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

Sadia Alam, Md Farhan Ishmam, Navid Hasin Alvee,
Md Shahnewaz Siddique, Abu Raihan Mostofa Kamal, Md Azam Hossain
Department of Computer Science and Engineering, Islamic University of Technology
{sadiaalam,farhanishmam,navidhasin,shahnewaz,raihan.kamal,azam}@iut-dhaka.edu

Abstract

The widespread availability of code-mixed data in digital spaces can provide valuable insights into low-resource languages like Bengali, which have limited annotated corpora. Sentiment analysis, a pivotal text classification task, has been explored across multiple languages, yet code-mixed Bengali remains underrepresented with no large-scale, diverse benchmark. Code-mixed text is particularly challenging as it requires the understanding of multiple languages and their interaction in the same text. We address this limitation by introducing BnSentMix, a sentiment analysis dataset on code-mixed Bengali comprising 20,000 samples with 4 sentiment labels, sourced from Facebook, YouTube, and e-commerce sites. By aggregating multiple sources, we ensure linguistic diversity reflecting realistic code-mixed scenarios. We implement a novel automated text filtering pipeline using fine-tuned language models to detect code-mixed samples and expand code-mixed text corpora. We further propose baselines using machine learning, neural networks, and transformer-based language models. The availability of a diverse dataset is a critical step towards democratizing NLP and ultimately contributing to a better understanding of code-mixed languages.

1 Introduction

In the rapidly evolving digital landscape, code-mixing has become increasingly prevalent, particularly in multilingual societies. Code-mixing is the phenomenon of alternating between two or more languages within a single conversation or sentence (Thara and Poornachandran, 2018). Code-mixing can occur in various forms, including intra-sentential switching, where words from different languages appear within the same sentence, and intra-word switching, where elements from other languages combine to form a single word (Stefanich et al., 2019; Litcofsky and Van Hell, 2017).

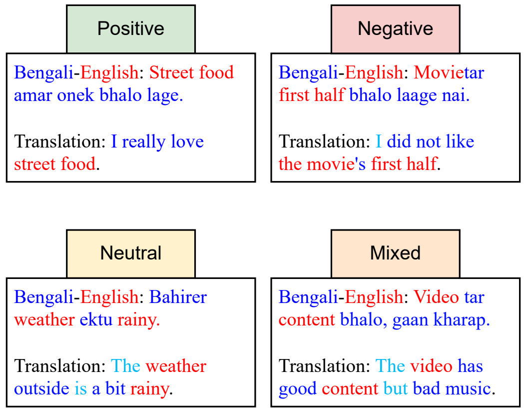

Figure 1: Examples of the four sentiment labels from our code-mixed Bengali-English dataset BnSentMix and the corresponding English translations. Red represents English words, blue represents Bengali words written in English alphabets, and cyan represents implicit words in the code-mixed text.

Intra-sentential switching is more frequently observed in colloquial settings. One significant yet understudied domain of code-switching is Bengali-English code-mixed text.

We consider Fig. 1 where the sentences are examples of Bengali-English intra-sentential switching. Intra-word switching is observed in the negative sentiment example. Here, Movietar is considered a single word, whereas the Bengali sub-word tar indicates possession. We also observe several words in the transliterated text that are not explicitly written in the code-mixed text. These implicitly defined words increase the challenges in processing code-mixed Bengali-English texts.

With over 250 million native speakers globally, Bengali is the seventh most spoken language in the world but remains a low-resource language in terms of research. While typing texts, Bengali speakers often use Bengali-English code-mixed terms to express their thoughts in writing. Despite the preva-

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 68–77
January 20, 2025.

Dataset	#Samples	#SL	#DS	Filtering	#Baselines	PA
Hindi (Joshi et al., 2016)	3.9k	3	1	Manual	10	✓
Bengali (Mandal et al., 2018)	5k	3	1	Manual	5	✗
Tamil (Chakravarthi et al., 2020b)	15.7k	5	1	`langdetect`	10	✓
Malayalam (Chakravarthi et al., 2020a)	6.7k	5	1	`langdetect`	10	✓
Persian (Sabri et al., 2021)	3.6k	3	1	Keywords search	3	✓
Swiss (Pustulka-Hunt et al., 2018)	963	3	1	Manual	7	✗
BnSentMix (Ours)	**20k**	**4**	**3**	**mBERT**	**14**	✓

Table 1: Comparison of the number of samples, #SL: Sentiment Labels, #DS: Data Sources, filtering method, number of baselines, and PA: Public Availability of various code-mixed (with English) sentiment analysis datasets.

lence of code-mixed text on social media platforms, e-commerce sites, and other digital spaces, there remains a notable scarcity of resources to analyze and process such data.

Sentiment analysis, the computational study of people's opinions, sentiments, emotions, and attitudes expressed in written language, plays a critical role in various applications, including social media monitoring, customer feedback, market research, and public opinion analysis (Wankhade et al., 2022). While substantial progress has been made in monolingual sentiment analysis (Medhat et al., 2014; Birjali et al., 2021), the complexities introduced by code-mixed texts present unique challenges that current models struggle to address (Barman et al., 2014). This is particularly true for Bengali-English code-mixed texts (Chanda et al., 2016), which have not received adequate attention in existing research.

Table 1 highlights the limitations of Bengali-English code-mixed sentiment analysis datasets compared to other Indic-English code-mixed datasets. The only available Bengali dataset (Mandal et al., 2018) is limited to $5k$ samples, 3 sentiment labels, a single data source, 5 baselines, and is not publicly available. The existing language detection tools also have severe limitations in filtering code-mixed Bengali-English. Tools like `langdetect`[1] and Bengali phonetic parser[2] designed for general language identification and code-mixed Bengali identification struggled with the spelling nuances of code-mixed text.

Addressing these challenges, our contribution can be summarized:

- We present, BNSENTMIX, a novel Bengali-English code-mixed dataset comprising 20,000 samples and 4 sentiment labels for sentiment analysis. Data has been curated from YouTube, Facebook, and e-commerce platforms to encapsulate a broad spectrum of contexts and topics.
- Following the intricacies of code-mixed test, visualized in Fig. 1, we propose a novel automated code-mixed text detection pipeline using fine-tuned language models, reaching an accuracy of 94.56%.
- We establish 11 baselines including classical machine learning, neural network, and pre-trained transformer-based models, with BERT achieving accuracy and F1 score of 69.5% and 68.8% respectively.

2 Related Work

2.1 Code-Mixing

Code-mixed data can be the source of several text classification tasks (Thara and Poornachandran, 2018) with sentiment analysis (Mahadzir et al., 2021) being one of the most popular ones. Other natural language processing tasks (NLP) on code-mixed data include hate speech detection (Sreelakshmi et al., 2020), translation (Gautam et al., 2021), part of speech tagging (Vyas et al., 2014), emotion classification (Ameer et al., 2022), language identification (Mandal and Singh, 2018), and speech synthesis (Sitaram and Black, 2016). Researchers also incorporate training data augmentation (Gupta et al., 2021; Rizvi et al., 2021) and code-mix word embeddings (Pratapa et al., 2018) to process code-mixed texts.

2.2 Sentiment Analysis

The significance of sentiment analysis has grown with the rise of social media, prompting extensive

[1] https://pypi.org/project/langdetect/
[2] https://github.com/porimol/bnbphoneticparser

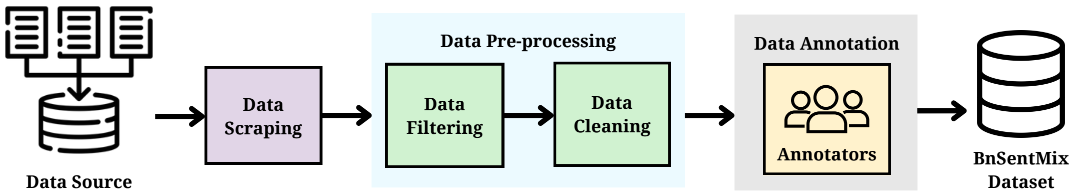

Figure 2: Dataset creation pipeline of the BNSENTMIX dataset.

research on monolingual corpora. Studies explored various languages, including English (Hu and Liu, 2004; Wiebe et al., 2005; Jiang et al., 2019), Russian (Rogers et al., 2018), German (Cieliebak et al., 2017), Norwegian (Mæhlum et al., 2019), several Indian languages (Agrawal and Awekar, 2018; Rani et al., 2020), and Bengali (Fahim, 2023; Kabir et al., 2023). Multilingual sentiment analysis (Dashtipour et al., 2016; Pustulka-Hunt et al., 2018) gained popularity with the recent advancements in multilingual language models (Devlin et al., 2019; Conneau et al., 2020).

2.3 Code-Mixing in Bengali

Bengali is often code-mixed with English (Chanda et al., 2016) and Hindi (Raihan et al., 2023). In Bengali-English code-mixing, English tokens are commonly used alongside romanized or transliterated Bengali (Shibli et al., 2023; Fahim et al., 2024), which is often back-transliterated before processing (Haider et al., 2024). Sentiment analysis on code-mixed Bengali has limited studies, either using small private datasets (Mandal et al., 2018) or performed in a multilingual setting (Patra et al., 2018). Data augmentation techniques have also been explored to enhance code-mixed sentiment analysis datasets in Bengali (Tareq et al., 2023). Emotion detection, a task similar to sentiment analysis, has also been studied in the context of code-mixed Bengali (Raihan et al., 2024).

3 BNSENTMIX Dataset

The BNSENTMIX data has been collected from multiple data sources to reflect realistic code-mixed texts commonly found in digital spaces. We labeled the data using four distinct sentiments: the commonly used positive, negative, and neutral sentiments, as well as a *mixed* sentiment. As illustrated in Fig. 1, the mixed sentiment represents instances where both positive and negative sentiments are conveyed within different parts of the text. We decided to include the mixed label because the associated sentences are frequently observed in everyday texts and cannot be correctly classified under the traditional sentiment labels.

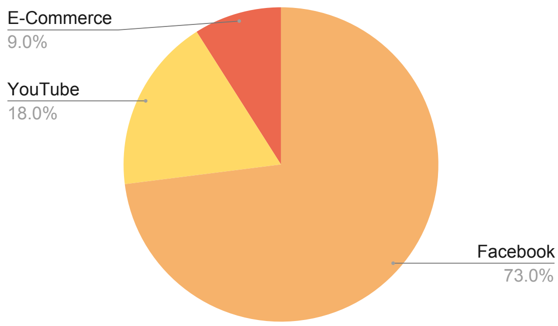

Figure 3: Composition of data sources of the BNSENTMIX dataset.

3.1 Data Sourcing

We collected extensive user-generated content from YouTube comments, Facebook comments, and e-commerce site reviews. These data sources were chosen for their high engagement rates and diverse linguistic input. YouTube comments were scraped using the YouTube API. We used `Facepager`[3] to extract comments from public Facebook posts, pages, and groups. `Selenium`[4] was employed to mimic human browsing behavior on e-commerce sites to scrape product reviews. We amassed over 3 million samples of user-generated content, forming the foundation for our dataset and subsequent analysis. Fig. 3 illustrates the composition of the aforementioned data sources.

3.2 Data Cleaning

We discard samples with four words or less and samples containing external URLs. Redundant whitespaces, special characters, and non-ASCII characters including emojis and emoticons are also removed. Consequent sequences of punctuation symbols are reduced to single instances. The English words are downcased unless they appear at

[3]https://github.com/strohne/Facepager
[4]https://selenium-python.readthedocs.io/

the beginning of the sentence. However, we did not correct any form of typing or grammatical errors in our dataset to ensure the trained model is robust for practical scenarios. The data cleaning procedure has been formally described in Algo. 1.

Algorithm 1 Clean Text

Require: $text \leftarrow$ Input text
Ensure: $text \leftarrow$ Preprocessed text
$text \leftarrow text.lower()$ {Convert to lowercase}
$text \leftarrow$ Remove all special characters except "?", ",", "!", and "."
$text \leftarrow$ Reduce consecutive sequences of punctuations to a single instance
$text \leftarrow$ Remove all non-ASCII characters
$text \leftarrow$ Remove extra white spaces
$text \leftarrow$ Capitalize the first letter after each period (.)
return $text$

3.3 Data Filtering

We construct a novel Bengali-English code-mix detection dataset and fine-tune pre-trained language models to automatically filter code-mixed Bengali-English. Detecting these texts can pose significant challenges: (i) rule-based methods struggle with intra-word switching (ii) romanized Bengali or English samples may be incorrectly classified as code-mixed text by automated methods, and (iii) samples from a third language often bypass the filtering process. Our approach addresses these challenges by incorporating pre-trained language models, which excel in intricate text detection settings. Algo. 2 outlines the data filtering pipeline.

3.3.1 Code-mix Detection Dataset

The fine-tuning dataset for code-mixed Bengali-English detection comprises 3 data sources. We incorporate the Dakshina dataset (Roark et al., 2020) which has a rich collection of Southeast Asian languages, including many Bengali-English code-mixed sentences. Secondly, we utilized a Kaggle English dataset[5] consisting of a wide range of English words and extended with a third source Mandal and Singh (2018). By integrating these diverse sources, we curated a comprehensive dataset of $100k$ words, ensuring a balanced mix of Bengali, English, and code-mixed Bengali-English words. To maintain the linguistic purity of code-mixed Bengali-English, we exclude sentences containing words that are neither English nor Bengali, e.g. Hindi words.

Algorithm 2 Detect Code-mixed Bengali

Require: $S \leftarrow$ List of sentences
Require: $model \leftarrow$ Pre-trained mBERT model
Require: $tokenizer \leftarrow$ Pre-trained mBERT tokenizer
Ensure: $pred \leftarrow$ Predicted class label (0 or 1)
$b\_count \leftarrow 0$
$w\_count \leftarrow 0$
for each $sent$ in S **do**
  $words \leftarrow$ split($sent$)
  for each w in $words$ **do**
    $w \leftarrow$ preprocess(w)
    if w is empty **then**
      continue
    end if
    $w\_count \leftarrow w\_count + 1$
    $inputs \leftarrow$ tokenize(w)
    $outputs \leftarrow$ model($inputs$)
    $pred\_class \leftarrow$ argmax($outputs$)
    if $pred\_class == 1$ **then**
      $b\_count \leftarrow b\_count + 1$
    end if
  end for
end for
if $w\_count < 4$ **then**
  return 0
end if
$b\_percent \leftarrow b\_count/w\_count$
if $b\_percent \geq 0.3$ **then**
  return 1
else
  return 0
end if

3.3.2 Code-mix Detection Results

We evaluate 3 pre-trained models – the multilingual models, mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), and the Bengali-English model BanglishBERT (Bhattacharjee et al., 2022). Table 2 reveals mBERT showing substantially higher accuracy and F1 score in code-mixed Bengali-English detection. We argue that the pre-trained multilingual capabilities of mBERT effectively handled the nuances of code-mixed Bengali-English text.

[5]https://www.kaggle.com/datasets/rtatman/english-word-frequency

Model	Acc(%)	F1 Score
XLM-RoBERTa	89.60	0.8985
BanglishBERT	90.56	0.8961
mBERT	**94.56**	**0.9403**

Table 2: Comparison of the accuracy and F1 score of the code-mixed Bengali-English detection methods.

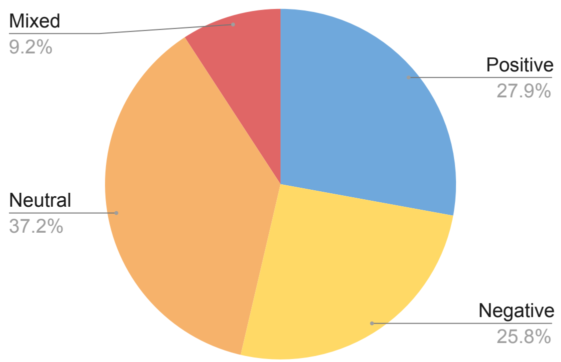

Figure 4: Distribution of sentiment labels in the BNSENTMIX dataset.

3.4 Data Annotation

Each sample in our dataset has been annotated twice by two different annotators to ensure generalized sentiment is conveyed. In cases where the two independent annotations did not match, a third annotator would break the tie. To perform data annotation, we recruited 64 annotators who had been provided hourly monetary compensation. The data annotators have at least a high-school degree (equivalent to Grade 12 education) and are familiar with social media and digital spaces. The annotators were asked to re-label the same 250 samples to measure inter-annotator agreement. We measured the agreement score using Cohen's Kappa $\kappa = 0.86$, indicating substantial agreement.

3.5 Dataset Statistics

Fig. 4 visualizes the label composition of the annotated dataset. An overview of the key statistics of the annotated dataset is shown in table-3. We split the dataset into $[70 : 15 : 15]$ training, validation, and test splits i.e. 14,000, 3,000, and 3,000 samples respectively.

4 Methodology and Experimental Setup

4.1 Baseline Models

We evaluate 11 baselines encompassing traditional machine learning models, recurrent neural network variants, and transformer-based pre-trained language models, observed in table 4. All the pre-trained models were fine-tuned on our dataset.

4.2 Evaluation Metrics

We use classification accuracy and F1-score for model evaluation – both well-known metrics for text classification (Hossin and Sulaiman, 2015).

Statistic	Value
Mean Character Length	62.77
Max Character Length	1985
Min Character Length	14
Mean Word Count	11.65
Max Word Count	368
Min Word Count	4
Unique Word Count	37734
Unique Sentence Count	20000

Table 3: Key statistics of the BNSENTMIX dataset.

4.3 Implementation Details

The models were trained on NVIDIA Tesla P100 GPUs with 16GB of memory. We followed the Huggingface implementation (Wolf et al., 2019) for the pre-trained language models. All the models utilized Adam Optimizer (Kingma and Ba, 2014) with a training batch size of 32. The training configuration used most of the default hyperparameters. Logistic Regression, RNN, and LSTM models used the learning rate of $1E{-}5$ while the BERT-family language models used the learning rate of $1.5E{-}6$. The training time for each epoch varied from 8 to 13 minutes.

5 Results and Analysis

5.1 Performance Evaluation

Table 4 highlights the performance of the 11 baselines with BERT achieving the best performance in terms of both accuracy and F1 score. We now analyze the category-wise model performance.

5.1.1 Machine Learning (ML) Models

The ML models provide simple baselines and achieve considerably high accuracy, with the Sup-

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 4: Performance of the proposed baselines based on accuracy, precision, recall, and F1 score.

port Vector Machine (SVM) (Vapnik, 1995) achieving accuracy and F1 score on par with larger transformer-based models like BanglishBERT. The other two ML baselines Logistic Regression (Cox, 1958) and Random Forest (Breiman, 2001) achieve satisfactory performance with relatively simpler architectures. These ML baselines can be effective in resource-constrained scenarios.

5.1.2 Recurrent Neural Networks (RNNs)

RNN (Hopfield, 1982) underperformed compared to the other baselines. On the contrary, the performance of Long Short-Term Memory (LSTM) models (Hochreiter and Schmidhuber, 1997) was significantly higher in terms of both accuracy and F1 score. We argue that the long-term textual dependencies and the impact of vanishing and exploding gradients limited the performance of the RNN models.

5.1.3 Transformer-based Models

The best performance is achieved by the BERT model (Devlin et al., 2019) pre-trained on an English corpus. The BERT model is closely followed by the multilingual models XLM-RoBERTa (Conneau et al., 2020) and mBERT (Devlin et al., 2019). We hypothesize that the low proportion of Bengali text in the multilingual pre-training corpus does not

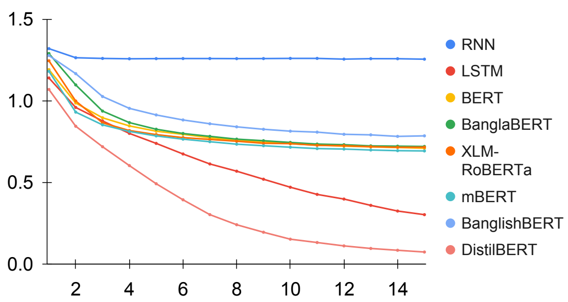

Figure 5: Comparison of epoch-wise training loss of the established baselines.

provide any significant advantage in code-mixed Bengali classification tasks.

In contrast, English pre-trained models like BERT exhibit better understanding of the linguistic intricacies of English words used in code-mixed Bengali, thereby producing better performance than other multilingual and Bengali models. Similarly, the Bengali language models BanglaBERT (Bhattacharjee et al., 2022) and BanglishBERT (Bhattacharjee et al., 2022) are trained on Bengali and Bengali-English corpora respectively. Code-mixed Bengali uses English tokens and hence, the pre-training on Bengali tokens does not provide any significant advantage. The lighter version of BERT, DistilBERT (Sanh et al., 2019) produces comparable but slightly worse results.

5.2 Training Loss Analysis

Figure 5 illustrates the training loss across 15 epochs for the baselines. We observe that all models converge before reaching the 15^{th} epoch. The only exception is the LSTM model which shows a slight indication of being benefited by additional training epochs. Excluding DistilBERT, the other BERT family models converged relatively faster in the earlier epochs. For most models, training for 5-8 epochs is appropriate to prevent overfitting.

6 Conclusion

We introduce BNSENTMIX, a novel sentiment analysis dataset tailored for code-mixed Bengali-English. Our work opens several potential research avenues for code-mixed Bengali. Researchers can explore other tasks, such as hate speech, offensive language, and abusive content detection on code-mixed data. Our work addresses a significant gap for low-resource languages and sets a new standard for sentiment analysis in code-mixed Bengali-English.

Code & Data Availability

Our code and dataset are publicly available[6] under the Creative Commons Attribution 4.0 International (CC BY 4.0). Any form of private data or personal identification information has been removed from the dataset to prevent privacy violations.

Limitations

The label distribution of BNSENTMIX dataset is slightly imbalanced with only 9.2% samples labeled as mixed sentiment which can affect the performance of the model in classifying mixed sentiments. Further error analysis for each sentiment label can reveal the impact of imbalance on the overall performance of the model. We also acknowledge that the sentiment of the annotator can be a source of bias during data annotation, though each data sample has been annotated twice by two different annotators, and annotation conflicts have been resolved by a third annotator.

Ethical Statement

The hired data annotators were compensated significantly higher than the region's minimum wage. Each annotator was only given around 630 data samples with no time restrictions. This ensured that the annotator did not overwork during data annotation. Annotator sentiment is subject to long working hours and can affect sentiment labeling. To prevent this, we mandated five-minute breaks after every twenty-minute interval and provided refreshments upon request.

[6] https://github.com/Nishita2000/BnSentMix

Acknowledgements

Our work is supported by the Islamic University of Technology Research Seed Grants (IUT RSG) (Ref: REASP/IUT-RSG/2022/OL/07/012). We sincerely appreciate Mohammed Saidul Islam and Md Mezbaur Rahman for guidance and Nejd Khadija for proofreading our work.

References

Shivam Agrawal and Amit Awekar. 2018. No more beating about the bush: A step towards idiom handling for indian language NLP. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Iqra Ameer, Grigori Sidorov, Helena Gomez-Adorno, and Rao Muhammad Adeel Nawab. 2022. Multi-label emotion classification on code-mixed text: Data and methods. *IEEE Access*, 10:8779–8789.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the first workshop on computational approaches to code switching*, pages 13–23.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. 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, Seattle, United States. Association for Computational Linguistics.

Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st*

Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), pages 177–184, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Arunavha Chanda, Dipankar Das, and Chandan Mazumdar. 2016. Unraveling the english-bengali code-mixing phenomenon. In *Proceedings of the second workshop on computational approaches to code switching*, pages 80–89.

Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A twitter corpus and benchmark resources for german sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51, Valencia, Spain. Association for Computational Linguistics.

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

David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232.

Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad YA Hawalah, Alexander Gelbukh, and Qiang Zhou. 2016. Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8:757–771.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Md Fahim. 2023. Aambela at BLP-2023 task 2: Enhancing BanglaBERT performance for Bangla sentiment analysis task with in task pretraining and adversarial weight perturbation. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 317–323, Singapore. Association for Computational Linguistics.

Md Fahim, Fariha Tanjim Shifat, Fabiha Haider, Deeparghya Dutta Barua, MD Sakib Ul Rahman Sourove, Md Farhan Ishmam, and Md Farhad Alam Bhuiyan. 2024. BanglaTLit: A benchmark dataset for back-transliteration of Romanized Bangla. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14656–14672, Miami, Florida, USA. Association for Computational Linguistics.

Devansh Gautam, Prashant Kodali, Kshitij Gupta, Anmol Goel, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. Comet: Towards code-mixed translation using parallel monolingual sentences. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 47–55.

Abhirut Gupta, Aditya Vavre, and Sunita Sarawagi. 2021. Training data augmentation for code-mixed translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5760–5766.

Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Deeparghya Dutta Barua, Md Sakib Ul Rahman Sourove, Md Fahim, and Md Farhad Alam. 2024. Banth: A multi-label hate speech detection dataset for transliterated bangla. *arXiv preprint arXiv:2410.13281*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

John J Hopfield. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558.

Mohammad Hossin and Md Nasir Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. Association for Computing Machinery.

Qingqing Jiang, Lei Chen, Rui Xu, Xiao Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6279–6284, Hong Kong, China. Association for Computational Linguistics.

Aditya Joshi, Ameya Prabhu, Manish Shrivastava, and Vasudeva Varma. 2016. Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2482–2491, Osaka, Japan. The COLING 2016 Organizing Committee.

Mohsinul Kabir, Obayed Bin Mahfuz, Syed Rifat Raiyan, Hasan Mahmud, and Md Kamrul Hasan. 2023. BanglaBook: A large-scale Bangla dataset for sentiment analysis from book reviews. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1237–1247, Toronto, Canada. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kaitlyn A Litcofsky and Janet G Van Hell. 2017. Switching direction affects switching costs: Behavioral, erp and time-frequency analyses of intra-sentential codeswitching. *Neuropsychologia*, 97:112–139.

Nurul Husna Mahadzir et al. 2021. Sentiment analysis of code-mixed text: a review. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(3):2469–2478.

Soumil Mandal, Sainik Kumar Mahata, and Dipankar Das. 2018. Preparing bengali-english code-mixed corpus for sentiment analysis of indian languages. *ArXiv*, abs/1803.04000.

Soumil Mandal and Anil Kumar Singh. 2018. Language identification in code-mixed data using multichannel neural networks and context capture. *Preprint*, arXiv:1808.07118.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Stian Mæhlum, John Barnes, Lilja Øvrelid, and Erik Velldal. 2019. Annotating evaluative sentences for sentiment analysis: a dataset for norwegian. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 121–130, Turku, Finland. Linkoping University Electronic Press.

Braja Gopal Patra, Dipankar Das, and Amitava Das. 2018. Sentiment analysis of code-mixed indian languages: An overview of sail_code-mixed shared task@ icon-2017. *arXiv preprint arXiv:1803.06745*.

Adithya Pratapa, Monojit Choudhury, and Sunayana Sitaram. 2018. Word embeddings for code-mixed language processing. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3067–3072.

Ela Pustulka-Hunt, Thomas Hanne, Eliane Blumer, and Manuel Frieder. 2018. Multilingual sentiment analysis for a swiss gig. In *2018 6th International Symposium on Computational and Business Intelligence (ISCBI)*, pages 94–98. IEEE.

Md Nishat Raihan, Dhiman Goswami, Antara Mahmud, Antonios Anastasopoulos, and Marcos Zampieri. 2023. Sentmix-3l: A novel code-mixed test dataset in bangla-english-hindi for sentiment analysis. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 79–84.

Nishat Raihan, Dhiman Goswami, Antara Mahmud, Antonios Anastasopoulos, and Marcos Zampieri. 2024. Emomix-3l: A code-mixed dataset for bangla-english-hindi emotion detection. *arXiv preprint arXiv:2405.06922*.

Poonam Rani, Suryakanth Suryawanshi, Koustav Goswami, B. R. Chakravarthi, Tommaso Fransen, and John P. McCrae. 2020. A comparative study of different state-of-the-art hate speech detection methods for hindi-english code-mixed data. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, Marseille, France. European Language Resources Association (ELRA).

Mohd Sanad Zaki Rizvi, Anirudh Srinivasan, Tanuja Ganu, Monojit Choudhury, and Sunayana Sitaram. 2021. Gcm: A toolkit for generating synthetic code-mixed text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 205–211.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Işin Demirşahin, and Keith Hall. 2020. Processing South Asian languages written in the Latin script: the Dakshina dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 2413–2423.

Anna Rogers, Aleksei Romanov, Anna Rumshisky, Svitlana Volkova, Maksim Gronas, and Alexander Gribov. 2018. Rusentiment: An enriched sentiment analysis dataset for social media in russian. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 755–763, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nazanin Sabri, Ali Edalat, and Behnam Bahrak. 2021. Sentiment analysis of persian-english code-mixed texts. In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, pages 1–4. IEEE.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

GM Shahariar Shibli, Md Tanvir Rouf Shawon, Anik Hassan Nibir, Md Zabed Miandad, and Nibir Chandra Mandal. 2023. Automatic back transliteration of romanized bengali (banglish) to bengali. *Iran Journal of Computer Science*, 6(1):69–80.

Sunayana Sitaram and Alan W Black. 2016. Speech synthesis of code-mixed text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3422–3428.

K Sreelakshmi, B Premjith, and KP Soman. 2020. Detection of hate speech text in hindi-english code-mixed data. *Procedia Computer Science*, 171:737–744.

Sara Stefanich, Jennifer Cabrelli, Dustin Hilderman, and John Archibald. 2019. The morphophonology of intraword codeswitching: Representation and processing. *Frontiers in Communication*, 4:54.

Mohammad Tareq, Md Fokhrul Islam, Swakshar Deb, Sejuti Rahman, and Abdullah Al Mahmud. 2023. Data-augmentation for bangla-english code-mixed sentiment analysis: Enhancing cross linguistic contextual understanding. *IEEE Access*, 11:51657–51671.

S Thara and Prabaharan Poornachandran. 2018. Code-mixing: A brief survey. In *2018 International conference on advances in computing, communications and informatics (ICACCI)*, pages 2382–2388. IEEE.

Vladimir N. Vapnik. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. Pos tagging of english-hindi code-mixed social media content. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 974–979.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.