Enhancing Banglish Text Accuracy: NLP-Based Error Detection and Correction

MD Mohibur Zaman
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
md.mohibur.zaman@g.bracu.ac.bd

Rakesh Rakshit
Computer Science and Engineering
Brac university
Dhaka, Bangladesh
rakesh.rakshit@g.bracu.ac.bd

Priya Saha
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
priya.saha@g.bracu.ac.bd

Ehsanur Rahman Rhythm
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
ehsanur.rahman.rhythm@g.bracu.ac.bd

Humaion Kabir Mehedi
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

Annajiat Alim Rasel
Computer Science and Engineering
Brac University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—In our country, where both Bengali and English are widely spoken, people often mix the two languages for their communication, creating something called "Banglish." However, because Bengali and English have distinct conventions for word order, Banglish material frequently has errors that make it difficult to interpret. There is very little research being performed on Banglish misspell detection. Most of the work on Banglish is based on either sentiment analysis or analyzing different types of reviews based on different types of comments on social media. In our research, we propose a new way to make Banglish text better by using Natural Language processing (NLP) tools that are good with language to detect errors and corrections. In this study, we tried to show a new way to make Banglish text more accurate. For the research, a custom dataset is being created based on different types of classification models, such as Word2Vec,Multinomial Naive Bayes (MultinomialNB), Support Vector Machine (SVM), RandomForestClassifier, DecisionTreeClassifier. Apart from the classification, we tried to integrate word suggestions for misspelled words based on Levenshtein distance. Which overall will help people write better Banglish by finding and fixing mistakes.**

***Index Terms*—classification, correct, misspelled, word, Banglish, suggestion, model**

I. Introduction

A spell checker is a tool that fixes mistakes and suggests appropriate words in writing. It can find words that are spelled wrong or used in the wrong way. This tool is used in computer programs like word processors and web browsers, or in social media apps. People have been working on making spell checkers, and they're really good for languages like English, German, and Chinese. However, multilingual languages like banglish don't have as many tools for checking mistakes because there isn't much information available yet.

In Bangladesh, people are mixing the two languages together in their writing, and we call this "Banglish." It happens because many people know both languages.However, there are certain grammatical errors in Banglish that can occasionally make it difficult to comprehend because Bengali and English employ words in different sequences. Since Bengali and English have additional rules, when people mix them in Banglish, there are mistakes like wrong spellings, sentences that don't fit, and words used in the wrong way. These mistakes can make it hard to understand what's actually being said. This study is about using NLP tools to fix those mistakes in Banglish.

By creating many different Banglish words that show how people mix Bengali and English, this research prepares to teach more efficient words in Banglish. This is like a mix of different ways to understand Banglish, and will learn to find different types of mistakes in Banglish. These mistakes can be small, like when sentences don't sound right or when words are spelled wrong. They happen because Bengali words and English letters are mixed together.

This research can be really helpful in many ways. First, it gives people a useful suggestion to write better Banglish words, so they can communicate more clearly when they use both languages. Second, the things we learned from making this tool can also help improve other computer tools for languages that mix together, like Banglish does. This may make communication better for many languages that are a mix of different ones.

In the next parts of this paper, we will talk about how we did our research, and what happened when we tried it. This research is our way of helping make multilingual language technology and computer understanding of these languages better. At the same time, we're trying to solve the problems that people using Banglish face when they write.

II. Literature Review

There have been many types of research conducted on increasing text accuracy by spell checking and error correction using NLP.

Author Maja Mitreska,Kostadin Mishev and Monika Simjanoska of article [1]showed the comparison between three models in typo correction accuracy. Models they used were

Multilingual BERT, DistilBERT and XLM-RoBERTa. They choose multilingual BERT as it supports multiple language models to compare with. They choose DistilBert and XLM-Roberta for better comparison. Moreover, to train the model they used the Croatian Language Dataset which has 14.7 million entries and also created a parallel synthetic dataset of misspelled words by using the noising function.The DistilBERT, on the other hand, was improved with a Subword BERT architecture, 256-batch size, and a lexicon of 100,000 words. The XLM-ROBERTa was finally adjusted with 2 epochs, 64 batch size, and a lexicon of 100,000 words.The test results demonstrate that the XLM-RoBERTa model had the highest accuracy, scoring 94.36% for the less corrupted dataset and 93.15% for the more corrupted dataset, respectively. Additionally, writers employed recall, a metric for word rate correction, to accurately assess the performance of the models during testing.XLM-RoBERTa again provides the best results after assessing the second measure, i.e., in the word rate correction section, it delivers around 86 percent as opposed to 80% and 72% for the other models, BERT and DistilBERT. Also, in precision and F1-score the XLM-RoBERTa surpasses both BERT and DistilBERT models. Overall their research shows for better word rate correction the best-performing model is XLMRoBERTa.

In the field of spell checking there is a problem that arises on a regular basis. To keep updated with the problem, the author of [2] article proposed a spelling error detection with the combination of Web search and the Statistical Language Models. Both offline and online n-grams are used to build the statistical language models.They made an effort to take use of the vast, dynamic information resources that individuals around the world update online. They used news articles to gather data for their research.It was divided into 10 categories.Their proposed method is for text based approach which is suitable for both online keyin and file operation.The author created two language models; the first is referred to as LM' and is made up of uni- and bi-grams. The second one, known as LM, is made up of trigrams, bigrams, and unigrams. They also combined web search with the LM' model. After the experiments, they illustrated some findings.Firstly, for both language models the precision rate is low. The LM'+Web Search got 0.33 precision and the traditional LM got 0.06 precision. This points out that in natural language processing new word detection is one of the hard problems.Secondly, in the case of the proposed method, recall rate is lower compared to the language model-based approach which is 0.63% in LM'+Web Search model and 0.78% in LM model. But in the case of the F1 score, the proposed model outperformed with 0.43 F1 score, whereas the LM model only had 0.11 F1 score.

[3] Bangla's intricate character system and rigid grammatical regulations make it challenging to comprehend. In this study, a hybrid strategy for actual word mistake detection and repair in Bangla is proposed. It combines two distinct methods: Bidirectional Long Short-Term Memory (LSTM), a specific kind of Recurrent Neural Network (RNN), and N-gram language models like bigram. As the current data sets have legitimate output but do not give input sets with genuine word errors, they first collected a Bangla data set from several sources, which is then further processed. After completing tokenization, the proposed approach creates bigram sequences from the data corpus. The Bidirectional LSTM model is then fed the sequences to forecast the precise word to replace the inconsistent term. Bidirectional LSTM recalls the forward and backward relationship between words, which improves the network's comprehension of context. This is a benefit of the suggested method. Additionally, the system uses word length matching to ensure that the output word length stays consistent with the original. The system has a highly encouraging performance, accurately communicating 82.86%.

Jodani [4] which is a Gujarati language spell checker and suggestion tool.In order to find misspelt words, it applied string similarity measurements.This spell checker also handles inflected words.This checkers accuracy shows 91.56% efficiency while Jodani algorithm used. When using Sara's spell checker, Jodani did better than Saras in terms of F-measure.Sara's F-measure was 50%, recall was 93%, and accuracy was 85%.When handling false positive situations, Jodani was effective.

Machine-generated documents and publication materials frequently contain errors stated by Xu et al in their paper [5]. However, several correction algorithms need to perform better for complicated problems, and hiring people to complete the task would be expensive. To address the issue, the prototype computer game Cipher was created. It motivates players to find textual flaws. Steganography, a fascinating gaming element, is used to achieve gamification. Users enjoy playing the game while providing valuable text annotations. 35 players tested the prototype during an evaluation experiment, producing 4,764 annotations. When the data had been filtered, the algorithm found typos that had been entered by hand as well as actual text mistakes that had existed before the texts were included in the game.

[6]He et al. said that the field of automatic marking of English texts has advanced quickly in recent years. It has progressively taken the role of teachers reading manuals and developed into a crucial instrument to lessen the strain of teaching. The two categories of grammatical errors in English authorship with the highest error rates are those involving verb consistency and verb tense, according to the literature currently in existence. So, the verb mistake detection findings can demonstrate the viability and efficiency of an automatic reading system. This research suggests a cyclic neural network-based approach for detecting grammar mistakes in English verbs. This article chose to utilize LSTM to model the labeled training corpus because it can successfully maintain the relevant context-relevant information during training. A crucial stage in automated reading is learning how to translate the textual information in English compositions into numerical values for later calculation. Most widely used tools employ the word bag paradigm, which involves encoding each word in accordance with its order in the dictionary. Although this encoding method is straightforward and uncomplicated, it also

leaves the vector without the text's sequence information and is vulnerable to dimensional disaster. As a result, the text in this study is encoded using a word embedding model, and the text data is sequentially mapped to a low-dimensional vector space. This keeps track of the text's position information and averts a dimensional catastrophe. The research that is being suggested gathers certain corpus samples and evaluates the proposed algorithm against Jouku and Bingguo.

In the research of Yeh et al. [7]he said that Mandarin is not an easy language to learn for foreigners. Children had to spend extra time learning, despite Mandarin being their mother tongue. The following problems are what make learning difficult. The term is first affected by Hieroglyphics. Hence, a character can express meanings in its own right, but a word takes on a new semantic meaning. Second, the grammar in Mandarin has a flexible rule system and unique usage. In addition, there are four types of common grammatical errors: absent, redundant, selective, and disorder. In this paper, they presented a LSTM based framework for recurrent neural networks (RNN-LSTM). It has the ability to identify the many types of writing errors made by foreign students. The word vector and part-of-speech vector provide the basis for the characteristics. The test data revealed that their method outperformed the competition in terms of detection level recall, even reaching a score of 0.9755. This is due to the fact that they offer a wider range of options for error detection.

Kumar et al. said in their study [8] a spell checker's function has expanded recently, and it is now also used to recommend potential fixes for the spelling errors that are found. Tamil is one of the oldest still-spoken languages in use today, and it has a very rich grammatical structure. For communication and information transfer to be successful, grammar is essential. Yet, the researchers find it difficult to understand the language's grammar and the traditional teaching methods. Natural language processing (NLP), which combines computers with language, offers a solution to this issue. In this study, an advanced NLP technique is utilized to identify misspelled terms in Tamil text and to offer suggestions for probable correct words as well as the likelihood that each word will appear in the corpus. The suggested model employs the Tamil-specific minimum edit distance (MED) algorithm to offer the correct alternatives for misspelled words. The incorrectly spelled word and all of its variants are placed in a distance matrix. Dynamic programming is used to identify the words that need the fewest changes to be corrected for misspellings and to recommend the most suitable replacement terms.

Spell checks are now commonplace in the majority of word processing programmes said by Abdulrahman et al.in their article [9]. They help us write more clearly in a variety of digital contexts. However, the Kurdish language, which is still thought to have fewer resources than other languages, does not yet have any well-known and reliable spell checkers. Based on instructional texts written in the Persian/Arabic script, they developed a language model for the Kurdish (Sorani) dialect. For the spell-checking algorithm, they primarily use a probabilistic approach and our language model with Stupid Backoff smoothing. In this paper, they check words and contexts for spelling problems. When a word is misspelled, the spell checker offers a list of corrections. The findings reveal an F1 score of 43.33%, 88.54

Zhao et al. said in their paper [10] that their system was submitted to the shared challenge on Chinese grammatical error diagnosis at the 2nd Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA-2) (CGED). They employ a statistical machine translation technique that has been used for various comparable tasks. Here, they investigate various translation models, such as models that are based on hierarchical phrases and on syntax, and assess corpus-augmentation. Lastly, they demonstrate variants utilizing various fusions of these variables.

From the research paper [11] of Sun et al., it shows that a growing body of research asserts that deep neural networks are fragile when faced with intentionally manufactured hostile samples. However, it is uncertain how the models would function in real-world situations when hostile natural adversarial occurrences frequently occur rather than intentionally malicious ones. This study systematically investigates the resilience of BERT, the most advanced Transformer-style NLP model, in handling noisy input, notably accidental keyboard errors. The results of much research on sentiment analysis and benchmarks for question responding show that typos in different words of a phrase do not affect each other equally. When compared to inserting, deleting, etc., mistakes in informative words cause more serious harm; mistyping is more harmful; and humans and machines have varied strengths when it comes to identifying adversary attacks.

III. Proposed methodology

A. Dataset

In the interest of our study, we developed a dataset of our own. There was a lack of resources available for the Banglish words data. We find some datasets of different types of food review comments and some sentiment analysis data. But there is a possibility of missing many words that we use frequently.So, we manually give input of different Banglish words to create the data set.Most of the words are from the daily conversations we make.Since it's possible that many of the items were duplicated more than once, we removed the duplicates from the data set as part of the pre-processing.As it was created by giving input manually, because of that the entries are given input cautiously. As a result, only removing the duplicates was enough to clear the data and make it suitable to train the model. The dataset contains a list of words.In the dataset, both correctly-spelled and misspelled words are included. Our dataset contained overall 1505 entries. The correctly spelled words are labeled as 1 and the misspelled words are labeled as 0. Which will help our models to give proper classification.

B. Architecture

In the initial phase of developing our spell checking system, we have adopted a classification-based approach to

effectively differentiate between correctly spelled and misspelled words. To accomplish our objective, we have evaluated and selected several classification models, including Multinomial Naive Bayes (MultinomialNB), Support Vector Machine (SVM), RandomForestClassifier, DecisionTreeClassifier, and Word2Vec-based classification models.To ensure the reliability of our model assessment, we have meticulously divided our dataset into two distinct subsets: a training set and a test set. This separation involves allocating 80% of the dataset to the training set and reserving the remaining 20% for the test set.For the SVM model, we applied a Vectorizer to execute the extraction of essential features from the dataset. Subsequently, by including the extracted features we trained the SVM model.In the case of the Word2Vec classification model, we initiated the process by tokenizing the words present in the dataset. This includes the breaking down of words into individual units to build a foundation for subsequent analysis. By using this tokenized characterization, we proceeded to train the Word2Vec model, which further be a factor in our ability to accurately differentiate words as either correctly spelled or misspelled. Moreover, by applying the classification models MultinomialNB, SVM, RandomForestClassifier, DecisionTreeClassifier, and Word2Vec we were able to effectively categorize words based on their spelling accuracy. After classifying the words we integrate a method to make suggestions based on user inputs. For this implementation, at first we take word input from the user and our system will match the user input with our respective dataset. If the word is spelled correctly it will give output as the word is already correctly spelt. On the other hand, for misspelled words we tried to calculate the Levenshtein distance. This distance measurement will check which word from the dataset is close enough to the wrong spelled word. Our system will suggest some words which will be sorted by calculated minimum edit distance.

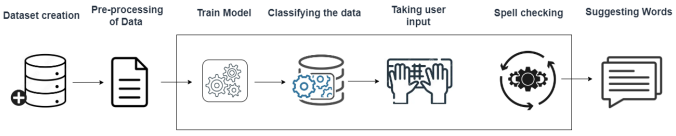

Fig. 1. The methodology pipeline

C. Spell checking

Spell checking: In our system after classification the spell checking is integrated. This feature will help to suggest words based on the wrong word given by the user. It's a step by step method which will help to execute this whole method

- Takes input; the inputted word is going to check with words available in the word dictionary.
- If matched, the result is true.
- If not matched, it prepares a list of words where each word is compared with the word dictionary using the Levenshtein edit distance algorithm.
- If similarity with any of the word dictionaries is 1.0 then the word will be marked as correct and it will return the word.
- If its similarity less than 1.0 then the word is added to the suggestion list.
- Step3 will be proceed after words preparation where the list of suggestions will be returned based on maximum similarity value.

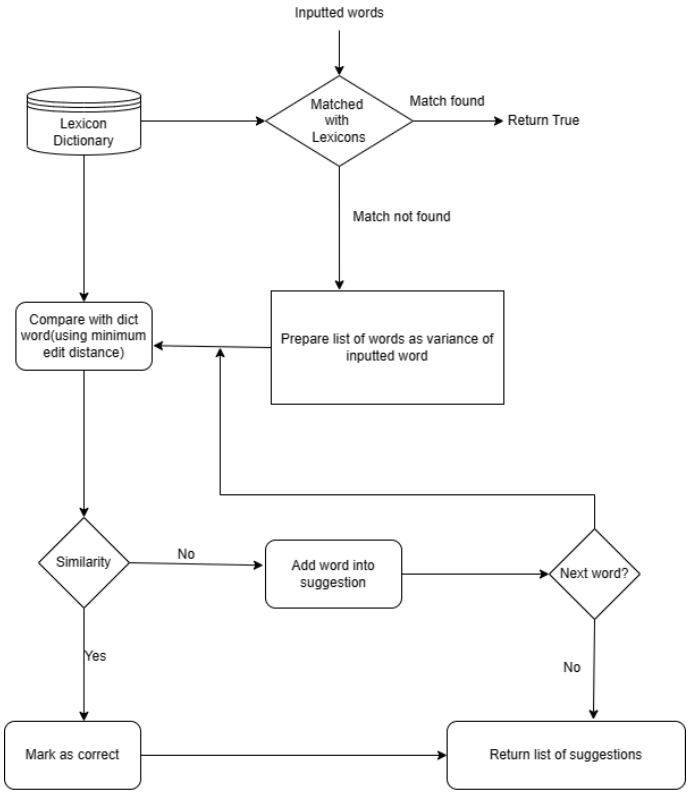

Fig. 2. Flowchart of the spell checking

IV. Experimental Results and Analysis

A. Evaluation Criteria

The precision and recall rate are the parameters which are used to measure the performance of a model. It also helps to acknowledge if its outcome is good or bad, and other issues.Both of these parameters are defined as follows. Precision can be determined as the ratio of the retrieval of relevant documents to the total count of documents retrieved. The retrieval of relatable documents demonstrates the actual count of system judges as the count of errors and the total documents which are retrieved denote system judges or evaluates based on the errors. In the context of recall, the retrieval of relevant documents herein denote as a representation the actual number of system judgements, equating them with the number of errors. The total amount of documents obtained signify system judgments as errors. The suggested system's performance is assessed using the F1 measure as the single value in the following equation based on precision and recall.

The F1 score, which is utilized as a single number in the following equation to assess the performance of our suggested method, is derived based on the precision and recall.

B. Experimental result

In this part, we contrast the various classification models that we employed for our study. On the specially created dataset, we classified the incorrectly spelt and properly spelt words using five different classification models, as previously mentioned. The dataset was used to train each of the models. The system checked each of the words from the dataset and trained the models respectively. After training the test was performed with the 20% data from the dataset and the precision, recall and F1 score being measured. Apart from this parameter the accuracy rate is also being calculated to evaluate the performance of the models. According to the findings of the experiment, we can find the result of precision, recall, F1 score and accuracy in the five models are really close to each other. There is not much result difference between the models. The main reason behind this issue is that the dataset is really small. Most of the time the result depends on the dataset. Apart from that, even in this small dataset the models give a good output. Most of the models precision rate is near 76% and F1 score is 86%. Whereas, the Multinomial Naive Bayes performed a bit better than the other models,with a precision rate of 77% and F1 score of 85%. Also in case of accuracy Multinomial Naive Bayes have a 77% rate. Which is relatively better than the other models.

TABLE I
RESULTS OF THE MODELS

	Precision	***Recall***	***F1 score***	***Accuracy***
Multinomial Naive Bayes (MultinomialNB)	77%	100%	87%	77% [a]
Support Vector Machine (SVM)	76%	99%	86%	76% [a]
Random Forest Classifier	76%	99%	86%	76% [a]
Decision Tree Classifier	76%	99%	86%	76% [a]
Word2Vec	76%	99%	86%	76% [a]

CONCLUSION

To conclude, making Banglish text more accurate using computer tools is a big step forward in how we use multilingual languages with technology. In our country, the way people mix Bengali and English words in Banglish showed us that we needed to help fix mistakes. In this paper, an NLP-based spell checker is proposed for suggesting spelling mistakes in the words of Banglish. This research is significant for improving communication and writing in mixed languages, as well as in other multilingual languages worldwide. We accurately classified the correct and misspelled data using the models. This spell checker shows us about 77% accuracy on multinomial naive bayes which was highest among all models .Also, by calculating the minimum edit distance we successfully integrated the word suggestion on wrong words in our system. In future work we plan to include a more complex model. Also, as our dataset is a bit small, we planned to increase the size of the dataset. By integrating this it will enhance the accuracy more in future.

REFERENCES

[1] M. Mitreska, K. Mishev, and M. Simjanoska, "Nlp-based typo correction model for croatian language," in *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, 2022, pp. 942–947.

[2] J.-F. Yeh, G.-H. Wu, S.-Y. Wang, C.-K. Yeh, and Y.-Y. Wang, "Statistical language models for spelling error detection with web search new word acquisition," in *2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 2021, pp. 1–2.

[3] M. Jahan, A. Sarker, S. Tanchangya, and M. Yousuf, *Bangla Real-Word Error Detection and Correction Using Bidirectional LSTM and Bigram Hybrid Model*, 01 2021, pp. 3–13.

[4] H. Patel, B. Patel, and K. Lad, "Jodani: A spell checking and suggesting tool for gujarati language," in *2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, 2021, pp. 94–99.

[5] L. Xu and J. Chamberlain, "Cipher: A prototype game-with-a-purpose for detecting errors in text," in *Workshop on Games and Natural Language Processing*. Marseille, France: European Language Resources Association, May 2020, pp. 17–25. [Online]. Available: https://aclanthology.org/2020.gamnlp-1.3

[6] Z. He, "English grammar error detection using recurrent neural networks," *Scientific Programming*, vol. 2021, pp. 1–8, 2021.

[7] J.-F. Yeh, T.-W. Hsu, and C.-K. Yeh, "Grammatical error detection based on machine learning for mandarin as second language learning," in *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, 2016, pp. 140–147.

[8] P. Kumar, A. Kannan, and N. Goel, "Design and implementation of nlp-based spell checker for the tamil language," in *Presented at 1st International Electronic Conference on Applied Sciences*, vol. 10, 2020, p. 30.

[9] R. Abdulrahman and H. Hassani, "A language model for spell checking of educational texts in kurdish (sorani)," in *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, 2022, pp. 189–198.

[10] Y. Zhao, M. Komachi, and H. Ishikawa, "Improving chinese grammatical error correction with corpus augmentation and hierarchical phrase-based statistical machine translation," in *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*, 2015, pp. 111–116.

[11] L. Sun, K. Hashimoto, W. Yin, A. Asai, J. Li, P. Yu, and C. Xiong, "Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert," *arXiv preprint arXiv:2003.04985*, 2020.