

# BdRegionText: resource creation and evaluation for Bangla regional text classification with machine learning

Babe Sultana<sup>1,2</sup>, S. M. Mirajul Hoque<sup>3</sup>, Md Gulzar Hussain<sup>4</sup>, Mohammad Nurul Huda<sup>2</sup>

<sup>1</sup>Department of Computer Science and Engineering, Faculty of Science and Engineering, Green University of Bangladesh, Dhaka, Bangladesh

<sup>2</sup>Department of Computer Science and Engineering, Faculty of Science and Engineering, United International University, Dhaka, Bangladesh

<sup>3</sup>Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka, Bangladesh

<sup>4</sup>School of Software, Nanjing University of Information Science and Technology, Nanjing, China

## Article Info

### Article history:

Received Oct 2, 2024

Revised Apr 16, 2025

Accepted Jul 3, 2025

### Keywords:

Bangla regional text

BoW

DT

Random forest

SGD

SVM

TF-IDF

## ABSTRACT

Regional text analysis acknowledges the cultural diversity encompassed by a language. It offers insights into the authentic ways people communicate, promoting cultural awareness and genuineness in communication. This research paper delves into the classification of Bangla regional text using machine learning (ML) algorithms. Consequently, this study compiles a dataset comprising 2,573 sample texts in four distinct regional Bangla dialects (Chittagong, Rangpur, Barishal, and Noakhali). We focused on these dialects because they were more readily available on the internet than others. The primary objective is to identify synthesized Bangla text and assign appropriate categories. The categorization process focuses on a regional language authored by Bengali individuals, aiming to ascertain its authenticity and using ML techniques named decision tree (DT), stochastic gradient descent (SGD), support vector machine (SVM), and random forest (RF) to check how well categorization worked and also handled the issue of slight imbalance in the dataset. As there is limited prior research in this domain, we compare our work with the existing studies available, and we have employed various popular feature extraction techniques for text classification in natural language processing (NLP), specifically TF-IDF, CountVectorizer, and bag of words (BoW). Our comparative analysis indicates that an aggregation of term frequency-inverse document frequency (TF-IDF) and CountVectorizer outperforms BoW in terms of performance. Among the ML techniques we applied, the RF algorithm yielded the utmost accuracy of 79.15% and a mean accuracy of 79.47%.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Babe Sultana

Department of Computer Science and Engineering, Faculty of Science and Engineering

Green University of Bangladesh

Dhaka, Bangladesh

Email: [babecse@gmail.com](mailto:babecse@gmail.com)

## 1. INTRODUCTION

The Bengali language holds a prominent global standing, frequently acknowledged as among the most extensively used languages across the world. It ranks as the fifth most spoken language [1]. With time, the Bengali language has undergone several transformations in terms of its pronunciation and usage across various

regions. Typically, after birth, we instinctively adopt the language spoken by our mother or those around us as our own, without relying on grammatical rules and regulations. Every country exhibits variations based on its regional mother tongue. In contrast to the language we use while studying or at educational institutions, we often gravitate towards our regional language when communicating at home or with family members. The comfort we experience when speaking our local language and expressing our thoughts is incomparable. Across the 64 districts of Bangladesh, approximately 55 distinct regional languages are spoken [2]. It can be observed on social media that Bengali people tend to write in their regional languages. Based on the November 2017 report from BTRC, it was determined that Bangladesh has an estimated 25 to 30 million users on Facebook [3]. The current Facebook audience size for Bangladesh as of January 2023 stands at 43.25 million [4]. The prevalence of communicating in regional languages is evident through the creation of Facebook pages where individuals from specific areas interact using their respective local languages. Dialogues for dramas and movies are now being crafted in regional languages, and this trend extends to creating poems, songs, and compositions in these languages. As we mentioned, there are many regional languages in Bangladesh depending on the region, in this research, we focused on four such regional languages: Chittagong, Noakhali, Barisal, and Rangpur.

In the contemporary world, social media has become an indispensable element of virtually everyone's daily life. Individuals have a fondness for utilizing it to express their feelings and ideas. As a result, a substantial amount of textual data is accessible through these online resources, often appearing disorganized and untidy. The text data written in regional languages can be utilized to develop various natural language processing (NLP) tools. These tools might encompass tasks like word classification and sentiment analysis, social platform monitoring, and spam detection in the regional language context. Numerous researchers have engaged with text classification challenges using machine learning (ML) algorithms within the Bangla language domain. These endeavors have spanned diverse areas including news classification like Bangla news classification [5], Bengali accent classification [6], Bangla hate speech detection [3], Arabic news classification [7], multi-class sports news classification [8], sentiment analysis [9]-[11], multi-class textual emotion classification [12], Spam message detection [13], Mobile SMS spam message filtering [14], and Cyberbullying detection [15].

However, to the best of our knowledge, only two research studies have been conducted on regional Bangla text classification to date. Although numerous Bangla language corpora have been developed in various categories, such as MONOVAB [16], Vacaspati [17], Okkhor [18], Banglasenti [19], and BanglaLM [20], there is a noticeable gap when it comes to the creation of a corpus for regional Bangla text. Despite the limited resources of available datasets encompassing multiple Bangla regional texts, we took the initiative to create our dataset named BdRegionText by collecting data from various online sources. Our objective is to address this gap by creating a corpus specifically tailored to regional text, with plans to expand and improve it in the future. A research paper [2] undertook a study involving Bangla regional speech recognition, focusing on 7 regional dialects within the Bangla domain. However, despite such endeavors, limited research has been identified on Bangla regional text analysis. Very recently, two regional text datasets have been found, named Vashantor [21] in November 2023 and Bhashamul [22] in March 2024. Vashantor is essentially standard Bangla text translated into five specific regional dialects, and they mentioned that they aimed to translate the standard Bangla text to regional dialects and detect the region accurately.

However, we know that we often use English words when talking or texting in real-world language usage, even alongside regional texts. In our research, we focus on real-world language usage-how humans talk or text. As for Bhashamul, it is a result of a competition aimed at the exact IPA transcription of different dialects of Bengali. They used a real-world language format, but they mentioned that the minimum sentence length is 1, which is a challenging issue in accurately identifying which region the sentence belongs to. Such a short length does not provide sufficient information to represent the actual region and does not offer any semantic meaning that would allow for meaningful sentence analysis or context identification. As for our target, we aim to collect real-world usage texts that involve the context and meaning of the text. Every instance represents a complete sentence, and the text is diverse as it is collected from various sources, rather than being selected text simply translated into particular regions. This approach allows for a more accurate analysis of real-world perspectives. Here is Table 1, which provides an overall summary of different regional text datasets and their corresponding research analysis.

Table 1. Comparison of existing Bangla regional text dataset

Paper	Year	Language	Region	Contributions	Limitations
Rakshitha <i>et al.</i> [23]	2021	Hindi	Telugu, Tamil, Malayalam, Hindi, and Kannada	They scrape tweets from Twitter using its APIs, and later assign different sentiment scores to the customer reviews using TextBlob.	They used five different regional texts, which were collectively labeled as positive, negative, and neutral sentiments, but they did not specify what these positive sentiments signify in each particular region.
Faria <i>et al.</i> [21]	2023	Bangla	Chittagong, Noakhali, Sylhet, Barishal, Myensingh	Vashantor: a total of 32,500 sentences, incorporating Bangla, Banglish, and English, representing five distinct regional Bangla dialects.	Lacking linguistic diversity, the text followed a pattern translated into specific regional dialects, such as converting standard Bangla text into Chittagong region dialect, which did not reflect real-world language usage.
Islam <i>et al.</i> [22]	2024	Bangla	Rangpur, Narail, Tangail, Narsingdi, Kishoreganj	Bhashamul: the dataset includes 30,311 training samples and 8,941 test samples, representing six distinct regions.	The minimum sentence length is 1, making it challenging to accurately identify which region the sentence belongs to, as such short length does not provide sufficient information to represent the actual region.

The process of placing a group of documents into predetermined categories is known as text classification. This process is automated, performing the classification automatically. Automatic text classification frequently employs ML techniques. While numerous research papers address text classification issues across various domains, there is a relatively limited number of research papers focused on regional text classification. In this research paper [24], the authors dealt with the Marathi document classification problem. Some ML algorithms have been applied to the data set for Marathi document classification such as support vector machine (SVM), Naïve Bayes (NB), KNN, and Ontology. However, the research did not delve into assessing these algorithms' accuracy or performance in detail, thereby leaving their effectiveness unexplored. Shiravale *et al.* [25] applies the Stroke Width Transform and SVM algorithm to detect regions of this text within scene images (Indian Street). Despite the focus on image analysis, the study concluded that the SVM proved an efficient classifier for distinguishing between text and non-text elements. Rostam and Malim [26] emphasized the interconnected nature of the Quran and Hadith by incorporating both sources during the testing and training phases. Various classification methods, including NB, SVM, and KNN, were employed along with term weighting techniques, specifically term frequency-inverse document frequency (TF-IDF). Among these methods, SVM demonstrated superior performance compared to others. If we talk about the Bengali language limited works have been done on regional text classification in the Bengali language. Recently there has been work on regional speech classification which we have already mentioned [2]. Haque *et al.* [9] tackled multi-class sentiment classification in the context of Bengali social media comments. They employed the TF-IDF feature extraction technique before implementing the ML algorithms. Specifically, they applied random forest (RF), stochastic gradient descent (SGD), and decision tree (DT) algorithms. Notably, they discovered that RF and SGD exhibited the same performance, yielding an accuracy of 84.40%. Both of these algorithms outperformed the DT method. They introduced a novel CLSTM architecture, which combines convolutional neural network (CNN) and long short-term memory (LSTM) layers. This architecture was contrasted with six ML baseline models. Impressively, their proposed model achieved an accuracy of 85.8%, signifying a notable advancement in sentiment classification.

Additionally, Zhang [27] explored the application of LSTM with existing datasets from online resources, although not specifically in the Bengali language. They evaluated the performance of LSTM and demonstrated that the CNN-LSTM hybrid architecture achieved an accuracy of 93.61%. Indeed, variations in accuracy for the same architecture can arise due to differences in the characteristics of the datasets being used. With small datasets, the risk of overfitting is higher, and a complex architecture like CNN-LSTM could worsen this issue. Moreover, it requires more computational resources, which could be challenging with limited data and computational power. Since our created dataset is relatively small in size, it might not be suitable for our research work. The research project [28] at hand focuses on the development of an Arabic text classification model employing a range of algorithms and for feature extraction they have used 3 different models TF-IDF, bag of words (BoW), and character level. Among these methods are CNN, SVC, linear SVC, logistic regression (LR), Bernoulli Naïve Bayesian (BNB), Multinomial Naïve Bayesian (MNB), and SGD.

Notably, the CNN model achieved exceptional accuracy, surpassing state-of-the-art ML methods with a 98% accuracy. However, for Arabic, it's worth mentioning that the study does not delve into the issue of data augmentation, which remains a significant challenge for addressing data imbalances.

Additionally, enhancing text classification accuracy using a hybrid ensemble technique presents a challenging prospect in Arabic text classification. Wang *et al.* [29] introduces a novel guide network (GUDN) that incorporates label kind reinforcement techniques relying on basically label semantics. This strategy aids in the fine-tuning of pre-trained models for classification tasks. The experimental findings illustrate that GUDN surpasses existing state-of-the-art approaches when applied to the Eurlex-4k dataset and yields competitive outcomes on many other extensively used datasets. The use of the label reinforcement technique improves performance and effectively closes the semantic gap.

Ensemble learning entails a strategy within ML where predictions from numerous individual models are generated to produce a more potent and precise unified prediction. Among the ensemble methods, the RF algorithm is particularly prominent. In the domain of text classification, the application of the RF algorithm to imbalanced datasets has become a standard practice. More *et al.* [30] utilized a RF classifier to address the challenge of imbalanced big data classification. This choice allowed the classifier to effectively manage situations where the data distribution is uneven, leading to optimal classification accuracy. Padurariu and Breaban [31], the experiments conducted by the authors revealed that straightforward text presentations, such as BoW or TF-IDF, yield superior results for smaller datasets with significant imbalances when compared to more intricate embeddings like Glove, doc2vec, or FastText. Concerning the classifiers employed, the linear models, including LR and linear kernel SVM, exhibit a higher bias in situations of substantial class imbalance when contrasted with DT. This bias is more evident when using the TF-IDF representation and less pronounced when utilizing BoW. Jalal *et al.* [32], the author has focused on the RF algorithm to address text classification challenges. However, the author also introduces an innovative and enhanced RF algorithm. This novel approach combines both bootstrapping and random subspace methods concurrently, resulting in improved performance compared to the conventional RF algorithm. Due to the lack of resources, the number of Chittagong regional text samples is much higher than the other regional text samples. This situation has led to an imbalanced dataset, potentially favoring the performance of the RF algorithm. The following key statements are a summary of this work's main contributions:

- This research introduces BdRegionText, a textual Bangla dataset from different regions of Bangladesh, containing 2,573 texts from four regions: Chittagong, Barishal, Rangpur, and Noakhali.
- This dataset represents the authentic way of communication from various regions, where people commonly use their regional texts to interact with others.
- Using the BdRegionText dataset, this research compares the performance of popular text classification feature extraction techniques: TF-IDF, CountVectorizer, and BoW. It also applies four popular ML algorithms- DT, SVM, SGD, and RF-for Bangla regional text classification.
- Finally, this research demonstrates a comparative analysis of the results with other existing datasets in this domain better to illustrate the motivation and significance of this work.

The details description of the corpus creation process with the sample text format and other relevant things are discussed in section 2. Section 3 outlines our proposed research framework, while section 4 delves into the performance analysis of ML algorithms applied to our framework and also discusses our research findings. Section 5 offers overall discussion and insights into the future directions of our work.

## 2. CORPUS DESCRIPTION

Social media and online platforms have become some of the most important tools for easily understanding human thoughts, suggestions, reflections, and more through users' comments, posts, and other interactions. They also tend to express their interactions or reflections in regional text, which we refer to as real-world usage language, rather than formal language like pure standard English or their mother tongue. They may use a local or regional language in a code-mixed format, which is commonly used in everyday communication. However, Bengali-speaking individuals can frequently be observed commenting or sharing their thoughts using their respective regional languages on social media. In this rapidly growing trend, if we analyze online platforms, it is evident that people are happy and feel proud to text in their regional dialects. In this context, we aim to create a corpus for regional text in Bangladesh. The dataset that we have created can be found here [33], named

“BdRegionalText”, which reflects the main contribution of this research work. As there are limited resources available in the domain of Bangla regional datasets, up until September 2024, only 2 available datasets and corresponding research works have been found. We may consider this as the third Bangla regional text dataset, but in the context of real-world usage format, where the semantic meaning and full text are appropriately analyzed without using the same pattern text and merely translating it into multiple regions, it can be considered the first Bangla regional text dataset.

## 2.1. Data collection environment

As we already mentioned, the motivation of our research is to create a corpus for Bangla regional text that is used in regular communication, referred to as real-world usage text format. Therefore, our corpus creation or data collection environment is focused on online platforms, specifically online social media. The dataset was created by collecting text data from numerous online social platforms such as Facebook, YouTube, and various online resources, including comments, poems, and song lyrics. The data collection process was managed through human interaction. When we collected the data, we first analyzed the text to ensure it was genuinely from that region, rather than simply labeling it based on assumptions. For example, if a comment was written in the Chittagong dialect, we would also examine the replies to the comment to see if anyone pointed out spelling or language issues, which helped verify the regional authenticity.

Additionally, we focused on region-specific Facebook groups, where people from the same region typically post or comment using their local dialect. Collecting text from these groups made our approach more efficient and accurate. And we have also included some texts in this dataset, such as poems or song lyrics that are written in specific regional dialects. Overall, the dataset purely represents the selected four regional texts with full semantic meaning and ensures a complete text format.

## 2.2. Dataset overview

As previously mentioned, we selected four different regions from Bangladesh to create the dataset. A brief description of our dataset is provided in Table 2. The instances from the Chittagong region are comparatively higher than those from other regions due to their availability. We found many Facebook groups dedicated to the Chittagong region, where people are very active and eager to express their thoughts and reflections using their regional text, which resulted in a higher count of instances from this region.

Table 2. Dataset description table

Dataset description	
Platform	Facebook, Youtube, Instagram, Poem, Song Lyrics
Type of data	Text
Language	Bangla
Region	Chittagong, Rangpur, Noakhali, Barishal
Number of Chittagong regional text sample	1,167
Number of Barishal regional text sample	338
Number of Rangpur regional text sample	313
Number of Noakhali regional text sample	755

In Table 3, we provide samples of Bengali text from four distinct regions-Chittagong, Barishal, Rangpur, and Noakhali-collected from various sources. In this table, we have included the original Bengali text as spoken by people in their native dialects, along with the corresponding English translations for global understanding.

Table 3. Four regional Bengali text sample with Bangla meaning and english translation

Region	Regional text	Bangla meaning	English meaning
Chittagong	আঁত কসি গম ন লাগরে!!!!!! ☺	আমার কিছু ভালো লাগছে না !	I do not feel good!
Barishal	মুই বাইততে আইছি, তামেরা কডো কমেমে আছো??	আমি বাড়তি আসছি তামেরা কে কথোষ আছো?	I am in home, where are you all?
Rangpur	ক্যামনে আচনে বাহরো?>>>>	কমেন আছনে আপনারা?	How are you all?
Noakhali	আননে ভালো আছনে?	আপনি ভালো আছনে?	Are you fine?

### 3. METHOD

This study focuses on exploring system frameworks. In the absence of an available dataset, we gathered four regional Bengali text data and stored them. Subsequently, the textual data underwent preprocessing, which involved tokenization, punctuation removal, elimination of Bangla stop words, and eventual storage for feature extraction. Following this, features were extracted using TF-IDF, CountVectorizer, and BoW from the text, serving as input for the machine-learning techniques. Figure 1 illustrates the graphical overview of the process.

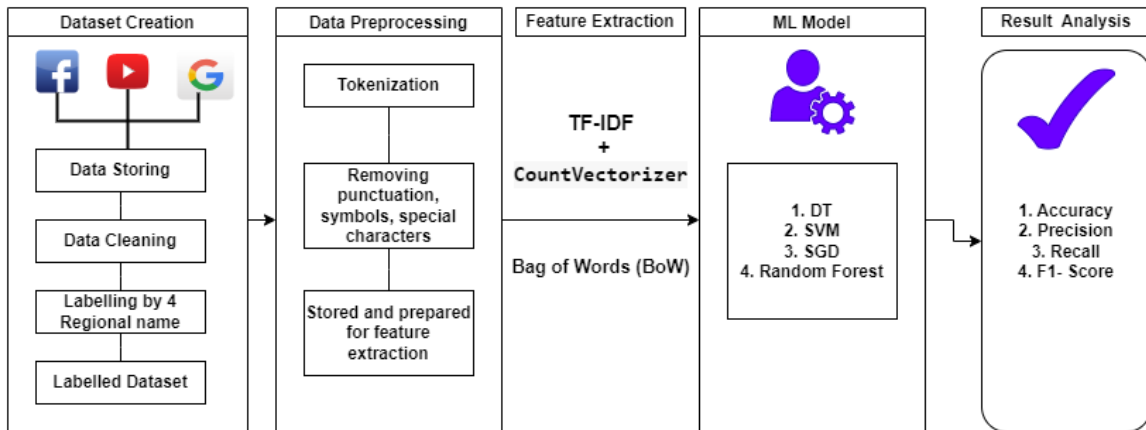


Figure 1. Framework overview: the framework encompasses dataset creation, data preprocessing, feature extraction, ML model, and result analysis

#### 3.1. Data preprocessing

Data preprocessing aims to reduce the textual noise, thereby enhancing the effectiveness of the classifier. These preprocessing steps cleanse the text data, rendering it suitable for utilization. The tokenization process is then used to extract features from text documents by disassembling the order of characters or words. Text documents should be divided into tokens to obtain the features because sentences in them are made up of a series of characters or words. Breaking down the text into sentences and then words is known as tokenization. During this process, less relevant symbols like !, x, >, <, \$, %, ~, and numerical values, are omitted since they hold minimal significance for classification purposes. We have collected data from various online resources which were noisy informal text that contained unnecessary symbols, emoji, and numerical values. To render this data suitable for utilization within the classifier model, a preprocessing step is imperative. The Table 4 displays the processed text after applying data preprocessing techniques to the original texts.

Table 4. Original text and processed text of Bangla regional text dataset

Region	Regional text	Bangla meaning
Chittagong	আঁতত কসি গম ন লাগরে!!!!!!! ☺	আঁতত কসি গম ন লাগরে
Barishal	মুই বাইতত, আইছ, তমেৱা কজে কমেমমে আছো???	মুই বাইতত, আইছ, তমেৱা কজে কমেমমে আছো
Rangpur	ক্যামনে আচনে বাহরো?>>>>	ক্যামনে আচনে বাহরো
Noakhali	আননে ভালা আছনেনি??	আননে ভালা আছনেনি

#### 3.2. Feature extraction

Feature extraction plays a pivotal role in text classification. Raw data cannot be directly input into ML because these models exclusively process numerical values. To facilitate model training, text inputs must be transformed into numeric features through encoding, effectively translating the text into numerical representations. In our study, we employed various text representation methods, including TF-IDF, CountVectorizer, and BoW. We applied ML algorithms to assess their comparative performance.

### 3.2.1. TF-IDF and CountVectorizer

Vectorization is the common process used to convert a set of text documents into numerical feature vectors. Using the CountVectorizer technique, a set of text documents is transformed into a matrix that counts the token instances in each document. The metric known as the TF-IDF is created when the ideas of term frequency (TF) and inverse document frequency (IDF) combine.

$$TF - IDF(t, d, D) = tf(t, d) * idf(t, D)$$

Here,

- $TF(t, d)$  = represents the term frequency of term  $t$  in document  $d$ , calculated as the number of times  $t$  appears in  $d$ .
- $IDF(t, D)$  = represents the IDF of the term  $t$  across the entire document collection  $D$ , calculated as the logarithm of the ratio of the total number of documents in  $D$  to the number of documents containing term  $t$ .

### 3.2.2. Bag of Words (BoW)

BoW, short for the BoW, stands as a foundational and extensively applied method within the realm of NLP. Its core concept centers on the notion that, in a document, the arrangement and organization of words bear less significance compared to the frequency of individual words. BoW's methodology revolves around the development of a lexicon containing all distinct terms present in a specific collection of documents. It proceeds to express each document as a numeric vector, where this representation tallies the instances of each term within the document, shaping a vector with many zero entries, with each dimension corresponding to one of the unique words in the lexicon. The CountVectorizer method is used for the BoW model, where a vector of word occurrence frequencies is considered for each document, and the vocabulary size is  $N$ . An  $N$ -dimensional vector is transformed from each document, where each entry corresponds to the count of a specific word in the document. The frequency representation is:

$$V_d = (f_1, f_2, \dots, f_N)$$

where,

- The feature vector for document  $d$  is  $V_d$ .
- The frequency (raw count) of the  $i$ -th in the document  $d$  is  $f_i$ .
- The total number of unique words (vocabulary size) across all documents is  $N$ .

## 3.3. Classifier model

Different machine-learning algorithms have been used for evaluation as we have created the dataset. We chose the RF algorithm, DT, SVM, and SGD among the options because of their higher accuracy compared to others. This means, we applied various traditional ML algorithms and found that these 4 performed or showed higher accuracy than the rest.

### 3.3.1. Decision tree

DT is a renowned supervised ML algorithm. DT constructs a tree-like structure by iteratively splitting the dataset based on features. This yields a sequence of decision rules that govern the classification process. The procedure builds a DT model based on feature values to simulate human decision-making. Nodes in the model represent features, branches represent decision rules, and leaves represent results or class labels. The best feature selection process for splitting this algorithm is Entropy or Gini impurity \Gini Index.

- Entropy: entropy represents the amount of information required to characterize a given dataset precisely. When a dataset is completely homogeneous, meaning all elements are identical, the entropy value is 0. Conversely, if the dataset is evenly split among different classes, entropy reaches its maximum value of 1. The entropy mathematically is written as:

$$Entropy = - \sum_{i=1}^n P_i * \log(P_i)$$

- Gini impurity \Gini index: the Gini index quantifies the degree of inequality within a dataset, ranging from 0 to 1. A Gini Index of 0 indicates complete homogeneity, meaning all elements belong to the same class, while a value of 1 represents maximum diversity, signifying that the elements are entirely different. It is calculated as the sum of the squared probabilities of each class in the dataset. It is illustrated as:

$$GiniIndex = 1 - \sum_{i=1}^n P_i^2$$

After selecting the best feature, the dataset is partitioned into subsets based on the selected feature, and then the best feature selection and data splitting are repeated until the stop condition is met. Stop conditions include all samples belonging to the same class, reaching the maximum depth for the generated tree, or each node having a minimum number of samples.

### 3.3.2. SGD

SGD was also employed. It is an iterative optimization algorithm primarily used for training ML models, particularly in scenarios involving large-scale datasets. It is an expansion of the gradient descent approach, which uses a randomly chosen subset (or a single instance) of the training data to iteratively update the model weights. The update rule is:

$$\theta_{t+1} = \theta_t - \eta \cdot L_i(\theta)$$

where,

- The learning rate is  $\eta$ .
- From a single training sample  $i$  the computed gradient is  $L_i(\theta)$ .

However, at the time, SGD's parameters upgradation process uses a single sample, which helps to avoid local minima and adopts better generalization due to its randomness in updation. For big data problems, it is computationally efficient, and compared to batch gradient descent, it also converges more quickly, especially in online learning scenarios.

### 3.3.3. Support vector machine

Incorporating SVM, a potent classification algorithm, further enriched our evaluation. SVM aims to determine a hyperplane that maximally separates classes within the feature space, all the while preserving a margin between these classes. The support vectors, or data points closest to the margin, are pivotal in this procedure. The decision function is:

$$f(x) = w^T x + b$$

where,

- The weight vector is  $w$  to the hyperplane.
- The input is  $x$ .
- The offset or bias term is  $b$ .

The margin refers to the distance between the hyperplane and the closest support vectors. In the case of a hard margin, the goal is to find the optimal hyperplane that maximizes this distance while ensuring a strict separation between the classes. However, when the data is not perfectly separable, a soft margin is used, which introduces slack variables. This approach allows some misclassifications while maintaining a balance between maximizing the margin and minimizing classification errors. For real-world datasets, we know that perfect separation is not possible always and here we also used a soft margin in our research where the regularization parameter  $C$  value is 1.0.

### 3.3.4. Random forest

Lastly, we harnessed the power of RF, an ensemble learning technique. During training, RF constructs multiple DT and amalgamates their predictions, culminating in heightened accuracy, improved generalization, and resilience against over-fitting. It is a preferred option for many real-world applications in classification and regression problems due to its robustness, accuracy, and versatility in handling different types of data. For classification, the most frequent class among all trees is the final prediction. For classification, the final prediction is:

$$\hat{y} = \text{mode}(T_1(x), T_2(x), \dots, T_m(x))$$



where,

- The mode function identifies the class that appears the most frequently among those predictions.
- The final prediction  $\hat{y}$  is the majority vote from all the individual tree predictions.
- From each of the  $m$  decision trees for a given input  $x$ ,  $T_1(x), T_2(x), \dots, T_m(x)$  represent the predictions.

The RF classifier performs well even in cases where the connection between features is complex or non-linear, and it has the advantage of being resistant to overfitting, particularly for big datasets with numerous features. When a dataset has missing values and noisy data, RF can handle these issues and automatically measure feature importance. It is the best option because RF uses class weights and bootstrap sampling to handle imbalanced data. On datasets with imbalances, it usually works well and requires little fine-tuning.

#### 4. PERFORMANCE EVALUATION

This section represents the environment of implementation, detailing the performance evaluation metrics that we have chosen to evaluate our model. We have also presented the performance analysis of various ML algorithms after applying various feature extraction techniques, discussing their impacts and comparing their results in this section. Additionally, the performance comparison of existing work with ours is also represented here.

##### 4.1. Environmental setup

- Operating system: Windows 10
- Processor: Intel(R) Core(TM) i5-4300M CPU @ 2.60 GHz
- RAM: 8 GB
- IDE: Google Colab
- Programming language: Python

##### 4.2. Train and test data split

The resulting dataset has been divided into the “training dataset” and “testing dataset” using the Python module from “scikit-learn” (sklearn). A ratio of 80:20 was maintained when dividing the dataset, with 20% of the data used for model testing, and the remaining 80% used for training the model.

##### 4.3. Performance evaluation metrics

Metrics for performance evaluation are quantitative measurements that are used to evaluate ML models’ accuracy and efficacy. These metrics offer information about how well a model predicts results or categorizes data. A comparative study of several ML algorithms, including RF, DT, SVM, and SGD, has been carried out in this part. The details and corresponding subsequent points outline the specifics of the performance evaluation metrics we have employed:

- Only a few of the performance metrics used to evaluate these algorithms were accuracy, precision, recall, and F1-score. To calculate these evaluation metrics, we employed the concepts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

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

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - score = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

- As part of our analysis, to ensure a robust assessment of the algorithms' performance, we applied k-fold cross-validation techniques to partition the training dataset into k smaller subsets. This approach allows us to comprehensively evaluate the performance of the ML algorithms across multiple iterations. K-fold cross-validation involves dividing the training dataset into k equally sized or nearly equal subsets. For our experiments, we used a k-value of 10. During k-fold cross-validation, the techniques involve training the algorithms k times, with each iteration using k-1 of the subsets for training and the remaining subset for validation. Lastly, the average and standard deviation of the performance metrics across all k iterations are calculated.

#### 4.4. Result analysis and discussion

As previously stated, the proposed model employed an 80% training and 20% testing data split. To account for the dataset's imbalanced distribution, the evaluation utilized a weighted F1-score metric. Table 5 and Table 6 presents the evaluation results for all models applied to the four regional Bengali texts for both feature extraction. We employed the P, R, and F to represent precision, recall, and F1-scores, respectively.

Table 5. Evaluation results for various ML techniques using TF-IDF + CountVectorizer feature extraction on the test set

Model	TF-IDF + CountVectorizer											
	Chittagong			Barishal			Rangpur			Noakhali		
	P	R	F	P	R	F	P	R	F	P	R	F
DT	0.75	0.84	0.80	0.70	0.61	0.65	0.84	0.91	0.87	0.69	0.59	0.64
SVM	0.73	0.92	0.81	0.81	0.45	0.58	0.86	0.75	0.80	0.77	0.64	0.70
SGD	0.78	0.89	0.83	0.69	0.62	0.66	0.88	0.89	0.89	0.78	0.64	0.70
RF	0.77	0.91	0.83	0.78	0.61	0.68	0.89	0.98	0.93	0.78	0.61	0.68

Table 6. Evaluation results for various ML models using BoW feature extraction on the test set

Model	Bag of Words (BoW)											
	Chittagong			Barishal			Rangpur			Noakhali		
	P	R	F	P	R	F	P	R	F	P	R	F
DT	0.72	0.79	0.75	0.56	0.61	0.58	0.70	0.91	0.79	0.67	0.47	0.55
SVM	0.72	0.92	0.81	0.55	0.44	0.49	0.90	0.81	0.85	0.75	0.52	0.61
SGD	0.77	0.88	0.82	0.66	0.61	0.63	0.76	0.89	0.82	0.77	0.58	0.66
RF	0.75	0.89	0.81	0.66	0.64	0.65	0.83	0.93	0.88	0.79	0.53	0.64

The overall effectiveness of the aforementioned machine-learning algorithms is shown in Table 7. The test data's projected labels were assessed and contrasted with the actual labels. The results are displayed in this table along with the accuracy, and weighted average of precision, recall, and F1-score. This comparison enabled us to assess the performance of each model comprehensively.

Table 7. Overall performance analysis table for ML models concerning weighted average score

Model	TF-IDF + CountVectorizer				Bag of Words (BoW)			
	Accuracy (%)	Precision (wt.Avg)	Recall (wt.Avg)	F1-score (wt.Avg)	Accuracy (%)	Precision (wt.Avg)	Recall (wt.Avg)	F1-score (wt.Avg)
DT	74.35	0.74	0.74	0.74	68.40	0.68	0.68	0.68
SVM	75.73	0.77	0.76	0.75	72.58	0.73	0.73	0.71
SGD	78.10	0.78	0.78	0.78	75.73	0.76	0.76	0.75
RF	79.15	0.79	0.79	0.78	75.92	0.76	0.76	0.75

Even though the generated dataset is relatively small due to resource availability, the results provide a significant foundation for advancing Bangla regional text classification. This development ushers in a new era for researchers in this field, offering full semantic information and showcasing the linguistic diversity inherent in complete text formation.

**4.4.1. Impacts of using feature extraction TF-IDF and CountVectorizer**

Both the DT algorithm and SGD achieve an accuracy of 74.35 % and 78.10% accordingly, yet the SGD exhibits higher values for precision and recall compared to the DT. And SVM achieves 75.73% which is better than DT. Moreover, RF demonstrates superior performance in comparison to the other algorithms, achieving an accuracy of 79.15%.

Now, Figure 2 is the confusion matrix for all of the models used with TF-IDF and CountVectorizer. The confusion matrix for a particular model is represented by each Figures 2(a)-(d), which shows the proportion of properly and wrongly identified cases in the test set using feature extraction TF-IDF and CountVectorizer. In this Figure 2(a) we can see that Rangpur is handled more accurately by DT, but it has trouble distinguishing between closely related dialects, especially Chittagong and Noakhali. Compared to the other models, SVM accurately identifies Chittagong (215) and Noakhali (96) more often, offering a balanced classification across regions. Although SGD frequently produces high-level metrics (precision, recall), it has a bias in favor of some classes, most notably Chittagong, which causes misclassifications in other areas. But overall separation is improved by RF Figure 2(d), particularly for Rangpur (56 correct out of 57) and Chittagong (212 correct). With the fewest misclassifications in every region, RF has the highest accuracy out of the four models.

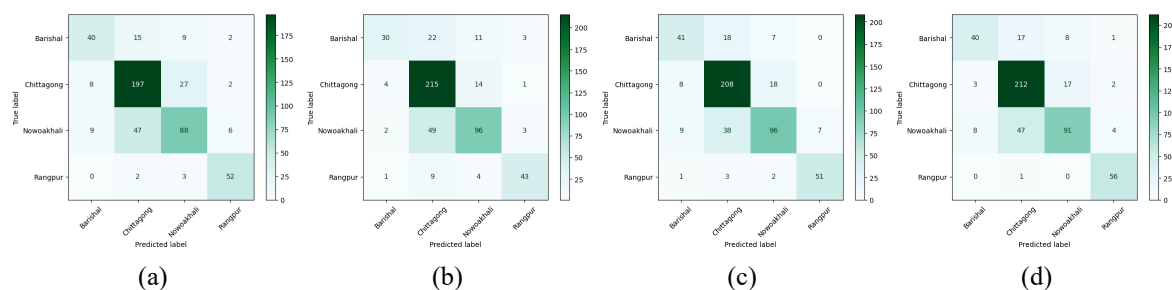


Figure 2. Confusion matrix for; (a) DT, (b) SVM, (c) SGD, and (d) RF models used with TF-IDF and CountVectorizer

**4.4.2. Impacts of using feature extraction BoW**

In this study, the DT algorithm and SGD exhibit an accuracy of 68.40% and 75.73% sequentially when applying the BoW technique. Interestingly, the SGD outperforms the DT in precision, recall, and F1-score, demonstrating its effectiveness. Additionally, the RF algorithm surpasses all other methods, achieving an impressive accuracy of 75.92% with BoW.

The results are presented in Figure 3 confusion matrix for all of the models used with BoW. After analyzing all Figures 3(a)-(d) it is stated that Chittagong and Rangpur are handled more effectively by DT with BoW than by the other classes, while Noakhali and, to a lesser degree, Barishal are difficult to distinguish from one another. But DT with BoW performs similarly for the Rangpur dialect but performs poorly for the Chittagong dialect compared to TF-IDF. With a small amount of misunderstanding between Barishal (4) and Noakhali (15), the SVM model successfully detects 215 Chittagong instances when we look at Figure 3(b). SVM is generally very good at differentiating between Chittagong and Rangpur, but it still has trouble differentiating between Barishal and Noakhali.

Then, as we can see in Figure 3(c), the SGD model correctly labeled Noakhali dialects 87 times, although 43 were misclassified as Chittagong and 11 as Barishal. This shows that these dialects overlap, with Rangpur being largely recognized correctly (51), and a few misclassifications as Chittagong (4) and Barishal (2). All things considered, SGD does a great job at identifying Chittagong and Rangpur, but it still has trouble telling Barishal and Noakhali apart from Chittagong. Among DT, SVM, and SGD, the RF confusion matrix produces the best accuracy. In this case, it is very accurate (53 correct), with only 4 cases incorrectly categorized as Chittagong. With the ability to correctly discriminate the majority of samples and improve upon the misclassifications shown in DT, SVM, and SGD, RF exhibits the most balanced performance overall.

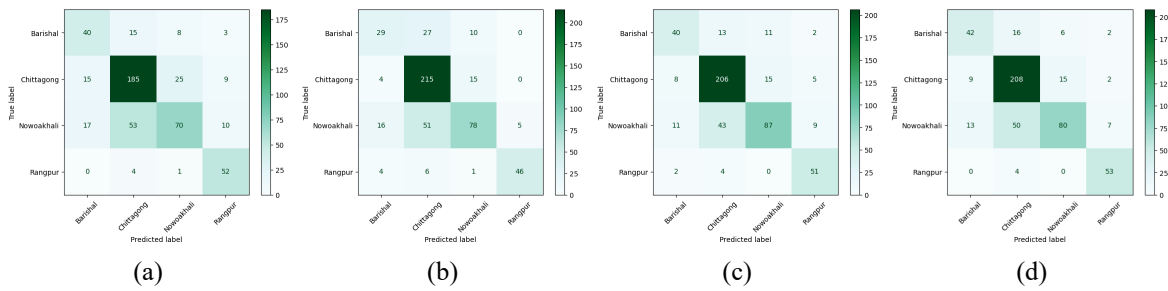


Figure 3. Confusion matrix for (a) DT, (b) SVM, (c) SGD, and (d) RF models used with BoW

#### 4.4.3. Impacts of K-fold cross validation

Using the TF-IDF and BoW feature extractor, we used the K-fold cross-validation technique with  $K = 10$  to validate the performance of the aforementioned machine-learning algorithms. The mean accuracy for each fold was computed and presented in Table 8. Notably, the RF algorithm demonstrated a higher mean accuracy with 79.47%. This higher mean accuracy underscores the model's proficiency in accurately performing regional text classification. The sample text from the Chittagong region appears to be comparatively higher than the other three regional Bangla texts Barishal, Rangpur, and Noakhali. Notably, the RF algorithm excels in handling imbalanced datasets, which might explain its better performance compared to other algorithms.

Table 8. 10 fold cross-validation on ML models with different feature extraction

Model	TF-IDF + CountVectorizer	Bag of Words (BoW)
	Mean accuracy (%)	Mean accuracy (%)
DT	75.72	73.75
SVM	76.91	76.88
SGD	78.41	77.05
RF	79.47	77.26

#### 4.4.4. Performance comparison with others relevant Bangla regional text dataset

We have already mentioned the details of the available dataset we are working with for the Bangla regional text. As we consider Vashantor, it is a benchmark dataset for the Bangla region and also a large dataset, but it primarily consists of translated actual text to the selected region. To evaluate this dataset, they used Bangla-BERT-base and mBERT and found that mBERT achieved an accuracy of 84.36% and Bangla-BERT-base achieved an accuracy of 85.86%, respectively. After that, the second dataset is Bhashamul, which is the result of a competition aimed at the exact IPA transcription of different dialects of Bengali. Their dataset is also considered large, including 30,311 training samples and 8,941 test samples, with a minimum instance length of 1 as well. They used transformer-based models like ByT5, mT5, BanglaT5, and mt5 to evaluate and used word error rates as evaluation metrics. They found a public score WER of 1.995% and a private score WER of 2.072% as results. The third dataset is from our research, which is comparatively low in size but considered real-world usage text with full semantic meaning. For the imbalanced dataset, the RF model performed well with an accuracy of 79.15% and a mean accuracy of 79.47%, creating a new era and adding research guidance for the Bangla regional text-domain. Table 9 represents a critical analysis of the performance of those existing works in comparison to our research. If efficiency is taken into account the conventional ML models employed in this study are more computationally efficient than transformer-based models like BERT, which demand substantial computer resources for training and fine-tuning. Because of this, they are a good option for environments with limited resources, where quicker training times and fewer resources are essential. While transformer-based models like BERT provide state-of-the-art performance, traditional ML models such as SVM and RF offer a simpler and more accessible approach to text classification. These models are easier to implement and require fewer hyperparameter adjustments, which can be beneficial when time and expertise are limited. Finally, we can state that the models employed in this study SGD, SVM, RF, and DT offer a more interpretable and computationally efficient method while maintaining competitive accuracy on comparatively smaller datasets. This study shows that conventional ML techniques are still a formidable competitor in the sector, providing a workable and expandable solution for challenges involving the classification of regional texts.

Table 9. Performance comparison analysis with related works used Bangla regional text datasets

Paper	Dataset	Feature engineering approaches	Classifier	Evaluation metrics	Result
Paper [21]	Vashantor	BanglaT5 and mT5 transformer models	Bangla-bert-base, mBERT	Character error rate (CER), word error rate (WER), BLEU score, METEOR score, and accuracy	mBERT accuracy = 84.36%, Bangla-bert-base accuracy = 85.86%
Paper [22]	Bhashamul	District guided tokens (DGTs), tokenizer adjustments, byte-based tokenization, and multi-head self-attention	Transformer-based models ByT5, umt5, BanglaT5, mt5	Word error rates	Public score WER = 1.995%, Private score WER = 2.072 %
This work	BdRegional Text	TF-IDF, CountVectorizer, BOW	RF, DT, SVM, and SGD	Accuracy, precision, recall, F1-score, mean accuracy	Accuracy = 79.15%, mean accuracy = 79.47%

### 5. CONCLUSION AND FUTURE DIRECTION

In this paper, we’ve addressed the significance of using the Bengali regional language for speaking or writing and also we’ve highlighted the inclination towards researching this Bengali regional language and also explored the role of ML algorithms in this domain. By observing the well-being of online social networks, we can comprehend the growing trend of texting and communicating in the regional Bangla language. This realization aids us in forming our dataset for analysis. The results of this study could have a big impact on real-world applications including sentiment analysis, chatbots that use regional languages, and better language tools for local dialect communication. Although we have done preliminary work in this field, later on, it is possible to do better work by increasing the number of data sizes. We didn’t employ any techniques to balance the dataset since it’s not highly imbalanced. Our future objective is to expand our work to include more regional dialects and develop a model to balance the dataset. Due to resource constraints, we focused on four regions: Chittagong, Noakhali, Rangpur, and Barishal. This research paper will serve as a foundation for future researchers interested in Bengali regional text classification. It aims to spark interest in this area and offer innovative ideas for further enhancements. Future studies may produce more precise and broadly applicable models by combining a wider range of regional dialects and bigger datasets.

### FUNDING INFORMATION

This section should describe sources of funding agency that have supported the work. Authors should state how the research described in their article was funded, including grant numbers if applicable. Include the following (or similar) statement if there is no funding involved: Authors state no funding involved.

### AUTHOR CONTRIBUTIONS STATEMENT

This work was supported in part by the Center for Research, Innovation, and Transformation of Green University of Bangladesh.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Babe Sultana	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				✓
S. M. Mirajul Hoque	✓	✓		✓	✓	✓		✓	✓	✓	✓			
Md Gulzar Hussain				✓	✓	✓			✓		✓	✓	✓	
Mohammad Nurul Huda				✓	✓	✓				✓		✓	✓	

- C : Conceptualization      I : Investigation      Vi : Visualization
- M : Methodology          R : Resources          Su : Supervision
- So : Software              D : Data Curation      P : Project administration
- Va : Validation            O : Writing - Original Draft      Fu : Funding acquisition
- Fo : Formal analysis      E : Writing - Review & Editing

## CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## DATA AVAILABILITY

The supporting data of this study are openly available in “BdRegionText: Bangladeshi Regional Dataset” at <http://doi.org/10.17632/g6zgd66cg5.1>, reference number [33].




## REFERENCES

- [1] F. Alam, S. M. M. Habib, D. A. Sultana, and M. Khan, “Development of annotated Bangla speech corpora,” *2nd Workshop on Spoken Language Technologies for Under-Resourced Languages, SLTU 2010*, pp. 35–41, 2010.
- [2] P. S. Hossain, A. Chakrabarty, K. Kim, and M. J. Piran, “Multi-label extreme learning machine (MLELMs) for Bangla regional speech recognition,” *Applied Sciences*, vol. 12, no. 11, p. 5463, May 2022, doi: 10.3390/app12115463.
- [3] A. K. Das, A. Al Asif, A. Paul, and M. N. Hossain, “Bangla hate speech detection on social media using attention-based recurrent neural network,” *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 578–591, Apr. 2021, doi: 10.1515/jisys-2020-0060.
- [4] S. Dixon, “Leading countries based on Facebook audience size as of January 2023,” *Statista*, 2023.
- [5] M. G. Hussain, B. Sultana, M. Rahman, and M. R. Hasan, “Comparison analysis of Bangla news articles classification using support vector machine and logistic regression,” *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 21, no. 3, p. 584, Jun. 2023, doi: 10.12928/telkomnika.v21i3.23416.
- [6] S. M. S. I. Badhon, H. Rahaman, F. R. Rupon, and S. Abujar, “Bengali accent classification from speech using different machine learning and deep learning techniques,” *Advances in Intelligent Systems and Computing*, vol. 1248, pp. 503–513, 2021, doi: 10.1007/978-981-15-7394-1\_46.
- [7] I. Jamaledyn, R. El ayachi, and M. Biniz, “An improved approach to Arabic news classification based on hyperparameter tuning of machine learning algorithms,” *Journal of Engineering Research*, vol. 11, no. 2, p. 100061, Jun. 2023, doi: 10.1016/j.jer.2023.100061.
- [8] A. Barua, O. Sharif, and M. M. Hoque, “Multi-class sports news categorization using machine learning techniques: resource creation and evaluation,” *Procedia Computer Science*, vol. 193, pp. 112–121, 2021, doi: 10.1016/j.procs.2021.11.002.
- [9] R. Haque, N. Islam, M. Tasneem, and A. K. Das, “Multi-class sentiment classification on Bengali social media comments using machine learning,” *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 21–35, Jun. 2023, doi: 10.1016/j.ijcce.2023.01.001.
- [10] A. S. Neogi, K. A. Garg, R. K. Mishra, and Y. K. Dwivedi, “Sentiment analysis and classification of Indian farmers’ protest using twitter data,” *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100019, Nov. 2021, doi: 10.1016/j.ijime.2021.100019.
- [11] A. Alshamsi, R. Bayari, and S. Salloum, “Sentiment analysis in english texts,” *Advances in Science, Technology and Engineering Systems*, vol. 5, no. 6, pp. 1638–1689, Dec. 2020, doi: 10.25046/AJ0506200.
- [12] T. Parvin and M. M. Hoque, “An ensemble technique to classify multi-class textual emotion,” *Procedia Computer Science*, vol. 193, pp. 72–81, 2021, doi: 10.1016/j.procs.2021.10.008.
- [13] M. M. Uddin, M. Yasmin, M. S. H. Khan, M. I. Rahman, and T. Islam, “Detecting Bengali spam SMS using recurrent neural network,” *Journal of Communications*, vol. 15, pp. 325–331, 2020, doi: 10.12720/jcm.15.4.325-331.
- [14] S. M. Abdulhamid *et al.*, “A review on mobile SMS spam filtering techniques,” *IEEE Access*, vol. 5, pp. 15650–15666, 2017, doi: 10.1109/ACCESS.2017.2666785.
- [15] R. Ghosh, S. Nowal, and G. Manju, “Social media cyberbullying detection using machine learning in bengali language,” *International Journal of Engineering Research & Technology (IJERT)*, vol. 10, no. 5, 2021, doi: <https://doi.org/10.17577/IJERTV10I0500083>.
- [16] S. K. Banshal, S. Das, S. A. Shammi, and N. R. Chakraborty, “MONOVAB: an annotated corpus for Bangla multi-label emotion detection,” *ArXiv*, 2023, [Online]. Available: <https://arxiv.org/abs/2309.15670v1>.
- [17] P. Bhattacharyya, J. Mondal, S. Maji, and A. Bhattacharya, “VACASPATI: a diverse corpus of Bangla literature,” in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 1118–1130, doi: 10.18653/v1/2023.ijcnlp-main.72.
- [18] M. Banik, M. J. R. Rifat, J. Nahar, N. Hasan, and F. Rahman, “Okkhor: a synthetic corpus of Bangla printed characters,” *Advances in Intelligent Systems and Computing*, vol. 1288, pp. 693–711, 2021, doi: 10.1007/978-3-030-63128-4\_53.
- [19] H. Ali, M. F. Hossain, S. B. Shuvo, and A. Al Marouf, “BanglaSenti: a dataset of Bangla words for sentiment analysis,” in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Jul. 2020, pp. 1–4, doi: 10.1109/ICCCNT49239.2020.9225565.
- [20] M. Kowsher, M. J. Uddin, A. Tahabilder, M. R. Amin, M. F. Shahriar, and M. S. I. Sobuj, “BanglaLM: data mining based Bangla corpus for language model research,” *Proceedings of the 3rd International Conference on Inventive Research in Computing Applications, ICIRCA 2021*, pp. 1435–1438, 2021, doi: 10.1109/ICIRCA51532.2021.9544818.
- [21] F. T. J. Faria, M. Bin Moin, A. Al Wase, M. Ahmmed, M. R. Sani, and T. Muhammad, “Vashantor: a large-scale multilingual benchmark dataset for automated translation of Bangla regional dialects to Bangla language,” *ArXiv*, 2023.
- [22] S. Islam, S. Ahmmed, and S. H. Mustakim, “Transcribing bengali text with regional dialects to ipa using district guided tokens,” *ArXiv*, 2024, doi: <https://doi.org/10.48550/arXiv.2403.17407>.
- [23] K. Rakshitha, R. H M, M. Pavithra, A. H D, and M. Hegde, “Sentimental analysis of Indian regional languages on social media,” *Global Transitions Proceedings*, vol. 2, no. 2, pp. 414–420, Nov. 2021, doi: 10.1016/j.gltp.2021.08.039.
- [24] P. Bolaj and S. Govilkar, “Text classification for marathi documents using supervised learning methods,” *International Journal of Computer Applications*, vol. 155, no. 8, pp. 6–10, 2016, doi: 10.5120/ijca2016912374.
- [25] S. S. Shiravale, S. S. Sannakki, and R. Jayadevan, “Text region identification in Indian street scene images using stroke width transform and support vector machine,” *SN Computer Science*, vol. 2, no. 5, p. 357, Sep. 2021, doi: 10.1007/s42979-021-00745-y.




- [26] N. A. P. Rostam and N. H. A. H. Malim, "Text categorisation in Quran and Hadith: overcoming the interrelation challenges using machine learning and term weighting," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 6, pp. 658–667, Jul. 2021, doi: 10.1016/j.jksuci.2019.03.007.
- [27] Y. Zhang, "Research on text classification method based on LSTM neural network model," in *2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, Apr. 2021, pp. 1019–1022, doi: 10.1109/IPEC51340.2021.9421225.
- [28] A. Y. Muaad et al., "An effective approach for Arabic document classification using machine learning," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 267–271, Jun. 2022, doi: 10.1016/j.glt.2022.03.003.
- [29] Q. Wang, J. Zhu, H. Shu, K. O. Asamoah, J. Shi, and C. Zhou, "GUDN: a novel guide network with label reinforcement strategy for extreme multi-label text classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 4, pp. 161–171, 2023, doi: 10.1016/j.jksuci.2023.03.009.
- [30] A. S. More, D. P. Rana, I. Agarwal, and S. Vallabhbai, "Random forest classifier approach for imbalanced big data classification for smart city application domains," *International Journal of Computational Intelligence & IoT*, vol. 1, no. 2, 2018.
- [31] C. Padurariu and M. E. Breaban, "Dealing with data imbalance in text classification," *Procedia Computer Science*, vol. 159, pp. 736–745, 2019, doi: 10.1016/j.procs.2019.09.229.
- [32] N. Jalal, A. Mehmood, G. S. Choi, and I. Ashraf, "A novel improved random forest for text classification using feature ranking and optimal number of trees," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 2733–2742, Jun. 2022, doi: 10.1016/j.jksuci.2022.03.012.
- [33] B. Sultana, S. M. M. Hoque, M. G. Hussain, and M. N. Huda, "Bdregiontext: Bangladeshi regional dataset," 2024, [Online]. Available: <https://data.mendeley.com/datasets/g6zgd66cg5/1>.

## BIOGRAPHIES OF AUTHORS






**Babe Sultana**    was born in, Cox's Bazar, Bangladesh. She received her B.Sc. degree in Computer Science and Engineering from Green University of Bangladesh (GUB) in 2018. She is pursuing her MSc in CSE from United International University and working as a Lecturer, at Dept. of CSE, Green University of Bangladesh. Also, her publication was "Multi-mode Project Scheduling with Limited Resource and Budget Constraints" published at the International Conference on Innovation in Engineering and Technology (ICIET) on 27-28 December 2018 and she got best paper award and IEEE best paper award at this conference. Her research interests include theory of optimization, NLP, ML, and renewable and sustainable energy. She can be contacted at email: [babecse@gmail.com](mailto:babecse@gmail.com).






**S. M. Mirajul Hoque**    was born in, Chittagong, Bangladesh. He received his B.Sc. degree in Computer Science and Engineering from Bangladesh University of Business and Technology (BUBT). At present, he has been working as an Instructor (ICT), at Dhaka Commerce College, Dhaka, Bangladesh. His research interest includes NLP, web engineering, and database. He can be contacted at email: [miraz.bubtce@gmail.com](mailto:miraz.bubtce@gmail.com).



**Md Gulzar Hussain**    was born in Nimnagar, Dinajpur City, Bangladesh. He is pursuing his Ph.D. in the School of Software at Nanjing University of Information Science and Technology, Nanjing, China. He received his Master of Engineering degree (2024) in Computer Application Technology from Changzhou University, Changzhou, China, and his Bachelor of Science degree (2018) in Computer Science and Engineering from the Green University of Bangladesh, Dhaka, Bangladesh. He worked as a lecturer in the Department of Computer Science and Engineering, Green University of Bangladesh, Dhaka, Bangladesh, from 2019 to 2022. He is affiliated with IEEE as a student member and IAER as a professional member. His research interests include transfer learning, NLP, text mining, and topic modeling. He can be contacted at email: [gulzar.ace@gmail.com](mailto:gulzar.ace@gmail.com).



**Mohammad Nurul Huda**    received a Ph.D. degree in automatic speech recognition (ASR) from Toyohashi University of Technology, Aichi, Japan, and a degree from the CSE Department, BUET. He is currently a professor and the head of the Department of CSE, at UIU, Bangladesh. He is one of the Senior Director and an AI and NLP expert with eGeneration Ltd., a leading software company in the area of NLP, ML, and AI. He is especially involved in machine learning, reinforcement learning, speech analysis and recognition, sentiment analysis, Bangla spell and grammar checker, international Bangla morphology, Bangla similarity measure, Bangla document classification, artificial intelligence, computational linguistics, pattern classification, and NLP. He has more than 140 international research articles in related fields of which more than 78 are SCOPUS-indexed articles. He is the elected Vice-President (Academic) of the BCS (2021–2023). He can be contacted at email: [mnh@cse.uui.ac.bd](mailto:mnh@cse.uui.ac.bd).