Codewithzichao@DravidianLangTech-EACL2021: Exploring Multilingual Transformers for Offensive Language Identification on Code Mixing Text

Zichao Li
School of Software and Microelectronics, Peking University, China
lizichao@pku.edu.cn

Abstract

This paper describes our solution submitted to shared task on Offensive Language Identification in Dravidian Languages. We participated in all three of offensive language identification. In order to address the task, we explored multilingual models based on XLM-RoBERTa and multilingual BERT trained on mixed data of three code-mixed languages. Besides, we solved the class-imbalance problem existed in training data by class combination, class weights and focal loss. Our model achieved weighted average F1 scores of 0.75 (ranked 4th), 0.94 (ranked 4th) and 0.72 (ranked 3rd) in Tamil-English task, Malayalam-English task and Kannada-English task, respectively.

1 Introduction

Offensive language identification is a research field that has received increasing attention in recent years. Especially with the rise of social media platforms, it is essential to identify offensive language on code-mixed social media texts. It is a challenging task to identify offensive language on social media texts. Additionally, a lot of work has been done for offensive language identification in languages like English, Greek or Spanish (Zampieri et al., 2019; Pitenis et al., 2020; Ranasinghe and Zampieri, 2020), but little work has been done for offensive language identification of code-mixed text in Dravidian languages.

Shared task on Offensive Language Identification in Dravidian Languages (Tamil-English, Malayalam-English and Kannada-English) has changed this situation. The goal of this shared task is to identify offensive language on code-mixed text in Dravidian languages. The code-mixed text is collected from social media platforms. It is a comment or post level multilingual classification task that given a comment or post in code-mixed Tamil-English language, Malayalam-English language and Kannada-English language, systems have to classify it into Not-offensive, offensive-untargeted, offensive-targeted-individual, offensive-targeted-group, offensive-targeted-other, or Not-in-indented-language.

In this paper, we explore multilingual transformers on code-mixed text for offensive language identification in Dravidian languages. Inspired by the conclusion that lexical overlap among different languages can be helpful to improve the performance of the model in a single language given in (Pires et al., 2019), we combine the training data of three languages and trained our multilingual model on the mixed data. Besides, we solve the class-imbalance problem existed in training data by class combination, class weights and focal loss. Finally, we use adversarial training to further improve performance of our model. With these approaches, we achieved 4th Rank in Tamil-English, Malayalam-English task and 3rd Rank in Kannada-English task.

2 Data

We used the data provided by the organizers of shared task on Offensive Language Identification in Dravidian Languages (Chakravarthi et al., 2021, 2020b; Hande et al., 2020; Chakravarthi et al., 2020a), which have been annotated well at comment or post level. The numbers of Tamil, Malayalam and Kannada training data are 35139, 16010 and 6217, respectively. The statistics of data are shown in Table 1, Table 2 and Table 3.

There are four methods used to preprocess the data as follow:

- **Data combination:** Inspired by the conclusion that lexical overlap among different languages can be helpful to improve the performance of the model in a single language given in multilingual BERT, we combined the train-

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 164–168

Class	Train	Dev	Test
Not-offensive	25425	3193	3190
Offensive-Untargetede	2906	356	368
Offensive_Targeted_Insult_Individual	2343	307	315
Offensive_Targeted_Insult_Group	2557	295	288
Offensive_Targeted_Insult_Other	454	65	71
not_Tamil	1454	172	160
Count	35139	4388	4392

Table 1: Statistics of Tamil-English language dataset.

Class	Train	Dev	Test
Not-offensive	14153	1779	1765
Offensive-Untargetede	191	20	29
Offensive_Targeted_Insult_Individual	239	24	27
Offensive_Targeted_Insult_Group	140	13	23
Offensive_Targeted_Insult_Other	0	0	0
not_Malayalam	1287	163	157
Count	16010	1999	2001

Table 2: Statistics of Malayalam-English language dataset.

ing data of the three languages and trained our model on the mixed data.

- **Noise removal:** Emojis and extra blanks in the code-mixed data are removed in advance. The experimental results show that removing these noise can improve the performance of our model.

- **Class combination:** We combined classes named not Tamil, not Malayalam and not Kannada which have few numbers into one class named Not in indented language, which is one of the ways to alleviate the class-imbalance problem. Finally there are 6 classes in training data.

- **Tokenization:** Texts are tokenized using the sentencepiece toolkit[1] and converted to the corresponding IDs through the vocabulary of XLM-RoBERTa (Conneau et al., 2020).

3 System

In this section, we first present our model for offensive language identification task, and then introduce the solution which is used to solve the class-imbalance problem and improve the robustness and generalization of our model.

[1]https://github.com/google/sentencepiece

3.1 Model Architecture

Our model is mainly divided into three layers: encoding layer, pooling layer and prediction layer.

3.1.1 Encoding Layer

Given a sentence $X = \{x_1, x_2, ..., x_n\}$, which $x_i \in R^d$ and n is the length of a sentence, they are fed into an encoder to obtain the contextual representation for each word of the sentence:

$$[h_1, h_2, ..., h_n] = Encoder([x_1, x_2, ..., x_n]), \quad (1)$$

where the $Encoder$ can be XLM-RoBERTa or multilingual BERT.

3.1.2 Pooling Layer

Inspired by (Reimers and Gurevych, 2019), we use the average-over-time pooling $G \in R^d$ of the output of the last layer of the pre-trained model as the sentence embedding instead of the [CLS] embedding:

$$G = AvgPool([h_1, h_2, ..., h_n]). \quad (2)$$

3.1.3 Prediction Layer

To classify each sentence, we put representation of each sentence into the softmax layer:

$$P(y_i|X) = softmax(W_G G + b). \quad (3)$$

Class	Train	Dev	Test
Not-offensive	3544	426	427
Offensive-Untargetede	212	33	33
Offensive_Targeted_Insult_Individual	487	66	75
Offensive_Targeted_Insult_Group	329	45	44
Offensive_Targeted_Insult_Other	123	16	14
not_Kannada	1522	191	185
Count	6217	777	778

Table 3: Statistics of Kannada-English language dataset.

Language	Model	Precision	Recall	F1-score	Rank
Tamil	Multilingual-BERT	0.74	0.75	0.74	
	XLM-RoBERTa	0.73	0.75	0.74	
	Our submission (Ensemble)	**0.74**	**0.77**	**0.75**	4
Malayalam	Multilingual-BERT	0.92	0.93	0.92	
	XLM-RoBERTa	0.93	0.94	0.93	
	Our submission (Ensemble)	**0.93**	**0.94**	**0.94**	4
Kannada	Multilingual-BERT	0.65	0.70	0.67	
	XLM-RoBERTa	0.70	0.74	0.71	
	Our submission (Ensemble)	**0.70**	**0.75**	**0.72**	3

Table 4: Official results and ablations of our model for Tamil, Malayalam and Kannada languages on the test datasets.

3.2 Class-Imbalance Problem

As mentioned in the second section, there is a serious class-imbalance problem in the training data. We have taken three ways to solve this problem as follow:

- **Class combination:** We combine the three classes named not-Tamil, not-Malayalam, and not-Kannada into one class named Not-in-indented-language, which helps to alleviate the class-imbalance problem.

- **Adjustment to class weights:** We count the frequency of each class: $P = \{p_i\}_{i=1}^{6}$, and then use the reciprocal of the log of frequency as class weights: $W = \{\frac{1}{log(p_i)}\}_{i=1}^{6}$. This can reduce the loss of classes which have large numbers and increase the loss of classes which have few numbers, so that the model can pay more attention to the classes which have few numbers (King and Zeng, 2001).

- **Focal loss:** Proposed by (Lin et al., 2017), focal loss has been proved to be able to deal with the class-imbalance problem. Experimental results prove that it can improve the performance of our model.

Hyper-parameters	Value
Batch size	8
Dropout rate	0.2
Gradient clipping	0.25
Epoch	80
Learning rate	2e-5

Table 5: Hyper-parameters of our model.

3.3 Adversarial Training

Adversarial training is an important way to enhance the robustness and generalization of neural networks. We adopt FGM (Goodfellow et al., 2015) method to interfere with the embedding layer, so as to further improve the performance of our model.

4 Experiment and Results

4.1 Experimental Settings

We use Pytorch (Paszke et al., 2017) and HuggingFace's transformers (Wolf et al., 2020) to implement our model. We use XLM-RoBERTa and multilingual BERT as encoders for text and combine XLM-RoBERTa and multilingual BERT model with the highest weighted average F1 score on the development dataset. We apply dropout with a

80% keep probability. We optimize the loss using AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate at 2e-5. We use mixed precision training based on Apex library[2]. We list all hyper-parameters in Table 5. We conduct the experiments on NVIDIA Tesla T4 GPUs. Our code is available at Github[3].

4.2 Results and Ablations

All teams were ranked by the weighted average F1 score. Table 4 shows results of our ensemble model on all three of languages. Our model achieved 0.75 weighted average F1 score (ranked 4th) in Tamil-English task, 0.94 weighted average F1 score (ranked 4th) in Malayalam-English task and 0.72 weighted average F1 score (ranked 3rd) in Kannada-English task.

In addition, the ablation results are also shown in Table 4. The results of the ensemble model on the test dataset of three languages are improved. The weighted average F1 score of the ensemble model on the test data is 0.01 higher than that of multilingual BERT and XLM-RoBERTa in Tamil-English task. The weighted average weighted average F1 score of the ensemble model on the test data is 0.02 higher than that of multilingual BERT and 0.01 higher than that of XLM-RoBERTa in Malayalam-English task. The result of the ensemble model on the test data is 0.05 higher than that of multilingual BERT and 0.01 higher than that of XLM-RoBERTa in Kannada-English task.

5 Conclusion

This paper describes our solution submitted to shared task on Offensive Language Identification in Dravidian Languages (Tamil-English, Malayalam-English, and Kannada-English). First of all, we use multilingual models trained on mixed data of three code-mixed languages to obtain better performance in all three of code-mixed languages. Secondly, we use class combination, class weights and focal loss to solve the class-imbalance problem existed in training data. Finally, we use adversarial training to further improve the performance of our model. We achieved 4th Rank in Tamil-English, Malayalam-English task and 3rd Rank in Kannada-English task. In future research, we will further consider the differences among different languages to further improve performance of our model.

[2]https://github.com/NVIDIA/apex

[3]https://github.com/codewithzichao/Multilingual-Transformers

References

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan V, Elizabeth Sherly, and John Philip McCrae. 2021. Findings of the shared task on Offensive Language Identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572.

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, Barcelona, Spain (Online).

Gary King and Langche Zeng. 2001. Logistic regression in rare events data. *Political Analysis*, 9(2):137–163.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

I. Loshchilov and F. Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Adam Paszke, S. Gross, Soumith Chintala, G. Chanan, E. Yang, Zachary Devito, Zeming Lin, Alban Desmaison, L. Antiga, and A. Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.