Short Story Genre Prediction: A Study Driven By Transformer And Sequential Models

by

Sadman Shakib
24141169
Md Talha Ibne Hasan
24141170
Mim Raihan
24341098
Azanuzzaman Bhuiya
24141232

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
June 2025.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Sadman Shakib 24141169	Md Talha Ibne Hasan 24141170
Mim Raihan 24341098	Azanuzzaman Bhuiya 24141232

Approval

The thesis/project titled "Short Story Genre Prediction: A Study Driven By Transformer And Sequential Models" submitted by

1. Sadman Shakib (24141169)
2. Md Talha Ibne Hasan (24141170)
3. Mim Raihan (24341098)
4. Azanuzzaman Bhuiya (24141232)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June, 2025.

Examining Committee:

Supervisor:
(Member)

Md Tanzim Reza
Senior Lecturer
Department of Computer Science and Engineering
BRAC University

Co-Supervisor:
(Member)

Farig Yousuf Sadeque
Associate Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

A huge number of genres that represent a variety of themes, styles, and narrative structures are represented in short stories. Genre detection in short stories is an integral task for analysis, recommendation systems, content organization, and research. Therefore, the topic we have decided to work on is Short Stories Genre Detection using NLP techniques. The goal is to identify the genre of stories so that the audience may quickly determine whether or not the story is appropriate for them. Our approach is based on feeding the model as much big data as possible and maximizing efficiency by emphasizing on non automatically marked-up text features. Specifically, we experimented with models such as LSTM, Bi-LSTM, and transformer-based architectures like BanglaBERT, DistilBERT, and RoBERTa for genre classification tasks. After research, BanglaBERT showed the best results while the sequential models showed the least improved results. Among the five models evaluated, BanglaBERT achieved the highest performance with an accuracy of 72% , significantly outperforming both sequential and transformer-based models. RoBERTa followed in 2nd place showing strong results especially for certain classes. DistilBERT placed last among the BERT models. Moreover, Bi-LSTM showed better performance than LSTM with an F1-score of 60%, compared to LSTM. Overall, transformer models, particularly BanglaBERT and RoBERTa, performed significantly better compared to the LSTM-based models.

Keywords: NLP; Genre detection; BanglaBERT; LSTM; Bi-LSTM; DistilBERT; RoBERTa

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis has been completed without any major interruption. Secondly, to our supervisor and co-supervisor sir, for their kind support and advice in our work. They helped us whenever we needed help. And finally, to our parents, without their support, it may not have been possible. With their kind support and prayer, we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

A short story is a brief version of a story which helps to evoke an emotional response within the limitations of work. The exact length of short stories may vary but generally. Since the length is shorter than a typical story, it consists of fewer characters as well as specific events rather than creating multiple events. The history of short stories is seen in ancient Greece.[1]Aesop's fable which was written in 6th BCE and Panchatantra was written around 600 CE. During the Renaissance Miguel de Cervantes (author of "Don Quixote") and Geoffrey Chaucer began to explore short prose narratives that contributed to the expansion of the short story[2]. The pioneer of modern short stories Edgar Allan Poe. Poe put his effort into the significance of the single effect and unity in storytelling. His contributions to "The Tell-Tale Heart" and "The Fall of the House of Usher" are noteworthy[3]. Today, the short stories are not limited to books. Online publications, blogs, and digital books have given this section of literature a new life. Short stories provide a dynamic form of literary expression. The story includes profound truths and complex emotions within a limited scope. It is very difficult to understand the metaphor or the gist of a story. With the advancement of technology, artificial intelligence is helping the world to enhance efficiency without any limits. Natural Language Processing helps to categorize text, analyze sentiments, and understand the metaphors of these stories. NLP can work on the sentiments expressed in a short story of joy, sadness, anger, or surprise. This helps in understanding the emotional tone and mood of the narrative. By identifying recurring themes and topics, NLP can provide insights into the central ideas and motifs of a short story. This can be useful for literary analysis and academic research. With the help of NLP, we can identify the genre of short stories.

1.2 Problem Statement

Short stories have long been a vital component of literature, providing concise yet powerful narratives that explore a wide range of themes and human experiences. Various works have already been done on finding genre in English Literature but working on Bangla Short Stories is a new field to work on. The dataset is limited as well as Bangla language contains more letters than English which makes it difficult to depict true emotion. Identifying the central themes in short stories is essential

for various applications, including literary analysis, educational purposes, and enhancing digital library search functionalities. Conventional ways of genre detection are heavily favored on manual analysis by experts. The procedure is both time-consuming and personalized. With NLP's emergence, there is huge potential for automation and detection of theme detection and therefore, making it more convenient to operate. One of key focuses is to solve the genre detection of stories where multiple or more than one genres co-exist. Seldom do we see short stories with one main theme or genre as the main genre often encompasses a large number of sub-genres underneath it. Due to dataset story overlap, ambiguity, complexity of genre detection, a common misconception of having only one main theme is noticed in the research papers but that is totally not accurate. Our research will be heavily focused on solving the problem of hybrid genre classification prevalent in short stories. Furthermore, we will also focus on catching thematic nuances and contextual dependency as it is essential for detecting and setting the plot of the stories.
We will try to work on genre classification by using LSTM. As manual classification takes so much time, it would save time. We will focus on ensuring less computational time. In a library, if books are not placed according to their genre , book readers will lose their interest to read books, rather they will lose their energy to find books according to their interest.In different kinds of previous paper, many research had already been done on genre classification but we will try to attain the optimum accuracy and computational time.
In the paper by Vargas et al. (2017) [4], one of the main problems is the limited size of the dataset used for evaluation which consisted of only 21 Russian stories manually translated to English. A larger and more diverse dataset could provide a more comprehensive assessment of the system's performance and generalizability. Additionally, the paper may benefit from a more in-depth discussion on the scalability of the approach to handle a wider range of narrative styles and complexities beyond the stories included in the dataset. Similarly, in the paper by Pal et al.(2014) [5], only 159 Bangla stories were tokenized and taken for research. Hence, there is a limit of dataset size which is not suitable when trying to process different and diverse short stories with varying contexts. Short stories come in various styles, genres, and narrative structures. A large dataset ensures that different writing styles, tones, and thematic expressions are adequately represented. This helps the model generalize well across different types of short stories. Moreover, in the paper by Sims et al.(2019) [6] , there is a limited dataset scope as all the papers used were published before 1923 which results in a lack of generalization of findings related to contemporary and recent literature. Furthermore, the research paper did not go into information about the limitations and shortcomings of the annotation process which could affect the correctness of the prediction. In addition, in the paper by Berbatova (2019) [7], no definite operating procedure had been set or defined for determining the assessment of content based recommendations. Numerous NLP methods had been mentioned and used but had not been properly compared. Furthermore, case studies, incidents and examples showing how NLP methods used in real life for enhancing content based recommender systems would improve the quality of the paper and may add further information on the dataset provided.
In the paper by Stamatatos et al.(2003) [8] word frequency is considered the main basis for text detection. A theme of quantity over quality is noticed in the research of the paper. Word frequency alone is often insufficient because it fails to

capture contextual meaning, semantic relationships, and syntactic nuances. High-frequency words, such as stop words and function words, carry little semantic value. Rare words, despite their low frequency, can be highly informative. Contextual understanding is crucial, as words can have multiple meanings (polysemy) or be synonyms with different frequencies. Advanced techniques like word embeddings (e.g., Word2Vec, BERT) and TF-IDF consider both context and semantic similarity, providing a deeper and more accurate understanding of language than frequency counts alone.

1.3 Research Objectives

Manual classification takes a long time and is extremely challenging. While categorizing manually, errors are most likely to happen. Our motive is also driven by the necessity of advancement in nlp. Our research will also help in improving content recommendation systems. In addition, It will help instructors in customizing lessons by classifying materials for reading based on genre. If a reader is given content according to his/her taste and choices, It will make his/her reading experience far better. It can be done by our classification research. The significance of Bangla story genre detection lies in its potential to streamline the categorization of literary content, enhancing the reading experience by aligning readers with materials that match their interests. When selecting a book, readers can choose with an overview of what to expect from it given the variety of genres. When anyone reads a book labeled as mysterious or when readers open a romantic or adventurous book, they predict specific features from previous books in the same genre, such as character connections, literary techniques, and story patterns. Our research is not only limited in classification work,but also it will also help commercially. In addition, we will try to focus on attaining best possible accuracy. Genre classification in Bangla presents significant challenges due to the lack of extensive corpora and the intricacies of the language. Furthermore, essential pre-processing tools, such as stemmers and part-of-speech taggers, are scarce, complicating NLP tasks.

Chapter 2

Literature Review

2.1 Literature Review

[9] The paper consisted of finding a method to categorize and to find the genre of the book from the short description given in the book. Two lazy methods were proposed which are kNN based method and the topic vector-based method. The dataset was collected from the Goodreads website which had the genre mentioned in each book. There were 160794 books in the dataset and the genres mentioned were 506. 13 most popular genres were chosen to get only the base categories. The number was arbitrary. 90 percent of the data was trained and 10 percent of the data was tested in the algorithms. Data preprocessing included lowercasing the data, removal of punctuation, and splitting of the data. Naive Bayes and Doc2Vec and Averaged Doc2Vec approaches were tested. It was seen that Doc2Vev and Averaged Doc2vec worked better in terms of categorizing than the naive Bayes baseline. It was because Doc2Vec could capture context and thus was more effective because it was a paragraph vector-based model. Because of not having an NLP-based approach, this process became more universal.
[10] Biradar et al (2019) researched classifying the genre of the book from the book title and book cover using NLP and image processing. The research was done based on both the title and the cover. It had three parts. They were image features, title features, and classification. Data preprocessing was done by resizing and normalizing the book cover image. A glove model was used for extracting the features. Word-word co-occurrence matrix was used. For classification, a multinomial logistic regression model was used for the research. The dataset contained 4176 training images from amazon.com. Keras model was used to load a pre-pre-trained exception model. The results were seen better when combining both image processing and classification rather than using only one. When combining both, the accuracy was 87.2 %. When only the image was used, it became 67.4 % and when only the title was used, it became 86.2 %.
This research [11] was about identifying themes using various machine-learning techniques. Extracting a theme from a text was a very useful method that could be automated rather than done manually. The techniques mentioned were binary classification, multiclass classification, feature extraction, etc. In this research the theme detector was made using Linear SVM and Multinomial Naive Bayes. Data was split for training and testing. Feature extraction was done using TF-IDF. The dataset was built using different texts from Google Scholar. Data pre-processing was done

by removing stop words, digits, punctuation marks, and so on. Also, words with less than 4 letters were removed. Splitting the data was the next step of theme identification. 80 % data was used for training and 20 % for testing. TF-IDF approach was used for feature extraction. A linear SVM with a soft margin and multinomial naive Bayes was used for the classification part. Accuracy was tested based on the confusion matrix where it was seen that using Linear SVM gave an accuracy of 100 % whereas multinomial naive Bayes gave an accuracy of 98.50 %.
[12] Qorich et al. (2023) researched and analysed the sentiments in the review of Amazon. The paper proposed a CNN model and it was compared with several word embedding models. Kims CNN model was suggested. It had three convolutional layers, three connected layers, and a max pooling layer. Different representations of the embedding layer were done in this paper. The dataset used was taken from Kaggle which contained Amazon reviews. The dataset contained 4000000 reviews from customers. 70 % of the data was trained, 20 % of the data was validated and 10 % of the data was tested. Data preprocessing was done by removing noisy texts, regular texts, etc. The accuracy of the proposed CNN model was 90 %. It was also seen that the stemming algorithm affected the accuracy significantly. It was 87.04 % with stemmer and 84.94 % without it.
[13] Cen et al. (2020) researched analyzing sentiments in IMDB movie reviews using some NLP techniques. The NLP techniques used here are LSTM, CNN, and RNN. It was then compared with SVM and RNTN. This research would help the users to get particular recommendations and the producers for their market research and so on. The first model mentioned in the research was a convolutional neural network. It contained three layers which are a polling layer, a fully connected layer and a convolutional layer. The second model mentioned was RNN. Taking sequence data as input, it recursively recursed in the evolution direction. LSTM was the third model explained which was developed from RNN. In this research, the model of CNN had 128 input neurons. LSTM had 128 LSTM blocks and an output layer with 2 neurons. Tanh function was the activity function. Baidus PaddlePaddle Deep Learning framework was used to conduct the experiments. A public IMDB review of 50,000 datasets was used for this research. From the research, it was seen that CNN got the highest accuracy of 88.22% while RNN got 68.64% and LSTM got 85.32%.
[14] The present research investigates how different machine learning algorithms could be used to identify book genres depending on their titles. Five models were used for this research. The dataset was taken from Amazon which consists of 2207.575 book tiles including 32 genres. Book Titles were processed by doing Tokenization,vectorization using word to vec ,also with padding and one hot encoding. The study cames to the conclusion that while CNNs perform well in predicting genre-related word patterns, LSTM models were still better for this task.LSTM model works with the accuracy of 65% and Cnn with the accuracy of 63%.By adding attention methods to the LSTM outputs to identify the most important components of an input, the models could be improved. Experimenting with other architectures including Support Vector Machines (SVM), could also be effective.Though the CNN and LSTM models performed well, it was likely they might ignore subtle context which could have been better caught in longer texts.
[15] The increasing diversity of text collections has made the automatic detection of text genres essential in computational linguistics, demanding precise categorization algorithms that go beyond traditional wisdom. The foundational work of aca-

demics like Aristotle, Butcher, Dubrow, Fowler, Frye, Hernadi, Hobbes, Staiger, and Todorov forms the basis of this field of study. they also investigated genre as a classification based on communicative purposes and formal properties. Biber's multi-dimensional approach in computational linguistics classified texts along textual dimensions through the identification of linguistic elements that distinguished genres by factor analysis. In their investigation of structural cues for genre detection using the Brown Corpus, Karlgren and Cutting experienced limitations with imprecise categories and the necessity for tagged texts. They showed that surface cues are capable of identifying text genres by modeling genre detection using logistic regression techniques and neural networks. The ability of the system to autonomously identify genre offers useful uses in information retrieval, word-sense disambiguation, parsing, and part-of-speech tagging, enabling enhanced precision and customized search outcomes for individual users. The study of Kessler et al. is a noteworthy achievement, emphasizing the relevance of genre in computer linguistics and establishing the feasibility of genre recognition using readily available features.
[16] Lee et al. (2002) research on automatic genre classification by particular specified features from training data that had been genre- and subject-classified.They worked on genre classification by implementing the tf,idf method.To compute how many documents related to the category contain the term, how perfectly the term is spread out among the class, and how discerning the term is among different categories are important features. Here the hypothesis is a good genre revealing words that will appear in many texts.Here they also used a naive Bayes classifier approach. They used web documents including 7828 Korean documents and 7615 English documents. Half documents were used for training and the other half for testing. Effectiveness in this paper was counted with recall and precision scores. Finally, they balanced the tf and df ratio to make efficient selection for each genre class. The common strategy extends precision by increasing thresholds when precision is low. In total, Word statistics (df ratios, discrimination formulas) had been used to determine the genre. As here two kinds of language usage ensure the method's effectiveness. They only tested this method on a fixed set of genres which they decided earlier. So it is not sure how effective would be their method on other genres and document types. Also they could use more advanced method for more accurate result like Svm or any other machine learning model which can handle more complex datasets.
[17] Agarwal et al. (2021) proposed employing character networks for sorting diverse book genres. They have compared graphs with the same character sets utilizing the eigenvectors of the Laplacian matrices of the feature graphs to bring attention to the feature interactions. Their dataset is from Fanfiction.net. This gives them a chance to think about the same character set from multiple viewpoints. The text is preprocessed through tokenization, lemmatization, and part-of-speech tagging with the goal of extracting character networks. Name entity recognition is used to make connections among different occurrences of characters. Consequently character interaction matrix is made, which shows how frequently pairs of characters interact.Then character interaction graphs are made by these matrices. They also used different types of machine learning algorithms like Random forest, support Vector machine(SVM), Gaussian process classifier(Gpc), Gradient Boosting Gaussian naive bayes,adaboost, Multi layer perceptron (MLP) for classification. All of them are used on the dataset to see how accurate they are in classification. Their method

acquired a 67.4% level of accuracy and an 80.5% F1 score.Their research builds on the new proposal that emphasizes the necessity of character interactions.
[18] Pepa et al.(2024) developed a number of diagnostic properties to determine existing explainability techniques and use this list to compare diverse techniques on text classification tasks and neural network architectures which includes CNN, LSTM, and Transformer models. The author compared saliency scores from explainability techniques with human annotations to investigate the relationship between model performance and the alignment of its rationales with human reasoning. Gradient-based explanations generally perform best across tasks and architectures. The evaluation across three text classification tasks with human-annotated salient tokens revealed the explainability techniques perform best with CNN models on e-SNLI and IMDB datasets. LSTM and Transformer models worked with less efficiency.From the result, the Transformer performs best overall but its complex architecture challenges explainability tools more than simpler models like CNN. Detailed results regarding model performance variations on specific datasets are provided in the supplementary material.
[19] Duarte et al.(2023) gave a review of semi-supervised learning (SSL) for text classification that identified key idioms, text representations, tasks, datasets, algorithms, domains, and SSL techniques used in the field. The authors examined 157 articles from 2017 to 2022 which highlighted the common use of 22 benchmark datasets and various text representation methods that included BoW, TF-IDF, Word2Vec, fastText, GloVe, BERT, DistilBERT, ALBERT, and ELMo. They worked on SSL techniques to determine the percentage of data. It was practically impossible to recommend a specific classifier for all problems. So, the author provided efficiency based paths and promoted the use of diverse datasets which included those in non-English languages. The paper outlined current research trends and emphasized SSL's potential to reduce annotation costs while achieving competitive results. The tables provided detailed information on dataset specifics and results. The limitations were having fewer non-english resources and the difference of evaluation matrices which limited the proper evaluation of the result.
[20] Wilmot et al.(2020) suggested a hierarchical language model that encoded stories and computed surprise and uncertainty reduction to predict suspense and also achieved near-human accuracy when experimented against human-annotated short stories. This model used Generative Pre-Training based on psycholinguistic concepts of human language processing. Less engaging datasets like ROC Cloze, utilized the WritingPrompts corpus that contained around 300k creatively written short stories from the /r/WritingPrompts subreddit. It was split into 90 % training, 5 % development, and 5 % test sets. The model's ability to predict suspenseful events is also checked on the movie synopsis. Modern neural network models can implicitly represent text meaning and effectively analyze or generate narrative fiction. It laid the groundwork for future research and applications in modeling suspense computationally within NLP systems. More sophisticated models could enhance this by incorporating coreference and structured event representations.
[21] Qasim et al.(2022) explored the application of using three datasets. The first one was fake news of COVID-19, tweets regarding COVID-19, and extremist as well as nonextremist. These datasets were preprocessed by removing links, converting text into lowercase, lemmatizing, and eliminating punctuation. This preprocessing step ensured the cleaning of special characters, hyperlinks, and short words. The models

were examined on accuracy, precision, recall, and F1-score. It outperformed traditional machine learning methods. RoBERTa-base achieved the highest accuracy (99.66%) on the fake news dataset while Bart-large excelled with 98.83% accuracy on the English tweets dataset. The study highlighted the superior performance of conventional methods using LLM method and suggested future research directions in this domain. The paper had some flaws as the data was not available until the defense for this paper.

[22] Abburi et al. (2023) presented a system which can identify the writings between Artificial Intelligent and human-written text. The paper also suggested text-to-specific language models using datasets from the AuTexTification shared task. The datasets were in both English and Spanish. The research used a collective neural model by combining probabilities from various pre-trained large language models (LLMs) fed into a Traditional Machine Learning (TML) classifier. The first task was the model which achieved macro F1 scores of 0.733 (fifth place) for English and 0.649 (thirteenth place) for Spanish. In the case of the model attribution task, it was the most efficient with macro F1 scores of 0.625 and 0.65. LLMs such as BERT, DeBERTa, RoBERTa, and XLM-RoBERTa were fine-tuned for each subtask and language. The study demonstrated the effectiveness of the ensemble LLM approach in model attribution. They suggested further enhancements for the binary classification task.

[5] Pal et al.(2014) aimed to generate a dataset known as "Anubhuti" specially designed for analyzing the emotion of Bengali short stories in order to understand the flow of emotion of Bengali writers. Three linguistically qualified Bengali speakers were selected manually to annotate the datasets and the class of the given dataset was ensured as there was a "high-annotator agreement" among the three candidates. For the classification of emotions, a host of deep learning and machine learning models were applied to the dataset. Genres like Romance, Mystery, and Horror were included in the 159 Bengali Short Stories and the annotation of it was done by Bengali speakers with at least 15 years of formal education. Hence linguistic accuracy and relevance of the stories were ensured. A total of 30,000 annotated sentences were counted from the 159 Bengali short stories and a high annotator agreement of .807 based on Fleiss Kappa was achieved.

[4] Vargas et al. (2017), with a focus on long stories, had an objective to represent and extract narration-based data and information from input received from natural language. English-skilled translators translated all 21 Russian stories from the dataset. Based on the conclusion, the graphs were successfully extracted from a natural language which were also unannotated. This approach could be applied across various time periods as the story graphs accurately represented events in different story sections. There were discussions of using the graphs as inputs for generation of stories and maps. Furthermore, data was supplied on sentence and word counts noting that the corrected dataset included 914 sentences totaling 18,126 tokens.

[7] Berbatova (2019) researched natural language processing (NLP) techniques for content-based recommender systems for books on the GoodReads platform. The main goal of the research was to enhance book suggestions by leveraging textual data. Furthermore, the main objective of the paper was to improve the content based filtering methods by relying heavily on metadata and textual data. The dataset taken for the paper was from the GoodBooks where 10k datasets were for testing and training. 5,976,479 ratings made up the dataset and around 80 % of

the total was used for training while the remaining 20 % was utilized for testing. Since the dataset was biased towards a higher rating, a "like" was rated towards four stars or more. In conclusion, the paper was focused heavily on NLP techniques extraction of syntactic and lexical features and the best result obtained from the paper was a recall of 98 % using the publication year and language as textual and metadata features.

[6] The paper aims to introduce a new dataset of literary events and investigate the challenges of event detection within novels. It seeks to bridge the gap between existing event detection systems focused on contemporary news and the complexities of analyzing events in literary texts. The study compares models for predicting events and applies them to a corpus of novels to uncover insights into the role of events in shaping literary narratives and explore potential correlations between event distribution and authorial prestige. The main methodology involves annotating events within a dataset of 100 novels and applying various models for event detection. Featured models and neural models optimized for event trigger detection are compared. Specifically, the study explores the effectiveness of a bidirectional LSTM model with BERT token representations for predicting events in literary texts. The dataset consists of approximately 210,532 tokens from 100 literary works published before 1923, including canonical novels and popular genre fiction. Annotations for events were conducted based on guidelines adapted from the ACE 2005 annotation guidelines for events. The dataset contains 7,849 annotated events and is freely available for public use. The outcome of the paper reveals that the best-performing model, a bidirectional LSTM with BERT token representations, achieves an F1 score of 73.9 for event detection in literary texts. By applying this model to the corpus of novels split by prestige and popularity, the study identifies statistically significant differences in the distribution of events for novels categorized by prestige.

[8] Stamatatos et al.(2000) presented a method for rapidly and accurately identifying text genres and the method in use is inspired by an authorship attribution technique that uses the frequency of most common words(Burrows, 1992). However, unlike that approach, the frequency of the most common words was used across the entire written language. Stamatos et al.(2000) tested the method using a section of the Wall Street Journal corpus and found that the most frequent words from the British National Corpus (BNC), are more effective at distinguishing text genres than the aforementioned method. Additionally, the frequency of common punctuation marks significantly improves text categorization accuracy, especially when the training data is limited. In multivariate models, reducing the number of necessary parameters is crucial for achieving reliable results and lowering computational costs. Parts of the WSJ corpus which were classified into four low-level genres that can be grouped into two higher-level genres were used in the research. Stylistic consistency at both levels was captured by the automated classification model which was heavily reliant on discriminant analysis of the most frequent common words.

[23]Literary computing presents a variety of obstacles for Natural Language Processing (NLP) due to the unstructured nature of literature compared to typical datasets like Wikipedia articles. Literature often encompasses complex themes within long narratives, averaging over 200,000 words, unlike the 100-1,000 words found in conventional document modeling. Worsham et al. (2018) focused on identifying the correct literary genre, like Adventure or Love, based on the title and content of a literary work. A solution would benefit organizations providing literary services, such

as libraries and online bookstores, enhancing decision-making for library scientists and improving recommendation systems. Additionally, this research contributes to the broader understanding of literature, themes, and genre, ultimately aiding in the comprehension of full-length narratives with multiple themes which often coincide in a single narrative. After training, models were evaluated on test records using accuracy and F1-Measure. CNN-Kim had the highest performance among deep learning models at 75% accuracy. In contrast, the HAN model struggled, and the LSTM performed the worst, showing an improvement when trained on smaller text slices. XGBoost outperformed all models, achieving 84% accuracy and 81% F1-Measure. While results were lower than other studies, the research found better accuracy with the first 5,000 words of books and suggested that using a voting algorithm on unique subtexts could improve genre classification. [24]The problem of genre identification, especially distinguishing between fiction and non-fiction, has gained attention in recent years, though most studies focus on longer texts. There has been little research on classifying shorter texts, like paragraphs, which are common online and helpful for tasks like breaking news detection and opinion mining.Tools for genre identification in shorter texts can help filter misleading content on platforms like social media and product reviews. The dataset used for this research consists of paragraphs from the Brown corpus, specifically those containing 5 or 6 sentences. Additionally, preprocessing involved tagging and parsing with advanced tools, followed by extracting various linguistic features to create vector representations. The classification task employed logistic regression and two deep learning models, CNN and BERT. The study also explored the model's applicability in British English using the Baby BNC corpus, applying transfer learning from the Brown corpus. A binary classification task was set up to categorize paragraphs into fiction and non-fiction genres. Results showed that the 1D-CNN model outperformed logistic regression by 2% accuracy, while BERT achieved state-of-the-art accuracies of 97% and 98% for the Brown and Baby BNC corpora, respectively.However, deep learning models are still hard to interpret. In contrast, logistic regression revealed insights into genre differences, showing that fiction often has more character-level diversity and lower lexical density, along with significant variations in word order between genres. [25]Document classification in Bangla is challenging due to the scarcity of large corpora and the complexity of the language. Additionally, basic pre-processing resources like stemmers and part-of-speech taggers are limited for Bangla, making NLP tasks difficult.In this study, data from 2015 to early 2020 was collected from five major Bangladeshi news portals, resulting in around 640,000 news articles and titles categorized into ten different genres. The study used traditional machine learning, advanced neural networks, and attention-based models with various feature extraction techniques to compare their performance. Data was collected from five leading Bangladeshi news sources: Prothom Alo, Bangladesh Pratidin, Kaler Kantho, BBC Bangla, and BD-News24, spanning from 2015 to early 2020. This resulted in approximately 6,40,000 news articles and titles, which were extracted using Python's urllib and BeautifulSoup libraries to scrape the text from the news sites. The LSTM model achieved best out of all models for both news article and title classification, with F1-scores of 0.97 and 0.91, respectively. Attention-based models also performed similarly well. In contrast, the multi-layer perceptron had the lowest precision, recall, and F1-score. [26]This paper uses machine learning and deep learning models to classify Bengali novels by genre based on their titles and snippets.Two datasets were created: one

with titles across nine genres and another with snippets from three genres. Various models were tested, including logistic regression, SVM, random forests, RNN, LSTM, CNN, and BERT. BERT achieved the highest accuracy, reaching 90% for the snippet dataset and 77% for the title dataset. The dataset consists of 12,925 Bengali book titles across nine genres and 452 book excerpts in three genres, collected from Rokomari.com. Three classifiers—Random Forest, Logistic Regression, and Multinomial Naive Bayes—achieved the highest accuracy of 54% using the Title dataset for classical machine learning models. For the excerpt dataset, Multinomial Naive Bayes reached an accuracy of 81%. Afroze et al. (2024) proposed future plans include increasing the dataset size to enhance the performance of the approach used [27]NLP is a computational approach to analyzing text, excelling in tasks like Semantic Search, Machine Translation, and Sentiment Analysis. It's seen as a key future technology, though challenges remain for its compatibility with emerging tech. The paper covers its benefits, limitations, and research opportunities. NLP is a key research area in AI and computer science, focusing on enabling communication between humans and computers.NLP's development has gone through several phases: early stages in the 1950s, significant advances from the 1950s to 1970, slower progress from 1971 to 1993, and a revival from 1994 onwards due to increased computing power and the commercialization of the Internet. Key milestones include the introduction of neural networks for NLP by Yoshua Bengio in 2001, multitasking neural networks by Ronan Collobert in 2008, Word2Vec by Tomas Mikolov in 2013, and sequence-to-sequence learning by Ilya Sutskever in 2014. The field continues to evolve with applications spanning various disciplines, offering research opportunities in refining existing algorithms and developing new techniques for NLU and NLG. The most common and noticeable problem faced during the review was that most of the papers had small datasets which were not diversified enough to cover all intricacies of the annotated and unannotated natural language taken as input. Taking the papers above for example, it was seen that Biradar et al. (2019) [10] used only 4176 training images from Amazon, and a public IMDB review dataset of 50,000 entries was used by Cen at al.(2020) [13]. Similarly, other studies used datasets with limited scope, such as a fixed set of genres (Lee et al., 2002) [16] or specific platforms like Goodreads (Biradar et al., 2019; Berbatova, 2019) [7]. Overfitting and inaccurate accuracy is noticed during the outcome detection as seen in the paper by Vargas et al. (2017). [4] Moving on to the second point, preprocessing steps like uppercasing, lowercasing, punctuation removal, and stemming were done. However, there was a lack of certain features to capture hidden characteristics. For example, simple methods such as word-word co-occurrence matrices or basic TF-IDF may not be sufficient to capture complex patterns in the text, as highlighted in (Lee et al., 2002; Cen et al., 2020). [16] [13] Furthermore, Naive Bayes and Doc2Vec may not fully leverage deeper linguistic features, impacting the effectiveness of classification. Finally, the results of some of the models are not satisfactory which indicates issues of optimization and model selection. For instance, the LSTM model achieved an accuracy of 65%, and the CNN model 63% in genre prediction from book titles [14]. Additionally, there is often a lack of comparison with more advanced or alternative models that could potentially yield better results. The research typically sticks to a few traditional models without exploring a wider range of architectures or hybrid approaches that might improve the accuracy and robustness of the model.

Chapter 3

Work Plan

The Work Plan for this thesis consists of three main phases. The first phase contains topic selection, collection of other literary works related to the topic and performing a literature review. Also, that phase contained the models that were expecting to be used. The second phase of the thesis contains preparing the dataset while pre-processing the dataset for model evaluation, and evaluating the dataset using a model. Moreover, the plan is to to work with Bangla Short Stories and collect the dataset manually. All the models were used for finding accuracies, F1 scores, precision and recall. The third phase of this thesis will be fine-tuning the models so that accuracy and efficiency is increased. Among all the models, BanglaBERT showed the best results compared to the other models.The dataset will be collected in a large volume with more features. It will also include the improvement of the dataset Also, the results will be analyzed deeply to complete the thesis.

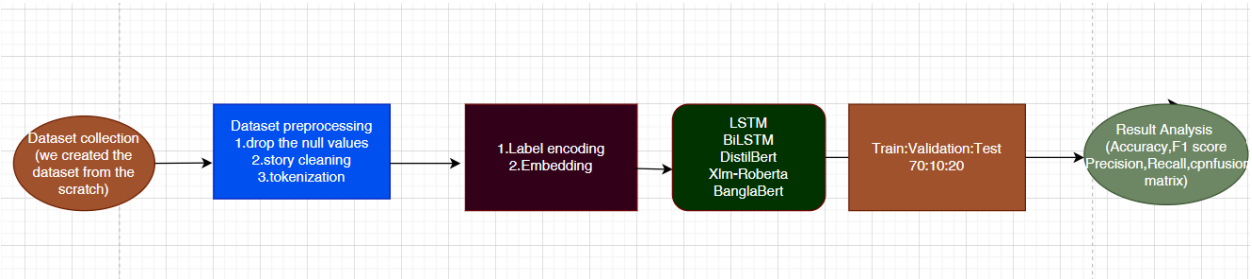

Figure 3.1: WorkPlan

In all the phases a report creation is necessary to show the findings and analysis of the thesis. The diagram gives us a structural representation of our work plan. The workflow consists of several steps to predict the genre. Each step indicates the flow of the process starting with the dataset collection to predicting the genre of the short stories. The initial step of the work plan is dataset collection. The collected data is raw and unprocessed. In order to prepare the data for analysis, data preprocessing is necessary. Data preprocessing helps to filter the data by removing duplicates and missing values to meet our research criteria. All the models are used for the preprocessed data and the data has been divided into three types which are training, validation and test data. The dataset is trained with training data and the rest is tested to enhance the efficiency of the model. The last step helps in predicting the genre by using the trained and tested data. Based on the result, further improvements will be made to increase the efficiency of our model.

Chapter 4

Methodology

4.1 Dataset Collection

The accurate genre detection of stories is massively dependent on the dataset collected. If data collection and implementation are flawed, it is difficult to carry out a meaningful study and therefore, determine the genre of the story provided. Our one of the source for the dataset is the dataset titled "Short Stories with Sentiment Analysis," which was compiled by researchers at Brunel University London [28] in English. From where we took selective short stories. Moreover, additional stories were collected from a host of websites with original stories from Bengali writers and authors. A diverse collection of Bangla short stories compiled through different websites like BdEbooks [29], Matrubharti [30], TagoreWeb [31]. The collection of such stories contained the author's name and the full story in Bangla. Furthermore, the collection of this portion was done manually. The stories were collected from a number of such verified websites. Legal and ethical boundaries were maintained in accordance with copyright regulations.

4.2 Translation and Summarisation

The collected dataset is divided into two datasets depending on language, as the dataset consists of both Bangla and English stories. At first, keeping the Bangla stories aside, the English stories were taken into consideration. If the English stories are translated into Bangla first, the dataset creates a lot of noise since the automation of translating large datasets into Bangla is still in an early state. As a result, the English datasets were summarized first and then translated into Bangla. The Bangla dataset is only summarized and then merged with the English translation manually. Summarisation was used instead of full stories as models like Bangla BERT, DistilBERT, etc, are quite limited due to the capacity of computers.

4.3 Dataset Description

The total number of story summaries is 5447. From the dataset mentioned above, we formed our own main dataset consisting of 2 columns. The two columns consisted of the summarisation of the story in Bangla and the genre. The translations are in Bengali, with each row containing a concise summary and its associated genre label.

The dataset consists of two main columns:
সারাংশ (Summary): A Bengali translation of the story's plot, condensed into a few sentences.
Genre: The story's primary genre, labeled in English (e.g., Fiction, Thriller, Romance, Adventure, Political, etc.).

4.4 Data Annotation and Guideline

4.4.1 Data Annotation Process

The genres were narrowed down to eight categories. A total of four native Bengali speakers manually annotated the data. The genre has been decided via a three-person voting system where each story gets three votes and the majority wins. The eight genre classifications consist of the following genres: fiction, thriller, war, science fiction, political, children's book, satire, and motivational.

Bangla Summary	1st Person Vote	2nd Person Vote	3rd Person Vote	Final Genre
একটি গ্রামে দুইজন মানুষ বাস করত, যাদের নাম ছিল ক্লাউস...ক্লাউস তখন নিজের দাদীকে মেরে ফেলে টাকা আয় করার চেষ্টা করে কিন্তু ব্যর্থ হয়।	Thriller	Thriller	War	Thriller
স্টুয়ার্ট ম্যাকগ্রেগর আফ্রিকার জঙ্গলে তার...হাচিসনের পিঠের ভয়াবহ উল্কি দ্বারা তাড়িত হন।	Thriller	Adventure	Adventure	Adventure
এই গল্পটি জো কলিন্স সম্পর্কে, একজন লম্বা, মানুষ, যিনি গৃহযুদ্ধে লড়াই করতে যান...গল্পটি আত্মত্যাগ, বিশ্বস্ততা এবং যুদ্ধের প্রবীণদের মুখোমুখি কঠিন বাস্তবতাগুলো তুলে ধরে।	Political	War	War	War
হ্যারি লেনক্স তার নিরানন্দ ছোট শহরে বিরক্ত ও অনাগ্রহী ছিল, যতক্ষণ না সে বেল মর্গান...কাজে আরও বেশি জড়িত হতে উদ্বুদ্ধ করে।	Fiction	Fiction	Adventure	Fiction

Table 4.1: Example of Three-Way Voting for Genre Selection

Furthermore, we have also verified that summaries are assigned the proper genre via an Asst. Professor of Bangla Literature. The main intention was to ensure that proper genre classification is given to the summaries and verification of our classification. In the case of a tie, we went to the specialist to help break the tie and specify the genre. For instance,

Bangla Summary	1st Person Vote	2nd Person Vote	3rd Person Vote	Final Genre
বর্ণনাকারী, টম বেইলি, তার রিভারমাউথ শহরে বেড়ে ওঠার জীবনের□উদ্বেগের সাথে অপেক্ষা করে।	Adventure	Political	Fiction	Adventure
নেক আগে, এক দরিদ্র কৃষক ও তার বৃদ্ধা মা শান্তিতে বাস করতেন□বৃদ্ধদের জ্ঞানকে সম্মন জানানো হল।	Children's Book	Fiction	Adventure	Children's Book

Table 4.2: Breaking the tie via the specialist's vote

4.4.2 Data Annotation Guidelines

Fiction: Fiction refers to literature that is created from the imagination rather than being based entirely on true events. Fiction may or may not include characters, settings, plots, and themes that are not based on reality. Imagination is the key here, and fantasy elements, along with literary fiction, play an important role. What separates fiction from science fiction is that in science fiction, elements of fantasy are highly influenced by technology and potential technological advancements in the future. Dataset samples like "গল্পটি ১৯৫০-এর দশকে কাল্পনিক বার্সলি শহরে মার্টিনমাস মেলার পটভূমিতে রচিত...প্রচেষ্টার মধ্যে সংঘাতকে অনুসন্ধান করে" with words like "কাল্পনিক" indicate that the sample in question is a fiction based sample since the word "কাল্পনিক" literally means imaginary.

Thriller: Thrillers are stories designed to provoke intense excitement, anxiety, and suspense. Elements of surprise, along with suspenseful cliffhangers, high-risk risk high-reward situations, and devastating consequences are often part of the criteria of thriller stories. Thrillers usually feature morally complex characters and incorporate plot twists and turns that challenge the reader's expectations. In order to be labeled a thriller genre sample, the sample must include fast-paced narratives with high stakes. In the sample "কটি গ্রামে দুইজন মানুষ...মেরে ফেলে টাকা আয় করার চেষ্টা করে কিন্তু ব্যর্থ হয়।' two brothers are engulfed in a bitter rivalry which results in the murder of their cattle and their grandmother. Hence, a high stake is introduced in this sample along with a murder mystery between two suspects - the brothers. Moreover, one of the brothers tries to acquire land through murder and deceitfulness. Hence, this helps in labelling a thriller.

War: For a story to be considered part of the war genre, it must center around military conflict or its direct consequences between regimes, countries, states, or kingdoms. The setting typically includes battlefields, military camps, or war-torn environments. Characters are often soldiers, commanders, or civilians affected by the chaos. These stories explore not only physical confrontations but also the emotional, psychological, and moral impact of war. In the criteria of war genre samples, words like civil war, bloodshed, battle, warfront, trenches, combat, tactics, squad, infantry,

artillery, shelling, etc were put emphasis on. Bangla words like 'গৃহযুদ্ধে", "লড়াই", "মারা", "সংগ্রাম" etc in the sample "এই গল্পটি জো কলিন্স সম্পর্কে, একজন লম্বা, স মেইনের মানুষ, যিনি গৃহযুদ্ধে...কঠিন বাস্তবতাগুলো তুলে ধরে।" met the criteria of war story and hence, upon such keywords, the labeling was done for war genre samples.

Science Fiction: The principles and presence of scientific advancements, inventions, future technologies, time travel, alien life, etc, make up the definition of the science fiction genre. Technological advancements and science, in combination with fiction, form science fiction genres. The term "এই শিশুটির মস্তিষ্কের বিশেষ ক্ষমতা মেশিনের চেয়েও অনেক বেশি কিছু" explains unusual condition of a child who is more advanced that artificial intelligence. Bangla words like "আধুনিক", "বিজ্ঞানী", "মহাকাশ", "টাইম মেশিন", "রোবট" etc usually gives the sign of sci-fi literature. Moreover, language related to science-fiction helps in differentiating between fantasy-based fiction and science-based fiction.

Political: Political fiction centers around the workings of power, government, and ideology. For a story to be political, it must engage with issues such as governance, civil rights, corruption, or revolution. For example: "তাকে ঘিরে চলে রাষ্ট্রীয় নজরদারি, জেরা ও বন্দিত্ব" term explains a political situation a person is going through. Characters typically include politicians, activists, or ordinary individuals caught in the tides of political change. Most common words found in samples with the political label were "রাজনীতি", "প্রতিবাদ", "রাষ্ট্র".

Children's Book: Children's books are written for kids, consisting of fun or meaningful stories using easy words and short sentences. For example, the term "হাতি টিপু প্রতিদিন জঙ্গলের গাছপালাকে রঙ করত যাতে সকলে খুশি থাকে" enhances the imagination and creativity of children. Words like "রঙ", "খুশি" provide a vibrance and lack violence. The stories often teach lessons about being kind, brave, or curious. Common topics include friends, family, good behaviour, and the colourful world. These books are usually short and have clear, happy endings with a moral.

Satire: Satire is a type of genre that uses humor and exaggeration to make fun of people, society, or rules. The main goal is not just to make the reader laugh, but to show what is wrong or silly about certain ideas or actions. For example: "মন্ত্রীর কুকুর অসুস্থ হওয়ায় পুরো হাসপাতালে মানুষের চিকিসা বন্ধ রেখে কুকুরের জন্য বিশেষ ওয়ার্ড খুলে দেওয়া হলো।" Here the term "মানুষের চিকিসা বন্ধ রেখে কুকুরের জন্য বিশেষ ওয়ার্ড" indicates exaggaration and silly action. Satire usually has hidden messages and plots that reflect real-life instances. Moreover, hypocrisy, absurdity, and flaws of certain laws and rules are made a mockery by writers through satire. For this dataset, all funny stories, including comedy, satire, etc, are grouped under satire or some form of satire.

Motivational: Stories written where the main characters go through perseverance and are encouraged not to give up are motivational stories. As the story goes on, the character changes in a good way, learning more about themselves and becoming stronger. The story often teaches values like bravery, hope, and never giving up. The overall feeling is positive, and the message is meant to make readers feel inspired and ready to do better in their own lives. The part "রহিম তখন চুপচাপ পরিশ্রম করে দেখিয়ে দিল" shows the dedication of a person which brings motivation to the readers. Words like "বারবার", "চেষ্টা", "জয়", "স্বপ্ন", "প্রমাণ" etc show up in motivational stories which implies

never giving up. Motivational stories also have a tendency to show the story of underdogs and how one can build oneself from nothing.

4.5 Exploratory Data Analysis

Our dataset consists of 5447 summarised stories across 12 annotated genres. Fiction with 1014 samples, Thriller with 777 samples, and Children's Book with 641 samples dominate the dataset. On the other hand, genres like Biography with 14 samples and Adventure with 119 samples are underrepresented. Therefore, Class imbalance is clearly visible as shown in the bar chart. Analysing the text length, although fiction and thriller genres were the most popular genres, children's books and science fiction had very high word counts in spite of their lower frequency. Furthermore, satire had the least number of words among the genres considered after pre-processing but ranked fourth in terms of genre frequency. These exploratory analyses helped to form decisions regarding preprocessing and balancing.

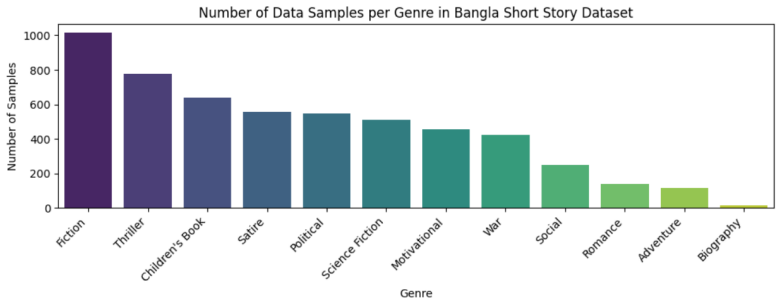

Figure 4.1: Number of samples in each class

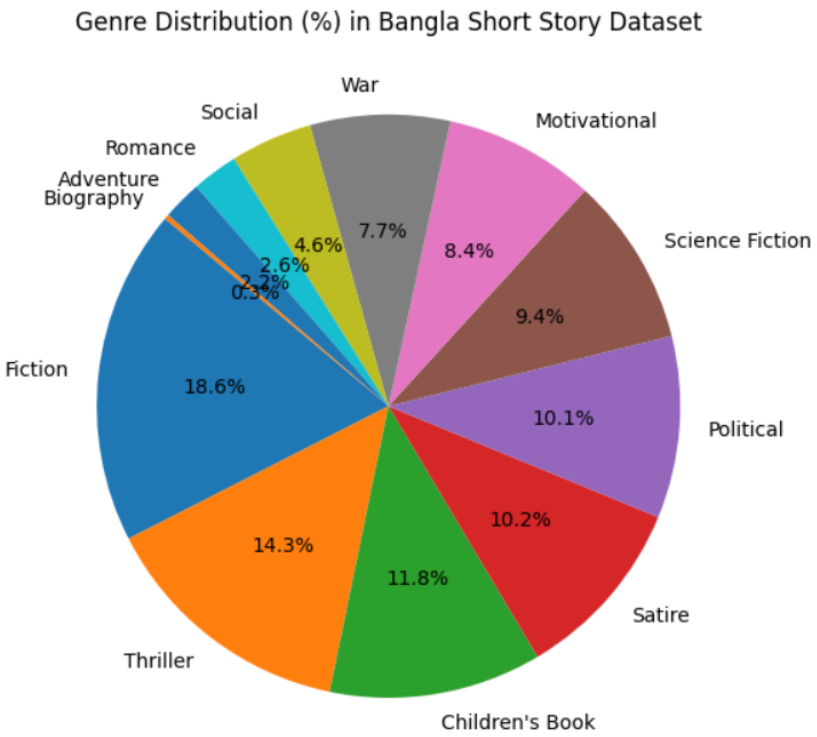

Figure 4.2: Percentage of samples in each class

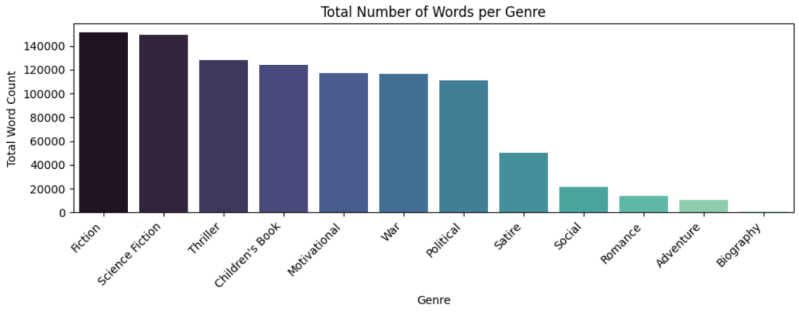

Figure 4.3: Total number of words per class

4.6 Data Preprocessing

4.6.1 Cleaning and Balancing the Dataset

Eight genres were taken, with the threshold at over 400 stories, to balance the dataset. As a result, stories with genres like adventure, romance, etc, were dropped. Comparing the ratio of the genres in the main dataset, it was found that the clipped genres fell under only 5% of the total sample size. A total of 4927 stories were annotated from the created dataset in the following frequency distribution shown in the bar graph.

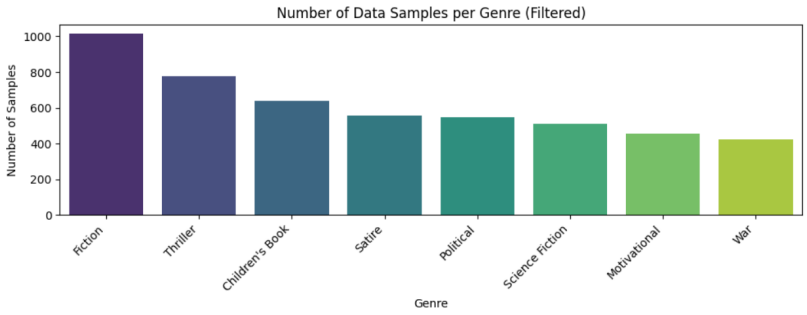

Figure 4.4: Number of data in each class after processing

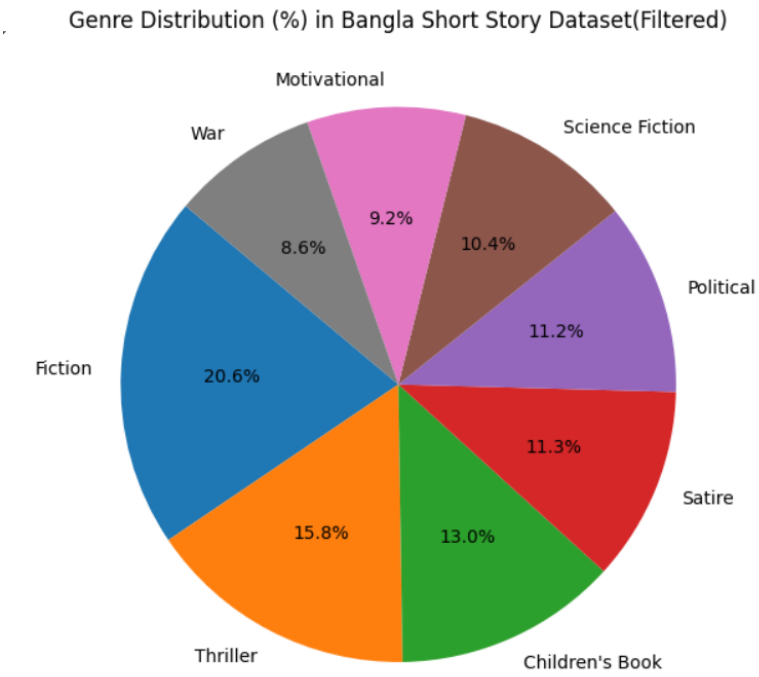

Figure 4.5: Percentage of data in each class after processing

4.6.2 Removal of Unwanted Characters

The cleaning begins by eliminating elements like emojis, special characters (like @, , *, +, =, etc.), and URLs. Only valid Bengali letters, numbers, and punctuation marks are kept. Additionally, line breaks (\n, \r) are stripped out to ensure consistency, and the main focus is put on linguistic content.

4.6.3 Removal of Redundant White Spaces

All unnecessary white spaces are eliminated from the text as a last cleaning step. This consists of cutting leading or trailing spaces and combining several spaces into one. By doing this, the summaries are made free of unusual spacing.

4.6.4 Genre Encoding

Machine learning models mainly operate on numerical data, and for this reason, genre labels were encoded into numerical values from 0 to 7 for our 8 classes. This encoding helps in training the models efficiently. The genres with encoded numerical labels are given in the table below.

Genre	Numerical Label
Children's Book	0
Fiction	1
Motivational	2
Political	3
Satire	4
Science Fiction	5
Thriller	6
War	7

Table 4.3: Genre and their Numerical Labels

4.6.5 Tokenization

Tokenization is the process of separating data and breaking it into smaller tokens. In our case, we used the pretrained tokenizer associated with each model. Tokenization is an important step in preprocessing. Tokenization would separate the statement "সিঁড়িতে পায়ের শব্দ শোনা যাচ্ছে" for instance into individual terms like ["সিঁ", "ড়িতে", "পায়ের", "শব্দ", "শোনা", "যা", "চ্ছে"] .Bangla sentences could be wrongly segmented due to a lack of enough tokenization, especially when compound characters are present. In a Wrong Tokenization method, the compound characters will be separated, but in a real meaning, it should not be separated. For a word like "বই", it shouldn't get separated into different tokens. Proper tokenization should have one token for this word. This underlines how important it is to use a tokenizer developed specifically for Bangla that is familiar with the character structure of the language and protects the integrity of words.

4.7 Data Spliting

The dataset is split into training, validation, and testing sets in a ratio of 70:10:20. The training set is used for training the model, whereas the validation set is used to tune hyperparameters. Finally, the testing set is used to evaluate the performance on unseen data. The split was stratified to maintain balanced class distribution across all sets.

Chapter 5

Model

5.1 Sequential Models

5.1.1 LSTM: Long Short Term Memory

A Recurrent Neural Network (RNN) has some limitations, for which Long-Short-Term Memory (LSTM) is used. RNN fails to depend on long-term dependencies. For tasks like genre detection, LSTM is well suited as it captures sequential data very well. LSTM can preserve context and can understand the sequence of even the Bangla language. Bangla Short Stories may contain many types of text. It may contain some contextual nuances, cultural references, and linguistic patterns. LSTM can capture these patterns and it can understand the progression and the tone of the story. It can analyze sequences of words and so can classify genres more accurately. It can maintain the flow of the narrative. Thus, LSTM is well suited for Bengali short story genre detection. LSTM contains memory cells so it doesn't have issues like vanishing gradient of RNN. RNN thus remembers relevant information and can use it for letter use. LSTM contains a conveyor belt-like structure called cell state, which runs through the entire sequence. It is controlled by gates and gates control which information to remember and which to forget. It can maintain the flow of the narrative.
The structure of LSTM is discussed below:
1.Input Layer: Each word or token of a sentence is taken as input in the input layer. They also contain an embedding layer generally. The embedding layer mainly converts words to numerical vectors.
2.Hidden Layers: LSTM contains some hidden layers that contain many gates performing various functions.

- Input Gate: This gate takes new information and controls which part of the new information it should store and remember.
- Forget Gate: This gate controls and decides which information it should forget from the past.
- Output Gate: This gate controls which information should pass and go to the next step.

3. Output Layer: After processing is done with the help of LSTM layers, output goes to the dense layer for class

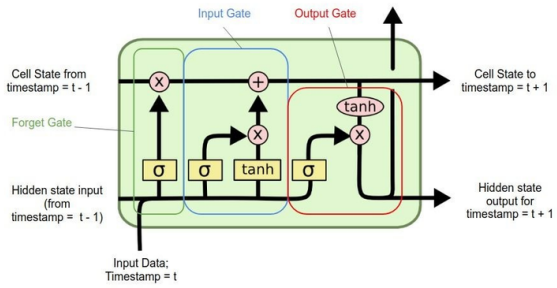

Figure 5.1: LSTM [32]

From the figure, the green box is the forget gate. It takes the previous hidden state and the current input data passes them through the help of sigmoid activation, and it gives an output value between 0 and 1. When the value comes closer to 0 it means to forget the information and when it comes closer to 1 it means to keep the information. The blue box is the input gate. Its sigmoid activation decides which value should be updated and the tanh function creates new values. The red box is the output gate which uses sigmoid activation to decide which part goes to cell state to output. Cell state is passed through tanh activation.

Data Word Embeddings

Words must be converted to numbers in order to be used as input in a neural network. Vectors are short lists of integers that are created using word embeddings. Word relationships and meaning are communicated through the arrangement of these numbers. Similar words group together in this space and the connections may be seen in the variations between the vectors.
In our model we use Bangla GloVe Word Embeddings [33]. For natural language processing uses, the GloVe Embedding layer is a crucial component of the process that converts integer-encoded word indices into dense vector representations. In our model, we have used Embedding with 39M tokens and it is a 300-dimensional vector.

Regularization and Optimization

We added dropout layers in between the LSTM layers to improve our LSTM model's performance and reduce overfitting. it can stop a portion of neurons at random which, during training. Furthermore, we used a custom attention layer. We used 10 epochs to train the model. By lessening the model's reliance on a specific group of features, this technique enables greater generalization. Additionally, early pausing was used to track the validation loss during the training procedure, If following a specific amount of epochs the loss does not improve, quit the process. This technique increases the utilization of resources while still maintaining the model's effectiveness on unidentified information. Our task in this research is basically to classify multi-class, So a categorical cross-entropy loss function is used to construct the model. We use the Adam optimizer to ensure effective convergence, where we put the learning rate of 0.0003.
To enhance the performance of our LSTM model and mitigate overfitting, we incorporated dropout layers of 0.5 between the LSTM layers, which randomly deactivate a percentage of neurons during training. This strategy fosters better generalization

by reducing the model's dependency on any specific set of features. We also employed early stopping to monitor the validation loss throughout the training process, allowing us to halt training when the loss fails to improve for a designated number of epochs. This technique not only helps preserve the model's effectiveness on unseen data but also optimizes resource usage. The model is compiled with the categorical cross-entropy loss function, appropriate for multi-class classification tasks. we use the Adam optimizer to ensure effective convergence, where we put the learning rate of 0.0003.

5.1.2 BiLSTM

Because the genre of different short stories is subtle, a model that can capture intricate language patterns and contextual relationships is necessary. We chose a Bidirectional Long Short-Term Memory (BiLSTM) neural network as the foundation of our method to tackle this problem. BiLSTMs are great in recognising the connections connecting words in a single sentence through taking account of both the context that comes before (forward) and after (reverse). Because sentimental language frequently relies on tiny contextual clues and word choices that can only be fully comprehended in the context of the adjacent text, this reversible analysis is crucial to genre detection. Bidirectional enables the model to take into account

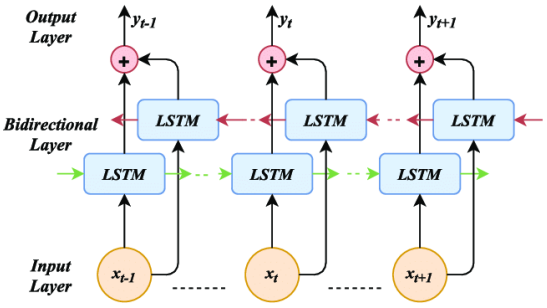

Figure 5.2: BiLSTM [34]

the significance and potential genre of a particular word when evaluating its previous and succeeding terms. Moreover, BiLSTMs have been configured to deal with sequence that have various widths by standard, which is a vital feature forour collection. The length of the dataset varies significantly, ranging from within 200 words on a maximum of 5000 words on average. The willingness of the BiLSTM is for managing this Variability ensures that there are both short-term relationships between words that are close.

Word Embedding

Words must be converted to numbers in order to be used as input in a neural network. Vectors are short lists of integers that are created using word embeddings. Word relationships and meaning are communicated through the arrangement of these numbers. Similar words group together in this space and the connections may be seen in the variations between the vectors.
We employed similar embedding as we did for LSTM which is Bangla GloVe Embedding [33] with 39M tokens and a 300 dimensional vector. GloVe's emphasis on global co-occurrence might be useful in identifying more extensive semantic links.

Network Architecture: A Hierarchical BiLSTM Model

the core component of a BiLSTM is the LSTM cell, which uses its internal memory to keep track of information over time, making it ideal for handling sequences like text.
In a BiLSTM, there are two LSTM cells operating simultaneously: one processes the input from the beginning to the end, while the other processes it from the end to the beginning. The outputs from both directions are combined, creating a complete representation of each word in its surrounding context.
The LSTM cells use a unique mechanism that allows them to decide what information to keep or discard, helping the model remember important details across a single sentence or multiple sentences.

Regularization and Optimization

The model uses dropout for regularization, setting 50% of neurons to zero during training to prevent overfitting. The Adam optimizer adjusts the learning rate using gradient information, combining features of momentum and RMSProp. This makes it suitable for handling sparse gradients in language tasks. The final softmax layer converts outputs to probabilities for multi-class classification.

5.2 Transformer Models

5.2.1 DistilBERT

The implemented model is based on DistilBert. The variant base-multilingual-cased is used, and it is implemented using the TFDistilBertForSequenceClassification class. The main objective of DistilBert is to reduce the computational cost but retain the capabilities of BERT's language understanding. 104 languages are supported by the multilingual cased version, including Bengali.

Architecture

The architecture of DistilBert is similar to RoBERTa, with fewer parameters, resulting in faster speed. It is a compressed version, and so instead of 12, it has 6 transformer encoder layers. It removes components like the token-type embeddings and next sentence prediction (NSP) objective. DistilBERTa is trained using a masked language modeling (MLM) objective, where the smaller model learns to mimic the behavior of the larger RoBERTa by matching its output representations. This results in a faster and lighter model that is well-suited for deployment in resource-constrained environments while maintaining competitive accuracy on many natural language processing tasks.

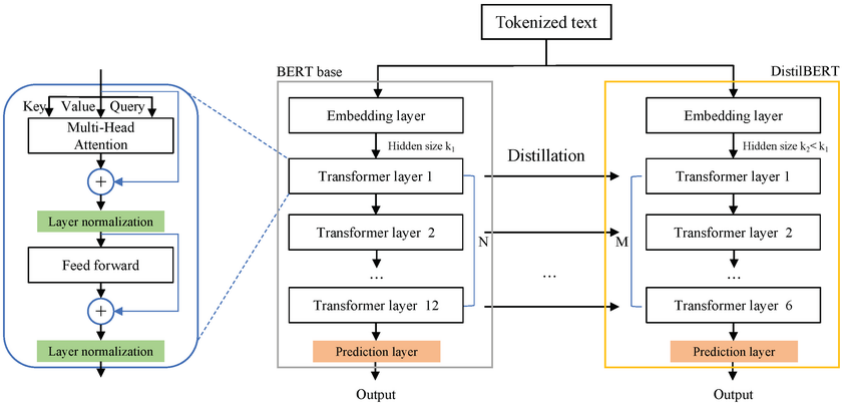

Figure 5.3: Architecture of DistillBERT [35]

Input Representation

Each input consists of a Bangla text sequence labeled with a genre. Inputs are preprocessed using the AutoTokenizer, which is included with the model that is pre-trained. The Tokenization contained Truncation and Padding, which is similar to BanglaBERT. For each sequence, `input_ids` and `attention_mask` were generated. These are passed into the model along with the target labels, which are encoded.

Compilation and Optimization

The optimization is done using the Adam Optimizer. The learning rate of the optimizer is 3e-5, and with an epsilon of 1e-08, along with gradient clipping. Batch size is set to 16. Sparse Categorical Crossentropy is preferred due to 8-class classification with integer labels as 0,1,2, respectively. A TensorFlow-compatible DistilBERT model is initialized with a classification head suited for multi-class prediction, determined by the number of unique genre labels.

5.2.2 XLM-RoBERTa-Base

Another model implemented is XLM-RoBERTa-base which has similar characteristics like BanglaBERT. BanglaBERT is trained specially on Bangla but XLM-RoBERTa model is pre-trained on 2.5TB of filtered CommonCrawl data of multiple languages which includes Bangla as well. The XLM-RoBERTa-base is initialized using the TFXLMRobertaForSequenceClassification class.

Architecture

The architecture of XLM-RoBERTa-base is similar to BanglaBERT with some tweakings in hyperparameters. It is using only the masked language modeling (MLM) objective without next sentence prediction. It is composed of 12 transformer encoder layers where each layer has 12 attention heads and a hidden size of 768 which makes it a total of around 270 million parameters. The model uses a SentencePiece tokenizer with a 250,002 token vocabulary and byte-level encoding, allowing it to process diverse languages, including low-resource ones like Bangla. Pretrained on

the massive CC100 dataset (2.5TB), XLM-R achieves strong performance on cross-lingual tasks by leveraging dynamic masking and a deep contextual understanding of multilingual text.

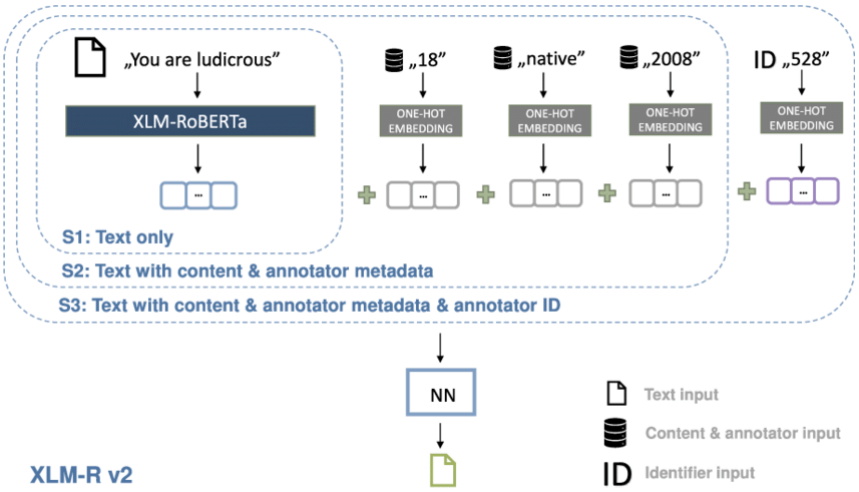

Figure 5.4: Architecture of RoBERTa [36]

Data Processing

The dataset is tokenized using the AutoTokenizer from csebuetnlp/banglabert. The tokenizer performs truncation after 512 tokens. Padding is also done at a fixed length. The tokenized data is converted to TensorFlow format using the `model.prepare_tf_dataset`

Hyperparameters

Adam optimizer is used having a learning rate of 1e-5 and with a gradient clip norm of 1.0. The loss function was SparseCategoricalCrossentropy. Max sequence length is 512, and also contains a gradient clip norm of 1.

5.2.3 BanglaBERT : Bidirectional Encoder Representations from Transformers

In order to focus on linguistic structure, cultural expressions, and contexts, it is important to choose a model that can detect short stories in Bengali. BanglaBERT is a language model based on transformers that is trained on Bangla dataset. The model is made in BERT architecture and known as one of the first BERT models for the Bangla language. Classifying genre is not only about finding attention words but also it necessitates the flow of sentence, contexts and techniques. The deep contextual understanding of BanglaBERT's allows it to adapt to these complications and classify genres with high accuracy.

Tokenization

AutoTokenizer is placed to structure input data effectively. The tokenization is divided into three tokens. [CLS], [SEP], and [PAD] work for sequences of models.

[CLS] is used at the beginning for classification tasks and as the overall representation of the sequence. [SEP] is used at the end of each sentence to maintain the coherence. [PAD] ensures uniform length of sequence and adds padding if necessary. The Auto Tokenizer with truncation and padding is enabled for a maximum sequence length of 512.

Architecture

BanglaBERT is a transformer model that features only encoder transformer architecture. It is made of 168M parameters in order to optimize for understanding Bangla text. It takes the most advantage of the multi headed self attention mechanism that helps to differentiate words and sub words in a sentence. Using the GeLU activation function, a position-wise feed-forward network refines feature extraction by introducing non-linearity.

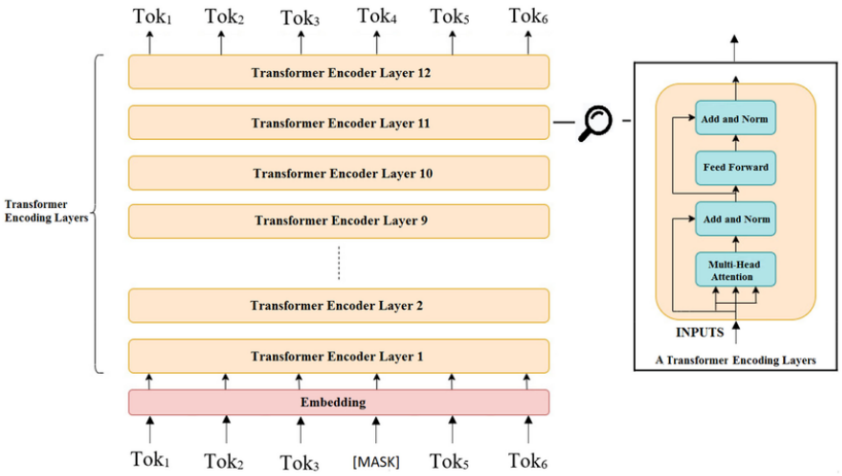

Figure 5.5: Architecture of BERT model [37]

BanglaBERT has 12 layers of encoder where each is composed of 768 hidden units to form a deep transformer model. Diverse linguistic patterns across Bangla literature and modern text is learnt effectively to understand genre indicators that allows it to classify with high accuracy.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Figure 5.6: Attention Score

Here, Q, K and V are the concatenation of query, key and value vector respectively.

Fine-Tuning

As we are classifying the genre, we need to add FFN over BanglaBERT. A fine-tuned version of BanglaBERT for sequence classification includes a Pooler that selects

the [CLS] token's representation. This token is passed through both a linear and non-linear activation function (tanh), helping BanglaBERT to understand complex relationships in Bangla text. A classification head is attached to the pooled output from the final transformer layer. This head maps the encoded representation to a probability distribution over the 8 genre classes, which is based on the 8 genres.

Optimization

The model is compiled using the Adam Optimizer. Learning rate is 3e-5 and with a gradient clip norm of 1.0. The loss function was SparseCategoricalCrossentropy. Max sequence length is 512, and also contains a gradient clip norm of 1 and batch size of 4. The model is then trained for 10 epochs on the converted training and validation datasets. Metrics like accuracy are tracked during training.

Chapter 6

Result

6.1 LSTM

In our work, we evaluated the lstm model performance using eight unique class.Our model achieved an F1 score of .56 after 10 epochs of training . While it is performing well on certain classes like 0 (children's book) and 5 (science fiction) with high recall and precision value. On the other hand it faces troubles classifying other classes like 1,3, and 7 which have lower precision and recall.It is representing that model is affected by the class distribution of dataset.

	Precision	Recall	F1 Score	Support
Children's Book	0.65	0.78	0.71	128
Fiction	0.41	0.56	0.47	198
Motivational	0.79	0.49	0.61	91
Political	0.49	0.47	0.48	89
Satire	0.45	0.45	0.45	111
Science Fiction	0.91	0.73	0.81	102
Thriller	0.57	0.54	0.55	155
War	0.53	0.33	0.41	84
accuracy			0.56	958
macro avg	0.60	0.54	0.56	958
weighted avg	0.58	0.56	0.56	958

Table 6.1: LSTM Classification Metrics for Eight classes

Moreover the confusion matrix represents multiple misclassifications between class 7,6 and 1. In additional the low recall values also reflects this as well .From the confusion matrix, class 1 (fiction) is misclassified with class 0 (children's book), 4 (satire), and 6 (thriller) multiple times.Similarly, class 5 (science fiction) is also missclassified for 41 times with class 1 (Fiction) which illustrates LSTM model finds it difficult to tell the difference between genres that are similar.

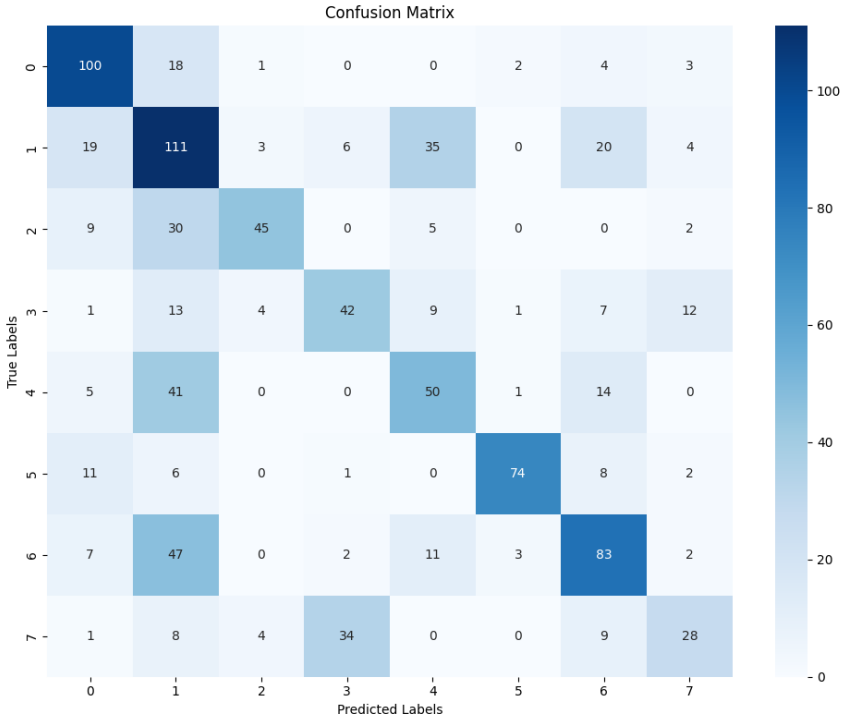

Figure 6.1: Confusion Matrix for LSTM model

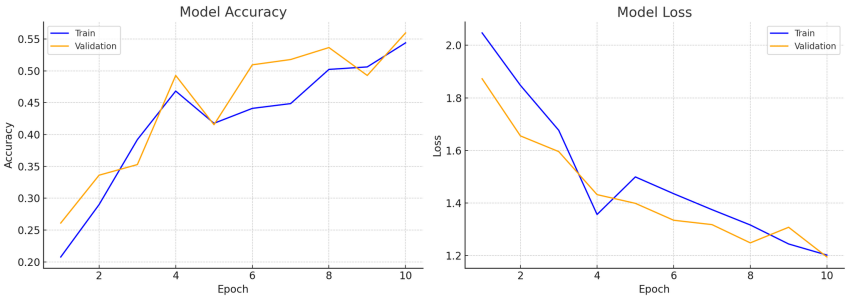

Figure 6.2: Training and Validation Accuracy and Loss Curves of LSTM

6.2 BiLSTM

In the research, alongside LSTM, Bi-LSTM was also used to correctly predict the genre of Bangla short stories. Performance was evaluated using eight genre classification. In the case of Bi-LSTM, epoch number was 10 and the model achieved an F1 score of 60%, which is comparatively better than LSTM's F1 score but still not as desired.

	Precision	Recall	F1 Score	Support
Children's Book	0.70	0.80	0.75	128
Fiction	0.49	0.42	0.45	198
Motivational	0.71	0.54	0.61	91
Political	0.58	0.64	0.61	89
Satire	0.47	0.59	0.52	111
Science Fiction	0.86	0.79	0.83	102
Thriller	0.54	0.65	0.59	155
War	0.73	0.44	0.55	84
accuracy			0.60	958
macro avg	0.63	0.61	0.61	958
weighted avg	0.61	0.60	0.60	958

Table 6.2: BILSTM Classification Metrics for Eight classes

Similar to that of LSTM, Bi-LSTM performed well in class 0 and class 5 which are the genres children's book and science fiction respectively. Although the other classes had over 50% on their F1 score, suggesting a potential trade-off between correctly identifying positive cases and minimizing false positives, discrepancy was seen in each genre's precision, recall and F1-score. Comparing the confusion matrix, for class 1 (Fiction), around 45% of the true label is correct and around another 45% , Bi-LSTM gets the wrong output genre consisting of class 4 (Satire) and Class 6 (Thriller). Similar incidents can be noticed in the case of war genre (class 7) with roughly 45% of the time getting true labels and more than the other half getting wrong predictions.

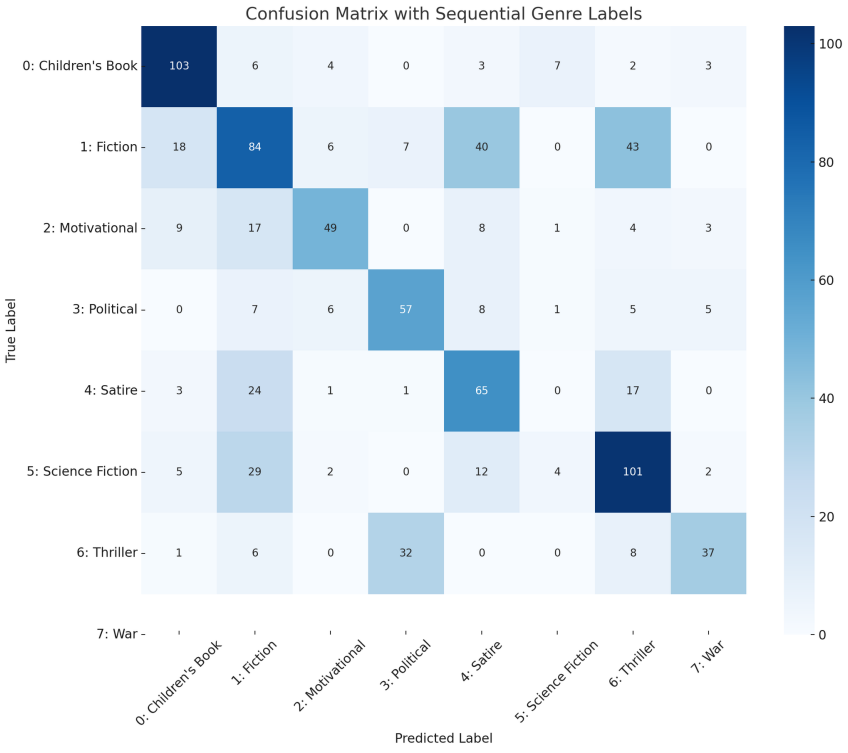

Figure 6.3: Confusion Matrix for BILSTM model

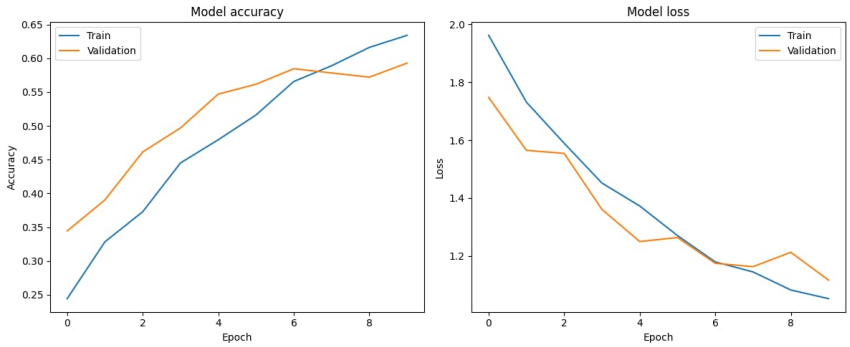

Figure 6.4: Training and Validation Accuracy and Loss Curves of BILSTM

6.3 DistilBERT

Moving on to transformers, DistilBERT shows improvement compared to the sequential models . Epoch number for DistilBert was 10 and the F1 score for distilBERT is 63% which is more accurate than the LSTM models. Science Fiction or Class 5 had the highest F1-score at 85% while class 0 ,2, and 7 had over or around 60% F1 score. Class 0,2 and 7 are respectively Children's Book, Motivational and War genre classes. Class 3 (Political) had the second highest recall score at 72% while also not being largely confused with all other genres. Furthermore, the discrepancy between each genre and their F1-score, precision and recall was comparatively less compared to the language transformation models.

	Precision	Recall	F1 Score	Support
Children's Book	0.73	0.66	0.70	128
Fiction	0.51	0.58	0.55	198
Motivational	0.63	0.56	0.59	91
Political	0.63	0.72	0.67	89
Satire	0.66	0.50	0.57	111
Science Fiction	0.84	0.85	0.85	102
Thriller	0.55	0.59	0.57	155
War	0.63	0.62	0.62	84
accuracy			0.63	958
macro avg	0.65	0.64	0.64	958
weighted avg	0.63	0.63	0.63	958

Table 6.3: DistilBERT Classification Metrics for Eight classes

Looking at the confusion matrix, all genres had a recall higher than 50%. Similar to that of Bi-LSTM, Fiction (Class 1) had the lowest recall and the other two lowest were class 2 and 4 which are motivational and satire respectively. Satire is highly confused with Fiction and Thriller as both of these classes were wrongly predicted in 40% of Satire's samples . Furthermore, a chunk of fiction classes were identified as political, satire and thriller classes.

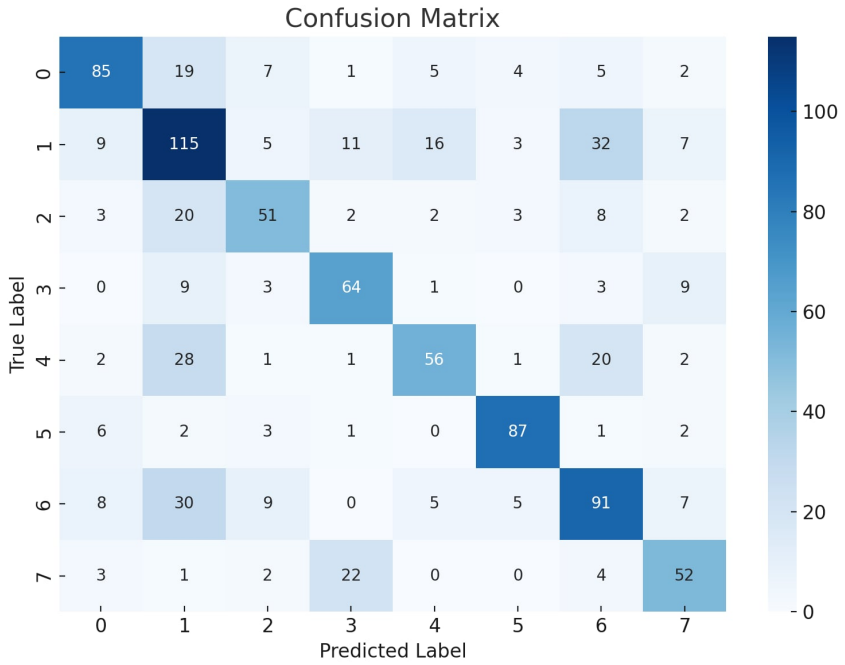

Figure 6.5: Confusion Matrix for DistilBERT model

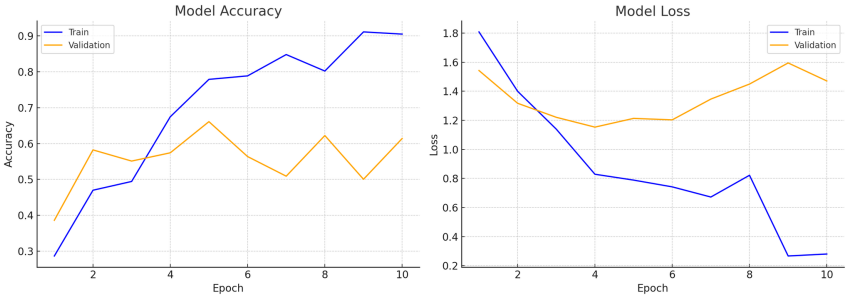

Figure 6.6: Training and Validation Accuracy and Loss Curves of DistilBERT

6.4 XLM-RoBERTa-Base

After DistilBERT, RoBERTa was used in the research and it further improved on certain aspects compared to DistilBERT. RoBERTa had the second highest F1-score among all the models that were tested at 67%. Similar to the former models used in the paper, science fiction or class 5 had the highest recall and F1-score. But in the case of precision, children's book class 0 had the highest precision at a rate of nearly 80%. Class 0 (Children's Book) and Class 5 (Sci-Fi) had the highest F1-score. Both Fiction and Thriller had the lowest F1- score which is a common theme between the RoBERTa and DistilBERT.

	Precision	Recall	F1 Score	Support
Children's Book	0.79	0.76	0.77	128
Fiction	0.54	0.67	0.60	198
Motivational	0.74	0.54	0.62	91
Political	0.71	0.71	0.71	89
Satire	0.75	0.53	0.62	111
Science Fiction	0.77	0.90	0.83	102
Thriller	0.61	0.58	0.60	155
War	0.67	0.69	0.68	84
accuracy			0.67	958
macro avg	0.70	0.67	0.68	958
weighted avg	0.68	0.67	0.67	958

Table 6.4: XLM-RoBERTa Classification Metrics for Eight classes

Examining the confusion matrix, for satire and motivation , around 47% was incorrectly predicted for both the classes. A huge portion of Satire was misclassified with Class 2 (Motivational) and Class 6 (Thriller). Class 2 is the most ambiguous as the model predicted at least 1 occurrence in the remaining 7 classes while identifying class 2 (Motivational). Also, class 5 or Science Fiction is the least ambiguous in this case.

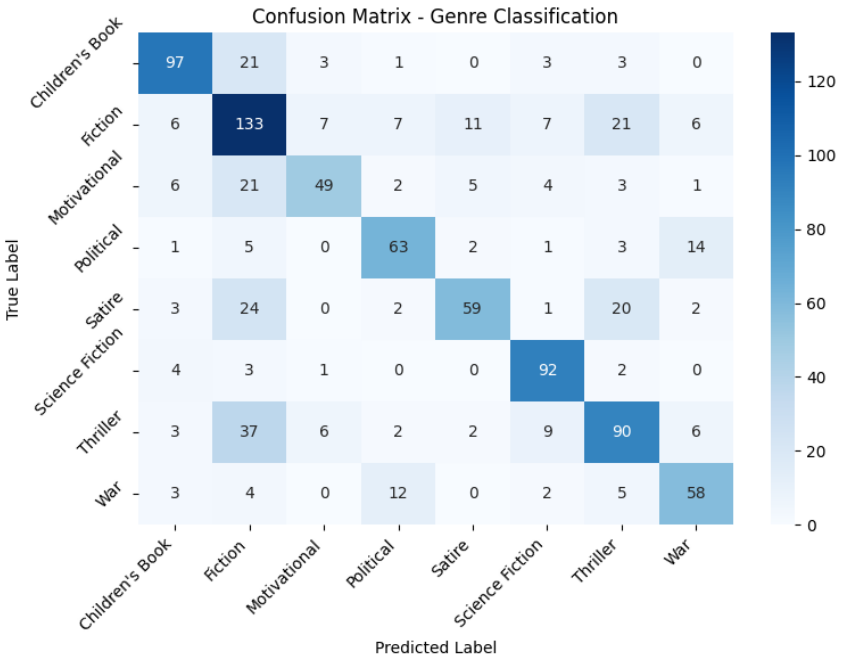

Figure 6.7: Confusion Matrix for XLM-RoBERTa model

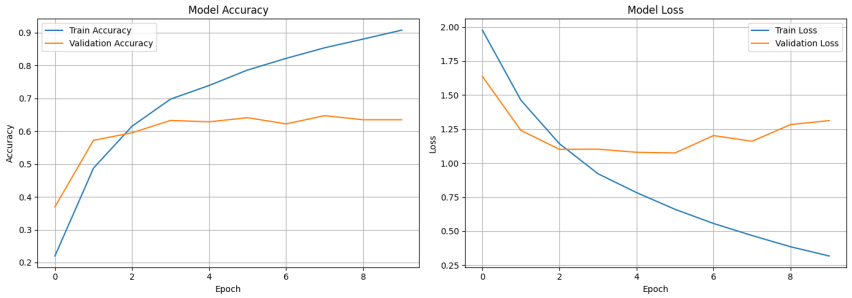

Figure 6.8: Training and Validation Accuracy and Loss Curves of XLM-RoBERTa

6.5 BanglaBERT

BanglaBERT has shown a remarkable improvement over the previously evaluated sequential architectures. The model is trained for 10 epochs and achieved an overall accuracy of 72%. This demonstrates more balanced and accurate performance compared to not only sequential models but also other variants of transformer models. Class 5 (Science Fiction) exhibited the highest F1-score at 85% which indicates a strong ability to classify texts with minimum confusion. Meanwhile, Class 1(Fiction) has been misclassified as Class 4 (Satire) and Class 6 (Thriller) as Class 1 (Fiction) 25 and 32 times respectively.

	Precision	Recall	F1 Score	Support
Children's Book	0.78	0.84	0.81	128
Fiction	0.63	0.67	0.65	198
Motivational	0.71	0.60	0.65	91
Political	0.78	0.66	0.72	89
Satire	0.67	0.77	0.71	111
Science Fiction	0.86	0.88	0.87	102
Thriller	0.77	0.59	0.67	155
War	0.67	0.85	0.75	84
accuracy			0.72	958
macro avg	0.73	0.73	0.73	958
weighted avg	0.73	0.72	0.72	958

Table 6.5: BanglaBERT Classification Metrics for Eight classes

In the classification report, the precision of Class 3 (Political) and Class 6 (Thriller) is 78% and 86% respectively. Class 5 (Science Fiction) has the highest precision, recall and F1 Score. Furthermore, Class 7 (War) has achieved the second highest recall score which is 72%. Class 0 (Childrens' book) has the second highest F1 score of 81%. Class 1 (Fiction) has the lowest score in precision and f1-score and Class 6 (Thriller) has the least for recall which is 59%.

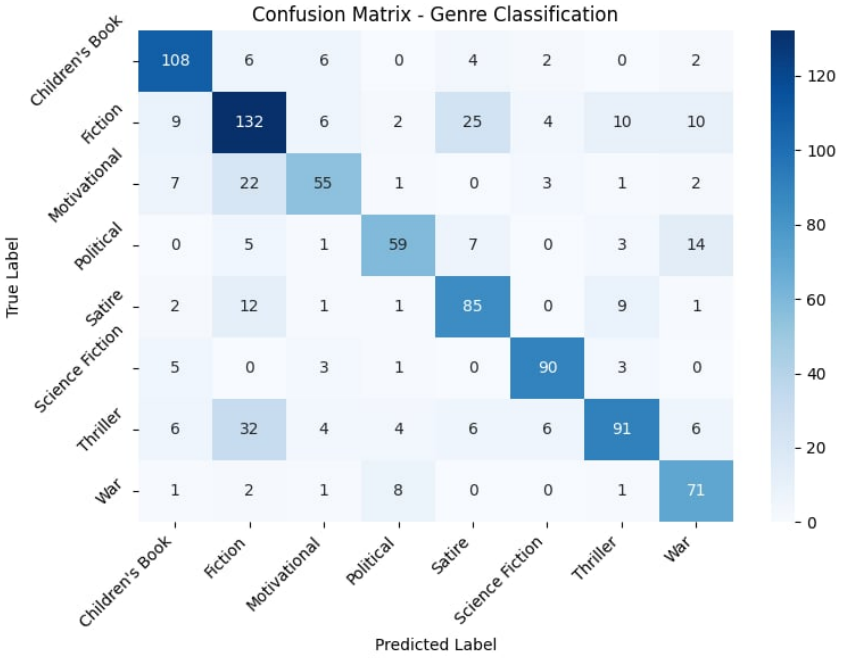

Figure 6.9: Confusion Matrix for BanglaBERT model

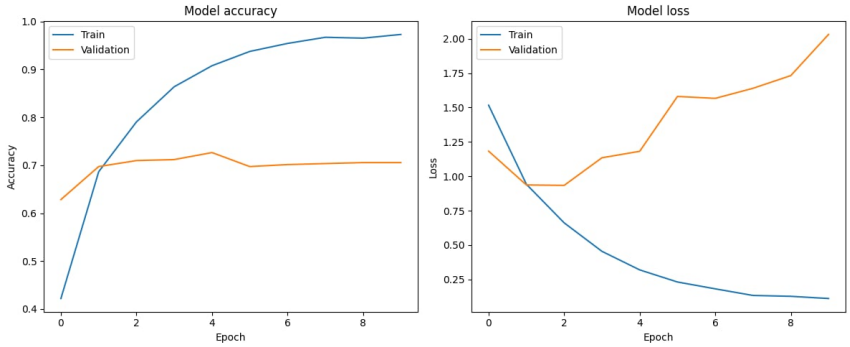

Figure 6.10: Training and Validation Accuracy and Loss Curves of BanglaBERT

The distribution of precision, recall and F1-scores shows that the precision of the model is 73%, recall is 72% and F1-score is 72%. In comparison with the other sequential and transformer models we have discussed, it is seen that the scores in each class have increased significantly.

6.6 Comparative Analysis and Model Selection

Model	Accuracy	Precision	Recall	F1 Score
LSTM	0.56	0.58	0.56	0.56
Bi-LSTM	0.60	0.61	0.60	0.60
DistilBERT	0.63	0.63	0.63	0.62
XLM-RoBERTa	0.67	0.68	0.67	0.66
BanglaBERT	0.72	0.73	0.72	0.71

Table 6.6: Classification Metrics for Eight classes for the Models

After comparing all the models, it is clear that BanglaBERT performs the best and it outperfroms others on important metrics. Our experiment shows the importance of transformer based models as they perform significantly better than sequential models for classification based problems. Thus it shows the importance of contextual information for classification based tasks. Ultimately, we selected BanglaBERT as the primary model for our further analysis because of its higher scores macro F1 score, which indicates its ability to classify across all 8 genres.

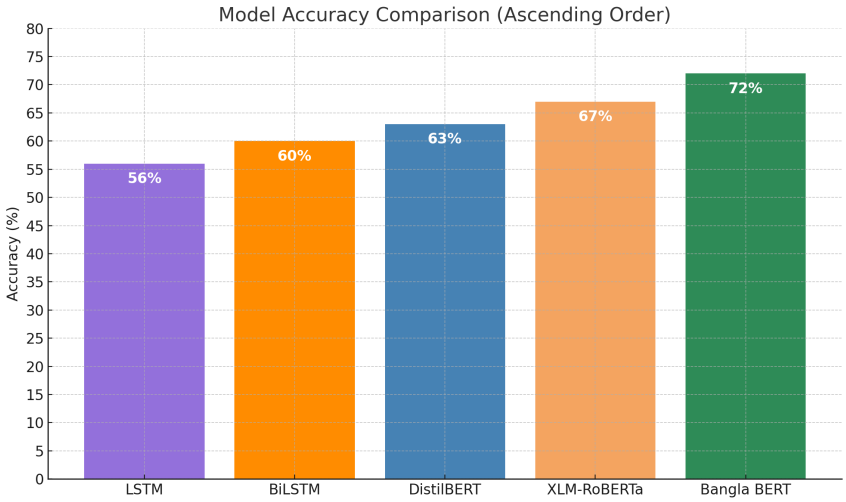

Figure 6.11: Accuracy Comparison Among all the Models

Chapter 7

Result Analysis

7.1 Ablation Study:Effect of Model Architecture and Training Parameters

To understand the significance of various design choices, we performed an ablation study. We tried to understand that effect of model architecture, pretraining and training hyperparameters like batch size and learning rate. This will help us identify the factors which contributes significantly to model performance.

7.1.1 BanglaBERT Vs Multilingual BERT

While comparing BanglaBERT and Multilingual BERT (mBERT) for genre classification for Bangla writing, BanglaBERT scored an accuracy of 72%, beating mBERT's 63%.This shows that BanglaBERT scored better since it was developed specifically on Bangla language data, while mBERT was developed on various languages concurrently and is not proficient in Bangla as BanglaBert.This result validates the notion that domain-specific pretraining provides significant advantages in tasks such as text classification, especially in diverse languages like Bangla.

Model	Accuracy	Precision	Recall	F1 Score
BanglaBERT	0.72	0.73	0.72	0.71
Multilingual BERT	0.63	0.65	0.63	0.63

Table 7.1: Comparision between BanglaBERT and Multilingual BERT

7.1.2 Pretrained Vs Non-Pretrained Models

We have conducted a comparative study by training the same architecture from scratch. The accuracy dropped significantly from 72% to 20% which shows the importance of pre-training for accuracy. We compared BanglaBERT which is initialized with pretrained weights against the same architecture with randomly initialized weights. The pretrained model achieved a validation accuracy of approximately 71%. This huge difference in accuracy indicates that the model has not learned any meaningful patterns and shows the importance of pre-training in BanglaBERT.

Model	Accuracy	Precision	Recall	F1 Score
Pretrained BanglaBERT	0.72	0.73	0.72	0.71
Nonpretrained BERT	0.20	0.04	0.21	0.07

Table 7.2: Comparision between Pretrained and Non-Pretrained Models

7.1.3 Batch Size Variation

A larger batch size gives a better approximation of the true gradient. So, even if the learning rate is unchanged, the updates become more stable and accurate which leads to better convergence.With batch size jump from 2 to 4, the optimizer moves more consistently toward a better minimum at the same learning rate. Larger batch sizes provide more stable and accurate estimates of the gradient during backpropagation.

Batch Size	Accuracy	Precision	Recall	F1 Score
4(BanglaBERT)	0.72	0.73	0.72	0.71
2(BanglaBERT)	0.70	0.72	0.70	0.69

Table 7.3: Comparision between variation of Batch Size of BanglaBERT

7.1.4 Learning Rate Variation on DistillBERT

Even though the batch size stayed the same, increasing the learning rate from 0.00003 to 0.00005 made the model take larger steps during training. As a result,the optimal weights were most probably missed during learning.Hence, the model didn't learn well as it missed on important features and accuracy dropped from 63% to 55%.

Learning Rate	Accuracy	Precision	Recall	F1 Score
3e-5	0.63	0.63	0.63	0.62
5e-5	0.55	0.57	0.55	0.55

Table 7.4: Comparision between the learning rate of DistillBERT

7.1.5 Freezing a Layer in DistillBERT

Freezing the DistilBERT backbone means the model couldn't update its language understanding. Therefore, only the classification head learned and since the backbone remained unchanged, DistilBERT could not adapt according to the required criteria. As a result, accuracy dropped from 63% to 28% as DistilBERT could not learn.The frozen backbone failed to capture task-specific features in Bangla and limited the classifier's ability to make accurate predictions. Here is the metrics after freezing a layer.

Accuracy	Precision	Recall	F1 Score
0.28	0.22	0.28	0.18

Table 7.5: Freezing a layer in DistillBERT

7.1.6 Hyperparameter Tuning in XLM-RoBERTa

We tried different combination of hypertuning to observe the difference on model performance.
In this observation, we tuned the learning rate to 5e-5 and dropout layer to 0.3 and thus we got an accuracy of 21%. This indicates underfitting

Learning Rate	Dropout	Accuracy
5e-5	0.3	0.21

Table 7.6: Tuning Learning Rate and Dropout Layer

In the the next observation, we changed the learning rate to 3e-5 and kept batch size to 16. Dropout was at default which is 0.1. This led to an accuracy of 61%

Learning Rate	Batch Size	Accuracy
3e-5	16	.61

Table 7.7: Tuning Learning Rate and Batch Size

In this observation, we tried changing the batch size from 16 to 8 and learning rate from 3e-5 to 1e-5. Whereas, if we keep the same learning rate while changing the batch size, we get different accuracy.

Learning Rate	Batch Size	Accuracy
1e-5	8	.62
1e-5	16	.67

Table 7.8: Tuning Learning Rate and Batch Size

7.2 Data Augmentation Leading to Overfitting and Poor Generalization

We applied data augmentation to compare between non-augmented(original) and augmented data. Our main goal is to evaluate whether augmentation improves the performance of the model. Two main methods were used in the augmentation procedure to expand the Bangla story dataset's volume.The first method was synonym substitutions, which consisted of substituting original words with their semantically linked equivalent words using a predefined vocabulary of Bangla words. This increased vocabulary diversity without changing the story's main message. By using data augmentation the volume of the dataset increased from 5k to around 9k. In the second method, the original story was rephrased without modifying the true genre label. Substitution example is given below:

Original Word	Synonym Word
চোখ	দৃষ্টি
ভূত	অতিপ্রাকৃত সত্তা
আকাশ	নভোমণ্ডল
মন্দ	অস□
জ্ঞান	বোধ
স্বাধীনতা	মুক্তি

Table 7.9: Example of Original and Synonym Word

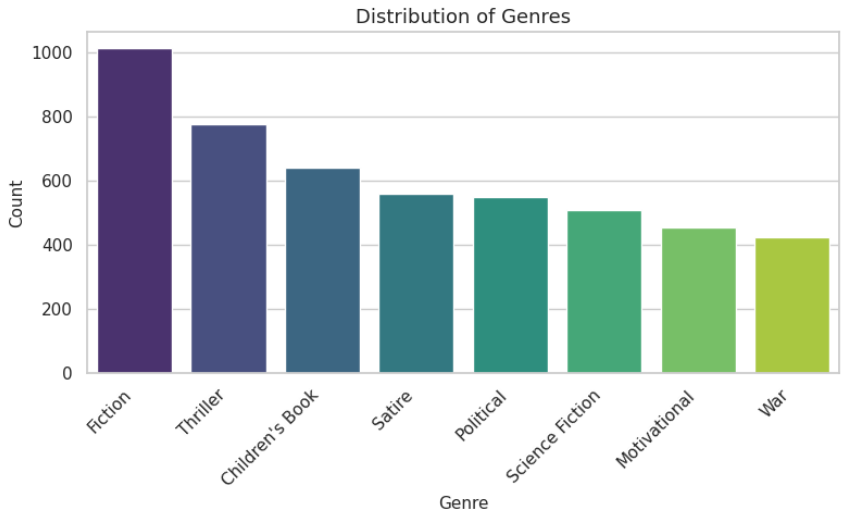

Figure 7.1: Before Augmentation

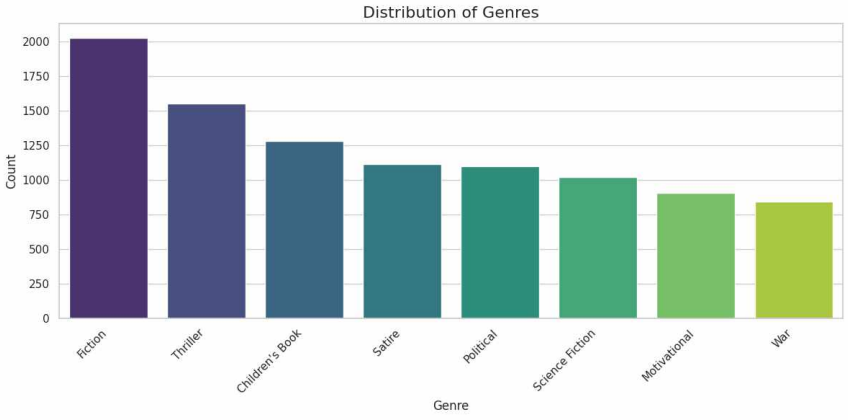

Figure 7.2: After Augmentation

Augmentation was implied on the dataset to proliferate dataset sample size from 5000 to around 9000. Mainly synonym based approach for limited words was used

in this case. As a result, overfitting was seen as accuracy of the BanglaBERT was around 89%. In both cases of training and test sets, overfitting was prevalent and hence the model had memorized the dataset. In conclusion, the augmented data quality was not good enough to prevent overfitting which led to the memorization of the dataset. keep To improve on data augmentation quality, proper synonym replacement and sentence shuffling along with back translation must be applied.

7.3 Result Analysis

We analysed few misclassified labels to identify the pattern or change of misclassification. We presented a few examples and analysed them to understand the decision making behavior of the model.

Table 7.10: Analysis of Classification and Misclassification

Bangla Summary	True Genre	Predicted Genre	Explanation
শার্লট এম ইয়ং-এর "হাউ ওয়ান ম্যান হ্যাজ সেভড এ হোস্ট" শিরোনামের উদ্দীপনাময় গল্পে, ব্যক্তিগত সাহস ও আত্মত্যাগের শক্তি ইতিহাস জুড়ে উজ্জ্বল হয়ে ওঠে, দেখায় কিভাবে একজন মানুষের অটুট নিষ্ঠা পুরো সেনাবাহিনী ও জাতির ভাগ্য পরিবর্তন করতে পারে। গল্পটি শুরু হয় খ্রিস্টপূর্ব ৫০৭ সালে প্রাচীন রোমে, যখন সদ্য মুক্ত শহরটি তার বিতাড়িত রাজাদের প্রত্যাবর্তনের মুখোমুখি হয়, যাদের সমর্থনে ছিলেন ইত্রুরিয়ান বাহিনীর নেতা লার্স পোরসেনা। শত্রুরা জানিকুলাম দুর্গ দখল করে দ্রুত রোমের দিকে অগ্রসর হলে, শহরটির টিকে থাকার জন্য শুরু হয় কঠিন সংগ্রাম।	Motivational(2)	War(7)	The model predicted ”war” instead of ”motivational” due to the strong presence of military context and historical battle imagery. Bangla words like "আত্মত্যাগ", "সাহস", "সেনাবাহিনী", "দুর্গ" mean sacrifice, bravery, army, and fort respectively, and are frequently associated with war-related texts. The references to Lars Porsena, ancient Rome, and Etruscan forces likely contributed to the model's classification.

Bangla Summary	True Genre	Predicted Genre	Explanation
একটি ছেলে তার মায়ের আর্থিক অবস্থার প্রতি অসন্তোষ অনুভব করে এবং ঘোড়দৌড়ের বিজয়ীদের পূর্বাভাস দেওয়ার এক অতিপ্রাকৃত উপায় আবিষ্কার করে। মায়ের ভালোবাসা পাওয়ার আকাঙ্ক্ষায় সে নিজেকে চূড়ান্ত পরিণতির দিকে ঠেলে দেয়। গল্পটি বস্তুবাদ এবং আবেগগত অবহেলার সমালোচনা..	Science Fiction(5)	Fiction(1)	The excerpt describes the boy's ability to predict race winners as an "অতিপ্রাকৃত উপায়," meaning supernatural powers, which aligns more with fiction than science fiction. Indicators of sci-fi like technological advancements, inventions, or technological use has not been mentioned in the excerpt, and more focus has been given on the social impact of the family. Hence, the model leans towards fiction rather than science fiction.
যখন লেখক প্রচুর বই পড়ে, সাক্ষা□কার নিয়ে এবং ঘটনাকে নিজস্ব বিচারবুদ্ধি সাপেক্ষে বিশ্লেষণ করেন, তখন বোঝাই যায় লেখক একটি উদ্দেশ্যকে সামনে রেখে এগুচ্ছেন। পুরো বইটা খুঁটিয়ে খুঁটিয়ে না পড়লে 'উদ্দেশ্য' নির্ণয় বড় কঠিন। মুজিব বাহিনী থেকে গণবাহিনীঃ ইতিহাসের পুনর্পাঠ "শিরোনামের ৬০০ পৃষ্ঠার তথ্যপূর্ণ বইটি লিখেছেন আলতাফ পারভেজ। লেখকের রাজনৈতিক একটি মতাদর্শ রয়েছে তার স্বীকারোক্তি দিয়ে নিজেই লিখেছেন তিনি জাসদ ছাত্রলীগ করতেন। চার যুবনেতার বিএলএফ মুক্তিযুদ্ধের মাঝামাঝিতে যার নাম মুজিব বাহিনী আর জাসদের...	Political(3)	War(7)	The true class is political, but the model predicted war because the excerpt contains several strong Bangla keywords like "মুক্তিযুদ্ধ", "বিএলএফ", "যুদ্ধের সময়", and "সশস্ত্র সংগঠন", which are closely associated with war-related themes and the setting is during a war time scenario. Even though the story is on the political ideologies of the writer, the model does not put as much emphasis on politically themed words like "রাজনৈতিক", "মতাদর্শ", and hence fails to predict the genre correctly.

Bangla Summary	True Genre	Predicted Genre	Explanation
গ্রামীণ জীবনের স্নিগ্ধ উষ্ণতায়, মাত্র ছয় বছর বয়সী ছোট সনি, অজান্তেই তার পরিবারের আধ্যাত্মিক যাত্রার কেন্দ্রবিন্দু ও দিকনির্দেশক হয়ে ওঠে। সব ধর্মের প্রতি উন্মুক্ত এক পরিবারে বেড়ে ওঠা সনির বাবা-মা মূলত একজন প্রেসবিটারিয়ান ও একজন মেথডিস্ট তাদের ছেলের ধর্মীয় ভবিষ্য□কে অবাধ রেখে দেন, যাতে সে নিজে যে বিশ্বাসে আকৃষ্ট হয়, সেটিই গ্রহণ করতে পারে। তারা তাকে বিভিন্ন গির্জায় নিয়ে যান মেথডিস্ট পুনর্জাগরণ, কঠোর প্রেসবিটারিয়ান উপদেশ, এমনকি নাটকীয় ব্যাপটিস্ট নদীতীরের বাপ্তিস্ম কিন্তু কোনোটিই সনির আত্মাকে নাড়া দে ...	Children's Book(0)	Fiction(1)	The excerpt discusses আধ্যাত্মিক যাত্রা, ধর্মীয় ভবিষ্য□, and includes terms like প্রেসবিটারিয়ান, মেথডিস্ট, and বাপ্তিস্ম concepts that may seem too complex for children's literature. The main intention of the story is through the lens of a six-year-old and his viewpoint on religion and spirituality. These topics may seem hard to fathom for most children, which is why the model wrongly predicted it to be a work of fiction.
দ্য সিগনাল ম্যান একটি ভৌতিক গল্প, যেখানে এক রেলওয়ে সিগনালম্যান তার পোস্টের কাছে রহস্যময় সতর্কবার্তা ও ভূতের আবির্ভাবে আতঙ্কিত হন। বর্ণনাকারী সিগনালম্যানের কাছে যান, যিনি তার অস্বস্তিকর অভিজ্ঞতা ভাগ করে নেন তিনি একটি ছায়াময় অবয়ব দেখেন, যা মর্মান্তিক রেল দুর্ঘটনার আগে উপস্থিত হয়। সিগনালম্যান এই অতিপ্রাকৃত দর্শনে গভীরভাবে বিচলিত হন এবং বিপর্যয় ঠেকাতে নিজেকে অসহায় মনে করেন। ভূতের বার্তা ও আসন্ন বিপদের অনুভূতিতে তিনি ক্রমশ মানসিকভাবে ভেঙে পড়েন। গল্পটি এগোতে থাকলে, বর্ণনাকারী জানতে পারেন যে স ...	Thriller(6)	Thriller(6)	The Signal-Man was correctly classified as a thriller because of suspense and psychological elements which are typical of the genre. Words like "ভৌতিক গল্প" set an eerie tone from the start, but the tension is built more through the sense of imminent danger. Phrases such as "রহস্যময় সতর্কবার্তা," "ছায়াময় অবয়ব," and "মর্মান্তিক রেল দুর্ঘটনা" suggest that the character is trapped in a sequence of foreboding events beyond his control. The signalman's emotional state expressed through "বিপর্যয় ঠেকাতে নিজেকে অসহায় মনে করেন" adds psychological depth, a key ingredient in thriller narratives. Moreover, there is a constant feeling of imminent danger through phrases like "আসন্ন বিপদের অনুভূতি,". Hence, the excerpt is correctly identified as a thriller due to its core elements.

Bangla Summary	True Genre	Predicted Genre	Explanation
অনেক দিন আগে, এক দয়ালু ও ন্যায়পরায়ণ মালিক বাস করতেন, যার অনেক দাস ছিল। অন্যান্য নিষ্ঠুর মালিকদের মতো নয়, তিনি তার দাসদের মঙ্গল নিয়ে গভীরভাবে চিন্তা করতেন তাদের ভালোভাবে খাওয়াতেন ও পরাতেন, তাদের শক্তির উপযোগী কাজ দিতেন, সদয়ভাবে কথা বলতেন এবং কোনো ক্ষোভ পোষণ করতেন না। দাসরাও তাকে ভালোবাসত এবং তার ন্যায়পরায়ণতা ও সহানুভূতির জন্য প্রশংসা করত, এমন একজন মালিকের সেবা করতে পেরে তারা সত্যিই নিজেদের ভাগ্যবান মনে করত। এই সম্প্রীতি ও সদিচ্ছা শয়তানকে ক্রুদ্ধ করল, সে এই শান্তি নষ্ট করার জন্য আলেব নাম ...	Motivational(2)	Motivational(2)	The model correctly classified the excerpt as motivational, as it emphasizes themes of kindness, justice, gratitude, and inspiration, which are central to motivational narratives. From the beginning, the description of the master as "দয়ালু ও ন্যায়পরায়ণ" sets a positive and uplifting tone. His caring for his slaves, feeding them, assigning work, speaking kindly, etc., all point towards a positive and uplifting attitude. Phrases like "মঙ্গল নিয়ে গভীরভাবে চিন্তা করতেন," "সদয়ভাবে কথা বলতেন," , "প্রশংসা করত" and "নিজেদের ভাগ্যবান মনে করত" indicate a grateful and positive attitude. Hence, the model correctly identifies the sample as motivational work.

Chapter 8

Conclusion

8.1 Future Work

For future work, a handful of sections can be improved upon. Primarily, the size of the dataset has to be increased. Roughly 5 thousand-plus story samples were used in the dataset, which is quite limited. Moreover, not all genres have an equal number of samples. Hence, improvement on the dataset must include not only increasing the number of stories and sample size but also increasing the sample size of the genres with the lowest number of samples. Secondly, augmentation was used to increase the size of the dataset to around 9000, which is quite low. Also, augmentation resulted in pverfitting. Hence, in the future, priority should be given to improving the augmentation process by ensuring proper synonym replacement, back-translation, and sentence shuffling. Thirdly, obstacles were identified while running the state-of-the-art language models. A lack of computational resources was noticed while running the dataset on transformer models, which limited our findings and made inefficient use of time. Therefore, to be more efficient and thorough, computers with higher-powered processors should be used for improvement. Finally, the number of genres must be expanded to ensure that most stories fall under all the classifications mentioned in the research work. Genres like biography, fable, comedy, romance, etc, should be included to make the model more robust, accurate, and diverse.

8.2 Conclusion

In this modern age of science, automation is in every sphere of our life, and so it is in literary analysis. Detecting the theme of a short story is often time-consuming manually, and thus we explored and addressed the challenges in automating the detection of a theme in a short story. Detecting the genre of a short story can be difficult, as the endings are sometimes abrupt. By using neural networks like LSTM and BI LSTM and transformer-based models like BERT, RoBERTa, and DistilBERT, we analyzed the architecture of each of these models to see which one would work better for our research. In the future, we have to collect more datasets if possible, and learn about more models so that we can practically implement more models to detect the genre of a short story and give more accuracy. Our research aims to accurately detect the genre of Bangla short stories, and by analyzing different models of NLP and processing the dataset, we can do so. Our research provides

a foundation that can help literary scholars, educators understand the genre and predict with full accuracy.

Bibliography

[1] D. T. Vassiliades, ?Greeks and Buddhism: Historical contacts in the development of a universal religion,? The Eastern Buddhist, jourvol 36, number 1/2, pages 134–183, 2004.

[2] D. Palomo, ?Chaucer, Cervantes, and the Birth of the Novel,? Mosaic: A Journal for the Interdisciplinary Study of Literature, jourvol 8, number 4, pages 61–72, 1975.

[3] R. KOPLEY, ?A Tale by Poe,? Bloom's Modern Critical Interpretations, page 173, 2009.

[4] J. Valls-Vargas, J. Zhu and S. Ontanón, ?Towards automatically extracting story graphs from natural language stories,? inWorkshops at the Thirty-First AAAI Conference on Artificial Intelligence 2017.

[5] A. Pal and B. Karn, ?Anubhuti–An annotated dataset for emotional analysis of Bengali short stories,? arXiv preprint arXiv:2010.03065, 2020.

[6] M. Sims, J. H. Park and D. Bamman, ?Literary event detection,? inProceedings of the 57th Annual Meeting of the Association for Computational Linguistics 2019, pages 3623–3634.

[7] M. Berbatova, ?Overview on NLP techniques for content-based recommender systems for books,? inProceedings of the Student Research Workshop Associated with RANLP 2019 2019, pages 55–61.

[8] E. Stamatatos, N. Fakotakis and G. Kokkinakis, ?Text genre detection using common word frequencies,? inCOLING 2000 Volume 2: The 18th International Conference on Computational Linguistics 2000.

[9] A. Sobkowicz, M. Kozłowski and P. Buczkowski, ?Reading Book by the Cover—Book Genre Detection Using Short Descriptions,? inMan-Machine Interactions 5 A. Gruca, T. Czachórski, K. Harezlak, S. Kozielski and A. Piotrowska, editors, Cham: Springer International Publishing, 2018, pages 439–448, ISBN: 978-3-319-67792-7.

[10] G. Biradar, R. JM, A. Varier and M. Sudhir, ?Classification of Book Genres using Book Cover and Title,? june 2019, pages 72–723. DOI: 10.1109/ICISGT44072.2019.00031.

[11] S. H. Jayady and H. Antong, ?Theme Identification using Machine Learning Techniques,? Journal of Integrated and Advanced Engineering (JIAE), jourvol 1, number 2, pages 123–134, 2021.

[12] M. Qorich and R. El Ouazzani, ?Text sentiment classification of Amazon reviews using word embeddings and convolutional neural networks,? The Journal of Supercomputing, jourvol 79, number 10, pages 11 029–11 054, 2023.

[13] P. Cen, K. Zhang and D. Zheng, ?Sentiment analysis using deep learning approach,? J. Artif. Intell, jourvol 2, number 1, pages 17–27, 2020.

[14] E. Ozsarfati, E. Sahin, C. J. Saul and A. Yilmaz, ?Book genre classification based on titles with comparative machine learning algorithms,? pages 14–20, 2019.

[15] B. Kessler, G. Nunberg and H. Schütze, ?Automatic detection of text genre,? arXiv preprint cmp-lg/9707002, 1997.

[16] Y.-B. Lee and S. H. Myaeng, ?Text genre classification with genre-revealing and subject-revealing features,? inProceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval 2002, pages 145–150.

[17] D. Agarwal, D. Vijay andothers, ?Genre classification using character networks,? in2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) IEEE, 2021, pages 216–222.

[18] P. Atanasova, ?A diagnostic study of explainability techniques for text classification,? inAccountable and Explainable Methods for Complex Reasoning over Text Springer, 2024, pages 155–187.

[19] J. M. Duarte and L. Berton, ?A review of semi-supervised learning for text classification,? Artificial Intelligence Review, jourvol 56, number 9, pages 9401–9469, 2023.

[20] D. Wilmot and F. Keller, ?Modelling suspense in short stories as uncertainty reduction over neural representation,? arXiv preprint arXiv:2004.14905, 2020.

[21] R. Qasim, W. H. Bangyal, M. A. Alqarni and A. Ali Almazroi, ?A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification,? Journal of healthcare engineering, jourvol 2022, number 1, page 3 498 123, 2022.

[22] H. Abburi, M. Suesserman, N. Pudota, B. Veeramani, E. Bowen and S. Bhattacharya, ?Generative ai text classification using ensemble llm approaches,? arXiv preprint arXiv:2309.07755, 2023.

[23] J. Worsham and J. Kalita, ?Genre identification and the compositional effect of genre in literature,? inProceedings of the 27th international conference on computational linguistics 2018, pages 1963–1973.

[24] A. Kazmi, S. Ranjan, A. Sharma and R. Rajkumar, ?Linguistically Motivated Features for Classifying Shorter Text into Fiction and Non-Fiction Genre,? inProceedings of the 29th International Conference on Computational Linguistics 2022, pages 922–937.

[25] R. Rahman, ?A benchmark study on machine learning methods using several feature extraction techniques for news genre detection from bangla news articles & titles,? inProceedings of the 7th International Conference on Networking, Systems and Security 2020, pages 25–35.

[26] A. Afroze, K. Dutta, S. Sadik, S. Khanam, R. Rab and M. A. Rahim, ?Empirical Text Analysis for Identifying the Genres of Bengali Literary Work,? Journal of Advances in Information Technology, jourvol 15, number 5, 2024.

[27] A. A. Abro, M. S. H. Talpur and A. K. Jumani, ?Natural language processing challenges and issues: A literature review,? Gazi University Journal of Science, pages 1–1, 2023.

[28] B. U. London, 4000 stories with sentiment analysis dataset, e, march 2019. url: https://brunel.figshare.com/articles/dataset/4000_stories_with_sentiment_analysis_dataset/7712540.

[29] BDeBooks, Best Bangla Short Story PDF Collection - Bangla eBooks, en-US. url: https://bdebooks.com/genres/short-story/?fbclid=IwY2xjawK_2DxleHRuA2FlbQIxMABicmlkETFqaEZqc2xmY2RncnJCbDlJAR586mtp2BhVLf4U_mRXY8qmhZq_s6qB_u6f2YfrxqmXpSCRFoSi4mKQJHL2qw_aem_iuSKtQD_ukC_MRby4e24Qg.

[30] en. url: https://www.matrubharti.com/stories/bengali/short-stories?fbclid=IwZXh0bgNhZW0CMTEAAR6DeoJoIcZnX_dilPoe662IRa67Avvjubpwexd-WKCjSyg50xiHQ8eT8ie7bQ_aem_QDpwibjUASRh30_68AyqiA.

[31] url: https://tagoreweb.in/Stories?fbclid=IwY2xjawK_1xZleHRuA2FlbQIxMABicmlkETE0cjh_4WRQFoRQVEdg_ZL-Brltdfsm6YO28vojgEXrxVyvbv6fp2VmrM5g_aem_zSX_3h61MnDuAjUYmVTV-A.

[32] L. Zhang, X. Meng, G. Tian and Q. Duan, ?Stability assessment of forestry plant power system based on improved long short-term memory network,? Journal of Physics: Conference Series, jourvol 2703, page 012 037, february 2024. DOI: 10.1088/1742-6596/2703/1/012037.

[33] url: https://huggingface.co/sagorsarker/bangla-glove-vectors/blob/main/bn_glove.39M.300d.zip.

[34] K. Resiandi, Y. Murakami and A. H. Nasution, ?Neural Network-Based Bilingual Lexicon induction for Indonesian ethnic languages,? Applied Sciences, jourvol 13, number 15, page 8666, july 2023. DOI: 10.3390/app13158666. url: https://www.mdpi.com/2076-3417/13/15/8666.

[35] Quantpedia, ?Top Models for Natural Language Understanding (NLU) Usage,? en, Medium, july 2023. url: https://quantpedia.medium.com/top-models-for-natural-language-understanding-nlu-usage-380ecbfccb5a.

[36] J. Kocoń, A. Figas, M. Gruza, D. Puchalska, T. Kajdanowicz and P. Kazienko, ?Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach,? Information Processing Management, jourvol 58, page 102 643, september 2021. DOI: 10.1016/j.ipm.2021.102643.

[37] A. Noorian, A. Harounabadi and M. Hazratifard, ?A sequential neural recommendation system exploiting BERT and LSTM on social media posts,? Complex Intelligent Systems, jourvol 10, august 2023. DOI: 10.1007/s40747-023-01191-4.