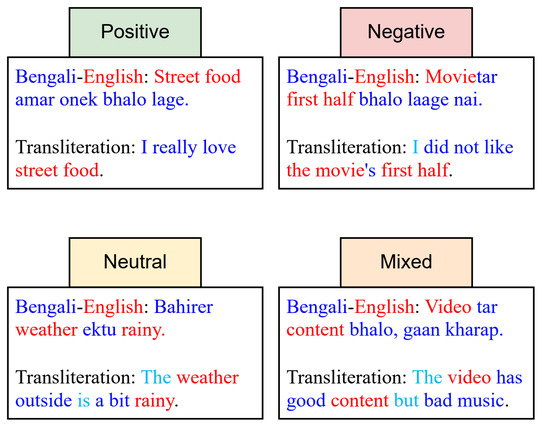
BnSentMix: A Diverse Bengali-English Code-Mixed Dataset for Sentiment Analysis

Sadia Alam Md Farhan Ishmam Navid Hasin Alvee Md Shahnewaz Siddique
Md Azam Hossain Abu Raihan Mostofa Kamal

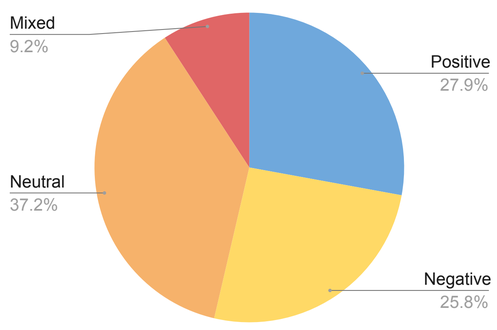
Network and Data Analysis Research Group, Computer Science and Engineering, Islamic University of Technology

Background

- **Code-mixing/Code-Switching:** Alternating between two or more languages within a single conversation or sentence.
- **Intra-sentential Switching:** Foreign words appear within the same sentence.
- **Intra-word Switching:** Foreign word elements (*e.g.* sub-words) are combined to form a single word.



(a) Examples of the four sentiment labels – Blue: Bengali, Red: English, and Cyan: Implicit Words.



(b) Distribution of sentiment labels. Mixed sentiment represents the presence of both positive and negative sentiments in different parts of the text.

Motivation

- Code-mixing with English is common in colloquial text written by Bengalis but has limited research and resources in Bengali NLP.
- Code-mixing can be challenging due to inter and intra-word mixing.
- No existing automated text filtering method for code-mixed Bengali.
- Current Code-mixed Bengali-English Sentiment Analysis (CBESA) datasets are not publicly available and limited to 5k samples only.
- Existing research did not evaluate multilingual language models on CBESA.

Contribution

1. We present, BnSentMix, a CBESA dataset comprising 20,000 samples and 4 sentiment labels.
2. Dataset has been curated from YouTube, Facebook, and E-commerce platforms to encapsulate a broad spectrum of contexts and topics in a realistic setting.
3. We propose a novel automated code-mixed text detection pipeline using fine-tuned language models, reaching an accuracy of 94.56%.
4. We establish 11 baselines on CBESA, including classical machine learning, neural network, and pre-trained transformer-based language models.

BnSentMix Dataset

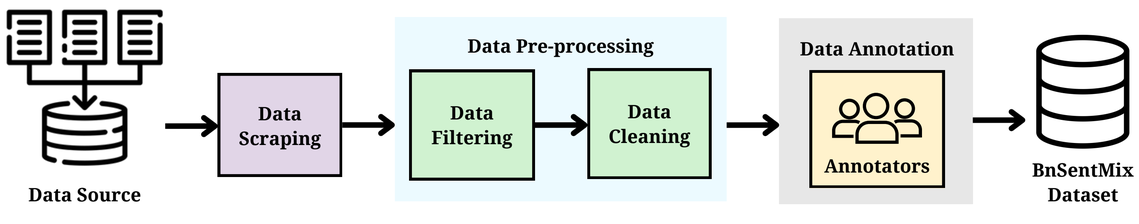


Figure 2. Dataset creation pipeline.

i **Scrapping:** 3 million user-generated data samples have been scrapped from online platforms using the YouTube API, Facepager, and Selenium.

ii **Filtering:** Automated data filtering using fine-tuned mBERT [3].

iii **Cleaning:** Discarded samples with four words or less and URLs. Removed whitespaces, special characters, emojis, and emoticons. Consequent sequences of punctuations have been reduced to a single instance.

iv **Annotation:** Each sample has been annotated by 2 annotators and third annotator is assigned for tie-breaking.

Automated Code-Mixed Text Filtering

We fine-tune pre-trained language models on a novel Bengali-English code-mix detection dataset and use the fine-tuned model to filter the code-mixed text for annotation.

Dataset: We collect and label text from 3 data sources: Dakshina [5], Kaggle English word dataset [6], Mandal and Singh [4]. The corpus includes 100k words with a balanced mix of Bengali, English, and code-mixed words.

Language Models: We evaluate 3 pre-trained models – the multilingual models, mBERT [3] and XLM-RoBERTa [2], and the Bengali-English model BanglishBERT [1].

Model	XLM-RoBERTa		BanglishBERT		mBERT	
	Acc	F1	Acc	F1	Acc	F1
Score	0.896	0.898	0.906	0.896	0.946	0.940

Table 1. Performance of the fine-tuned language models on code-mixed text detection.

Baselines

Machine Learning (ML) Models: The ML Models offer simple baselines with the SVM achieving accuracy and F1 score on par with larger transformer-based models. The ML baselines can be effective in resource-constrained settings.

Recurrent Neural Network (RNN) Variants: The base RNN variant relatively underperformed while the Long Short Term Memory (LSTM) model significantly outperformed the base RNN.

Pre-trained Language Models: The pre-trained language models are primarily transformer-based architectures which are pre-trained on either English, Bengali, or multilingual corpus. BERT [3] achieves the best performance, closely followed by XLM-RoBERTa [2] and mBERT [3].

Model	Validation				Test			
	Acc	Precision	Recall	F1	Acc	Precision	Recall	F1
Machine Learning Models								
Logistic Regression	0.668	0.656	0.668	0.662	0.667	0.614	0.667	0.639
Random Forest	0.672	0.661	0.672	0.666	0.648	0.635	0.648	0.641
SVM	0.694	0.676	0.694	0.685	0.660	0.637	0.660	0.648
Recurrent Neural Network Variants								
RNN	0.406	0.308	0.406	0.350	0.401	0.352	0.401	0.375
LSTM	0.678	0.670	0.678	0.674	0.670	0.657	0.670	0.663
Multilingual Language Models								
XLM-RoBERTa	0.726	0.709	0.726	0.717	0.698	0.642	0.698	0.669
mBERT	0.726	0.713	0.726	0.719	0.694	0.675	0.694	0.684
Bangla Language Models								
BanglaBERT	0.721	0.668	0.721	0.693	0.698	0.642	0.698	0.669
BanglishBERT	0.694	0.715	0.694	0.704	0.686	0.653	0.686	0.669
English Language Models								
DistilBERT	0.701	0.694	0.701	0.697	0.672	0.665	0.672	0.668
BERT	0.727	0.710	0.724	0.717	0.695	0.683	0.694	0.688

Table 2. Performance of the proposed baselines based on accuracy, precision, recall, and F1 score.

Future Directions

- The dataset is slightly imbalanced with only 9.2% mixed sentiment samples, which can affect the performance in classifying mixed sentiments. Further error analysis can be conducted to reveal the impact of imbalance on overall performance.
- Back-transliterating and classifying using Bengali language models can be a plausible direction due to the limited CBESA resources.
- The recent large language models using zero-shot prompting, in-context learning, or translation prompting can potentially produce good results in code-mixed tasks.

Conclusion

BnSentMix establishes a novel dataset and text detection method for sentiment analysis of code-mixed Bengali-English.

Acknowledgments

Our work is supported by the Islamic University of Technology Research Seed Grants (IUT RSG). We sincerely appreciate Mohammed Saidul Islam and Md Mezbaur Rahman for guidance and Nejd Khadija for proofreading our work.

Authors' Note: *We honor the brave souls of the July student movement, reflecting on their courage, resilience, and fight for justice.*

References

[1] Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar.
BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla.
In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327. Association for Computational Linguistics, July 2022.

[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov.
Unsupervised cross-lingual representation learning at scale.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics, July 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
BERT: Pre-training of deep bidirectional transformers for language understanding.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, June 2019.

[4] Soumil Mandal and Anil Kumar Singh.
Language identification in code-mixed data using multichannel neural networks and context capture.
In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 116–120, 2018.

[5] Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall.
Processing south asian languages written in the latin script: the dakshina dataset.
In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2413–2423, 2020.

[6] Rachael Tatman.
English word frequency dataset, 2017.
Accessed: 2025-01-15.