A Comparative Study on different Machine Learning Approaches for Categorizing Bangla Documents

Abu Jafar Md Jakaria
Department of Computer Science and Engineering
Metropolitan University
Sylhet, Bangladesh

Rajarshi Roy Chowdhury
Department of Computer Science
American International University-Bangladesh
Dhaka, Bangladesh

Jaima Jaman Konia
Department of Computer Science and Engineering
Metropolitan University
Sylhet, Bangladesh

Debashish Roy
Faculty of Applied Science and Technology
Humber College
Toronto, Canada

Nishat Tasnim Ahmed Meem
Department of Computer Science and Engineering
Shahjalal University of Science and Technology
Sylhet, Bangladesh

ABSTRACT

Document categorization (DC) is a pivotal technique employed to efficiently ascertain the category of a document within a reasonable timeframe. It is essential for efficient information retrieval, organization, and analysis, which enables quick identification of relevant documents, facilitates effective search functionality, and streamlines decision-making processes. In this paper, a comparative analysis of nine well-known supervised machine learning (ML) approaches, including random forest (RF), k-nearest neighbors (KNN), support vector machine (SVM), decision tree (DT), bernoulli naïve-bayes (BNB), complement naïve-bayes (CNB), multinomial naïve-bayes (MNB), bagging (BC), and logistic regression (LR), is presented, demonstrating how each algorithm performs on various metrics for automatic Bengali document categorization, thereby highlighting significant differences in their classification accuracy and computational efficiency. Feature selection plays a crucial role in enhancing classification performances alongside the choice of classifier. Normalized term frequency-inverse document frequency (TF-IDF) is utilized to systematically evaluate the effectiveness of various classification techniques across eight distinct categories, highlighting the significant impact of the feature optimization approach. Experimental results have shown that SVM, BC, and LR exhibited significantly higher accuracy than the other methods, with SVM achieving 92.76%, BC reaching 92.64%, and LR attaining 92.26%, respectively, when tested on the Bangla newspaper dataset, highlighting their superior performance in automatic document categorization within this context. These findings underscore the effectiveness of SVM, BC, and LR in the context of Bengali document categorization, demonstrating their ability to consistently deliver high accuracy rates.

Keywords

Document Categorization, Machine Learning, Term Frequency-Inverse Document Frequency, Bengali Document, Natural Language Processing.

1. INTRODUCTION

Document categorization (DC) is a fundamental process used to classify documents into predefined categories, whether manually in the field of library science or automatically in computer and information science, and it spans various media formats such as texts, images, and videos [1], [2]. Its significance extends notably into natural language processing (NLP) [3], where DC serves as a foundational component, with major search engines like Google and Yahoo heavily relying on it for efficient information retrieval, indexing, and organization of vast amounts of data across the web. Beyond search engines, DC finds application in diverse fields such as spam filtering [4], online news filtering [5], [6], and social media analytics [7], showcasing its versatility and critical role in enhancing information management, content filtering, and large-scale data analysis across various digital platforms. In the realm of machine learning (ML), three primary approaches: supervised, unsupervised, and semi-supervised methods, are employed for document classification, each offering unique strengths and applications depending on the availability of labeled data and the specific requirements of the task at hand. The multilayered role of DC underscores its importance as a basis technology in information processing, enabling precise classification and retrieval of digital content across a wide range of platforms and applications, from search engines to content recommendation systems.

The increasing reliance on the internet has led to a vast accumulation of data online, intensifying the need for effective methods to classify and categorize text documents based on their content. As a result, extensive research has been conducted in the field of document categorization using various ML techniques, with a significant focus on English text [8], [9], considering its importance in global digital communication. However, comparatively fewer studies have explored DC for Bangla, although it is the 7th most spoken language in the world [10], [11], highlighting a critical gap in addressing the linguistic diversity in text categorization research. ML techniques are applied not only for document categorization [5], [8], [9] but also extend to various domains, such as Internet of Things (IoT) device classification [12], [13], [14], [15], network traffic analysis [16], [17], detection of malicious traffic in network [18], [19], recommendation systems [20], and advancements in medical science [21], [22], where ML approaches play a key role in diagnostics, predictive analytics, and personalized treatment.

In this study, different supervised ML techniques are employed to categorize Bengali documents, utilizing a range of methods to generate a function from labeled training data that maps inputs to outputs based on example input-output pairs, thereby enhancing the accuracy and reliability of document classification. The techniques explored in this study for Bengali document categorization encompass k-nearest neighbors (KNN), support vector machine (SVM), naive bayes: multinomial naive bayes (MNB), bernoulli naive bayes (BNB) and complement naive bayes (CNB), random forest (RF), logistic regression (LR), bagging (BC), and decision tree (DT), all of which are applied under the framework of supervised ML that relies on generating a function from labeled training data to effectively map inputs to outputs. A publicly available dataset, such as the Bangla newspaper dataset [23], is utilized for evaluating the performance of the proposed models, providing a standardized benchmark for assessing their effectiveness in categorizing Bengali text documents. Additionally, normalized term frequency-inverse document frequency (TF-IDF) [20], [24] is employed, as the feature selection method, in the proposed approach to systematically evaluated the efficacy of various classification techniques across eight distinct categories, including sports, bangladesh, international, entertainment, economy, opinion, technology, and life-style, thereby revealing the substantial impact of feature optimization on enhancing overall classification accuracy and improving the effectiveness of the document categorization process. TF-IDF measures a term's importance in a document based on how often it appears in that document and how rare it is across all documents, thereby providing a weighted measure that highlights terms that are significant in specific contexts while diminishing the relevance of more common terms.

The primary objective of this study is to compare various supervised ML approaches to identify the most effective method for categorizing Bangla text documents based on accuracy. This involves using the TF-IDF approach to identify effective feature sets and enhance the evaluation process. By assessing these techniques in the context of Bengali document categorization, the aim is to advance language-specific document classification methodologies and offer insights that could enhance the accuracy and efficiency of text categorization for less commonly studied languages.

The remainder of this paper is organized as follows: **Section 2** reviews related work, **Section 3** covers the feature engineering process, and **Section 4** describes the various ML approaches. **Section 5** details the proposed methodology for Bengali document categorization, including dataset preparation and the proposed model. **Section 6** presents the experimental results and discussion, and **Section 7** concludes with a summary and suggestions for future research.

2. RELATED WORK

ML approaches are broadly utilized across several languages, with a substantial body of literature focusing on text and document categorization, particularly in English [1], [8], [9], [25], while significant research also extends to other languages such as Arabic [26], [27], Chinese [9], Japanese [28], and Hindi [29], as well as major European languages [30], [31] including French, German, and Spanish. Table 1 presents a brief overview of some existing works.

Bijalwan et al. [25] introduced a DC model utilizing numerous ML approaches, such as KNN, naive bayes (NB), and Term-graph, aimed at improving text classification accuracy. Their experimental results demonstrated that the proposed model achieved over 98% accuracy using the Reuters-21578 dataset, which encompasses five distinct categories, including people, places, exchange, organization, and topics. Among the approaches, KNN outperformed both NB and Term-graph in terms of classification performance, underscoring the versatility of supervised learning approaches and their significance in achieving high accuracy across diverse linguistic contexts. In reference [32], the authors compared three major document categorization approaches alongside the highest average similarity over retrieved documents (HASRD) method to evaluate classification effectiveness. Using the Reuters-21578 dataset, KNN (unary) demonstrated superior precision at 88.51%, outperforming the other approaches in terms of classification accuracy.

Table 1. Existing works on document categorization

Source	Problem	Language	No of Categories
[25]	Document categorization Text categorization	English	5
[32]		English	5
[33]		English	4
[34]		English	22
[35]		Bangla	12
[36]		Bangla	9
[10]		Bangla	5
[37]		Bangla	12

Lie et al. [33] performed a comparative analysis of several ML classifiers for text classification tasks, concluding that the SVM classifier demonstrated significant superiority over the other ML models, including KNN and NB. SVM classifier achieved 86.25% of F1-score on the Reuters-21578 dataset, which encompassed four categories. The researchers, in reference [34], aiming to classify cricket sports news, employed different ML classifiers, including SVM, C4.5, and NB. Their approach impressively reached 99% accuracy in the plain case, bypassing the use of feature selection methods, on the SGSC sports news corpus, which consists of 22 categories including cricket and swimming.

Saiful et al. [35] presented a ML model for classifying Bengali text documents, focusing on three classifiers: SVM, NB, and stochastic gradient descent (SGD), to evaluate their effectiveness. The study also incorporated two feature selection methods: chi-square distribution and normalized TF-IDF into the classification process to enhance the performance of the models. Their experimental results revealed that the model using the TF-IDF approach, achieved an impressive F1-score of over 92% on a Bengali document corpus, whilst this dataset compiled from articles published in Bangladeshi newspapers. In reference [36], the authors introduced a ML model for categorization Bangla text documents, integrating the word2vec word embedding technique and the SGD for statistical analysis. The model achieved an accuracy of over 93% when applied to the Bangla document corpus using the SVM classifier, showcasing its effectiveness in text classification. The dataset, which includes nine categories such as accidents, crime, sports, and politics, was generated from a variety of online platforms including websites, blogs, newspapers, and online books.

Mandal and Sen [10] investigated four supervised ML classifiers, including C4.5, KNN, NB, and SVM, for the categorization of Bangla documents. The BD corpus, which was collected from various online platforms like prothom-alo.com, bdnews24.com, and bbc.co.uk/Bengali. Their approach achieved an accuracy of over 89% on the BD corpus, which consists of five categories: business, sports, health, technology, and education. In reference [37], the authors achieved an impressive accuracy of over 93% using the NB classifier on a Bangla dataset that encompasses twelve distinct categories, including accident, opinion, and crime.

A review of key existing works reveals that many researchers have explored various ML classifiers and feature selection techniques for text classification across multiple languages, including English and Bangla. However, despite Bangla being the seven most spoken language in the world, its representation in text classification research remains limited. This limitation highlights the need for more extensive exploration of Bangla language datasets in order to enhance classification performance and contribute to the growing body of knowledge in multilingual text classification.

3. FEATURE ENGINEERING

Feature extraction involves simplifying data to reduce the dimensionality of data to minimize the resources required for describing large datasets [12], [38], which assists addressing the challenges created by the abundance of variables, such as high memory and computational demands, as well as the risk of overfitting classification algorithms. Feature extraction methods aim to mitigate issues related to high-dimensional data by constructing variable combinations that accurately represent the data, and many ML experts consider well-optimized feature extraction essential for effective model construction. In this study, TF-IDF [20], [24] was utilized for feature selection to enhance the representation of textual features. In the realm of data retrieval, TF-IDF functions as a quantitative metric used to assess the importance of a word within a document set or corpus, and is computed by multiplying two key components: term frequency (TF), which measures how often a term appears in a document, and inverse document frequency (IDF), which evaluates how unique or rare the term is across the entire corpus [24].

TF quantifies how often a term appears within a document, and in longer documents, certain terms may occur more frequently than in shorter ones, thus assigning greater significance to frequently occurring terms within the same document. In contrast, IDF evaluates the importance of a term by considering how common (or rare) it is across the entire corpus, addressing the limitation of TF, which treats all terms with equal importance, potentially overemphasizing less significant terms that occur frequently. To address this issue, the following formula is used to reduce the significance of common terms while amplifying that of rare ones:

$$TF(t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d} \qquad (1)$$

$$IDF(t) = log\left(\frac{N}{df+1}\right) \qquad (2)$$

$$TF - IDF(t,d)\ =\ TF(t,d)\ *\ IDF(t) \qquad (3)$$

where *N* denotes the total number of documents within a given corpus, *d* represents a document, and *df* refers to the number of documents containing the term *t*. TF-IDF remains one of the most widely used term-weighting schemes in contemporary applications, with a 2015 survey indicating that 83% of text-based recommender systems in digital libraries depend on TF-IDF for ranking and retrieving relevant content [39].

4. MACHINE LEARNING ALGORITHMS

In this section, a brief overview of different ML classifiers [40], [41] is presented. The selected algorithms are as follows: KNN, naive bayes (multinomial NB, bernoulli NB, and complement NB), SVM, DT, RF, LR, and BC.

4.1 K-Nearest Neighbor

K-nearest neighbor (KNN) algorithm is recognized as one of the simplest lazy ML techniques, with its nonparametric nature meaning it makes no assumptions about the underlying data distribution. In this context, *K* represents the number of nearest neighbors, usually chosen as an odd integer [42]. In reference [43], the researchers explained that when classifying a new document, the system identifies the *K* nearest neighbors from the training documents and uses their categories to weigh potential classifications. A significant drawback of the KNN algorithm is its computational intensity, as it requires comparing a test document with every sample in the training set. Additionally, the algorithm’s performance is highly dependent on two key factors: selecting an appropriate similarity function and determining the optimal value for the parameter *K*.

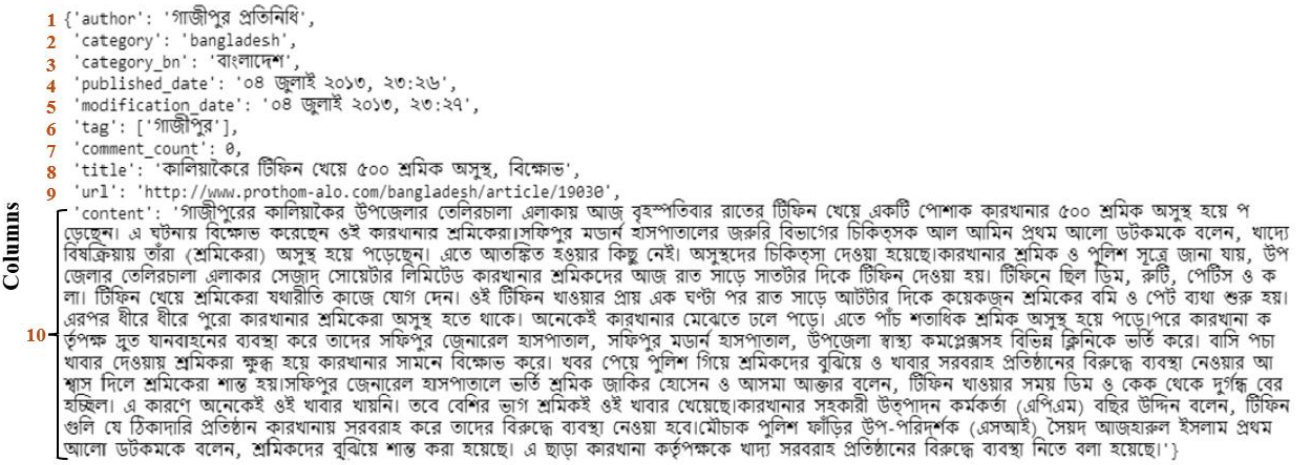
4.2 Naïve Bayes

Naive bayes (NB) represents a straightforward classification method grounded in Bayes’ theorem, serving as a basic probabilistic classifier. In this study, three variations of NB classifiers are employed [44], [45], including MNB, BNB, and CNB.

- Multinomial naive bayes (MNB) [44], [46] is a probabilistic classification algorithm frequently applied in text classification tasks, where word frequency within documents is taken into account, and it operates under the assumption that features are drawn from a multinomial distribution, making it particularly effective with text data represented as word frequency vectors.
- Bernoulli naive bayes (BNB) [44] is a variant closely related to MNB, but it differs in that its predictors are binary variables, representing whether a specific feature, such as a word, is present or absent in a document. Unlike MNB, which accounts for word frequency within a document, BNB focuses exclusively on the presence or absence of each feature, making it well-suited for datasets with binary features or scenarios where feature occurrence is prioritized over frequency.
- Complement naive bayes (CNB) [45] is an adaptation of the traditional MNB algorithm, specifically designed to handle imbalanced datasets by leveraging statistics derived from the complement of each class to calculate the model's weights. By focusing on the complement of class statistics, CNB effectively addresses the challenges posed by imbalanced data distributions, offering a more tailored solution that enhances performance in scenarios where class imbalances are significant.

4.3 Support Vector Machine

Support Vector Machine (SVM) [29] is a supervised ML algorithm capable of handling both classification and regression tasks, although it is predominantly used for classification. The SVM algorithm works by mapping each data point as a coordinate in an *n*-dimensional space, where *n* represents the number of features, and then identifies a hyperplane that optimally separates the different classes for accurate classification. In this study, linear SVM (LSVM) is employed, a feature highlighted by Huang and Kecman [47] which signifies that the creation of an SVM model using LSVM scales linearly with the size of the training dataset, thereby demonstrating efficient utilization of central processing unit (CPU) time.



```
{'author': 'গাজীপুর প্রতিনিধি',
 'category': 'bangladesh',
 'category_bn': 'বাংলাদেশ',
 'published_date': '০৪ জুলাই ২০১৩, ২৩:২৬',
 'modification_date': '০৪ জুলাই ২০১৩, ২৩:২৭',
 'tag': ['গাজীপুর'],
 'comment_count': 0,
 'title': 'কালিয়াকৈরে টিফিন খেয়ে ৫০০ শ্রমিক অসুস্থ, বিক্ষোভ',
 'url': 'http://www.prothom-alo.com/bangladesh/article/19030',
 'content': 'গাজীপুরের কালিয়াকৈর উপজেলার তেলিরচালা এলাকায় আজ বৃহস্পতিবার রাতের টিফিন খেয়ে একটি পোশাক কারখানার ৫০০ শ্রমিক অসুস্থ হয়ে প
ড়েছেন। এ ঘটনায় বিক্ষোভ করেছেন ওই কারখানার শ্রমিকেরা।সফিপুর মডার্ন হাসপাতালের জরুরি বিভাগের চিকিত্সক আল আমিন প্রথম আলো ডটকমকে বলেন, খাদ্যে
বিষক্রিয়ায় তাঁরা (শ্রমিকেরা) অসুস্থ হয়ে পড়েছেন। এতে আতঙ্কিত হওয়ার কিছু নেই। অসুস্থদের চিকিত্সা দেওয়া হয়েছে।কারখানার শ্রমিক ও পুলিশ সূত্রে জানা যায়, উপ
জেলার তেলিরচালা এলাকার সেজাদ সোয়েটার লিমিটেড কারখানার শ্রমিকদের আজ রাত সাড়ে সাতটার দিকে টিফিন দেওয়া হয়। টিফিনে ছিল ডিম, রুটি, পেটিস ও ক
লা। টিফিন খেয়ে শ্রমিকেরা যথারীতি কাজে যোগ দেন। ওই টিফিন খাওয়ার প্রায় এক ঘণ্টা পর রাত সাড়ে আটটার দিকে কয়েকজন শ্রমিকের বমি ও পেট ব্যথা শুরু হয়।
এরপর ধীরে ধীরে পুরো কারখানার শ্রমিকেরা অসুস্থ হতে থাকে। অনেকেই কারখানার মেঝেতে ঢলে পড়ে। এতে পাঁচ শতাধিক শ্রমিক অসুস্থ হয়ে পড়ে।পরে কারখানা ক
র্তৃপক্ষ দ্রুত যানবাহনের ব্যবস্থা করে তাদের সফিপুর জেনারেল হাসপাতাল, সফিপুর মডার্ন হাসপাতাল, উপজেলা স্বাস্থ্য কমপ্লেক্সসহ বিভিন্ন ক্লিনিকে ভর্তি করে। বাসি পচা
খাবার দেওয়ায় শ্রমিকরা ক্ষুব্ধ হয়ে কারখানার সামনে বিক্ষোভ করে। খবর পেয়ে পুলিশ গিয়ে শ্রমিকদের বুঝিয়ে ও খাবার সরবরাহ প্রতিষ্ঠানের বিরুদ্ধে ব্যবস্থা নেওয়ার আ
শ্বাস দিলে শ্রমিকেরা শান্ত হয়।সফিপুর জেনারেল হাসপাতালে ভর্তি শ্রমিক জাকির হোসেন ও আসমা আক্তার বলেন, টিফিন খাওয়ার সময় ডিম ও কেক থেকে দুর্গন্ধ বের
হচ্ছিল। এ কারণে অনেকেই ওই খাবার খায়নি। তবে বেশির ভাগ শ্রমিকই ওই খাবার খেয়েছে।কারখানার সহকারী উত্পাদন কর্মকর্তা (এপিএম) বছির উদ্দিন বলেন, টিফিন
গুলি যে ঠিকাদারি প্রতিষ্ঠান কারখানায় সরবরাহ করে তাদের বিরুদ্ধে ব্যবস্থা নেওয়া হবে।মৌচাক পুলিশ ফাঁড়ির উপ-পরিদর্শক (এসআই) সৈয়দ আজহারুল ইসলাম প্রথম
আলো ডটকমকে বলেন, শ্রমিকদের বুঝিয়ে শান্ত করা হয়েছে। এ ছাড়া কারখানা কর্তৃপক্ষকে খাদ্য সরবরাহ প্রতিষ্ঠানের বিরুদ্ধে ব্যবস্থা নিতে বলা হয়েছে।'}
```

Fig 1: A sample of the Bangla newspaper dataset in JSON format

4.4 Decision Tree

Decision tree (DT) [48] is composed of a series of discrete rules organized in a tree-like structure, resembling a flowchart, where the root node sits at the top, internal nodes represent features or attributes, decision rules are illustrated by branches, and the outcomes are denoted by leaf nodes. The effectiveness of a DT classifier [49] is largely determined by the quality of its construction from the training data, as the process involves beginning at a root node and progressively splitting the dataset into subsets based on feature values, thereby creating sub-trees through iterative divisions until leaf nodes are formed.

4.5 Random Forest

Random Forest (RF) classifier [50], [51] functions as a composite learning method used for both classification and regression tasks, leveraging its strength in managing numerous individual DTs within its framework, where each tree contributes a prediction for a specific class to the collective decision-making process. The final prediction of the RF model is determined by aggregating the votes from each DT within the ensemble, with the class receiving the most votes emerging as the model's overall prediction, thereby benefiting from the combined accuracy of multiple trees. This classifier has been utilized across various domains [15], [52], [53], [54], [55], such as network traffic classification, malicious traffic detection, and document categorization, due to its efficient and accurate classification performance.

4.6 Bagging

Bagging (bootstrap aggregating or BC), a technique developed by Leo Breiman [56], is designed to improve the performance of ML classification algorithms. Acting as a meta-estimator, BC [57] fits base classifiers on various subsets of the original dataset and aggregates their individual predictions, usually through voting or averaging, to produce a final output. This method is frequently used to reduce the variance of a black-box estimator, like a DT, by introducing randomness into the model-building process and forming an ensemble, which helps mitigate overfitting through averaging or voting techniques.

5. METHODOLOGY

5.1 Dataset and Preprocessing

The Bangla newspaper dataset, sourced from Kaggle [23] and containing 437,948 news samples across 32 categories, was initially provided in JavaScript object notation (JSON) format, but was subsequently converted into comma-separated values (CSV) format to facilitate easier processing and analysis. A sample of the dataset in JSON format is depicted in Figure 1.

Initially, the original dataset exhibited a shape of (437948, 10), after conversion to CSV format, its shape transformed to (437948, 3). The initial dataset comprised ten columns namely: ‘author’, ‘category’, ‘category-bn’, ‘published date’, ‘modification date’, ‘tag’, ‘comment count’, ‘title’, ‘url’, and ‘content’. Subsequently, seven columns were removed out of ten columns, resulting in the retention of the ‘category’, ‘title’, and ‘content’ columns. The ‘content’ column exhibited 78 instances of not a number (NaN) value. In this study, these NaN values were removed and altered the dataset’s shape to (437870, 3).

Table 2. A list of 32 categories along with the number of samples from the Bangla newspaper dataset

SN	Category	No of Samples	SN	Category	No of Samples
1.	Bangladesh	232,500	17.	Special-supplement	859
2.	Sports	49,002	18.	Kishoralo	497
3.	International	30,855	19.	Trust	443
4.	Entertainment	30,461	20.	Protichinta	170
5.	Economy	17,245	21.	-1	123
6.	Opinion	15,699	22.	Nagorik-kantho	83
7.	Technology	12,114	23.	Chakri-bakri	75
8.	Life-style	10,831	24.	Tarunno	40
9.	Education	9,721	25.	Mpaward1	17
10.	Durporobash	7,402	26.	22221	11
11.	Northamerica	6,990	27.	Facebook	10
12.	Pachmisheli	3442	28.	Events	2
13.	We-are	2,999	29.	Diverse	2
14.	Onnoalo	2,700	30	Demo-content	2
15.	Roshalo	2,602	31.	Bs-events	1
16.	Bondhushava	971	32.	AskEditor	1

The Bangla newspaper dataset consists of a total of 32 different categories. A list of 32 categories and their corresponding sample counts from the Bangla newspaper dataset are presented in Table 2. It has been observed that a significant portion of the categories (24 out of 32, specifically from serial numbers 9 to 32) contain fewer than 10,000 instances. To avoid the adverse effects of data imbalance on the proposed model's performance, these categories, including education, special-supplement, askeditor, and others, have been excluded from the dataset. Table 3 offers a concise overview of the cleaned dataset used in this experiment, which comprises training and testing sample data across a total of 8 distinct categories.

Table 3. A list of 8 selected categories along with the number of samples from the Bangla newspaper dataset

SN.	Category	No of Samples	No of Training Samples	No of Testing Samples
1.	Bangladesh	232,500	186,140	46,360
2.	Sports	49,002	39,190	9,812
3.	International	30,855	24,705	6,150
4.	Entertainment	30,461	24,233	6,228
5.	Economy	17,245	13,860	3,385
6.	Opinion	15,699	12,625	3,074
7.	Technology	12,114	9,650	2,464
8.	Life-style	10,831	8,562	2,269
	Total	398,707	318,965	79,742

5.2 Proposed Model

In Figure 2, the abstract design of the proposed ML methodologies for categorizing Bangla documents is depicted. The process begins with the acquisition of the Bangla newspaper dataset from a publicly accessible repository such as Kaggle [23]. Subsequently, data preprocessing is carried out, which includes converting the dataset from JSON to CSV format and performing data cleaning to mitigate noisy data and address missing values (NaN), thereby ensuring the availability of clean data for further analysis.

The study employs a feature engineering process using TF-IDF to extract a suitable feature set that has the potential to improve the overall performance of classification and categorization tasks. A total of nine distinct ML models are utilized, with 80% of the dataset designated for training and the remaining 20% for testing, ensuring a balanced evaluation of each model's predictive capabilities. Additionally, five-fold cross-validation is applied to further optimize model performance, ensuring a rigorous assessment of model generalization across multiple data splits. Standard performance metrics and training time are utilized to evaluate the performance of each ML model, providing a comprehensive framework for determining the most effective approach for the categorization of Bangla documents.

5.3 Results and Discussion

The performance of the proposed Bengali document categorization model has been evaluated using Google Collab, a free online service. To improve processing time (reduce training time), a graphics processing unit (GPU), specifically the NVIDIA Tesla K80, was used. All tasks were performed using the Python programming language. The model's performance was evaluated using a Bangla newspaper dataset [23], which was divided into training and testing datasets in 80:20 ratio. Thus, the training dataset contains a total of 318,965 samples, while the testing dataset contains 79,742 samples.

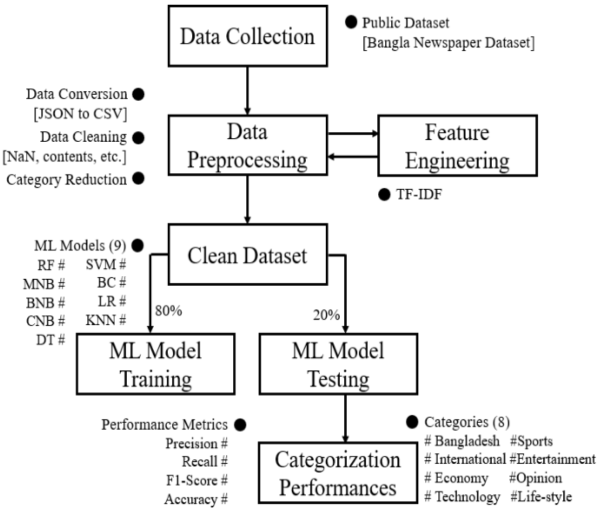


Fig 2: The proposed ML approaches for categorizing Bangla documents

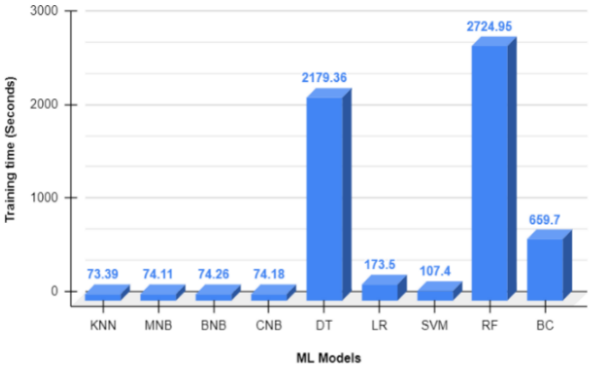


Fig 3: Approximate training time required using different ML models

Figure 3 presents the approximate training times required by various ML classifiers, including KNN, naive bayes (MNB, BNB, and CNB), DT, RF, LR, SVM, and BC, for the categorization of the Bangla newspaper dataset. Among these classifiers, tree-based models such as DT and RF showed the highest training times, taking 2179.36 seconds and 2724.95 seconds, respectively, which were significantly longer than those of the other ML models. In contrast, KNN required the least amount of training time, with a duration of 73.39 seconds, while the three NB classifiers exhibited nearly identical training times, ranging from 74.11 to 74.26 seconds. Additionally, the difference in training time between LR and SVM was minimal, with only a 66.2 second gap, whereas the BC model required a relatively longer time of 659.7 seconds for training on the same dataset.

All the experimental results are illustrated in Figure 4, the performance metrics for each individual ML model, including standard classification measures, are presented for comparison. The highest classification accuracy, 92.76%, is obtained using the SVM classifier, which significantly outperforms all other models, including LR and BC, which achieve close values of 92.26% and 92.64%, respectively. In contrast to the top-performing models, the MNB classifier exhibits the lowest accuracy at 76.62%, whereas tree-based classifiers such as DT and RF demonstrate similar performances. Although KNN marginally outperforms the MNB classifier by just 0.13%. However, overall classification performances are varied significantly among different NB classifiers.

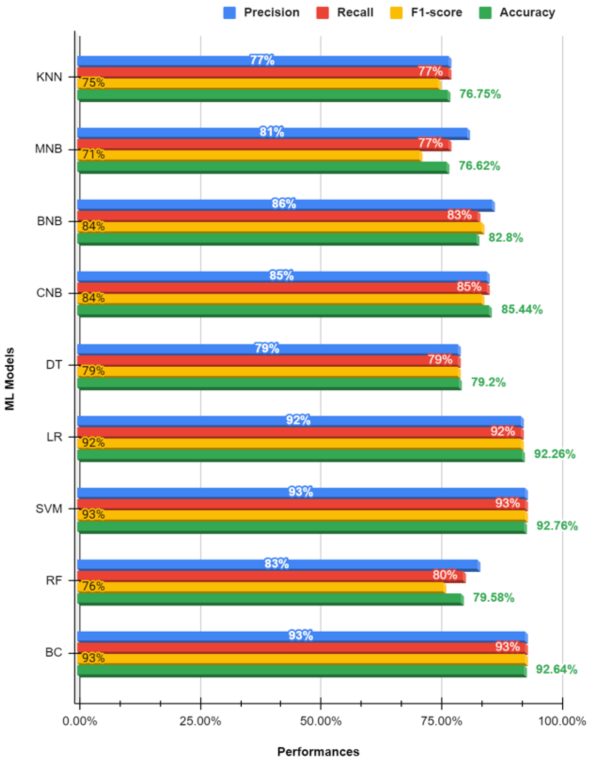


Fig 4: Classification performances using different ML models

These results are significant as it's demonstrated the SVM classifier superior ability to categorize Bangla language text more accurately than other models, indicating its robustness in managing the complexities of the language. Moreover, the variations in accuracy across classifiers highlight the inherent challenges in categorizing Bangla, emphasizing the importance of selecting and optimizing models carefully to enhance classification performance.

6. CONCLUSION AND FUTURE WORK

A detailed comparative analysis of nine supervised ML techniques is explored to distinguish their effectiveness in various aspects of performance and training time. This investigation revealed notable differences among these approaches, whilst some ML models exhibited swift training times, yet their performance did not meet expectations. Conversely, certain models displayed slow training processes and failed to deliver satisfactory performance outcomes. From the experiments three classifiers are identified that exhibited a balance between training times and performance metrics. Consequently, experimental results have shown that SVM, BC, and LR outperforms compared to other ML models such as KNN, RF, MNB, BNB, CNB, and DT. Specifically, SVM, BC, and LR demonstrated superior performance and relatively faster training times compared to their counterparts. Moreover, this analysis revealed that while SVM, LR, and BC exhibited comparable accuracy scores, SVM emerged as the preferred choice over LR and BC. Although their accuracy scores are similar, the subtle differences in performance suggest that SVM is better suited for managing the classification tasks being examined. Overall, the results underscore the significance of considering both training efficiency and performance metrics when selecting appropriate supervised learning techniques. The prominence of SVM, BC, and LR suggests their suitability for tasks requiring robust classification capabilities within a reasonable training timeframe.

For future work, it is required to explore advanced optimizations and hybrid approaches that integrate the strengths of SVM, BC, and LR, with the aim of enhancing classification performance and training efficiency further and expanding the proposed approach to include diverse datasets and novel ML techniques to comprehensively assess their robustness and adaptability across various contexts.

7. ACKNOWLEDGMENTS

The authors are profoundly grateful to the Department of Computer Science and Engineering, Metropolitan University, Sylhet, Bangladesh and the Department of Computer Science, American International University-Bangladesh (AIUB), Dhaka, Bangladesh for supporting this research work.

8. REFERENCES

[1] M. Z. Afzal *et al.*, "Deepdocclassifier: Document classification with deep Convolutional Neural Network," in *13th international conference on document analysis and recognition*, IEEE, 2015, pp. 1111–1115.

[2] H. Borko and M. Bernick, "Automatic Document Classification," *Journal of the ACM*, vol. 10, no. 2, pp. 151–162, 1963.

[3] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimed Tools Appl*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023, doi: 10.1007/s11042-022-13428-4.

[4] Q. Ouyang, J. Tian, and J. Wei, "E-mail Spam Classification using KNN and Naive Bayes," 2023.

[5] R. Evans, D. Jackson, and J. Murphy, "Google News and Machine Gatekeepers: Algorithmic Personalisation and News Diversity in Online News Search," *Digital Journalism*, vol. 11, no. 9, pp. 1682–1700, 2023, doi: 10.1080/21670811.2022.2055596.

[6] C.-H. CHAN Aixin SUN, E. Peng LIM, E. Peng, and C.-H. Chan Aixin Sun Ee-Peng Lim, "Automated online news classification with personalization," 2001. [Online]. Available: https://ink.library.smu.edu.sg/sis_research/913

[7] P. Melville, V. Sindhwani, and R. D. Lawrence, "Social Media Analytics: Channeling the Power of the Blogosphere for Marketing Insight," *Proc. of the WIN*, vol. 1, no. 1, pp. 1–5, 2009, [Online]. Available: http://www.universalmccann.com

[8] H. Abburi, M. Suesserman, N. Pudota, B. Veeramani, E. Bowen, and S. Bhattacharya, "Generative AI Text Classification using Ensemble LLM Approaches," Sep. 2023, [Online]. Available: http://arxiv.org/abs/2309.07755

[9] Y. Wei, "Chinese and English text classification techniques incorporating CHI feature selection for ELT cloud classroom," *Open Computer Science*, vol. 14, no. 1, Jan. 2024, doi: 10.1515/comp-2024-0007.

[10] A. K. Mandal and R. Sen, "Supervised Learning Methods for Bangla Web Document Categorization," *International Journal of Artificial Intelligence & Applications*, vol. 5, no. 5, pp. 93–105, Sep. 2014, doi: 10.5121/ijaia.2014.5508.

[11] M. Habibullah, M. S. Islam, F. T. Jahura, and J. Biswas, "Bangla Document Classification Based on Machine Learning and Explainable NLP," *2023 6th International Conference on Electrical Information and Communication Technology, EICT 2023*, 2023, doi: 10.1109/EICT61409.2023.10427766.

[12] R. R. Chowdhury, A. C. Idris, and P. E. Abas, "Internet of Things Device Classification using Transport and Network Layers Communication Traffic Traces," *International Journal of Computing and Digital Systems*, vol. 12, no. 1, pp. 545–555, 2022, doi: 10.12785/ijcds/120144.

[13] R. R. Chowdhury, A. C. Idris, and P. E. Abas, "Internet of things: Digital footprints carry a device identity," in *AIP Conference Proceedings 2643*, 2023, p. 40003. doi: 10.1063/5.0111335.

[14] R. R. Chowdhury, S. Aneja, N. Aneja, and E. Abas, “Network Traffic Analysis based IoT Device Identification,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Aug. 2020, pp. 79–89. doi: 10.1145/3421537.3421545.

[15] R. R. Chowdhury, A. C. Idris, and P. E. Abas, “Identifying SH-IoT devices from network traffic characteristics using random forest classifier,” *Wireless Networks*, 2023, doi: 10.1007/s11276-023-03478-3.

[16] R. R. Chowdhury and P. E. Abas, “A survey on device fingerprinting approach for resource-constraint IoT devices: Comparative study and research challenges,” Nov. 01, 2022, *Elsevier B.V.* doi: 10.1016/j.iot.2022.100632.

[17] M. Miettinen, S. Marchal, and N. Asokan, “IoT Sentinel: Automated Device-Type Identification for Security Enforcement in IoT,” *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pp. 2177–2184, 2017, doi: 10.1109/ICDCS.2017.284.

[18] M. Hasan, Md. M. Islam, M. I. I. Zarif, and M. M. A. Hashem, “Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches,” *Internet of Things*, vol. 7, p. 100059, Sep. 2019, doi: 10.1016/j.iot.2019.100059.

[19] H. A. Khattak, M. A. Shah, S. Khan, I. Ali, and M. Imran, “Perception layer security in Internet of Things,” *Future Generation Computer Systems*, vol. 100, pp. 144–164, 2019, doi: 10.1016/j.future.2019.04.038.

[20] D. Roy, R. R. Chowdhury, A. Bin Nasser, A. Azmi, and M. Babaeianjelodar, “Item recommendation using user feedback data and item profile,” in *AIP Conference Proceedings 2643*, 2023, p. 40008. doi: 10.1063/5.0111349.

[21] Md. S. Azam, A. Rahman, S. M. H. S. Iqbal, and Md. T. Ahmed, “Prediction of Liver Diseases by Using Few Machine Learning Based Approaches,” *Australian Journal of Engineering and Innovative Technology*, pp. 85–90, Oct. 2020, doi: 10.34104/ajeit.020.085090.

[22] M. Kumar, S. K. Khatri, and M. Mohammadian, “Breast Cancer Classification Approaches - A Comparative Analysis,” *Journal of Information Systems and Telecommunication*, vol. 11, no. 1, pp. 1–11, Apr. 2022.

[23] Z. Al Nazi, “Bangla Newspaper Dataset,” Kaggle. Accessed: Sep. 02, 2024. [Online]. Available: https://www.kaggle.com/datasets/furcifer/bangla-newspaper-dataset

[24] S. Qaiser and R. Ali, “Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents,” *Int J Comput Appl*, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/ijca2018917395.

[25] V. Bijalwan, V. Kumar, P. Kumari, and J. Pascual, “KNN based machine learning approach for text and document mining,” *International Journal of Database Theory and Application*, vol. 7, no. 1, pp. 61–70, Jan. 2014, doi: 10.14257/ijdta.2014.7.1.06.

[26] M. Alhawarat and A. O. Aseeri, “A Superior Arabic Text Categorization Deep Model (SATCDM),” *IEEE Access*, vol. 8, pp. 24653–24661, 2020, doi: 10.1109/ACCESS.2020.2970504.

[27] H. M. Noaman, S. Elmougy, A. Ghoneim, and T. Hamza, “Naive Bayes Classifier based Arabic document categorization,” in *The 7th International Conference on Informatics and Systems (INFOS)*, 2010. Accessed: Sep. 02, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5461819/authors#authors

[28] F. Peng, D. Schuurmans, and S. Wang, “Language and Task Independent Text Categorization with Simple Language Models,” in *Proceedings of HLT-NAACL*, 2003, pp. 110–117.

[29] S. Puri and S. P. Singh, “Hindi Text Document Classification System Using SVM and Fuzzy,” *International Journal of Rough Sets and Data Analysis*, vol. 5, no. 4, pp. 1–31, Sep. 2018, doi: 10.4018/ijrsda.2018100101.

[30] C. H. A. Koster and J. G. Beney, “Phrase-based Document Categorization revisited,” in *PaIR ’09: Proceedings of the 2nd international workshop on Patent information retrieval*, ACM Digital Library, 2009, pp. 49–56.

[31] P. Mcnamee and J. Mayfield, “Character N-Gram Tokenization for European Language Text Retrieval,” 2004.

[32] V. Tam, A. Santoso, and R. Setiono, “A Comparative Study of Centroid-Based, Neighborhood-Based and Statistical Approaches for Effective Document Categorization,” in *International Conference on Pattern Recognition*, IEEE, Aug. 2002.

[33] Z. Liu, X. Lv, K. Liu, and S. Shi, “Study on SVM compared with the other text classification methods,” in *2nd International Workshop on Education Technology and Computer Science, ETCS 2010*, 2010, pp. 219–222. doi: 10.1109/ETCS.2010.248.

[34] T. S. Zakzouk and H. I. Mathkour, “Comparing text classifiers for sports news,” *Procedia Technology*, vol. 1, pp. 474–480, 2012, doi: 10.1016/j.protcy.2012.02.104.

[35] M. S. Islam, F. Elahi, M. Jubayer, and S. I. Ahmed, “A Comparative Study on Different Types of Approaches to Bengali document Categorization,” in *International Conference on Engineering Research, Innovation and Education*, Jan. 2017. [Online]. Available: http://prothom-alo.com,

[36] M. R. Hossain and M. M. Hoque, “Automatic Bengali Document Categorization Based on Word Embedding and Statistical Learning Approaches,” in *International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, IEEE, Feb. 2018.

[37] F. Quadery, A. Al Maruf, and T. Ahmed, “Semi Supervised Keyword Based Bengali Document Categorization,” in *3rd International Conference on Electrical Engineering and Information & Communication Technology*, IEEE, Sep. 2017, p. 139. doi: 10.1109/CEEICT.2016.7873040.

[38] T. Verdonck, B. Baesens, M. Óskarsdóttir, and S. vanden Broucke, “Special issue on feature engineering editorial,” *Mach Learn*, vol. 113, no. 7, pp. 3917–3928, Jul. 2024, doi: 10.1007/s10994-021-06042-2.

[39] J. Beel, B. Gipp, S. Langer, and C. Breitinger, “Research-paper recommender systems: a literature survey,” *International Journal on Digital Libraries*, vol. 17, no. 4, pp. 305–338, Nov. 2016, doi: 10.1007/s00799-015-0156-0.

[40] F. Mozaffari, I. R. Vanani, P. Mahmoudian, and B. Sohrabi, “Application of Machine Learning in the Telecommunications Industry-Partial Churn Prediction by using a Hybrid Feature Selection Approach,” *Journal of Information Systems and Telecommunication*, vol. 11, no. 4, pp. 331–346, Mar. 2023.

[41] K. Jindal and R. Aron, “A Hybrid Machine Learning Approach for Sentiment Analysis of Beauty Products Reviews,” *Journal of Information Systems and Telecommunication*, vol. 10, no. 37, pp. 1–10, Dec. 2022, doi: 10.52547/jist.15586.10.37.1.

[42] B. Trstenjak, S. Mikac, and D. Donko, “KNN with TF-IDF based framework for text categorization,” in *Procedia Engineering*, Elsevier Ltd, 2014, pp. 1356–1364. doi: 10.1016/j.proeng.2014.03.129.

[43] S. Jiang, G. Pang, M. Wu, and L. Kuang, “An improved K-nearest-neighbor algorithm for text categorization,” *Expert Syst Appl*, vol. 39, no. 1, pp. 1503–1509, Jan. 2012, doi: 10.1016/J.ESWA.2011.08.040.

[44] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, “Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification,” in *International Conference on Automation, Computational and Technology Management (ICACTM)*, IEEE, 2019.

[45] B. Seref and E. Bostanci, “Performance comparison of naïve bayes and complement naïve bayes algorithms,” in *Proceedings - 2019 6th International Conference on Electrical and Electronics Engineering, ICEEE 2019*, Institute of Electrical and Electronics Engineers Inc., Apr. 2019, pp. 131–138. doi: 10.1109/ICEEE2019.2019.00033.

[46] M. Abbas, K. Ali, A. Jamali, K. Ali Memon, and A. Aleem Jamali, “Multinomial Naive Bayes Classification Model for Sentiment Analysis,” *IJCSNS International Journal of*

Computer Science and Network Security, vol. 19, no. 3, p. 62, 2019, doi: 10.13140/RG.2.2.30021.40169.

[47] S. Ghosh, A. Dasgupta, and A. Swetapadma, "A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification," in *International Conference on Intelligent Sustainable Systems (ICISS)*, IEEE, 2019.

[48] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[49] S. Suthaharan, *Machine Learning Models and Algorithms for Big Data Classification*, vol. 36. in Integrated Series in Information Systems, vol. 36. Boston, MA: Springer US, 2016. doi: 10.1007/978-1-4899-7641-3.

[50] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[51] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random Forests and Decision Trees," *International Journal of Computer Science Issues*, vol. 9, no. 5, pp. 272–278, 2012.

[52] S. Waskle, L. Parashar, and U. Singh, "Intrusion Detection System Using PCA with Random Forest Approach," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2020, pp. 803–808.

[53] A. K. Mishra and B. K. Ratha, "Study of Random Tree and Random Forest Data Mining Algorithms for Microarray Data Analysis," *International Journal on Advanced Electrical and Computer Engineering (IJAECE)*, vol. 3, no. 4, pp. 5–7, 2016, Accessed: Jun. 18, 2021. [Online]. Available: http://www.irdindia.in/journal_ijaece/pdf/vol3_iss4/2.pdf

[54] R. Jehad and S. A.Yousif, "Fake News Classification Using Random Forest and Decision Tree (J48)," *Al-Nahrain Journal of Science*, vol. 23, no. 4, pp. 49–55, Dec. 2020, doi: 10.22401/ANJS.23.4.09.

[55] R. R. Chowdhury, A. C. Idris, and P. E. Abas, "Device identification using optimized digital footprints," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 1, pp. 232–240, Mar. 2023, doi: 10.11591/ijai.v12.i1.pp232-240.

[56] L. Bbeiman, "Bagging Predictors," 1996.

[57] C. D. Sutton, "Classification and Regression Trees, Bagging, and Boosting," 2005, *Elsevier*. doi: 10.1016/S0169-7161(04)24011-1.