

YNU@Dravidian-CodeMix-FIRE2020: XLM-RoBERTa for Multi-language Sentiment Analysis

Xiaozhi Ou, Hongling Li*

School of Information Science and Engineering, Yunnan University, Kunming, 650500, Yunnan, P.R. China

Abstract

This article describes the system that our team submitted to the Dravidian-CodeMix-FIRE 2020. The purpose of this task is to identify the sentiment polarity of the code-mixed dataset of Dravidian (Malayalam-English and Tamil-English) comments/posts collected from social media. Our system is based on a pre-trained multi-language model XLM-RoBERTa, and uses the K-folding method to ensemble and aims to solve the sentiment analysis problem of multilingual code-mixed across language models. We participate in the tasks of two code-mixed languages (Malayalam-English and Tamil-English), our system achieves the best F-Score of 0.74 in Malayalam-English (Ranks 1/28), and we rank third in Tamil-English with an F-Score of 0.63.

Keywords

code-mixed, multi-language, XLM-RoBERTa, K-folding, Dravidian-CodeMix-FIRE 2020

1. Introduction

The development of social media (such as Blogs, Twitter and Facebook) has created many new opportunities and challenges for information access and language technology. Although the current language technology is mainly built for English, non-native English speakers will combine English with other languages when using social media. Mixed language, also known as code-mixed, is a norm in a multilingual society. This linguistic phenomenon poses a great challenge to traditional NLP fields, such as sentiment analysis, machine translation and text summarization, etc. Those systems trained on single-language data have poor performances on code-mixed data because of the complexity of code-switching at different linguistic levels in text.

In recent years, researchers have turned their attention to the sentiment analysis task in code-mixed social media texts [1]. first shared task on Sentiment Analysis in Dravidian Code-Mixed Text (Dravidian-CodeMix-FIRE 2020) [2, 3]. The purpose of this task is to identify the sentiment polarity of the code-mixed dataset of comments/posts in Dravidian languages (Malayalam-English and Tamil-English) collected from social media [4, 5]. This is a polarity classification task at the message-level. Given a comment/posts, the system must classify the comment/posts as Positive, Negative, Mixed feelings, Neutral, or Non-Malayalam/Non-Tamil. In this competition, we focus on developing a viable solution for the code-mixed sentiment analysis field, and we participated in the task of two code-mixed languages (Malayalam-English and Tamil-English). We use a multi-language pre-training model called XLM-RoBERTa [6], which not only inherits the XLM training method but also draws on the ideas of RoBERTa. During the training process, we merge the official provided training set and validation set and randomly scramble them as the training data set to train the model. Finally, we employ the K-fold


FIRE 2020: Forum for Information Retrieval Evaluation, December 16-20, 2020, Hyderabad, India

EMAIL: xiaozhiou88@gmail.com (X. Ou); honglingli66@126.com (H. Li*)

ORCID: 0000-0001-6043-2348 (X. Ou)



© 2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

approach for ensemble [7]. Our model is available on GitHub: https://github.com/Ouxiaozhi/YNU_TEAM-IN-dravidian-codemix-.

The rest of this article is organized as follows. Section 2 introduces related work. Section 3 describes the data and approaches. Section 4 presents the experimental results. Finally, our conclusion and future work are presented in Section 5.

2. Related Work

For the past two decades, sentiment analysis has been an active area of research in academia and industry. In recent years, sentiment analysis has mostly focused on the study of a single language. Some related shared tasks of the organization, such as the detection of offensive language in German ¹ [8], the detection of hate speech in Italian ² [9], and organized of Semeval 2019 shared task 6 OffensEval ³ – Identifying and Categorizing Offensive Language in Social Media [10]. At present, the study of multilingualism has become a new upsurge, and some related tasks organized recently have attracted a large number of researchers. For example, the Semeval 2019 shared task 5 HatEval ⁴ – Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter [11]. Semeval 2020 shared task 12 OffensEval 2 ⁵ – Multilingual Offensive Language Identification in Social Media [12], and the shared tasks of HASOC (2019) and HASOC (2020) ⁶ – Hate Speech and Offensive Content Identification in Indo-European Languages [13]. Code-Mixed is a common phenomenon in multilingual communities, and there is a growing demand for sentiment analysis of a large number of code-mixed social media texts. In recent years, some shared tasks related to code-mixed have been launched, such as the Semeval 2020 shared task 9 ⁷ – Sentiment Analysis for Code-Mixed Social Media Text [14]. Four series of tasks on Mixed Script Information Retrieval was organized at the Forum for Information Retrieval Evaluation (FIRE) [15]. Three workshops on Computational Approaches to Linguistic Code-Switching (CALCS) were also held [16].

Some researchers try to analyze sentiment from code-mixed text. Chittaranjan et al. [17] tried word-level recognition of code-mixed data to classify emotions. Sharma et al. [18] tried to perform a shallow analysis of the code-mixed data obtained from online social media. Bojanowski et al. [19] proposed a word representation model based on skip-gram to classify the emotion of tweets. Giatsoglou et al. [20] trained a hybrid system based on dictionary-based document vectors, word embeddings, and word polarity to classify the sentiment of tweets.

3. Data and Approaches

3.1. Data description

In order to run the experiment, we use the datasets in two languages (Malayalam-English [4] and Tamil-English [5]) provided by the organizer, the data mainly comes from YouTube video comments. The dataset contains all three types of code-mixed sentences: Inter-Sentential switch, Intra-Sentential switch and Tag switching. Among them, the Malayalam-English dataset contains 6,739 comments,

¹<https://projects.fzai.h-da.de/iggsa/>

²<http://www.di.unito.it/~tutreeb/haspeede-evalita20/index.html>

³<https://competitions.codalab.org/competitions/20011>

⁴<https://competitions.codalab.org/competitions/19935>

⁵<http://alt.qcri.org/semeval2020/index.php?id=tasks>

⁶<https://hasocfire.github.io/hasoc/2019/index.html>

⁷<https://competitions.codalab.org/competitions/20654>

Table 1
Data distribution statistics

Languages	Train	Validation	Test	Total
Malayalam-English	4,717	674	1,348	6,739
Tamil-English	11,335	1,260	3,149	15,744

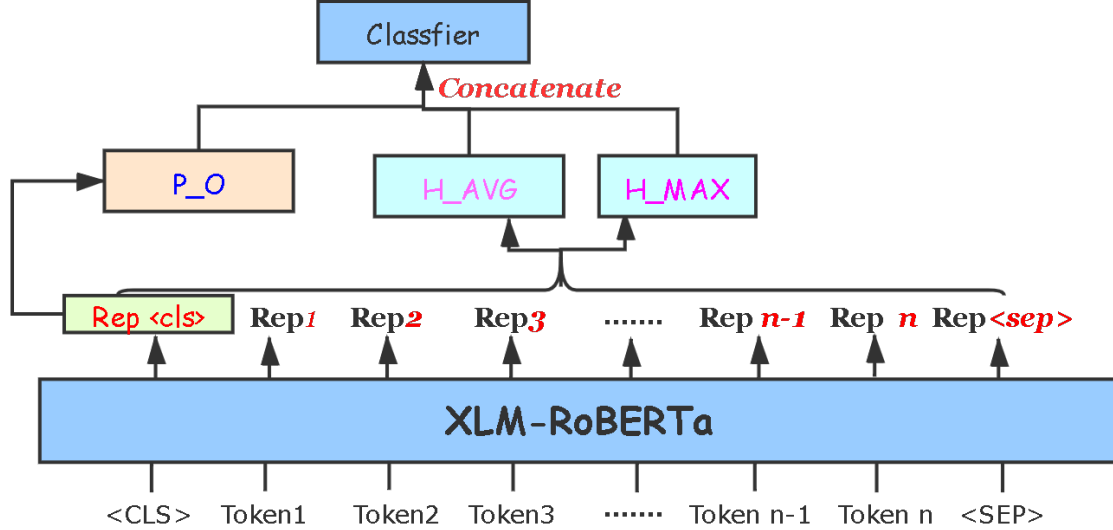


Figure 1: The overall architecture diagram of the model (P_O is the pooler output. H_AVG and H_MAX are the average-pooling and max-pooling of hidden-states sequences of the last layer output of the XLM-RoBERTa model, respectively.)

and the Tamil-English dataset contains 15,744 comments. Table 1 shows the distribution of the training set, validation set, and test set for the two languages.

In our experiment, we merge the training set and validation set released by the organizer, and randomly shuffled their order. For both languages, we adopt this method to process the data. Finally, we get a new training dataset of Malayalam-English with 5,391 comments and a new training dataset of Tamil-English with 12,595 comments. In the experimental run, we employ a cross-validation idea, the K-fold ensemble method, to improve the overall classification performance of the model.

3.2. Approaches

Inspired by the success of the multilingual model, XLM-RoBERTa has greatly expanded the amount of multilingual training data used in unsupervised MLM pre-training compared with previous work, and has reached the latest level in both monolingual and cross-lingual benchmarks [6]. As shown in Figure 1, our model is implemented based on the XLM-RoBERTa multilingual model. First, we get the pooler output (P_O), and obtain the sequence of hidden states on the output of the last layer of XLM-RoBERTa. Then, we obtain H_AVG through average-pooling and H_MAX by max pooling. Finally, we concatenate P_O, H_AVG and H_MAX into the Classifier.

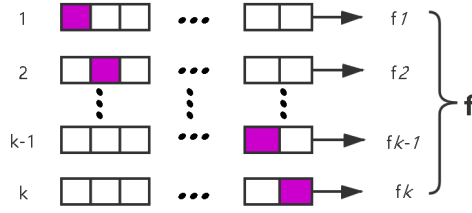


Figure 2: The K-fold ensemble method

3.3. K-folding ensemble

In this paper, in order to improve the overall classification performance of the model, we employed a K-fold ensemble method. The design idea of this method comes from K-fold cross-validation. The source data is randomly divided into K parts, the K-1 subset is used for training, the remaining subset is used as the validation set, and then repeated K times. Finally, the K results are accumulated and averaged to obtain the final output. The purpose of K-fold ensemble is to train different datasets during each fold training process and extract different features in the model feature extraction process, so as to further improve the generalization ability of the model. The K-fold ensemble method is shown in Figure 2.

4. Experiment results

4.1. Experiment setting

In our experiment, we did not preprocess the data. Our model is implemented based on Pytorch. We use XLM-RoBERTa base as our pre-training model, which contains 12 layers. We use the 8-fold cross-validation, and the Maximum sentence length is 160. We use the learning rate of $3e-5$, CrossEntropy Loss, and Adam as the optimizer. To save GPU memory, the batch size is set to 4 and the gradient accumulation step is set to 4 so that the gradient is accumulated 4 times each time a sample is input, and then the backpropagation update parameters are performed.

4.2. Results analysis

The primary evaluation metric for the task is F-Score toward the positive class as a trade-off between Precision (P) and Recall (R) [21]. Table 2 shows the 8-fold cross-validation results of our experiment on our validation set. (best regression heads for model and language are in **bold**).

First of all, from Table 2 we can conclude that the model performance of the Malayalam-English task is better than Tamil-English. Secondly, the correct selection of the regression head does help to obtain better performance. For the Malayalam-English final submission, we select the XLM-RoBERTa model with the P_O & H_MAX & H_AVG regression head. Based on the test data, it show an F-Score of 0.74 (P = 0.74, R = 0.74). That is higher than our 8-fold cross-validation result by 0.064, ranking No.1 in the competition leaderboard. For the Tamil-English final submission, we select the XLM-RoBERTa model with the P_O & H_MAX & H_AVG regression head. Based on the test data, it show an F-Score of 0.63 (P = 0.61, R = 0.67). That is higher than our 8-fold cross-validation result by 0.016, ranking 3rd in the competition leaderboard.

Table 2

8-fold cross-validation results for language, models and specific regression head on our validation set (weighted F-Score).

Language	Models	P_O	H_MAX	H_AVG	P_O & H_MAX & H_AVG
Malayalam-English	M-Bert	0.6315	0.6456	0.6378	0.6427
	XLM-RoBERTa	0.6556	0.6598	0.6632	0.6759
Tamil-English	M-Bert	0.5682	0.5734	0.5917	0.5796
	XLM-RoBERTa	0.5961	0.5986	0.6109	0.6143

5. Conclusions and Future Work

This article introduces our overall idea and specific plan for participating in the Dravidian-CodeMix-FIRE 2020 sharing task, aiming to identify the sentiment polarity of the code-mixed datasets annotated in Dravidian languages (Malayalam-English and Tamil-English) collected from social media. We use a multilingual pre-training model based on XLM-RoBERTa for classification, and use K-fold method for ensemble. Our results demonstrate that multilingual models perform well in code-mixed datasets, and we suggest that code-mixed NLP practitioners consider at least one of the XLM-RoBERTa variants when selecting language models for their NLP systems. At present, the importance of breaking the English-centric NLP research has been widely discussed, and we believe that the research of non-English languages will increase. We believe that the best models in the future can not only learn from different fields but also from different languages.

References

- [1] B. R. Chakravarthi, Leveraging orthographic information to improve machine translation of under-resourced languages, Ph.D. thesis, NUI Galway, 2020.
- [2] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, E. Sherly, J. P. McCrae, Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text, in: Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2020). CEUR Workshop Proceedings. In: CEUR-WS. org, Hyderabad, India, 2020.
- [3] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, E. Sherly, J. P. McCrae, Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text, in: Proceedings of the 12th Forum for Information Retrieval Evaluation, FIRE '20, 2020.
- [4] B. R. Chakravarthi, N. Jose, S. Suryawanshi, E. Sherly, J. P. McCrae, 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), European Language Resources association, Marseille, France, 2020, pp. 177–184. URL: <https://www.aclweb.org/anthology/2020.sltu-1.25>.
- [5] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, J. P. McCrae, 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), European Language Resources association, Marseille, France, 2020, pp. 202–210. URL: <https://www.aclweb.org/anthology/2020.sltu-1.28>.

- [6] K. Pant, T. Dadu, Cross-lingual inductive transfer to detect offensive language, arXiv preprint arXiv:2007.03771 (2020).
- [7] B. Wang, Y. Ding, S. Liu, X. Zhou, Ynu_wb at hasoc 2019: Ordered neurons lstm with attention for identifying hate speech and offensive language., in: FIRE (Working Notes), 2019, pp. 191–198.
- [8] J. M. Struß, M. Siegel, J. Ruppenhofer, M. Wiegand, M. Klenner, et al., Overview of germeval task 2, 2019 shared task on the identification of offensive language (2019).
- [9] C. Bosco, D. Felice, F. Poletto, M. Sanguinetti, T. Maurizio, Overview of the evalita 2018 hate speech detection task, in: EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, volume 2263, CEUR, 2018, pp. 1–9.
- [10] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), arXiv preprint arXiv:1903.08983 (2019).
- [11] V. Basile, C. Bosco, E. Fersini, N. Debra, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, et al., Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter, in: 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 54–63.
- [12] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020), arXiv preprint arXiv:2006.07235 (2020).
- [13] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, in: Proceedings of the 11th Forum for Information Retrieval Evaluation, 2019, pp. 14–17.
- [14] P. Patwa, G. Aguilar, S. Kar, S. Pandey, S. PYKL, B. Gambäck, T. Chakraborty, T. Solorio, A. Das, Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets, arXiv preprint arXiv:2008.04277 (2020).
- [15] S. Banerjee, K. Chakma, S. K. Naskar, A. Das, P. Rosso, S. Bandyopadhyay, M. Choudhury, Overview of the mixed script information retrieval (msir) at fire-2016, in: Forum for Information Retrieval Evaluation, Springer, 2016, pp. 39–49.
- [16] G. Molina, F. AlGhamdi, M. Ghoneim, A. Hawwari, N. Rey-Villamizar, M. Diab, T. Solorio, Overview for the second shared task on language identification in code-switched data, arXiv preprint arXiv:1909.13016 (2019).
- [17] G. Chittaranjan, Y. Vyas, K. Bali, M. Choudhury, Word-level language identification using crf: Code-switching shared task report of msr india system, in: Proceedings of The First Workshop on Computational Approaches to Code Switching, 2014, pp. 73–79.
- [18] S. Sharma, P. Srinivas, R. C. Balabantaray, Text normalization of code mix and sentiment analysis, in: 2015 international conference on advances in computing, communications and informatics (ICACCI), IEEE, 2015, pp. 1468–1473.
- [19] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.
- [20] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, K. C. Chatzisavvas, Sentiment analysis leveraging emotions and word embeddings, Expert Systems with Applications 69 (2017) 214–224.
- [21] C. Goutte, E. Gaussier, A probabilistic interpretation of precision, recall and f-score, with implication for evaluation, International Journal of Radiation Biology Related Studies in Physics Chemistry Medicine 51 (2005) 952–952.