UJNLP at SemEval-2020 Task 12: Detecting Offensive Language Using Bidirectional Transformers

Yinnan Yao
School of Information Science and Engineering
University of Jinan
yaoyinnan@foxmail.com

Nan Su
School of Information Science and Engineering
University of Jinan
me@sunan.me

Kun Ma*
Shandong Provincial Key Laboratory of Network Based Intelligent Computing
University of Jinan
ise_mak@ujn.edu.cn

Abstract

In this paper, we built several pre-trained models to participate SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media. In the common task of Offensive Language Identification in Social Media, pre-trained models such as Bidirectional Encoder Representation from Transformer (BERT) have achieved good results. We preprocess the dataset by the language habits of users in social network. Considering the data imbalance in OffensEval, we screened the newly provided machine annotation samples to construct a new dataset. We use the dataset to fine-tune the Robustly Optimized BERT Pretraining Approach (RoBERTa). For the English subtask B, we adopted the method of adding Auxiliary Sentences (AS) to transform the single-sentence classification task into a relationship recognition task between sentences. Our team UJNLP wins the ranking 16^{th} of 85 in English subtask A (Offensive language identification) .

1 Introduction

With the explosive growth of data generated by online social network users, the malicious content mixed in the information published by users has brought great challenges to the detection of hate speech and offensive language. Due to the difference between social media and traditional media, users are allowed to post information at will. The emergence of large amounts of data invalidates manual review, resulting in a large number of methods that use machine learning for automatic classification. Therefore, SemEval 2020 released OffensEval2 (Zampieri et al., 2020), which compared to OffensEval (Zampieri et al., 2019b) increased the size of the dataset and added multilingual. Multilingual data sets have been added in OffensEval2, includes Arabic (Mubarak et al., 2020), Danish (Sigurbergsson and Derczynski, 2020), English (Rosenthal et al., 2020), Greek (Pitenis et al., 2020), and Turkish (Çöltekin, 2020). This task of English is divided into three subtasks: offensive language recognition, automatic classification of attack types, and attack target recognition.

Our method for this task of English is based on BERT and RoBERTa. The organizer provided a large number of samples marked by the machine and included the confidence of each sample. Because deep neural networks rely on large-scale datasets, the results obtained when using only the OLID dataset are not ideal. Due to equipment limitations and considering the error of machine labeling samples, we did not use all data for training. Based on the OLID dataset (Zampieri et al., 2019a), we use the data with a confidence higher than threshold to expand the dataset to make the number of positive and negative samples equal to avoid the problems caused by the balance of dataset. For subtask B, we add Auxiliary Sentences(AS) to transform the single-sentence classification task into a problem of relationship recognition between sentences. We compared the effect of BERT and RoBERTa with the NSP task removed, and BERT is better for the relationship between sentences.

We compare BERT and RoBERTa models. Finally concluded that the RoBERTa model can achieve better results on large-scale datasets. Due to time constraints, we only participated in and submitted the

https://github.com/yaoyinnan/OffensEval

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 2203–2208
Barcelona, Spain (Online), December 12, 2020.

English subtask A. Our team UJNLP wins the 16th place (out of 85) in English subtask A. After the competition, we tested the effect of the method on subtask B and subtask C.

2 Related work

In recent years, many researchers and research institutions have done a lot of research on insulting language, hate speech and other offensive language. The English task of OffensEval2 consists of three subtasks. Subtask A aims to detect offensive language as not offensive (NOT) or offensive (OFF). Subtask B aims to classify the offensive language as targeting specific entities (TIN) or not targeting specific entities (UNT). Subtask C aims to determine whether the goal of the offensive position is individual (IND), group (GRP) or unknown (OTH). In addition to research in the OffensEval, other researchers also studied Chinese (Su et al., 2017) and Slovenian (Fišer et al., 2017) and German (Wiegand et al., 2018).

One of the benefits of transfer learning is that it can learn effectively from limited labeled data. The Bidirectional Encoder Representation from Transformer (BERT) model proposed by Google AI Language team (Devlin et al., 2018) is pre-trained using a large number of corpora from different sources . NULI used linear model, LSTM and BERT, and finally chose the best BERT (Liu et al., 2019). They also pointed out that the BERT model performed best in subtask A, and achieved the first place in subtask A of SemEval-2019 Task 6. In subtasks B and C, the dataset distribution is less smooth and the amount of data is less, so the effect is not as good as A. In contrast to other models, BERT uses a two-way representation to take advantage of the left and right context and deepen the understanding of the sentence by capturing long-term dependencies between the parts of the sentence (Wu et al., 2019).

Kumar et al. (2019) believe that it is very important to preprocess the words. Only in this way can the words form a sentence and conform to the normal grammatical structure. Aglionby et al. (2019) first detects words before training the model. If it is an unknown word, it first uses a word segmentation tool to process it, and then performs error correction operations if it is not in the dictionary. In addition, they also adopted a deep learning algorithm based on attention, which combines BiLSTM and emoji attention. It has been proved that this processing method is very effective.

3 System overview

3.1 dataset

The OffensEval2020 dataset available to participants contains 13240 tweets from OLID dataset and about 10 million tweets on subtask A marked by the machine. Considering the limitations of machine performance and the error of machine labeling, we selected the samples with confidence greater than 0.88 in the machine labeling dataset and expanded them on the basis of OLID dataset to form a balanced dataset with a similar number of samples. The counts of various labels in the adjusted dataset are shown in Table 1, Table 2 and Tabel 3. The second row in the table is OLID data, the third row is the data we extended, and the fourth row is the data that combines these two items. The table 4 showed classes statistics of OffensEval2 test dataset.

	NOT	OFF	TOTAL
OLID	8,840	4,400	13,240
EXPAND	18,267	22,638	40,905
TOTAL	27,107	27,038	54,145

Table 1: dataset: Subtask A

	TIN	UNT	TOTAL
OLID	3,876	524	4,400
EXPAND	23,162	26,202	49,364
TOTAL	27,038	26,726	53,764

Table 2: dataset: Subtask B

	IND	GRP	OTH	TOTAL
OLID	2,407	1,074	395	3,876
EXPAND	9,469	10,802	11,481	31,752
TOTAL	11,876	11,876	11,876	35,628

Table 3: dataset: Subtask C

A	Test	B	C	Test
NOT	2,807	TIN	IND	580
OFF	1,080	TIN	OTH	80
		TIN	GPR	190
		UNT	-	572
ALL	3,887			1,422

Table 4: Dataset statistics

3.2 Models

BERT A large number of researchers participating in OffensEval2019 have used BERT to achieve good results on three English subtasks A, B and C. BERT uses the same multi-head transformer structure introduced in (Vaswani et al., 2017) and bidirectional representation to capture the long-distance dependencies between the various parts of the sentence in the context, so that the sentence has a deeper understanding. BERT has Next Sentence Prediction (NSP) task, which makes it very suitable for the task of relationship recognition between sentences.

RoBERTa RoBERTa uses more data for training and uses a larger batch. In order to capture the relationship between sentences, BERT uses the Next Sentence Prediction(NSP) task as the pre-training target task, but RoBERTa believes that the judging standard of the NSP task is too simple and remove it. In addition, RoBERTa has researched and obtained pre-trained models in different language environments (Conneau et al., 2019; Martin et al., 2019), which will provide good help for the processing of various tasks in multilingual.

3.3 Methodology

For subtask A, because we have a dataset with a large amount of data, and the subtask A is biased towards the understanding of shallow text features, we use the dataset to fine-tune RoBERTa and use it as the final submission plan. For subtask B, it is difficult to mine the shallow features of text to achieve the desired effect. Since the BERT model has a better effect on improving the relationship between sentences, Sun et al. (2019) classifies the single sentence classification task and adds the auxiliary sentence to a double sentence relationship judgment task. We use this scheme to add an Auxiliary Sentences(AS) "Do posts contain the target audience of profanity?" To the sample of subtask B, and fine-tune it on the BERT model to achieve an improvement in effect, and RoBERTa does not have this improvement. For subtask C, we only used RoBERTa for fine-tuning, but more experiments will be conducted in the future.

4 Experimental setup

4.1 Data Pre-processing

The organizer has pre-processed the samples in the dataset. Users and links are replaced with standard tags @USER and URL. We mainly pre-process emoticons and morphological reduction.

Emoji substitution. In order to preserve the semantic and emotional information contained in emojis, we used an online emoji project on github . This project can map emoticons to phrases, so that we can handle these contents more conveniently.

Lemmatization There are often many wrong grammatical forms in the data published on online social media. In order to ensure the standardization of embedding, we use the WordNetLemmatizer module provided by NLTK to process incomplete words.

Misc. We convert all text to lower case. Continuous "@USER" is limited to a maximum of three to reduce redundancy.

https://github.com/google-research/bert
https://github.com/pytorch/fairseq
https://github.com/carpedm20/emoji

4.2 Experiments

In the subtask A and subtask C, we use RoBERTa-base. The Transformer has 12 layers with a size of 768 and 12 self-attention heads. Moreover, the softmax classification head is added on top of the pre-trained language model. The dataset is divided into 90% of the training set and 10% of the development set. We used the pre-processed training dataset as input to fine-tune the classification model. The hyper-parameters used in our fine-tuning training are as follows: The sentence length is 64, the batch size is 128, the learning rate is 1e-5, the weight decay is 1e-4, and the epochs are 10. For subtask B, we use added Auxiliary Sentence(AS) dataset to fine-tune BERT-base-cased, other hyperparameters are consistent.

5 Results and Analysis

5.1 Results

The macro F1 is used as a formal metric for all subtasks involved in this task. The RoBERTa model we use has a F1 of 0.9128 for the subtask A (@CodaLab). Outside of the competition, the result of using BERT (AS) to test subtask B is 0.6376 and subtask C test using the RoBERTa model was 0.6158. We show the results of each of the three subtasks in the table 5 ,Tabel 6 and Tabel 7 Accuracy and average macro-F1. In the table header, P represents the precision rate and R represents the recall rate.

5.2 Analysis

For subtask B, we use the data with added auxiliary sentences for fine-tuning. Because of the target task NSP in BERT, added the auxiliary sentences could get better results. Using the same data to fine-tune RoBERTa, the effect is reduced due to the increase in the price increase error after the auxiliary sentence. because we refer to the OLID test dataset that is inconsistent with the test set sample distribution of OffensEval2 to construct the verification set, this leads to the problem of serious deviation of the prediction result category. The comparison of the results of each model is shown in Table 8.

	NOT	OFF		P	R	F1
NOT	2538	269	2807	.99	.90	.95
OFF	21	1059	1080	.80	.98	.88
	2559	1328	3887	.90	.94	.91

Table 5: Test set: Subtask A (RoBERTa)

	TIN	UNT		P	R	F1
TIN	794	56	850	.68	.93	.79
UNT	370	202	572	.78	.35	.49
	1,164	258	1,422	.73	.64	.64

Table 6: Test set: Subtask A (BERT(AS))

	IND	GRP	OTH		P	R	F1
IND	533	29	18	580	.84	.91	.88
GRP	60	123	7	190	.71	.65	.68
OTH	40	22	18	80	.42	.23	.29
	633	174	43	850	.66	.60	.62

Table 7: Test set: Subtask A (RoBERTa)

System	F1 (macro)	Accuracy
BERT	0.6150	0.6892
BERT(AS)	**0.6376**	0.7004
RoBERTa	0.6335	0.7011
RoBERTa(AS)	0.6174	0.6920

Table 8: Methodology for Subtask A

6 Conclusion

In this paper, we introduced the results of the UJNLP team participating in SemEval-2020 Task 12. Due to time limitation, we only submitted the results of subtask A. Outside of the competition, we tested the effect of our scheme on subtask B and subtask C. We noticed that the task in OffensEval2019 has a class imbalance problem, which will have a significant impact on system performance. In order to offset the impact of various imbalances, we screened the data marked by the machine combined with the OLID dataset to build a balanced dataset. We use RoBERTa as the basis and use the dataset for fine-tuning. For subtask B, we add auxiliary sentences to convert single sentence classification tasks into sentence relationship recognition tasks, which to achieve good results on BERT. Our model ranked 16th in subtask A and achieved a result of macro-F1 of 0.9128. The current research results show that subtask A has

achieved better results, while subtask B and subtask C still have problems to be solved. In the future, we will further use the proposed model for research on subtasks B and C.

Acknowledgements

The author wishes to thank the organizers of the SemEval task for putting together this valuable event. This work was supported by the National Natural Science Foundation of China (61772231), the Industry-Academy Cooperative Education Project of Ministry of Education (201801002030 & 201702185051), the Shandong Provincial Natural Science Foundation (ZR2017MF025), the Project of Shandong Provincial Social Science Program (18CHLJ39), the Science and Technology Program of University of Jinan (XKY1734 & XKY1828), and the Project of Independent Cultivated Innovation Team of Jinan City (2018GXRC002). The corresponding author of this paper is Kun Ma.

References

Guy Aglionby, Christopher Davis, Pushkar Mishra, Andrew Caines, Eleni Giannakoudaki, Marek Rei, Ekaterina Shutova, and Paula Buttery. 2019. Camsterdam at semeval-2019 task 6: Neural and graph-based feature extraction for the identification of offensive tweets.

Çağrı Çöltekin. 2020. A Corpus of Turkish Offensive Language on Social Media. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*. ELRA.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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*.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in slovene. In *Proceedings of the first workshop on abusive language online*, pages 46–51.

Ritesh Kumar, Guggilla Bhanodai, Rajendra Pamula, and Maheswara Reddy Chennuru. 2019. bhanodaig at semeval-2019 task 6: Categorizing offensive language in social media. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 547–550.

Ping Liu, Wen Li, and Liang Zou. 2019. Nuli at semeval-2019 task 6: transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020. Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive Language Identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA.

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

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive Language and Hate Speech Detection for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*. ELRA.

Hui-Po Su, Zhen-Jie Huang, Hao-Tsung Chang, and Chuan-Jie Lin. 2017. Rephrasing profanity in chinese text. In *Proceedings of the First Workshop on Abusive Language Online*, pages 18–24.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv preprint arXiv:1903.09588*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.

Zhenghao Wu, Hao Zheng, Jianming Wang, Weifeng Su, and Jefferson Fong. 2019. Bnu-hkbu uic nlp team 2 at semeval-2019 task 6: Detecting offensive language using bert model. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 551–555.

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 NAACL*.

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*.