Empirical Text Analysis for Identifying the Genres of Bengali Literary Work

Ayesha Afroze, Kishowloy Dutta, Sadman Sadik, Sadia Khanam, Raqeebir Rab *, and Mohammad Asifur Rahim

Department of Computer Science and Engineering, Ahsanullah University of Science and Technology (AUST), Dhaka, Bangladesh
Email: ayeshaafrozeaust@gmail.com (A.A.); kishowloydatta016@gmail.com (K.D.); sadmansadikhasan@gmail.com (S.S.); sadiakhanamarni111@gmail.com (S.K.); raqeebir.cse@aust.edu (R.R.); mohammadasifurrahim@gmail.com (M.A.R.)
*Corresponding author

***Abstract*—Digital books and internet retailers are growing in popularity daily. Different individuals prefer various genres of literature. Categorizing genres facilitates the discovery of books that match a reader's tastes. The assortment is the process of categorizing or genre-classifying a book. In this paper, we categorize books by genre using a variety of traditional machine learning and deep learning models based on book titles and snippets. Such work exists for books in other languages but has not yet been completed for Bengali novels. We have developed two types of datasets as a result of data collection for this research. One dataset includes the titles of Bengali novels across nine genres, while the other includes book snippets from three genres. For classification, we have employed logistic regression, Support Vector Machines (SVM), random forest classifiers, decision trees, Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and Bidirectional Encoder Representations from Transformers (BERT). Among all the models, BERT has the highest performance for both datasets, with 90% accuracy for the book excerpt dataset and 77% accuracy for the book Title dataset. With the exception of BERT, traditional machine learning models performed better in the Snippets dataset, whereas deep learning models performed better in the Titles dataset. Due to the quantity and the number of words present in the dataset, the performance varied.**

***Keywords*—genre, Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), Bidirectional Encoder Representations from Transformers (BERT), Support Vector Machines (SVM), Natural Language Processing, Book Snippets, Recurrent Neural Networks (RNN)**

I. Introduction

Books are a significant part of our lives, with reading being a popular pastime for many people. Reading books sparks critical thinking and self-awareness, vital for personal growth amidst a thriving global book market [1]. The global book market is expanding, with 30% of the internet population reading books daily, according to a country analysis by global market researchers [2].

Books have transformed from papyrus scrolls to digital e-books, mirroring the evolution of the market through technological advancements in history.

Since purchasing and reading have shifted to digital platforms, the automatic classification of books into various genres is crucial. A genre describes the types of novels that belong to a particular category. The various genres cater to distinct audiences to satisfy a variety of needs. A genre refers to the types of books that fall into that category [3]. They are categorized by style, tone, time period, target audience, and numerous other factors [4]. Various genres cater to different audiences, meeting a variety of demands.

According to an article in "The Daily Sun" published on 21 February 2023 [5], Bengali is the fifth most-spoken native language and the seventh most-spoken language by the total number of speakers in the world. According to the article, "The 100 Most Spoken Languages Worldwide", approximately 260 million people speak Bengali, making it the seventh most-spoken language in the world, with around 22 million native speakers [5].

A substantial number of these individuals engage in online novel reading. Our research paper aims to enhance the reading experience of this audience by employing classical machine learning and deep learning models to classify books into various genres. Research on the classification of genres in Bengali literature books is currently lacking. This paper addresses this gap by categorizing various Bengali literary works based on their titles and snippets.

While research works on genre classification existed for other languages [6–8], none specifically addressed Bengali literature. Faced with the absence of a suitable dataset, we undertook the task of creating one from scratch. Our dataset includes titles and excerpts from Bengali books of various genres. Rokomari.com [9] was selected as our data collection platform due to its extensive collection of Bengali books, allowing us to gather snippets. One dataset encompasses 12,925 Bengali book titles spanning nine distinct genres, while the other

Manuscript received October 24, 2023; revised November 30, 2023; accepted January 25, 2023; published May 16, 2024.

comprises 452 Bengali book excerpts categorized into three genres. In our research, we utilized various machine learning algorithms for classification, such as logistic regression, Support Vector Machine (SVM), Random Forest Classifiers (RFC), decision trees, Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and Bidirectional Encoder Representations from Transformers (BERT). BERT exhibited superior performance compared to all other models for both datasets, achieving an impressive accuracy rate of 90% for the book snippets dataset and 77% for the book Titles dataset. The findings suggest that conventional machine learning models generally outperformed the snippet dataset, except for BERT. On the other hand, deep learning models demonstrated superior performance on the Title dataset.

The structure of the paper is as follows: The introduction lays out the rationale and objectives guiding our research. Section II consists of literature reviews, providing a comprehensive background for our study Section III focuses on our generated dataset, and Section IV goes into great details of the models we developed. Section V analyzes the results of our models and makes some informative observations. Section V concludes with a thorough wrap-up that includes closing remarks, a conclusion, and a discussion on potential prospects for future research

II. Literature Review

In this section, we have examined academic literatures that are relevant to our research. A summary of previous studies researching genre categorization using various approaches can be found in this section.

A. Classification Based on Title

Ozsarfati *et al.* [10] introduced a methodology for book genre classification based on titles using different machine learning algorithms, including RNN, Gated Recurrent Units (GRU), LSTM, Bi-LSTM, CNN, and Naive Bayes. The dataset comprised 207,575 samples, with each title associated with one of 32 distinct genres. This dataset was sourced from Amazon's library. Notably, the LSTM model has the highest accuracy among the algorithms used. This superior performance is attributed to its capability to retain memory over long-term dependencies. Gupta *et al.* [7] proposed an automated genre classification of books using Machine Learning (ML) algorithms focusing on Natural Language Processing (NLP) and dimensionality reduction techniques. This method includes collecting a large amount of text data from books, cleaning and preprocessing the data, transforming tokens using WordNet and, applying Principal Component Analysis (PCA), and using AdaBoost for genre prediction. Labeled and unlabeled data were used in training. The AdaBoost classifier was used to improve the accuracy of the decision tree by reducing bias and variance. The model had 81.18% accuracy on labeled data and 92.88% after using unlabeled data, making the technique scalable for applications beyond book genres, such as predicting genres for news articles and blogs. Shiroya *et al.* [8] tried to classify books by genre using machine learning algorithms and text classification techniques using a customized data set. Two different datasets were used for experimental purposes. The first one is The CMU Book Summary dataset, extracted from Wikipedia and Freebase matched metadata such as author, title, and genre. The second dataset was created from data extracted from various websites containing books translated from Gujarati and Hindi into English. Data pre-processing includes cleaning and abstract cleaning. Feature extraction is performed using TfidfVectorizer, and machine learning algorithms are trained and evaluated. The first dataset had accuracy results of 2.68%, 9.53%, and 7.27% in K-Nearest Neighbor (KNN), Logistic Regression (LR), and SVM, respectively. The result of the second dataset was 45.45% accurate in both KNN and LR, while SVM accuracy was 54.54%. The results showed that SVM outperformed K-NN and LR in accuracy and processing speed. Finn *et al.* [11] showed ways of learning to classify documents according to genre. Two sorts of genre classification tasks are done: if an article is subjective or objective and if a review is positive or negative. Three distinct feature-sets—Bag-of-Words (BOW), Part-of-Speech statistics (POS), and Text Statistics (TS)—are explored as independent views of the dataset, with the C4.5 decision tree algorithm as the primary learning tool. BOW performs well within a single subject domain but struggles with domain transfer. POS excels in both single domain and domain transfer, particularly for subjectivity classification. No single feature set stands out as universally superior for both genre classification tasks. Kim *et al.* [12] showed how to examine the variations of prominent features in genre classification, which was performed on six classes of documents: academic monographs, books of fiction, business reports, minutes, periodicals, and thesis. Here, two types of data sets were used: RAGGED Dataset (I) and KRYS I Dataset (II). Three different elements were used here: style, image, and Rainbow. The SVM, NB, and Rainbow Forest methods were applied. NB was the best method, showing better accuracy for the image feature in both data sets. SVM and Random Forest are better for the style feature. Style RF showed the best overall recall rate. Ostendorff *et al.* [6] introduced a methodology for enhancing Bidirectional Encoder Representations from Transformers (BERT) by incorporating knowledge graph embeddings and additional metadata. This approach improved the accuracy of standard BERT models, with an increase of up to four percentage points. The study encompassed the analysis of four widely-used datasets. The input length was restricted to 300 tokens to optimize GPU memory consumption. A dedicated preprocessing phase was employed to generate non-text features. Subsequently, the three representations were concatenated and fed into a Multi-layer Perceptron (MLP) featuring two layers, each with 1024 units and a Rectified Linear Unit (RELU) activation function. The final classification was executed in the output layer, where each unit in the SoftMax output layer corresponded to a specific class label. Evaluation of the model's performance utilized a micro-averaged F1-score. Notably,

the setup incorporating BERT-German with metadata features and author embeddings (1) outperformed all other configurations, achieving an F1-score of 87.20 for Task A and 64.70 for Task B. When focusing solely on precision scores, the configuration of BERT-German with metadata features (2), excluding author embeddings, demonstrated the highest performance. The study's findings underscored the significance of incorporating task-specific information, such as author names and publication metadata, in significantly enhancing the classification process compared to a text-only strategy. Table I shows the summary of literature reviews related to genre classification-based book titles.

TABLE I. OVERVIEW OF THE RELEVANT WORKS BASED ON TITLE

Author	Dataset	Method	Findings
E. Ozsarfati *et al.* [10]	207,575 samples from Amazon's library and categorized into 32 genres.	Applied algorithms: RNN, GRU, LSTM, Bi-LSTM, CNN, and NB.	LSTM achieved highest accuracy (due to long-term memory).
S. Gupta *et al.* [7]	Dataset from The Project Gutenberg eBook of Encyclopedia of Needlework	ML algorithms, NLP, Wordnet, PCA, TF- IDF, AdaBoost.	AdaBoost improved Decision Tree. 81.1% accuracy on labeled data, improving to 92.885 with the inclusion of unlabeled data.
P. Shiroya *et al.* [8]	CMU Book Summary dataset and a custom dataset from translated books.	KNN, SVM, LR and TfidfVectorizer for Feature extraction.	SVM outperformed KNN and LR in accuracy and processing speed for both datasets.
A. Finn *et al.* [11]	Movie review from MRQE site and restaurant reviews from Zagat survey site.	C4.5 Decision Tree algorithm, BOW, POS statistics & TS feature sets.	Classified articles as subjective / objective, positive / negative.
Y. Kim *et al.* [12]	RAGGED Dataset (I) and KRYS I Dataset (II).	Genre classification on six document classes using SVM, NB, and Rainbow Forest methods.	NB exhibited superior accuracy for image features, while SVM and Random Forest excelled for style features. Style RF demonstrated the highest overall recall rate.
M. Ostendorff *et al.* [6]	Dataset of 20,784 German books.	Augmented BERT models with knowledge graph embeddings and additional metadata, MLP.	Enhanced BERT accuracy up to 4%. Highlighted the importance of integrating task-specific information in classification

B. Classification Based on Snippet

Battu *et al.* [13] predicted movie genres using ratings and synopses in multiple languages, including Hindi, Telugu, Tamil, etc. They mined data from seven websites and pre-processed the data to group the genres into classes using ratings and genre as data points. They combined the data and divided it into two portions for training and testing, each containing 80% and 20% of the total data. Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) were used for character embedding. Support Vector Machine (SVM), random forest, and a hybrid model were used in the study.

Saputra *et al.* [14] introduced a text classification model aimed at identifying the genre of Indonesian films using synopses. The study incorporated CNN, RNN, and LSTM for analysis. In the pre-processing phase, various techniques were applied, including Repeat Character and Spell Normalization, Text Tokenization, Text Stemming, POS Tagging, Special Characters Removal, and Stop Words Removal. Feature extraction involved the utilization of TF-IDF and Bag-of-Words. The SVM classification algorithm and TF-IDF extraction demonstrated optimal accuracy and F1-scores following the training data (45%).

Ertugrul *et al.* [15] utilized Bi-LSTM to classify movies based on storyline summaries. They sampled data uniformly by genre in a document-level categorization challenge, obtaining 6,360 movies and 22,278 sentences. Bi-LSTM was applied to classify multiclass movie genres from plot summaries, encompassing class labels such as thriller, horror, comedy, and drama. They labeled their approach as a document-level technique, using the entire plot summary for training without sentence separation. They also trained a general RNN model using sentence- and document-level approaches. A comparative analysis with a Bi-LSTM model trained using a document-level approach showed that, especially with limited data, specifying the movie genre using phrases outperformed the overall plot summary of the recurrent neural network.

Portolese *et al.* [16] investigated using text-based traits derived from movie summaries for multilabel movie genre classification. Synopses were extracted from the Movie Database (TMDb) website, and all movies were categorized into 12 classes, including adventure, action, comedy, crime, drama, fantasy, horror, mystery, romance, science fiction, thriller, and war. Combinations were evaluated through 5-fold cross-validation, and final classification metrics were averaged across folds. The best experiment yielded averaged scores for the 12 genres in the dataset: precision of 57.61%, recall of 53.36%, and an F1-score of 54.80%.

While numerous studies have categorized genres across different disciplines and languages, there has been limited focus on the classification of genres in Bengali literature. This research proposes a methodology for classifying genres in Bengali literature based on book titles and snippets. The aim is to address the existing gap in scholarly literature and establish a structured framework for genre classification, leveraging the inherent characteristics found within book titles and snippets. This approach seeks to contribute to a more nuanced understanding and systematic categorization of Bengali literature's diverse genres. Table II shows the summary of literature reviews related to genre classification-based book snippets.

TABLE II. OVERVIEW OF THE RELEVANT WORKS BASED ON SNIPPET

Author	Dataset	Method	Findings
V. Battu *et al.* [13]	Data collected from seven different websites for seven languages	CNN and RNN for character embedding, SVM, Random Forest, and a hybrid model	Division of data: 80% training and 20% testing. CNN & RNN were used for character embedding. SVM, Random Forest, and a hybrid model were employed.
A. C. Saputra *et al.* [14]	Dataset from IMDB Indonesia Movies website	CNN, RNN, and LSTM, Various pre-processing techniques (Repeat Character, Spell Normalization, etc.), TF-IDF and BOW for Feature Extraction.	SVM with TF-IDF extraction achieved optimal accuracy and F1-scores after training data (45%).
A. M. Ertugrul *et al.* [15]	Movie Lens Dataset, Total of 6,360 movies and 22,278 sentences	Bi-LSTM applied for multiclass genre classification.	Document-level and sentence -level approaches are compared. Specification of movie genre using phrases outperformed using the overall plot summary.
G. Portolese *et al.* [16]	Extracted synopses from TMDb website and 12 movie classes	Four Classifier-Decision Tree, Extra Trees, Random Forest, MLP. TF-IDF, Embeddings	5-fold cross-validation was used for evaluating combinations. Averaged precision of 57.61%, recall of 53.36%, and F1-score of 54.80% were achieved.

III. DATA ACQUISITION AND PREPARATION

To facilitate data collection, we opted for an online platform due to its practicality and convenience. Rokomari.com [9] has been identified as the primary online platform for purchasing Bengali books. The platform was chosen for data collection because of its extensive library of 200,000 books classified into various genres. This classification greatly facilitated the process of organizing datasets. Another reason for choosing this platform was its ability to allow online reading of specific book sections, which created a dataset comprising book snippets. We generated two datasets for our research purposes. These datasets are as follows:

- Dataset of book titles.
- Dataset of book snippets.

Data Acquisition for Book Titles: The dataset comprises a collection of 12,925 Bengali book titles. These books encompass nine distinct genres: Humor and Entertainment, Biographies, Memories and Interviews, Philosophy, Law and Justice, History and Tradition, Self Help, Motivational and Meditation, Travel, Rhymes, Poems and Recitation, and Science Fiction.

TABLE III. DATA DISTRIBUTION AMONG DIFFERENT GENRES OF BOOK TITLE DATASET

Genre	No of Books
History and Tradition	2669
Self-Help, Motivational and Meditation	1666
Biographies, Memories, and Interviews	1396
Humor And Entertainment	1279
Travel	1274
Philosophy	1248
Rhymes, Poems, Recitation	1222
Sci-Fi	1097
Law and Justice	1074

The book titles were extracted from the website [9]. Beautiful Soup, a Python library, was employed for the purpose of data scraping. This library is specifically designed to extract data from HTML or XML files. Initially, we collected data on various genres individually and categorized them based on the genres available [9]. In this way, we gathered data about all nine of these genres. Subsequently, the entirety of the gathered data was consolidated and organized into a file adhering to the CSV (Comma-Separated Values) format. Table III shows data distribution among different genres of book Title datasets, and Fig. 1 is a sample from the Title dataset.

	A	B
1	Title	Genre
2	নেফারতিতি	Biographies, Memories & Interviews
3	বোদ্ধ ধর্ম দর্শন ভূমিকা	philosophy
4	প্রশাসনিক আইনের রূপরেখা	law and justice
5	একাত্তরের কার্টুন	history and tradition
6	বাঙালির মুক্তির ইতিহাস বঙ্গবন্ধু থেকে বঙ্গকন্যা	Biographies, Memories & Interviews
7	ইট দ্যাট ফ্রগ	Self-Help, Motivational and Meditation
8	মোল্লা নাসিরুদ্দীন	HumorAndEntertainment
9	ঢাকা জেলার ভূমিব্যবস্থা ও ভূমিরাজস্ব প্রশাসনিক ইতিহাস	history and tradition
10	বিজ্ঞানের দর্শন	philosophy
11	আপনার সাফল্য আপনার ক্যারিয়ার	Self-Help, Motivational and Meditation
12	দেশ দেশান্তর ২য় খণ্ড	travel
13	বাংলাদেশের ভূমি ব্যবস্থাপনা	law and justice
14	ডুব সাঁতার	Rhyms, poems, recitation
15	তাজউদ্দীন আহমদের ডায়েরি ১৯৪৭৪৮ ১ম খণ্ড	Biographies, Memories & Interviews
16	নির্বাচিত সহস্র প্রেমপ্রবচন	philosophy
17	রামরাজ্য ও মার্কসবাদ	philosophy
18	প্রাচীন সভ্যতার ইতিকথা	history and tradition
19	মুক্তিযুদ্ধে পুলিশের ভূমিকা ২য় খণ্ড	history and tradition
20	দেওয়ানী ও ফৌজদারী রিভিশন বিষয়ক আইন ১ম ২০০১	law and justice
21	মুক্তিযুদ্ধে সর্বহারা পার্টির ভূমিকা	history and tradition
22	ঢাকা টু কাশ্মীর	travel
23	বঙ্গবন্ধুকে নিয়ে ছড়া কাব্য	Rhyms, poems, recitation
24	উনিশ শতকে পূর্ববঙ্গের গরীবদের জীবন	history and tradition
25	দিব্যবাণী	philosophy
26	উলুবনে মুক্তা	HumorAndEntertainment
27	নিজেকে অতিক্রম করার কথা	Self-Help, Motivational and Meditation
28	একাত্তরের দিনগুলি	history and tradition
29	পাশ্চাত্য দর্শন আধুনিক ও সাম্পসারিত কাল নিউজ	philosophy
30	জল ও ডাঙার কবিতা	Rhyms, poems, recitation
31	তাছাওয়ফ তত্ত্ব বা তরিকত দর্পণ	philosophy
32	বাংলাদেশের রাজনৈতিক ইতিহাস	history and tradition
33	ধর্ম জীবন	Self-Help, Motivational and Meditation

Fig. 1. Sample from the Title dataset.

Data Acquisition for Book Snippets: The term "book snippets" refers to concise segments or excerpts extracted from books. To compile a collection of snippets, we took advantage of the opportunity to peruse a selection of pages from books exclusively available through Rokomari.com [9]. We read those pages and captured snapshots of them, from which we extracted the text. We used an online program called Image to Text to extract text from images. Subsequently, we categorized the snippets based on the categories of Portolese *et al.* [16]. The distribution of data among several genres of book Title datasets is presented in Table IV, and Fig. 2 is a sample from the Snippet dataset.

TABLE IV. DATA DISTRIBUTION AMONG DIFFERENT GENRES OF BOOK SNIPPET DATASET

Genre	No of Books
History	2669
Travel	1274
Sci-Fi	1097

	Title	Genre	Snippet
2	হায়েনার খাঁচায় অদম্য জীবন	History and Tradition	পূর্ব বাংলার ঘরে ঘরে উনসওরের গণ অভ্যুত্থান কখনো ভাবেনি পাক-ইরা ভেঙ্গে হবে খান খান। টুঙ্গিপাড়া থেকে উঠে এলো মজিবের কণ্ঠস্বর আকাশে বাতাসে উড়ছিলো জয় বাংলার ঝড়। মুক্তির যন্ত্রনা বুকে নিয়ে কাঁপছিল মানবতা বজ্রকণ্ঠে থেকে উড়ে এলো বাঙালিরা স্বাধীনতা । উত্তাল মার্চের সেই ভাষণেই ছিল, নির্দেশনা চূড়ান্ত যুদ্ধের জন্য তৈনী-হয়ে তুলেছিল ফণা । ঘরের ভেতরে শত্রুরাও ছিল, ছিল রাজাকার মুক্তির যোদ্ধারা, যুদ্ধজয়ের হয়েছিল একাকার । স্বাধীন হবেই দেশ, বুকে এক স্বপ্ন ছিল আঁকা লাল সবুজের দেশ হবে, পাবো নতুন পতাকা । একাত্তর-মার্চ ছিল আমাদের অহংকার মাস বাঙালিরা নয় মাস কেরছিল, স্বাধীনতার চাষ ।
3	বিজ্ঞান কল্প গল্প পিঁপড়াবতা	Science Fiction	সে বুঝতে পারে, চার্লস কোয়াল নামের পৃথিবীবাসী এই জ্যোতির্বিদ সবাইকে জানিয়ে দেবে, শনি ও ইউরেনাস এর মধ্যবর্তী কক্ষপথে একটা গ্রহ রয়েছে, যা এতদিনেও কেউ দেখেনি। টেরিস্টরিয়াল প্ল্যানেট বুধ, শুক্র, মঙ্গল এবং জায়ান্ট প্ল্যানেট বৃহস্পতি, শনি, ইউরেনাস, নেপচুন অনেক আগে আবিষ্কার করা হলেও এটি অনেকের চোখেই পড়েনি। পড়েনি মানে এরা চিরবাসীরা মানুষের চোখকে ফাঁকি দিতে পেরেছে। বৃহস্পতির চেয়েও বড় মাপের এই গ্রহ চিরণকে চিরবাসীর নেতা কিরণ লোকচক্ষুর অন্তরালে রেখে এতদিন যাবৎ পরিচালনা করে আসছেন। অন্যান্য সব গ্রহ আবর্তিত হয় নিজের ইচ্ছামতো। গ্রহবাসী প্রাণীর এতে কোনো নিয়ন্ত্রণ নেই। কিন্তু এই চিরণবাসী তাদের নেতা কিরণের নেতৃত্বে এক হাজার তিনশো বত্রিশ বছর ধরে পরিচালিত হয়ে আসছে। এরা ইচ্ছেমতো গ্রহের আকার-আকৃতি পরিবর্তন করতে পারে
	পর্যটন নগরী ঢাকা	Travel	ঢাকা থেকে প্রায় ২৫ কি. মি. দূরে সাভার থানার নবীনগরে জাতীয় স্মৃতিসৌধ অবস্থিত। এ স্থানটি নির্বাচনের অন্যতম কারণ ছিল দেশ স্বাধীন হওয়ার পর এখানে অনেক গণকবর আবিষ্কৃত হয়েছিল । বাংলাদেশের স্বাধীনতা যুদ্ধকে চিরস্মরণীয় করে রাখার জন্য এ সৌধ নির্মাণ করা হয়। অধিগ্রহণকৃত ৮৪

Fig. 2. Sample from the Snippet dataset.

Genre selection and labeling: One of the variables influencing the selection of the Rokomari.com [9] website was its genre labeling feature. The books on the website are systematically classified into numerous genres. After completing the data scraping process, we manually assigned labels to the obtained data. The selection of many literature genres was based on the rationale that each encompasses a substantial collection of over a thousand novels, making them suitable for both machine learning and deep learning algorithms.

IV. METHODOLOGY

This section discussed the approaches employed to identify genres in Bengali literature. In this study, two datasets were utilized. One dataset comprises a collection of book titles, while the second dataset consists of excerpts or snippets from books. The workflow remains the same across both datasets. Both datasets were analyzed using classical machine learning models and deep learning techniques. The classification models include:

- Classic Machine Learning Techniques—Decision Tree, Random Forest, K-Nearest Neighbor, Naive Bayes, Support Vector Machine, and Logistic Regression classifier.
- Deep Learning Techniques—Recurrent Neural Networks (RNN), Long Short-Term Memory Networks (LSTM), Convolution Neural Networks (CNN), and Bidirectional Encoder Representations from Transformers (BERT).

The sequential processes of the proposed methodology are illustrated in Fig. 3.

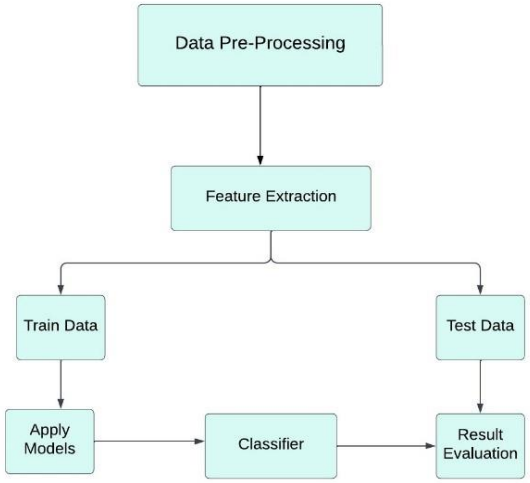

Fig. 3. Flowchart of proposed methodology.

A. Data Preprocessing

Preparing raw data to be acceptable for a machine learning model is known as data pre-processing, a crucial and initial phase in constructing a machine learning model. Data are often characterized by noise, missing values, and suboptimal formatting, rendering them unsuitable for the direct application of machine learning models. Data pre-processing is a crucial stage in the data analysis process, which involves cleaning and transforming raw data to improve its quality and prepare it for use in a machine learning model. This procedure plays a crucial role in improving the accuracy and efficacy of the model, as depicted in Fig. 4.

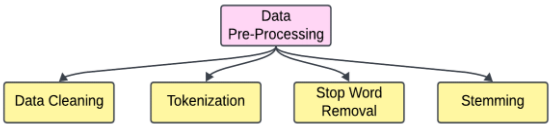

Fig. 4. Steps of data pre-processing.

1) Data cleaning

Data cleaning involves preparing unprocessed textual data for analysis by eliminating irrelevant content, managing missing data, and rectifying inconsistencies. We identified and eliminated replicated records within your dataset to prevent biasing analytical outcomes. To handle missing information, we removed rows containing null values. We used Regex for the removal of unnecessary characters, digits, emojis, and punctuations to ensure cleaner data.

A sample from the Snippet dataset is presented in Table V. In Table V, the “Snippet” column comprises the original data, while the “Snippet_After_Cleaning” column contains the cleaned data. The data originally included digits and punctuations, which were eliminated after the cleaning process.

TABLE V. DATA SAMPLE CONTAINING RAW DATA AND CLEANED DATA

Snippet	Snippet_After_Cleaning
পুলগুকসা প্রথম নির্মিত হয় ৫২৮ খ্রিস্টাব্দে রাজা পোপহাং কর্তৃক। পরবর্তীতে রাজা কিয়ংদক-এর শাসনামলে (৭৫১ খ্রি.) তার প্রধানমন্ত্রী কর্তৃক মন্দিরটির সংস্কার কাজ করা হলে এটির নাম হয় "পুলগুকসা' । এর অর্থ পরিচ্ছন্ন ও স্বস্তিময় দেশ গড়ার আশা । শিল্পা সাম্রাজ্যের স্থায়িত্ ও সমৃদ্ধির এক মূর্ত প্রতীক এ পুলগুকসা। এটির অভ্যন্তরে মোট ৮০টি কাঠের ভবন ছিল বিধায় এটিকে গ্র্যান্ড টেম্পল" বলা হতো। কোরীয় ও চোসান রাজতৃকাল ছিল এ মন্দিরের জন্য কঠিন সময় । তখন যুদ্ধ-বিপ্রহ লেগেই থাকতো এবং বৌদ্ধধর্মকে অবদমনের চেষ্টা চালানো হতো । ১৫৯৩ খরস্টাব্দে জাপান কর্তৃক কোরিয়া দখল করা হলে এর সকল কাঠের অবকাঠামো ভস্মীভূত হয়। মন্দিরটি এর হত গৌরব ফিরে পায় ১৯৬৯ থেকে ১৯৭৩ খ্রিস্টাব্দে যখন এটির ব্যাপক পুনঞ্জির্মাণ কর্ম সাধিত হয় । মন্দির অভ্যন্তরে রয়েছে মিশ্র সংস্কৃতির বিভিন্ন স্বর্গরাজ, প্রাণী ও বুদ্ধের অনেক মূর্তি, আছে হাজার বছরের কোরিয়ান স্থাপত্য ও ভাস্কর্য নিদর্শন ।	পুলগুকসা প্রথম নির্মিত হয় খ্রিস্টাব্দে রাজা পোপহাং কর্তৃক পরবর্তীতে রাজা কিয়ংদকএর শাসনামলে খ্রি তার প্রধানমন্ত্রী কর্তৃক মন্দিরটির সংস্কার কাজ করা হলে এটির নাম হয় পুলগুকসা এর অর্থ পরিচ্ছন্ন ও স্বস্তিময় দেশ গড়ার আশা শিল্পা সাম্রাজ্যের স্থায়িত্ ও সমৃদ্ধির এক মূর্ত প্রতীক এ পুলগুকসা এটির অভ্যন্তরে মোট টি কাঠের ভবন ছিল বিধায় এটিকে গ্র্যান্ড টেম্পল বলা হতো কোরীয় ও চোসান রাজতৃকাল ছিল এ মন্দিরের জন্য কঠিন সময় তখন যুদ্ধবিপ্রহ লেগেই থাকতো এবং বৌদ্ধধর্মকে অবদমনের চেষ্টা চালানো হতো খরস্টাব্দে জাপান কর্তৃক কোরিয়া দখল করা হলে এর সকল কাঠের অবকাঠামো ভস্মীভূত হয় মন্দিরটি এর হত গৌরব ফিরে পায় থেকে খ্রিস্টাব্দে যখন এটির ব্যাপক পুনঞ্জির্মাণ কর্ম সাধিত হয় মন্দির অভ্যন্তরে রয়েছে মিশ্র সংস্কৃতির বিভিন্ন স্বর্গরাজ প্রাণী ও বুদ্ধের অনেক মূর্তি আছে হাজার বছরের কোরিয়ান স্থাপত্য ও ভাস্কর্য নিদর্শন

TABLE VI. DATA SAMPLE CONTAINING RAW DATA AND CLEANED DATA

Snippet_After Cleaning	Snippet_After Tokenization
পুলগুকসা প্রথম নির্মিত হয় খ্রিস্টাব্দে রাজা পোপহাং কর্তৃক পরবর্তীতে রাজা কিয়ংদকএর শাসনামলে খ্রি তার প্রধানমন্ত্রী কর্তৃক মন্দিরটির সংস্কার কাজ করা হলে এটির নাম হয় পুলগুকসা এর অর্থ পরিচ্ছন্ন ও স্বস্তিময় দেশ গড়ার আশা শিল্পা সাম্রাজ্যের স্থায়িত্ ও সমৃদ্ধির এক মূর্ত প্রতীক এ পুলগুকসা এটির অভ্যন্তরে মোট টি কাঠের ভবন ছিল বিধায় এটিকে গ্র্যান্ড টেম্পল বলা হতো কোরীয় ও চোসান রাজতৃকাল ছিল এ মন্দিরের জন্য কঠিন সময় তখন যুদ্ধবিপ্রহ লেগেই থাকতো এবং বৌদ্ধধর্মকে অবদমনের চেষ্টা চালানো হতো খরস্টাব্দে জাপান কর্তৃক কোরিয়া দখল করা হলে এর সকল কাঠের অবকাঠামো ভস্মীভূত হয় মন্দিরটি এর হত গৌরব ফিরে পায় থেকে খ্রিস্টাব্দে যখন এটির ব্যাপক পুনঞ্জির্মাণ কর্ম সাধিত হয় মন্দির অভ্যন্তরে রয়েছে মিশ্র সংস্কৃতির বিভিন্ন স্বর্গরাজ প্রাণী ও বুদ্ধের অনেক মূর্তি আছে হাজার বছরের কোরিয়ান স্থাপত্য ও ভাস্কর্য নিদর্শন	পুলগুকসা', 'প্রথম', 'নির্মিত', 'হয়', 'খ্রিস্টাব্দে', 'রাজা', 'পোপহাং', 'কর্তৃক', 'পরবর্তীতে', 'রাজা', 'কিয়ংদকএর', 'শাসনামলে', 'খ্রি', 'তার', 'প্রধানমন্ত্রী', 'কর্তৃক', 'মন্দিরটির', 'সংস্কার', 'কাজ', 'করা', 'হলে', 'এটির', 'নাম', 'হয়', 'পুলগুকসা', 'এর', 'অর্থ', 'পরিচ্ছন্ন', 'ও', 'স্বস্তিময়', 'দেশ', 'গড়ার', 'আশা', 'শিল্পা', 'সাম্রাজ্যের', 'স্থায়িত্', 'ও', 'সমৃদ্ধির', 'এক', 'মূর্ত', 'প্রতীক', 'এ', 'পুলগুকসা', 'এটির', 'অভ্যন্তরে', 'মোট', 'টি', 'কাঠের', 'ভবন', 'ছিল', 'বিধায়', 'এটিকে', 'গ্র্যান্ড', 'টেম্পল', 'বলা', 'হতো', 'কোরীয়', 'ও', 'চোসান', 'রাজতৃকাল', 'ছিল', 'এ', 'মন্দিরের', 'জন্য', 'কঠিন', 'সময়', 'তখন', 'যুদ্ধবিপ্রহ', 'লেগেই', 'থাকতো', 'এবং', 'বৌদ্ধধর্মকে', 'অবদমনের', 'চেষ্টা', 'চালানো', 'হতো', 'খরস্টাব্দে', 'জাপান', 'কর্তৃক', 'কোরিয়া', 'দখল', 'করা', 'হলে', 'এর', 'সকল', 'কাঠের', 'অবকাঠামো', 'ভস্মীভূত', 'হয়', 'মন্দিরটি', 'এর', 'হত', 'গৌরব', 'ফিরে', 'পায়', 'থেকে', 'খ্রিস্টাব্দে', 'যখন', 'এটির', 'ব্যাপক', 'পুনঞ্জির্মাণ', 'কর্ম', 'সাধিত', 'হয়', 'মন্দির', 'অভ্যন্তরে', 'রয়েছে', 'মিশ্র', 'সংস্কৃতির', 'বিভিন্ন', 'স্বর্গরাজ', 'প্রাণী', 'ও', 'বুদ্ধের', 'অনেক', 'মূর্তি', 'আছে', 'হাজার', 'বছরের', 'কোরিয়ান', 'স্থাপত্য', 'ও', 'ভাস্কর্য', 'নিদর্শন'

2) Tokenization

Tokenization refers to the procedure of partitioning a given text into smaller units, commonly referred to as tokens. Tokenization is the process through which raw data are converted into a coherent and intelligible sequence of data elements. We used IndicNLP library for tokenization. Once the dataset was cleaned, we proceeded with its tokenization. After tokenization, a list of tokens is obtained, representing individual words or meaningful units from the text.

Table VI shows a sample of the tokenized data resulting from the preprocessing steps applied to the cleaned dataset. In Table VI, "Snippet_After Cleaning" column contains the cleaned data, while the "Snippet_After Tokenization" column comprises the tokenized data.

3) Stop word removal

After tokenization, we removed the stop words. Stop words are devoid of significant contextual meaning. We made a list of such stop words and removed them from the datasets to enhance the focus on essential information. The exclusion of stop words facilitated a more refined dataset.

In Table VII, a sample of the data derived after the removal of the stop words is showcased. In Table VII, the "Snippet_After Tokenization" column exhibits the tokenized data, while the "Snippet_After StopWordRemoval" column portrays the result after removing stop words. Words like 'ও', 'এবং' etc. got removed.

TABLE VII. DATA SAMPLE CONTAINING RAW DATA AND CLEANED DATA

Snippet_After Tokenization	Snippet_After StopWord Removal
পুলগুকসা', 'প্রথম', 'নির্মিত', 'হয়', 'খ্রিস্টাব্দে', 'রাজা', 'পোপহাং', 'কর্তৃক', 'পরবর্তীতে', 'রাজা', 'কিয়ংদকএর', 'শাসনামলে', 'খ্রি', 'তার', 'প্রধানমন্ত্রী', 'কর্তৃক', 'মন্দিরটির', 'সংস্কার', 'কাজ', 'করা', 'হলে', 'এটির', 'নাম', 'হয়', 'পুলগুকসা', 'এর', 'অর্থ', 'পরিচ্ছন্ন', 'ও', 'স্বস্তিময়', 'দেশ', 'গড়ার', 'আশা', 'শিল্পা', 'সাম্রাজ্যের', 'স্থায়িত্', 'ও', 'সমৃদ্ধির', 'এক', 'মূর্ত', 'প্রতীক', 'এ', 'পুলগুকসা', 'এটির', 'অভ্যন্তরে', 'মোট', 'টি', 'কাঠের', 'ভবন', 'ছিল', 'বিধায়', 'এটিকে', 'গ্র্যান্ড', 'টেম্পল', 'বলা', 'হতো', 'কোরীয়', 'ও', 'চোসান', 'রাজতৃকাল', 'ছিল', 'এ', 'মন্দিরের', 'জন্য', 'কঠিন', 'সময়', 'তখন', 'যুদ্ধবিপ্রহ', 'লেগেই', 'থাকতো', 'এবং', 'বৌদ্ধধর্মকে', 'অবদমনের', 'চেষ্টা', 'চালানো', 'হতো', 'খরস্টাব্দে', 'জাপান', 'কর্তৃক', 'কোরিয়া', 'দখল', 'করা', 'হলে', 'এর', 'সকল', 'কাঠের', 'অবকাঠামো', 'ভস্মীভূত', 'হয়', 'মন্দিরটি', 'এর', 'হত', 'গৌরব', 'ফিরে', 'পায়', 'থেকে', 'খ্রিস্টাব্দে', 'যখন', 'এটির', 'ব্যাপক', 'পুনঞ্জির্মাণ', 'কর্ম', 'সাধিত', 'হয়', 'মন্দির', 'অভ্যন্তরে', 'রয়েছে', 'মিশ্র', 'সংস্কৃতির', 'বিভিন্ন', 'স্বর্গরাজ', 'প্রাণী', 'ও', 'বুদ্ধের', 'অনেক', 'মূর্তি', 'আছে', 'হাজার', 'বছরের', 'কোরিয়ান', 'স্থাপত্য', 'ও', 'ভাস্কর্য', 'নিদর্শন'	পুলগুকসা', 'প্রথম', 'নির্মিত', 'হয়', 'খ্রিস্টাব্দে', 'রাজা', 'পোপহাং', 'কর্তৃক', 'পরবর্তীতে', 'রাজা', 'কিয়ংদকএর', 'শাসনামলে', 'প্রধানমন্ত্রী', 'কর্তৃক', 'কর্তৃক', 'মন্দিরটির', 'সংস্কার', 'কাজ', 'মন্দিরটির', 'সংস্কার', 'কাজ', 'করা', 'করা', 'হলে', 'এটির', 'নাম', 'হয়', 'হলে', 'এটির', 'নাম', 'হয়', 'পুলগুকসা', 'এর', 'অর্থ', 'পরিচ্ছন্ন', 'ও', 'পুলগুকসা', 'অর্থ', 'পরিচ্ছন্ন', 'স্বস্তিময়', 'দেশ', 'গড়ার', 'আশা', 'শিল্পা', 'স্বস্তিময়', 'দেশ', 'গড়ার', 'আশা', 'সাম্রাজ্যের', 'স্থায়িত্', 'ও', 'সমৃদ্ধির', 'শিল্পা', 'সাম্রাজ্যের', 'স্থায়িত্', 'সমৃদ্ধির', 'এক', 'মূর্ত', 'প্রতীক', 'এ', 'পুলগুকসা', 'মূর্ত', 'প্রতীক', 'এ', 'পুলগুকসা', 'এটির', 'অভ্যন্তরে', 'মোট', 'টি', 'কাঠের', 'এটির', 'অভ্যন্তরে', 'মোট', 'কাঠের', 'ভবন', 'ছিল', 'বিধায়', 'এটিকে', 'গ্র্যান্ড', 'ভবন', 'ছিল', 'বিধায়', 'এটিকে', 'টেম্পল', 'বলা', 'হতো', 'কোরীয়', 'ও', 'গ্র্যান্ড', 'টেম্পল', 'বলা', 'হতো', 'চোসান', 'রাজতৃকাল', 'ছিল', 'এ', 'কোরীয়', 'চোসান', 'রাজতৃকাল', 'মন্দিরের', 'জন্য', 'কঠিন', 'সময়', 'তখন', 'মন্দিরের', 'জন্য', 'কঠিন', 'সময়', 'যুদ্ধবিপ্রহ', 'লেগেই', 'থাকতো', 'এবং', 'তখন', 'যুদ্ধবিপ্রহ', 'লেগেই', 'বৌদ্ধধর্মকে', 'অবদমনের', 'চেষ্টা', 'থাকতো', 'বৌদ্ধধর্ম', 'অবদমনের', 'চালানো', 'হতো', 'খরস্টাব্দে', 'জাপান', 'চেষ্টা', 'চালানো', 'হতো', 'খরস্টাব্দে', 'কর্তৃক', 'কোরিয়া', 'দখল', 'করা', 'হলে', 'জাপান', 'কর্তৃক', 'কোরিয়া', 'দখল', 'এর', 'সকল', 'কাঠের', 'অবকাঠামো', 'হলে', 'সকল', 'কাঠের', 'অবকাঠামো', 'ভস্মীভূত', 'হয়', 'মন্দিরটি', 'এর', 'হত', 'ভস্মীভূত', 'হয়', 'মন্দিরটি', 'হত', 'গৌরব', 'ফিরে', 'পায়', 'থেকে', 'গৌরব', 'ফিরে', 'পায়', 'খ্রিস্টাব্দে', 'খ্রিস্টাব্দে', 'যখন', 'এটির', 'ব্যাপক', 'এটির', 'ব্যাপক', 'পুনঞ্জির্মাণ', 'কর্ম', 'পুনঞ্জির্মাণ', 'কর্ম', 'সাধিত', 'হয়', 'সাধিত', 'হয়', 'মন্দির', 'অভ্যন্তরে', 'মন্দির', 'অভ্যন্তরে', 'রয়েছে', 'মিশ্র', 'মিশ্র', 'সংস্কৃতির', 'বিভিন্ন', 'স্বর্গরাজ', 'সংস্কৃতির', 'বিভিন্ন', 'স্বর্গরাজ', 'প্রাণী', 'প্রাণী', 'বুদ্ধের', 'মূর্তি', 'হাজার', 'ও', 'বুদ্ধের', 'অনেক', 'মূর্তি', 'আছে', 'বছরের', 'কোরিয়ান', 'স্থাপত্য', 'হাজার', 'বছরের', 'কোরিয়ান', 'স্থাপত্য', 'ভাস্কর্য', 'নিদর্শন' 'ও', 'ভাস্কর্য', 'নিদর্শন'

4) Stemming

Following the removal of extraneous elements from our dataset, we proceeded with the use of stemming techniques. Stemming refers to the process of eliminating

a segment of a word or reducing it to its fundamental stem or root form. The titles and snippets underwent stemming using a Bangla stemmer after the removal of the stop words.

Table VIII presents a snapshot of the data following stop word elimination. The "Snippet_After StopWordRemoval" column displays the data without stop words, whereas the "Snippet_After Stemming" column illustrates the output post stemming. Here for example if we take the word 'খ্রিস্টাব্দে', it got converted to 'খ্রিস্টাব্দ' after stemming.

TABLE VIII. DATA SAMPLE CONTAINING RAW DATA AND CLEANED DATA

Snippet_After Stop Word Removal	Snippet_After Stemming
পুলগুকসা', 'প্রথম', 'নির্মিত', 'হয়', 'খ্রিস্টাব্দে',	পুলগুকসা', 'নির্মিত', 'হয়', 'খ্রিস্টাব্দ',
'রাজা', 'পোপহাং', 'কর্তৃক', 'পরবর্তীতে', 'রাজা',	'রাজা', 'পোপহাং', 'কর্তৃক', 'পরবর্তী',
'কিয়ংদকএর', 'শাসনামলে', 'খ্রি', 'তার',	'রাজা', 'কিয়ংদকএর', 'শাসনামলে',
'প্রধানমন্ত্রী', 'কর্তৃক', 'মন্দিরটির', 'সংস্কার',	'প্রধানমন্ত্রী', 'কর্তৃক', 'মন্দির', 'সংস্ক',
'কাজ', 'করা', 'হলে', 'এটির', 'নাম', 'হয়',	'কাজ', 'করা', 'হলে', 'এটি', 'নাম',
'পুলগুকসা', 'এর', 'অর্থ', 'পরিচ্ছন্ন', 'ও',	'হয়', 'পুলগুকসা', 'অর্থ', 'পরিচ্ছন্ন',
'স্বস্তিময়', 'দেশ', 'গড়ার', 'আশা', 'শিল্পা',	'স্বস্তিময়', 'দেশ', 'গড়', 'আশা', 'শিল্পা',
'সাম্রাজ্যের', 'স্থায়িত্', 'ও', 'সমৃদ্ধির', 'এক',	'সাম্রাজ্য', 'স্থায়িত্', 'সমৃদ্ধির', 'মূর্ত',
'মূর্ত', 'প্রতীক', 'এ', 'পুলগুকসা', 'এটির',	'প্রতীক', 'পুলগুকসা', 'এটি',
'অভ্যন্তরে', 'মোট', 'টি', 'কাঠের', 'ভবন', 'ছিল',	'অভ্যন্তরে', 'মোট', 'কাঠ', 'ভবন',
'বিধায়', 'এটিকে', 'গ্র্যান্ড', 'টেম্পল', 'বলা',	'ছিল', 'বিধায়', 'এটি', 'গ্র্যান্ড',
'হতো', 'কোরীয়', 'ও', 'চোসান', 'রাজতৃকাল',	'টেম্পল', 'বলা', 'হত', 'কোরীয়',
'ছিল', 'এ', 'মন্দিরের', 'জন্য', 'কঠিন', 'সময়',	'চোসান', 'রাজতৃকাল', 'মন্দির',
'তখন', 'যুদ্ধবিপ্রহ', 'লেগেই', 'থাকতো', 'এবং',	'কঠিন', 'তখন', 'যুদ্ধবিপ্রহ', 'লাগ',
'বৌদ্ধধর্মকে', 'অবদমনের', 'চেষ্টা', 'চালানো',	'থাক', 'বৌদ্ধধর্ম', 'অবদমন', 'চালানো',
'হতো', 'খরস্টাব্দে', 'জাপান', 'কর্তৃক',	'হত', 'খরস্টাব্দ', 'জাপান', 'কর্তৃক',
'কোরিয়া', 'দখল', 'করা', 'হলে', 'এর', 'সকল',	'কোরিয়া', 'দখল', 'হলে', 'সকল',
'কাঠের', 'অবকাঠামো', 'ভস্মীভূত', 'হয়',	'কাঠ', 'অবকাঠামো', 'ভস্মীভূত', 'হয়',
'মন্দিরটি', 'এর', 'হত', 'গৌরব', 'ফিরে', 'পায়',	'মন্দির', 'হত', 'গৌর', 'ফির', 'পায়',
'থেকে', 'খ্রিস্টাব্দে', 'যখন', 'এটির', 'ব্যাপক',	'খ্রিস্টাব্দ', 'এটি', 'ব্যাপক',
'পুনঞ্জির্মাণ', 'কর্ম', 'সাধিত', 'হয়', 'মন্দির',	'পুনঞ্জির্মাণ', 'কর্ম', 'সাধিত', 'হয়',
'অভ্যন্তরে', 'রয়েছে', 'মিশ্র', 'সংস্কৃতির',	'মন্দির', 'অভ্যন্তর', 'মিশ্র', 'সংস্কৃতি',
'বিভিন্ন', 'স্বর্গরাজ', 'প্রাণী', 'ও', 'বুদ্ধের',	'বিভিন্ন', 'স্বর্গরাজ', 'প্রাণী', 'বুদ্ধ',
'অনেক', 'মূর্তি', 'আছে', 'হাজার', 'বছরের',	'মূর্তি', 'হাজার', 'বছর', 'কোরিয়ান',
'কোরিয়ান', 'স্থাপত্য', 'ও', 'ভাস্কর্য', 'নিদর্শন'	'স্থাপত্য', 'ভাস্কর্য', 'নিদর্শন'

5) Data balancing

The dataset for the book titles is imbalanced. An imbalanced dataset tends to classify all the data as belonging to the majority class. We balanced our dataset using the Tomek-Link undersampling technique to address this issue. The Tomek-Link technique eliminates samples of the majority class that are the closest neighbors of samples from the minority class. Tomek connections represent pairings of instances from opposing classes close to one another. We removed the majority of these Tomek-Links class samples. This contributes to a better decision boundary and results in a more balanced dataset.

B. Feature Extraction

After pre-processing, the next step is feature extraction. Feature extraction is a fundamental method that is pivotal in enhancing our understanding of the contextual aspects we are addressing. Once the initial text has been cleaned, it must be converted into its features for use by the models.

Due to the inherent limitations of machine learning algorithms in processing textual data, it is currently not feasible to input text data into any machine learning method, as it exclusively understands numerical data. Feature extraction is a technique used to convert textual data into numerical representations, allowing machine learning models to comprehend and process such data. In their work, Mauni *et al.* [17] demonstrated that extracting a set of features with efficient algorithms reduces the dimensions of the feature space and eliminates redundant features from the model.

Feature Extraction for Classic Machine Learning Models: Term Frequency Inverse Document Frequency (TF-IDF) has been employed to extract features in the context of traditional machine-learning models. The TF-IDF measure is a comprehensive score that quantifies the ability of a certain word to distinguish and identify a document effectively. The greater the TF-IDF value of a word, the more unique and uncommon its occurrence. This measure considers each word's importance and is computationally inexpensive.

Feature Extraction for Deep Learning Models: word embedding has shown notable efficacy in increasing comprehension of textual information. Word vectorization approaches, such as TF-IDF and BOW, rely on the frequency of words for their implementation. The contextual meaning of a sentence becomes obscured when its frequency is quantified. In contrast, BERT utilizes transfer learning to derive contextualized word embeddings. This paper used the BERT word embedding technique as a deep learning approach for information retrieval. The Bangla BERT model, which is pre-trained using Bengali Wikipedia [18], performed the best with the proposed model architecture for deep learning.

C. Genre Classification with Classical Machine Learning Models

In this paper, we applied five classic machine learning techniques as predictive models to classify the genre of Bangla books based on their titles and snippets. TF-IDF was employed for feature selection due to its superior performance with the aforementioned models. Multinomial Naive Bayes (NB), Random Forest (RF), and Logistic Regression (LR) achieved the highest accuracy for the Title dataset, while Multinomial Naive Bayes had the highest accuracy for the excerpt's dataset.

D. Genre Classification with Deep Learning Models

In this study, four deep-learning models were employed for classification. They are:

- Long Short-Term Memory (LSTM)
- Convolutional Neural Network (CNN)
- Recurrent Neural Network (RNN)
- Bidirectional Encoder Representations from Transformers (BERT)

Model Description Long Short-Term Memory (LSTM): An embedding layer was employed for the purpose of embedding, and the model architecture included an LSTM layer with 128 units. To mitigate the issue of overfitting, recurrent dropout was implemented. Additionally, to ensure that the output of all sequences was obtained rather than solely the final one, the return sequences were preserved. Finally, the output is passed through a dense

layer, which yields the classification. We chose to utilize the SoftMax activation function for our model. The parameters pertaining to the layers of this model can be comprehensively inferred from Fig. 5 shown herein.

Layer (type)	Output Shape	Param #
embedding_6 (Embedding)	(None, 600, 150)	15000000
spatial_dropout1d (SpatialDropout1D)	(None, 600, 150)	0
lstm_6 (LSTM)	(None, 128)	142848
dropout_2 (Dropout)	(None, 128)	0
dense_13 (Dense)	(None, 9)	1161

Total params: 15144009 (57.77 MB)
Trainable params: 15144009 (57.77 MB)
Non-trainable params: 0 (0.00 Byte)

Fig. 5. Training parameters of LSTM model.

Convolutional Neural Network (CNN): In our study, we employed a convolutional neural network architecture based on deep learning techniques to perform the task of intent classification for textual commands. While Convolutional Neural Networks (CNNs) are commonly associated with computer vision tasks, CNN kernels are crucial in discerning pertinent patterns within textual input. The 1D convolution layer is responsible for generating a convolution kernel that is applied to the layer's input along a single spatial dimension. The pooling layer employed in our research was Maxpool1D, with pool dimensions measured to be 3 units. The flattened layer is utilized to modify the dimensional shape of the outputs. Finally, the output is passed through a dense layer, which yields the classification. We chose to utilize the SoftMax activation function for our model. The parameters pertaining to the layers of this model can be comprehensively inferred from Fig. 6 shown herein.

Layer (type)	Output Shape	Param #
embedding_7 (Embedding)	(None, 600, 150)	15000000
conv1d_4 (Conv1D)	(None, 598, 150)	67650
max_pooling1d_1 (MaxPooling1D)	(None, 199, 150)	0
dense_17 (Dense)	(None, 199, 20)	3020
flatten_4 (Flatten)	(None, 3980)	0
dense_18 (Dense)	(None, 13)	51753

Total params: 15122423 (57.69 MB)
Trainable params: 15122423 (57.69 MB)
Non-trainable params: 0 (0.00 Byte)

Fig. 6. Training parameters of CNN model.

Recurrent Neural Network (RNN) The Bidirectional Recurrent Neural Network (RNN) has been employed in our study. Recurrent Neural Networks (RNNs) are prevalent in Natural Language Processing (NLP), demonstrating a relatively high level of accuracy and efficiency in language acquisition. The layers of our RNN model are as follows: The embedding layer serves as the initial layer in a neural network model, responsible for mapping word tokenizers to a vector representation with a specified number of dimensions, known as word embedding. The spatial dropout layer is utilized to mitigate overfitting by selectively dropping nodes.

The value of 0.4 represents the probability that the nodes must be dropped. The bidirectional layer is a Recurrent Neural Network (RNN) layer containing long short-term memory (LSTM) units with a dimensionality of 128. Finally, the output is passed through a dense layer, which yields the classification. The parameters pertaining to the layers of this model can be comprehensively inferred from Fig. 7 presented below.

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 600, 150)	15000000
spatial_dropout1d (SpatialDropout1D)	(None, 600, 150)	0
bidirectional (Bidirectional)	(None, 600, 256)	286720
bidirectional_1 (Bidirectional)	(None, 256)	395264
dense (Dense)	(None, 9)	2313

Total params: 15,684,297
Trainable params: 15,684,297
Non-trainable params: 0

Fig. 7. Training parameters of RNN model.

Bidirectional Encoder Representations from Transformers (BERT): DistilBERT were employed for the purpose of categorization. DistilBERT is a compact, efficient, cost-effective, and lightweight Transformer model trained through the process of distillation, using a BERT base as its source. The model exhibits a 40% reduction in the number of parameters compared to Bert-base-uncased. Additionally, it demonstrates a 60% increase in computational efficiency while maintaining a performance level of over 95% of BERT's results, as evaluated on the GLUE language understanding benchmark. BERT preprocessing was employed to preprocess the training data in our study. DistilBERT was fine-tuned using Ktrain, a software library that serves as a lightweight wrapper for the TensorFlow Keras framework. DistilBERT is fine-tuned using a learning rate of 8e-5. The parameters of the layers of this model can be inferred in detail from Fig. 8 presented below.

Layer (type)	Output Shape	Param #
distilbert (TFDistilBertMainLayer)	multiple	66362880
pre_classifier (Dense)	multiple	590592
classifier (Dense)	multiple	2307
dropout_19 (Dropout)	multiple	0

Total params: 66,955,779
Trainable params: 66,955,779
Non-trainable params: 0

Fig. 8. Training parameters of BERT model.

Parameters and hyperparameters have been employed in deep learning models. These are displayed in Table IX. In this context, we have listed the optimizers, activation functions, loss functions, and epochs associated with LSTM, CNN, RNN, and BERT models.

TABLE IX. PARAMETER AND HYPERPARAMETER OF DEEP LEARNING MODELS

Model	Optimizer	Activation function	Loss function	Epoch
LSTM	ADAM	SoftMax	Cross-entropy	50
CNN	ADAM	SoftMax,ReLU	Cross-entropy	20
RNN	ADAM	SoftMax	Cross-entropy	20
BERT	ADAMW	GELU	Cross-entropy	10

V. RESULT ANALYSIS

We will discuss the performance of our models. For this purpose, we have employed various evaluation metrics, including accuracy, precision, F1-score, and recall. We employ both Machine Learning and Deep Learning algorithms in order to classify the genres of books. The identical models were utilized for both of our data sets.

Each data set is divided into training and testing portions in order to train the model and evaluate its performance. We have employed repeated stratified 10-fold cross-validation to estimate the performance of machine learning algorithms. Repeated stratified k-fold cross-validation yields a mean result across all folds from all repetitions, providing a more precise estimate of the performance. Stratification helps maintain the proportion of samples for each category.

A. Result Analysis for the Title Dataset

Table X presents the performance metrics for various machine learning models applied to the Title dataset, namely accuracy, macro F1-score, precision, and recall. Based on the findings presented in Table IV, it can be observed that the accuracy and F1-score of all the models exhibit a high degree of similarity.

TABLE X. RESULT OF EVALUATION METRICS FOR CLASSIC MACHINE LEARNING MODELS FOR TITLE DATASET

Classifiers	Accuracy	Precision	Recall	F1-score
Multinomial NB	54%	57%	52%	53%
Random Forest	54%	56%	52%	53%
Logistic Regression	54%	59%	52%	54%
Decision Tree	53%	56%	51%	52%
SVM	53%	57%	50%	52%

The classification accuracy achieved by the Multinomial Naive Bayes, Random Forest, and Logistic Regression models is 54%. It may be asserted that Logistic Regression demonstrated superior performance, as evidenced by its highest accuracy and F1-score. The scores of all models exhibit a high degree of similarity. Fig. 9 illustrates the performance graph of many classical machine-learning algorithms.

Figs. 10 and 11 displays the ROC (receiver operating characteristic curve) curve, which illustrates the performance of our classical machine learning models on the Title dataset. As genre classification is a multiclass classification scenario, the ROC curve utilizes the macro-average technique to present an overall evaluation.

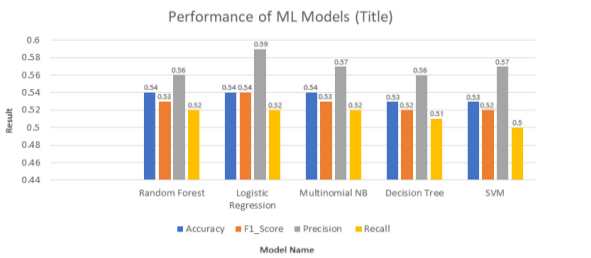

Fig. 9. Performance of classic machine learning models using Title dataset.

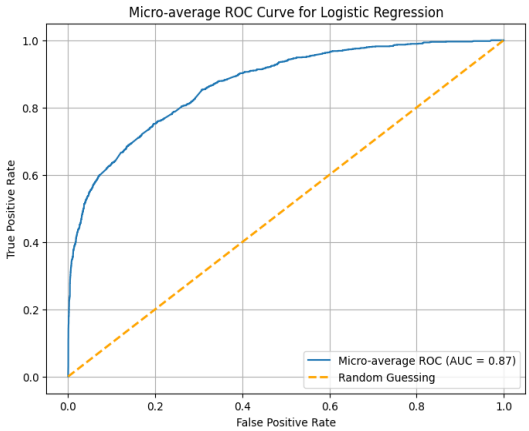

Fig. 10. ROC curve of logistic regression using Title dataset.

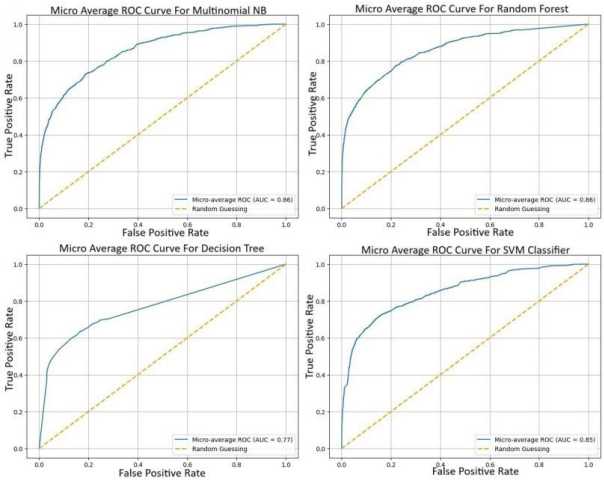

Fig. 11. ROC curve of multinomial NB, random forest classifier, decision tree classifier, SVM classifier using Title dataset.

Table XI presents the performance metrics, including accuracy, macro F1-score, precision, and recall, of various deep learning models applied to the Title dataset. BERT performs better than other deep neural network models, exhibiting an accuracy rate of 77%.

TABLE XI. RESULT OF EVALUATION METRICS FOR DEEP LEARNING MODELS FOR TITLE DATASET

Classifiers	Accuracy	Precision	Recall	F1-score
LSTM	66.19%	67.45%	65.52%	66.04%
CNN	52.80%	56.04%	53.16%	53.41%
RNN	58.84%	63.47%	58.98%	59.55%
BERT	77%	77%	76%	77%

The performance of this model shows significant improvement compared to the other models. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) exhibit superior performance in subsequent order. Among the various deep learning models, it can be observed that the Convolutional Neural Network (CNN) exhibits somewhat lower performance compared to other models. Fig. 12 displays the performance graph of deep learning models.

Fig. 13 exhibits the ROC curve, depicting the performance of RNN and LSTM models on the Title dataset. Since genre classification involves a multiclass scenario, the ROC curve employs the macro-average technique to offer a comprehensive assessment.

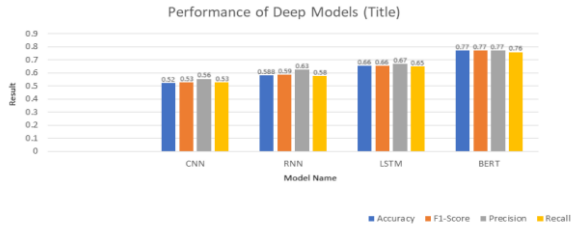

Fig. 12. Performance of deep learning models using Title dataset.

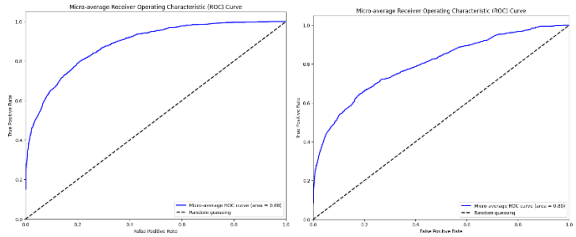

Fig. 13. ROC curve of LSTM, RNN using Title dataset.

B. *Result Analysis for the Snippet Dataset*

Table XII lists the precision, recall, accuracy, and macro F1-score of various machine learning models for the dataset. Multinomial NB produced the highest performance in the snippet dataset, yielding an average of 81% for all scores.

TABLE XII. RESULT OF EVALUATION METRICS FOR CLASSIC MACHINE LEARNING. MODELS FOR SNIPPET DATASET

Classifiers	Accuracy	Precision	Recall	F1-score
Multinomial NB	81%	82%	81%	81%
Random Forest	80%	81%	80%	80%
Logistic Regression	78%	79%	78%	78%
Decision Tree	77%	78%	77%	77%
SVM	77%	78%	77%	77%

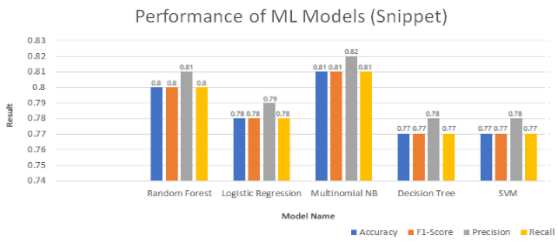

Fig. 14. Performance of classic machine learning models using snippet dataset.

Random Forest produces the second-best result, with nearly 80% for all scores. Fig. 14 provides a visual representation of the performance graph.

Figs. 15 and 16 showcase the ROC curve, delineating the performance of our classical machine learning models on the snippet dataset. Given the multiclass nature of genre classification, the ROC curve employs the macro-average technique to provide a comprehensive evaluation.

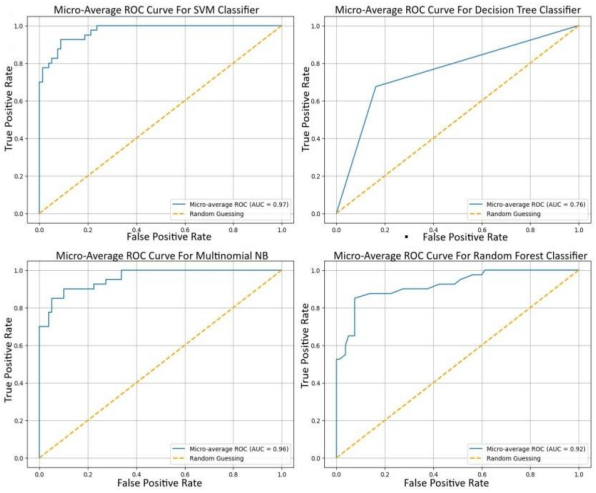

Fig. 15. ROC curve of SVM, decision tree classifier, multinomial NB, random forest classifier using snippet dataset.

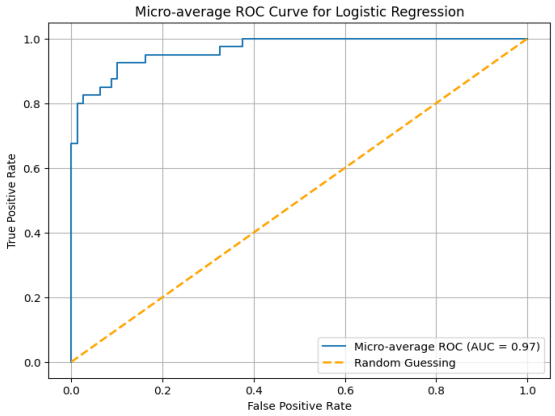

Fig. 16. ROC curve of logistic regression using snippet dataset.

Table XIII gives the Deep Model outcome analysis for the Snippet's dataset. In the snippet dataset BERT outperforms all other deep models, achieving an accuracy rate of 95%. Due to the quantity of the dataset, other deep learning models perform poorly. Deep learning algorithms require a significant amount of data for training. BERT demonstrates strong performance due to its utilization as a pre-trained model. The performance graph depicting the outcomes of several deep machine learning models is presented in Fig. 17.

TABLE XIII. RESULT OF EVALUATION METRICS FOR DEEP LEARNING MODELS FOR SNIPPET DATASET

Classifiers	Accuracy	Precision	Recall	F1-score
LSTM	53.66%	66.85%	55.34%	52.29%
CNN	52.44%	54.98%	54.31%	52.44%
RNN	53.66%	66.85%	55.34%	52.29%
BERT	90%	93%	94%	94%

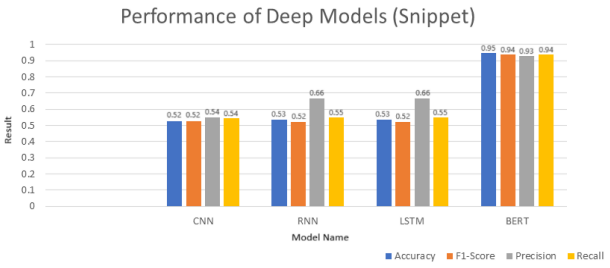

Fig. 17. Performance of deep learning models using snippet dataset.

Figs. 18 and 19 displays the ROC curve, which illustrates the performance of CNN, RNN and LSTM models on the snippet dataset. As genre classification is a multiclass classification scenario, the ROC curve utilizes the macro-average technique to present an overall evaluation.

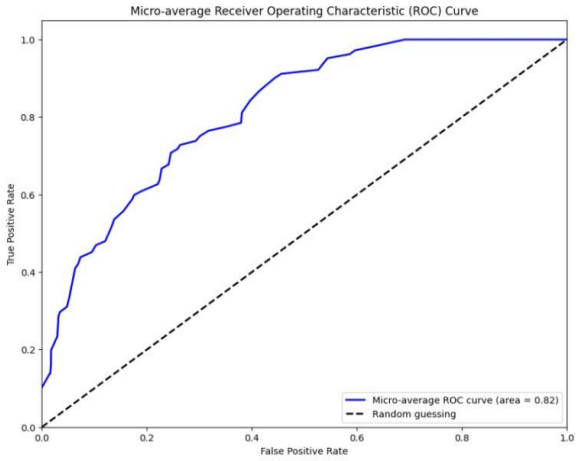

Fig. 18. ROC curve of CNN using snippet dataset.

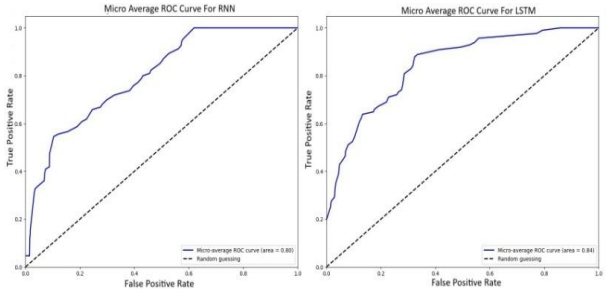

Fig. 19. ROC curve of RNN, LSTM using snippet dataset.

C. Result Comparison

Based on our result analysis, it has been found that with the exception of BERT, machine learning models exhibit superior performance compared to deep learning models in the context of the snippet dataset. On the contrary, in the context of the Title dataset, it can be observed that deep learning models exhibit better performance compared to machine learning models. Thus, BERT demonstrates superior performance on both datasets.

Improved performance is observed when utilizing the snippet dataset for both standard machine learning models and BERT. This may be attributed to the fact that working with snippets allows for a higher number of words per sample, enhancing the training process for our models. However, when employing the Title dataset, each sample provides only one to two words for training our models. It is widely acknowledged that text data classification models tend to achieve better results when sufficient words are available to effectively differentiate between various categories.

Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNNs) have accuracy levels ranging from 52% to 66%. The reason for the suboptimal performance of these models might be the insufficiency of the dataset in terms of its size. Although the Title dataset contains 12,925 samples, the length of each sample is very short because titles are typically brief. Eran *et al.* [7] used 207,575 samples of English-language book title data in their research and obtained an accuracy of 55.40% using Naive Bayes, 65.58% for LSTM, 55.91% for RNN, and 63.10% for CNN. Due to text lengths, the machine learning models performed better with the snippets dataset than with the Title dataset. Fig. 20 is a graph depicting the efficacy of all models in both datasets.

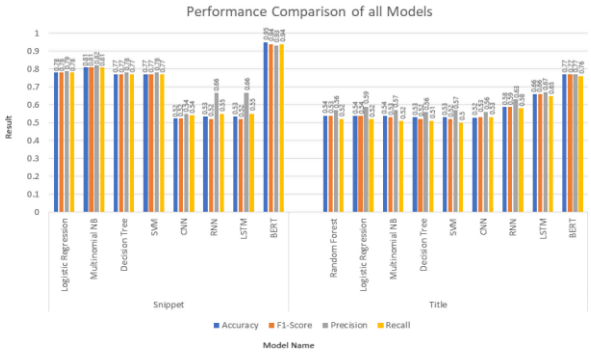

Fig. 20. Performance comparison among all models.

VI. Conclusion and Future Works

The primary objective of our research was to present a technique for the automated detection of genres in Bengali literary works. To achieve this goal, two datasets were generated: One dataset comprised the titles of books from nine distinct genres, whereas the second dataset consisted of book snippets from three different genres. Machine learning and neural network models were developed for both datasets. In the domain of neural networks, employing the Title dataset, we attained the highest accuracy of 77% through the utilization of BERT. Using the snippet dataset, we achieved a peak accuracy of 95% with the implementation of BERT. Three classifiers—Random Forest, Logistic Regression, and Multinomial Naive Bayes—achieved the highest accuracy of 54% for classical machine learning models using the Title dataset. Using the excerpt dataset and Multinomial Naive Bayes, we achieved an accuracy of 81%. In the future, there are plans to increase the size of our dataset to enhance the performance of our proposed approach.

Conflict of Interest

The authors declare no conflict of interest.

Authors Contribution

Ayesha, Kishowloy, Sadman, and Raqeebir initiated the research idea. Ayesha, Kishowloy, and Sadia conducted the background study. Ayesha, Kishowloy, Sadman, and Sadia were involved in the meticulous process of data acquisition and preparation. Ayesha undertook data processing and the implementation of classical machine learning techniques, while Kishowloy specialized in the implementation of deep learning techniques. Ayesha and Kishowloy conducted thorough comparisons of results. Asifur, Ayesha and Sadia were responsible for the composition and formatting of the paper. Raqeebir provided supervision, particularly focusing on the methodology, implementation, and writing aspects of the research paper. The contributions of the authors extended beyond the specific tasks mentioned above. All authors made significant contributions to the execution of the research paper and had approved the final version.

References

[1] A. Karadeniz and R. Can, "A research on book reading habits and media literacy of students at the faculty of education," *Procedia-Social and Behavioral Sciences*, pp. 4058–4067, 2015.

[2] What People Read around the World. Studying in Switzerland. [Online]. Available: https://studyinginswitzerland.com/what-people-read-around-the-world/

[3] C. R. Miller, "Genre as social action," *Quarterly Journal of Speech*, vol. 70, no. 2, pp. 151–167, 1984

[4] What Are Book Genres? A Detailed Guide. (8 February, 2022). [Online]. Available: https://becomeawritertoday.com/what-are-book-genres

[5] C. P. Biswas. (2023). Bengali Language and Market Economy. [Online]. Available: https://www.daily-sun.com/printversion/details/675239

[6] M. Ostendorff, P. Bourgonje, M. Berger, and J. Moreno-Schneider, G. Rehm, and B. Gipp, "Enriching BERT with knowledge graph embeddings for document classification," arXiv preprint, arXiv:1909.08402, 2019.

[7] S. Gupta, M. Agarwal, and S. Jain, "Automated genre classification of books using machine learning and natural language processing," in *Proc. 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, 2019, pp. 269–272. doi: 10.1109/CONFLUENCE.2019.8776935

[8] S. Gupta, M. Agarwal, and S. Jain, "Automated genre classification of books using machine learning and natural language processing," in *Proc. 2019 IEEE 9th International Conference on Cloud Computing, Data Science & Engineering*, 2019, pp. 269–272.

[9] Rokomari. (2023). [Online]. Available: https://www.rokomari.com/book

[10] E. Ozsarfati, E. Sahin, C. J. Saul, and A. Yilmaz, "Book genre classification based on titles with comparative machine learning algorithms," in *Proc. 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, 2019.

[11] A. Finn and N. Kushmerick, "Learning to classify documents according to genre," *Journal of the American Society for Information Science and Technology*, vol. 57, pp. 1516–1518, 2006.

[12] Y. Kim and S. Ross, "Examining variations of prominent features in genre classification," in *Proc. the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, Waikoloa, HI, USA, 2008, pp. 132–132. doi: 10.1109/HICSS.2008.157

[13] V. Battu, V. Batchu, R. R. R. Gangula, M. M. K. R. Dakannagari, and R. Mamidi, "Predicting the genre and rating of a movie based on its synopsis," in *Proc. the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong, Association for Computational Linguistics, 2018.

[14] A. C. Saputra, A. B. Sitepu, Stanley, P. W. P. Y. Sigit, P. G. S. A. Tetuko, and G. C. Nugroho, "The classification of the movie genre based on synopsis of the Indonesian film," in *Proc. 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, Yogyakarta, Indonesia, 2019, pp. 201–204. doi: 10.1109/ICAIIT.2019.8834606

[15] A. M. Ertugrul and P. Karagoz, "Movie genre classification from plot summaries using bidirectional LSTM," in *Proc. 2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, Laguna Hills, CA, USA, 2018, pp. 248–251. doi: 10.1109/ICSC.2018.00043

[16] G. Portolese and V. D. Feltrim, "On the use of synopsis-based features for film genre classification," in *Proc. National Meeting of Artificial and Computational Intelligence (ENIAC)*, 2018, doi: 10.5753/eniac.2018.4476

[17] H. Z. Mauni, T. Hossain, and R. Rab, "Classification of underrepresented text data in an imbalanced dataset using deep neural network," in *Proc. 2020 IEEE Region 10 Symposium (TENSYMP)*, Dhaka, Bangladesh, 2020, pp. 997–1000. doi: 10.1109/TENSYMP50017.2020.9231021

[18] Sagorsarker/bangla-bert-base. Hugging Face. [Online]. Available: https://huggingface.co/sagorsarker/bangla-bert-base