YNU-HPCC at WASSA 2023: Using Text-Mixed Data Augmentation for Emotion Classification on Code-Mixed Text Message

Xuqiao Ran, You Zhang, Jin Wang, Dan Xu and **Xuejie Zhang**
School of Information Science and Engineering
Yunnan University
Kunming, China
Contact: rxq@mail.ynu.edu.cn, yzhang0202@ynu.edu.cn

Abstract

Emotion classification on code-mixed texts has been widely used in real-world applications. In this paper, we build a system that participates in the WASSA 2023 Shared Task 2 for emotion classification on code-mixed text messages from Roman Urdu and English. The main goal of the proposed method is to adopt a text-mixed data augmentation for robust code-mixed text representation. We mix texts with both multi-label (track 1) and multi-class (track 2) annotations in a unified multilingual pre-trained model, i.e., XLM-RoBERTa, for both subtasks. Our results show that the proposed text-mixed method performs competitively, ranking first in both tracks, achieving an average Macro F_1 score of 0.9782 on the multi-label track and of 0.9329 on the multi-class track.

1 Introduction

Emotion classification is a fundamental task in natural language processing (NLP). The main purpose is to identify the emotions in a written text that potentially represents the author's mental state. Compared with single-label emotion classification, multi-label emotion classification is more difficult to determine all possible emotions instead of only one emotion in a given text. Accordingly, multi-label classification has shown wide applications, such as health care and e-learning (Maxwell et al., 2017).

With the rapid growth of the Internet, linguistic code-mixed culture has become one of the most prominent communication approaches. Code-mixed texts represent texts written by two or more languages, simultaneously. According to Ameer et al. (2022), more than half of Europeans use code-mixed texts in communication. Thus, providing a more accurate judgment about the potential emotional state of such code-mixed texts is essential for various real-world applications, such as author profiling and sentiment analysis (Santosh et al., 2013; Ahmed et al., 2015).

MCEC
Text:Yaro phr Huda and Mara ki *birthdays* ka kya *plan* hai? *I am excited*:D
Label: -Emotions: joy -One-hot: [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0]
MLEC
Text: *please* jaldi aa jao *we are missing you*
Label: -Emotions: Love, Joy, Trust -One-hot: [0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0]

Figure 1: Examples of code-mixed text messages with multi-class and multi-label emotions. *Italic* and red words present the English language in code-mixed texts. A total of 12 emotional labels are listed in order: neutral, anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust. Note that *neutral* means no explicit emotions.

Toward this issue, WASSA 2023 proposes a shared task for emotion classification on code-mixed (with Roman Urdu and English) text messages, consisting of two tracks. 1) Track 1: Multi-label Emotion Classification (MLEC). 2) Track 2: Multi-class Emotion Classification (MCEC). Code-mixed texts are given for both tracks. The MLEC requires a system to classify such texts as *neutral* or multi-label (one or more) emotions in given texts while MCEC requires the system for *neutral* or only one emotion that best presents the mental state of the author, as shown in Figure 1.

We participate in both tracks on the shared tasks and found that the main challenges are twofold: 1) Code-mixed texts consist of bilingual languages in a text; 2) Multiple labels are annotated for a code-mixed text. To address these problems, we provide a system that utilizes a text-mixed data augmentation method to handle two tracks at the same time. Initially, we build a shared pair-mixed corpus in a random combination between MLEC

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 611–615
July 14, 2023 ©2023 Association for Computational Linguistics

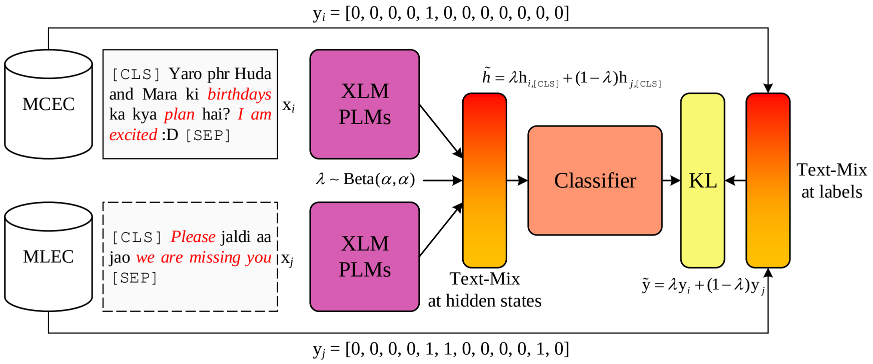
y_i = [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0]
MCEC
[CLS] Yaro phr Huda and Mara ki *birthdays* ka kya *plan* hai? *I am excited* :D [SEP]
x_i
XLM PLMs
$\tilde{h} = \lambda h_{i,[CLS]} + (1-\lambda) h_{j,[CLS]}$
$\lambda \sim \mathrm{Beta}(\alpha, \alpha)$
Classifier
KL
Text-Mix at labels
MLEC
[CLS] *Please* jaldi aa jao *we are missing you* [SEP]
x_j
XLM PLMs
Text-Mix at hidden states
$\tilde{y} = \lambda y_i + (1-\lambda) y_j$
y_j = [0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0]

Figure 2: The overview of the proposed model.

and MCEC training samples. Next, hidden states and annotated labels in a paired sample are individually mixed in an interpolation method by the same weights. Note that, hidden states in paired texts are generated from multilingual pre-trained language models (PLMs) (Qiu et al., 2020), i.e., XLM-RoBERTa (Conneau et al., 2019), that can align word representations from multilingual tokens and semantics from multilingual sentences. Finally, the predicted probabilities over multiple emotions from the system are converged with mixed annotated labels in a Kullback-Leibler (KL) (Eguchi and Copas, 2006) divergence loss function. Consequently, the proposed system can be shared for both tracks in one training phase. Extensive experiments are conducted to investigate the effect of the proposed method and the best submissions reveal that our system ranks first in both MLEC and MCEC tracks.

The remainder of this paper is constructed as follows. A line of related works is provided in section 2. A detailed description of the proposed system is introduced in section 3. Experimental results are analyzed in section 4. Finally, conclusions are drawn in section 5.

2 Related Work

Emotion classification is a challenging NLP task that aims to automatically classify text documents into one or more predefined emotion categories. It has long been of interest to researchers in areas such as sentiment analysis, opinion mining, and social media analysis.

Recent studies have explored the use of different deep learning architectures, such as Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2017) and Recurrent Neural Networks (RNNs) (Mikolov et al., 2010), to improve the performance of emotion classification. Other studies have focused on the development of pretraining models such as BERT (Devlin et al., 2019), ALBERT (A Lite BERT) (Lan et al., 2019) and RoBERTa (Liu et al., 2019), which have been shown to achieve state-of-the-art results on various NLP tasks, including emotion classification.

A growing body of research has focused on the use of monolingual and cross-lingual models to improve emotion classification. Monolingual methods are based on training models on large amounts of data from a single language, while cross-lingual models make use of data from multiple languages to learn more robust representations. Cross-lingual models such as XLM (Barbieri et al., 2022) have shown promise in many NLP tasks, including emotion classification.

Due to the scarcity of low-resource data, data augmentation is essential. Mixup (Zhang et al., 2017) is a simple and effective data augmentation method, which can significantly improve the effect in multiple fields such as image, text, speech, recommendation. Different variants of mixing methods have also been designed in the different space, the cutMix (Yun et al., 2019) method takes a different approach, instead of interpolating two samples from a numerical perspective, but from the spatial perspective of the image, it cuts a random rectangular area on one picture to another picture to generate a new picture. Manifold mixup (Verma et al., 2018) extends mixup, and extends the mixing of input data (raw input data) to the output mixing of the intermediate hidden layer.

3 Methodology

In this section, we will describe our system that participated in WASSA 2023 shared task on emotion classification on code-mixed text messages. As shown in Figure 2, the proposed models consist of four parts, including pair-mixed corpus, sentence encoder, text-mixed interpolation, and classifier. Before introducing the proposed model, we describe the shared task 2 in advance.

3.1 Emotion Classification on Code-mixed Text

Given a text code-mixed text $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ where N represents the text length. Regarding MCEC and MLEC, it requires a system $f_\theta(\hat{\mathbf{y}}|\mathbf{x})$ to predict ground-truth emotions $\mathbf{y}_c \in \mathbb{R}^C$ and $\mathbf{y}_l \in \mathbb{R}^C$ where θ represents the whole parameters in the system and C is the total number of emotional labels.

3.2 Pair-mixed Corpus

In this paper, we propose a unified system for both MCEC and MLEC tasks, simultaneously. Hence, we initially mix both corpora in a random combination. In detail, for each pair of training sample $((\mathbf{x}_i^c, \mathbf{x}_j^l), (\mathbf{y}_i^c, \mathbf{y}_j^l))$, we random select a code-mixed SMS message from MCEC training set $\mathcal{D}_c$ and another one from MLEC $\mathcal{D}_l$, where ($\mathbf{x}_i^c$, $\mathbf{y}_i^c$) $\in \mathcal{D}_c$, ($\mathbf{x}_j^l$,$\mathbf{y}_j^l$) $\in \mathcal{D}_l$, $i \in [1 : |\mathcal{D}_c|]$, and $j \in [1 : |D_l|]$ Consequently. a mixed corpus $\mathcal{D}_m = \sum_{i,j}((\mathbf{x}_i, \mathbf{x}_j), (\mathbf{y}_i, \mathbf{y}_j))$ is generated.

3.3 Multilingual Text Encoders

Due to code-mixed texts comprised bilingual languages, aligning word embedding and semantic cross-lingual sentences or phrases is critical for robust text representations. To this end, we adopt a cross-lingual PLM, XLM-RoBERTa, as the sentence encoder to encode paired texts into hidden spaces, formulated as:

$$\mathbf{h}_i, \mathbf{h}_j = \mathbf{XLM}((\mathbf{x}_i, \mathbf{x}_j))\,, \tag{1}$$

where($\mathbf{x}_i$, $\mathbf{x}_j$) $\sim \mathcal{D}_m$, $\mathbf{h}_i, \mathbf{h}_j \in \mathbb{R}^{N \times d}$; d is the dimensionality of hidden states.

3.4 Text-mixed Interpolation

To further mix up $(\mathbf{x}_i, \mathbf{y}_i)$ and $(\mathbf{x}_j, \mathbf{x}_j)$, a mixed interpolation method is proposed. Based on sentence representation $\mathbf{h}_i$ and $\mathbf{h}_j$, we choose `[CLS]` (a special token in PLMs) to represent the global sentence representation mixed in hidden spaces (denoted as $\tilde{h} \in \mathbb{R}^d$), as well as annotated labels (denoted as $\tilde{\mathbf{y}} \in \mathbb{R}^C$), as shown in Figure 2.

$$\begin{aligned} &\tilde{h} = \lambda \mathbf{h}_{i,\texttt{[CLS]}} + (1-\lambda)\mathbf{h}_{j,\texttt{[CLS]}} \\ &\tilde{\mathbf{y}} = \lambda \mathbf{y}_i + (1-\lambda)\mathbf{y}_j \\ &\lambda \sim \mathbf{Beta}(\alpha, \alpha) \\ &\lambda = \mathbf{max}(\lambda, 1-\lambda) \end{aligned}\;, \tag{2}$$

where $\lambda \in [0, 1]$ is generated from Beta distribution with a hyper-parameter α .

Dataset	Instances
Train	9530
Dev	1191
Test	1191

Table 1: Data distribution.

3.5 Training Objective and Inference Strategy

In this section, we introduce the proposed system in training and inference phases.

Training objective. To predict emotional probabilities $\hat{\mathbf{y}} \in \mathbb{R}^C$, we apply a Multi-Layer Perceptron (MLP) to encode mixed textual hidden states $\tilde{h}$:

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{MLP}(\tilde{h}) \\ &= \mathbf{Linear}(\tanh(\mathbf{Dropout}(\tilde{h}))) \end{aligned}\,, \tag{3}$$

where $\mathbf{Linear}(\cdot), \mathbf{tanh}(\cdot)$ and $\mathbf{Dropout}(\cdot)$ are fully-connected layer, hyperbolic tangent activation function, and dropout function, respectively. The loss function between predicted probabilities and mixed labels is KL divergence (Eguchi and Copas, 2006) for the system training:

$$\mathcal{L} = \mathbf{KL}(f(\hat{\mathbf{y}}|(\mathbf{x}_i, \mathbf{x}_j); \theta)||\tilde{\mathbf{y}}). \tag{4}$$

Inference strategy. In the inference phase, MCEC and MLEC test datasets are separately fed into the system for individual purposes. Note that there are not mixed procedures in the inference phase. For MCEC task, the argmax function is used to predict one emotion; for MLEC task, a threshold score of 0.5 is used to predict all most possible emotions, formulated as:

$$\begin{aligned} &\hat{\mathbf{y}}_c = \mathbf{argmax}(\mathbf{softmax}(\hat{\mathbf{y}})) \\ &\hat{\mathbf{y}}_l = \mathbf{threshold}(\mathbf{sigmoid}(\hat{\mathbf{y}}), 0.5) \end{aligned}. \tag{5}$$

4 Experimental Results

In this section, extensive experiments were conducted for both MCEC and MLEC tracks.

Model	MCEC				MLEC		
	Mac-P	Mac-R	Mac-F_1	Acc	Mac-F_1	Mic-F_1	Acc
BERT	0.70	0.70	0.69	0.70	0.58	0.64	0.57
mBERT	0.71	0.70	0.70	0.70	0.59	0.66	0.59
mBERT+Text-mixed	0.93	0.92	0.92	0.92	0.79	0.85	0.79
RoBERTa	0.71	0.70	0.70	0.70	0.50	0.63	0.56
XLM-RoBERTa	0.73	0.73	0.73	0.73	0.57	0.66	0.59
XLM-RoBERTa+Text-mixed	**0.94**	**0.93**	**0.93**	**0.93**	**0.82**	**0.86**	**0.80**

Table 2: Performance on Dev dataset (for both tracks). **Boldface** figures mean the best results.

Team	MLEC		MCEC	
	Mac-F_1	Rank	Mac-F_1	Rank
YNU-HPCC†	0.9869	1	0.9329	1
CTcloud	0.9833	2	0.8917	2
wsl&zt	0.9464	3	0.7359	3
Baseline‡	0.8347	-	0.7014	-

Table 3: Official results from the shared task leader board. Team† and Team‡ present our team name and the official baseline, respectively.

4.1 Datasets

The Internet is the most prominent source in promoting global, linguistic code-mixed culture. In South Asian community and particularly in Pakistan, code-mixed (English and Roman Urdu) text became a preferable script for Facebook comments/posts, tweets, and daily communication using SMS messages. The shared task organizers made available the dataset from (Ameer et al., 2022). Table 1 reported the detailed datasets in statistics.

4.2 Evaluation Metrics

To evaluate the performance of participant systems, the official competition provides Micro $\mathbf{F}_1$ (Mic-$\mathbf{F}_1$), Macro $\mathbf{F}_1$ (Mac-$\mathbf{F}_1$), and Accuracy (Acc) for track 1 and Mac-$\mathbf{F}_1$, Mac-Precision (Mac-P), Mac-Recall (Mac-R), and Acc for track 2.

4.3 Implementation Details

Hyper-parameters. All sentence is tokenized by XLM-RoBERTa based tokenizer with a maximum length of 90. Sentence encoder is `twitter-xlm-roberta-base-sentiment` PLM with the dimensionality of 768. α in Beta distribution is set as 0.75. Dropout ratio in MLP module is set as 0.2. For optimization, AdmW is adopted with learning rate of 2e-5 and batch size of 64. The code of this paper is availabled at `https://github.com/linsongisgood/wassa2023`.

Baselines. To investigate the effect of the proposed method, several baseline models are introduced, including BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) and their variants in cross-lingual versions.

4.4 Results and Analysis

Comparative Dev results were reported in Table 2. Due to PLMs that transfer generic language performance learned from a large pretrained corpus into downstream tasks, several PLMs achieved competitive results on both tracks. It can be observed that, with the cross-linguistic pretraining phase, PLMs such as mBERT (Pires et al., 2019) and XLM-RoBERTa (Conneau et al., 2019) outperformed monolingual PLMs such as BERT and XLM-RoBERTa, respectively. This phenomenon demonstrated that aligning cross-linguistic word representation and semantics is crucial to generate robust representation on code-mixed texts.

Furthermore, we found that the introduction of the text-mixed data augmentation method gained more performance on both tracks. A possible reason may be that the combination of multi-class and multi-label corpora improved the generalization capability of the system. Note that the proposed text-mixed method facilitates the shared system simultaneously performing both tracks during one training phase. Table 3 showed our best submissions with official results and ranks, revealing the effectiveness of the proposed system.

5 Conclusions

In this paper, we proposed described our system submission WASSA 2023 shared task 2 in emotion classification. Our system utilizes the text-mixed method and cross-lingual PLMs for robust representation of code-mixed texts. As a result, our system won the first rank in both tracks.

In the future, we will explore the text-mixed method applied to large amounts of unlabeled code-mixed texts for better performance.

Limitations

The limitations of this work can be concluded into three points: 1) The data in the test set is relatively small, so it cannot more accurately reflect the effectiveness of the method proposed in this paper. We believe that tuning the model on a larger dataset can help improve the performance of the model. 2) Due to device performance limitations,we did not experiment with larger models. In our experiment,we only tested the method with models like XLM-RoBERTa, mBERT and BERT. Its performance with larger models is not known. 3) We did not perform an extensive hyperparameter search, which might further improve the model's performance.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61966038 and 62266051, and the Yunnan Postdoctoral Science Foundation under Grant Nos. C615300504048. The authors would like to thank the anonymous reviewers for their constructive comments.

References

Khaled Ahmed, Neamat El Tazi, and Ahmad Hany Hossny. 2015. Sentiment analysis over social networks: an overview. In *2015 IEEE international conference on systems, man, and cybernetics*, pages 2174–2179.

Iqra Ameer, Grigori Sidorov, Helena Gomez-Adorno, and Rao Muhammad Adeel Nawab. 2022. Multi-Label Emotion Classification on Code-Mixed Text: Data and Methods. *IEEE Access*, 10:8779–8789.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. *2022 Language Resources and Evaluation Conference, LREC 2022*, pages 258–266.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and AI Language. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Shinto Eguchi and John Copas. 2006. Interpreting kullback–leibler divergence with the neyman–pearson lemma. *Journal of Multivariate Analysis*, 97(9):2034–2040.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut, and Google Research. 2019. ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS.

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.

Andrew Maxwell, Runzhi Li, Bei Yang, Heng Weng, Aihua Ou, Huixiao Hong, Zhaoxian Zhou, Ping Gong, and Chaoyang Zhang. 2017. Deep learning architectures for multi-label classification of intelligent health risk prediction. *BMC bioinformatics*, 18:121–131.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

K Santosh, Romil Bansal, Mihir Shekhar, and Vasudeva Varma. 2013. Author profiling: Predicting age and gender from blogs. *Notebook for PAN at CLEF*, 2013(2).

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, I oannis M itliagkas, D avid L opez-Paz, and Yoshua Bengio. 2018. Manifold Mixup: Better Representations by Interpolating Hidden States. *36th International Conference on Machine Learning, ICML 2019*, 2019:11196–11205.

Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Junsuk Choe, and Youngjoon Yoo. 2019. CutMix: Regularization strategy to train strong classifiers with localizable f eatures. *Proceedings of the IEEE International Conference on Computer Vision*, 2019:6022–6031.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.