A Deep Learning–Based Framework for Dataset Creation and Sentiment Classification of English–Bengali Code-Mixed Texts

Dalia Barua
School of Computer Applications, Faculty of Technology and Science, Lovely Professional University, Jalandhar-Delhi GT Road, Phagwara, Punjab, India
bdalia24@gmail.com (corresponding author)

Tarandeep Singh Walia
School of Computer Applications, Faculty of Technology and Science, Lovely Professional University, Jalandhar-Delhi GT Road, Phagwara, Punjab, India
taran_walia2k@yahoo.com

Received: 12 October 2025 | Revised: 1 November 2025 and 14 November 2025 | Accepted: 17 November 2025

 DOI: https://doi.org/10.48084/etasr.15475

ABSTRACT

Datasets form the foundation for most Natural Language Processing (NLP) operations, such as sentiment analysis, summarization, and translation. Lack of big and diverse datasets poses a critical problem for low-resource languages like Bengali. The problem is aggravated in code-mixed environments where users tend to intermix Bengali and English on web platforms like social media, video comments, and product reviews on e-commerce websites. It is very difficult to conduct sentiment analysis of such Bengali–English code-mixed reviews, as there are no high-quality, large-scale datasets available. To address this challenge, the present research introduces a novel synthetic English–Bengali code-mixed product review dataset, specifically curated for sentiment analysis in the Bangladeshi e-commerce domain. The dataset was generated through a new deep learning–driven pipeline integrating part-of-speech tagging, translation, transliteration, and a two-tier ensemble-based sentiment annotation framework to ensure linguistic diversity and contextual realism in code-mixed expressions. The dataset, En–Bn–Code–Mixed–Two–Class–Sentiment–Dataset–, is available on the Hugging Face repository, with a total of 100,000 reviews, balanced for both positive and negative sentiments. The quality and reliability of the generated dataset were validated through quantitative, linguistic, and statistical analyses, including translation, transliteration, and sentiment analysis evaluation metrics. The proposed two-tier ensemble sentiment annotation approach, implemented through deep learning models, meta-learning models, and majority voting, achieved 93.00% accuracy, 92.16% precision, 94.00% recall, and 93.07% F1-score. This work not only provides a useful resource for code-mixed Bengali–English NLP but also establishes a scalable methodology for low-resource and multilingual text processing, opening up further possibilities of more comprehensive and inclusive studies on sentiment analysis.

Keywords-code-mixed; sentiment analysis; low-resource; dataset; deep-learning; ensemble; majority voting

I. INTRODUCTION

NLP focuses on computational analysis and generation of human language. It has found extensive applications in tasks such as machine translation, sentiment analysis, information extraction, and text summarization. Good quality datasets, powerful models, and proper evaluation metrics are a combined driving force behind state-of-the-art language-based artificial intelligence and are important to NLP performance [1-4].

Sentiment analysis datasets are collections of text data with sentiment tags—positive, negative, or neutral—and serve as fundamental resources for training and testing machine learning models in sentiment classification tasks. Regarding sentiment analysis in online business, these datasets are highly valuable for automating the detection of customer opinions and feedback, enabling businesses to assess customer satisfaction, improve products, and make data-driven decisions. While English-language sentiment resources are abundant, languages with fewer resources, such as Bengali, do not have enough material, especially on English–Bengali code-mixed text, which is so popular on social media and e-commerce websites [5]. Addressing these challenges for datasets in low-resource code-mixed areas, this research aims to develop an English (En), Bengali (Bn) code-mixed product review dataset with sentiment annotation. Creating this resource is crucial to improving sentiment analysis models in minority languages to

deliver more accurate and complete sentiment detection in multilingual and code-mixed environments.

Traditional monolingual data neglect linguistic variation and, therefore, limit the performance of single-language-trained models. To address this challenge, code-mixed data have been proposed that comprise multiple languages to enable models to recognize language-switching behavior, lexical loans, and colloquial turns of expression typical of user-generated content, such as product reviews [6]. The development of a large-scale, annotated code-mixed corpus of English–Bengali code-mixed can only enable researchers to improve sentiment classification algorithms, which could be applied in real-world scenarios like product recommendation systems and customer feedback analysis [7]. These algorithms may provide a balanced coverage of positive and negative emotion while attempting to have similarity in certain characteristics to real language-like lexical variation, transliteration mismatch, and natural code-mixing patterns. To fulfill this requirement this study built a large-scale synthetic English-Bengali code-mixed e-commerce sentiment analysis dataset titled En–Bn–Code–Mixed–Two–Class–Sentiment–Dataset that captured natural language blending and domain-specific vocabulary. The dataset was constructed using a novel approach, which includes NLP tasks-POS tagging, translation, transliteration, and sentiment label annotation.

Research on multilingual processing and sentiment analysis has come a long way on a wide variety of subjects, ranging from machine translation to emotion classification. A statistical and rule-based hybrid approach was evaluated in [8] to solve structural variation in languages between English-Bengali. The model demonstrated good translation performance for morphologically rich languages like Bengali. Similarly, authors in [9] presented a new steganography method of Bengali text transliteration for secure communication through encoded messages. Their study laid the ground for the potential use of transliteration in bridging language and technology, as well as for the capacity of language technologies to evolve.

For sentiment analysis, authors in [10] brought forward context sensitivity and accuracy, showing that sentiment analysis is not necessarily emotional detection in actual real-world applications in marketing and business intelligence. Subsequently, authors in [11] examined zero-shot learning for Indonesian large language models in sentiment analysis and concluded that transformers can generalize cross-low-resource language sentiment tasks without having to be trained specifically. Authors in [12] had previously established that the machine learning classifiers Naive Bayes (NB) and Support Vector Machine (SVM) worked well for sentiment analysis from Twitter material.

Authors in [13] experimented with sentiment analysis on online shopping sites, discovered patches of multilingual weakness, and proposed a deep learning-based approach for accuracy enhancement. Authors in [14] introduced GMM-augmented N-gram LSTM, which is a combination of Gaussian Mixture Models with N-gram-based LSTM to identify fine-grained patterns in language and improve the precision of sentiment classification. Authors in [15] studied the usage of SVM and NB classifiers in an ensemble, illustrating how combining these traditional machine learning models might be more effective than using a single classifier. Authors in [16] developed an ensemble hybrid deep model for combining different architectures of different neural networks to counter the detection of faint emotional expressions. Similarly, authors in [17] also built an ensemble classification model for Twitter sentiment analysis that effectively addressed the issues with short, informal, and noisy textual data. Authors in [18] continued in this direction further with the use of SETAR, a stacking ensemble learning method making use of RoBERTa and hybrid feature representation to improve the performance and efficiency of Thai sentiment analysis.

Since ensemble learning architectures have become prevalent, several researchers have given more attention to the majority voting-based techniques for achieving higher accuracy and interpretable sentiment classification outputs. Authors in [19] proposed a majority voting technique for product review sentiment classification based on the outcome of several classifiers to make their outputs more stable and homogeneous. Authors in [20] applied a majority voting technique to disaster tweet classification and referenced it for application in real-time social media for improving decision quality. Finally, authors in [21] presented the Synthetic Minority Over-sampling Technique (SMOTE) to modify imbalanced data and improve the performance of machine learning models for sentiment analysis. Their evaluation showed that SMOTE significantly improved data quality, which enhanced the machine learning and majority voting-based sentiment analysis approach.

The key contributions of the present work are:

- A novel synthetic English–Bengali code-mixed sentiment dataset was introduced; rigorous quantitative, linguistic, and qualitative analyses were also conducted to ensure that it is consistent, reliable, correct, and effective.
- An innovative approach was presented, incorporating deep learning-based POS tagging, translation, and transliteration for generating code-mixed product reviews.
- A 20-line batch rule was implemented, which correlates with the translation and transliteration, as well as the natural mixing of the reviews.
- A new ensemble-based approach to sentiment annotation was introduced that provides refined binary sentiment labels, combining SiEBERT, DistilBERT, stacked ML models, and the majority voting technique.

The present research, combined with previous studies related to translation, transliteration, and sentiment analysis, significantly enhances the progression of synthetic code-mixed sentiment dataset development. The study prioritizes linguistic purity, scalability, and consistency to ensure that the dataset is suitable for code-mixing, multilingual, and low-resource workloads.

II. DATASET DESCRIPTION

A. Quantitative Data Analysis

The En–Bn–Code–Mixed–Two–Class–Sentiment–Dataset is a synthetic dataset intended to replace the low-resource

Bangladesh e-commerce sentiment dataset. It consists of En, Bn, English-Bengali (En-Bn) code-mixed, and English Roman Bengali (En-Rm Bn) code-mixed reviews, while it is representative of the multilinguality prevalent on South Asian e-commerce platforms.

The code-mixed product review dataset was created using a subset of 100,000 reviews from the Amazon product review dataset and spans reviews of numerous food products, including sweeteners and creamers, hot chocolate, cooking oil, as well as candy and sweets. It provides excellent sentiment analysis coverage and is efficient for code-mixed studies as well as multilingual NLP model training. The code-mixed dataset includes four columns, as presented in Table I.

The dataset is ready for computational analysis and modeling as well, for each column is in order and complete, and there are no missing values. The dataset includes a total of 12,560 unique products. As shown in Table II, the number of reviews for each product can vary significantly.

TABLE I. COLUMN DISTRIBUTION OF EN–BN–CODE–MIXED–TWO–CLASS–SENTIMENT–DATASET

Id	Product Id	Code-mixed-text	Sentiment
Unique identifier for each review	Identifier for the product being reviewed	The review text in English, Bengali, or code-mixed form	Sentiment label for the review (positive/negative)

TABLE II. REVIEW DISTRIBUTION BASED ON THE UNIQUE PRODUCT

Id	Product Id	Review count
2	B00002N8SM	38
3	B00002NCJC	2
4	B00002Z754	2
5	B00005V3DC	3
6	B00006IDJO	1
7	B00006IDJU	2
8	B000084DVR	2
9	B000084E1U	1
10	B000084E6V	177

B. *Linguistic Analysis*

The dataset consists of four distinct text categories—15,000 lines for English, 15,000 lines for Bengali, 35,000 lines for English–Bengali, and 35,000 lines for English–Roman Bengali—and comprises code-mixed as well as monolingual texts. As depicted in Table III, the prevalence of code-mixing in the dataset is substantial, which accounts for approximately 70% of all the reviews being code-mixed. This ratio is an estimate of real user behavior in multilingual environments such as Bangladesh, where users tend to mix English with their native language when communicating electronically. The distribution of the general frequency of words in the dataset into three categories: English, Bengali, and Roman Bengali can be observed from a scatter plot graph in Figure 1. Among these three, English words are predominant (6,870,500) and form the largest cluster of words. The remaining words are divided into Bengali and Roman Bengali. The balanced word distribution supports cross-lingual learning while preserving multilingual diversity and consistency.

TABLE III. TEXT CATEGORY DISTRIBUTION OF EN-BN CODE-MIXED TWO-CLASS SENTIMENT DATASET

Text category	Percentage
English	15%
Bengali	15%
English-Bengali	35%
English-Roman Bengali	35%

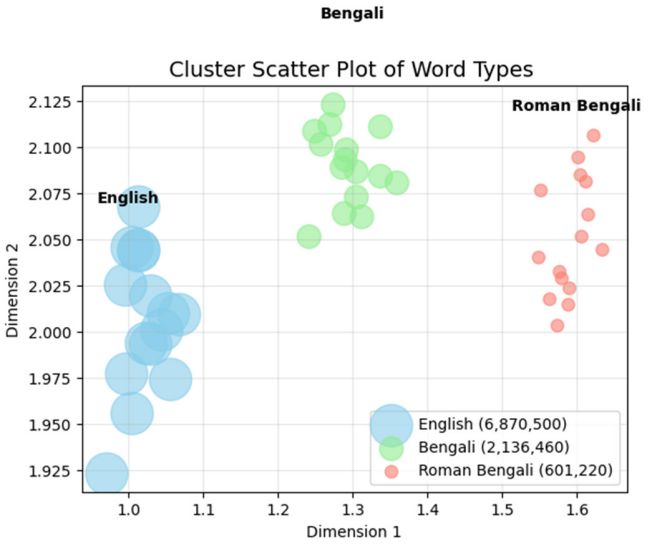

Fig. 1. Scatter plot for the word frequency distribution.

C. *Code-Mixed Sentiment Distribution*

Sentiment distribution presents the language characteristics of the En–Bn–Code–Mixed–Two–Class–Sentiment–Dataset in terms of examples of sentiment annotation on different kinds of texts, as placed in the dataset. Table IV presents a sample distribution of positive and negative sentiments in monolingual and code-mixed comments, illustrating how sentiment is expressed differently in English, Bengali, English–Bengali, and English–Roman Bengali texts.

TABLE IV. EXAMPLE SENTIMENT DISTRIBUTION/ANNOTATION BY TEXT TYPE

Text type	Sentiment	Example text
English	Positive	"This chocolate tastes amazing and melts perfectly!"
Bengali	Negative	"এই তেলের গন্ধ একদম ভালো নয়।"
English–Bengali	Positive	"Candy টা delicious, আমি অনেক like করেছি!"
English–Roman Bengali	Negative	"Ei sweet ta amar pochondo hoyni, khub sugary chhilo."

The decomposition provides a difference in sentiment expression across languages and categories of code-mixing. For example, English reviews comprise straightforward lexical emotion expressions, whereas Bengali and code-mixed reviews have native words, transliteration, and lexical mixing that affect the sentiment expression. The decomposition allows for the comprehension of linguistic patterns more effectively, which are essential for the creation of effective multilingual sentiment classification algorithms.

Also, sentiment polarity distribution, as portrayed in Figure 2, illustrates that positive reviews constitute 79.3% and negative reviews constitute 20.7%, indicating that most users gave positive feedback for the product reviews. Overall, the dataset is linguistically balanced, diverse, and reflective of real-world code-mixing, serving as a reliable benchmark for multilingual sentiment analysis.

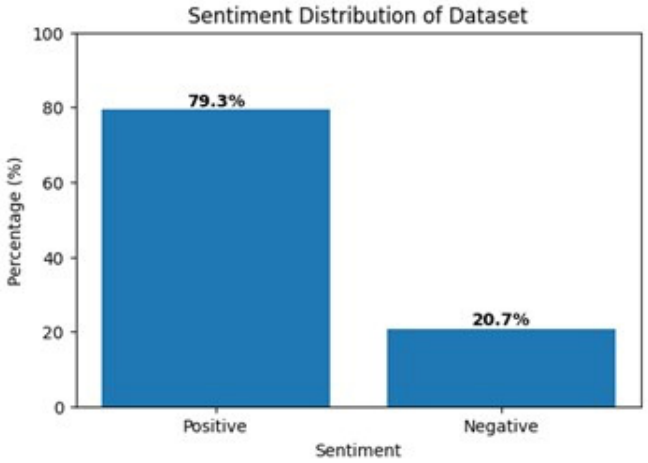

Fig. 2. Sentiment polarity distribution of En–Bn–Code–Mixed–Two–Class–Sentiment–Dataset.

D. Source Datasets

The work develops and evaluates an English–Bengali code-mixed sentiment corpus using diverse and well-defined datasets. The primary data source is the Amazon Product Reviews dataset [22, 23], available at Kaggle, which contains 500,000 English product reviews tagged with ratings. The Dakshina dataset [24] was used for transliteration to maintain linguistic parity. Additionally, the Amazon Review Polarity dataset [25] was employed in this study for sentiment analysis evaluation, applied to Amazon English product reviews.

III. PROPOSED METHODOLOGY

The proposed methodology is divided into three steps to build a solid framework for code-mixed sentiment analysis. It begins with data cleaning as the first step. The second step is dedicated to a rule-based, systematic approach for generating English–Bengali code-mixed reviews. The third step focuses on sentiment annotation of the created corpus to support downstream sentiment classification tasks. These three stages generate linguistically diverse data and provide the sentiment annotation to enable the training and assessment of the model, as illustrated in Figure 3.

A. Data Cleaning

The Amazon raw product review English dataset was required to be preprocessed first for further transformation. To standardize the text and remove noise, the dataset was subjected to:

- Removing extra whitespaces
- Lower casing the data
- Handling repeated characters or emojis.

This preprocessing ensures that the reviews are clean and consistent, allowing for accurate tokenization and language identification in code-mixed scenarios.

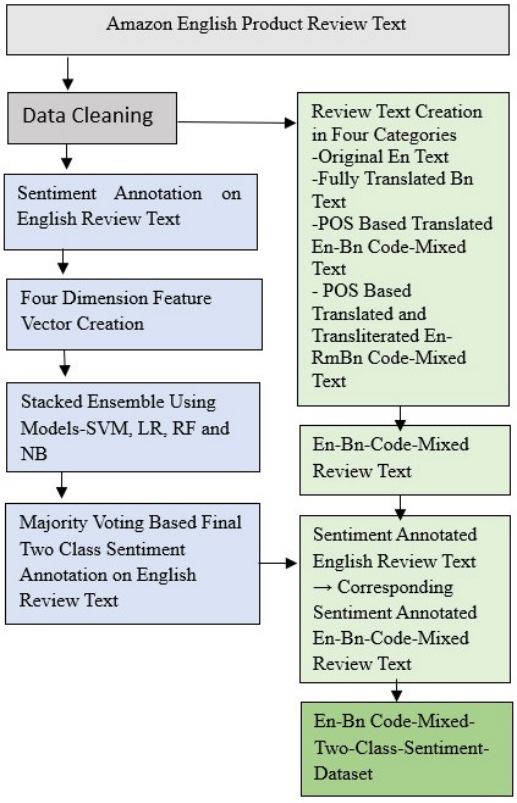

Fig. 3. Proposed approach for creating En–Bn–Code–Mixed–Two–Class–Sentiment–Dataset.

B. Code-Mixed Data Creation

The dataset generation provides balanced coverage of various types of text with linguistic heterogeneity and less redundancy. The operation is split into consecutive 20-line batches, performed multiple times (a maximum of 5000 iterations) in order to make the dataset substantial. Each batch tries to provide proportional coverage of En, Bn, POS-based En-Bn, and POS-based English–Roman Bengali (En-Rm Bn). The derived text distribution over each 20-line batch is:

- The first type of text in each 20-line set consists of distinct English sentences from the source dataset verbatim. Three reviews are selected in each set (denoted as $En = 3$) to act as embodiments of English content. This preserves the syntactic and semantic correctness of the source reviews and also provides a baseline against which code-mixed forms can be compared.
- The second is end-to-end Bengali translations of the English reviews. Three reviews simultaneously ($Bn = 3$) are translated using Google Neural Machine Translation (GNMT), which benefits from a system having a neural network-based system in place for high-quality, context-

aware translation. The Bengali sentences, therefore, maintain the meaning and nuances of the original English text and mix well in the multilingual dataset.

- The third is a POS-based English–Bengali code-mixed type and contains seven sentences in a batch (En-Bn = 7). Selective translation is utilized in words tagged as adjectives, adverbs, conjunctions, nouns, and verbs. POS tagging is performed by employing a pre-trained BERT-based POS tagger and yields strong and accurate syntactic annotation. Words of the selected POS categories are translated from English into Bengali through Google NMT, and other words are left in English. Such an approach produces naturalistic bilingual code-mixing highly similar to patterns occurring in real user-generated data.
- The fourth class is POS-based English–Roman Bengali text with seven sentences per batch (En-Rm Bn = 7). This is done by transliterating the Bengali items of the previous POS-based code-mixed sentences into Roman script. Transliteration is carried out using a specially fine-tuned M2M100 model, trained specifically for Bengali-to-Roman Bengali translation, with proper handling of complex consonant clusters (jukto borno) and vowel diacritics. M2M100 was used because of its non-English-oriented, multilingual sequence-to-sequence model that supports phonological and contextual consistency in transliteration. This produces natural Romanized code-mixed sentences that look very similar to typing styles commonly seen in digital channels.

C. Sentiment Annotation

Manual and deep learning-based approaches were carried out for sentiment annotation over the code-mixed dataset. Expert annotators manually annotated utilizing pre-defined guidelines and cross-checked for consistency, which were then used as ground truth for assessment. Meanwhile, a new pre-trained transformer-based stacking ensemble, including SiEBERT and DistilBERT, was used in automatically generating sentiment labels, with corresponding performance being assessed utilizing the Amazon Polarity Dataset.

The SiEBERT (RoBERTa-based) model is a pretrained model trained on a large English sentiment corpus, which enables more complex sentiment polarity understanding. For output, it generates probability values for positive-negative binary classes. DistilBERT is the faster, lighter version of BERT with all of its language comprehension included. BERT works very effectively for sentiment analysis, and DistilBERT offers a less heavy model with reduced computational cost. DistilBERT also has a simpler but complementary architecture than RoBERTa. The former generates a sentiment class probability distribution over each review in the dataset, which can be output along with other models for ensembled prediction.

For the ensemble model, probability outputs from SiEBERT and DistilBERT were used to construct a new feature set by concatenation. Each transformer produces a two-dimensional probability vector since there are two sentiment classes (positive and negative). Thus, for each review, a 4-dimensional feature vector is created:

$$[\text{negSiEBERT}, \text{posSiEBERT}, \text{negDistilBERT}, \text{posDistilBERT}]$$

With this newly created feature space, some baseline classifiers—SVM, Logistic Regression (LR), Random Forest (RF), and NB were individually trained as meta-models in the stacking ensemble. All classifiers learned how to map the probability, based on features, to the true sentiment labels. Among these, SVM demonstrated strong performance through its capacity to form optimal decision boundaries in small probability spaces and strong accuracy in binary sentiment classification.

To further enhance the resilience of the ensemble, voting was also employed to gather predictions from all the meta-classifiers (SVM, LR, RF, NB). Simultaneously, each classifier voted for a single sentiment label alone, and the final prediction was made by examining the class that received the most votes. The voting contributes to enhancing decision-making by reducing the effect of individual-model errors and biases. The algorithm for code-mixed data creation and sentiment tagging pipeline is:

```
Algorithm 1: code-mixed data creation
and sentiment tagging pipeline

Require: Original English Review Dataset
𝒟 = {r_1, r_2, ..., r_N}
Ensure: Code-mixed dataset 𝒟′with
sentiment labels ŷ_1, ŷ_2, ..., ŷ_N
Stage 1: Code-Mixed Data Creation
2.  Split 𝒟into batches of 20 reviews.
3.  for each batch do
4.    Assign batch-wise distribution:
       En = 3 (Original English)
       Bn = 3 (Fully Bengali via Google
NMT)
       En − Bn = 7 (POS-based English-
Bengali code-mixed)
       En − Rm Bn = 7 (POS-based English-
Roman Bengali transliteration)
5.    for each review r_i in batch do
6.       if review assigned to En:
7.          r_i′ ← r_i  ▷ Original English
.       else if review assigned to Bn:
9.          r_i′ ← T_B(r_i)  ▷ Google NMT full
translation
10.       else if review assigned to
POS-based En-Bn:
11.          Select tokens w_i via BERT-
based POS tagger (JJ, RB, CC, NN, VB).
12.          r_i′ ← T_POS(r_i)  ▷ Google NMT
for selected tokens
13.       else if review assigned to
POS-based En-Rm Bn:
14.          r_i′ ← 𝒯_TR(T_POS(r_i))  ▷ Fine-tuned
M2M100 transliteration after POS-based
translation
15.       end if
```

16. end for
17. end for
18. Construct final code-mixed corpus:
$\mathcal{D}' = \{r_1', r_2', \ldots, r_N'\}$

Stage 2: Transformer-Based Feature Extraction

20. for each review r_i' in $\mathcal{D}'$ do
21. Compute transformer-based sentiment probabilities:
$P_{\text{SiEBERT}}(r_i') = [P_{pos}^{SiEBERT}, P_{neg}^{SiEBERT}]$
$P_{\text{DistilBERT}}(r_i') = [P_{pos}^{DistilBERT}, P_{neg}^{DistilBERT}]$
22. Concatenate feature vectors:
$X_i = [P_{pos}^{SiEBERT}, P_{neg}^{SiEBERT}, P_{pos}^{DistilBERT}, P_{neg}^{DistilBERT}]$
23. end for

Stage 3: Stacking Ensemble

25. Train meta-classifiers $\{f_{SVM}, f_{LR}, f_{RF}, f_{NB}\}$ on (X_i, y_i).
26. Each classifier predicts:
$\hat{y}_k = f_k(X_i), k \in \{SVM, LR, RF, NB\}$

Stage 4: Ensemble Fusion (Majority Voting)

28. for each sample i do
29. $\hat{y}_i = \text{mode}(\hat{y}_i^{SVM}, \hat{y}_i^{LR}, \hat{y}_i^{RF}, \hat{y}_i^{NB})$
30. end for

The workflow algorithm effectively generates Text-CM English–Bengali code-mixed sentiment datasets through a systematic, multi-stage process.

D. Experimental Setup

After collecting 100,000 Amazon product reviews from a source dataset of 500,000 reviews, a small batch size facilitated efficient processing and consistent quality during the development of the dataset. The 100,000 reviews were split into 20 subsets of 5,000 reviews, which were processed in 20-line batches, and this was done over 250 iterations. This approach allowed the systematic generation of code-mixed text and the corresponding sentiment annotations, which formed the two components of the experiment. In the present study, GNMT was installed via a Google translation-based Python library. The BERT-based pretrained POS tagger and the RoBERTa-based SiEBERT and DistilBERT models were employed through Hugging Face transformer libraries. For the Bengali–Roman Bengali transliteration, a fine-tuned, M2M100 model was trained through the Dakshina Bengali–Roman Bengali parallel corpus by using Fairseq and Transformers with an 80:20 split, optimized by Adam. All the utilized transformer-based models, both zero-shot and fine-tuned, including meta-learning ensemble models, were carried out in the PyTorch–scikit-learn GPU environment for optimal performance and reproducibility.

IV. RESULTS AND DISCUSSION

A. Evaluation Metrics for Code-Mixed Data Creation

The pre-trained BERT-based approach performs well on the code-mixed English-Bengali corpus in the task of POS tagging. The performance of the model, as observed by token-level accuracy and F1-score, is very close to 94% and 93%, respectively. This shows the strength of the model in identifying the parts of speech in code-mixed sentences. High accuracy in POS tagging helps to support translation and transliteration tasks. In English to Bengali translation, Google Translator performed reasonably well, as presented in Table V. This is supported by its BLEU and ROUGE scores. BLEU-1 through BLEU-4 range from 39.24% to 63.18%, and ROUGE-L is 66.54%, showing that the model retains primitive lexical matches and some phrase structures but tends to drop longer n-gram strings or subtle syntactic structures.

The Loss versus Epoch graph, as displayed in Figure 4, of the M2M100 fine-tuning process shows loss initially dropping from 1.21 to 0.25 and gradually reaching 0.09. Based on this, it could be concluded that the model translated steadily, converged, and improved over time in transliteration. The Tuned M2M100 model, as presented in Table V, utilized for transliteration, shows much improvement in generating accurate Romanized Bengali outputs. The model obtains a BLEU-1 of 90.73%, a BLEU-4 of 70.56%, and a ROUGE-L of 91.00%, demonstrating its ability to generate coherent sequences with high n-gram overlap and fluency. The strong performance across BLEU metrics ensures that the transliteration task achieves local word-level accuracy as well as sequential dependencies over longer lengths. In general, these findings show the strength of combining a pre-trained model for POS tagging, neural machine translation, and a multilingual sequence-to-sequence model, fine-tuned for transliteration to support richer code-mixed, cross-script text processing.

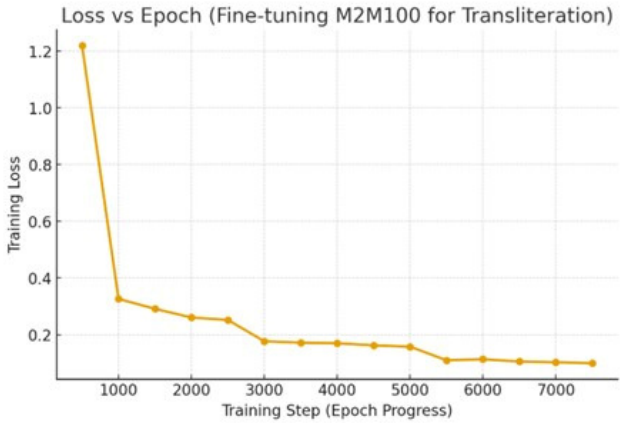

Fig. 4. Loss versus Epoch graph for the fine-tuning process of the M2M100-based transliteration.

B. Lexical Normalization for Code-Mixed Dataset

For acquiring lexical propriety and orthographic coherence on a large-scale English‑Bengali code-mixed corpus (≈100,000 reviews), a two-stage Levenshtein-distance-based spell normalization process was employed. In the first stage, normalization was applied after GNMT produced Bengali text from English sentences. Minor normalization was unnecessary, as outputs from GNMT were morphologically and

orthographically accurate with good-quality translation and little noise.

The transliteration and translation quality were qualitatively confirmed by BLEU and ROUGE-L scores, as presented in Table VI. These results directly validate the minimal amount of spell normalization work that had to be done−both GNMT and M2M100 produced high-quality, linguistically consistent outputs with less than 2% of the total tokens requiring correction in the whole 100k dataset.

As observed in Table VI, the severity of correction was also reduced further after transliteration, suggesting that both GNMT and M2M100 outputs were effectively error-free on the lexical level. Thus, the normalization process behaved like the final polish layer, attaining spelling uniformity and eliminating minor irregularities across the 100k-sample dataset.

C. Evaluation Metrics for Code-Mixed Data Quality

To measure the lexical and structural fidelity of the generated bilingual corpus, BLEU scores were computed for three sequence sets: English–Bengali code-mixed, English–Roman Bengali code-mixed, and fully Bengali translations. The results are explained in Table VII.

TABLE V. EVALUATION METRICS FOR EN TO BN GOOGLE TRANSLATION AND FINE-TUNED M2M100 BASED TRANSLITERATION (BN TO RM BN)

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
Google translator (En to Bn)	63.18	53.31	45.70	39.24	66.54
Fine-tuned M2M100 (transliteration)	90.73	85.40	76.89	70.56	91.00

TABLE VI. SPELL NORMALIZATION STAGES AND CORRECTION COUNTS (ON 100,000 REVIEWS)

Stage no.	Normalization stage description	Target token type(s)	Unique words corrected	Percentage of total tokens corrected	Relative correction intensity
1	After Google translation (GNMT)	Bengali	12,450	1.24%	(Minimal)
2	After transliteration (Bn to Roman)	Transliterated Bengali tokens	5,980	0.60%	(Very minimal)

TABLE VII. EVALUATION METRICS FOR TRANSLATED SEQUENCES

Sequence type	BLEU-1	BLEU-2	BLEU-3	BLEU-4
English–Bengali code-mixed	93.75	90.36	86.89	82.93
English–Roman Bengali code-mixed	89.55	86.35	82.82	79.35
Fully Bengali	40.07	31.12	22.50	15.19

For English–Bengali code-mixed strings, BLEU-1 is 93.75%, BLEU-2 is 90.36%, BLEU-3 is 86.89%, and BLEU-4 is 82.93%, indicating very high lexical overlap and great syntactic faithfulness. Similarly, the English–Roman Bengali code-mixed strings received BLEU-1: 89.55%, BLEU-2: 86.35%, BLEU-3: 82.82%, and BLEU-4: 79.35%, indicating that both sequences create a well code-mixed sequence with minimal loss of quality.

In contrast, BLEUs for the complete Bengali test sequences were lower, at 40.07% for BLEU-1, 31.12% for BLEU-2, 22.50% for BLEU-3, and 15.19% for BLEU-4. This is expected since the morphological complexity of Bengali is higher, and no English lexical anchors were present, and hence naturally, n-gram overlap with the reference sequences would be lower.

Overall, these measurement tools confirm that the code-mixed data, whether written in Bengali script or Romanization, are highly lexically similar and syntactically fluent, substantiating the efficacy of the bilingual blending approach.

D. Semantic Similarity Evaluation

Finally, to verify semantic consistency, a Sentence Transformer-based cosine similarity was quantified between the original English sentences and their code-mixed generated versions. The average cosine similarity of 0.7811 substantiates that the code-mixed dataset retained the semantic meaning of the original English reviews with the addition of bilingual variation successfully.

E. Evaluation of Sentiment Annotation

As shown in Table VIII, the sentiment scores confirm that transformer-based and ensemble techniques perform well on the English test set. If run separately, the performance of the RoBERTa (SiEBERT) model is excellent, with an accuracy of 93% and an F1-score of 93.07%, which validates its proficiency to detect subtle sentiment cues in code-mixed sentences. The performance of the DistilBERT model is moderate with an accuracy of 86.5% and an F1-score of 85.86%, which shows that while precision is high, 90.11%, it misses some cases of sentiment due to low recall.

Various ensemble techniques were attempted to achieve robustness and reliability. Stacked ensembles employ SVM, LR, and NB classifiers. The majority voting ensemble achieved a performance close to that of the standalone RoBERTa model, 93% accuracy, and an F1-score of 93.07%, demonstrating that averaging many models achieves high performance with the additional advantage of stability across various test cases. RF-based ensemble lost slightly in accuracy (92%) and F1-score (92.08%), but otherwise performed similarly.

The proposed two-tier ensemble sentiment analysis model performed within the same range as the RoBERTa-based SiEBERT (accuracy 93%, F1-score 93.07%), with greater stability and reliability in handling noisy, code-mixed text. Unlike a single model, the ensemble approach is designed to mitigate overfitting and improve robustness by integrating multiple classifiers to counteract biases in a more balanced way. Given that the system operates under a principle of

majority voting, it guarantees consistency in performance while surpassing minimal expectations applicable to sentiment analysis, particularly in analyzing e-commerce reviews and social media, which require a system that can handle complex and noisy linguistic data.

TABLE VIII. PERFORMANCE EVALUATION FOR SENTIMENT ANNOTATION

Model /Metric	Accuracy	Precision	Recall	F1-score
RoBERTa (SiEBERT)	93.00	92.16	94.00	93.07
DistilBERT	86.50	90.11	82.00	85.86
Ensemble (SVM)	93.00	92.16	94.00	93.07
Ensemble (LR)	93.00	92.16	94.00	93.07
Ensemble (RF)	92.00	91.18	93.00	92.08
Ensemble (NB)	93.00	92.16	94.00	93.07
Majority voting ensemble	93.00	92.16	94.00	93.07

F. Validation and Comparison of Sentiment Annotation

The proposed sentiment annotation method is evaluated in comparison to [26]. In [26], sentiment labeling was created via emotion-dominated manual annotation based on Shaver's hierarchical emotion model, where the five core emotions: Happiness, Love, Sadness, Fear, and Anger were clustered into positive and negative. Consensus among the four annotators was measured through Cohen's Kappa, which resulted in an average $\kappa \approx 0.62$. While this method is emotion-oriented, rooted psychologically, and sentiment identification is grounded in human emotions, it is still limited by human subjectivity and fundamentally lacks the potential to scale to large multilingual datasets.

On the other hand, in this research, manual and deep learning-based sentiment annotation is used to label the English–Bengali code-mixed text. The proposed method reflects strong predictive consistency with an accuracy of 93.00%, a precision of 92.16%, a recall of 94.00%, and an F1-score of 93.07%. To assess reliability in the sentiment labeling, Cohen's Kappa coefficients (κ) were evaluated for the model alongside the two human annotators. The validation results are presented in Table IX, which shows near-perfect reliability for all comparison pairs.

TABLE IX. KAPPA AGREEMENT VALIDATION FOR SENTIMENT ANNOTATION

Comparison pair	Agreement (%)	Interpretation
Annotator 1 versus Annotator 2	85.00%	Almost perfect agreement
Proposed Model versus Annotator 1	81.00%	Almost perfect agreement
Proposed Model versus Annotator 2	84.00%	Almost perfect agreement
Average agreement	82.50%	Strong agreement

The inter-rater agreement scores Kappa (κ) of greater than 0.80 indicate that the sentiment predictions generated from the ensemble model are in perfect agreement with the human annotators. This illustrates that the model's sentiment predictions are consistent with human expectations, reproducible, and reflect human polarity. The deep learning–based stacking ensemble provides automated, linguistically flexible sentiment annotation for English–Bengali code-mixed data, reducing subjective bias compared to prior emotion-based sentiment annotation methods.

V. CONCLUSION

The current study attempts to fill the gap concerning the unavailability of reliable multilingual sentiment resources by generating a new synthetic English–Bengali code-mixed product review sentiment dataset. To create the dataset, a combination of a rule-based system and a transformer-based approach was used, along with selective translation, transliteration, and full translation techniques to mirror real bilingual user reviews. To construct a realistic and context-rich dataset for machine learning–based sentiment analysis, various styles of code-mixing and the different syntactic and semantic patterns of code-mixing across languages were preserved.

To create a new ensemble-based sentiment annotation technique and improved binary sentiment labels, the new approach integrates transformer models with stacked machine learning classifiers via majority voting. To ensure the correctness and effectiveness of the proposed dataset, a quantitative and qualitative analysis was performed on the created English-Bengali code-mixed text. Additionally, the statistical Cohen's Kappa agreement interpretation analysis was also employed to provide evidence of the consistency and reliability of the sentiment annotation approach and comparison to previous related studies. The comparison showed that the proposed framework gained improved linguistic diversity, annotation accuracy, and classification performance in the text. The results of this study not only bridge the English–Bengali code-mixed resource gap but also establish a reproducible and scalable framework for dataset generation and sentiment analysis performance in low-resource multilingual settings. The methodologies proposed, along with evaluation strategies, enable future research related to multilingual sentiment modeling and cross-lingual resource development.

ACKNOWLEDGMENT

The authors gratefully acknowledge the School of Computer Applications, Department of Technology and Science, Lovely Professional University, Punjab, India, for their valuable support and assistance, which greatly helped in the development of this PhD research project.

DATA AVAILABILITY AND ETHICS

The developed En–Bn–Code–Mixed–Two–Class–Sentiment–Dataset is publicly available on Hugging Face at: https://huggingface.co/datasets/DaliaBarua/En-Bn-Code-Mixed-Two-Class-Sentiment-Dataset/viewer. The Amazon Product Reviews dataset is publicly available at: https://www.kaggle.com/datasets/arhamrumi/amazon-product-reviews. The Dakshina dataset is publicly available at: https://github.com/google-research-datasets/dakshina, and the Amazon Review Polarity dataset is publicly available at: https://www.kaggle.com/datasets/bhavikardeshna/amazon-customerreviews-polarity.

All datasets are publicly available and used under their respective licenses. No personal or sensitive data have been used, and all ethical standards have been followed in the creation and analysis of the datasets.

REFERENCES

[1] E. D. Liddy, "*Natural Language Processing*," in *Encyclopedia of Library and Information Science*, 2nd ed. New York City, NY, USA: Marcel Dekker Inc, 2001.

[2] K. R. Chowdhary, "Natural Language Processing," in *Fundamentals of Artificial Intelligence*, New Delhi, India: Springer India, 2020, pp. 603–649.

[3] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural Language Processing: State of The Art, Current Trends and Challenges," 2017, https://doi.org/10.48550/ARXIV.1708.05148.

[4] T. Young, D. Hazarika, S. Poria, and E. Cambria,"Recent Trends in Deep Learning Based Natural Language Processing [Review Article]," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, Aug. 2018, https://doi.org/10.1109/MCI.2018.2840738.

[5] S. Alam, M. F. Ishmam, N. H. Alvee, M. S. Siddique, M. A. Hossain, and A. R. M. Kamal, "BnSentMix: A Diverse Bengali-English Code-Mixed Dataset for Sentiment Analysis," arXiv, 2024, https://doi.org/10.48550/ARXIV.2408.08964.

[6] B. R. Chakravarthi *et al.*, "DravidianCodeMix: Sentiment Analysis and Offensive Language Identification Dataset for Dravidian Languages in Code-Mixed Text," *Language Resources and Evaluation*, vol. 56, no. 3, pp. 765–806, Sept. 2022, https://doi.org/10.1007/s10579-022-09583-7.

[7] M. Tareq, Md. F. Islam, S. Deb, S. Rahman, and A. A. Mahmud, "Data-Augmentation for Bangla-English Code-Mixed Sentiment Analysis: Enhancing Cross Linguistic Contextual Understanding," *IEEE Access*, vol. 11, pp. 51657–51671, 2023, https://doi.org/10.1109/ACCESS.2023.3277787.

[8] S. Bal, S. Mahanta, L. Mandal, and R. Parekh, "Bilingual Machine Translation: English to Bengali," in *Proceedings of International Ethical Hacking Conference 2018*, vol. 811, M. Chakraborty, S. Chakrabarti, V. E. Balas, and J. K. Mandal, Eds. Singapore: Springer Singapore, 2019, pp. 247–259.

[9] M. Khairullah, "A Novel Steganography Method using Transliteration of Bengali Text," *Journal of King Saud University - Computer and Information Sciences*, vol. 31, no. 3, pp. 348–366, July 2019, https://doi.org/10.1016/j.jksuci.2018.01.008.

[10] J. Hartmann, M. Heitmann, C. Siebert, and C. Schamp, "More than a Feeling: Accuracy and Application of Sentiment Analysis," *International Journal of Research in Marketing*, vol. 40, no. 1, pp. 75–87, Mar. 2023, https://doi.org/10.1016/j.ijresmar.2022.05.005.

[11] Hidayaturrahman and I. Prawira, "Leveraging Zero-Shot Learning in Large Language Models for Sentiment Analysis: A Comparative Study on the Indonesian Language," in *2024 International Conference on Informatics, Multimedia, Cyber and Information System*, Jakarta, Indonesia, Nov. 2024, pp. 614–619, https://doi.org/10.1109/ICIMCIS63449.2024.10956237.

[12] A. Hasan, S. Moin, A. Karim, and S. Shamshirband, "Machine Learning-Based Sentiment Analysis for Twitter Accounts," *Mathematical and Computational Applications*, vol. 23, no. 1, p. 11, Feb. 2018, https://doi.org/10.3390/mca23010011.

[13] H. Huang, A. A. Zavareh, and M. B. Mustafa, "Sentiment Analysis in E-Commerce Platforms: A Review of Current Techniques and Future Directions," *IEEE Access*, vol. 11, pp. 90367–90382, 2023, https://doi.org/10.1109/ACCESS.2023.3307308.

[14] K. D. S. Devi, V. Sireesha, C. Sudha, M. Ravisankar, and P. D. K. Reddy, "A Novel Approach to Sentiment Analysis using GMM-Enhanced N-gram LSTM Networks," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23068–23073, June 2025, https://doi.org/10.48084/etasr.10640.

[15] K. Korovkinas and P. Danėnas, "Support Vector Machine (SVM) and Naïve Bayes Classification Ensemble Method for Sentiment Analysis," *Baltic Journal of Modern Computing*, vol. 5, no. 4, Dec. 2017, https://doi.org/10.22364/bjmc.2017.5.4.06.

[16] K. L. Tan, C. P. Lee, K. M. Lim, and K. S. M. Anbananthen, "Sentiment Analysis With Ensemble Hybrid Deep Learning Model," *IEEE Access*, vol. 10, pp. 103694–103704, 2022, https://doi.org/10.1109/ACCESS.2022.3210182.

[17] Ankit and N. Saleena, "An Ensemble Classification System for Twitter Sentiment Analysis," *Procedia Computer Science*, vol. 132, pp. 937–946, 2018, https://doi.org/10.1016/j.procs.2018.05.109.

[18] P. Thiengburanathum and P. Charoenkwan, "SETAR: Stacking Ensemble Learning for Thai Sentiment Analysis Using RoBERTa and Hybrid Feature Representation," *IEEE Access*, vol. 11, pp. 92822–92837, 2023, https://doi.org/10.1109/ACCESS.2023.3308951.

[19] D. Moldovan, "A Majority Voting Framework for Reliable Sentiment Analysis of Product Reviews," *PeerJ Computer Science*, vol. 11, p. e2738, Feb. 2025, https://doi.org/10.7717/peerj-cs.2738.

[20] D. S. Krishna, G. Srinivas, and P. V. G. D. Prasad Reddy, "Disaster Tweet Classification: a Majority Voting Approach using Machine Learning Algorithms," *Intelligent Decision Technologies*, vol. 17, no. 2, pp. 343–355, May 2023, https://doi.org/10.3233/IDT-220310.

[21] F. Suandi *et al.*, "Enhancing Sentiment Analysis Performance Using SMOTE and Majority Voting in Machine Learning Algorithms," in *Proceedings of the 7th International Conference on Applied Engineering*, vol. 251, L. Lumombo, A. Rahmi, S. Suwarno, N. Ardi, and D. E. Kurniawan, Eds. Dordrecht: Atlantis Press International BV, 2024, pp. 126–138.

[22] Z. Qin, K. Dong, and B. Xie, "What Affects Customers Online Shopping Behavior, Research that Applied Machine Learning to Amazon Product Reviews," *Applied and Computational Engineering*, vol. 76, no. 1, pp. 48–64, July 2024, https://doi.org/10.54254/2755-2721/76/20240564.

[23] Y. Xiao, C. Qi, and H. Leng, "Sentiment Analysis of Amazon Product Reviews Based on NLP," in *2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering*, Changsha, China, Mar. 2021, pp. 1218–1221, https://doi.org/10.1109/AEMCSE51986.2021.00249.

[24] B. Roark *et al.*, "Processing South Asian languages written in the Latin script: the Dakshina dataset," in *12th Language Resources and Evaluation Conf*erence, Marseille, France, 2020, pp. 2413–2423.

[25] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *NIPS'15: Proceedings of the 29th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA, Dec. 2015, pp. 649–657.

[26] M. R. A. Rashid, K. F. Hasan, R. Hasan, A. Das, M. Sultana, and M. Hasan, "A comprehensive dataset for sentiment and emotion classification from Bangladesh e-commerce reviews," *Data in Brief*, vol. 53, p. 110052, Apr. 2024, https://doi.org/10.1016/j.dib.2024.110052.

AUTHOR PROFILES

Dalia Barua is a Ph.D. scholar in Computer Applications at Lovely Professional University, India. She holds an M.Sc. in Information Technology from Jahangirnagar University and a B.Sc. in Computer Engineering from the American International University–Bangladesh. Her current research focuses on text-based language processing.

Tarandeep Singh Walia is an Associate Professor in Computer Applications at Lovely Professional University, India. He holds an MCA and Ph.D. from IKG Punjab Technical University, has over 14 years of teaching experience, and his research in AI, programming, and natural language processing is reflected in his publications.