

Bangla Sentence Function Classification: Corpus Development, Model Benchmarking, and Interpretability

Swapnil Kundu Argha*, Abdullah Al Shafi†, Rowzatul Zannat‡, Shoumik Barman Polok§, Abdul Muntakim¶, Jannatul Ferdousi||, M.A. Moyeen**

Department of CSE, Khulna University of Engineering & Technology, Bangladesh * † § **

Institute of ICT, Khulna University of Engineering & Technology, Bangladesh † **

Department of CSE, Daffodil International University, Bangladesh ‡

Department of Computer Science, Kennesaw State University, United States ¶ ||

swapnilkundu01@gmail.com*, abdullah@iict.kuet.ac.bd†, w.rzrowza@gmail.com‡, polokbarman874@gmail.com§, amuntaki@students.kennesaw.edu¶, jannatulferdousi3110@gmail.com||, moyeen.kuet@gmail.com**

Abstract—Automatic sentence function identification is important for many downstream natural language processing (NLP) applications such as dialogue systems, text-to-speech synthesis, and machine translation. However, benchmark resources for Bangla sentence function classification remain limited. To mitigate this gap, this paper introduces a corpus of 10,000 Bangla sentences, manually annotated into four functional categories, namely declarative, interrogative, imperative, and exclamatory. The corpus is nearly balanced across the four classes, with high annotation reliability reflected by a Fleiss’ Kappa of 0.82. Furthermore, we evaluate multiple feature representations, including Bag-of-Words (BoW), TF-IDF, and Word2Vec, with several classical machine learning classifiers. In addition, two heterogeneous ensemble models, namely Single-Level Ensemble (SLE) and Double-Level Ensemble (DLE), are utilized to improve classification performance. Experimental results show that TF-IDF consistently outperforms Word2Vec, likely due to its ability to emphasize discriminative lexical cues associated with sentence functions, particularly given the relatively small corpus used to train Word2Vec. The DLE model with TF-IDF features achieves the best performance with accuracy and macro-F1 of 0.95, demonstrating the effectiveness of sparse lexical representations and heterogeneous ensemble learning for this task. Further cross-validation confirms the robustness of the approach, while LIME-based interpretability provides insights into model predictions. The developed corpus and model benchmarking establish strong baselines for Bangla sentence function classification.

Index Terms—sentence classification, ensemble learning, word embeddings, interpretability

I. INTRODUCTION

With about 250 million native speakers, mostly in Bangladesh and West Bengal of India, Bangla is one of the most frequently spoken languages in the world. Compared to languages like English or Chinese, Bangla is still relatively deficient in automated language processing resources [1]. But automated systems require solutions to understand the functional expression of large amount

of textual data as digital Bangla content has rapidly expanded its presence through various platforms [2].

Identifying sentence functions requires an evaluation of whether a sentence functions as declarative, interrogative, imperative, or exclamatory. The different functional categories allow a speaker to deliver his intended expression while controlling how listeners will understand the information delivered. For example, the declarative “সে বাড়িতে যায় (He goes home)” expresses a statement, while the interrogative “ভিতরে আসতে পারি? (May I come in?)” functions to obtain permission. Similarly, the imperative “গোলমাল করিও না। (Don't make a fuss.)” delivers a command, while the exclamatory “শৈশবের স্মৃতি কত মধুর! (How sweet childhood memories are!)” expresses strong feelings.

Reliable functional classification of a sentence has broad implications across various downstream natural language processing (NLP) applications. Dialogue systems [3] rely on the identification of the type of query to generate appropriate responses. Machine translation [4] benefits from preserving the intent of the sentence in multiple languages to maintain communicative accuracy. Sentiment and emotion detection can benefit from recognizing exclamatory and other functional cues, while stance detection can use the sentence function to better capture the communicative intent of the speaker [5]. Text-to-speech synthesis uses functional cues to produce suitable intonation patterns [6]. In short, without accurate understanding of sentence-level communicative function, higher-level NLP systems may fail to capture speaker intent, leading to reduced effectiveness in real-world language understanding.

The main contributions of this work are threefold.

- We present a novel Bangla sentence function corpus with 10,000 manually annotated sentences across four functional categories. To ensure high-quality labeling of the corpus, a reliable multi-annotator annotation framework with agreement validation is designed. Furthermore, the corpus is released as a publicly

available resource¹ to support reproducible research.

- A comprehensive benchmarking study for sentence function classification is conducted on various feature representations and classical machine learning models. We also benchmark two heterogeneous ensemble architectures to effectively combine diverse classifiers.
- We further incorporate interpretability and error analysis, providing insights into model behavior and linguistic challenges.

II. LITERATURE REVIEW

Early Bangla text classification research relied mainly on traditional machine learning methods due to the lack of large-scale datasets. Bijoy et al. [1] used ensemble models such as Random Forest and XGBoost for sentence classification, achieving 96.39% accuracy, while Das et al. [7] applied Decision Tree models for structural sentence classification with 93.72% accuracy. Furthermore, [8] introduced a large Bangla news dataset and showed that Word2Vec embeddings outperformed TF-IDF-based approaches, achieving a 0.96 F1-Score.

With the rise of deep learning, Sikder et al. [9] proposed a hybrid CNN-BiLSTM model that achieved 98% F1-Score, demonstrating the effectiveness of neural architectures over traditional methods. Recently, Al Shafi et al. [2] introduced BiST, a Bangla-English bilingual corpus for sentence structure and tense classification, with experiments showing the superiority of dual-encoder models over multilingual baselines.

III. PROPOSED METHOD FOR SENTENCE FUNCTION CLASSIFICATION

The complete workflow of our proposed method is shown in Fig. 1, which includes the sequence of corpus creation, data preprocessing, feature extraction, model training, evaluation, and interpretation.

A. Development of Bangla Sentence Function Corpus

The construction of a high-quality corpus is essential for robust text classification. Fig. 2 presents the workflow of the corpus creation process, including data collection, preprocessing, annotation, and quality verification.

Data Collection and Preprocessing: We prioritized collecting Bangla sentences that demonstrate different usage patterns through formal and informal usage across various online platforms which users can access without restrictions. Several popular Bengali newspapers serve as our primary sources that deliver high-quality formal content. To create a wider linguistic range, we also added Bengali blogs and online magazines and youtube comments together with digitized storybooks and novels which exist in the public domain. We have used automated

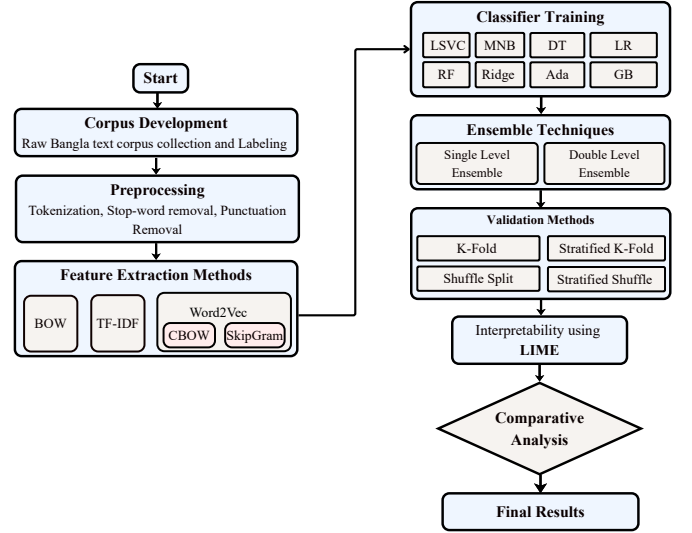


Figure 1: Overview of the proposed system.

web scraping techniques through Python scripts that used BeautifulSoup² to collect the initial raw sentences.

The collected text then underwent a rigorous preprocessing pipeline that includes removing HTML tags and non-Bengali characters, sentence segmentation using a Bengali sentence tokenizer, deduplication, and noise filtering to maintain focus on well-formed sentences.

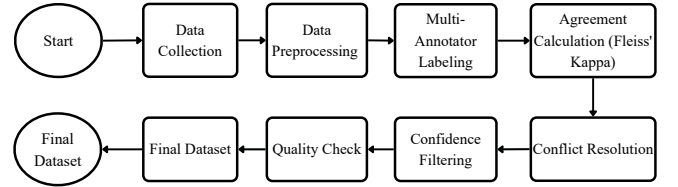


Figure 2: Workflow of the corpus development process.

Annotation Guidelines: Based on standard linguistic theory, we define four mutually exclusive functional classes, namely declarative, interrogative, imperative, and exclamatory for annotators.

A sentence was annotated as declarative if it conveyed a statement, fact, or opinion without expressing a query, command, or strong emotion. For example, ‘আমি আজ স্কুলে গিয়েছিলাম। (I went to school today.)’. In contrast, sentences were annotated as interrogative when they were intended to seek information or clarification, often characterized by the presence of interrogative markers such as কি, কেন, কোথায়, কবে, and কিভাবে. For instance, ‘‘তুমি কোথায় যাচ্ছ? (Where are you going?)’’.

Sentences expressing commands, requests, suggestions, or instructions were categorized as imperative. ‘দরজা

¹https://github.com/AbdullahRatulk/Bangla_Sentence_Function_Classification_Corpus

²<https://pypi.org/project/beautifulsoup4/>

বন্ধ করো। (Close the door.)’ would be an example of such a sentence. Meanwhile, sentences conveying strong emotions, such as surprise, joy, anger, or admiration, were labeled exclamatory. An example would be ‘আহা! কী সুন্দর দৃশ্য! (What a beautiful view!)’.

Algorithm 1 Corpus Annotation with Fleiss’ Kappa

Require: Sentences $S = \{s_1, \dots, s_n\}$

Ensure: Final labels $F = \{f_1, \dots, f_n\}$

```

1: Define annotators  $P = \{a_1, a_2, a_3\}$  and label sets  $\mathcal{F}$ 
2: for all  $s_i$  and  $a_j$  do
3:   Annotator assigns  $f_{ij} \in \mathcal{F}$ ,
4: end for
5: for all  $s_i$  do
6:   Compute item agreement  $A_i$ 
7: end for
8:  $\bar{A} = \frac{1}{n} \sum_{i=1}^n A_i$ 
9: Compute expected agreement  $A_e$ 
10: Fleiss’ Kappa,  $\kappa = \frac{\bar{A} - A_e}{1 - A_e}$ 
11: for all  $s_i$  do
12:    $f_i \leftarrow \arg \max_{f \in \mathcal{F}} \sum_{j=1}^3 \mathbf{1}(f_{ij} = f)$ 
13:   if no clear majority emerges then
14:     consensus discussion among annotators
15:   end if
16: end for
return Annotated corpus with final labels  $F$ 

```

Corpus Annotation: The multi-annotator framework which describes the corpus annotation process operates according to the steps which Algorithm 1 specifies. Three annotators assess each sentence in the corpus by applying their designated function labels from the defined function label set in annotation guideline. The reliability of the annotations is evaluated through inter-annotator agreement (IAA) assessment which uses Fleiss’ Kappa method. The final label for each sentence comes using majority voting. The final label assignment process requires annotators to hold a consensus discussion when a situation arises without a distinct majority. The resulting labels collectively form the finalized annotated corpus.

Final Corpus: The final corpus was created from the original collection of approximately 24,000 Bangla sentences from which 10,000 sentences underwent preprocessing and annotation to establish a reliable benchmark for sentence function classification.

B. Data Preprocessing

The raw Bangla sentences undergo a series of cleaning and preprocessing steps before feature extraction to reduce noise while preserving linguistic cues essential for function classification. The cleaning phase applies sequential operations, including the elimination of hashtags and special characters that do not carry linguistic meaning, space normalization, etc. Common stop words for Bangla³ are

removed with a critical exception made for wh-question words such as কী (what), কেমন (how), কখন (when), কোথায় (where), as these words are essential for identifying interrogative and/or exclamatory sentences. All punctuation marks are removed to ensure that punctuation does not serve as a direct cue for sentence function, forcing the classification model to rely on lexical and structural cues instead. A Bangla-aware custom tokenizer splits text on whitespace and returns only non-empty tokens after all cleaning steps, with the same tokenizer used across all feature extractors for fair comparison. In addition, the four function labels are encoded into integer IDs from 0 to 3 for model training.

C. Feature Extraction

We employ multiple feature extraction techniques to represent the preprocessed text as numerical vectors, namely Bag of Words (BoW), TF-IDF, and Word2Vec.

BoW [7] counts the frequency of each word or n-gram in a document while ignoring word order and grammar, implemented with CountVectorizer using configurable n-gram ranges. On the other hand, TF-IDF [1] weights terms by their frequency in a document and their rarity across the corpus according to the mathematical formulation where TF represents term frequency in a document and IDF represents inverse document frequency across the corpus. For example, the term কী (what) appearing in an interrogative document receives high TF in that document and moderate IDF, resulting in high TF-IDF weight that marks it as a useful feature for identifying interrogatives.

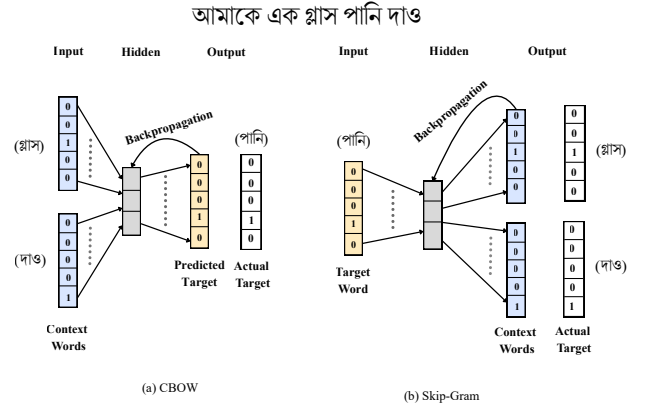


Figure 3: Comparison of (a) CBOW and (b) Skip-Gram architectures where CBOW predicts the target word from surrounding context words while Skip-Gram predicts context words from the target word.

In contrast, Word2Vec [8] is a two-layer neural network trained on the corpus to capture semantic relationships using two different architectures as shown in Fig. 3: (i) Continuous Bag of Words (CBOW) which predicts the target word from its surrounding context and performs well for frequent words, and (ii) Skip-gram which predicts

³<https://github.com/stopwords-iso/stopwords-bn>

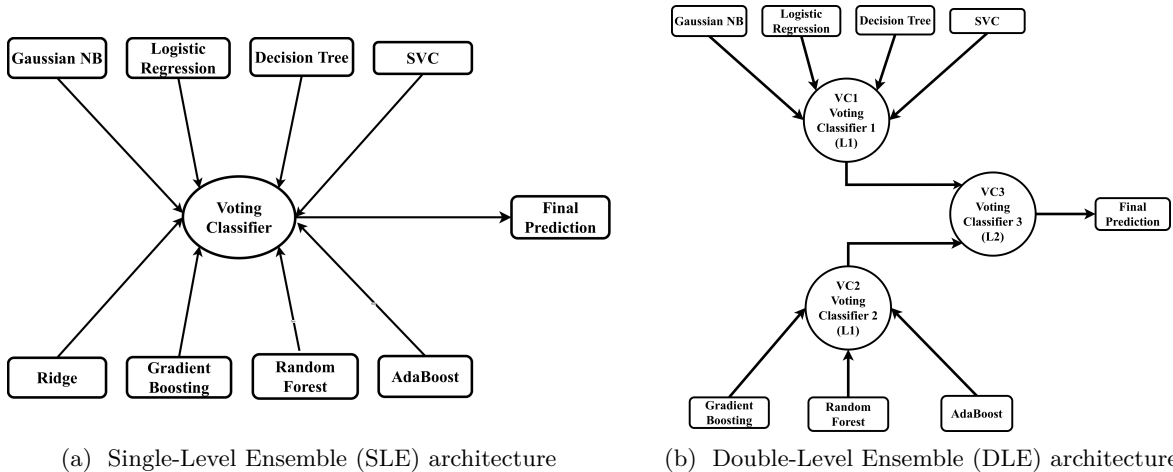


Figure 4: Architecture of the two heterogeneous ensemble models: (a) SLE where multiple classical classifiers feed into a Voting Classifier for final prediction; (b) DLE with two levels of voting classifiers leading to final prediction.

context words from the target word and tends to capture rare words and semantic relationships better.

D. Classification

We evaluate multiple classifiers spanning classical ML and heterogeneous ensemble approaches.

Classical ML Classifiers: The classical classifiers trained using the extracted features include Decision Tree (DT), Multinomial Naive Bayes (MNB), Logistic Regression (LR), Linear Support Vector Classifier (LSVC), Ridge classifier, Gradient Boosting (GB), AdaBoost (Ada), and Random Forest (RF).

Heterogeneous Ensemble Classifiers: We design two heterogeneous ensemble architectures, namely Single Level Ensemble (SLE) and Double Level Ensemble (DLE), to leverage the complementary strengths of different classical classification algorithms. Fig. 4 presents a detailed overview of both architectures.

SLE encompasses all classical ML classifiers operating on fixed feature vectors. These classifiers then feed into a voting classifier to produce the final classification output. On the other hand, DLE implements a more sophisticated hierarchical voting mechanism with two distinct aggregation levels. In the first level, the base classifiers are organized into two groups based on their underlying algorithmic characteristics. Group 1 consists of MNB, LR, DT, and LSVC, whose outputs are combined into voting classifier 1 (VC1). Group 2 comprises the homogeneous ensemble methods including RF, GB, and AdaBoost, whose predictions are aggregated into voting classifier 2 (VC2). In the second level, the outputs from VC1 and VC2 are fed into a third voting classifier (VC3) that performs the final aggregation to produce the ultimate prediction. This hierarchical structure with two layers of voting enables the model to capture different levels of decision making, first within similar classifier types and then across the two broad categories of classifiers.

It is to be noted that the voting classifiers use soft voting, where the final prediction is obtained by averaging the predicted probabilities of all base classifiers and selecting the class with the highest score.

Cross-Validation and Hyperparameter Tuning:

Robust evaluation is ensured through multiple cross-validation (CV) strategies including (i) K-fold where data is split into K folds with the model training on (K-1) folds and validating on 1 fold rotating K times, (ii) stratified K-fold which maintains original class proportions in each fold, (iii) shuffle split with multiple random train-validation splits, and (iv) stratified shuffle split combining random splitting with stratification. Grid search with 5-fold CV is used to optimize hyperparameters for each model with macro-F1 score as the objective metric, and after identifying the best parameters, models are refitted on the full training set.

E. Interpretability using LIME

The research employs Local Interpretable Model-agnostic Explanations (LIME) [10] to analyze the model’s prediction results. This local approximation method enables us to identify which words or phrases have the greatest impact on model’s prediction.

IV. EXPERIMENTAL SETUP AND RESULT ANALYSIS

Table I: Class distribution and sentence length statistics for the Bangla Sentence Function Corpus.

Function	Total	Length Bins (Tokens)			Avg. Len
		0-8	9-14	15+	
Declarative	2,487	881	932	674	10.78
Interrogative	2,532	895	902	735	11.02
Imperative	2,511	845	833	833	11.29
Exclamatory	2,470	804	833	833	12.52
Total	10,000	3,425	3,500	3,075	11.40

Table II: Comparative analysis of various feature representation and classifiers based on performance evaluation. All the metrics except accuracy are macro-averaged.

Metric	Feature	NB	LSVC	DT	LR	RF	Ridge	Ada	GB	SLE	DLE
Accuracy	BOW	0.86	0.93	0.84	0.93	0.94	0.90	0.88	0.94	0.94	0.94
	TF-IDF	0.87	0.94	0.85	0.94	0.94	0.92	0.89	0.94	0.94	0.95
	CBOW	0.83	0.90	0.81	0.89	0.91	0.89	0.86	0.91	0.93	0.92
	SkipGram	0.84	0.91	0.82	0.90	0.92	0.90	0.87	0.92	0.93	0.94
Precision	BOW	0.86	0.93	0.84	0.94	0.94	0.91	0.88	0.94	0.94	0.95
	TF-IDF	0.87	0.94	0.85	0.94	0.94	0.92	0.89	0.94	0.94	0.95
	CBOW	0.83	0.90	0.81	0.89	0.91	0.89	0.86	0.91	0.94	0.93
	SkipGram	0.84	0.91	0.82	0.90	0.92	0.90	0.87	0.92	0.93	0.94
Recall	BOW	0.86	0.93	0.84	0.93	0.94	0.90	0.88	0.94	0.94	0.95
	TF-IDF	0.87	0.94	0.85	0.94	0.94	0.92	0.89	0.94	0.94	0.95
	CBOW	0.83	0.90	0.81	0.89	0.91	0.89	0.86	0.91	0.93	0.93
	SkipGram	0.84	0.91	0.82	0.90	0.92	0.90	0.87	0.92	0.93	0.94
F1-Score	BOW	0.86	0.93	0.84	0.93	0.94	0.90	0.88	0.94	0.94	0.94
	TF-IDF	0.87	0.94	0.85	0.94	0.94	0.92	0.89	0.94	0.94	0.95
	CBOW	0.83	0.90	0.81	0.89	0.91	0.89	0.86	0.91	0.94	0.93
	SkipGram	0.84	0.91	0.82	0.90	0.92	0.90	0.87	0.92	0.93	0.94

Table III: Cross-validation performance evaluation of SLE and DLE models (mean $\pm$ std). Here, we set K = 5.

CV techniques	Features	SLE		DLE	
		Accuracy	Macro-F1	Accuracy	Macro-F1
K-Fold	BoW	0.94 $\pm$ 0.012	0.94 $\pm$ 0.012	0.95 $\pm$ 0.009	0.95 $\pm$ 0.009
	TF-IDF	0.94 $\pm$ 0.011	0.94 $\pm$ 0.011	0.95 $\pm$ 0.008	0.95 $\pm$ 0.008
	CBOW	0.92 $\pm$ 0.018	0.92 $\pm$ 0.018	0.93 $\pm$ 0.015	0.93 $\pm$ 0.015
	SkipGram	0.93 $\pm$ 0.015	0.93 $\pm$ 0.015	0.94 $\pm$ 0.012	0.94 $\pm$ 0.012
Stratified K-Fold	BoW	0.94 $\pm$ 0.010	0.94 $\pm$ 0.010	0.95 $\pm$ 0.007	0.95 $\pm$ 0.007
	TF-IDF	0.94 $\pm$ 0.009	0.94 $\pm$ 0.009	0.95 $\pm$ 0.006	0.95 $\pm$ 0.006
	CBOW	0.92 $\pm$ 0.015	0.92 $\pm$ 0.015	0.93 $\pm$ 0.012	0.93 $\pm$ 0.012
	SkipGram	0.93 $\pm$ 0.012	0.93 $\pm$ 0.012	0.94 $\pm$ 0.010	0.94 $\pm$ 0.010
Shuffle Split	BoW	0.94 $\pm$ 0.018	0.94 $\pm$ 0.018	0.95 $\pm$ 0.014	0.95 $\pm$ 0.014
	TF-IDF	0.94 $\pm$ 0.016	0.94 $\pm$ 0.016	0.95 $\pm$ 0.012	0.95 $\pm$ 0.012
	CBOW	0.92 $\pm$ 0.024	0.92 $\pm$ 0.024	0.93 $\pm$ 0.020	0.93 $\pm$ 0.020
	SkipGram	0.93 $\pm$ 0.020	0.93 $\pm$ 0.020	0.94 $\pm$ 0.016	0.94 $\pm$ 0.016
Stratified Shuffle Split	BoW	0.94 $\pm$ 0.014	0.94 $\pm$ 0.014	0.95 $\pm$ 0.011	0.95 $\pm$ 0.011
	TF-IDF	0.94 $\pm$ 0.013	0.94 $\pm$ 0.013	0.95 $\pm$ 0.010	0.95 $\pm$ 0.010
	CBOW	0.92 $\pm$ 0.020	0.92 $\pm$ 0.020	0.93 $\pm$ 0.017	0.93 $\pm$ 0.017
	SkipGram	0.93 $\pm$ 0.017	0.93 $\pm$ 0.017	0.94 $\pm$ 0.014	0.94 $\pm$ 0.014

A. Inter-Annotator Agreement

The Fleiss’ Kappa (κ) value for the developed corpus is 0.82, which is treated as ‘Almost Perfect’ in the standard, indicating a high level of inter-annotator agreement and strong consistency in the labeling process.

B. Corpus Statistics

As shown in Table I, each of the four functional classes contributes approximately 2,500 sentences, ensuring a near-equal distribution that prevents bias towards any single function. The three sentence length analysis bins which include 0-8 tokens and 9-14 tokens and 15+ tokens show that all functional classes use a combination of brief sentences, intermediate sentences, and extended sentences to create a corpus through which we can train models that can handle different input lengths.

C. Experimental Setup and Evaluation Metrics

The final curated dataset was partitioned into two subsets using stratified sampling, a training-validation set of 8000 sentences, and an unseen test set of 2000 sentences for the final unbiased evaluation. Model performance is assessed using standard metrics, including accuracy, precision, recall, and F1-Score.

D. Result Analysis

Table II presents the comprehensive performance comparison across four feature representations and eight classifiers. Linear models, including LSVC, Ridge, and LR demonstrate strong performance across most feature types. RF and GB emerged as the top performers, RF particularly excelled with TF-IDF features reaching 0.94 accuracy and F1-Score. DLE achieved 0.95 on the TF-IDF features, while SLE maintained 0.94 on most metrics.

In addition, the superior performance of TF-IDF over Word2Vec can be attributed to the discriminative lexical

Table IV: Representative genuine misclassification cases with DLE on TF-IDF features.

Text	Actual	Predicted	Primary Error Source
বাংলাদেশের এই প্রাকৃতিক দৃশ্য কত মনোমুগ্ধকর! (How mesmerizing this natural scenery of Bangladesh is!)	Exclamatory	Declarative	Exclamatory sentence with declarative syntax.
এই নীতিমালার কার্যকারিতা নিয়ে কোনো সন্দেহ আছে? (Is there any doubt about the effectiveness of this policy?)	Interrogative	Declarative	Rhetorical question interpreted as statement.
সংশ্লিষ্ট কর্তৃপক্ষকে দ্রুত প্রয়োজনীয় ব্যবস্থা গ্রহণ করতে হবে। (The relevant authorities must take necessary measures promptly.)	Declarative	Imperative	Obligation modality interpreted as directive.
গবেষণা অনুযায়ী জলবায়ু পরিবর্তনের প্রভাব ক্রমেই বৃদ্ধি পাচ্ছে। (According to research, the impact of climate change is increasing.)	Declarative	Exclamatory	Strong evaluative tone misinterpreted as exclamation.

markers of sentence functions, such as interrogative words, directive, and exclamatory expressions, etc. TF-IDF emphasizes such class-specific terms, whereas Word2Vec captures broader semantic relationships and may not preserve these functional cues as explicitly, particularly when trained on a relatively small corpus. Thus, sparse lexical weighting is well suited to this task.

The heterogeneous ensembles are also tested against several different CV methods to confirm their robustness, and the results are shown in Table III. The results confirm the stability of our ensemble approaches. Both SLE and DLE reach 92-95% accuracy with standard deviations between 0.6% and 2.4% for all feature representation.

Table IV presents representative examples of such misclassified sentences which highlight common challenges. Exclamatory sentences are frequently misclassified as declarative when exclamation expressions are underweighted. The use of directive verbs creates problems that lead to difficulties in determining between declarative and imperative sentence structures. Questions that contain embedded commands are sometimes treated as direct orders.

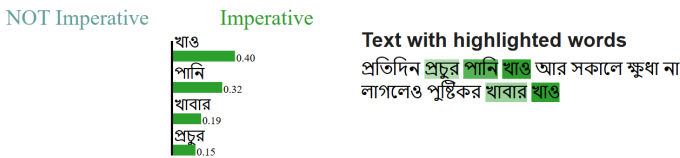


Figure 5: Interpreting model decisions through LIME, visualizing feature importance for predicted sentence class.

Fig. 5 provides an example of clear and intuitive illustration of the model’s decision-making process using LIME, highlighting the contribution of individual words toward the predicted Imperative class. Key token ‘খাও’ is assigned higher importance scores, which aligns well with their semantic association with actions and directives. This indicates that the model is able to capture meaningful lexical cues relevant to imperative sentence construction. Similar patterns were observed for the other classes.

V. CONCLUSION

The research presents a high-quality Bangla sentence function corpus together with a study that compares traditional learning methods and heterogeneous ensemble

learning techniques. The experimental findings demonstrate that TF-IDF which represents data in sparse form continues to perform effectively for the current task. The ensemble methods demonstrate superior performance stability among all tested models whereas DLE produces the highest overall performance results. The analysis of errors shows difficulties that arise from linguistic uncertainty which exists between different types of sentences. The newly developed corpus together with its benchmark tests offers valuable resources which enable future researchers to advance Bangla NLP studies. The future direction would be using transformer-based language models and larger datasets to enhance both classification accuracy and system resilience.

REFERENCES

- [1] M. H. I. Bijoy, M. Hasan, A. N. Tusher, M. M. Rahman, M. J. Mia, and M. Rabbani, “An automated approach for bangla sentence classification using supervised algorithms,” in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2021, pp. 1–6.
- [2] A. Al Shafi, S. K. Argha, M. Moyeen, A. Muntakim, and S. B. Polok, “Bist: A gold standard bangla-english bilingual corpus for sentence structure and tense classification with inter-annotator agreement,” in *SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL: Towards Inclusivity and Equality: Language Resources and Technologies for Under-Resourced and Endangered Languages@ LREC 2026*. European Language Resources Association (ELRA), 2026, pp. 143–152.
- [3] Z. Yi, J. Ouyang, Z. Xu, Y. Liu, T. Liao, H. Luo, and Y. Shen, “A survey on recent advances in llm-based multi-turn dialogue systems,” *ACM Computing Surveys*, vol. 58, no. 6, pp. 1–38, 2025.
- [4] S. Castilho and R. Knowles, “A survey of context in neural machine translation and its evaluation,” *Natural Language Processing*, vol. 31, no. 4, pp. 986–1016, 2025.
- [5] A. Al Shafi, M. M. Islam, S. I. Hossain, and K. M. A. Hasan, “Stancemoe: Mixture-of-experts architecture for stance detection,” in *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources @ LREC 2026*. European Language Resources Association (ELRA), 2026, pp. 244–251.
- [6] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He *et al.*, “Naturalspeech: End-to-end text-to-speech synthesis with human-level quality,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 6, pp. 4234–4245, 2024.
- [7] 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 *International Conference on Deep Learning, Artificial Intelligence and Robotics*. Springer, 2022, pp. 36–52.

- [8] M. M. Islam and M. T. Alam, "Bard: Bangla article classification using a new comprehensive dataset," in *2018 International Conference on Bangla Speech and Language Processing (ICB-SLP)*, 2018.
- [9] J. Sikder, P. Chakraborty, U. K. Das, and K. Dhar, "A hybrid approach for bengali sentence validation," *Artificial Intelligence Review*, vol. 57, no. 11, p. 316, 2024.
- [10] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, and G. Menegaz, "A perspective on explainable artificial intelligence methods: Shap and lime," *Advanced Intelligent Systems*, vol. 7, no. 1, p. 2400304, 2025.