Will_go at SemEval-2020 Task 9: An Accurate Approach for Sentiment Analysis on Hindi-English Tweets Based on Bert and pseudo Label Strategy

Wei Bao[1],Weilong Chen[2], Wei Bai[3], Yan Zhuang[4],Mingyuan Cheng[5], Xiangyu Ma[6]
[1,6]Southeast University
[2,3,4]University of Electronic Science and Technology of China
[5]Renmin University of China
{willinseu,chenweilong921,cellurbw,delecisz}@gmail.com
{mundane0827,xma2182}@gmail.com

Abstract

Mixing languages are widely used in social media, especially in multilingual societies like India. Detecting the emotions contained in these languages, which is of great significance to the development of society and political trends. In this paper, we propose an ensemble of pseudo-label based Bert model and TFIDF based SGDClassifier model to identify the sentiments of Hindi-English (Hi-En) code-mixed data. The ensemble model combines the strengths of rich semantic information from the Bert model and word frequency information from the probabilistic ngram model to predict the sentiment of a given code-mixed tweet. Finally, our team got an average F1 score of 0.686 on the final leaderboard, and our codalab username is will_go.

1 Introduction

The rapid development of modern social media makes it possible for people to express their opinions almost at any time. Detecting the sentiments of these views can roughly analyze the social and economic development of the user's current region, the psychological characteristics, interests and hobbies of people of different ages, so as to facilitate the company to better deliver information. It is also possible to judge the emotion and psychological characteristics of a certain type of users according to their statements, so as to eliminate some adverse factors to the society. Many companies put efforts to concern customer feedback to achieve better product optimization, and then win the trust of consumers.

However, in social media, speech is not required to conform to certain norms, so there are differences in language and format. In some multilingual societies, when different languages are mixed, it is even more difficult to detect the sentiments. In India and Spain, a number of bi-lingual hybrids produced many texts. The process of switching text between two or more languages is called code-mixing and a significant part is the mixing of the native language and English. The difference of language habits makes the problem even harder. A quite number of people prefer to use nonstandard words like coooool rather than cool, thx instead of thanks, which causes obstacles.

Code mixing has always been one of the most important directions of natural language processing, and there is a lot of valuable research and a lot of good results in language identifying, POS tagging and Named Entity Recognition of code-mixed from a lot of researchers (Bali et al., 2014; Kumar et al., 2018). LSTM model shows great results in the sentiment analysis of code-mixing (Prabhu et al., 2016). However, the noise of the code mixed data exists.

The model we propose in this paper combines multi-sample-dropout and pseudo label based on the BERT. Besides, TDIDF is also adopted to get better performance. We compare different models to get a higher F1 and the results show that the model BERT with pseudo label performs well when the size of batch is 16. When the n-gram ranges from 1 to 3 helps the model more accurate. Overall solutions can be seen in (Patwa et al., 2020)

The rest of the paper is organized as follows. The overview of sentiment analysis in code-mixed data is shown in section 2. And in section 3, we explain the details of our model and the corresponding frames. Section 4 shows the results of our models.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1348–1353
Barcelona, Spain (Online), December 12, 2020.

2 Related Works

The phenomenon of mixed-use of mixing languages in the world is becoming more and more frequent. Although most people use English on social media, many of them are mixed with languages other than English. This poses a great challenge to traditional natural language processing. The combination of mixing languages and codes brings some invisible difficulties to natural language processing(NLP) tasks, such as word-level language identification, part-of-speech tagging, dependency parsing, machine translation and semantic processing. In recent years, there have been many studies aimed at mixing language processing, and they have made significant contributions to mixing languages or mixing code processing.

In terms of language recognition, Banerjee et al. (Banerjee et al., 2014), Mandal et al. (Mandal et al., 2015), Barman et al. (Barman et al., 2014), Das and Gamback (Das and Gambäck, 2015) have done researches on mixing codes processing. And Vyas et al. (Vyas et al., 2014), Jamatia et al. (Jamatia et al., 2015), Gupta et al. (Gupta et al., 2017) took a big step in the part-of-speech tagging. Wang et al. (Wang et al., 2018) proposed three strategies to improve the multilingual translation framework quality of multilingual neural translation models. However, none of these works can effectively solve the problem of sentiment analysis in code mixed data.

With the rapid development of the Internet and the emergence of various social media, sentiment analysis has become increasingly important. The importance of sentiment analysis or opinion mining for the entire business and society has expanded from computer science to management science and social science, such as marketing, finance, political science, communication, health, and even history. Since the beginning of 2000, sentiment analysis has become one of the most active research areas in the field of NLP. In recent years, there has been a breakthrough in deep learning. The sentiment analysis task in the NLP field has gradually introduced this method and has formed many best results in the industry. Zheng and Xia (Zheng and Xia, 2018) adopted the LSTM model based on context2target attention, and conducted targeted sentiment analysis by capturing the most important words in the left and right context. Joshi et al. (Prabhu et al., 2016) used sub-word-level representation in the LSTM architecture, which produced the most advanced results compared to other traditional machine learning models and models based on word polarity. Traditional deep learning models often have certain limitations in processing speed and accuracy. So we came up with a model that combines multi-sample dropout and BERT-based TFIDF, and the model performs very well.

3 Methodology

Our model is based on the Bidirectional Encoder Representation from Transformer (BERT) and Frequency-Inverse Document Frequency (TF-IDF) for detecting sentiments. Bert can learn the semantic relationship in the text, and TFIDF technology can better learn the relationship between word frequency and label. The difference between the two provides theoretical support for the model integration we use later. In addition, in the Bert model, we adopted pseudo label and multi-sample dropout, both of which make the model accuracy much improved. Pseudo label was adopted during the training. We divide the training data into 5 folds and use the cross-validation for improving the training. A multi-sample dropout was adopted to avoid overfitting and accelerate training and improve generalization. Our model was conducted with the following details.

3.1 Bert Model with Multi-Sample Dropout and pseudo Label Strategy

Bert model is a commonly used pre-training model technique in the NLP problem recently, but on the premise of the mixed Hindi-English language of this track, it puts forward higher requirements for the Bert model. In this track, we tried the traditional English Training model and multilingual Bert model, from the perspective of the model on the test set, the traditional Bert model is better. The introduction of the two is shown as follows.

BERT is proposed in the google research team (Devlin et al., 2018). This is a method of pre-training language representation that trains a general language understanding model on a large number of textual data, and then uses this model to perform the desired NLP task, the structure is shown in Figure 1. Two

steps, pre-training and fine-tuning, are contained. During former one, unlabeled texts is trained. All the parameters are initialized in the pre-training process and then fine-tuned in the detecting sentiments in the latter period. In this task, we did experiments with different models, BERT has higher learning rates than other models on detecting sentiments and chooses BERT as our basic model.

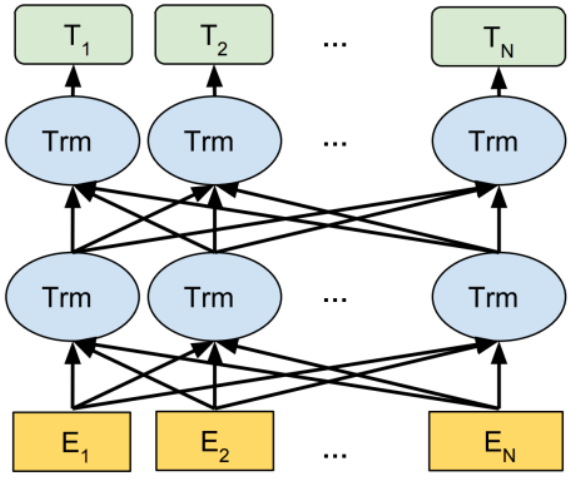

Figure 1: Bidirectional transformer architectures of BERT

Multilingual BERT aims to treat all target languages with the same model and weight (Pires et al., 2019). Multilingual BERT builds a vocabulary that includes all the target languages to avoid different words in different languages. The structure of the Multilingual BERT encoding part is the same as BERT and is shared among all languages. Task in the process of the training is still the BERT masked language model in which each sample is a single paragraph, all language interval training in the form of multitasking which does help in migrating across languages. BERT is more frequently used when certain language is more than others.

Dropout is commonly and regularly used in deep neural network(DNN). Dropout randomly ignores some neural to avoid overfitting. Multi sample dropout creates multiple dropout samples and then averages the loss of all the samples to get the final loss while in the traditional dropout selects a random set of samples from the input for each round of training (Inoue, 2019). This approach simply replicates parts of the training network after the dropout layer and shares the weights between those replicated full connection layers, without the need for new operators. This proposed method can be easily implemented in to our language model. After adding multi sample dropout to BERT, the model stability improved a lot.

Pseudo-label technology is a common semi-supervised learning method for expanding data. After the model is trained with the labeled training set, the label of the test set is predicted. We call part of the test set data and the predicted labels as pseudo data, and then they are added to the original training set for training. Labeled data often means high cost and difficulty to obtain, but unlabeled data is large and cheap. The pseudo label plays an important role in improving the accuracy of decision boundaries and the robustness of models.

3.2 SGDClassifier Model with Char-level and Word-level Combined TFIDF Method

TF-IDF is a commonly used weighting technique for information retrieval and data mining, used to evaluate the importance of a word (Ramos and others, 2003). TF indicates the frequency of entries in the document. The main idea of IDF is: if there are fewer documents containing the term t, the IDF is larger, which means that the term t has a good ability to distinguish between categories, suitable for classification. We used TF-IDF in the classification process and continuously fine-tune parameters. Finally, we chose word n-gram = 1-3 and char n-gram = 4-6, and concatenate them to achieve a good performance, the flow is shown in Figure 2.

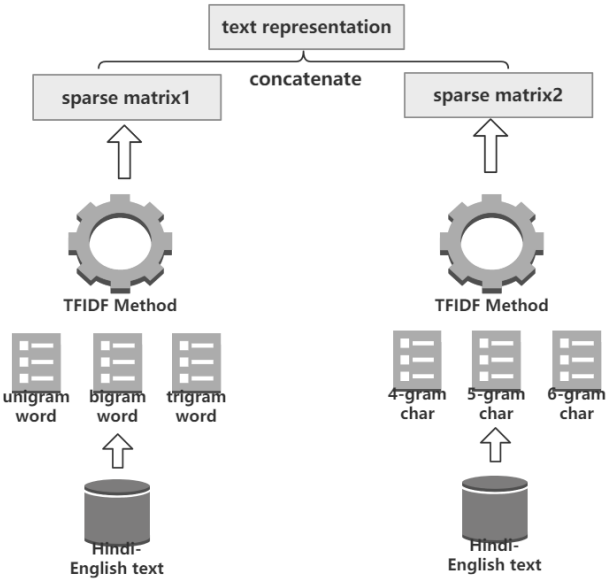

Figure 2: Flow of text representation based on TFIDF

SGDClassifier refers to the method of stochastic gradient descent for classification. SGD is mainly used in large-scale sparse data problems, often used in text classification and natural language processing (Kabir et al., 2015). Each iteration randomly extracts one sample from the training set. In the case of extremely large sample sizes, it may not be necessary to extract all samples to obtain a model with a loss value within an acceptable range.

3.3 Ensemble Methodology

The ensemble is a machine learning method that integrates a variety of basic models or weak classifiers to complete the final decision, which can improve the accuracy of various machine learning tasks. Here, we used the blending method, trained different base models with disjoint data, and averaged (weighted) their outputs (Grady et al., 1999). Blending linearly merges the results of the learned basic learner without cross-validation of k times to obtain the stacker feature, so it is simpler and can prevent information leakage.

4 Experiments and Results

The final model results are shown in Table 1. Although the corpus (Patwa et al., 2020) is a Hindi-English mixed language, compared with multilingual Bert, the model pre-trained by English still performs better, because the English words still account for a large proportion in the dataset. We use a batch size of 16 and fine-tune for 7 epochs over the code-mixed data and the fine-tuning learning rate is 5e-5. The ensemble model refers to the blending of Bert model with pseudo label and TFIDF model with a weight of 1: 1, which was developed after exhausting the leaderboard.

Approach	**Negative-Class F1 Score**	**Neutral-Class F1 Score**	**Positive-Class F1 Score**	**Leaderboard F1 Score**
Multilingual BERT	0.690	0.636	0.734	0.685
SGDClassifier with TFIDF	0.691	0.634	0.737	0.686
Bert	0.696	0.668	0.783	0.715
Bert with pseudo label	0.738	0.658	0.787	0.725
Ensemble	0.746	0.667	0.789	0.731

Table 1: Test set results for different models

5 Conclusion

Sentiment analysis is one of the important means for us to understand society, especially the processing of multiple languages. In this paper, we used Bert, Multilingual Bert, SGD classifier with TF-IDF and ensemble method to analyze the sentiment of English documents mixed with India. The Bert model with pseudo label strategy performed best,and the ensemble model reached 0.686 in the leaderboard. The results of the competition showed that our model has a good effect on the sentiment analysis of mixing languages. However, there are still some deficiencies in model optimization and parameter setting. In the future, we hope to be able to process documents mixed in more languages, optimize and extend the method to more aspects to learn more rich features in various languages.

References

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126.

Somnath Banerjee, Alapan Kuila, Aniruddha Roy, Sudip Kumar Naskar, Paolo Rosso, and Sivaji Bandyopadhyay. 2014. A hybrid approach for transliterated word-level language identification: Crf with post-processing heuristics. In *Proceedings of the Forum for Information Retrieval Evaluation*, pages 54–59.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the first workshop on computational approaches to code switching*, pages 13–23.

Amitava Das and Björn Gambäck. 2015. Code-mixing in social media text: the last language identification frontier?

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Joseph Grady, Todd Oakley, and Seana Coulson. 1999. Blending and metaphor. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pages 101–124.

Deepak Gupta, Shubham Tripathi, Asif Ekbal, and Pushpak Bhattacharyya. 2017. Smpost: parts of speech tagger for code-mixed indic social media text. *arXiv preprint arXiv:1702.00167*.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *arXiv preprint arXiv:1905.09788*.

Anupam Jamatia, Björn Gambäck, and Amitava Das. 2015. Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages. Association for Computational Linguistics.

Fasihul Kabir, Sabbir Siddique, Mohammed Rokibul Alam Kotwal, and Mohammad Nurul Huda. 2015. Bangla text document categorization using stochastic gradient descent (sgd) classifier. In *2015 International Conference on Cognitive Computing and Information Processing (CCIP)*, pages 1–4. IEEE.

Upendra Kumar, Vishal Singh, Chris Andrew, Santhoshini Reddy, and Amitava Das. 2018. Consonant-vowel sequences as subword units for code-mixed languages. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Soumik Mandal, Somnath Banerjee, Sudip Kumar Naskar, Paolo Rosso, and Sivaji Bandyopadhyay. 2015. Adaptive voting in multiple classifier systems for word level language identification. In *FIRE workshops*, pages 47–50.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December. Association for Computational Linguistics.

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

Ameya Prabhu, Aditya Joshi, Manish Shrivastava, and Vasudeva Varma. 2016. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. *arXiv preprint arXiv:1611.00472*.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142. Piscataway, NJ.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. Pos tagging of english-hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 974–979.

Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2955–2960.

Shiliang Zheng and Rui Xia. 2018. Left-center-right separated neural network for aspect-based sentiment analysis with rotatory attention. *arXiv preprint arXiv:1802.00892*.