Analysis of Bangla Transformation of Sentences using Machine Learning

Rajesh Kumar Das, Samrina Sarkar Sammi, Khadijatul Kobra, Moshfiqur Rahman Ajmain, Sharun Akter khushbu, Sheak Rashed Haider Noori

Daffodil International University, 1341 Dhaka, Bangladesh
rajesh15-13032, samrina15-12532, khadijaatul15-12319, moshfiqur15-14090, sharun.cse, {@diu.edu.bd}
drnoori@daffodilvarsity.edu.bd

Abstract. In many languages, various language processing tools have been developed. The work of the Bengali NLP is getting richer day by day. Sentence pattern recognition in Bangla is a subject of attention. Additionally, our motivation was to work on implementing this pattern recognition concept into user-friendly applications. So, we generated an approach where a sentence(sorol, jotil and jougik) can be correctly identified. Our model accepts a Bangla sentence as input, determines the sentence construction type, and outputs the sentence type. The most popular and well-known six supervised machine learning algorithms were used to classify three types of sentence formation: Sorol Bakko (simple sentence), Jotil Bakko (complex sentence) and Jougik Bakko(compound sentence). We trained and tested our dataset, which contains 2727 numbers of data from various sources. We analyzed our dataset and got accuracy, precision, recall,f1-score and confusion matrix. We get the highest accuracy with the decision tree classifier, which is 93.72%.

Keywords: Formation of sentence · Bangla NLP · Supervised Machine learning · Decision Tree Classifier · Sorol Bakko · Jotil Bakko · Jougik Bakko

1 Introduction

Bangla is the world's eighth most pronounced language, ranking eighth among all languages. Since Bangla is a language with a very rich morphological structure, it is very difficult to implement every guideline for word construction necessary to create the stemmer that can accurately mark out all the root words in Bangla. It is difficult. Bangladeshi people speak Bangla as their mother tongue and Indian as a second language. As a result, sentence identification in Bangla texts has become a difficult problem in modern times. There are many studies on sentence classification in English, which is very different from Bangla. About 261 million people speak Bengali worldwide. In Bangla, there are three types of sentences based on sentence structure, and those are Sorol Bakko (simple sentence), Jotil Bakko (complex sentence) and Jougik Bakko(compound sentence). Sorol Bakko

contains only one verb and one or more subjects. In Jotil Bakko, two sentence blocks or clauses are joined together to form one sentence. One clause is independent, and that's called the main clause, and the other is a subordinate clause, which is dependent. Jougik Bakko or Compound sentence is a combination of two or more sorol bakko or jotil bakko. In this paper, we will categorize sorol, jotil, and jougik bakko using machine learning algorithms. The use of this model will identify if a sentence is a Sorol, Jotil, or Jougik sentence. The text carries a lot of important information, but it becomes very difficult to extract that information manually from plenty of texts. Natural Language Processing and Machine Learning make this task quite easy. It automatically extracts a lot of important information, and it gets the job done in less time and at a lower cost. Businesses can, and still do, benefit greatly by building systems that can work and extract information from text automatically. A data set's data is analyzed and classified using natural language processing, and the classification is carried out in NLP. Multi-level classification in addition to binary level classification, are two different types of text categorization. All throughout this paper, we will discuss multi-level text classification as our dataset contains three different sorts of text. The types are Sorol bakko, Jotil bakko, and Jougik bakko. Accurate data and the amount of data play a big role in building and training a model. If there is a lot of data, it will be easy to understand the formulation or structure of a model, and then it will give a more accurate result. So it is necessary to collect junk-free and proper data through the data cleaning process. We have used six machine learning algorithms Random Forest, Logistic Regression, Decision Tree Classifier, KNN, SVM, and SGD. From the model, the result we get is a unique output from our dataset.

In this classification, we compared input data with training data. If the dataset and data were more accurate, then we would get more accuracy, so we cleaned the dataset. In this process, we train the model, and after that, we give input to the model to predict its label. Thus, it'll compare the input with the training data and give the nearest result. We'll know about this briefly in this paper.

2 Literature Review

Shetu et al [1] developed an algorithm that can determine whether the sentence was written in sadhu basha or chorito basha. This has helped in identifying Guruchandali Dosh, a typical grammar mistake in written Bangla. Two types of data were collected. One of them was Sadhu Data, while the other one was Cholito Data. They collected data from daily newspapers and popular Bengali novels.

Bijoy et al [2] collected data from different kinds of sources such as Bangla Blog, Conversation and Story etc. To get the results, their workflow was data processing, tokenization using the Countervector, and data cleaning. They use various classification techniques after processing the data for the feeding model to anticipate the sentence and deliver very accurate decision-making. Random

Forest and XGBoost produce the highest accuracy of 96.39% in parallel in their approach to analyze large amounts of data.

Čandrlić et al [3] purposed of developing a system model and methods for converting textual knowledge into relational databases was the focus of their work. This study used a one-way knowledge node approach, where the links connecting the nodes matter only when they move from one node to another. This work implements the technology developed and makes use of the research findings to turn natural language sentences into an enhanced and formalized record.

According to Dhar et al. [4], different classifiers performed differently when it came to categorizing Bangla texts. Eight separate domains or text categories containing a total of 8000 Bangla text documents were gathered from various web news sources for the experiment. Two weighting systems based on word association and term aggregation were utilized as experimental feature extraction techniques.

Over 1,000 product reviews and opinions were gathered by Shafin et al [5]. Therefore, based on Bangla comments and reviews, their goal was to ascertain consumer opinions around products, particularly product assessments both favorable and negative. KNN, decision trees, support vector machines (SVM), random forests, and logistic regression were among the categorization methods they employed. With a maximum accuracy of 88.81%, SVM surpassed all other algorithms.

Al-Radaideh et al [6] evaluated and developed effective classification strategies to create rule-based classifiers for medical free texts written in Arabic using association rule mining techniques. In addition to examining the impact of rule pruning on the rule generation stage, this study will look at the impact of integrating categorization and association rules in the domain of Arabic medical texts.

Bolaj et al [7] proposed an effective ontology-based and supervised learning method for categorizing Marathi text. It automatically classifies and detects patterns in various document types by combining data mining, natural language processing (NLP), and machine learning techniques. The system accepts Marathi documents as input. Article preprocessing includes validation of the input, tokenization, elimination of stopwords, stemming, and morphological analysis.

Dhar et al [8] investigated methods to classify Bangla text documents from an open web corpus using both traditional features and machine learning techniques. Here, they combine a dimensionality reduction approach (40% of TF) with the TF-IDF feature to increase the accuracy of the overall lexical matching process and determine the domain categories or classes of text documents.

Islam et al [9] used three supervised learning strategies: Support Vector Machines (SVM), Naive Bayes (NB), and Stochastic Gradient Descent (SGD). These types of algorithms are used for comparing Bengali document categorization. In this research, they used two alternative feature selection strategies in an effort to examine the effectiveness of those three classification algorithms.

Sen et al [10] analyzed 75 BNLP research publications in detail and divided them into 11 categories, including word group, part-of-speech tagging, sentiment analysis, fraud and forgery detection, information extraction, machine translation, named entity identification, and question-answering systems. Language processing and recognition, word meaning clarification, and summarization. They discuss the drawbacks of BNLP and present and potential future trends while describing traditional machine learning and deep learning techniques utilizing a variety of datasets.

Tuhin et al [11] proposed two machine learning methods to extract sentiment from Bangla texts. The six different emotional classes were happiness, sadness, kindness, excitement, anger, and fear. The Topic Approach and the Naïve Bayes Classification Algorithm used for topical techniques yielded the best results on both levels. They used a hand-compiled data corpus of over 7,500 sentences as a learning resource.

Das et al [12] presented an overall opinion mining system and an effective automated opinion polarity identification technique based on features for determining the polarity of phrases in documents. As a result of the evaluation, the accuracy was 70.04% and the recall was 63.02%.

Hasan et al [13] evaluated sentiment in Bangla texts. Contextual valence analysis is employed. They used SentiWordNet to prioritize each word and WorldNet to determine the part-of-speech meaning of each word. They determine the overall positivity, negativity, and neutrality of a sentence in relation to its overall meaning. Using valence analysis, they have developed a unique approach to determining emotions from Bangla texts.

Uddin et al [14] created a Long Short-Term Memory (LSTM) neural network which was used to analyze negative Bangla texts. They collected objectionable tweets in Bangla from Twitter, as well as other bad Bangla phrases collected through Google Forms. In this study, four hyperparameters of the LSTM model were tuned in three steps.

Hassan et al [15] collected data from YouTube, Facebook, Twitter, internet news sources, and product review websites. Their data set contained 9337 post samples, where 72% of the textual data was held for Bangla text, and the remaining 28% of the samples were for Romanized Bangla text. Three types of fully connected neural network layers were used. Their model was based on RNN and more specifically used the neural network LSTM.

Sarker et al [16] described an effort to develop a Bengali closed-ended appropriate question-and-answer system. They looked at various sources to collect data, such as crowdsourcing, social media, and manual generation. For both the questions and the papers, they had to gather information through various methods. They gathered questions from Shahjalal University of Science and Technology students (SUST) and the official website of the university. There they found a list of frequently asked questions by applicants.

Monisha et al [17] used machine learning-based techniques to categorize the questions. They have used four different types of machine learning algorithms, such as SVM, Naive Bayes, Decision Trees, and Stochastic Gradient Descent.

Moreover, stochastic gradient descent classifiers perform better on datasets. 25% of the dataset was utilized for testing, and 75% for training questions. The overall count of SUST-related questions in their question dataset is 15355 in Bengali.

Khan et al [18] implemented an effective QA system. This approach has 60% accuracy. In this study, they used WordNet to experiment with extracting the exact answer from the data set and reducing the complexity of replacing pronouns with the most appropriate nouns. They used over 50 sentence pairs as their dataset.

Urmi et al [19] provided a contextual similarity-based method, based on an N-gram language model, to identify stems or root forms in Bangla. They used a 6-gram model for their phylogeny identification process, which improved the accuracy of the corpus by 40.18%. About 1,593,398 sentences in the test corpus cover various topics such as news, sports, blogs, websites, business journals, and magazines.

Ahmad et al [20] described using Bengali word embeddings to address the problem of Bengali document classification. Using the K-means technique, they cluster the Bengali words based on their vector representations. They completed the classification task using the Support Vector Machine (SVM) and received an F1 score of about 91%. In three steps, their task was completed. Firstly, for each word in a corpus, create word embeddings. Secondly, reduce the vector dimension. Finally, make word-representing vector clusters.

Rahaman et al [21] presented a two-phase automatic hand sign and written Bangla sign recognition using a Bangla language modeling algorithm. They have developed an approach for modeling in the Bangla language that finds all hidden characters. In this experiment, the system recognizes words with an average accuracy of 93.50% in BdSL, compo digits with an average accuracy of 95.50%, and sentences with an average accuracy of 90.50%.

Haque et al [22] provided the "Subject-Verb Relational Algorithm" as the algorithm whose objective is to determine the sentence's main verb's validity from a semantic perspective. specific subjects when translating with a machine. With 600 sentences, they tested the subject-verb relationship algorithm. For 598 sentences, the system produced accurate results. During testing, the algorithm's accuracy was 99.67%.

Islam et al [23] discussed how deep learning may be used to generate Bangla text. Special kinds of RNN (Recurrent Neural Network) and Long Short-term Memory are used in their study. They used LSTM in their article for the Bangla Text Generator. Online Prothom Alo's website provided them with a corpus of 917 days' worth of newspaper text.

Abujar et al [24] used an extraction method to summarize a text in Bangla. They have suggested that analytical models can be used to summarize Bangla text. For the purpose of summarizing Bangla texts, this paper introduced a novel approach to sentence grading. When their technique was evaluated, the system displayed good accuracy.

Dhar et al [25] described forth two hypotheses: (a) the word length of medical texts is longer than that of other texts when measured in terms of characters, and

(b) the length of medical texts‘ sentences, when measured in terms of their word counts, is longer than that of other texts. They gathered data from a Bangla daily newspaper's online corpus of Bangla news articles. The sample set for their investigation includes texts from the five textual categories of politics, business, sports, medicine, and legal.

Hamid et al [26] suggested a technique to distinguish transliterated Bangla sentences from interrogative statements in Bangla. Explore rule-based techniques, supervised learning approaches, and deep learning approaches to find solutions. With the use of machine learning methods including Support Vector Machine, k-Nearest Neighbors, Multilayer Perceptron, and Logistic Regression, they were able to attain accuracy levels of 91.43%, 75.98%, 92.11%, and 91.68%, respectively.

Razzaghi et al [27] described the parsing and classification of questions for FAQs using machine learning. With over 3000 questions on the Internet, we've compiled FAQs and non-FAQs about sports, foods, and computers (especially the Internet). They demonstrated the 80.3% accuracy of the SVM with Naive Bayes.

Wang et al [28] suggested a multilayer sub-neural network using a separate structural match for each mode exists. It is used to transform features in various modes into features in the same mode. SQuAD, Wiki Text, and NarrativeQA are three corpus databases from which experimental data was collected.

Oliinyk et al [29] detected successfully propaganda indicators in text data and aims to develop a machine learning model, preliminary data processing and feature extraction techniques, and binary classification tasks. The Grid Search cross-validation algorithm and the Logistic Regression model have been used to perform categorization.

Alian et al [30] investigated techniques for identifying paraphrases in English and Arabic texts and provide an overview of previous work that they have proposed. For recognizing paraphrases in English, better results were obtained using WordNet-based measures. Accuracy of deep learning by statistics function yields the highest accuracy.

Mohammad et al [31] used a number of text processing, feature extraction, and text categorization processes. For the Arabic language, lexical, syntactic, and semantic aspects are extracted to overcome the shortcomings and constraints of the available technologies. They crawled Twitter data using the Twitter Streaming API. More than 8000 tweets were gathered. Scikit-learn, a Python-based machine learning library, was used to create the SVR model.

Lamba et al [32] provided a survey of numerous plagiarism detection methods applied to various languages. They used NLP techniques. The effectiveness of plagiarism detection has been investigated using a wide variety of text preparation approaches.

Ngoc et al [33] discovered several straightforward features that let them conduct both operations on Twitter data with a level of efficiency that is highly competitive. It's interesting to note that they also support the importance of

applying word alignment methods from machine translation assessment measures to the overall performance of these activities.

2.1 Comparison:

Reference	Dataset Quantity	Performance evaluation	Lack of scope
Md. Hasan Imam Bijoy et.al.[2] introduced an automated approach for Bangla sentence classification	Collected bangla data from different kinds of sources such as BanglaBlog, Conversation and Story etc.	Random Forest and XGBoost produce the highest accuracy of 96.39%.	
Md. Musfique Anwar et. al. [34] implements a technique using context -sensitive grammar rules with all types of Bangla sentences.	sentence is taken as input of the parsing system +Sentence	28 decomposition rules and 90% success rate in all cases	sentences composed of idioms and phrases are beyond the scope Of the paper, mixed sentence are out of discussion
Pooja Bolaj et. al. presents an [7] efficient Marathi text classification system using Supervised Learning Methods and Ontology based classification.	Set of Marathi text documents + Sentence	Naïve Bayes (NB), Modified K-Nearest Neighbor (MKNN) and Support Vector Machine (SVM) algorithms used+ output is set of classified Marathi documents as per the class label.	
Qasem A. Al-Radaideh et. al. [6] proposed rule-based classifier for Arabic medical text		the ordered decision list strategy out performed other methods, with an accuracy rate of 90.6%.	
Lenin Mehedy et.al.[35] introduced an approach of bangla syntax analysis		Proposed all Bangla sentences, including complicated, compound,exclamatory, and optative ones, are accepted under these rules' context-free use.	

Reference	Dataset Quantity	Performance evaluation	Lack of scope
Bidyut Das et. al. [36] proposed a novel system for generating simple sentences from complex and compound sentences	sentence	Modified Stanford Dependency (MSD) and Simple Sentence Generation (SSG) Algorithm+ as per judgment of five human linguistic experts the system's accuracy is 91.102%	The system sometime generates incomplete sentences
Parijat Prashun Purohit et. al. [37] proposes a semantic analyzer that can semantically parse the Bangla sentences.	1120 sentence + sentence	Complex sentence (word length 5) give 100% accuracy and Compound sentence (word length 5) give 100% accuracy (accuracy varies with word length)	
K. M. Azharul Hasan et. al. [38] described the detection of Semantic Errors from Simple Bangla Sentences		The approach is for simple sentences of the form SOV, although the classification of nouns and verbs may be applied for various types of Bangla phrases, including complicated and compound sentences as well as sentences with numerous verbs.	
Ayesha Khatun et.al. [39] proposed Statistical Parsing of Bangla Sentences by CYK Algorithm	Collected 2025 different kinds of sentences and word lengths from different bangla sites	The average accuracy of a simple sentence is 92.75%, but the average accuracy of a complex and compound sentence is 83.75% and 76.66%, respectively. Accuracy depended on the length of word and the number of parsed sentences.	increase the system's efficiency by expanding the probabilistic context -free grammar's existing vocabulary and grammar.
Parijat Prashun Purohit et.al. [40] propose a framework for the semantic analyzer that can parse the Bangla sentence semantically	2120 sentences of various word length	Simple ,complex & compound sentence(word length 5) give 98.67% ,100% & 100% accuracy respectively (accuracy varies with word length)	

3 Methodology

The measures required to implement the proposed "Simple, Complex, And Compound Sentence Detection Using Machine Learning for the Bangla Language" are listed below.

3.1 Dataset Preparation:

There are two key stages to the dataset preparation process.

- Data Collection
- Preprocessing of Data

3.2 Data Collection:

Our dataset gathered information from several blogs, Facebook, and the SSC (Secondary School Certificate) Bangla 2nd Paper book. We gathered about 2727 data from the above mentioned sources and the dataset was one we had built in. Three separate classes—Simple, Complex, and Compound—are included in our dataset. Our dataset contains 2 columns- the first column contains sentences that we collected from different resources and the last one contains types of sentences.

Sentence	TypeOfSentence
সন্ধ্যায় পাখিরা বাসায় ফেরে	সরল বাক্য
সন্ধ্যা হয় এবং পাখিরা বাসায় ফেরে	যৌগিক বাক্য
যখন সন্ধ্যা হয় তখন পাখিরা বাসায় ফেরে	জটিল বাক্য
দর্শক মাত্রই আশ্চর্য হয়েছে	সরল বাক্য

Fig. 1. Simple, complex, and compound form data in Bangla.

3.3 Preprocessing of Data:

Data Preprocessing is the process of transforming unstructured and raw data into useful sets of information so that analysis using data mining can operate as anticipated. Preprocessing of data is a necessary step since even after data collection because certain mistakes might still exist. A specific technique can be used to address these issues. The accomplishment of a data analysis project is closely correlated with how well or poorly data preprocessing was done.

To produce clean data, procedures such as cleaning duplicate data, removing punctuation, label encoding, dataset separation, and tokenization must be carried out. Because punctuation is often used in sentences, removing it gives us a clean dataset. When label encoding, the text is first labeled, followed by an instantaneous label, and then processing the outcome using a large dataset. Model comprehension requires the use of LabelEncoder by the Python scikit-learn library. Categorization is a feature of the Label-Encoder, which transfers data into columns and reloads encoded text data all at once. Tokenization of simple, complex, and compound forms is required for classifiers in a Bengali dataset. After that, the text has been split, and to prepare the text for classifiers preprocessing is carried out using tokenization.

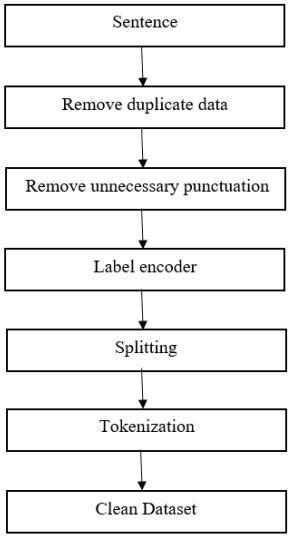

Fig. 2. Preprocessing method The totals for all categories of clean data are displayed below.

We obtained 2703 data after cleaning the dataset.

Type of Sentence	**Cleaned Sentence**
Simple	904
Complex	902
Compound	897

3.4 Data Vectorization or Distribution:

A great usefulness provided by the Python scikit-learn library is CountVectorizer. It's utilized to change a particular sentence into a vector depending on how frequently (count) each word appears across the whole text. The size of the n-grams we want to utilize will be specified by the ngram_range parameter, thus 1, 1 would give us unigrams (n-grams made up of only one word), while 1-3 would give us n-grams made up of one to three words.

- Unigram: We pass the value of n=1 to the n-grams function to produce unigram or 1-grams and also calculate the word frequency of the words.
- Bigram: We pass the value of n=2 to the n-grams function to produce bi-grams or 2-grams and also calculate the word frequency of the words.
- Tigram: We pass the value of n=3 to the n-grams function to produce tri-grams or 3-grams and also calculate the word frequency of the words.

Sentence	Unigram	Bigrams	Trigrams
অন্ধকার হয়ে আসছে এখনো আমরা হোটেলে পৌঁছাইনি	('অন', 1), ('ধক', 1), ('হয', 1)	('অন ধক', 1), ('ধক হয', 1), ('হয আসছ', 1)	('অন ধক হয', 1), ('ধক হয আসছ', 1), ('হয আসছ এখন', 1)

Fig. 3. Example of n-gram distribution

Here, we have determined the value for the quantity of simple, complex, and compound datasets' documents, words, and unique words and visualized them with the figure:

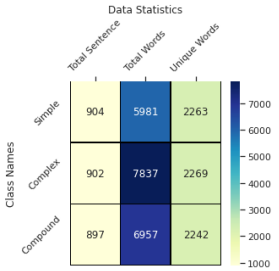

Fig. 4. Statistics between class name within sentence

4 Model and performance:

4.1 Proposed Model:

A range of supervised and unsupervised models were available thanks to machine learning. Using six of the most pertinent classification techniques in ML supervised model, we categorized Bengali texts as simple, complex, and compound. There are several types of machine learning algorithms, including support vector machine, Logistic Regression, Decision Tree, Random Forest, KNN (K-nearest neighbor) and SGD (Stochastic Gradient Descent).

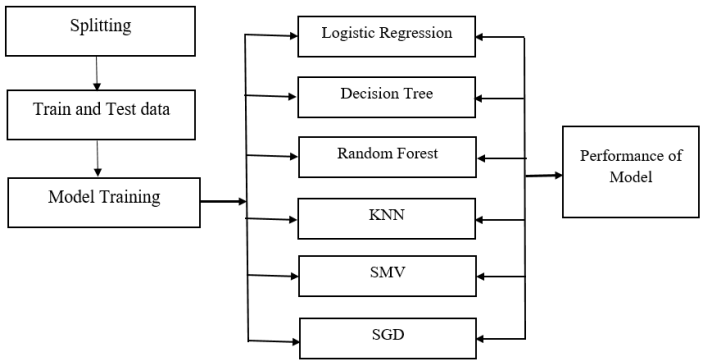

Fig. 5. Working form input

4.2 Model Performance:

Based on our results in the classification of Bengal data text, we give a short statement below.

Logistic regression Classifier: Logistic regression is also a supervised algorithm used to categorize dependent variables. LR is used to express the connection between dependent and independent variables. We get 84.84% accuracy with our dataset using the LR algorithm.

Decision Tree Classifier: For solving classification problem Decision tree classifier is mostly used and also for regression problems. Decision tree organizes a series of roots in a tree structure. From all other algorithms, the decision tree predicts too much accurately and the score is 93.72%.

Random Forest Classifier: A classification system made up of several decision trees is called the random forest classifier, which deals with high-dimensional data. Since we are just utilizing a portion of the input in our model, we can easily accommodate hundreds of characteristics, which makes training our model faster than training decision trees. We get a good accuracy of 91.68%.

KNN Classifier: K-nearest neighbor algorithm is used for regression and also for machine learning in text processing. It classifies data by assuming the similarity between new data and available data. Include new information in the category that is much more similar to the previous existing category. Through our KNN model, we get an accuracy of 73.94%.

SVM Classifier: Support Vector Machine is a powerful algorithm because it does not require much training data to start giving accurate results. SVMs perform the classification test by drawing a hyperplane that is a line in 2D or a plane in 3D in such a way that the categories can be differentiated by that line. Using SVM, our dataset gets 85.95% accuracy.

SGD Classifier: The optimization procedure of stochastic gradient descent is frequently used in machine learning applications to identify the model parameters that best match the expected and actual outputs. It is an imprecise method. Through this method, we get 87.06% accuracy.

Here, we demonstrate our noteworthy results for Table 2 with accuracy, precision, recall, and f1-score.

Table 1. Significant results for each classifier

Model Name	Accuracy	Precision	Recall	F1-score
Logistic regression	84.84%	84.84	84.84	84.84
Decision Tree	93.72%	93.72	93.72	93.72
Random Forest	91.68%	91.68	91.68	91.68
KNN	73.94%	73.94	73.94	73.94
SMV	85.95%	85.95	85.95	85.95
SGD	87.06%	87.06	87.06	87.06

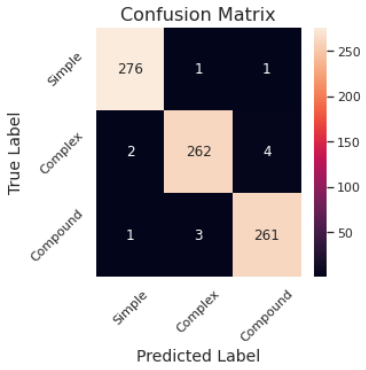

Fig. 6. Confusion Matrix on prediction

5 Result and Discussion:

We implemented numerous classification methods, including KNN, decision trees, SGD, support vector machine (SVM), random forests, and logistic regression in machine learning models using the simple, complex and compound form, and the results are quantifiable. All other algorithms were surpassed by the decision tree, which had an accuracy of up to 93.72

There are many work related to simple ,complex and compound sentence. Many of them similar to ours are of different languages. Some of the Bangla language works are available related to this but not exactly same as our work. They worked in syntex analysis, simple to complex and compound, generate bangla simple ,complex, compound to english simple, complex and compound etc. But our work is the first sentence classification(simple, complex, compound) work in Bangla Language.

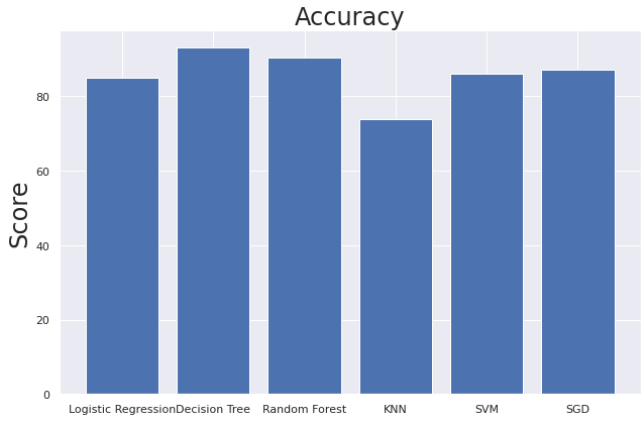

Fig. 7. Accuracy bar graph for classification of Bengali text

6 Conclusion

There are some analogous works in Bangla, although they differ somewhat from our work. However, we are the first to classify sentences in Bangla as simple, complex, or compound using Bangla NLP. This study can be viewed as the first significant step in novel extensibility initiatives. In this paper, we suggest a model development classification approach based on Decision Tree, Random Forest, Logistic Regression, KNN, SVM, and SVG using a simple, complex and compound dataset with over two thousand data points. Data were compared with training data and input data. Experimental findings demonstrate the proposed strategy's effectiveness and productivity. While doing this work, we came across a lot of

suggestions that may be considered for further development of the suggested system. Utilizing these techniques, association rules are created by using more precise stemming algorithms. Different pruning techniques can be used to affect classification accuracy. In addition, we will work on converting these ideas into user-friendly programs that are called applications. We will expand our dataset to achieve more accuracy. In the future, we also intend to apply Deep Learning techniques to enhance system performance.

References

1. Shetu, S.F., Saifuzzaman, M., Parvin, M., Moon, N.N., Yousuf, R. and Sultana, S., 2020, July. Identifying the writing style of bangla language using natural language processing. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.
2. Bijoy, M.H.I., Hasan, M., Tusher, A.N., Rahman, M.M., Mia, M.J. and Rabbani, M., 2021, July. An Automated Approach for Bangla Sentence Classification Using Supervised Algorithms. In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.
3. Čandrlić, S., Katić, M.A. and Pavlić, M., 2019. A system for transformation of sentences from the enriched formalized Node of Knowledge record into relational database. Expert Systems with Applications, 115, pp.442-464.
4. Dhar, A., Mukherjee, H., Dash, N.S. and Roy, K., 2018, October. Performance of classifiers in bangla text categorization. In 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET) (pp. 168-173). IEEE.
5. Shafin, M.A., Hasan, M.M., Alam, M.R., Mithu, M.A., Nur, A.U. and Faruk, M.O., 2020, December. Product review sentiment analysis by using NLP and machine learning in Bangla language. In 2020 23rd International Conference on Computer and Information Technology (ICCIT) (pp. 1-5). IEEE.
6. Al-Radaideh, Q.A. and Al-Khateeb, S.S., 2015. An associative rule-based classifier for Arabic medical text. International Journal of Knowledge Engineering and Data Mining, 3(3-4), pp.255-273.
7. Bolaj, P. and Govilkar, S., 2016. Text classification for Marathi documents using supervised learning methods. Int. J. Comput. Appl, 155(8), pp.0975-8887.
8. Dhar, A., Dash, N.S. and Roy, K., 2018. Application of tf-idf feature for categorizing documents of online bangla web text corpus. In Intelligent Engineering Informatics (pp. 51-59). Springer, Singapore.
9. Islam, M., Jubayer, F.E.M. and Ahmed, S.I., 2017. A comparative study on different types of approaches to Bengali document categorization. arXiv preprint arXiv:1701.08694.
10. Sen, O., Fuad, M., Islam, M.N., Rabbi, J., Masud, M., Hasan, M.K., Awal, M.A., Fime, A.A., Fuad, M.T.H., Sikder, D. and Iftee, M.A.R., 2022. Bangla Natural Language Processing: A Comprehensive Analysis of Classical, Machine Learning, and Deep Learning Based Methods. IEEE Access.
11. Tuhin, R.A., Paul, B.K., Nawrine, F., Akter, M. and Das, A.K., 2019, February. An automated system of sentiment analysis from Bangla text using supervised learning techniques. In 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS) (pp. 360-364). IEEE.
12. Das, A. and Bandyopadhyay, S., 2010. Phrase-level Polarity Identification for Bangla. Int. J. Comput. Linguistics Appl., 1(1-2), pp.169-182.

13. Hasan, K.A. and Rahman, M., 2014, December. Sentiment detection from bangla text using contextual valency analysis. In 2014 17th International Conference on Computer and Information Technology (ICCIT) (pp. 292-295). IEEE.
14. Uddin, A.H., Dam, S.K. and Arif, A.S.M., 2019, December. Extracting severe negative sentence pattern from bangla data via long short-term memory neural network. In 2019 4th International Conference on Electrical Information and Communication Technology (EICT) (pp. 1-6). IEEE.
15. Hassan, A., Amin, M.R., Al Azad, A.K. and Mohammed, N., 2016, December. Sentiment analysis on bangla and romanized bangla text using deep recurrent models. In 2016 International Workshop on Computational Intelligence (IWCI) (pp. 51-56). IEEE.
16. Sarker, S., Monisha, S.T.A. and Nahid, M.M.H., 2019, September. Bengali question answering system for factoid questions: A statistical approach. In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-5). IEEE.
17. Monisha, S.T.A., Sarker, S. and Nahid, M.M.H., 2019, May. Classification of bengali questions towards a factoid question answering system. In 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT) (pp. 1-5). IEEE.
18. Khan, S., Kubra, K.T. and Nahid, M.M.H., 2018, December. Improving answer extraction for bangali q/a system using anaphora-cataphora resolution. In 2018 International Conference on Innovation in Engineering and Technology (ICIET) (pp. 1-6). IEEE.
19. Urmi, T.T., Jammy, J.J. and Ismail, S., 2016, May. A corpus based unsupervised Bangla word stemming using N-gram language model. In 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV) (pp. 824-828). IEEE.
20. Ahmad, A. and Amin, M.R., 2016, December. Bengali word embeddings and it's application in solving document classification problem. In 2016 19th international conference on computer and information technology (ICCIT) (pp. 425-430). IEEE.
21. Rahaman, M.A., Jasim, M., Ali, M. and Hasanuzzaman, M., 2020. Bangla language modeling algorithm for automatic recognition of hand-sign-spelled Bangla sign language. Frontiers of Computer Science, 14(3), pp.1-20.
22. Haque, M. and Huda, M.N., 2016, May. Relation between subject and verb in Bangla Language: A semantic analysis. In 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV) (pp. 41-44). IEEE.
23. Islam, M.S., Mousumi, S.S.S., Abujar, S. and Hossain, S.A., 2019. Sequence-to-sequence Bangla sentence generation with LSTM recurrent neural networks. Procedia Computer Science, 152, pp.51-58.
24. Abujar, S., Hasan, M., Shahin, M.S.I. and Hossain, S.A., 2017, July. A heuristic approach of text summarization for Bengali documentation. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-8). IEEE.
25. Dhar, A., Dash, N.S. and Roy, K., Weighing Word Length and Sentence Length as Parameters for Subject Area Identification in Bangla Text Documents.
26. Hamid, M.M., Alam, T., Ismail, S. and Rabbi, M., 2018, September. Bangla Interrogative Sentence Identification from Transliterated Bangla Sentences. In 2018 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-6). IEEE.
27. Razzaghi, F., Minaee, H. and Ghorbani, A.A., 2016, October. Context free frequently asked questions detection using machine learning techniques. In 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 558-561). IEEE..

28. Wang, D., Su, J. and Yu, H., 2020. Feature extraction and analysis of natural language processing for deep learning English language. IEEE Access, 8, pp.46335-46345.
29. Oliinyk, V.A., Vysotska, V., Burov, Y., Mykich, K. and Fernandes, V.B., 2020. Propaganda Detection in Text Data Based on NLP and Machine Learning. In MoMLeT+ DS (pp. 132-144).
30. Alian, M. and Awajan, A., 2020, April. Paraphrasing identification techniques in English and Arabic texts. In 2020 11th International Conference on Information and Communication Systems (ICICS) (pp. 155-160). IEEE.
31. Mohammad, A.S., Jaradat, Z., Mahmoud, A.A. and Jararweh, Y., 2017. Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features. Information Processing & Management, 53(3), pp.640-652.
32. Lamba, H. and Govilkar, S., 2017. A survey on plagiarism detection techniques for indian regional languages. Int. J. Comput. Appl, 975, p.8887.
33. Ngoc Phuoc An, V., Magnolini, S. and Popescu, O., 2015. Paraphrase identification and semantic similarity in twitter with simple features.
34. Anwar, M.M., Anwar, M.Z. and Bhuiyan, M.A.A., 2009. Syntax analysis and machine translation of Bangla sentences. International Journal of Computer Science and Network Security, 9(8), pp.317-326.
35. Mehedy, L., Arifin, N. and Kaykobad, M., 2003. Bangla syntax analysis: A comprehensive approach. In Proceedings of International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh (pp. 287-293).
36. Das, B., Majumder, M. and Phadikar, S., 2018. A novel system for generating simple sentences from complex and compound sentences. International Journal of Modern Education and Computer Science, 11(1), p.57.
37. Purohit, P.P., Hoque, M.M. and Hassan, M.K., 2014, December. Feature based semantic analyzer for parsing Bangla complex and compound sentences. In The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014) (pp. 1-7). IEEE.
38. Hasan, K.A., Hozaifa, M. and Dutta, S., 2014, December. Detection of semantic errors from simple Bangla sentences. In 2014 17th International Conference on Computer and Information Technology (ICCIT) (pp. 296-299). IEEE.
39. Khatun, A. and Hoque, M.M., 2017, February. Statistical parsing of Bangla sentences by CYK algorithm. In 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 655-661). IEEE.
40. Purohit, P.P., Hoque, M.M. and Hassan, M.K., 2014, October. An empirical framework for semantic analysis of Bangla sentences. In 2014 9th International Forum on Strategic Technology (IFOST) (pp. 34-39). IEEE.