

SRJ @ Dravidian-CodeMix-FIRE2020: Automatic Classification and Identification Sentiment in Code-Mixed Text

Ruijie Sun, Xiaobing Zhou*

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

Abstract

Sentiment analysis of the Code-Mixed text has received increasing research attention. To facilitate the researches on Code-Mixed text, the Sentiment Analysis of Dravidian Languages in Code-Mixed Text is proposed in FIRE 2020. This paper introduces the system submitted by the SRJ team. We participate in the Malayalam-English task and Tamil-English task. We use XLM-Roberta as our model. And abundant semantic information is obtained by extracting XLM-Roberta's hidden state. Our approach achieves the best results in both tasks with weighted F-scores of 0.74 and 0.65, respectively. Our code is available on GitHub(<https://github.com/lonelyjie323/HASOC>).

Keywords

Sentiment Analysis, Abundant Semantic Information, Code-Mixed Text, XLM-Roberta

1. Introduction

With the application of the Internet in our daily life, more and more people communicate through social networking platforms. The use of social media such as Facebook, Twitter, and YouTube has made the dissemination and communication of information more and more rapid. Therefore, the sentiment analysis of language in social networks has attracted more and more attention.

Different people tend to use mixed language to express their feelings. However, considering the low adoption of some languages and the non-native scripts used by some social media users, it becomes difficult to perform existing natural language processing tasks, so our study is to analyze the way Dravidian Languages express emotions in the mixed scenario of social media codes. Malayalam-English is one of the Dravidian Languages used in southern India[1]. There are about 38 million people in India and other countries/regions who speak Malayalam-English [2]. Tamil-English is the language spoken in Sri Lanka and Tamil diaspora in India[3]. But because of the peculiarities of the two languages, social media users tend to use Roman scripts that are easy to type in, so the under-resourced languages on social media are mostly a mixture of code.

In the YouTube comments in the survey, there are a lot of mixed data of Malayalam-English,


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

EMAIL: zhouxb@ynu.edu.cn (X. Zhou*)

ORCID: 0000-0003-2987-5743 (R. Sun)



© 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)

Tamil-English, and English codes. Our goal is to categorize the comments obtained by YouTube into positive, negative, neutral, mixed emotions, or not in the intended languages[4].

2. Related Work

In recent years, sentiment analysis has attracted the attention of a large number of industrial and academic researchers. With the acceleration of information dissemination, Code-Mixed has gradually become a common phenomenon in multilingual communities.

Asoka Chakravarthi et al.[5] proposed an improvement of word meaning translation by utilizing orthographic information to translate words in languages with insufficient resources, such as Dravidian languages. A word in a language may have more than one meaning, for this reason, R. padamara[6] also proposed a word-level translation method based on knowledge engineering for Tamil-English. There are two traditional approaches to solve the problem of sentiment analysis, such as lexicon-based or machine learning approaches[7]. In general, it is difficult to understand and analyze texts written in multiple languages. Veena P V et al.[8] developed a word-level language recognition system based on Code-Mixed for social media text, which was used to complete Tamil-English and Malayalam-English Code-Mixed Facebook reviews. Shashi Shekhar et al.[9] proposed a novel architecture combining a multichannel neural network (MNN) and quantum Bi-directional Long Short-Term Memory (QBLSTM).

For this task, the content of sentiment analysis related to Malayalam-English and Tamil-English is very few[10], because for Indian language, machine translation between India and English has certain difficulties[11], so the data is not easy to obtain, let alone carry out a sentiment analysis on it.

3. Methodology

3.1. Dataset

In the Malayalam-English task, the official organizers provide training set (4,851 comments / posts) and validation set (540 comments / posts). In our experiment, we combine training set and validation set, and label distribution is { positive: 2,246, negative: 600, neutral: 1,505, mixed emotions: 333, not-Malayalam: 707 }, which is an imbalance data set.

In the Tamil-English task, the official organizers provide training set (11,335 comments / posts) and validation set (1,260 comments / posts). In our experiment, we combine training set and validation set, and label distribution is { positive: 8,124, negative: 1,613, neutral: 677, mixed emotions: 1,424, not-Tamil:397 }, which is an imbalance data set.

3.2. XLM-Roberta with hidden state

Early work in the field of cross-language understanding has proved the effectiveness of multilingual masked language model (MLM) in cross-language understanding, but models such as XLM[12] and Multilingual BERT[13] (pre-trained on Wikipedia) are still limited in learning useful representations of low resource languages. XLM-Roberta[14] shows that the performance of cross-language transfer tasks can be significantly improved by using the large-scale

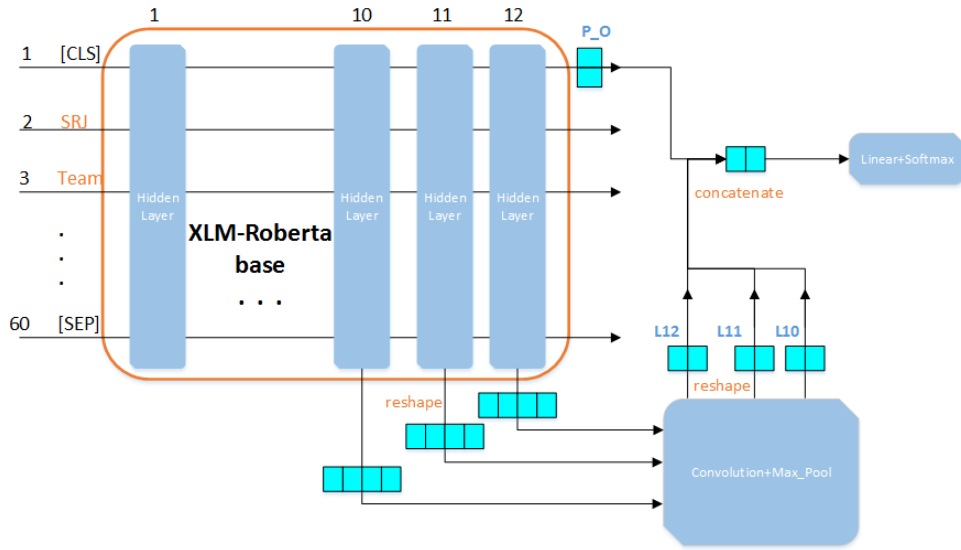


Figure 1: Our model (P_O is the pooler out of XLM-Roberta. It is obtained by its last layer hidden state of the first token of the sequence (CLS token) further processed by a linear layer and a tanh activation function. After extracting the hidden state of layer 12 of the XLM-Roberta, $L12$ is obtained through the convolution and max pooling. The same is true for $L11$ and $L10$.)

multi-language pre-training model. It can be understood as a combination of XLM and Roberta. It is trained on 2.5 TB of newly created clean CommonCrawl data in 100 languages. Because the training of the model in this task must make full use of the whole sentence content to extract useful semantic features, which may help to deepen the understanding of the sentence and reduce the impact of noise on data. Therefore, we use XLM-Roberta in this work.

In the classification task, the original output of XLM-Roberta is obtained through the last hidden state of the model. However, the output usually does not summarize the semantic content of the input. Recent studies have shown that abundant semantic information features are learned by the top hidden layer of BERT[15], which we call the semantic layer. In my opinion, the same is true of XLM-Roberta. Therefore, in order to make the model obtain more abundant semantic information features, we propose the following model, as shown in Figure 1. Firstly, we get P_O . Secondly, we extract the hidden state of the last three layers of XLM-Roberta and input them into CNN to get $L12$, $L11$, and $L10$. Finally, we concatenate P_O , $L10$, $L11$, and $L12$ into the classifier.

4. Experiments and Results

4.1. Preprocessing and Experiments Setup

In the experiment, we try to clean the text. But it does not achieve the desired results. So the text is not cleaned. In the Malayalam-English and Tamil-English tasks, the hyper-parameters are set to the same, and the best weight of the model is saved in the training. In this work, official organizers provide training sets and validation sets. We combine the official training set

Table 1

Precision, recall, F1-score and support for XLM-Roberta with P_O in Malayalam task and Tamil task (The validation set is the first fold in the 5-fold cross-validation)

XLM-Roberta with P_O in Malayalam task Validation set of 1-fold					XLM-Roberta with P_O in Tmail task Validation set of 1-fold			
	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
not-malayalam/ not-Tamil	0.78	0.79	0.78	141	0.67	0.70	0.68	73
Positive	0.70	0.75	0.72	450	0.75	0.88	0.81	1526
Negative	0.51	0.57	0.54	120	0.32	0.37	0.35	2879
unknown_state	0.61	0.63	0.62	301	0.25	0.13	0.17	122
Mixed_feelings	0.00	0.00	0.00	67	0.00	0.00	0.00	257
Macro avg	0.52	0.55	0.53	1079	0.40	0.42	0.40	2267
weighted acg	0.62	0.66	0.64	1079	0.58	0.67	0.62	2267

and the validation set to get the new data set, which is split into the new training set and the validation set by using the Stratified 5-Fold Cross-validation¹. Due to the imbalance of datasets, the Stratified 5-Fold Cross-validation ensures that the proportion of samples in each category in each fold data set remains unchanged.

For the XLM-Roberta, we use XLM-Roberta-base² pre-trained model, which contains 12 layers. We use Adam optimizer with a learning rate as 5e-5. The batch size is set to 32 and the max sequence length is set to 60. We extract the hidden layer state of BERT by setting the `output_hidden_states` is true. The model is trained in 10 epochs with a dropout rate of 0.2.

For the convolution layer, we use 2D convolution($nn.Conv2d^3$). The size of the convolution kernel is set to (3,4,5) and the number of convolution kernels is set to 256.

4.2. Results and Analysis

In this work, we find the limitations of P_O for sentiment analysis of Code-Mixed text in Dravidian languages. In the classification task, the original output of BERT is P_O . Chakravarthi et al.[2, 3] pointed out that Multilingual BERT fails to identify “Mixed feeling” class in the Malayalam-English task and Multilingual BERT fails to identify “Negative”, “Neutral”, “Mixed feeling” class in the Tamil-English task. In the same way, we just put P_O as the output of XLM-Roberta. The results are shown in Table 1. We can see that the results are not good when P_O is used as the output of Bert and XLM-Roberta. We think that just using P_O as the output will lose some effective semantic information. So we think that deep and abundant semantic features are effective for this work. We extract the hidden state of XLM-Roberta and we also discover that the performance of the model improves with the increase of the semantic layer. Table 2 shows the performance of our model at different semantic layers.

Table 3 shows our results on the test set. For two tasks, we only use the official training set

¹https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html-sklearn.model_selection.StratifiedKFold

²<https://huggingface.co/xlm-roberta-base>

³<https://pytorch.org/docs/stable/generated/torch.nn.Conv2d.html>torch.nn.Conv2d

Table 2

The performance of our model at different semantic layers (The validation set is the first fold in the 5-fold cross-validation)

	Malayalam Task Validation set of 1-fold	Tamil Task Validation set of 1-fold
	Macro/Weighted F1-score	Macro/Weighted F1-score
XLM-Roberta+Conv2d+L12	45.03/64.87	62.35/67.76
XLM-Roberta+Conv2d+L12+L11	46.49/65.55	64.56/68.56
XLM-Roberta+Conv2d+L12+L11+L10	47.52/66.78	65.78/69.78

Table 3

Results of Weighted F1 on Test set

Our Team	Malayalam Task			Tamil Task		
	Precision	Recall	Weighted F1-score	Precision	Recall	Weighted F1-score
SRJ	0.74	0.75	0.74	0.64	0.67	0.65

and validation set and do not use any external data. The hyper-parameters of the model are set to the same. In the Malayalam-English and Tamil-English tasks, our models all achieve the best performance.

5. Conclusion

In this work, we provide a baseline for sentiment analysis of Code-Mixed text in Dravidian languages (Malayalam-English and Tamil-English). We find the limitation of only using Pooler out as the output of BERT. To obtain deeper and more abundant semantic features, we extract the hidden layer state of XLM-Roberta, which is input into convolution and max pooling. The result shows that it is helpful to improve the performance of XLM-Roberta to obtain more abundant semantic information features by extracting the hidden state of XLM-Roberta.

References

- [1] M. Haridas, N. Vasudevan, G. J. Nair, G. Gutjahr, R. Raman, P. Nedungadi, Spelling errors by normal and poor readers in a bilingual malayalam-english dyslexia screening test, in: 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT), IEEE, 2018, pp. 340–344.
- [2] 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>.
- [3] 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>.

- [4] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, J. P. Sherly, Elizabeth 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.
- [5] B. R. Chakravarthi, Leveraging orthographic information to improve machine translation of under-resourced languages, Ph.D. thesis, NUI Galway, 2020.
- [6] R. Padmamala, Word level translation (tamil-english) with word sense disambiguation in tamil using ontnet, in: 2015 International Conference on Computing and Communications Technologies (ICCCT), IEEE, 2015, pp. 191–198.
- [7] O. Habimana, Y. Li, R. Li, X. Gu, G. Yu, Sentiment analysis using deep learning approaches: an overview, *Science China Information Sciences* 63 (2020) 1–36.
- [8] P. Veena, M. A. Kumar, K. Soman, An effective way of word-level language identification for code-mixed facebook comments using word-embedding via character-embedding, in: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2017, pp. 1552–1556.
- [9] S. Shekhar, D. K. Sharma, M. S. Beg, Language identification framework in code-mixed social media text based on quantum lstm—the word belongs to which language?, *Modern Physics Letters B* 34 (2020) 2050086.
- [10] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, J. P. Sherly, Elizabeth 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.
- [11] M. A. Kumar, B. Premjith, S. Singh, S. Rajendran, K. Soman, An overview of the shared task on machine translation in indian languages (mtil)–2017, *Journal of Intelligent Systems* 28 (2019) 455–464.
- [12] G. Lample, A. Conneau, Cross-lingual language model pretraining, *arXiv preprint arXiv:1901.07291* (2019).
- [13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).
- [14] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, *arXiv preprint arXiv:1911.02116* (2019).
- [15] G. Jawahar, B. Sagot, D. Seddah, What does bert learn about the structure of language?, 2019.