Empirical Text Analysis for Identifying The Genres of Bengali Literary Work

Ayesha Afroze[1], Kisholoy Dutta [2], Sadman Sadik [3], Sadia Khanam [4], Raqeebir Rab[5], and Mohammad Asifur Rahim [6]
[1,2,3,4,5,6]Department of Computer Science and Engineering
[1,2,3,4,5,6]Ahsanullah University Of Science and Technology (AUST), Dhaka, Bangladesh
Email- ayeshaafrozeaust,kishowloydatta016,sadmansadikhasan,sadiakhanamarni111@gmail.com
raqeebir.cse@aust.edu, mohammadasifurrahim@gmail.com

***Abstract*—Digital books and internet retailers are growing in popularity daily. Different individuals prefer various genres of literature. The categorizing of 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 it 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 amount of words present in the dataset, the performance varied.**

***Index Terms*—Genre, LSTM, CNN, BERT, book titles, SVM, Natural Language Processing, Book Snippets. RNN**

I. Introduction

Books continue to be a significant part of our lives, with reading being a popular pastime for many people[1]. In paper[1], it is shown that book reading is the most effective learning activity through which an individual can improve himself in terms of critical thinking, developing new and different perspectives, understanding himself and the world, and interpreting the events and situations that he will encounter. The global book market is growing[2]. According to a country analysis published by global market researchers, 30% of the global Internet population reads books every day or most days[3][2]. When it comes to reading, people have varying preferences. The growth of this market will be aided by presenting consumers with items that appeal to them. Throughout their history, books have undergone significant transformations. The market has evolved in many forms, from papyrus scrolls to the introduction of e-books with the advent of digital media and new technology. Since purchasing and reading have shifted to digital platforms, automatic classification of books into various genres is crucial. A genre describes the types of novels that belong to a particular category. The many genres cater to distinct audiences in order to satisfy a variety of needs. A genre refers to the types of books that fall into that category[4]. They are categorized by their style, tone, time period, target audience, and numerous other factors[6]. The various genres are geared towards different audiences to suit a variety of demands.

With approximately 300 million native speakers and another 37 million as second language speakers, Bengali is the fifth most-spoken native language and the seventh most-spoken language by the total number of speakers in the world[5][6][7]. Numerous of these individuals read novels online. Our research paper will enrich the reading experience of these individuals by classifying books into various genres using classic machine learning and deep learning models. In the Bengali literature, the classification of genres is deficient. Based on the titles and snippets of various Bengali literary works, we divide them into several genres in this paper. Such research works existed previously for other languages[3][4][5], but not for Bengali literature. Due to the lack of a suitable dataset, we were required to generate one from scratch. We produced the data set, which includes titles and excerpts from Bengali novels of various genres. We selected the online bookstore Rokomari.com[8] for data collection, since only on this platform we find a large selection of Bengali novels from which we could collect snippets. One dataset has a total of 12,925 Bengali book titles, encompassing nine distinct genres. In contrast, the other dataset consists of 452 Bengali book excerpts, which are categorized into three different genres. In our research, we employed various machine learning algorithms for classification purposes, including logistic regression, support vector machines (SVM), random forest classifiers, decision trees,

[1]https://studyinginswitzerland.com/what-people-read-around-the-world
[2]https://www.grandviewresearch.com/industry-analysis/books-market
[3]https://www.gfk.com/press/majority-of-internet-users-read-books-either-daily-or-at-least-once-a-week
[4]https://becomeawritertoday.com/what-are-book-genres
[5]Wikipedia contributors, "language Wikipedia, the free encyclopedia"
[6]https://www.cia.gov/the-world-factbook/countries/world/
[7]Ethnologue: Languages of the World https://www.ethnologue.com
[8]https://www.rokomari.com/book

recurrent neural networks (RNN), long short-term memory (LSTM), convolutional neural networks (CNN), and bidirectional encoder representations from transformers (BERT). BERT demonstrates superior performance compared to all other models for both datasets, with a remarkable accuracy rate of 90% for the book snippets dataset and 77% for the book titles dataset. The results indicate that conventional machine learning models generally outperformed on the dataset of snippets, with the notable exception of BERT. Conversely, deep learning models demonstrated superior performance on the dataset of titles.

The paper is organized as follows: The introduction describes the rationale and aims of our research. The relevant literature reviews of our research are addressed in Section 2. Our constructed dataset is discussed in Section 3, while our developed models are described in Section 4. The fifth section analyzed the results of our model simulations. Section 5 was concluded with a conclusion and discussion of future work.

II. Literature Review

In this section, we have examined the scholarly articles in relation to our research endeavors. This section provides an overview of previous research studies that have explored genre classification through the use of various methodologies. The text emphasizes the researchers' focus on their research methodology and the challenges they faced during the process. Furthermore, it highlights their successful resolution of these challenges by adopting novel recommended methodologies and models, resulting in improved accuracy and improved assessment metrics. The literature review section has been divided into two distinct segments. The first method of classification involves utilizing the title section, whereas the second method involves employing the snippet segment.

A. Classification based on title

Eran et al[7] proposed a book genre classification based on titles using different machine learning algorithms, which are RNN, GRU, LSTM, Bi-LSTM, CNN and Naive Bayes. The dataset contained 207575 samples of data, with each title corresponding to one of 32 different genres, and it was retrieved from Amazon's library. The LSTM model has the highest accuracy due to its ability to maintain memory over long-term dependencies. Gupta et al.[8] proposed an automated genre classification of books using ML algorithms and NLP. This method acquires a large number of words and transforms them into feature matrices using TF-IDF on labeled data. Labeled data was 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. Shiroya et al.[2] 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. The first dataset had accuracy results of 2.68%, 9.53%, and 7.27% in KNN, 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%. Finn at el.[9] Showed ways of learning to classify documents according to genre. Two sorts of genre classification tasks are done, first one is if an article is subjective or objective, and another one is if a review is positive or negative.

Kim at el.[10] 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 set 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. Here, 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. [11] provided a method for supplementing bidirectional encoder representations from transformers with knowledge graph embeddings and additional metadata in this work. They enhanced the accuracy of standard BERT models by up to four percentage points using this strategy. Four popular datasets were included in the analysis. They limited the input length to 300 tokens to reduce GPU memory consumption. A separate preprocessing phase was used to generate the non-text features. The three representations were then concatenated and fed into an MLP with two layers of 1024 units each and a RELU activation function. Finally, the output layer performed the classification. Each unit in the Softmax output layer corresponds to a class label. They used a micro-averaged F1 score for evaluation. With an F1 score of 87.20 for task A and 64.70 for task B, the setup using BERT-German with metadata features and author embeddings (1) outperformed all other setups. Looking only at the precision score, BERT-German with metadata features (2) without author embeddings performed best. Their findings demonstrated that including task-specific information such as author names and publication metadata significantly enhances the classification job when compared to a text-only strategy.

B. Classification based on snippet

Battu at et.[12] predicted movie genres using ratings and synopses in multiple languages, including Hindi, Telugu, Tamil, etc. Mining data from seven different websites, they pre-processed the data to group the genres into classes using rating and genre as data points. They combined the data and divided them into two portions for training and testing, each containing 80% and 20% of the total data. CNN and RNN were used for character embedding. SVM, random forest, and a hybrid model were used here. Saputra at el.[13] presented a text classification model to identify the genre of Indonesian film using synopsis. CNN, RNN, and LSTM were used in this study. For pre-processing, they have used different such as

Repeat Character and Spell Normalization, Text Tokenization, Text Stemming, POS Tagging, Special Characters Removal, Stop words Removal, etc. They have used TF-IDF and Bag-of-Words (BOW) for feature extraction. SVM classification algorithm with TF-IDF extraction was able to find the best accuracy and F1 scores after training data (45%). Using the Bi-LSTM network, Ertugrul at el.[14] tried to classify movies based on storyline summaries. They uniformly sampled data based on their genres in the document-level categorization challenge. For training, the data for sentence-level categorization is uneven. For the genre classification task, they received a total of 6.360 movies and 22.278 sentences. They used Bi-LSTM to accomplish multiclass movie genre classification from plot summaries, with the class labels thriller, horror, comedy, and drama. They labeled it a document-level technique when they used the entire plot summary for training without separating it into sentences. They trained a general RNN model using sentence-level and document-level approaches. They compared the results of their method with a Bi-LSTM model trained using a document-level approach. They observed that when data were limited, using phrases to specify the movie genre performed better than using the overall plot summary of the recurrent neural network.

The use of text-based traits derived from movie summaries for the purpose of multilabel movie genre classification was investigated by Portolese et al.[15]. The synopses were extracted from the Movie Database (TMDb) website. All movies were classified into 12 classes, namely adventure, action, comedy, crime, drama, fantasy, horror, mystery, romance, science fiction, thriller and war. All combinations were evaluated using 5-fold cross-validation and the final classification metrics were obtained by averaging the results from each fold. The results of the best experiment present the following averaged scores between the 12 genres present in the data set: precision of 57. 61%, recall of 53. 36%, and f1 score of 54. 80%. Several studies have been made to categorize genres across different disciplines and languages, however, there has been limited focus on the classification of genres in Bangla literature. This research presents a proposed methodology for classifying genres in Bangla literature based on book titles and snippets.

III. Data Acquisition and Preparation

To facilitate data collection, we opted for an online platform due to its practicality and convenience. Rokomari.com[17] has been identified as the primary online platform for the purchase of Bengali books. The platform was chosen for data collection because of its extensive library of 200,000 books. The books within its collection are classified into many genres. This eased the process of categorizing datasets. One additional rationale for choosing this platform was its capability of allowing online reading of certain sections of books, aiding in the creation of a dataset comprising book snippets. Two datasets have been generated for the purpose of our research. The two 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, namely 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. The book titles were extracted from the website[17]. Beautiful Soup, a Python library[9], 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 pertaining to various genres individually and subsequently categorized them based on the genres available on [17]. In this way, we gathered data pertaining to 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 I shows data distribution among different genres of book title datasets.

TABLE I
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.

Data Acquisition for Book Snippets The term "book snippets" refers to concise segments or excerpts extracted from books. In order to compile a collection of snippets, we took advantage of the opportunity to peruse a selection of pages from books that are exclusively available through Rokomari.com[17]. We read those pages and took snapshots of those pages. Then we extracted the texts from those images. To extract text from images, we employed an online program called Image to Text[10]. Subsequently, the snippets were categorized based on the categories of [17]. The distribution of data among several genres of book title datasets is presented in TableII.

[9]https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[10]https://www.imagetotext.info

TABLE II
DATA DISTRIBUTION AMONG DIFFERENT GENRES OF BOOK SNIPPET DATASET

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

Genre selection and labelling One of the variables influencing the selection of the Rokomari.com[17] website was its genre labeling feature. The books on the website have been systematically classified into many genres. Following the process of data scraping, we proceeded to manually assign labels to the obtained data. Many genres of literature exist, and the rationale behind their selection lies in the fact that each of these genres encompasses a substantial collection of over a thousand novels that are considered suitable for both machine learning and deep learning algorithms.

IV. METHODOLOGY

This section discussed the approaches employed for the identification of genres in Bengali literature. In this study, two data sets were utilized. One data set comprises a collection of book titles, while the second data set 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 Network (RNN), Long Short-Term Memory Networks (LSTM), Convolution Neural Network(CNN), Bidirectional Encoder Representations from Transformers (BERT)

The sequential processes of the proposed methodology are illustrated in Figure1.

A. Data Preprocessing

Preparing raw data to be acceptable for a machine learning model is known as data pre-processing. The initial and pivotal phase in constructing a machine learning model involves data pre-processing. Data are often characterized by noise, missing values, and suboptimal formatting, making them unsuitable for the direct application of machine learning models. Data pre-processing is an essential step in the data analysis pipeline, as it involves cleaning and transforming the raw data to enhance its quality and make it suitable for utilization in a machine learning model[11]. This procedure plays a crucial role in improving the accuracy and efficacy of the model. The process of data preprocessing is shown in Figure2. The very first step was to clean the data. Data cleaning is a crucial part when it comes to handling text data. Data contains many

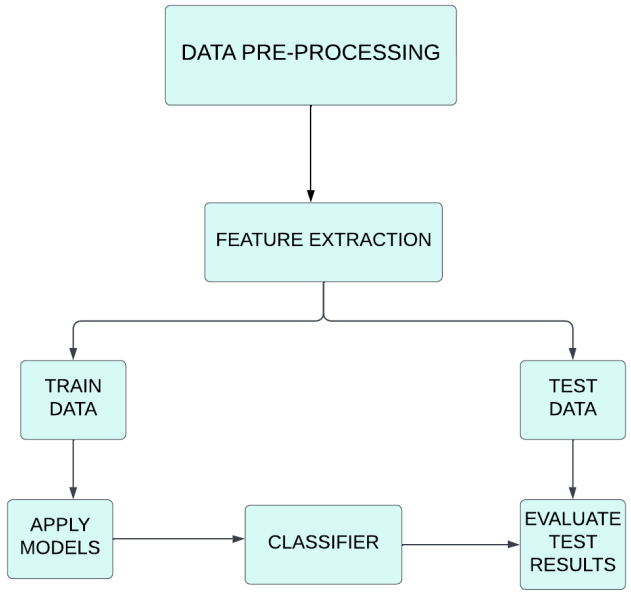

Fig. 1. Flowchart of proposed methodology

such things which are not useful and hamper the performance of the algorithms. Removal of these helps to enhance the performance of the models. We removed the punctuation marks along with different kinds of symbols and digits as they are insignificant in these models. Additionally, during the cleaning process, we eliminated the use of the English alphabet.

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[12]. Next, the stop words were eliminated. These words do not contribute any substantial information to our current model. The process of removing stop words helps to improve the visibility of significant information. 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.

Data Balancing

The data set for the title of the book is imbalanced. A dataset that is imbalanced tends to classify all the data as belonging to the majority class. To address this issue, we balanced our dataset. We utilize the Tomek-Link undersampling technique. Tomek-Link[13] to eliminate samples of the majority class that are closest neighbors of samples of the minority class. Tomek connections are pairings of instances from opposing classes

[11]https://towardsdatascience.com/nlp-in-python-data-cleaning-6313a404a470

[12]https://towardsdatascience.com/text-pre-processing-stop-words-removal-using-different-libraries-f20bac19929a

[13]https://towardsdatascience.com/imbalanced-classification-in-python-smote-tomek-links-method-6e48dfe69bbc

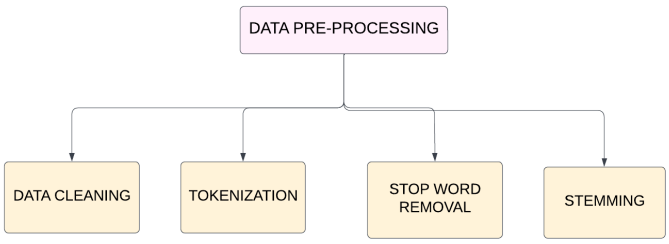

Fig. 2. Steps of data pre-processing

that are close to one another. We eliminated the majority of these Tomek-Links class samples. This contributes to a better decision boundary and more balanced data.

B. Feature Extraction

After pre-processing, the next step is feature extraction. Feature extraction is a fundamental method that plays a pivotal role in enhancing our understanding of the contextual aspects that we are addressing. After the initial text has been cleaned, it must be converted into its features in order to be utilized 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 that is used to convert textual data into numerical representations, allowing machine learning models to comprehend and process such data. Mauni at el.[16] demonstrated that extracting a set of features with efficient algorithms not only reduces the dimensions of the feature space but also eliminates redundant features from the model.

Feature Extraction for Classic Machine Learning Models Term Frequency Inverse Document Frequency of records (TF-IDF) has been employed as a method for extracting 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 effectively distinguish and identify a document. The greater the TF-IDF value of a word, the more unique and uncommon its occurrence. It takes into account the importance of each word and is computationally inexpensive.

Feature Extraction for Deep Learning Models The use of 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. In this paper, we used the BERT word embedding technique as a deep learning approach for information retrieval. The Bangla BERT model is which is pre-trained using Bengali Wikipedia[18]. This model performed the best with the proposed model architecture for deep learning.

C. Genere 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. Multimodal Naive Baised (NB), Random Forest (RF), and Logistic Regression (LR) have the highest accuracy for the title dataset, while Multimodal Naive Baised has the highest accuracy for the excerpts dataset.

D. Genere Classification with Deep Learning Models

In this study, four deep-learning models were employed for the purpose of 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. An LSTM layer was incorporated into the model architecture, consisting of 128 units. In order to mitigate the issue of overfitting, the utilization of 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 the dense layer, which will yield the classification. We have chosen to utilize the softmax activation function for our model. The parameters pertaining to the layers of this model can be comprehensively inferred from the Figure 3 shown herein.

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_4 (Embedding)     (None, 600, 150)          15000000

 spatial_dropout1d (SpatialD  (None, 600, 150)         0
 ropout1D)

 lstm (LSTM)                 (None, 128)               142848

 dropout (Dropout)           (None, 128)               0

 dense_6 (Dense)             (None, 9)                 1161

=================================================================
Total params: 15,144,009
Trainable params: 15,144,009
Non-trainable params: 0
```

Fig. 3. Training parameters of LSTM Model

Convolutional Neural Network (CNN) In our study, we employed a convolutional neural network architecture that is based on deep learning techniques to carry out the task of intent classification for textual commands. However, convolutional neural networks (CNNs) are commonly linked with computer vision tasks. Convolutional neural network (CNN) kernels play a crucial role in discerning pertinent patterns

inside textual input. The 1D convolution layer is responsible for generating a convolution kernel that is applied to the input of the layer along a single spatial dimension. The pooling layer employed in our research was Maxpool1D. The dimensions of the pool were measured to be 3 units. The flattened layer is utilized to modify the dimensional shape of the outputs. Finally, the output is passed through the dense layer, which will yield the classification. We have chosen to utilize the softmax activation function for our model. The parameters pertaining to the layers of this model can be comprehensively inferred from Figure 4 shown herein.

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 600, 150)	15000000
conv1d (Conv1D)	(None, 598, 150)	67650
max_pooling1d (MaxPooling1D)	(None, 199, 150)	0
dense_1 (Dense)	(None, 199, 20)	3020
flatten (Flatten)	(None, 3980)	0
dense_2 (Dense)	(None, 13)	51753

Total params: 15,122,423
Trainable params: 15,122,423
Non-trainable params: 0

Fig. 4. Training parameters of CNN Model

Recurrent Neural Network (RNN) The Bidirectional Recurrent Neural Network (RNN) has been employed in our study. The recurrent neural network (RNN) is a prevalent neural network structure employed in the field of natural language processing (NLP). The approach has demonstrated a relatively high level of accuracy and efficiency in language acquisition. The layers of our recurrent neural network (RNN) model were as: 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 embed dim. The spatial dropout layer is utilized in order to mitigate overfitting by selectively dropping nodes. The value of 0.4 represents the probability at which the nodes need to be dropped. The bidirectional layer is a recurrent neural network (RNN) layer containing long short-term memory (LSTM) units, and it has a dimensionality of 128. Finally, the output is passed through the dense layer, which will yield the classification. The parameters pertaining to the layers of this model can be comprehensively inferred from Figure 5 presented below.

Bidirectional Encoder Representations from Transformers (BERT) DistrilBERT was employed for the purpose of categorization. DistilBERT is a compact, efficient, cost-effective, and lightweight Transformer model that has been trained through the process of distillation, using a BERT base as its source. The model exhibits a reduction of 40% in the number of parameters compared to Bert-base-uncased. Additionally, it

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. 5. Training parameters of RNN Model

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. DistrilBERT was fine-tuned using Ktrain in our study. Ktrain is a software library that serves as a lightweight wrapper for the TensorFlow Keras framework. DistrilBERT is fine-tuned using a learning rate of 8e-5. The parameters of the layers of this model can be inferred in detail from Figure 7 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. 6. Training parameters of BERT Model

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

V. Result Analysis

We will discuss the performance of our models. We have employed various evaluation metrics, including accuracy, precision, F1 score, and recall, for this purpose. 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

TABLE III
PARAMETER AND HYPERPARAMETER OF DEEP LEARNING MODELS

Model	Optimizer	Activation Function	Loss fuction	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

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 IV presents the performance metrics, namely accuracy, macro F1-score, precision, and recall, for various machine learning models applied to the title dataset. Based on the findings presented in the table IV, it can be observed that the accuracy and F1-score of all the models exhibit a high degree of similarity. 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. Figure 7 illustrates the performance graph of many classical machine-learning algorithms.

TABLE IV
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%

Table V presents the performance metrics, including accuracy, macro F1 score, precision, and recall, of various deep learning models applied to the title dataset. BERT demonstrates superior performance compared to other Deep Neural Network models, exhibiting an accuracy rate of 77%. The performance of this model exhibits significant improvement in comparison to the other models. Recurrent Neural Networks (RNN) and Long Short-Term Memory(LSTM) exhibit superior

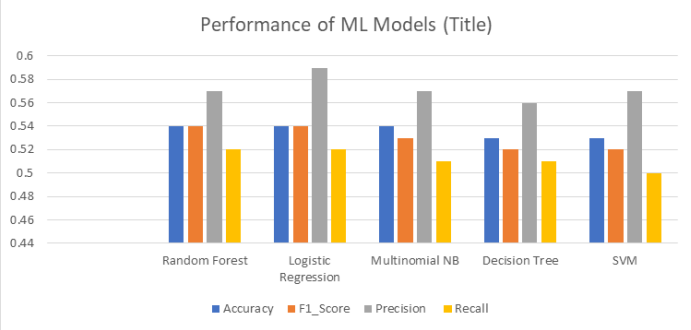

Fig. 7. Performance of Classic Machine Learning Models Using Title Dataset

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. Figure8 displays the performance graph of many traditional machine learning models.

TABLE V
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%

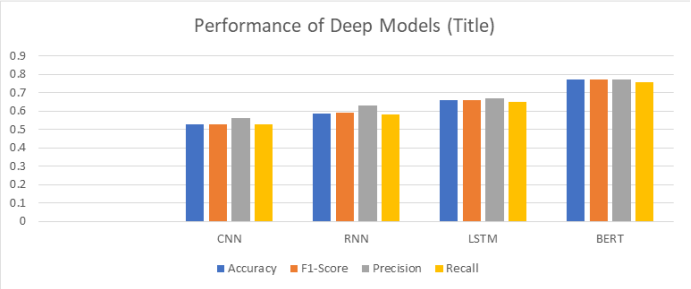

Fig. 8. Performance of Deep Learning Models Using Title Dataset

B. Result Analysis for the Snippet Dataset

Table VI lists the precision, recall, precision, and macro F1 score of various machine learning models for the data set. Multinomial NB produced the highest performance in the snippet data set, yielding an average of 81% for all scores. Random Forest produces the second-best result, with nearly 80% of all scores. Figure 9 provides a visual representation of the performance graph.

Table VII gives the Deep Model outcome analysis for the Snippet's dataset. In the snippet dataset, BERT outperforms all other deep models, with 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

TABLE VI
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%

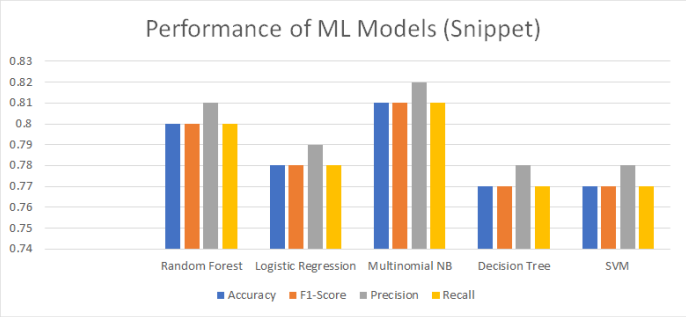

Fig. 9. Performance of Classic Machine Learning Models Using Snippet Dataset

of data to train. 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 Figure 10.

TABLE VII
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	95%	93%	94%	94%

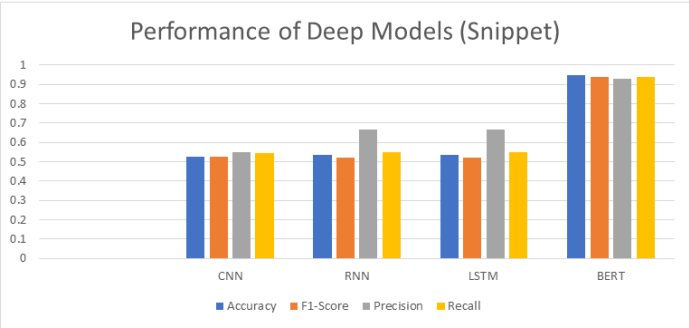

Fig. 10. Performance of Deep Learning Models Using Title Dataset

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, hence 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 there is a sufficient amount of words 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 12925 samples, the length of each sample is very short because titles are typically brief. Eran at el.[7] used 207575 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. Figure 11 is a graph depicting the efficacy of all models in both datasets.

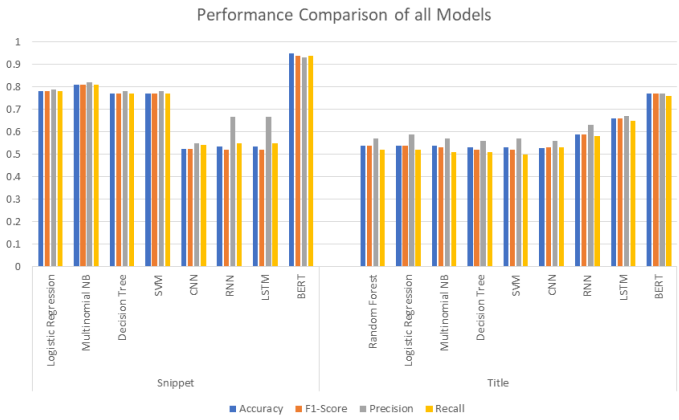

Fig. 11. Performance comparison among all models.

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. In order to achieve the stated goal, two datasets were first 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 have been developed for both datasets. Thus far, in the domain of neural networks, employing the title dataset, we have attained the highest accuracy of 77% through the utilization of BERT. By using the snippet dataset, we have 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 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 is a plan to increase the size of our dataset in order to enhance the performance of our proposed approach.

VII. Conflict of Interest

The authors declare no conflict of interest.

VIII. Authors Contribution

Each author has contributed equally in conducting the research, analyzing the data, and writing the paper. All authors had approved the final version.

References

[1] Abdulkerim Karadeniz and Remzi 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," 2022, [Accessed 24 September 2023].

[3] Ostendorff, Malte and Bourgonje, Peter and Berger, Maria and Moreno-Schneider, Julian and Rehm, Georg and Gipp, Bela, "Enriching bert with knowledge graph embeddings for document classification," arXiv preprint arXiv:1909.08402, 2019.

[4] S. Gupta, M. Agarwal and S. Jain, "Automated Genre Classification of Books Using Machine Learning and Natural Language Processing," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 269-272, doi: 10.1109/CONFLUENCE.2019.8776935..

[5] Shiroya, Parilkumar and Vaghasiya, Darshan and Soni, Meet and Panchal, Brijeshkumar, "Book genre categorization using machine learning algorithms (K-nearest neighbor, support vector machine and logistic regression) using customized dataset," Int. J. Comput. Sci. Mobile Comput, vol. 10, pp. 14-25, 2021.

[6] Editor, "What Are Book Genres? A Detailed Guide," 8 February 2022. [Online]. Available: https://becomeawritertoday.com/what-are-book-genres. [Accessed 24 September 2023].

[7] Ozsarfati, Eran and Sahin, Egemen and Saul, Can Jozef and Yilmaz, Alper, "Book genre classification based on titles with comparative machine learning algorithms," in 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), 2019.

[8] Gupta, Shikha and Agarwal, Mohit and Jain, Satbi, "Automated Genre Classification of Books Using Machine Learning and Natural Language Processing", in 2019 IEEE 9th International Conference on Cloud Computing, Data Science & Engineering, 2019, pp. 269-272.

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

[10] Y. Kim and S. Ross, "Examining Variations of Prominent Features in Genre Classification," Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), Waikoloa, HI, USA, 2008, pp. 132-132, doi: 10.1109/HICSS.2008.157.

[11] Malte Ostendorff and Peter Bourgonje and Maria Berger and Julian Moreno-Schneider and Georg Rehm and Bela Gipp, "Enriching BERT with Knowledge Graph Embeddings for Document Classification," arXiv preprint arXiv:1909.08402, 2019.

[12] A. C. Saputra, A. B. Sitepu, Stanley, P. W. P. Yohanes Sigit, P. G. Sarto Aji Tetuko and G. C. Nugroho, "The Classification of the Movie Genre based on Synopsis of the Indonesian Film," 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), Yogyakarta, Indonesia, 2019, pp. 201-204, doi: 10.1109/ICAIIT.2019.8834606.

[13] A. C. Saputra, A. B. Sitepu, Stanley, P. W. P. Yohanes Sigit, P. G. Sarto Aji Tetuko and G. C. Nugroho, "The Classification of the Movie Genre based on Synopsis of the Indonesian Film," 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), Yogyakarta, Indonesia, 2019, pp. 201-204, doi: 10.1109/ICAIIT.2019.8834606.

[14] A. M. Ertugrul and P. Karagoz, "Movie Genre Classification from Plot Summaries Using Bidirectional LSTM," 2018 IEEE 12th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 2018, pp. 248-251, doi: 10.1109/ICSC.2018.00043.

[15] Portolese, Giuseppe and Feltrim, Valéria Delisandra, "On the use of synopsis-based features for film genre classification," in Anais do XV Encontro Nacional de Inteligência Artificial e Computacional, 2018, doi=10.5753/eniac.2018.4476 .

[16] H. Z. Mauni, T. Hossain and R. Rab, "Classification of Underrepresented Text Data in an Imbalanced Dataset Using Deep Neural Network," 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 2020, pp. 997-1000, doi: 10.1109/TENSYMP50017.2020.9231021.

[17] rokomari, https://www.rokomari.com/book,year=2012,[Online; Accessed 24 September 2023].

[18] sagorsarker/bangla-bert-base, Hugging Face." https://huggingface.co/sagorsarker/bangla- bert-base, dec 2 2022. [Online; accessed 2023-12-10].