in Business, IT, and Education*. 5, 8

Raffel, C., Shazeer, N., and Roberts, A. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*. 26

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 73

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 67

Sanabila, H. R. and Manurung, R. (2014). Automatic wayang ontology construction using relation extraction from free text. In *LaTeCH@EACL*. 22

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108. 72, 101

Sarkanych, P., Fedorak, N., Holovatch, Y., MacCarron, P., Yose, J., and Kenna, R. (2022). Network analysis of the kyiv bylyny cycle - east slavic epic narratives. *Adv. Complex Syst.*, 25:2240007:1–2240007:25. 7

Sattar, A., Salwana, E., Nazir, M., and Ahmad, M. (2020). Comparative analysis of methodologies for domain ontology development: A systematic review. *International Journal of Advanced Computer Science and Applications*. 23, 28

Srijeyarankesh, S., Aishwarya, K., Nithyasri, L., and priya, M. S. (2022). A comprehensive study of mahabharat using semantic and sentiment analysis. In *13th International Conference on Natural Language Processing (ICON-2022)*. 20

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. 25, 29, 112

Sun, Q., Wang, H., Li, P., Ren, Z., Chen, J., and Ma, J. (2019). Attention-based neural keyphrase extraction. *arXiv preprint arXiv:1906.09861*. 27

Syamili, C. and Rekha, R. (2017). Developing an ontology for greek mythology. *Electron. Libr.*, 36:119–132. 6, 22, 28

Tesconi, M., Ronzano, F., Minutoli, S., Aliprandi, C., and Marchetti, A. (2010). Kafnotator: a multilingual semantic text annotation tool. In *The Second International Conference on Global Interoperability for Language Resources*, volume 1. 25

Tuamsuk, K., Chansanam, W., and Kaewboonma, N. (2018). Ontology of folktales in the greater mekong subregion. *Int. J. Metadata Semant. Ontologies*, 13:57–67. 22

Tyers, F. M., Sheyanova, M., and Washington, J. N. (2017). Ud annotatrix: An annotation tool for universal dependencies. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 13–16. 25

Ushio, A., Camacho-Collados, J., and Schockaert, S. (2023). Relbert: Embedding relations with language models. *ArXiv*, abs/2310.00299. 24, 29

Valls-Vargas, J., Ontanon, S., and Zhu, J. (2013). Toward character role assignment for natural language stories. In *Proceedings of the Ninth Artificial Intelligence and Interactive Digital Entertainment Conference*. 19

Valls-Vargas, J., Zhu, J., and Ontanon, S. (2014). Toward automatic character identification in unannotated narrative text. In *Proceedings of the Seventh Workshop in Intelligent Narrative Technologies*. 19

Valls-Vargas, J., Zhu, J., and Ontanon, S. (2015). Narrative hermeneutic circle: Improving character role identification from natural language text via feedback loops. In *Proceedings of the 24th International Conference on Artificial Intelligence*. AAAI Press. 19

Vandeck, N. J., Waltman, L., Noyons, E. C. M., and Buter, R. K. (2010). Automatic term identification for bibliometric mapping. *Scientometrics*, 82(3). 20

Varadarajan, U., Bagchi, M., Tiwari, A., and Satija, M. P. (2022). Genome: A generic methodology for ontological modelling of epics. *ArXiv*. 5, 8, 20, 22, 28

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need. 25, 27

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 353–355. Association for Computational Linguistics. 72

Wang, S. and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 90–94. 61

Wang, S., Thompson, L., and Iyyer, M. (2021). Phrase-bert: Improved phrase embeddings from bert with an application to corpus exploration. *ArXiv*, abs/2109.06304. 24, 28

Wang, Y. (2023). Topic classification for political texts with pretrained language models. In *Political Analysis*, pages 662–668. 21

Watrobski, J. (2020). Ontology learning methods from text - an extensive knowledge-based approach. In *International Conference on Knowledge-Based Intelligent Information & Engineering Systems*. 22, 28

Wei, J. and Santos, E. (2020). Narrative origin classification of israeli-palestinian conflict texts. In *The Florida AI Research Society*. 21

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. 25

Xenouleas, S., Tsoukara, A., Panagiotakis, G., Chalkidis, I., and Androutsopoulos, I. (2022). Realistic zero-shot cross-lingual transfer in legal topic classification. In *Proceedings of the 12th Hellenic Conference on Artificial Intelligence*. 21

Xu, H., Durme, B. V., and Murray, K. (2021). Bert, mbert, or bibert? a study on contextualized embeddings for neural machine translation. In *Conference on Empirical Methods in Natural Language Processing*. 24

Zhang, Y. and et al. (2020). Event detection using graph neural networks. In *Proceedings of the ACL Conference*, pages 234–245. 26

Zhou, W., Zhang, J., and Xu, Y. (2021). Abstractive document summarization with pre-trained language models. *arXiv preprint arXiv:2102.10920*. 26