DEEP LEARNING BASED SENTIMENT ANALYSIS IN A CODE-MIXED ENGLISH-BENGALI SOCIAL MEDIA CORPUS

Shadbin Akter | c00317205 | MSc in Applied Artificial Intelligence | SETU Carlow

Supervisor: Dr Rejwanul Haque | Department of Computing

Introduction

Code-mixed (CM) text — alternating between languages in a single utterance — is ubiquitous on social media among multilingual communities worldwide.

Bangla-English CM (BE-CM) text presents distinct NLP challenges:

- Heavy code-switching within single sentences
- Romanised Bangla (e.g. 'Ami bhalo achi')
- Non-standard, inconsistent spelling conventions
- Very limited high-quality annotated datasets

Bengali has 250M+ native speakers globally yet remains severely under-resourced in NLP. Existing sentiment tools fail on CM text.

mDeBERTa-v3 has been used on Tamil and Uzbek code-mixed tasks but never on BE-CM. LaBSE has been used for multilingual low-resource sentiment but never for BE-CM. No hybrid deep learning + classical ML model exists for this task. These are the gaps this work addresses.

Research Objective

To benchmark established transformer baselines (mBERT, XLM-R, BERT) and apply novel architectures (DeBERTa, adapter-BERT) and a hybrid deep learning + classical ML model to BE-CM sentiment analysis — none of which have been tested on this task.

Research Questions:

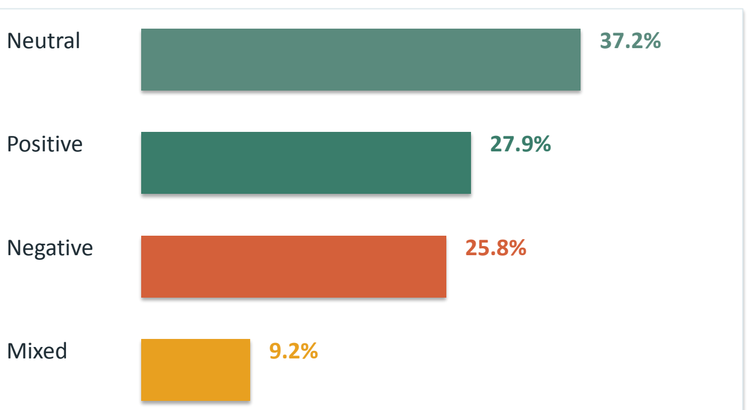
① Do DeBERTa and adapter-BERT outperform existing BERT-family baselines on BE-CM?

② Does a hybrid model (transformer embeddings + XGBoost / SVM / LR) outperform standalone transformers?

Dataset

20K	4	Multi
Samples	Labels	Sources

Label Distribution (BNSENTMIX)



BNSENTMIX (Alam et al., 2025)
Largest annotated BE-CM corpus — Facebook (73%), YouTube (18%), e-commerce (9%). Expert human annotation, Cohen's κ = 0.86.
Splits: Train 70% (14,000) · Val 15% (3,000) · Test 15% (3,000)

Literature Review

Language Identification & Datasets

Chanda et al. (2016): pioneered word-level LID for English-Bengali CM. Alam et al. (2025): BNSENTMIX — 20K samples, 4 labels, multi-source.

Classical Machine Learning

Sultana et al. (2024): RF + TF-IDF → 83% accuracy. Tareq et al. (2023): XGBoost + FastText → 87% weighted F1.

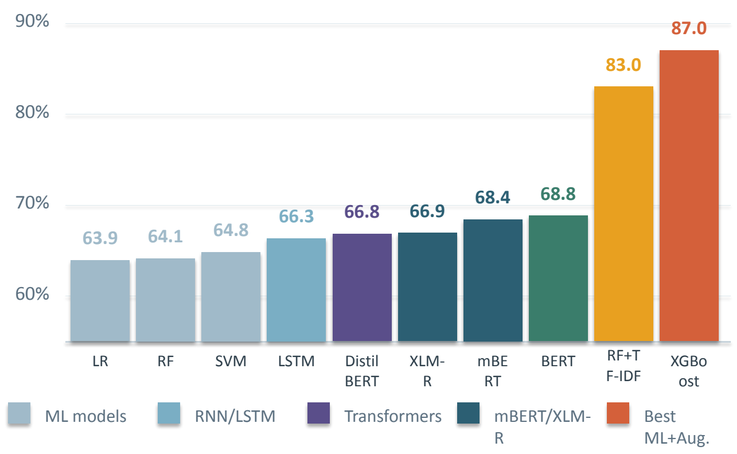
Transformer & Deep Learning Models

Alam et al. (2025): BERT → 68.8% F1 · mBERT → 68.4% F1 on BNSENTMIX. Raihan et al. (2023): Mixed-Distil-BERT — efficient distillation for CM.

Hybrid & New LLM Approaches

Yadav et al. (2022): normalisation + hybrid architectures improve CM performance. DeBERTa (disentangled attention) & adapter-BERT (parameter-efficient fine-tuning) — both architecturally motivated for BE-CM, both untested on this task.

Existing Baseline Performance on BE-CM Sentiment (BNSENTMIX Test F1)



Research Gap: mDeBERTa-v3 is untested on BE-CM (used only on Tamil/Uzbek CM tasks). LaBSE is untested on BE-CM (used only in African multilingual sentiment). No hybrid transformer + classical ML model exists for BE-CM sentiment analysis.

Methodology

① Data Preparation

BNSENTMIX 20K · Clean & normalise text · 70 / 15 / 15 split

② Text Preprocessing & Tokenisation

Tokenisation · Handle Romanised Bangla & Unicode · Remove noise

③ Model Fine-Tuning

Fine-tune Baselines - mBERT · XLM-R · BERT
Novel models - DeBERTa · adapter-BERT · mDeBERTa-v3 · LaBSE

④ Hybrid Model

Transformer embeddings → XGBoost · SVM · Logistic Regression.
Compare hybrid vs. standalone transformer performance.

⑤ Evaluation & Ablation Study

Accuracy · Macro F1 · Precision · Recall · Confusion Matrix

Technologies & Models

Model Comparison

Model	Type	F1 (Test)	Status
LR	ML	0.639	Existing baseline
SVM	ML	0.648	Existing baseline
LSTM	RNN	0.663	Existing baseline
DistilBERT	Transformer	0.668	Existing baseline
BERT	Transformer	0.688	Best existing baseline
mBERT	Multilingual	0.684	Baseline — used here
XLM-R	Multilingual	0.669	Baseline — used here
DeBERTa	Novel LLM	—	Untested on BE-CM
adapter-BERT	Novel LLM	—	Untested on BE-CM
Hybrid Model	DL + ML	—	Novel contribution

Libraries & Tools

Python	HuggingFace	PyTorch
scikit-learn	Jupyter	langdetect

Early Indications & Next Steps

Early Indications

Deep learning outperforms classical ML

BERT: 68.8% F1 vs RF: 64.1% vs LR: 63.9% on BNSENTMIX — confirms the advantage of contextual representations for BE-CM sentiment analysis.

Hybrid models show strong potential

Combining transformer embeddings with XGBoost/SVM/LR has shown gains on similar CM NLP tasks, suggesting measurable gains are achievable for BE-CM.

mDeBERTa-v3 & LaBSE are untested on BE-CM

Both used on other CM tasks (Tamil, Uzbek, African languages) — never applied to BE-CM sentiment. First application to BNSENTMIX is a clear, testable novelty.

BNSENTMIX enables reliable benchmarking

Largest annotated BE-CM corpus (Cohen's κ = 0.86, 3 sources) — ensures fair, reproducible comparison across all transformer and hybrid model conditions.

Next Steps

1. Fine-tune mDeBERTa-v3, DeBERTa, adapter-BERT and LaBSE on BNSENTMIX alongside mBERT, XLM-R and BERT baselines using identical hyperparameters.
2. Build hybrid model: extract [CLS] embeddings from best-performing transformer → train XGBoost, SVM and Logistic Regression classifiers. Compare all combinations.
3. Run systematic ablation study — remove each component one at a time. Evaluate Accuracy, Macro F1, Precision, Recall and Confusion Matrix across all conditions.
4. Benchmark all results against BNSENTMIX baselines (Alam et al., 2025 Table 4) to precisely quantify the novel contribution of each new model and the hybrid architecture.

References: 1. Alam et al. (2025). BNSENTMIX: Bengali-English Code-Mixed Dataset for Sentiment Analysis. LoResMT 2025. 2. Sultana & Mamun (2024). Cross-Lingual Augmentation for BE-CM Sentiment Analysis. IEEE ICEEICT. 3. Raihan et al. (2023). Mixed-Distil-BERT: Code-mixed Language Modeling for Bangla, English & Hindi. arXiv:2309.10272. 4. Tareq et al. (2023). Data-Augmentation for Bangla-English Code-Mixed Sentiment Analysis. IEEE Access. 5. Chanda et al. (2016). Unraveling the English-Bengali Code-Mixing Phenomenon. EMNLP.

6. Yadav et al. (2022). Normalisation of Spelling Variations in Code-Mixed Data. ICON 2022. 7. Nelatoori & Kommanti (2024). Sentiment analysis in Tamil code-mixed text using adapter-based mDeBERTa. Journal of Big Data, Springer. 8. Roy et al. (2026). Hybrid NER-Sentiment Model with mDeBERTa-v3 for code-mixed Uzbek text. MDPI Big Data. 9. Azime et al. (2025). Advancing sentiment analysis for low-resourced African languages — LaBSE ensemble. PMC / PLOS ONE.