PrecogIIITH@WASSA2023: Emotion Detection for Urdu-English Code-mixed Text

Bhaskara Hanuma Vedula[†] **Prashant Kodali**[†] **Manish Shrivastava**[†]
Ponnurangam Kumaraguru[†]
[†]International Institute of Information Technology Hyderabad
{vedula.hanuma, prashant.kodali}@research.iiit.ac.in
{m.shrivastava, pk.guru}@iiit.ac.in

Abstract

Code-mixing refers to the phenomenon of using two or more languages interchangeably within a speech or discourse context. This practice is particularly prevalent on social media platforms, and determining the embedded affects in a code-mixed sentence remains as a challenging problem. In this submission we describe our system for WASSA 2023 Shared Task on Emotion Detection in English-Urdu code-mixed text. In our system we implement a multiclass emotion detection model with label space of 11 emotions. Samples are code-mixed English-Urdu text, where Urdu is written in romanised form. Our submission is limited to one of the subtasks - Multi Class classification and we leverage transformer-based Multilingual Large Language Models (MLLMs), XLM-RoBERTa and Indic-BERT. We fine-tune MLLMs on the released data splits, with and without pre-processing steps (translation to english), for classifying texts into the appropriate emotion category. Our methods did not surpass the baseline, and our submission is ranked sixth overall.

1 Introduction

Emotion Detection, which involves understanding the emotion expressed in a given text or conversation, is a widely popular task in the field of natural language processing (Peng et al., 2022). While significant research has been conducted to identify emotions in monolingual languages, the prevalence of code-mixing, particularly on social media, has made this task more challenging. Code-mixing refers to the practice of switching between two or more languages within a single discourse. While classifying emotion as positive or negative is a relatively simple task (binary classification), accurately classifying emotions into 12 categories is comparatively more challenging, due to the higher number of class labels.

In this submission, we describe our methodology and results for our submission to the WASSA 2023 Shared Task on Multi-Class Emotion Classification on Code-Mixed text messages. The data for this task was collected as part of a study (Ameer et al., 2022) that aimed to address the lack of exploration in multi-label emotion classification within code-mixed text. Specifically, the study focused on English and Roman Urdu, a language combination commonly used by the South Asian community in social media posts, comments, tweets, and SMS messages. The study presents a large benchmark corpus of 11,914 code-mixed SMS messages, manually annotated for 12 emotions, including anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust, and neutral (no emotion). Organisers use this particular dataset for the shared task.

The WASSA-2023 shared task has two tracks: Track 1: Multi-Label Emotion Classification (MLEC), where participants had to classify code-mixed SMS messages as either neutral/no emotion, or assign one or more of the eleven emotions that best represented the author's mental state. In contrast, Track 2: Multi-Class Emotion Classification (MCEC), required participants to classify code-mixed SMS messages as either neutral or no emotion, or assign one of the eleven emotions that best represented the author's mental state.

In our submission, we participated in Track 2 of the shared task. We used transformer-based multilingual models, such as XLM Roberta (Conneau et al., 2019) and Indic Bert (Kakwani et al., 2020) to fine-tune on the given dataset. MLLMs are trained on multiple languages covering high and low resource languages. However, MLLMs are known to under perform on low-resource languages. To leverage a model's propensity to perform well for high resource language like English, we also translated the code-mixed sentences into English using the OpenAI API [1]. Fine-tuned XLM-R outperformed all our other approaches. The rest of the

[1]https://openai.com/blog/chatgpt

paper is organized as follows: Section 2 describes the related work; Section 3 describes the implementation in detail as well as the experimental setup; Section 4 covers the results of our experimentation; and we end with Section 5 discussing the implications, and limitations of our current submission along with possible avenues for future work.

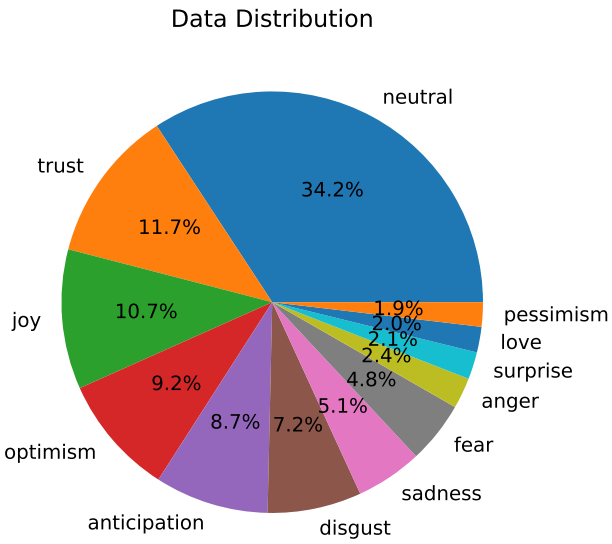

Figure 1: Pie chart showing the distribution of 12 emotions, anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust, and neutral, in the data. Neutral emotions have the highest percentage in the data, followed by a smaller percentage of other emotions.

2 Related Work

The rise of non-native English speakers on social media has led to increased interest in sentiment and emotion analysis of code-mixed data and several works have been done.

A study Ilyas et al. (2023) focused on Roman Urdu (UR) and English (EN) code-mixed text reveals the absence of a dedicated code-mixed emotion analysis corpus. To address this, the authors collect 400,000 sentences from social media, identify 20,000 UR-EN code-mixed sentences, and develop emotion detection guidelines. A large UR-EN-Emotion corpus is created, and experiments done by the authors demonstrate the effectiveness of CNN with GloVe embeddings and the improved use of the developed corpus.

Wadhawan and Aggarwal (2021) introduced a Hinglish dataset labeled for emotion detection and proposed a deep learning approach for detecting emotions in Hindi-English code-mixed tweets. The approach utilizes bilingual word embeddings from FastText and Word2Vec, as well as transformer models such as BERT, RoBERTa, and ALBERT. Experimental results show that the transformer-based BERT model achieves the highest accuracy of 71.43 percent, outperforming other models considered in the study.

Ghosh et al. (2023) has done research on Hindi-English code-mixed texts by creating an emotion-annotated Hindi-English dataset through annotations of the SentiMix benchmark dataset. The researchers propose a transformer-based multitask framework for sentiment detection and emotion recognition, utilizing the pre-trained XLMR model. Their multitask solution outperforms both single-task and multitask baselines, obviating the need for ensemble techniques and showcasing its efficiency and applicability in practical natural language processing (NLP) applications.

Set	Number of Samples
Train	9530
Test	1191
Val	1191

Table 1: Dataset statistics

3 System Description

In this section, we present details about the dataset, along with details about our experiments.

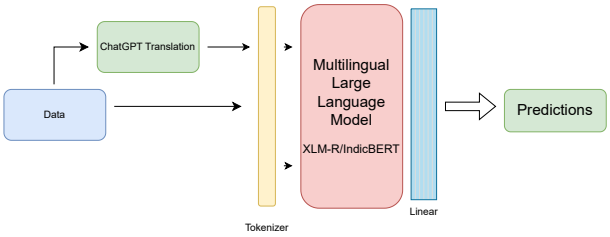

Figure 2: System Architecture for Multi Class Emotion classification using XLM-Roberta and Indic-BERT. Translated code-mixed sentences were also used as input using ChatGPT API.

3.1 Data

Dataset used in this shared task is a collection of texts that contain a mix of Roman Urdu and English language, along with corresponding labels indicating the emotions expressed in the text. This dataset was sampled from a benchmark corpus that was published in (Ameer et al., 2022) study. There were total 12 emotions in dataset which also included

no emotion. Data distribution over the 12 labels is visualized in Figure 1. The distribution of the data across training, validation, and test sets is tabulated in Table 1. Dataset is imbalanced with 34% of the samples being labeled as Neutral/no emotion. In the dataset released for the task, English-Urdu sentences were written in roman form. We translated the sentences to English with the intention to leverage MLLM's propensity to perform well in high-resource settings like English.

Using the MLLMs and the datasets we conducted three experiments to evaluate the effectiveness of different language models for code-mixed emotion classification.

3.2 Setup

We use MLLMs models available in Huggingface library (Wolf et al., 2020), and train them using PyTorch (Paszke et al., 2019). We use AdamW optimizer with default hyper-parameters for optimizing our network. We train each model for 50 epochs. For the ideal learning rate, we ran multiple runs with learning rates between [0.05, 0.0005]. Finally, we trained all our models with learning rate of 0.0005. We evaluate the model's performance using standard metrics such as accuracy, recall, precision, and F1 score, which are computed using the scikit-learn (Pedregosa et al., 2011) package.

3.3 Experiments

Fig. 2 shows the system architecture for the current submission. We describe the details pertaining to our experiments and models below. We have used transformer based encoder models - XLM-RoBERTa (Conneau et al., 2019) and Indic BERT (Kakwani et al., 2020).

XLM-Roberta Our initial experiments involved the use of XLM-Roberta. XLM-Roberta is an extension of the RoBERTa model, and it was trained on 2.5 terabytes of filtered Common Crawl data in 100 different languages.

Indic-BERT In our second experiment, we utilized Indic-BERT, another state-of-the-art language model that is specifically designed to handle code-mixed language data in the Indian subcontinent. Indic-BERT is based on the BERT architecture and was pre-trained on a large corpus of code-mixed text in 11 different Indian languages.

We choose these models because of their pre-training on multiple languages, including Indian languages. It is worth noting that romanized Hindi sentences were part of XLM-R's training corpus. Given the linguistic similarities between Hindi and Urdu, we hypothesize that the chosen models will perform well on the downstream task.

Translation: For our third experiment, we explored the use of translation to improve the performance of our emotion classification models. Specifically, we utilized ChatGPT, a language model that is capable of generating human-like text in multiple languages. ChatGPT is based on the GPT-3.5 architecture and was pre-trained on a massive corpus of diverse text. We translated code mixed sentences using OpenAI's official API. For translating the sentences we used the following prompt "Translate roman Urdu English code mixed "sentence" into English". An example of such translation is: "OK mae internet sae dekh or btata hoon" is translated to "OK, let me check on the internet and I will inform you". We, then, concatenated original data and translated English data before passing them into both the XLM-Roberta and Indic-BERT models for code-mixed emotion classification and conducted the experiment in the similar fashion described above.

We added a single MLP on top of MLLMs and trained the models using the PyTorch framework and validated it on the released validation set to determine the accuracy, precision, recall, and Macro F1 score. Predictions for test set made using the final checkpoint of the trained model were submitted for final evaluation. We describe the results for all our experiments in subsequent section.

4 Results

Table 2 presents the results of our experiments, including the F1 scores obtained by each model, allowing for easy comparison of their respective performances. The results indicate that XLM-Roberta achieved the highest F1 score among the models tested (with an F1 score of 0.60). On the other hand, Indic-BERT had the lowest F1 score among the models tested, with an F1 score of 0.54. These results demonstrate the superior performance of XLM-Roberta in the task of code-mixed emotion classification in the English-Urdu dataset. We conjecture that XLM-R's better performance can be attributed to the presence of romanized Hindi in its pre-training corpus.

Initially, we expected Indic-BERT to outperform XLM-Roberta, since IndicBERT is trained on only Indian languages (12 Indian languages), whereas XLM-R is trained on 100 languages across the

	Accuracy	Precision	Recall	Macro-F1 Score
XLM-Roberta	**0.67**	**0.67**	**0.67**	**0.60**
Indic-BERT	0.59	0.59	0.59	0.54
Translation - XLM-Roberta	0.63	0.63	0.63	0.57

Table 2: Performance Metrics Comparison of XLM-Roberta and Indic-BERT Models, and Translation using XLM-Roberta, in terms of Accuracy, Precision, Recall, and Macro-F1 Score. XLM-Roberta outperformed the other two models with highest F1 Score of 0.60.

world. However, the Indic-BERT model produced unsatisfactory results. This could be attributed to multiple reasons - incorrect spellings in the data due to romanization, IndicBERT's lack of familiarity with romanized Hindi/Urdu, making it challenging for the model to accurately capture the nuances of the emotions expressed in the text.

Initially, we hypothesized that translating the code-mixed text into English would result in better performance. However, the pre-trained models employed in our experiments failed to support this hypothesis. Surprisingly, the XLM-Roberta model outperformed the models that included translated texts. We also attempted to augment the dataset by appending translated texts to the original code-mixed data. However, this approach did not significantly improve the performance of the models. FI score obtained for this experiment was 0.57.

5 Conclusion

Based on our experimental results, XLM-Roberta has demonstrated the best performance among the approaches we tested. But our approaches couldn't match the performance of the baseline released as part of the shared task - fine-tuned mBERT.

Future Work Investigating effectiveness of ensemble methods for the task could be fruitful direction for future work. Continued pre-training on code-mixed corpora before fine-tuning on the task-specific dataset could also lead to better results. However, creating such code-mixed corpora for pre-training is non-trivial, and synthetic code-mixed corpora can be leveraged.

Limitations While the multilingual models employed in this study are capable of processing a range of languages, their performance is restricted when it comes to code-mixed sentences that feature a combination of Roman Urdu and English. This limitation suggests that the models may yield comparable results when dealing with similar language pairs. Additionally, the effectiveness of utilizing ChatGPT's API to translate code-mixed sentences into English has not been conclusively established, and thus, it remains uncertain whether this approach represents the optimal solution.

References

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

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. *CoRR*, abs/1911.02116.

Soumitra Ghosh, Amit Priyankar, Asif Ekbal, and Pushpak Bhattacharyya. 2023. Multitasking of sentiment detection and emotion recognition in code-mixed hinglish data. *Knowledge-Based Systems*, 260:110182.

Abdullah Ilyas, Khurram Shahzad, and Muhammad Kamran Malik. 2023. Emotion detection in code-mixed roman urdu - english text. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(2).

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., 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 EMNLP*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Sancheng Peng, Lihong Cao, Yongmei Zhou, Zhouhao Ouyang, Aimin Yang, Xinguang Li, Weijia Jia, and Shui Yu. 2022. A survey on deep learning for textual emotion analysis in social networks. *Digital Communications and Networks*, 8(5):745–762.

Anshul Wadhawan and Akshita Aggarwal. 2021. Towards emotion recognition in hindi-english code-mixed data: A transformer based approach. *CoRR*, abs/2102.09943.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.