MLInitiative@WILDRE7: Hybrid Approaches with Large Language Models for Enhanced Sentiment Analysis in Code-Switched and Code-Mixed Texts

Hariram Veeramani[1], Surendrabikram Thapa[2], Usman Naseem[3]
[1]UCLA, USA [2]Virginia Tech, USA [3]Macquarie University, Australia
[1]hariramveeramani@gmail.com, [2]sbt@vt.edu, [3]usman.naseem@mq.edu.au

Abstract

Code-switched and code-mixed languages are prevalent in multilingual societies, reflecting the complex interplay of cultures and languages in daily communication. Understanding the sentiment embedded in such texts is crucial for a range of applications, from improving social media analytics to enhancing customer feedback systems. Despite their significance, research in code-mixed and code-switched languages remains limited, particularly in less-resourced languages. This scarcity of research creates a gap in natural language processing (NLP) technologies, hindering their ability to accurately interpret the rich linguistic diversity of global communications. To bridge this gap, this paper presents a novel methodology for sentiment analysis in code-mixed and code-switched texts. Our approach combines the power of large language models (LLMs) and the versatility of the multilingual BERT (mBERT) framework to effectively process and analyze sentiments in multilingual data. By decomposing code-mixed texts into their constituent languages, employing mBERT for named entity recognition (NER) and sentiment label prediction, and integrating these insights into a decision-making LLM, we provide a comprehensive framework for understanding sentiment in complex linguistic contexts. Our system achieves competitive rank on all subtasks in the Code-mixed Less-Resourced Sentiment analysis (Code-mixed) shared task at WILDRE-7 (LREC-COLING).

Keywords: Code-switched language, Code-switched language, Sentiment analysis, Named entity recognition (NER), Large language models (LLMs)

1. Introduction

Informal communication constitutes a significant proportion of short text communications and online posts in our digital world (Tay, 1989). People tend to express themselves freely and spontaneously through various online platforms, ranging from social media to messaging apps. While some individuals stick to a single language when communicating, the use of two or more languages is also very common in informal communication. This phenomenon of code-mixing—mixing two or more languages within a single utterance—is common, especially in regions where closely related languages coexist (Thara and Poornachandran, 2018).

In code-mixing, individuals incorporate elements of different languages within their communication. This incorporation may occur for a multitude of reasons, including cultural affinity, linguistic convenience, or social dynamics (Lamabam and Chakma, 2016; Barman et al., 2014). For instance, individuals may switch between languages based on their proficiency, the context of the conversation, or the preferences of the people with whom they are conversing. By doing so, speakers can interact with ease and convey their intended messages more accurately. In non-English speaking and multilingual countries, code mixing is particularly prevalent due to the coexistence of multiple languages within the same socio-cultural space (Pratapa et al., 2018).

With a lot of code-mixed languages used, there is a need to automatically detect the sentiment of such code-mixed text important for various reasons (Kodali et al., 2022). Firstly, it allows for a deeper understanding of the emotions and opinions expressed by individuals in multilingual contexts. By accurately identifying sentiment, researchers and analysts can gain insights into the attitudes, preferences, and behaviors of diverse language communities. Secondly, sentiment analysis of code-mixed text enables businesses and organizations to effectively monitor and analyze customer feedback, social media trends, and public opinion in linguistically diverse markets. This information can inform marketing strategies, product development decisions, and customer relationship management efforts tailored to specific language communities (Joshi et al., 2016).

Moreover, sentiment analysis in code-mixed text can contribute to the development of more inclusive and culturally sensitive natural language processing (NLP) technologies (Chakravarthi et al., 2020). By recognizing and accounting for the linguistic nuances and variations present in multilingual communications, NLP models can better serve diverse user populations and facilitate more accurate language understanding and generation. Additionally, automatic sentiment detection in code-mixed text has implications for social and political analysis. By

WILDRE-7 Workshop, pages 66–72

analyzing sentiment patterns across different language groups, researchers can uncover insights into socio-political dynamics, cultural trends, and community sentiments, aiding in areas such as public policy formulation, cross-cultural communication, and conflict resolution.

In the seventh Workshop on Indian Language Data: Resources and Evaluation (WILDRE), a shared task on Code-mixed Less-Resourced Sentiment analysis was launched to address this issue. This shared task focuses on sentiment analysis in code-mixed data from less-resourced similar languages, particularly in language pairs and triplets of closely related Indo-Aryan languages spoken in eastern India. These languages include Magahi, Maithili, Bangla, and Hindi, along with English. The task aims to explore different machine learning and deep learning approaches to train models robust enough to perform well on the given training and validation datasets, thus providing insights into language representation and speakers' preferences in code-mixed settings. In this paper, we present our system description for this shared task. In our approach, we leverage named entity recognition, language decomposition, and large language models.

2. Related Works

The exploration of sentiment analysis in code-mixed text has been a subject of growing interest within the field of natural language processing (NLP), particularly due to the challenges and complexities associated with understanding and processing multilingual text. Several studies have laid the groundwork for addressing these challenges, providing valuable insights and methodologies for future research.

Pednekar and Saravanan (2023) addresses the scarcity of resources for sentiment analysis (SA) in mixed code languages by proposing the creation of a gold standard dataset. Their research aims to advance SA in underrepresented languages, highlighting the importance of high-quality datasets for evaluating SA models in languages with diverse code-mixing patterns (Pednekar and Saravanan, 2023).

Early work in the domain focused on identifying the occurrence and patterns of code-mixing across different linguistic contexts. A seminal study by Solorio and Liu (2008b) explored part-of-speech tagging for code-switched (a form of code-mixing) data. Their research highlighted the need for tailored NLP tools that can accurately process and understand the grammatical structures of mixed-language texts (Solorio and Liu, 2008a,b). Building upon these foundational insights, subsequent research has ventured into the sentiment analysis of code-mixed texts. For example, Joshi et al. (2016) developed algorithms that harness code-switching to improve sentiment analysis in bilingual text corpora. Their work underscored the potential benefits of leveraging the linguistic features inherent to code-switching for more nuanced sentiment detection.

In an effort to specifically address the challenges posed by code-mixed text in Indian languages, Barman et al. (2014) investigated code-mixing on Indian social media platforms. They created a corpus of code-mixed text and developed classification models that significantly improved the understanding of sentiment within these multilingual datasets.

The complexity of code-mixing and its implications for sentiment analysis have also been explored through competitions and shared tasks. For instance, the shared task on Sentiment Analysis for Dravidian Languages in Code-Mixed Text, as part of the Forum for Information Retrieval Evaluation (FIRE), has provided a platform for researchers to apply and evaluate various computational models on code-mixed datasets, leading to significant advancements in the field (Chakravarthi et al., 2020).

Similarly, Tho et al. (2020) provides a systematic literature review on code-mixed sentiment analysis using machine learning approaches. Their findings suggest that Support Vector Machine, Naïve Bayes, and Logistic Regression are the most common classifiers for this task, with Support Vector Machine exhibiting superior performance based on accuracy and F1 scores (Tho et al., 2020). Jin et al. (2023) offers a comprehensive review of text sentiment analysis methods and applications, exploring a variety of feature extraction and representation methods, including deep learning-based approaches. This review serves as a foundation for understanding the current status and development trends in SA (Jin et al., 2023). Similarly, Zucco et al. (2020) present a detailed study on sentiment analysis (SA) tools and methods for mining texts and social network data. Their analysis, based on objective criteria, highlights the importance of developing more advanced SA tools to enhance end-user experience (Zucco et al., 2020). Moreover, Habimana et al. (2020) review deep learning approaches for various SA tasks, suggesting that the future of SA models could benefit from incorporating advanced techniques such as BERT, sentiment-specific word embedding models, and attention mechanisms (Habimana et al., 2020).

3. Task Descriptions

We only participate in the first shared task, i.e. Code-mixed less-resourced sentiment analysis. This shared task aimed to address the complexities of sentiment analysis in code-mixed data from less-resourced similar languages, with a focus

on Magahi-Hindi-English, Maithili-Hindi, Bangla-English-Hindi, and Hindi-English language pairs and triplets. These languages, belonging to the Indo-Aryan language family and predominantly spoken in eastern India, present unique challenges due to their linguistic similarities and low-resource settings. An important aspect of this shared task was the introduction of an unlabelled test dataset for the code-mixed Maithili language (Maithili-Hindi-English) (Rani et al., 2024b). Participants were challenged to leverage the available training datasets from Magahi-Hindi-English, Maithili-Hindi, Bangla-English-Hindi, and Hindi-English to determine the sentiment of comments in this target language.

Participants were tasked with exploring different machine learning and deep learning approaches to train models on the training and validation data sets provided. The goal was to develop models robust enough to perform well on code-mixed language datasets, thus enhancing sentiment analysis capabilities in multilingual contexts.

3.1. Evaluation

The shared task on CodaLab employed standard evaluation metrics, primarily the average F1 score, to assess participating teams' models. The evaluation also included precision, recall, and F1 scores across sentiment classes for detailed analysis. Initially, teams accessed training and validation data, with test data and the Maithili test set later released. Two tracks were defined: one for determining polarity in code-mixed settings and another for sentiment analysis in code-mixed Maithili. Datasets were divided into train, validation, and test sets, with a 70:15:15 ratio. For the combined language pairs, training and validation sets were merged. Submitted models were evaluated based on their ability to predict sentiment labels on test data. Results, including F1 scores, precision, and recall, were provided to teams for analysis and discussion, offering insights into code-mixed sentiment analysis challenges and solutions.

4. Dataset

The dataset provided for the shared task comprised annotated code-mixed text in three language pairs: Magahi-Hindi-English, Bangla-English-Hindi, and Hindi-English. Each comment or sentence in the Magahi-Hindi-English and Hindi-English datasets was labeled with four sentiment categories: Positive, Negative, Neutral, or Mixed (Rani et al., 2024a). In contrast, the Bangla-English-Hindi dataset was labeled with three sentiment categories: Positive, Negative, or Neutral.

The Magahi-Hindi-English and Hindi-English datasets were collected from various YouTube channels and meticulously annotated with the assistance of native speakers of the respective languages. This ensured that the data accurately reflected the nuances of sentiment expression in these code-mixed contexts. Additionally, for the Bangla-English-Hindi dataset, the SentMix-3L dataset by Raihan et al. (2023) was utilized. This dataset provided a rich collection of code-mixed text in Bangla, English, and Hindi, offering valuable insights into sentiment analysis in a multilingual context.

Participants in the shared task were allowed to leverage external resources, provided they were openly available and could be used by other participants for research purposes. Proper citation and detailed information about any external dataset utilized were included in the system description paper submitted by participants. Overall, the dataset offered a diverse collection of code-mixed text across different language pairs and sentiment categories, enabling participants to develop and evaluate models robust enough to handle sentiment analysis in code-mixed data from less-resourced similar languages.

5. System Description

Our system for sentiment analysis in code-mixed texts employs a multi-step approach, leveraging the capabilities of large language models (LLMs) and the multilingual BERT (mBERT) model to accurately process and analyze sentiment in less-resourced, code-mixed language data. As shown in Figure 1, our methodology consists of the following steps:

5.1. Decomposition of Code-Mixed Language into Individual Languages

The first step in our approach involves decomposing the code-mixed language data into its constituent languages. This process is crucial for handling the intricacies of code-mixed texts and allows for more accurate subsequent analysis. We utilize three prominent LLMs: Mistral, LLama (Touvron et al., 2023), and Gemma, to perform this decomposition. By prompting these models with the specific languages present in the code-mixed text, as illustrated in Figure 1, we effectively separate Hindi-English code-mixed language into individual Hindi and English components. This decomposition aids in the further processing and understanding of the text.

5.2. Finetuning mBERT

We use mBERT for two major objectives: Named Entity Recognition (NER) and label prediction.

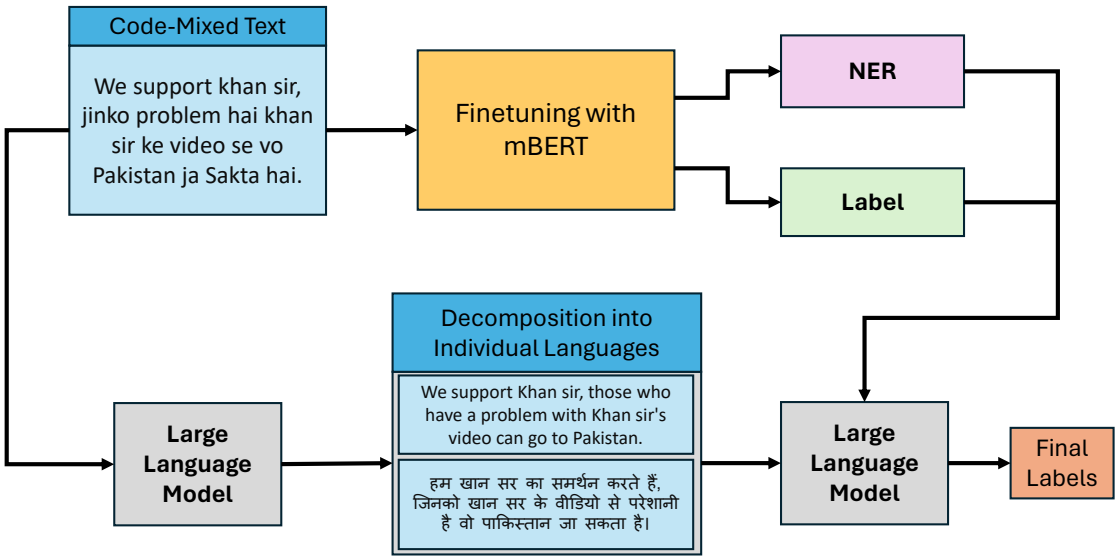

Figure 1: System Description of Our Approach

5.2.1. Use of mBERT for NER

For NER: We employ mBERT (Devlin et al., 2019), a model pre-trained on multiple languages, to perform NER on the decomposed language texts. NER is instrumental in identifying key entities within the text, providing valuable context that enhances the sentiment analysis process (Li et al., 2020). mBERT's ability to understand multiple languages makes it particularly suited for this task, enabling accurate entity recognition in all language components of the code-mixed text.

5.2.2. For Label Prediction

Additionally, we leverage mBERT for the prediction of sentiment labels in the test dataset. By fine-tuning mBERT with our training data, we are able to classify the sentiment of code-mixed texts effectively. The fine-tuned mBERT model is then used for inference on the test data, predicting the sentiment labels with a high degree of accuracy.

5.3. Large Language Models for Final Decision

In the final step of our approach, we integrate the outputs from the previous steps—including the NER results, sentiment labels from mBERT, and the decomposed language components—into a comprehensive input for a large language model. This LLM is tasked with making the final sentiment analysis decision. By providing the LLM with a holistic view of the text, including both the original code-mixed form and the derived insights from mBERT and language decomposition, we enable it to leverage all available information for conflict resolution and final sentiment classification. This step is crucial for resolving any discrepancies between the sentiment labels predicted by mBERT and the nuances captured through NER and language decomposition, ensuring a cohesive and accurate sentiment analysis outcome.

This multi-step approach leverages the complementary strengths of LLMs and mBERT, facilitating a nuanced and effective analysis of sentiment in code-mixed texts, particularly in the context of less-resourced languages. Through this methodology, we address the challenges posed by code-mixing and provide insights into the sentiments expressed in multilingual communities.

6. Results

Our participation in the first shared task—Code-mixed less-resourced sentiment analysis—yielded notable results across various language combinations, including Bangla-English, Hindi-English, Magahi-Hindi-English, and Maithili-Hindi-English. We evaluated our model's performance using three distinct Large Language Models (LLMs): Mistral, Llama, and Gemma, across the criteria of macro-averaged F1 score, precision, and recall. Table 1 summarizes our findings.

In the Bangla-English combination, Mistral outperformed its counterparts with a macro-averaged F1 score of 0.67, precision of 0.76, and recall of 0.68. This result indicates Mistral's superior capability in handling the intricacies of Bangla-English code-mixed texts. For the Hindi-English dataset, Gemma led with a macro-averaged F1 score of

Language Combination	LLM	Macro-Averaged F1	Macro-Averaged Precision	Macro-Averaged Recall
Bangla-English	Mistral	**0.67**	**0.76**	**0.68**
	Llama	0.34	0.58	0.41
	Gemma	0.64	0.66	0.63
Hindi-English	Mistral	0.31	**0.38**	**0.47**
	Llama	0.28	0.36	0.32
	Gemma	**0.34**	0.35	0.39
Magahi-Hindi-English	Mistral	0.23	**0.39**	**0.39**
	Llama	0.21	0.29	0.19
	Gemma	**0.26**	0.28	0.27
Combined*	Mistral	0.33	**0.40**	**0.38**
	Llama	0.26	0.33	0.28
	Gemma	**0.35**	0.36	0.36
Maithili-Hindi-English	Mistral	0.13	0.26	0.27
	Llama	**0.35**	**0.36**	**0.36**
	Gemma	0.24	0.26	0.31

Table 1: Performance of our model with different datasets. *The combined language represents Bangla-English, Hindi-English and Magahi-Hindi-English datasets altogether.

0.34, albeit Mistral showed better performance in terms of precision and recall. This suggests that while Gemma was more effective overall, Mistral was better at identifying relevant instances, albeit with a higher rate of false positives.

Magahi-Hindi-English texts, which represent a more challenging setting due to their triple-language mix, saw Gemma performing the best in terms of the F1 score. However, Mistral consistently showed higher precision and recall, indicating its effectiveness in accurately classifying sentiments in this complex language mix. When evaluating the combined dataset, which includes all language pairs, Gemma again demonstrated the highest F1 score, highlighting its robustness across multiple code-mixed settings. Mistral, however, maintained higher precision and recall scores, reinforcing its efficiency in identifying sentiment with greater accuracy.

Notably, the Maithili-Hindi-English combination presented a unique challenge. In this case, Llama achieved the best performance across all metrics, underscoring its effectiveness in dealing with the code-mixed Maithili language. This performance emphasizes the potential of LLMs in addressing sentiment analysis in less-explored language combinations. These results highlight the ability of our approach in leveraging LLMs for sentiment analysis in code-mixed texts. The varied performance across different models and language combinations shows the importance of model selection based on the specific linguistic characteristics of the target data. Our findings contribute to the broader understanding of sentiment analysis in multilingual contexts, especially within less-resourced languages.

7. Conclusion

In conclusion, our exploration of sentiment analysis in code-mixed and code-switched texts across less-resourced languages demonstrates the significant potential of leveraging Large Language Models (LLMs) such as Mistral, Llama, and Gemma. Our methodology, which intricately combines language decomposition, named entity recognition (NER), and sentiment classification, showcases a novel approach to navigating the complexities inherent in multilingual sentiment analysis. The results across various language combinations underscore the importance of model and technique selection tailored to the specific challenges posed by each language mix. Through this work, we not only contribute to the understanding of sentiment analysis in the context of code-mixed and code-switched languages but also highlight the importance of developing NLP tools that are inclusive of linguistic diversity. Our findings pave the way for future research to further refine these methods and expand the scope of sentiment analysis in multilingual and multicultural societies, ensuring that NLP technologies remain responsive to the nuances of human language and emotion.

References

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.

Bharathi Raja Chakravarthi, Ruba Priyadharshini,

Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P McCrae. 2020. Overview of the track on sentiment analysis for dravidian languages in code-mixed text. In *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 21–24.

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.

Olivier Habimana, Yuhua Li, Ruixuan Li, Xiwu Gu, and Ge Yu. 2020. Sentiment analysis using deep learning approaches: an overview. *Science China Information Sciences*, 63:1–36.

Yuxin Jin, Kui Cheng, Xinjie Wang, and Lecai Cai. 2023. A review of text sentiment analysis methods and applications. *Frontiers in Business, Economics and Management*, 10(1):58–64.

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.

Prashant Kodali, Anmol Goel, Monojit Choudhury, Manish Shrivastava, and Ponnurangam Kumaraguru. 2022. Symcom-syntactic measure of code mixing a study of english-hindi code-mixing. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 472–480.

Priyadarshini Lamabam and Kunal Chakma. 2016. A language identification system for code-mixed english-manipuri social media text. In *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, pages 79–83. IEEE.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70.

Chaitanya B Pednekar and P Saravanan. 2023. A study on different methods in sentiment analysis from text. In *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 1115–1122. IEEE.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553.

Md Nishat Raihan, Dhiman Goswami, Antara Mahmud, Antonios Anstasopoulos, and Marcos Zampieri. 2023. Sentmix-3l: A bangla-english-hindi code-mixed dataset for sentiment analysis. *arXiv preprint arXiv:2310.18023*.

Priya Rani, Gaurav Negi, Theodorus Fransen, and John P. McCrae. 2024a. Macms: Magahi code-mixed dataset for sentiment analysis. *arXiv preprint arXiv:2403.04639*.

Priya Rani, Gaurav Negi, Saroj Jha, Shardul Suryawanshi, Atul Kr. Ojha, Paul Buitelaar, and John P. McCrae. 2024b. Findings of the wildre shared task on code-mixed less-resourced sentiment analysis for indo-aryan languages. In *Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation @LREC-COLING-2024 (WILDRE-7)*, Turin, Italy. ELRA Language Resources Association (ELRA) and the International Committee on Computational Linguistics (ICCL).

Thamar Solorio and Yang Liu. 2008a. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981.

Thamar Solorio and Yang Liu. 2008b. Part-of-speech tagging for english-spanish code-switched text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060.

Mary WJ Tay. 1989. Code switching and code mixing as a communicative strategy in multilingual discourse. *World Englishes*, 8(3):407–417.

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.

Cuk Tho, Harco Leslie Hendric Spits Warnars, Benfano Soewito, and Ford Lumban Gaol. 2020. Code-mixed sentiment analysis using machine learning approach–a systematic literature review. In *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*, pages 1–6. IEEE.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open

and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Chiara Zucco, Barbara Calabrese, Giuseppe Agapito, Pietro H Guzzi, and Mario Cannataro. 2020. Sentiment analysis for mining texts and social networks data: Methods and tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(1):e1333.