2023 26th International Conference on Computer and Information Technology (ICCIT)
13-15 December, Cox's Bazar, Bangladesh

Navigating Bengali Linguistics: Insights from Machine and Deep Learning Perspectives for Categorization of Sentences

Md Minhazul Islam*, Tanbeer Jubaer†, Azmain Yakin Srizon‡, Md. Rakib Hossain§,
S. M. Mahedy Hasan¶, Md. Farukuzzaman Faruk‖ and A. F. M. Minhazur Rahman**
*†‡¶‖**Department of Computer Science & Engineering
§Department of Electronics And Telecommunication Engineering
Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh
E-mails: minannu.2001@gmail.com, tanbeerjubaer@gmail.com, azmainsrizon@gmail.com,
rakib.ete12.ruet@gmail.com, mahedy@cse.ruet.ac.bd, faarukuzzaman@gmail.com, m.r.saurov@gmail.com

***Abstract*—The paper presents an innovative exploration into the classification of Bengali sentences, an essential aspect of advancing natural language processing (NLP) for languages characterized by rich linguistic diversity. In this comprehensive study, various machine learning and deep learning methodologies were meticulously applied to navigate the intricate linguistic landscape of the Bengali language. The research draws upon and synthesizes insights from previous studies that have focused on aspects such as sentence attributes, simplification algorithms, context-free grammar applications, and grammar-rule based transformations. The core of this study is anchored in the deployment of three distinct vectorization approaches - BERT-based, Glove-based, and TF-IDF-based - for the nuanced classification of Bengali sentences into simple, complex, and compound categories. A series of classifiers were applied, with the BERT-based vectorization demonstrating exemplary proficiency, registering an impressive 98.07% overall test accuracy. This finding underscores the model's robustness and versatility, illuminating its potential applicability beyond the designated dataset to encompass bespoke, real-life datasets. The research is set against the backdrop of a recently published large-scale Bengali sentence dataset titled Bangla Transformation of Sentence Dataset (BTSD) consisting of 3793 samples, which has filled a significant void in the existing literature, enabling a more intricate and detailed analysis. The findings of this study not only contribute to the enriched understanding of Bengali sentence structures but also hold implications for the broader field of computational linguistics and NLP.**

***Index Terms*—Natural Language Processing, BERT, Deep Learning, Sentence Classification, Bengali Text Processing**

I. Introduction

The sophistication and diversity inherent in natural language processing (NLP) are prominently exhibited in the endeavor to decipher and categorize distinct sentence structures. Such an endeavor is markedly intricate in the context of languages characterized by extensive linguistic diversity, as exemplified by Bengali. The current study is anchored in the exploration of Bengali sentence classification.

A recent study [1] highlighted the significance of compatibility, proximity, and expectancy in enhancing the clarity of Bengali sentences. This research employed the Long Short Term Memory (LSTM) network and one-hot encoding to effectively analyze these sentence attributes. Another study [2] focused on the simplification of English and Bengali sentences to improve machine translation quality, utilizing the character level Neural Machine Translation (NMT) architecture for both sentence classification and simplification. A research [3] in the context of the Indonesian language offered insights applicable to Bengali, introducing an algorithm based on context-free grammar to distinguish between active and passive sentence constructions, enhancing plagiarism detection systems. An additional study [4] described a grammar-rule based approach, utilizing Modified Stanford Dependency (MSD) structures, to transform complex and compound sentences into simpler forms, establishing foundational grammatical identification processes.

This study employs BERT-based, Glove-based, and TF-IDF-based vectorizations for the detailed classification of Bengali sentences into simple, complex, and compound categories. Among various classifiers tested, the BERT-based approach demonstrated superior performance, with a notable 98.07% overall test accuracy, showcasing its adaptability to both designated and custom real-life datasets.

II. Literature Review

Research in Bengali sentence classification via machine and deep learning is sparse. Few existing studies offer insights into classifying sentences into simple, complex, and compound categories, indicating a noticeable gap in this field. Research employing six distinct machine learning algorithms was conducted on this topic[5]. The study yielded notable findings, with the most promising outcome achieved through the implementation of a decision tree classifier, resulting in an impressive accuracy rate of 93.72%.Another research conducted slightly different topic[6]. They measure sentence Readability by defined algorithm. A prominent study [7] used LSTM and Bi-LSTM models on the Bangla Transformation

979-8-3503-5901-5/23/$31.00 ©2023 IEEE

of Sentence Classification dataset [7], which comprises 3,793 annotated sentences. This dataset is pivotal for classifying Bengali sentences and is also utilized in our study. Another research [8] introduced a hybrid machine learning model for identifying Bengali cyberbullying on social media, a method similar to ours.

A different study [9] applied the Bidirectional Encoder Representations from Transformers [10] model to classify traffic injury types, using over 750,000 unique crash narrative reports, and achieved 84.2% accuracy and an AUC of 0.93 ± 0.06 per class. However, existing works, including [7], are limited by simplistic approaches and dataset errors. Our paper addresses this gap by employing a refined methodology using BERT for sentence classification. Traditional machine learning algorithms were considered but the fine-tuned BERT [10] model proved superior. This study aims to reveal complex patterns in Bengali language usage, fostering progress in natural language processing across various fields. One significant application of this research is in authorship attribution, highlighting the extensive utility of advanced Bengali sentence classification.

III. Materials and Methods

A. Dataset Description

The research is anchored on the 'Bangla Transformation of Sentence Dataset (BTSD),' [7] a meticulously assembled collection of sentences curated explicitly for this study. Contained within the repository as a raw data file titled "Bangla Transformation of Sentence Dataset (BTSD).xlsx", this dataset comprises 3793 sentences extracted from publicly available Facebook pages. The authors of the study [7] assert the dataset's reliability and appropriateness, attributing it to a rigorous curation process. However, a detailed examination revealed the presence of inaccurately annotated data within this dataset. Amendments have been made to rectify these inaccuracies, and a detailed account of these corrections will be presented in subsequent sections of this paper.

B. Dataset Division

The dataset employed in this study was segregated into three distinct subsets: training, validation, and testing, following a 70-15-15 partition scheme. Seventy percent of the data was designated as the training set, serving as the foundational data on which the model was trained. This enabled the model to assimilate and learn from labelled examples, grasping inherent patterns and relationships. The validation set played a pivotal role in the model's refinement and optimization during the training phase. This subset facilitated performance monitoring and model adjustments, preserving the integrity of the test set. The latter was exclusively reserved for evaluating the model's final performance and its generalization capabilities. The accuracy score, as delineated in the results section, was derived from this test set.

The dataset is categorized into three distinct classes, labeled as 0, 1, and 2, corresponding to simple, complex, and compound (সরল, জটিল ও যৌগিক) sentences respectively, as elucidated in Table I. Furthermore, Table II provides a detailed exposition of the dataset's distribution, offering insights into the class-wise allocation of samples across the training, validation, and test sets.

Table I: Classes under consideration

Class Name (In English)	Class Name (In Bengali)	Label
Simple	সরল	0
Complex	জটিল	1
Compound	যৌগিক	2

Table II: Dataset distribution: number of class-wise train, validation and test samples in the dataset

Label	Train Samples	Validation Samples	Test Samples
Simple	1079	231	227
Complex	831	169	174
Compound	745	169	168

Table III: Some examples of wrong labelling

Example Sentences	Label	Correct
দুঃখও আসলে বিপদ আসে	1 (জটিল)	0 (সরল)
যে সত্য কথা বলে তাকে সবাই ভালোবাসে	0 (সরল)	1 (জটিল)
তুমি মনে মনে যা কামনা করেছে তা সফল হোউক	2 (যৌগিক)	1 (জটিল)
আমাদের একটি পরিবার সমাবেশ ছিল	1 (জটিল)	0 (সরল)
তিনি পরীক্ষার জন্য কঠোর অধ্যয়ন করেছিলেন, তবুও তিনি ভাল পারফর্ম করেননি	1 (জটিল)	2 (যৌগিক)

Table IV: Some examples of sentences with duplicate words

Original Samples	Corrected Samples
সে যদি কঠিন কাজ করে, তবে সে সে পরিশ্রম সমর্পণ দ্বারা পরিস্কার পরিস্কার সাফল্য অর্জন করতে পারে পারে পারে পারে পারে	সে যদি কঠিন কাজ করে, তবে সে পরিশ্রম সমর্পণ দ্বারা পরিস্কার সাফল্য অর্জন করতে পারে
আমি যদি সমস্ত প্রতিষ্ঠানের জন্য উন্নত পরিচালনা পরিচালনা পদ্ধতি প্রয়োগ পারি পারি, তবে আমি আমি প্রতিষ্ঠানটির করতে করতে পারব।।।	আমি যদি সমস্ত প্রতিষ্ঠানের জন্য উন্নত পরিচালনা পদ্ধতি প্রয়োগ পারি, তবে আমি প্রতিষ্ঠানটির করতে পারব।
তুমি যদি প্রতিষ্ঠানের সকল দক্ষতা উন্নত করতে করতে পারো পারো প্রতিষ্ঠানের প্রতিষ্ঠানের নতুন প্রকল্পের প্রকল্পের জন্য রাজনৈতিক করতে করতে পারবে।।	তুমি যদি প্রতিষ্ঠানের সকল দক্ষতা উন্নত করতে পারো প্রতিষ্ঠানের নতুন প্রকল্পের জন্য রাজনৈতিক করতে পারবে।

Table V: Some examples of sentences with spelling mistakes (mistakes are marked red)

Example Sentences	Label
যতি না পড়ো তবে পাশ করবে না।	1 (জটিল)
চোরকে যতেষ্ট দড়ি দিলে নিজেকে ঝুলিয়ে ফেলবে	0 (সরল)

C. Preprocessing

Data of disarray and disorganization can precipitate inaccurate conclusions, underscoring the cardinal role of data preprocessing in data mining. This procedure is characterized by the excision of superfluous or redundant information, which does not contribute to the training process and poses a risk of generating confusion during the classification phase. The dataset [7] employed in this research is not devoid of

inaccuracies. In the course of data preprocessing, three categories of errors were identified: incorrect labeling, repetitive Words: The dataset contained instances of word duplication and orthographic errors.

Addressing and amending these issues was integral to the data refinement process. Table III provides exemplifications of incorrect labeling, with the interpretations of 0, 1, 2 being previously delineated in Table I. Bengali sentences adhere to foundational principles that define their categorization and structure [11]. Nonetheless, a scrutiny of the dataset revealed inconsistencies in the labeling of sentences. Consequently, a manual examination of each sentence was executed, leading to necessary amendments in labels. These rectifications were meticulously performed by individuals proficient in Bengali literature to ensure accuracy and consistency.

Table IV underscores a notable issue of word repetition within individual sentences. Such redundancies can inject noise, impeding the efficient training of the model. Steps were undertaken to amend this anomaly to enhance data quality. Additionally, orthographic errors were identified within the dataset. Initial model testing with this flawed data yielded inaccurate predictions. However, post-correction of these errors, a notable enhancement in prediction accuracy was observed upon reevaluation. While the exhaustive correction of all spelling errors within the dataset is a laborious task, diminishing the number of errors bolsters the training efficacy and performance of the model. Instances of such errors are cataloged in Table V.

D. Vectorization Approach

The dataset is presently constituted of pre-processed text. For classification purposes, it is requisite to convert this text data into vector form to facilitate processing by various models. This section delineates the transformation of text data into its vectorized counterpart. Several vectorization methodologies were assessed, culminating in the identification of three efficacious processes, which are expounded upon in subsequent subsections.

1) BERT-based Vectorization: The process of fine-tuning BERT[12] is delineated below.First step is achieved by tokenizing the review texts using the bert-base-multilingual-cased [12] for BERT-1 and sagorsarker/bangla-bert-base [13] for BERT-2 model and used BERT auto tokenizer to tokenize from these model. Post tokenization, special tokens [CLS] and [SEP] are affixed at the beginning and end of the word tokens respectively, with [CLS] being integral for classification tasks. Subsequently, these tokens are associated with their indexes in the tokenizer vocabulary. A uniform sequence length (with a maximum cap of 512) is established, with reviews being truncated or padded accordingly. Attention masks are then formulated to differentiate between actual and padded tokens, exemplified with a sequence length of 64.

Therefore, for the text, যে সত্য কথা বলে তাকে সবাই ভালোবাসে, Tokenized text will be: ['যে', 'সত', '##য', 'কথা', 'বলে', 'তাকে', 'সবাই', 'ভাল', '## ে', '##বাস', '## ে', '## ে'], Tokenized text with special token will be: [‘[CLS]’, 'যে',,'সত', '##য', 'কথা', 'বলে', 'তাকে', 'সবাই', 'ভাল', '## ে', '##বাস', '## ে', ‘[SEP]’] ‘input_ids’ will be: [101, 2060, 36079, 9294, 2085, 2080, 2271, 2553, 5477, 2094, 59368, 2094, 102], and ‘attention_mask’ will be: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1].

In the context of BERT tokenization, the “#” character denotes the inception of subword tokens, resulting from the subword tokenization process that fragments words into smaller constituent units. The BERT input representation is a composite of token embeddings, segmentation embeddings, and position embeddings.

2) GLOVE-based Vectorization: GloVe (Global Vectors for Word Representation) is a popular word embedding method that captures semantic meanings of words based on their co-occurrence information in a given corpus [14]. Mathematically, the objective function J to minimize is:

$$J = \sum_{i,j=1}^{V} f(X_{ij}) \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log(X_{ij}) \right)^2$$

where, V is the size of the vocabulary, w_i and $\tilde{w}_j$ are the word vectors for words i and j, b_i and $\tilde{b}_j$ are the biases for words i and j, X_{ij} represents the number of times word i occurs in the context of word j in the training corpus and f is a weighting function that assigns relatively lower weight to rare and frequent co-occurrences to avoid biases. The weighting function $f(x)$ is defined as

$$f(x) = \begin{cases} (x/x_{\max})^\alpha & \text{if } x < x_{\max} \\ 1 & \text{otherwise} \end{cases}$$

where $x_{\max}$ and α are hyperparameters. This weighting function ensures that the model doesn't overfit to extremely common word pairs.

GloVe vectorization transforms words into numerical vectors using statistical information from word co-occurrence within a corpus. A co-occurrence matrix is first built, and the GloVe model factorizes it by minimizing a specific objective function with optimization techniques like gradient descent. This process results in vector embeddings that encapsulate words' semantic and syntactic relationships. The study utilizes Bangla GloVe Vectorization [15], transforming a vast Bengali text corpus from Wikipedia articles and news content into vector representations using a 39 million parameter model. This model generates 1×300 dimensional vectors, converting the studied dataset into a numerical format suitable for machine learning applications.

3) TF-IDF-based Vectorization: TF-IDF (Term Frequency-Inverse Document Frequency) is a popular method for numerical representation of text in machine learning [16]. It assesses a word's importance in a document compared to a corpus. Term Frequency (TF) is the term's frequency in a document divided by the total terms. Inverse Document Frequency (IDF)[17] is the logarithm of total corpus documents divided by documents containing the term. The TF-IDF score, obtained by multiplying TF and IDF, represents a term in a document within the corpus. Documents are then represented

Table VI: Hyperparameter values for fine-tuning BERT

Algorithm Name	Hyperparameter Values	Accuracy
BERT-1, BERT-2	max_position_embedding = 512, batch_size = 16, epochs = 2, learning_rate = 3e-5, epsilon = 1e-08, clipnorm = 1.0, name_or_path = sagorsarker/bangla-bert-base, bert-base-multilingual-cased	97.01%, 98.07%

Table VII: Hyperparameter values for training Glove-based approach

Algorithm Name	Hyperparameter Values	Accuracy
SVC	C=10.0, kernel='rbf', degree=3, gamma='scale', random_state=42	89.63%
Cat Boost	Iterations = 500, Learning_rate = 0.1, Depth = 6, L2_leaf_reg = 3, Random_seed = 42, Verbose = 0	85.06%
Logistic Regression	Class_weight = None N_jobs = None	86.12%

Table VIII: Hyperparameter values for training TF-IDF-based approach

Algorithm Name	Hyperparameter Values	Accuracy
SVC	C=1, kernel='linear', degree=3 gamma='5', random_state=0	74.71%
Xgboost	n_estimators=10, max_depth=5 learning_rate=.0051, random_state=0	75.09%
CatBoost	Iterations: 500, learning_rate: 0.1 depth: 6	75.36%

as TF-IDF score vectors, suitable for various machine learning algorithms.

IV. Experimental Outcomes

A. Experimental Settings

Despite the time-intensive nature of training data with the BERT model [10], we limited our training to 2 epochs instead of the initially planned 10 epochs. Kaggle, with its computational resources featuring 29GB RAM and a P100 GPU with 16GB memory, was used for training, taking 329 seconds for 10 epochs. An observation was made that even after an extensive duration of training spanning 10 epochs, there was no significant increase in validation accuracy. Empirical evidence suggested that 2 epochs were adequate, leading to the decision to limit the training to this duration for final results. The Bangla BERT Base [13], a variant of the BERT [12] architecture tailored for the Bengali language, was employed for tokenization and modeling. We represent this version of BERT as BERT-2 in our study. An alternative tokenization version which is represented as BERT-1 in this study, 'bert-base-multilingual-cased', was also explored, maintaining consistency in all other parameters. The hyperparameter values for BERT fine-tuning are detailed in Table VI.

For the conventional machine learning models, two distinct vectorization strategies were tested: TF-IDF (Term Frequency-Inverse Document Frequency) and GloVe (Global Vectors for Word Representation). These vectorization methods were applied to a range of machine learning algorithms, including but not limited to, Random Forest Classifier, XGBoost Classifier, and Support Vector Classifier. The GloVe-based and TF-IDF-based approaches are elaborated in Table VII and Table VIII, respectively.

Table IX: Performance of different proposed models

Model Name	Acc.	Class	Pre.	Rec.	F1 Sco.
BERT-2	98.07%	Simple	0.97	1.00	0.98
		Complex	0.99	0.98	0.98
		Compound	0.99	0.96	0.98
BERT-1	97.01%	Simple	0.99	0.97	0.98
		Complex	0.99	0.97	0.98
		Compound	0.93	0.98	0.95
SVC (GLOVE)	89.63%	Simple	0.87	0.93	0.90
		Complex	0.93	0.89	0.91
		Compound	0.90	0.86	0.88
CatBoost (TF-IDF)	75.36%	Simple	0.68	0.89	0.77
		Complex	0.80	0.64	0.71
		Compound	0.86	0.67	0.76

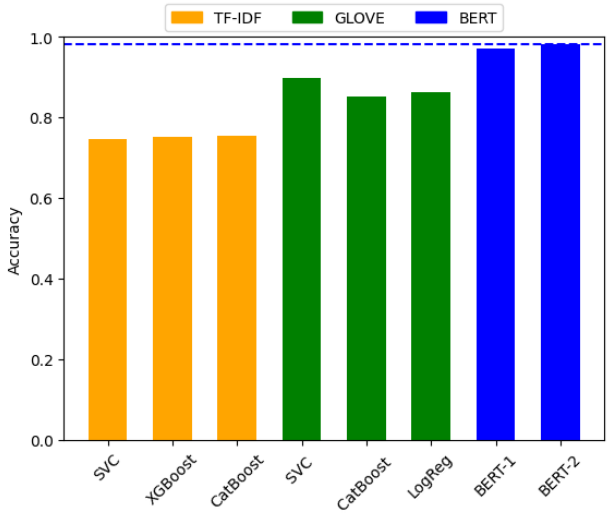

Figure 1: Overall comparison of models with different vectorization

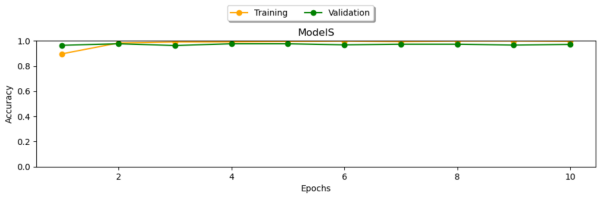

Figure 2: Training and validation accuracy for BTSD dataset

B. Obtained Results

This section delineates the evaluation results for classifiers that were based on TF-IDF and Glove vectorization methods, and encompasses models like Logistic Regression, Support Vector Classifier, Cat Boost Classifier, among others. The

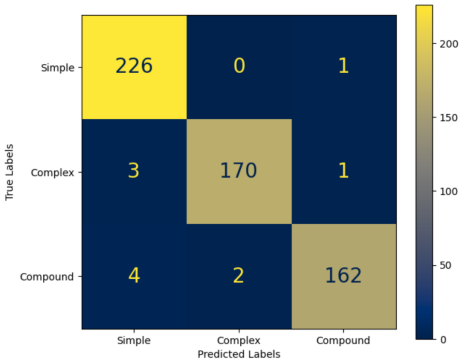

Figure 3: Confusion matrix for the BTSD dataset

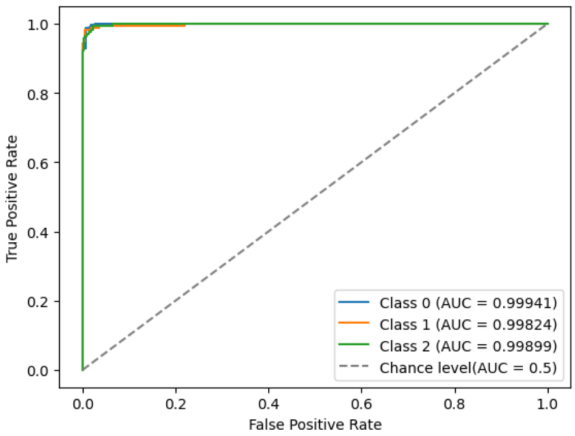

Figure 4: ROC curve for BTSD dataset

assessment of the fine-tuned BERT classification model is also detailed. The Bangla BERT Base [13] version and BERT Base Multilingual Cased [12] were employed for vectorization, owing to BERT's inherent vectorization proficiencies. Table IX exhibits the efficacies of premier models spanning three vectorization techniques. BERT-2, structured on Bangla BERT Base [13], manifested an exemplary accuracy of 98.07%. BERT-1, based on BERT Base Multilingual Cased, also registered substantial accuracy, attaining 97.01%. Within the realm of GLOVE vectorization, the Support Vector Classifier emerged as the most proficient, recording an accuracy of 89.63%. Conversely, the Cat Boost Classifier, under the TF-IDF vectorization paradigm, topped the charts albeit with a diminished accuracy of 75.36%. Figure 1 facilitates a comparative analysis of all the models, accentuating the preeminence of BERT-2. Figure 2 delineates the trajectory of training and validation accuracy for the BTSD [7] dataset over a span of 10 epochs. An anomaly observed was the subordination of training accuracy to validation accuracy during the inaugural epoch, a trend that reversed from the second epoch onwards, propelling the training accuracy to the vicinity of 100

The confusion matrix for the test data extracted from the BTSD dataset is illustrated in Figure 3. Of the 569 data

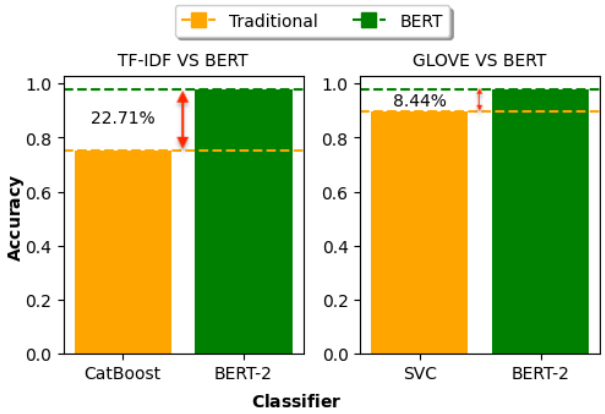

Figure 5: BERT vs. best of TF-IDF and GLOVE

samples constituting the test set, 558 were accurately classified by the BERT-2 model. Figure 4 unveils the ROC curve for the dataset. The Area Under the Curve (AUC) values bear testimony to the model's adeptness - 0.99941 for simple sentence class, 0.99824 for complex sentence, and 0.99899 for compound sentence, underscoring the model's precision in class discrimination.

C. Result Analysis and Discussion

Table IX elucidates the unparalleled accuracy of the BERT-based model compared to others, attributed to its adeptness at deciphering intricate contextual relationships inherent in language. BERT's bidirectional training and attention mechanisms facilitate an enriched comprehension of word contexts, enhancing its efficacy in diverse NLP tasks. Its extensive pre-training on copious datasets imbues it with an acute sensitivity to linguistic subtleties. The model's proficiency isn't confined to accuracy alone; its consistency is manifested when tested with structurally varied yet semantically identical sentences. BERT's aptitude to accurately classify such sentences underscores its robustness. In contrast, the Support Vector Classifier, employed in the GLOVE-based approach, while commendable, was not on par with BERT. The omission of words without vectors in the GLOVE dataset, even with its 39M parameters [15], was a conscious decision made during the study.

The diminished accuracy in the TF-IDF approach is attributable to the vectorizer's inability to encapsulate semantic intricacies effectively. The absence of potent Bengali stemmers resulted in inferior lemmatized forms, undermining this method's effectiveness. The persistence of low accuracy, despite the employment of Grid Search and increased epochs, culminated in the decision to confine the training to 2 epochs, curtailing the training duration substantially. The ROC curves, with AUC values of 0.9994, 0.9982, and 0.9989 for the three classes, exemplify the model's acumen in class differentiation and predictive prowess. These metrics corroborate the model's exemplary performance in classifying specific categories, underscoring its suitability for tasks entailing these classes. In

Table X: Example of custom dataset

Sentence	Label	Prediction
বাংলাদেশ একটি সুন্দর দেশ।	সরল	সরল
আমরা সবাই একটি মহান জাতির অংশ।	সরল	সরল
যখন বিপদ আসে, তখন দুঃখও আসে।	জটিল	জটিল
যে হিমালয়ে বাস করিতেন, সেই হিমালয়ের তিনি মিতা।	জটিল	জটিল
দুঃখ এবং বিপদ এক সাথে আসে।	যৌগিক	যৌগিক
এতক্ষণ অপেক্ষা করলাম কিন্তু গাড়ি পেলাম না।	যৌগিক	যৌগিক

Table XI: Example of exceptional cases

Sentence	Correct Label	Prediction
তুমি আসবে বলে অপেক্ষা করছি।	জটিল	সরল
আমি দেখা করব বলে এসেছি।	জটিল	সরল

a previous study [7], they used an LSTM-based model with an accuracy of 91.04%. Another study [5] achieved 93.72% using a decision tree model. Our model outperformed both with a 98.07% accuracy.

D. Performance on Custom Dataset

The creation of a custom dataset [18] aimed to evaluate the model's consistency, especially after addressing dataset errors highlighted in the Preprocessing section. This balanced dataset, comprised of 60 randomly collected examples of simple, complex, and compound sentences from the internet, initially yielded 100% accuracy. However, the integration of unique sentence structures and exceptional cases introduced an element of complexity, as illustrated in Table X. Among these, BERT accurately classified 59 sentences, faltering on a single exceptional instance, marking a limitation of this study.

A specific challenge arises in the Bengali language with sentences containing the word '**বলে**'. Typically, when used as an '**অনুসর্গ**' between two simple sentences, it denotes a complex sentence. However, this is not a universal rule. The term can also function as an infinitive verb, leading to a simple sentence classification, as the model predicts. Examples of this linguistic nuance are documented in Table XI. This limitation is likely attributable to the model's lack of extensive training on such specialized instances.

V. Conclusion

This study delves into the classification of Bengali sentences, utilizing an array of machine learning and deep learning methodologies. The BTSD [7] dataset was employed, and despite the challenges arising from limited resources compared to English, a model was developed that showcased an impressive 98% accuracy. However, a noted limitation was the model's struggle with exceptional cases due to inadequate training on such instances. The research offers significant contributions to Bengali sentence classification, yet there remains room for enhancement. Future studies could focus on enriching the dataset and encompassing all potential exceptional cases to bolster the model's accuracy universally. The implications of this research are profound, especially in educational and digital communication contexts. In education, the classification aids learners with tailored exercises for diverse sentence structures. In the digital sphere, it empowers chatbots with contextual responsiveness, elevating user experience. Linguists, educators, and digital platforms alike can leverage these classifications for refined content categorization, nuanced language studies, and enhanced user engagement, marking a significant stride in Bengali language processing.

References

[1] M. A. H. Chowdhury, N. Mumenin, M. Taus, and M. A. Yousuf, "Detection of compatibility, proximity and expectancy of bengali sentences using long short term memory," in *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, IEEE, Jan. 2021.

[2] S. K. Mahata, A. Garain, D. Das, and S. Bandyopadhyay, "Simplification of english and bengali sentences for improving quality of machine translation," *Neural Processing Letters*, vol. 54, pp. 3115–3139, Feb. 2022.

[3] D. Anggraini, A. Benny Mutiara, T. M. Kusuma, and L. Wulandari, "Identification of active and passive sentence for plagiarism detection," in *2017 Second International Conference on Informatics and Computing (ICIC)*, pp. 1–6, 2017.

[4] B. Das, , M. Majumder, and S. Phadikar, "A novel system for generating simple sentences from complex and compound sentences," *International Journal of Modern Education and Computer Science*, vol. 10, pp. 57–64, Jan. 2018.

[5] R. K. Das, S. S. Sammi, K. Kobra, M. R. Ajmain, S. A. khushbu, and S. R. H. Noori, "Analysis of bangla transformation of sentences using machine learning," in *Key Digital Trends in Artificial Intelligence and Robotics* (L. Troiano, A. Vaccaro, N. Kesswani, I. Díaz Rodriguez, I. Brigui, and D. Pastor-Escuredo, eds.), (Cham), pp. 36–52, Springer International Publishing, 2023.

[6] S. Chakraborty, M. T. Nayeem, and W. U. Ahmad, "Simple or complex? learning to predict readability of bengali texts," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, p. 12621–12629, May 2021.

[7] R. K. Das, M. Islam, and S. A. Khushbu, "Btsd: A curated transformation of sentence dataset for text classification in bangla language," *Data in Brief*, vol. 50, p. 109445, 2023.

[8] A. Akhter, U. K. Acharjee, M. A. Talukder, M. M. Islam, and M. A. Uddin, "A robust hybrid machine learning model for bengali cyber bullying detection in social media," *Natural Language Processing Journal*, vol. 4, p. 100027, 2023.

[9] A. H. Oliaee, S. Das, J. Liu, and M. A. Rahman, "Using bidirectional encoder representations from transformers (bert) to classify traffic crash severity types," *Natural Language Processing Journal*, vol. 3, p. 100007, 2023.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] M. M. Anwar, M. Z. Anwar, and M. A.-A. Bhuiyan, "Syntax analysis and machine translation of bangla sentences," *International Journal of Computer Science and Network Security*, vol. 9, no. 8, pp. 317–326, 2009.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] S. Sarker, "Banglabert: Bengali mask language model for bengali language understanding," 2020.

[14] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[15] S. Sarker, "Bengali glove pretrained word vector," 2019.

[16] A. Aizawa, "An information-theoretic perspective of tf–idf measures," *Information Processing Management*, vol. 39, no. 1, pp. 45–65, 2003.

[17] A. Aizawa, "An information-theoretic perspective of tf–idf measures," *Information Processing &amp Management*, vol. 39, pp. 45–65, Jan. 2003.

[18] T. Jubaer, "Custom dataset for bangla sentence classification," 2023.