My Boli: Code-mixed Marathi-English Corpora, Pretrained Language Models and Evaluation Benchmarks

Tanmay Chavan[1*], **Omkar Gokhale**[2*], **Aditya Kane**[2*], **Shantanu Patankar**[2*], and **Raviraj Joshi**[3]
Pune Institute of Computer Technology, L3Cube[1]
Georgia Institute of Technology, L3Cube[2]
Indian Institute of Technology Madras, L3Cube[3]
chavantanmay1402@gmail.com, ogokhale3@gatech.edu, adityakane1@gmail.com,
spatankar34@gatech.edu, ravirajoshi@gmail.com

Abstract

The research on code-mixed data is limited due to the unavailability of dedicated code-mixed datasets and pre-trained language models. In this work, we focus on the low-resource Indian language Marathi which lacks any prior work in code-mixing. We present L3Cube-MeCorpus, a large code-mixed Marathi-English (Mr-En) corpus with 10 million social media sentences for pretraining. We also release L3Cube-MeBERT and MeRoBERTa, code-mixed BERT-based transformer models pre-trained on MeCorpus. Furthermore, for benchmarking, we present three supervised datasets MeHate, MeSent, and MeLID for downstream tasks like code-mixed Mr-En hate speech detection, sentiment analysis, and language identification respectively. These evaluation datasets individually consist of manually annotated ~12,000 Marathi-English code-mixed tweets. Ablations show that the models trained on this novel corpus significantly outperform the existing state-of-the-art BERT models. This is the first work that presents artifacts for code-mixed Marathi research. All datasets and models are publicly released at https://github.com/l3cube-pune/MarathiNLP.

1 Introduction

The modern world has been engulfed by the presence of social media platforms like Twitter and Facebook (Salem and Mourtada, 2011). Moreover, websites like YouTube have witnessed considerable user interaction in the comments section of videos (Siersdorfer et al., 2010). These posts and comments closely reflect the thoughts of the general public. It is common for users from different linguistic backgrounds to discuss social, political, and other topics over such social media. This leads to users using a mixed language for communicating over social media platforms.

Code-mixing is known as the mixing of words from multiple languages while retaining the script of a single language. The Latin script is often used to encapsulate the terms of some languages. For example, a given text can be of the Marathi language written in Latin script, as opposed to the Devanagari script, which is the original script of the Marathi language (Joshi, 2022a). Code-mixed data is inherently difficult to process and analyze due to its linguistic complexity, variance in spelling and grammar, and long-tailed distribution of uncommon terms and phrases, which are often specific to the geography and demographics of the source location. It is observed that a large number of tweets, comments, and posts on social media are code-mixed in nature. Thus, with the advent of social media analytics, effectively analyzing code-mixed data has gained the utmost importance.

Marathi is a language which has its origins in Maharashtra, a state in India. Due to the state's geographic and demographic expanse, Marathi has evolved into a language with multiple varieties and dialects. Recently, there has been some focus on Marathi NLP based on the Devanagari script (Joshi, 2022b,a; Kulkarni et al., 2021; Patil et al., 2022; Litake et al., 2022). However, a large chunk of tweets, posts, and comments in Marathi are in code-mixed form. In spite of this, no efforts have been made to curate models and datasets pertaining to Marathi code-mixed data in the past. This work presents the following.

1. We release three supervised datasets (L3Cube-MeHate, MeLID, and MeSent) and one unsupervised dataset (L3Cube-MeCorpus)[1]. The unsupervised corpus comprises 5 million (70.9M tokens) Roman script code-mixed Marathi-English samples compiled from various sources. We further include 5M Devanagari sentences based on original text making it a 10 million (139.5M tokens) mixed-script MeCorpus.

[*] First author, equal contribution

[1] MarathiNLP

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023, pages 242–249

2. The supervised dataset contains labels for code-mixed Marathi-English (Roman script) hate classification, sentiment detection, and language identification. These datasets were manually annotated by native Marathi speakers.

3. Finally, we release a plethora of code-mixed MeBERT-based pre-trained and fine-tuned models for downstream tasks trained on these novel corpora. These models include MeBERT[2], MeBERT-Mixed[3], MeBERT-Mixed-v2[4], MeRoBERTa-Mixed[5], and MeRoBERTa[6]. The models suffixed as 'Mixed' were trained on full 10M MeCorpus while others were trained on 5M Roman MeCorpus. The supervised models include MeSent-RoBERTa[7], MeHate-RoBERTa[8], and MeLID-RoBERTa[9].

This work is a major milestone towards democratizing NLP for the Marathi language. Additionally, we present several ablations with fine-tuned models. This is the first work to present a large unsupervised corpus, multiple pre-trained models, and high-quality supervised datasets. This work is a strong foundation in the domain of Marathi and code-mixed Marathi NLP.

2 Related Works

The use of regional scripts, such as Devanagari, Gurmukhi, Bengali, etc., presents a significant challenge in India due to keyboards primarily designed for the Roman script and the population's familiarity with it. The demand for code-mix datasets and models tailored to regional languages has increased exponentially. These resources play a crucial role in enabling enhanced analysis and moderation of social media content that is code-mixed.

In the realm of language models, BERT-based architectures (Vaswani et al., 2017), including variations such as RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2019), have gained popularity due to their application in pre-training and fine-tuning on various tasks. Multilingual models like multilingual-BERT, XLM-RoBERTa (Conneau et al., 2019), and MuRIL(Khanuja et al., 2021) have specifically focused on data representations that are multilingual and cross-lingual in nature, offering improvements in accuracy and latency. However, these models are pre-trained on less than a hundred thousand real code-mix texts.

While previous research efforts have addressed code mixing in other Indian languages, the specific domain of code-mix Marathi remains largely unexplored. Notably, there is a scarcity of prior work and an absence of a dedicated code-mix Marathi dataset. However, other Indian languages have seen some notable contributions. For instance, Hande et al. (2020) presents KanCMD a code-mixed Kannada dataset for sentiment analysis and offensive language detection. Chakravarthi et al. (2021) and have released datasets encompassing Tamil-English and Malayalam-English code-mixed texts. Nayak and Joshi (2022) have made available HingCorpus, a Hindi-English code-mix dataset, and also open-sourced pre-trained models trained on code-mix corpora. Srivastava and Singh (2021) provide HinGE, a dataset for the generation and evaluation of code-mixed Hinglish text, and demonstrate techniques for algorithmically creating synthetic Hindi code-mixed texts.

In the realm of transliteration, there have been attempts to pre-train language models using transliterated texts. However, these models often underperform due to the rule-based nature of most transliteration techniques, which struggle to account for the diverse spelling variations present in real-life code-mixed texts (Santy et al., 2021).

3 MeCorpus - Pretraining Data Creation

We introduce MeCorpus, a new pre-training corpus of 10 million code-mixed Marathi sentences. This consists of 5M Roman and 5M Devanagari sentences. These sentences are extracted from the social media platforms YouTube and Twitter. We also used synthetic data obtained by transliterating Tweets written in the Devanagari script. The complete data collection process is illustrated in Figure 1.

3.1 Twitter data

A part of the pretraining corpus is obtained from the social networking site Twitter. We utilize `snscrape`, a scraper for social networking sites

[2]MeBERT
[3]MeBERT-Mixed
[4]MeBERT-Mixed-v2
[5]MeRoBERT-Mixed
[6]MeRoBERTa
[7]MeSent-RoBERTa
[8]MeHate-RoBERTa
[9]MeLID-RoBERTa

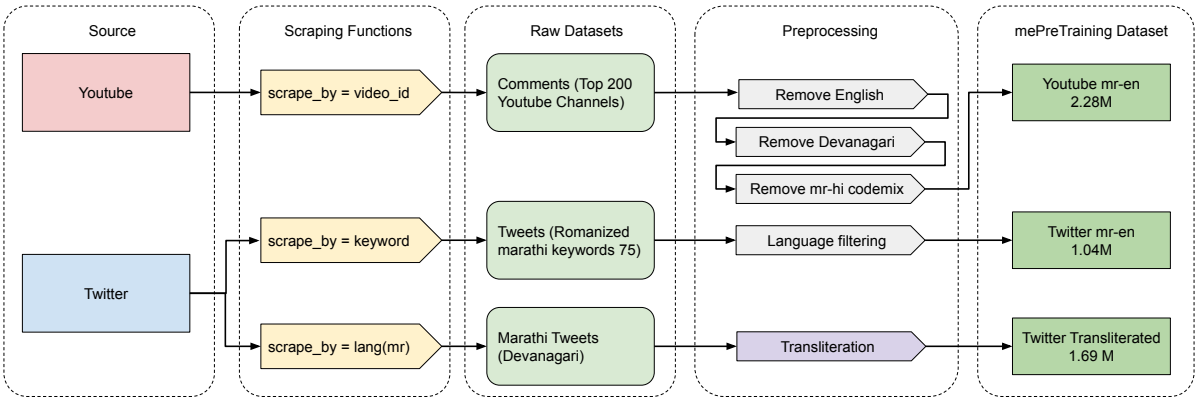

Figure 1: Dataset creation process for our 5 million code-mixed corpora. These 5 million examples are further transliterated to the Devanagari script, which results in a combined corpora of 10 million examples.

to scrape the data from Twitter. We use a keyword-based approach to curate the data. We use frequently used Marathi words as keywords and fetch all the tweets containing the given word. The list of seed keywords is generated by selecting a few common Marathi words and scraping tweets containing these words. Then we identify the most frequently occurring Marathi words in these tweets and add them to our list. The seed words are properly vetted before being added to the list. Proper care is taken to manually check that the word is predominantly exclusive to the Marathi language and doesn't occur in texts from other languages. We fetch a fixed number of tweets belonging to a certain keyword and discard the keyword if the tweets fail to satisfactorily meet the aforementioned conditions after manual verification. Otherwise, we scrape all the tweets containing the keyword and add them to our corpus. This manual verification process ensures that the curated data largely contains Marathi text.

The Twitter data amounts to over a million tweets. All of the tweets at least partly contain code-mixed Marathi. A significant number of the tweets exhibit code-switching between Marathi and English. A small portion of the tweets also contain code-switched Hindi-Marathi text. We anonymize the data before using it for pre-training. The username mentions are replaced with the '@USER' text. We also remove links and hashtags from the tweets. The Twitter corpus contains 50M tokens.

3.2 YouTube data

YouTube comments are an excellent source of Code-mixed Marathi data. We scrape all the comments from 200 Marathi YouTube channels using the `youtube-comments-scraper` library. This gave us a mix of English, Devanagari, and code-mixed Marathi sentences. We then removed the Devanagari and English comments to obtain the code-mixed Marathi data. This data was then pre-processed and used in our pre-training dataset. Devanagari words were identified and removed by checking their utf-8 encoding. We remove all comments that have more than 80% Devanagari words. This gives us comments that are either English or Marathi-English code-mixed. We used a fast text classifier to identify sentences that are English. We removed these sentences and were left with code-mixed Marathi sentences. Thus at the end of both filtering steps, we are left with 2,278,097 of the original 7,599,588 comments.

Source	Number of Sentences
Twitter tweets	1,037,659
Youtube comments	2,277,108
Transliterated-tweets	1,685,233
Total	**5,000,000**

Table 1: Source-wise split of the Roman script pretraining corpus (5M).

3.3 Transliteration

We scraped around 1.7 million Marathi Devanagari tweets from Twitter using the snscrape library. We then used the indic-trans Python library to transliterate 1,685,233 of these Devanagari tweets to Code-mixed Marathi and added them to our dataset.

4 Me Corpus - Transliteration

We created an additional 5 million Devanagari sentences and added them to the corpus. This was

done by transliterating the Tweets and YouTube comments mentioned in sections 3.1 and 3.2 respectively to the Devanagari script. We also used the scraped Devanagari tweets mentioned in section 3.3. This gave us a total of 5M Devanagari sentences. These 5M sentences were used to pre-train the multilingual models mentioned in the future sections.

5 MeEval - Downstream dataset creation

We aim to create a large dataset of code-mix Marathi-English data, annotated with sentiment and hatefulness. In this study, we selected a set of tweets from a larger corpus of 1,037,659 tweets obtained from the social media platform Twitter. Half of the tweets chosen were posted on Twitter before 2013, and the other half were posted after 2013. This helped to provide a more diachronic distribution of tweets, as the number of tweets posted in the past few years far outnumber the old tweets. The tweets were selected randomly apart from this criteria. This ensures a realistic representation of the sentiment, hate, and profanity distributions present in the real-world data across the past several years. For data annotation, we selected four annotators who are proficient in Marathi, Hindi, and English languages. All four annotators are native Marathi speakers and hold undergraduate-level proficiency in English. Any discrepancies discovered within the dataset were systematically resolved through collaborative consensus among the annotators.

The collected data was labeled according to three distinct categories: sentiment, hate, and language identification. We followed a set of guidelines while labeling the data to ensure the veracity of the annotation. We annotated the data after anonymizing it. This helped remove any bias or knowledge of the entity posting it. We also disregarded any additional information that could be inferred by us based on external context but is not apparent by reading the text by itself. Here, we outline the dataset statistics and annotation procedure. The dataset statistics are described in Table 3. A few annotated examples are presented in Table 2.

5.1 MeSent Dataset

The code-mixed Marathi-English sentiment data is termed the MeSent Dataset. Tweets expressing good or heartening emotions such as thankfulness, happiness, applause, and appreciation are labeled positive. Tweets expressing negative or disheartening emotions like strong dissent, disappointment, sorrow, derision, and hate are labeled negative. Plain facts, statements, and simple responses are labeled neutral. If a tweet contains conflicting emotions, the stronger emotion is chosen.

- +1 indicating a *positive* sentiment,
- -1 indicating a *negative* sentiment, and
- 0 indicating a *neutral* sentiment.

While annotating the data, we removed unsuitable and ambiguous tweets. Finally, we selected 4,000 tweets from each sentiment category, leading to the dataset containing a total of 12,000 tweets.

5.2 MeHate Dataset

For the hatefulness annotation, we labeled any tweets expressing strongly negative feelings such as insults, mockery, abuse, intimidation, and threats as *hateful*. Any tweet not containing such hateful content is labeled as *non-hateful*. We use 1 for hateful content and 0 for non-hateful content. The MeHate dataset contains 1384 hateful and 1384 non-hateful tweets, totaling 2768 tweets. We also release the full 12k labeled tweets with the majority of non-hate labels.

5.3 MeLID dataset

Additionally, a Language Identification (LID) dataset is created. Each word within the selected tweets is labeled based on its language as *Marathi*, *English*, or *Other*. The *Other* category contains invalid words, words from languages other than English or Marathi, and literals such as numbers and proper nouns. The MeLID dataset contains 11,814 tweets. For all three supervised datasets, we provide a pre-defined train, test, and validation split of 80:10:10.

6 Models trained on code-mixed MeCorpus

We train several well-known models on our novel pretraining corpora. In this section, we outline these models and their training details.

We used pre-trained BERT, RoBERTa, mBERT, MuRIL, and XLM Roberta as the base models and trained them on the novel MeCorpus using the Masked Language Modelling (MLM) objective. For MLM training, we train the models for two

Example Text	Hate (MeHate Dataset)	Sentiment (MeSent Dataset)	LID (MeLID Dataset)
Good morning sir mast tumhala bhetayche ahe	0 (Non-hateful)	1 (Positive)	["Good" : ENG, "morning" : ENG, "sir" : ENG, "mast" : MAR, "tumhala" : MAR, "bhetayche" : MAR, "ahe" : MAR]
nalayak jalavu nakos amhala	1 (Hateful)	-1 (Negative)	["nalayak" : MAR, "jalavu" : MAR, "nakos" : MAR, "amhala" : MAR]
Ssc cha decision lavkar ghya	0 (Non-hateful)	0 (Neutral)	["Ssc" : OTH, "cha" : MAR, "decision" : ENG, "lavkar" : MAR, "ghya" : MAR]

Figure 2: Table containing code-mixed examples with their annotations.

epochs at a learning rate of $2e-5$, with a weight decay of 0.01 and a mask probability of 0.15.

The monolingual models were pre-trained on the Roman 5M codemixed data mentioned in section 3. While the multilingual models were trained on the full 10M corpus (5M Roman + 5M Devanagari sentences) mentioned in section 4. The models pre-trained on mixed-script corpus are suffixed as 'Mixed'.

The resulting models were named similarly to the original models, prefixed with "**me**", which stands for **M**arathi-**E**nglish. Therefore, the models MeBERT, MeBERT-Mixed, MeBERT-Mixed-v2, MeRoBERTa-Mixed, and MeRoBERTa are the BERT, mBERT, MuRIL, XLM-RoBERTa, and RoBERTa models trained on the MeCorpus respectively. Note that these models are further "fine-tuned" on the MeCorpus using the MLM training objective.

7 Results

We fine-tune our MeBERT models on the MeSent, MeHate, and MeLID datasets as mentioned in section 5 and test them on the respective test data. The same process is repeated for their base models and a few state-of-the-art Marathi models like Indic-BERT (Kakwani et al., 2020), Marathi-Tweets-BERT (Gokhale et al., 2022), and Marathi Code-mixed Abusive MuRIL (Das et al., 2022). The results obtained from this are showcased in Table 2. It is observed that MeBERT-Mixed-v2 outperforms all other models on the MeHate evaluation set with an F1 score of 78.3%. For the sentiment analysis corpus MeSent, MeRoBERTa outperforms the others by obtaining an F1 score of 67.27%. Testing the models on the MeLID dataset, MeBERT-Mixed-v2 outperforms the other models by obtaining an F1 score of 88.6%. The newly pre-trained code-mixed MeBERT-based models consistently outperform their base models as well as the state-of-the-art Marathi models.

Model	**MeHate**	**MeSent**	**MeLID**
Indic-BERT	61.62	55.29	87.37
MahaTweets-BERT	63.18	57.59	87.38
Abusive-MuRIL	67.69	-	-
BERT	61.98	59.06	87.89
MeBERT	73.78	61.92	88.01
mBERT	66.80	60.25	87.64
MeBERT-Mixed	77.39	65.73	88.25
MuRIL	67.93	63.38	87.55
MeBERT-Mixed-v2	**78.3**	64.23	**88.6**
XLM-RoBERTa	64.66	61.06	87.54
MeRoBERTa-Mixed	78.07	67.17	87.42
RoBERTa	66.10	58.86	86.46
MeRoBERTa	77.85	**67.27**	88.41

Table 2: Macro F1 scores (in %) of models on the MeHate, MeSent, and MeLID datasets.

8 Conclusion

This work lays the necessary groundwork for future work on code-mixed Marathi. We introduce a novel pretraining corpus of 5 million code-mixed text examples. In addition to that, we present five new models trained on this code-mixed corpus. Furthermore, we present three supervised datasets of 12,000 tweets for hate classification, sentiment analysis, and language identification annotated by native Marathi speakers. We also present thorough ablations and show that our code-mixed MeBERT models outperform the previous state-of-the-art models by a considerable margin.

Limitations

A major problem while dealing with Romanized Marathi is the lack of a singular correct spelling of words. A Marathi word can be written in several ways in Marathi, all of which are equally valid and correctly convey meaning despite having significantly different spellings. Developing efficient approaches to tackle this issue will lead to a significant increase in performance on NLP tasks dealing with code-mixed languages. Our keyword-based scraping method uses words primarily from the western Maharashtra dialect of Marathi, which might not sufficiently represent samples from other Marathi dialects. Efforts to increase the dataset to include examples from other dialects will make the dataset more diverse and robust.

Ethics Statement

All of the data used in our experiments has been scraped by legal and valid means, adhering to the provided guidelines. We anonymized the data before usage to protect the privacy of the original authors of the data. This data might contain biases and thus must be used with care. This data also contains strong language which might be unsuitable for some applications. This data should be used only for research purposes and not for training any model for deployment.

Acknowledgments

This work was done under the L3Cube Pune mentorship program. We would like to express our gratitude towards our mentors at L3Cube for their continuous support and encouragement. This work is a part of the L3Cube-MahaNLP project (Joshi, 2022b).

References

Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Ratnasingam Sakuntharaj, Anand Kumar Madasamy, Sajeetha Thavareesan, Premjith B, Subalalitha Chinnaudayar Navaneethakrishnan, John P. McCrae, and Thomas Mandl. 2021. Overview of the HASOC-DravidianCodeMix Shared Task on Offensive Language Detection in Tamil and Malayalam. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation*. CEUR.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. Data bootstrapping approaches to improve low resource abusive language detection for indic languages. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, pages 32–42.

Omkar Gokhale, Aditya Kane, Tanmay Chavan, Shantanu Patankar, and Raviraj Joshi. 2022. Spread love not hate: Undermining the importance of hateful pretraining for hate speech detection. *arXiv preprint arXiv:2210.04267*.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. Kancmd: Kannada codemixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63.

Raviraj Joshi. 2022a. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 97–101.

Raviraj Joshi. 2022b. L3cube-mahanlp: Marathi natural language processing datasets, models, and library. *arXiv preprint arXiv:2205.14728*.

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 Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

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.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. 2021. L3cubemahasent: A marathi tweet-based sentiment analysis dataset. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 213–220.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Onkar Litake, Maithili Ravindra Sabane, Parth Sachin Patil, Aparna Abhijeet Ranade, and Raviraj Joshi. 2022. L3cube-mahaner: A marathi named entity recognition dataset and bert models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 29–34.

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. *arXiv preprint arXiv:1907.11692*.

Ravindra Nayak and Raviraj Joshi. 2022. L3cube-hingcorpus and hingbert: A code mixed hindi-english dataset and bert language models. *arXiv preprint arXiv:2204.08398*.

Hrushikesh Patil, Abhishek Velankar, and Raviraj Joshi. 2022. L3cube-mahahate: A tweet-based marathi hate speech detection dataset and bert models. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 1–9.

Fadi Salem and Racha Mourtada. 2011. Civil movements: The impact of facebook and twitter. *Arab Social Media Report*, 1.

Sebastin Santy, Anirudh Srinivasan, and Monojit Choudhury. 2021. BERTologiCoMix: How does code-mixing interact with multilingual BERT? In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 111–121, Kyiv, Ukraine. Association for Computational Linguistics.

Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. 2010. How useful are your comments? analyzing and predicting youtube comments and comment ratings. pages 891–900.

Vivek Srivastava and Mayank Singh. 2021. Hinge: A dataset for generation and evaluation of code-mixed hinglish text. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 200–208.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

A Appendix

Dataset	Sample Count	Average word count	Labels	Count per label
MeLID	11,813	12	Marathi	99,230
			English	26,290
			Other	10,870
MeSent	12,000	16	Positive	4,000
			Neutral	4,000
			Negative	4,000
MeHate	2,768	17	Non Hate	1,384
			Hate	1,384

Table 3: Statistics for benchmark datasets