KDELAB at SemEval-2020 Task 12: A System for Estimating Aggression of Tweets Using Two layers of BERT Features

Keisuke Hanahata, and Masaki Aono
Department of Computer Science and Engineering
Toyohashi University of Technology, Toyohashi, Aichi, Japan.
hanahata@kde.cs.tut.ac.jp and aono@tut.jp

Abstract

In recent years, with the development of social network services and video distribution services, there has been a sharp increase in offensive posts. In this paper, we present our approach for detecting hate speech in tweets defined in the SemEval- 2020 Task 12. Specifically, we describe our ensemble system, based on "BERT Large", having two parallel neural net component models, where one component model consists of MLP with input taken from the 23rd layer output from BERT, while another component model consists of bidirectional LSTM with input taken from the 1st layer of BERT.

1 Introduction

The number of offensive posts has been increasing rapidly due to the wide spread of social networking services such as Facebook, Twitter, and YouTube, where obvious and/or latent offensive texts are accompanied. It is necessary to filter the offensive postings because they can offend users and reduce their satisfaction with the service. Manual filtering is expensive, laborious, and time-consuming. Thus, the demand for automatic filtering has been increasing year by year.

Marcos et al. presented SemEval2020 Task12 (Zampieri et al., 2020), which includes three sub-tasks. Sub-task A is an estimate of the aggressiveness of a tweet, Sub-taskB is an estimate of whether a tweet is attacked, and Sub-task C is an estimate of the target of a tweet's attack.

2 Related Work

A number of studies have been conducted to estimate the aggressiveness of documents. One study classifies Facebook and Twitter posts into three classes: explicitly offensive, covertly offensive, and non-offensive (Kumar et al., 2018), and another identifies hate speech. At Kaggle, a competition was held to classify offensive comments into six classes: toxic, more toxic, obscene, threatening, insulting, and personal attack. In Task 6 of SemEval2019 (Zampieri et al., 2019b), a competition was held for Twitter posts using a dataset with hierarchical labeling focusing on the target of the attack.

3 Proposed Method

In the following section, we describe our proposed model for estimating the aggressiveness and the targets of tweets.

3.1 Preprocessing

We perform preprocess, given all incoming tweets. The atmark, hash, and "URL" tokens associated with user IDs are removed because they are considered to interfere with the conversion of tweets into features. The "USER" token is left out because it can be a clue to estimate the target of the attack. Many of the tweets contain emojis. Emojis can contain information about the attacker or the target of an attack. The emoji library is used to convert emojis into text.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 2030–2034
Barcelona, Spain (Online), December 12, 2020.

3.2 Feature extraction using BERT

We think that information related to word aggressiveness as well as word relatedness to other word is important for estimating the aggressiveness and target of a sentence. Thus, we transform tweets into features using a general-purpose language model, BERT (Devlin et al., 2019).

In our system, we use a pre-trained model "BERT-large Uncased" to obtain the features of the tweets. The above model consists of 24 Transformer layers in total. In each layer, input data is converted to a feature whose relationship between words is emphasized by the attention mechanism.

We describe the flow of feature extraction. First, the pre-processed tweets are fed into the BERT. We set the maximum length of the input word to 25. Each word is converted to a feature of 1024 dimensions, and the contextual features between words are gradually emphasized in the Transformer layer. The features obtained in each layer are 25×1024 dimensions.

3.3 Classification models

The features presented in the previous section are fed to the following two models as the main components of our proposed system. In a sense, our model can be regarded as an ensemble of two models.

3.3.1 MLP model

The BERT features are fed to the multi-layer perceptron and a classification model is constructed. We extract the output of the 23rd layer of the BERT model, which is considered to emphasize word-to-word attention. The reason why we do not use the output of the 24th layer is that it is likely to extract features that are biased against the BERT pre-training task. We take the total average of the obtained 25×1024-dimensional features, yielding 1024 dimensional features, by inserting "Average Pooling" layer prior to the multilayer perceptron to be trained. We consider the MLP component as one classification model.

3.3.2 Bidirectional LSTM model

Alternatively, the BERT features are fed to "Bidirectional-LSTM" to construct another classification model. Here we take the output of the first layer extracted from the BERT. The reason behind this idea is because we think such a feature, closer to the embedded representation of the word itself (25 x 1024 dimensions), might contribute to capture sequential orders of "words", appropriate for LSTM-like recurrent neural network.

3.3.3 Ensemble model

In our system, as described above, two component models are incorporated. The important part of making an ensemble model is how to connect them. After trial and error, we adopt element-wise sum of the class probabilities. Then, we ensemble the resulting class probabilities with high accuracy. The network diagram is shown in Figure 1.

4 Experiments and Evaluations

4.1 Task Description

This section describes the three types of tasks to tackle.

4.1.1 Sub-task A : Offensive language identification

This task estimates the aggressiveness of a tweet. We classify tweets into two classes: tweets with no aggression (NOT) and tweets with aggression (OFF).

4.1.2 Sub-task B : Automatic categorization of offense types

This task estimates whether a tweet is targeted for attack. We classify tweets into two classes: Tweets with an Attack Target (TIN) and Unattackable Tweets (UNT).

4.1.3 Sub-task C : Offense target identification

This task estimates the target of a tweet's attack. We classify tweets into three classes: Tweets targeted at individuals (IND) , Tweets targeted at the group (GRP) , Tweets targeted at others (OTH) .

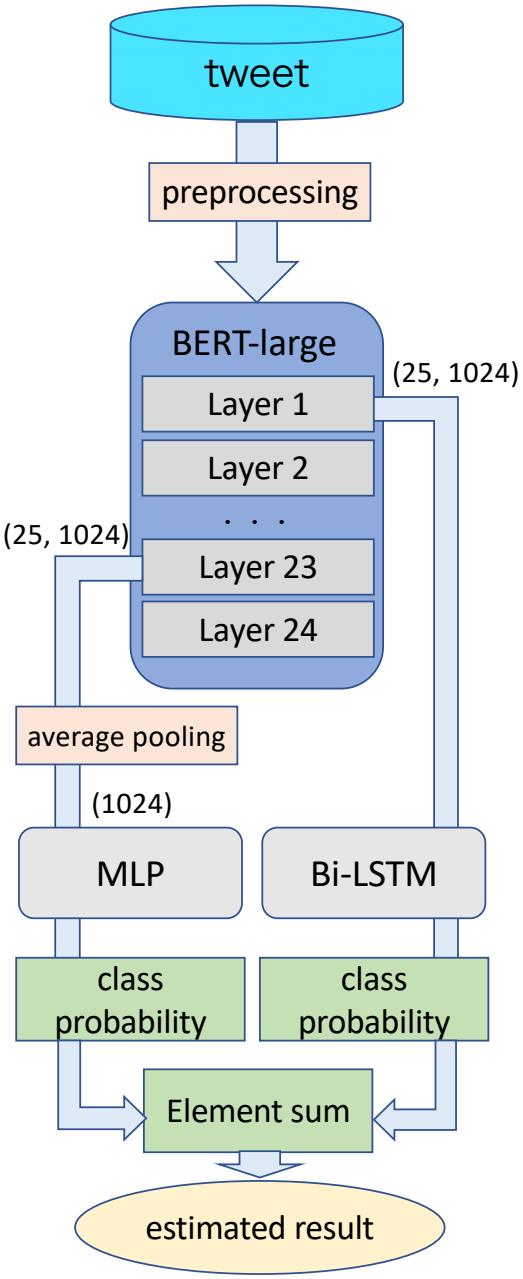

Figure 1: Diagram of our proposed neural network system

4.2 Dataset

The English dataset provided by SemEval2020 Task 12 (Rosenthal et al., 2020) was used to train the model. The data set provided was not labeled and was assigned a confidence score from unsupervised learning. Sub-tasks A and B are assigned one score because they are classified into two classes, and Sub-task C is assigned three scores because they are classified into three classes. Sub-tasks A and B were labeled to the dataset with thresholds 0.3 and 0.5, respectively. Sub-task C was labeled as the class with the highest score. In total, 9,075,418 tweets were provided for Sub-task A, while 188,973 tweets were provided for Sub-tasks B and C.

4.3 Evaluation Measures

Since the number of evaluation data is unbalanced between classes, we employ the macro-F1 value for the evaluation index, which is averaged over the whole class by calculating the F1 value for each class.

4.4 Experimental Results

In this section, we describe the results of experiments conducted using the proposed method. The results of participation in SemEval2020 are shown in the Table 1. For the effectiveness of the ensemble, SemEval2019 Task 6 training and test data (Zampieri et al., 2019a) were used as test data. The results are shown in the Table 2. Sub-tasks B and C were validated. In both Sub-tasks B and C, the ensemble model gave the best results.

We consider the F1 score for each class of Sub-task C. The results are shown in Table 3. In the GRP and OTH classes, the ensembles improved their accuracy.

Table 1: Results of SemEval2020 Task 12

Task	F1-score
Sub-task A	0.8653
Sub-task B	0.5638
Sub-task C	0.5720

Table 2: Ablation study

	accuracy	F1-score
MLP	0.7118	0.5457
Bi-LSTM	0.7108	0.5845
MLP+Bi-LSTM	0.7245	0.5927

Table 3: F1 Score by class

	IND	GRP	OTH
MLP	0.82	0.58	0.24
Bi-LSTM	0.82	0.61	0.32
MLP+Bi-LSTM	0.82	0.63	0.33

4.5 Conclution

In this paper, we described our approach to SemEval-2020 Task 12. The effectiveness of our ensemble model using features extracted from two different types of BERT layers was presented. In particular, the accuracy was improved in the GRP and OTH classes.

Future tasks include devising a new approach to the OTH class and applying fine-tuning to BERT.

Acknowledgments

The part of this research is supported by MEXT KAKENHI, Grant-in-Aid for Scientific Research (B), Grant Number 17H01746 and by grants from the KDDI Foundation.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A Large-Scale Semi-Supervised Dataset for Offensive Language Identification. In *arxiv*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. 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 (NAACL)*, pages 1415–1420.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval)*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.