Ferryman at SemEval-2020 Task 12: BERT-Based Model with Advanced Improvement Methods for Multilingual Offensive Language Identification

Weilong Chen, Peng Wang, Jipeng Li, Yuanshuai Zheng, Yan Wang, and Yanru Zhang*
University of Electronic Science and Technology of China
yanruzhang@uestc.edu.cn

Abstract

Indiscriminately posting offensive remarks on social media may promote the occurrence of negative events such as violence, crime, and hatred. This paper examines different approaches and models for solving offensive tweet classification, which is a part of the OffensEval 2020 competition(Zampieri et al., 2020; Zampieri et al., 2019b). The dataset is Offensive Language Identification Dataset (OLID)(Zampieri et al., 2019a), which draws 14,200 annotated English Tweet comments(Rosenthal et al., 2020). The main challenge of data preprocessing is the unbalanced class distribution, abbreviation, and emoji. To overcome these issues, methods such as hashtag segmentation, abbreviation replacement, and emoji replacement have been adopted for data preprocessing approaches. The main task can be divided into three sub-tasks, and are solved by Term Frequency–Inverse Document Frequency(TF-IDF) vectorizer, Bidirectional Encoder Representation from Transformer (BERT), and Multi-dropout respectively. Meanwhile, we applied different learning rates for different languages and tasks based on BERT and non-BERTmodels in order to obtain better results. Our team Ferryman ranked the 18th, 8th, and 21st with F1-score of 0.91152 on the English Sub-task A, Sub-task B, and Sub-task C, respectively. Furthermore, our team also ranked in the top 20 on the Sub-task A of other languages(Çöltekin, 2020; Sigurbergsson and Derczynski, 2020; Mubarak et al., 2020; Pitenis et al., 2020).

1 Introduction

With the continuous development of society, online social network (OSN) and microblog sites have attracted Internet users more than any other types of websites. Twitter, Facebook, and Instagram offer a growing variety of services that attract users from different cultures, religions, and interests around the world. The number of OSN users are getting larger, the variety of users and the contents uploaded are growing rapidly. The strong inclusiveness and free atmosphere of social networks provide a platform for all kinds of users to communicate, share, and discuss. Due to the argument caused by different backgrounds, cultures, and beliefs of users, malicious comments have been emerging all these years. The unrestrained release of offensive remarks on social media have negative impacts on the development of the whole society and may lead to outbreaks such as violence, crime, and hatred. The free and inclusive nature of OSN is easy for the widespread of negative effects among large number of users, since the comments and news are almost uncontrollable. Therefore, the detection and protection of offensive content on social media is particularly important.

Offensive comments are mostly related to religion, race, and gender, offending, insulting, or threatening others through a series of derogatory words. This is the main part of the objectionable content, which is widely distributed in social media. Many users have abused the power given by social networks to express their opinions excessively. Due to the diversity of language and expression on the Internet, the regulation of offensive content in OSN is extremely difficult to solve. Therefore, mining, processing, and automatic detection the information from the offensive content cannot be effectively achieved. The main task of this competition is to identify whether tweets in multiple languages on social media are offensive. This task has the following four challenges. 1. The given dataset is insufficient to train complex

*All the corresponding to Yanru Zhang.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1947–1952
Barcelona, Spain (Online), December 12, 2020.

models. 2. Different language features cause irregular sentences and out-of-speech vocabulary. 3. Alack of vocabulary leads to low detection accuracy of offensive content and difficulty in emoji recognition. 4. The distribution of target categories is uneven, and the test set and training set are inconsistent.

To achieve the goal of this task, we have adopted the term frequency–inverse document frequency (TF-IDF) vectorizer, bidirectional encoder representation from transformers (BERT), and Multi-dropout in our model. Meanwhile, we applied different learning rates to tasks in different languages based on bert and non-bert models in order to get better results. In the rest of this paper, we organize the content as follows: the related works of the hostile content will be described in Section 2.The data description, preprocessing details, and the methodology of our model are introduced in Section 3. The experimental results are discussed in Section 4. The work will be summarize at the end of this paper.

2 Related Work

The field of detecting offensive content have attracted increasingly interests in the recent years. Many previous works in this domain are strongly related with hate speech detection (Waseem and Hovy, 2016) and abusive languages of online statement (Nobata et al., 2016)(Waseem et al., 2017). Offensive comments detecting are tacked with several different approaches, such as feature engineering-based method, using specific bag-of-words, systems combined with several models.

Using the bag-of-words to detect hate speech is an effective way (Kwok and Wang, 2013). Other scholars utilize the word lists containing offenses as words or phrases and get outstanding experimental result (Bassignana et al., 2018).

As for the feature engineering-based approaches such as SVM, regression models, convolutional neural networks (CNN), and etc., there are some other attempts. Such as the system based on a K-max pooling CNN model (Wang et al., 2019) and the CNN model based on word2vec embedding (Gambäck and Sikdar, 2017).

For system combined with several models in detecting offensive tweet, Alessandro Seganti et al. constructed a system which system combines several models (LSTM, Transformer, OpenAI's GPT, Random forest, SVM) with various embeddings (custom, ELMo, fastText, Universal Encoder) together with additional linguistic features (number of blacklisted words, special characters, etc.) (Seganti et al., 2019). We also adapt ensemble strategy in our work.

It is worth mentioning that the latest Natural Language Processing (NLP) model Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) has attracted much attention. This model was trained on a large text corpus such as Wikipedia, which can be applied to various NLP tasks without changing its core archirtecture. Zhenghao Wu et al. used BERT to capture linguistic, syntactic and semantic features (Wu et al., 2019). Andraz Pelicon et al. used a fine-tuned BERT model to achieve offensive language identification (Pelicon et al., 2019). The existing implementation of BERT and the exceptional performance of this technique inspired our mind.

There are also several other superior methods such as the syntactic features method for identifying the targets and intensity of hate speech (Burnap and Williams, 2015) or approaches the task as topic modelling problem using Latent Dirichlet Allocation (Xiang et al., 2012)(Blei et al., 2003). We take these works as reference while constructing our own model.

3 Methodology and data

3.1 Data Description

A part of the offensive language recognition datasets are obtained by searching the keyword set in the Twitter API. The task given to the dataset contains "humiliate","conspiracy","ugly" and highly targeted phrases such as "he is", "it is", and "she is". Tweets containing these keywords tend to be more aggressive and directed. These are five different datasets collected according to different criteria but annotated with the same annotation scheme. The composition sentence contains different degrees of mixed using of uppercase and lowercase. Some acronyms and unrecognizable emoticons represent different semantics in different contexts and expressions.

The dataset of this competition is generally divided into training, development data and test data. The main task is composed of three different levels of subtasks. The next level of tasks is based on the results of the previous level tasks. Different tasks correspond to different data. The three subtasks are: Task A, Offensive language identification l Task B, Automatic categorization of offense types; Task C, Offense target identification. The overall solution logic is to determine whether it is offensive language, whether the attack language is targeted and the type of attack target, and gradually deepen and step by step.

3.2 Data Preprocessing

Before feeding the dataset to machine learning model, we take steps to pre-process the data. The core methods and strategies of preprocessing are as follows.

HashTag Segmentation - Rich data provided by tweets have been analyzed, clustered, and explored in a variety of studies. In addition, Hashtags have been used before as primary topic indicators to cluster tweets, In this task, we use Hashtags to preprocess the data and prove that it has an excellent effect on the clustering results.

Abbreviation Replacement - As we all know, people often use abbreviations to comment and express emotions. So we created a substitution dictionary for replacing abbreviations in the data, which turns out as an effective strategy. One typical example would be 'A-hole' is replaced as 'Ass-hole', which is obviously offensive in this case.

Emoji Replacement - We build substitution dictionaries according to the meaning of emojis, and most of emojis can be effectively replaced. For example, the middle-finger emoji in some cases can be replaced with offensive words. However, the emoji replacement may be different according to languages and cultures. Therefore, we have established different replacement dictionaries for different languages.

3.3 Methodology

According to the related research, we decided to train a large variety of different models and combined the advantages of them in ensembles.

TF-IDF - Term Frequency–Inverse Document Frequency (TF-IDF) is a statistical method which is adopted to evaluate the importance of an offensive word to a file in a corpus (Devlin et al., 2018). The importance of a word increases proportionally with the number of times it appears in the file, but at the same time decreases inversely with the frequency of its appearance in the corpus. TF-IDF vectorizer is lightweight and flexible, which can be easily embedded in other models.

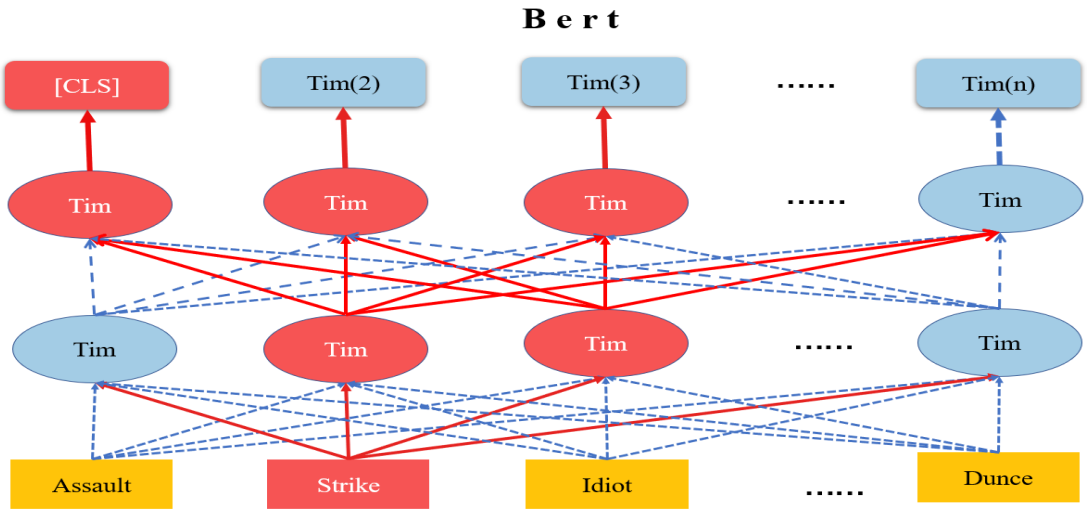

Figure 1: System model

BERT - Devlin et al.released BERT and achieve outstanding results on many NLP tasks.Because some offensive language is subtle, less ham-fisted, and sometimes cross sentence boundary, the model trained

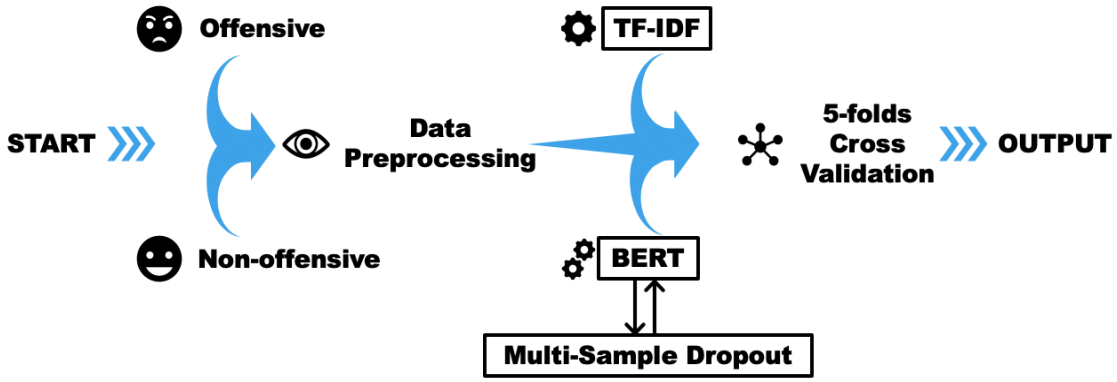

Figure 2: Experimental Process of the Task

for this task must make full use of the whole sentence . Since BERT has been trained on huge corpus from different sources, and it can be directly applied in proposed approaches.

Multi-Sample Dropout -Hiroshi Inoue presents an enhanced dropout technique, which is called multi-sample dropout, for both accelerating training and improving generalization over the original dropout (Inoue, 2019). The multi-sample dropout creates multiple dropout samples, which are randomly selected subsets created by original dropout from the input in each training iteration. This approach can significantly accelerate training and achieve lower error rates and losses for both the training and validation set. We combine the superiority of this method with BERT and facilitate effectiveness of our model.

When training the model, different learning rates are implemented to BERT and the fully connected layer are next to BERT before output. In the meanwhile, we utilize the 5-folds cross validation to obtain a more stable model. Finally we applied concat operation on the last five concept learning systems(CLS), the experimental process of this task is shown in Fig.2. As the result demonstrates, our model obtains an outstanding performance.

4 Result

For this task, our team compares two main methods TF-IDF vectorizer and BERT to deal with the dataset. The result is shown in table1, and it demonstrates that although TF-IDF vectorizer is light and flexible, its performance is far less than BERT. This is because TF-IDF vectorizer only uses the number of occurrences of a word in a file set to map the importance of the word, while the actual Tweet comments are implied in sentences, and sometimes even cross sentence boundaries. In contrast, BERT can make fully use of the entire sentence for recognition, since it receives training from a huge corpus of different sources, which can better reduce the impact of noise.

Methodology	Identification Accuracy
TF-IDF vectorizer	0.86533
BERT	0.91152

Table 1: Comparison of TF-IDF vectorizer and BERT

To further improve the performance, our group uses TF-IDF vectorizer as an auxiliary model and BERT as the main model. Experiments show that the hybrid model performs better than the single-BERT model. At the same time, we applied the Multi-Sample Dropout method to the above model, which greatly accelerated the training speed, reduced the error rate and loss, and made the model more stable and more Lupin.

Results of Sub-Task A, B and C are shown in table2.

Language	Sub-Task	Ranking	F1-Score
English	A	18th	0.91152
	B	8th	0.65764
	C	21th	0.58086
Greek	A	9th	0.822
Turkish	A	12th	0.773721906
Arabic	A	12th	0.85924
Danish	A	18th	0.752

Table 2: Results of Sub-Task A, B, C

5 Conclusion

In this work, we evaluate the performance of the models and methods we use in the three tasks of SemEval-2020 Task 6: Identifying and Categorizing Offensive Language in Social Media. We have used the RAdam optimizer, which makes the model have a very good performance for different learning rates. At the same time, we have used the last five layers of [CLS] tokens to make the model more able to capture higher-order semantic features. Methods and models like BERT, TFIDF all show good results in the given dataset and ranked the 18th, 8th, and 21st with F1-score of 0:91152 on the English Sub-task A, Sub-task B, and Sub-task C, respectively. In addition, we also achieved the top 20th rankings on the Sub-task A of other languages.

References

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *CLiC-it*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, March.

Pete Burnap and Matthew Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making: Machine classification of cyber hate speech. *Policy Internet*, 7, 04.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada, August. Association for Computational Linguistics.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *ArXiv*, abs/1905.09788.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *AAAI*.

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*.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Chang Yi. 2016. Abusive language detection in online user content. In *the 25th International Conference*.

Andraž Pelicon, Matej Martinc, and Petra Kralj Novak. 2019. Embeddia at SemEval-2019 task 6: Detecting hate with neural network and transfer learning approaches. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 604–610, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

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 Semi-Supervised Dataset for Offensive Language Identification.

Alessandro Seganti, Helena Sobol, Iryna Orlova, Hannam Kim, Jakub Staniszewski, Tymoteusz Krumholc, and Krystian Koziel. 2019. Nlpr@srpol at semeval-2019 task 6 and task 5: Linguistically enhanced deep learning offensive sentence classifier. *CoRR*, abs/1904.05152.

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.

Bin Wang, Xiaobing Zhou, and Xuejie Zhang. 2019. YNUWB at SemEval-2019 task 6: K-max pooling CNN with average meta-embedding for identifying offensive language. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 818–822, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. pages 78–84, 01.

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, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, page 1980–1984, New York, NY, USA. Association for Computing Machinery.

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*.