

SentMix-3L: A Bangla-English-Hindi Code-Mixed Dataset for Sentiment Analysis

Md Nishat Raihan*, Dhiman Goswami*, Antara Mahmud

Antonios Anastasopoulos, Marcos Zampieri

George Mason University, USA

{mraihan2, dgoswam, amahmud4, antonis, mzampier}@gmu.edu

Abstract

Code-mixing is a well-studied linguistic phenomenon when two or more languages are mixed in text or speech. Several datasets have been built with the goal of training computational models for code-mixing. Although it is very common to observe code-mixing with multiple languages, most datasets available contain code-mixed between only two languages. In this paper, we introduce SentMix-3L, a novel dataset for sentiment analysis containing code-mixed data between three languages Bangla, English, and Hindi. We carry out a comprehensive evaluation using SentMix-3L. We show that zero-shot prompting with GPT-3.5 outperforms all transformer-based models on SentMix-3L.

1 Introduction

Code-mixing and code-switching are very commonly observed in both text and speech. Code-mixing means the practice of using words from multiple languages within a single utterance, sentence, or discourse, and code-switching refers to the deliberate alteration between multiple languages within the same context (Thara and Poorachandran, 2018). The first case is spontaneous and the second case is purposeful. However, both are widely observed in bilingual and multilingual communities.

According to Anastassiou (2017), several factors are behind these two phenomena, which include social, convenience, linguistic, and cognitive reasons. Socially, this often serves as a sign of group identity which allows individuals to navigate multiple social and cultural affiliations. In terms of linguistics, it is a very common scenario to not be able to find any word for a specific term in one language, whereas another word from another language can help to communicate better. Additionally, there are several cases even in a monolingual community,

when Code-mixing might be the convenient way to express something.

In most occurrences, code-mixing is bilingual. In an early research, Byers-Heinlein and Lew-Williams (2013) states that, it is very likely that by the year 2035, over half of the children enrolled in kindergarten will have grown up speaking a language other than English. Another study conducted by Jeffery and van Beuningen (2020) shows that it is very common in European countries like Germany, Spain, and Italy to use bilingualism in practice. However, in cosmopolitan cities and areas like New York, London, Singapore, and others, code-mixing with three or even more languages is fairly common. This is also observed in countries and states like Luxembourg, Kolkata, and several South-East Indian regions where more than two languages are very commonly used on a daily basis.

Several research works have been conducted on building code-mixed datasets and performing several downstream tasks on such datasets. These datasets include both synthetic and natural ones. However, most of them are bilingual in nature. In this paper, we present SentMix-3L, a Bangla-English-Hindi dataset annotated for sentiment analysis.

The main contributions of our work are as follows:

- We introduce SentMix-3L, a novel three-language code-mixed test dataset with gold standard labels in Bangla-Hindi-English for the task of Sentiment Analysis, containing 1,007 instances.¹
- We provide a comprehensive experimental analysis with several monolingual, bilingual, and multilingual models on SentMix-3L.

We are presenting this dataset exclusively as a test set due to the unique and specialized nature of

*These two authors contributed equally to this work.

¹<https://github.com/LanguageTechnologyLab/SentMix-3L>

the task. Such data is very difficult to gather and requires significant expertise to access. The size of the dataset, while limiting for training purposes, offers a high-quality testing environment with gold-standard labels that can serve as a benchmark in this domain. Given the scarcity of similar datasets and the challenges associated with data collection, SentMix-3L provides an important resource for the rigorous evaluation of text-based models, filling a critical gap in multi-level Code-mixing research. In our experiments, we also prepare a synthetic train and a development dataset to evaluate several models.

2 Related Work

There have been some works conducted on Bangla-English, Hindi-English, and Bangla-Hindi Code-mixing and Code-switching separately. Most of them are case studies and surveys that show the common occurrences of Bangla-English (Alam, 2006; Hasan et al., 2015; Hossain and Bar, 2015; Begum and Haque, 2013; Mahbub-ul Alam and Quyyum, 2016), Hindi-English (Singh, 1985; Bali et al., 2014; Thara and Poornachandran, 2018) and Bangla-Hindi (Ali et al., 2019; Jose et al., 2020) Code-mixing in a wide variety of areas and situations.

Few works are done on sentiment analysis tasks for these types of cases. The work of Khan et al. (2022) presents a Bangla-English Sentiment Analysis dataset primarily related to COVID-19. Their dataset is called CoVaxBD and it contains 1113 samples. Their experiments show that the best result of the dataset is generated by BERT with a development accuracy of 97.3%. However, a lot of the data are purely in Bangla and they only experiment using the BERT and multilingual BERT model (Devlin et al., 2019) while only providing development accuracy as the performance metric. Another recent work by Tareq et al. (2023), consists of 18,074 Bangla-English Code-mixed data from online for the purpose of sentiment analysis. They augment the dataset using their own approach and achieve their best result of an 87% weighted F1 score by implementing XGBoost with Fasttext embedding. Their experiments lack the evaluation of transformer-based models.

An early work by Sitaram et al. (2015) focuses on Hindi-English Code-mixing dataset for Sentiment Analysis. However, it contains 345 data samples in total and only 180 of them are code-mixed.

They get their best results of 91.01% accuracy using Recursive Neural Tensor Network (RNTN). Joshi et al. (2016) compile a dataset of 3879 texts and get their best results using a subword-LSTM approach. Also, the work of (Yadav et al., 2020) includes a dataset of 6357 texts and Bi-LSTM helps them to get their optimal result. However, none of these works present how the transformer-based models perform on their datasets.

In summary, there are no works or datasets on sentiment analysis for code-mixed Bangla-English-Hindi altogether. SentMix-3L is a novel addition in this particular domain of research.

3 The *SentMix-3L* Dataset

In generating the dataset, we choose a controlled data collection method, asking the volunteers to freely contribute data in Bangla, English, and Hindi. This decision stems from several challenges of extracting such specific code-mixed data from the vast corpus available on social media or other online platforms. While the data are not rare, identifying and isolating them from large, unstructured corpora is a very labor-intensive and error-prone process. Our approach ensures data quality and sidesteps the ethical concerns associated with using publicly available online data. Such types of datasets are often used when it is very difficult to mine them from existing corpora. As examples, for fine-tuning LLMs on instructions and conversations, semi-natural datasets like Databricks (2023) and Nie (2023) have become popular.

Data Collection A group of 10 undergraduate students who are fluent in all 3 languages in all four language skills - listening, reading, writing, and speaking. We ask each of them to prepare 250 to 300 social media posts or tweets. They are allowed to use any language including Bangla, English, and Hindi to prepare posts on several daily topics like politics, sports, education, social media rumors, etc. We also ask them to switch languages if and wherever they feel comfortable doing it. The inclusion of emojis, hashtags, and transliteration are also encouraged. The students had the flexibility to prepare the data as naturally as possible. Upon completion of this stage, we filter 1863 samples that contain at least one word or subword from each of the three languages using langdetect (Mazzocchi, 2012) an open-sourced Python tool for language detection.

Data Annotation We annotate the dataset in two steps to prepare high-quality labels for the dataset. First, we recruit three students from social science, computer science, and linguistics as annotators who are also fluent in all 3 languages in all four language skills. They annotate all the 1863 samples with one of the three labels (Positive, Neutral, and Negative) with a raw agreement of 65.3%. We only take these 1182 data, where all three annotators agree on the labels. Second, we gather a second group of annotators consisting of two NLP researchers with the same level of fluency and skills. After their annotation, we calculate a raw agreement of 0.85, a Cohen Kappa score of 0.78 and only keep the data where both annotators agree. After the two stages, we end up with a total of 1007 data.

Label	No. of Data	Percentage
Positive	420	41.71%
Neutral	234	23.24%
Negative	353	35.05%
Total	1,007	100%

Table 1: Label distribution in SentMix-3L.

Dataset Statistics A detailed description of the dataset statistics is provided in Table 2. Since the dataset was generated by people whose first language is Bangla, we observe that the majority portion of the tokens in the dataset is in Bangla. It is also observed that all the data instances contain tokens from all three languages. There are several *Other* tokens in the dataset that are not from Bangla, English, or Hindi language. These tokens contain transliterated words and misspelled words as well as emojis and hashtags. We have Positive, Neutral, and Negative labels for the 1007 data. The label distribution is shown in table 1.

	All	Bangla	English	Hindi	Other
Tokens	89494	32133	5998	15131	36232
Types	19686	8167	1073	1474	9092
Avg	88.87	31.91	5.96	15.03	35.98
Std Dev	19.19	8.39	2.94	5.81	9.70

Table 2: SentMix-3L Data Card. The row *Avg* represents the average number of tokens with its standard deviation in row *Std Dev*.

Synthetic Train and Development Set We present SentMix-3L as a test dataset, hence for experimental purposes, we build a synthetic train

and development set that contains Code-mixing for Bangla, English, and Hindi. We originally take the Amazon Review Dataset (Ni et al., 2019) as seed data and pick 100K data instances randomly. The dataset labels are ratings on a 1 to 5 scale. We convert them into Positive (rating > 3), Neutral (rating = 3), and Negative (rating < 3) for our task. We carefully choose an equal number of instances for Positive, Neutral, and Negative labels. We then use two separate methodologies called *Random Code-mixing Algorithm* by Krishnan et al. (2021) and *r-CM* by Santy et al. (2021) to generate the synthetic Code-mixed dataset.

4 Experiments

Monolingual Models We use 5 monolingual models DistilBERT (Sanh et al., 2019), BERT (Devlin et al., 2019), BanglaBERT (Kowsher et al., 2022), roBERTa (Liu et al., 2019), HindiBERT (Nick Doiron, 2023) for this experiment. Here BanglaBERT is specifically trained on only Bangla and HindiBERT on only Hindi language. The rest of the models are trained in the English language only.

Bilingual Models BanglishBERT (Bhattacharjee et al., 2022) and HingBERT (Nayak and Joshi, 2022) is used as bilingual models which are trained on both Bangla-English and Hindi-English respectively thus effective for the purpose of code mixing tasks including where any two of these languages are involved.

Multilingual Models We use mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) as multilingual models which are respectively trained on 104 and 100 languages. These two models are very effective while we are working in a trilingual Bangla, Hindi and English domain. Moreover, we use IndicBERT (Kakwani et al., 2020) and MuRIL (Khanuja et al., 2021) which covers 12 and 17 Indian languages respectively including Bangla-English-Hindi. Thus these two models and the respective list of languages justify the inclusion of them for our targeted tri-lingual code-mixing task. We also perform hyper-parameter tuning while using all the models to prevent overfitting and ensure optimal F1 score.

Prompting We use prompting with GPT-3.5-turbo model (OpenAI, 2023) from OpenAI for this task. We use the API for zero-shot prompting

(see Figure 1) and ask the model to label the test set.

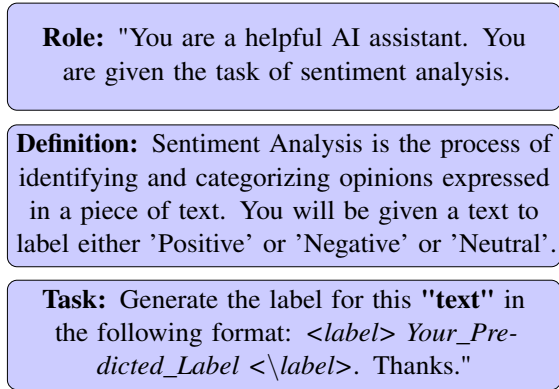


Figure 1: Sample GPT-3.5 prompt.

Additionally, we run the same experiments separately on synthetic and natural datasets splitting both in a 60-20-20 way for training, evaluating, and testing purposes.

5 Results

In this experiment, synthetic data is used as train set, and natural data is used as test set. The F1 scores of monolingual models range from 0.47 to 0.55 where BERT performs the best. Among the two bilingual models BanglishBERT scores 0.56 which is better than HingBERT. XLM-R is the best multilingual model with an F1 score of 0.59. On the other hand, a zero shot prompting technique on GPT 3.5 turbo performs the best with a 0.62 weighted F1 score. These results are available in Table 3.

Models	Weighted F1 Score
GPT 3.5 Turbo	0.62
XLM-R	0.59
BanglishBERT	0.56
mBERT	0.56
BERT	0.55
roBERTa	0.54
MuRIL	0.54
IndicBERT	0.53
DistilBERT	0.53
HindiBERT	0.48
HingBERT	0.47
BanglaBERT	0.47

Table 3: Weighted F-1 score for different models: training on synthetic, testing on natural data.

We perform the same procedure by using synthetic data for both the training and testing where

MuRIL and XLM-R with 0.77 F1 score are the best-performing models. These results are available in Table 4.

Models	Weighted F1 Score
MuRIL	0.77
XLM-R	0.77
mBERT	0.74
BanglishBERT	0.72
BERT	0.68
HingBERT	0.67
DistilBERT	0.66
roBERTa	0.66
IndicBERT	0.66
BanglaBERT	0.60
HindiBERT	0.59

Table 4: Weighted F-1 score for different models: training and testing on synthetic data.

5.1 Error Analysis

We observe *Other* tokens in almost 40% of the whole dataset, as shown in Table 2. These tokens occur due to transliteration which poses a challenge for most of the models since not all of the models are pre-trained on transliterated tokens. BanglishBERT did better than HingBERT since it recognizes both Bangla and English tokens and the total number of tokens for Hindi-English is less than Bangla-English tokens (see Table 3). Also, misspelled words and typos are also observed in the datasets, making the task even more difficult. Some examples are available in Appendix A which are classified wrongly by all the models.

6 Conclusion and Future Work

In this paper, we presented SentMix-3L, a Bangla-English-Hindi code-mixed offensive language identification dataset containing 1,007 instances. We also created 100,000 synthetic data in the same three languages for training. We evaluated various monolingual models on these two datasets. Our results show that prompting GPT3.5 generates the best result on SentMix-3L. When using synthetic data for both training and testing, multilingual models such as mBERT and XLM-R perform well. In the future, we would like to expand SentMix-3L so that it can serve as both training and testing data. Additionally, we are working on pre-training Bangla-English-Hindi trilingual code-mixing models for offensive language identification.

Acknowledgements

We thank the annotators who helped us with the annotation of SentMix-3L. We further thank the anonymous workshop reviewers for their insightful feedback. Antonios Anastasopoulos is generously supported by NSF award IIS-2125466.

Limitations

Although most datasets for the downstream tasks are scraped from social media posts in the real world, in our case these data instances are generated in a semi-natural manner, meaning that they were generated by people but not scraped from social media directly. This was done due to the complexity of extracting contents that contain a specific 3 language code-mixing in them. Also, the dataset is comparatively smaller in size, since it is costly to generate data by a specific set of people who are fluent in all 3 target languages.

Ethics Statement

The dataset introduced in this paper, which centers on the analysis of offensive language in Bangla-English-Hindi code-mixed text, adheres to the [ACL Ethics Policy](#) and seeks to make a valuable contribution to the realm of online safety. SentMix-3L serves as an important resource for opinion analysis of online content. Moreover, the contributors and annotators of the dataset are paid respectable remuneration for their efforts towards this dataset.

References

- Suraiya Alam. 2006. Code-mixing in bangladesh: A case study of non-government white-collar service holders and professionals. *Asian affairs*, 28(4):52–70.
- Muhammad Mooneeb Ali, Mazhar Iqbal Ranjha, and Sartaj Fakhar Jillani. 2019. Effects of code mixing in indian film songs. *Journal of Media Studies*, 31(2).
- Fotini Anastassiou. 2017. Factors associated with the code mixing and code-switching of multilingual children: An overview. *International Journal of Linguistics, Literature and Culture*, 4(3):13–26.
- Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In *Proceedings of CodeSwitch*.
- Most Tasnim Begum and Md Mahmudul Haque. 2013. Code mixing in the ksa: A case study of expatriate bangladeshi and indian esl teachers. *Arab World English Journal*, 4(4).
- 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 ACL*.
- Krista Byers-Heinlein and Casey Lew-Williams. 2013. Bilingualism in the early years: What the science says. *LEARNing landscapes*, 7(1):95.
- 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 ACL*.
- Databricks. 2023. [Dolly 2.0: An open source, instruction-following large language model](#). Accessed: 2023-09-10.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.
- Md Kamrul Hasan, Mohd Akhand, et al. 2015. Reviewing the challenges and opportunities presented by code switching and mixing in bangla. *Journal of Education and Practice*, 6(1):103–109.
- Didar Hossain and Kapil Bar. 2015. A case study in code-mixing among jahangirnagar university students. *International Journal of English and Literature*, 6(7):123–139.
- Jill V Jeffery and Catherine van Beuningen. 2020. Language education in the eu and the us: Paradoxes and parallels. *Prospects*, 48(3-4):175–191.
- Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. 2020. A survey of current datasets for code-switching research. In *Proceedings of ICACCS*.
- 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*.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. IndicNLPsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the ACL*.
- Mahamudur Rahaman Khan, SM Nawsad Rahmatullah, Md Fuadul Islam, Abu Raihan Mostofa Kamal, and Md Azam Hossain. 2022. Sentiment analysis of covid-19 vaccination in bangla language with code-mixed text from social media. In *Proceedings of ICECE*.

- Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. [Muril: Multilingual representations for indian languages](#).
- M Kowsher, Abdullah As Sami, Nusrat Jahan Protasha, Mohammad Shamsul Arefin, Pranab Kumar Dhar, and Takeshi Koshiba. 2022. Bangla-bert: transformer-based efficient model for transfer learning and language understanding. *IEEE Access*, 10:91855–91870.
- Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. 2021. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- Ahmad Mahbub-ul Alam and Shaima Quyyum. 2016. A sociolinguistic survey on code switching & code mixing by the native speakers of bangladesh. *Journal of Manarat International University*, 6(1):8–9.
- Daniele Mazzocchi. 2012. [langdetect: Language detection library](#). Python library.
- Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models. In *Proceedings WILDRE*.
- Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of EMNLP*.
- Nick Doiron. 2023. [hindi-bert \(revision aefac8e\)](#).
- Jianzhi Nie. 2023. [Awesome instruction datasets](#). Accessed: 2023-09-10.
- OpenAI. 2023. [Gpt-3.5 turbo fine-tuning and api updates](#). Accessed: 2023-08-28.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.
- Sebastin Santy, Anirudh Srinivasan, and Monojit Choudhury. 2021. Bertologicomix: How does code-mixing interact with multilingual bert? In *Proceedings of AdaptNLP*.
- Rajendra Singh. 1985. Grammatical constraints on code-mixing: Evidence from hindi-english. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 30(1):33–45.
- Dinkar Sitaram, Savitha Murthy, Debraj Ray, Devansh Sharma, and Kashyap Dhar. 2015. Sentiment analysis of mixed language employing hindi-english code switching. In *Proceedings of ICMLC*.
- 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*.
- S Thara and Prabakaran Poornachandran. 2018. Code-mixing: A brief survey. In *Proceedings of ICACCI*.
- Konark Yadav, Aashish Lamba, Dhruv Gupta, Ansh Gupta, Purnendu Karmakar, and Sandeep Saini. 2020. Bi-lstm and ensemble based bilingual sentiment analysis for a code-mixed hindi-english social media text. In *Proceedings of INDICON*.

A Examples of Misclassified Instances

Positive: জাস্ট গিভ এ পাবলিক স্পিচ এন্ড আই এম ওভার দ্য মুন। এইটা আমার জন্য খুবই বড় একটা ব্যাপার কারণ আমি always been terrified of public speaking But I pushed through the fear and delivered a speech and দর্শকরা খুব ভালভাবে গ্রহণ করেছে। उपलब्धि की भावना अद्भुत है और मुझे खुद पर बहुत गर्व है। yadi mujhase yah ho sakata hai, to tumase bhee ho sakata hai #publicspeaking #overcomingfears

Neutral: पिछले 30 मिनत बिना किसी पष्ठ इयां बताहेते मे उलबे मेरे साशेल मेडिया फेड के मध्यम से क्जेल कराने मे बिताये, अर इमानदारि से कहन त इयां बहत ভাল লাগা। कখনও कখনও আমাদের পৃথিবী থেকে বিচ্ছিন্ন এবং সংযোগে বিচ্ছিন্ন করার জন্য কিছুটা mindless scrolling প্রয়োজন। Of course, it's important to find balance and not let সাশ্যেল মিডিয়া consume too much of our time and energy. आप अपने दैनिक जीवन में सोशल मीडिया के उपयोग को कैसे नेविगेट करते हैं? Nicera mantabya apanara cinta seyara karuna! #Digitaldetox #mindfulness #socialmedia

Negative: ভারাক্রান্ত হৃদয়ে আমি শেয়ার করছি I loss of my beloved mini. দিস বরট সাে মাচ জয়, লাভ, এন্ড হ্যাপিনেস ইনটু মাই লাইফ, এন্ড ইটস অবসেস লিভস আ ভয়েড, কান্নত বে ফিলড উনহঁ খোনা অবিশ্বসনীয় রূপ সে দর্দনাক রহা है, और मैं खुद को दुख से उबरता हुआ पाता हूं। They were more than just a pet; তারা ছিল একটি লালিত সঙ্গী এবং নিঃশর্ত ভালবাসার উৎস। aaie yaadon ko sanjoen, unakee viraasat ka sammaan karen aur hamaare dvaara saajha kie gae pyaar mein taakat paen.