YUN111@Dravidian-CodeMix-FIRE2020: Sentiment Analysis of Dravidian Code Mixed Text

Yueying Zhu[a], Kunjie Dong[a]

[a]*School of Information Science and Engineering, Yunnan University, Yunnan, P.R. China*

Abstract

The use of social media has grown rapidly during the past few years, which provides a convenient platform for users to communicate by inserting their native language into English and results in a large amount of code mixed text. This way of communication is not only convenient but also reduces people's knowledge burden. They can express their opinions most simply and easily. However, it is quite difficult for non-native speakers to understand these code mixed texts. Therefore, it is important to analyze the sentiment expressed in these texts. This paper reports on our work in Dravidian- Codemix-Fire 2020[1]. We propose a sentiment analysis mBERT-based model, and use self-attention to assign a weight to the output of the BiLSTM, which further improve the analysis accuracy of the model. We test our model on the data sets released by the organizers, and the performance of our system is very close to the best system in the competition. We achieve the weighted average F1-scores of 0.73 and 0.64 in Malayalam and Tamil languages, respectively, and both rank 2^{nd}.

Keywords

Code mixed, Dravidian language, Sentiment analysis, Malayalam, Tamil

1. Introduction

There is an increasing number of people who will express their opinions on social media, including comments on a new movie, expectations, suggestions for a new product, or the responses to a new policy of the government. People will often use their most familiar native language mixed with English to form code mixed text because this relaxed communication way reduces their burden of language as a cognitive process. This is also why the code mixed is becoming more and more common on social media. But for those who are not the native speakers, fully understanding this code mixed texts can be a headache. Therefore, it is important to provide a systematic approach to fully dig the sentiment information given by the code mixed text.

Sentiment analysis is the task of determining subjective opinions or responses about a given topic. For the past two decades, it has been an active area of academic and industrial research. There is a growing need for sentiment analysis on social media code mixed text [1]. Code mixing is a common phenomenon in the multilingual community and the code mixed text is sometimes written in non-native scripts. Systems that train on monolingual data fail on code mixed data because of the complexity of switching code between different language levels in

[1] https://competitions.codalab.org/competitions/25215

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

EMAIL: kunjiedong@qq.com (K. Dong)

CEUR Workshop Proceedings (CEUR-WS.org)

text [2]. The complexities of this type of language include the presence of multilingual words, transliteration, spelling variations, and so on.

This common task proposes a new gold standard corpus for sentiment analysis of code mixed text in Dravidian (Malayalam-English and Tamil-English). This is the message level polarity classification task. Given a YouTube comment, the system must categorize it into negative, not-Malayalam (or not-Tamil), positive, unknown-state or mixed-feeling. We introduce a sentiment analysis mBERT-based model (mBERT means BERT multilingual model) and use self-attention to assign a weight to the output of the BiLSTM, which further improve the analysis accuracy of the model. We also test our model on the data sets released by the organizers, and the performance of our system is very close to the best system in the competition. We achieve the weighted average F1-scores of 0.73 and 0.64 in Malayalam and Tamil languages, respectively, and both rank 2^{nd}. Our code is available on GitHub[1]

2. Related Work

Code mixed text sentiment analysis seminal work was done by Joshi et al. [3] on Hindi text. The sub-word level representations in LSTM (Sub word-LSTM) architecture have been proposed by Prabhu et al. [4] on Hindi code mixed text. Shalini et al. [5] created the Kannada-English code mixed corpus and provided distributed representation methods. Jhanwar et al. [6] combined an ensemble model of character-n based LSTM and word-n based MBN to identify the sentiment of Hindi-English code mixed data. The multilayer perceptron model has been used to identify the polarity of the Bengali-English tweets by Ghosh et al[7]. Vijay et al. [8] proposed a supervised classification system that used a variety of machine learning techniques to detect sentiment. Lal et al. [9] proposed a hybrid architecture for the sentiment analysis task in English-Hindi code mixed data. In [10], the author proposed a POS tag method for code mixed social media text the recursive neural network-based language model (RNN-LM) architecture.

As mentioned above, there has been a lot of research on sentiment analysis in many different code mixed types of languages. [11] is the first task about the sentiment analysis of code mixed text in Dravidian language (Malayalam-English and Tamil-English).

3. Model Architecture

In this section, after a lot of experiments and comparisons that we choose the pre-trained methods. Which is the mBERT-based model (mBERT means BERT[2] multilingual model) [12]. It is a pre-trained model of the mBERT for multiple languages, because mBERT not only can predict and train the next sentence, it can also help process the logical relationship between two sentences. On the other hand, mBERT's high hidden layer can learn rich semantic information characteristics. Therefore, to obtain richer semantic information features [13], we also make use of the top hidden layer state output of the mBERT and feed it to BiLSTM. Then we give a weight for all hidden state output of the BiLSTM which helps in better classification of the polarity of sentiment in multiple sentiment bearing units, and the output of the attention model

[1] https://github.com/TroubleGilr/codemixed

[2] https://github.com/google-research/bert

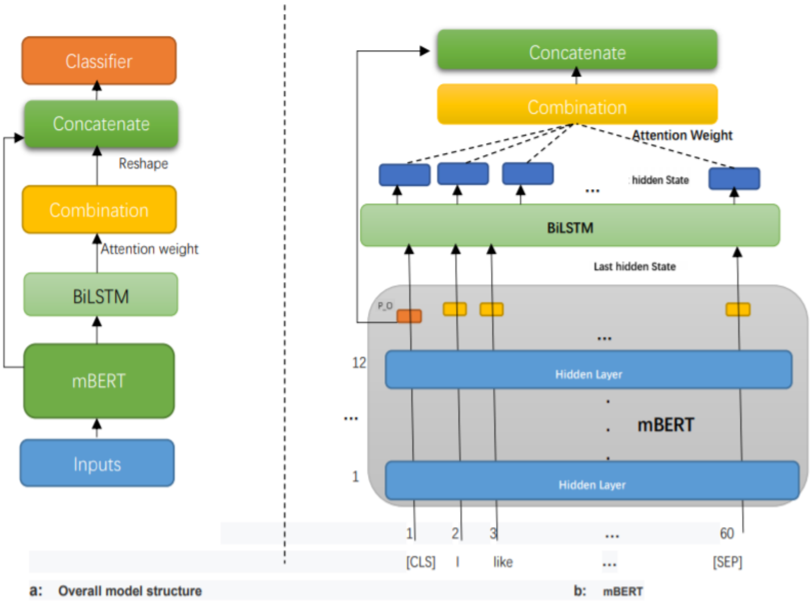

Figure 1: Schematic overview of the architecture

is a weighted sum of hidden representations at each hidden state. Finally, we concatenate the original output of mBERT with the output weighted representation vector.

The principle of weight attention

Formally, let C be the set of characters and T be the set of input statements. The sentence s ∈T can be made of characters by $[c_1, ..., c_l]$ where l is the length of input, and the l = 60 in our model. The i^{th} forward and backward hidden states of the BiLSTM are represented $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$, respectively. Concatenating $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ obtains the annotation k_i:

$$k_i = [\overrightarrow{h_i}; \overleftarrow{h_i}], (i = 1, 2, ..., l) \tag{1}$$

The attention weight a_i of k_i in the sentence s can be calculated by this following formula:

$$x = \exp(k_i{}^T k_l) \tag{2}$$

$$y = \sum_{j=1}^{l} \exp(k_i{}^T k_l) \tag{3}$$

$$a_i = \frac{x}{y} \tag{4}$$

And then we get the representation vector h by combination all the weighted outputs:

$$h = \sum_{i=1}^{l} a_i x k_i \tag{5}$$

Table 1
Class-wise statistics of the dataset for train, validation, and test set.

Languag	Split	Total	Neg	Pos	Not	Unknown	Mixed
Malayalam	Train	4851	549	2022	647	1344	289
	Validation	540	51	224	60	161	44
	Test	1348	138	565	177	398	70
Tamil	Train	11335	1448	7627	368	609	1283
	Validation	1260	165	857	29	68	141
	Test	3149	424	2075	100	173	377

Next, we do a reshape for h that let it keep the same dimensions with the P_O output of mBERT. Finally, we concatenate them into the classifier. The schematic overview of the model architecture was shown in a of Figure 1.

4. Data Description

The aim of this task is to identify the sentiment polarity of a code mixed dataset of comments/posts in Dravidian (Malayalam and Tamil) collected from social media. Comments/posts may contain more than one sentence, but the average sentence length of the corpus is 1. The datasets consist of YouTube comments labeled into one of the five classes:

Negative(Neg): Tweets contain obvious emotions such as sadness, dissatisfaction, loss, or offensive language. To express disgust or criticize some people.

Positive(Pos): A tweet that expresses happiness, satisfaction, praise for a person, group or country.

Not-Malayalam or Not-Tamil(Not): For the language of Malayalam or Tamil, if the sentence does not include either Malayalam or Tamil, then it is not Malayalam or Tamil.

Unknown-state(Unknown): Tweets that represent facts, provide news, or belong to advertisements. There's no obvious emotional expression.

Mixed-feeling(Mixed): Tweets explicitly or implicitly contain the user's emotions.

Most datas given are written in Roman script, which have the mixture of these forms of code-mixed sentences –Inter-Sentential switch, Intra-Sentential switch and Tag switching. The specific data set details are given in table 1. The various types of data are unbalanced from the table and more details about the dataset can be found in [14] and [15], and some of the processing of code mixed text can be seen in [16].

5. Experiments and Results

In this work, we use mBERT-based as our pre-training model. Before the training, we randomly shuffle the data and remove unwanted characters and emoticons (Generally, emoticons express specific sentiment, and we will consider introducing an emoticons system in our future work). Especially, the label of categorical sentiment values we encode as 0,1,2,3,4 to negative, not-Tamil(not-Malayalam), positive, unknown or mixed-feelings, respectively. This way is to give a

Table 2
Description of the results of the data sets given by the organizer

Language	Precision	Recall	F-Score	Rank
Malayalam-English	0.73	0.73	0.73	2
Tamil-English	0.63	0.67	0.64	2

numeric representation to the categorical sentiment data. Finally, we input the processed data into our model. And here mBERT uses the *WordPiece*[3] tool [17] for word segmentation and inserts special separators ([CLS], which separates each sample) and separator ([SEP], which separates different sentences in the sample) [18].

Here our model is implemented based on Pytorch. We use Adam optimizer with a learning rate of 1e-5 and *Cross-Entropy Loss*. The batch size is set to 8 and the *gradient accumulation steps* is set to 4. The epochs and maximum length of the sentence are 5 and 60, respectively.

The measurement to evaluate the participating system is the weighted average F1-scores. Table 2 shows the precision, recall and the weighted average F1-scores of our system. The final ranking is based on the weighted average F1-scores and our submit system obtains the weighted average F1-scores of 0.73 and 0.64 in Malayalam and Tamil languages, respectively.

6. Conclusion

With the increase of the social media text in popularity and influence, it is increasingly important to analyze the sentiment attached to the text. In this paper, we perform a sentiment analysis on Dravidian Languages (Malayalam-English and Tamil-English). We propose an mBERT-based multilingual processing model. We also give the weights for the hidden state output of the BiLSTM, and then we concatenate it with the original output of the mBERT. Our system achieves very satisfying performance.

On the whole, the article uses mBERT to represent the code-mixed Dravidian text, which has been feed to the BiLSTM (creates attention weighted vector representation of the vector). In the end, the output of BiLSTM and attention layer of mBERT are concatenated for the classification. This framework is well-combined with the help of both high level and low level. Through this model, rich sentiment information is extracted to improve the classification accuracy. In future work, we will consider incorporating emotional information into the classification system. On the other hand, we consider the depth of the model as far as the data allows.

Acknowledgments

I would like to thank the organizers for their hard work and the teachers for their help. Finally, I would like to thank the school for its support to my research and the future reviewers for their patient work.

[3] https://github.com/lovit/WordPieceModel

References

[1] 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.

[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: Proceedings of the 12th Forum for Information Retrieval Evaluation, FIRE '20, 2020.

[3] A. Joshi, P. Bhattacharyya, A. R. Balamurali, A fall-back strategy for sentiment analysis in hindi: a case study, in: Icon, 2010.

[4] A. Prabhu, A. Joshi, M. Shrivastava, V. Varma, Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text (2016).

[5] A. Joshi, P. Bhattacharyya, B. R, A fall-back strategy for sentiment analysis in hindi: a case study, 2010.

[6] M. G. Jhanwar, A. Das, An ensemble model for sentiment analysis of hindi-english code-mixed data (2018).

[7] S. Ghosh, S. Ghosh, D. Das, Sentiment identification in code-mixed social media text (2017).

[8] D. Vijay, A. Bohra, V. Singh, S. S. Akhtar, M. Shrivastava, Corpus creation and emotion prediction for hindi-english code-mixed social media text, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, 2018.

[9] Y. K. Lal, V. Kumar, M. Dhar, M. Shrivastava, P. Koehn, De-mixing sentiment from code-mixed text, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 2019.

[10] R. N. Patel, P. B. Pimpale, M. Sasikumar, Recurrent neural network based part-of-speech tagger for code-mixed social media text (2016).

[11] D. S. Nair, J. P. Jayan, R. R. R, E. Sherly, Sentima - sentiment extraction for malayalam, in: International Conference on Advances in Computing, 2014, pp. 1719–1723.

[12] G. Jawahar, B. Sagot, D. Seddah, What does bert learn about the structure of language?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[13] W. Dai, T. Yu, Z. Liu, P. Fung, Kungfupanda at semeval-2020 task 12: Bert-based multi-task learning for offensive language detection (2020).

[14] 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.

[15] 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.

[16] B. R. Chakravarthi, Leveraging orthographic information to improve machine translation of under-resourced languages, Ph.D. thesis, NUI Galway, 2020.

[17] B. C. Xu, L. Ma, L. Zhang, H. H. Li, M. C. Zhou, An adaptive wordpiece language model for learning chinese word embeddings, in: 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), 2019.

[18] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding (2018).