Identifying and Categorizing Offensive Language in Social Media

Version 1.0

Mohammadreza Tavasoli

Abstract

In recent years several papers studied the identification and classification of offensive and abusive content in social media such as Facebook and Twitter. This paper, we investigated the state of the art models for Identifying and Categorizing Offensive Language in Social Media. In this paper, we use a new dataset, the Offensive Language Identification Dataset (OLID).sentences in this dataset labeled based on they are offensive or non-offensive. Furthermore, In this dataset (only for the English Language) they categorize offensive tweets to some subclasses such as targets or offensive posts, or they do not target any person or group, which makes this dataset unique among other datasets that annotate offensive language. In this paper, we try to get state-of-the-art results based on tasks of the OLID dataset by using new deep learning models. We will examine the results of the trained network with various metrics.

1 Introduction

Easy access to the internet and especially social media, bring some new concern about offensive language that may use in these media. For this reason, several papers in NLP literature that try to identify and categorize offensive language in these posts. "In this paper, we want to study offensive language Identification and categorization based on the new dataset." The Offensive Language Identification Dataset (OLID) contains over 14,000 English tweets. It featured three sub-tasks. In sub-task A, the goal was to discriminate between offensive and non-offensive posts. In sub-task B, the focus was on the type of offensive content in the post. Finally, in sub-task C, systems had to detect the target of the offensive posts."[8]. We will take advantage of modern network architectures such as Bert, ELMO, GPT to address the tasks of these datasets. These new networks obtained state-of-the-arts results in many natural language processing tasks. We will examine the train network based on precision and recall and F1-score, and we will compare our results to the results of other papers.

2 Problem Definition

Three tasks considered for this dataset. Sub-task A:Offensive language detection. In this sub-task we have to separate offensive and non-offensive tweets. Sub-task B: We have to predict the type of offensive. Offensive tweets will classify to two categories in sub-task B: "Untargeted(UNT): Posts containing non- targeted profanity and swearing. Posts with general profanity are not targeted, but they contain non-acceptable language."[8] "Targeted Insult (TIN): Posts containing an in- sult/threat to an individual, group, or others" There are three labels in sub-task C for Target Insult tweets.[8] "• Individual (IND): Posts targeting an individ- ual. It can be a a famous person, a named individual or an unnamed participant in the conversation. Insults/threats targeted at individuals are often defined as cyberbullying."[8] "Group (GRP): The target of these offensive posts is a group of people considered as a unity due to the same ethnicity, gender or sexual orientation, political affiliation, religious belief, or other common characteristic. Many of the insults and threats targeted at a group correspond to what is commonly understood as hate speech."[8] "Other (OTH): The target of these offensive posts does not belong to any of the previous two categories, e.g., an organization, a situa- tion, an event, or an issue"[8]

3 Data

"OLID is a large collection of English tweets annotated using a hierarchical three-layer annotation

model. It contains 14,100 annotated tweets divided into a training partition of 13,240 tweets and a testing partition of 860 tweets. Additionally, a small trial dataset of 320 tweets was made available before the start of the competition."[8]. The dataset has @ user in front of each post, which we need to remove before the classification task. The new version of the dataset contains other languages such as Greek and Arabic, but it only annotated for the first task. Only the English language dataset contains three subtasks.(new version of the dataset is not available online now. I sent an email for organizers; they told me they would upload a new version of the dataset soon.)

4 Related Work

In the Internet era and the appearance of social media, offensive language is one of the big challenges of social media. In this order, lots of literature in the NLP area try to address this issue. There various kinds of offensive language, including cyberbullying, hate speech, and toxic comments [8]. We summarized some of them here. [6] used a deep learning method (Convolution Neural network) to classify hate speech into four classes: racism, sexism, both(racism and sexism), and not hate speech. They used ten-fold cross-validation to get the 78.3% F1 score as their best performance. As in many previous studies, hate speech and other kinds of offensive language did not differentiate. [3] trained a multi-class classifier to distinguish between hate speech, only offensive language, and those tweets which are not offensive or hate speech. Furthermore, [3] study when separating hate speech from other offensive languages is hard.[1] used deep learning methods such as LSTM for aggression detection, and they got the best results at Facebook share task on Aggression Identification[7]. They showed more having data in deep learning methods can get a better result than linear models such as NBSVM.[10] proposed a method to capture similarities and varieties between sub-tasks of hate speech, cyberbullying and online abuse. They studied actions that can take to get the best result on abusive detection of the desired task.[5] address the problem of high-dimensionality and sparsity, resulting in a more effective hate speech classifier.[2] proposed the Lexical Syntactic Feature (LSF) architecture to detect offensive content and potentially more offensive users. They achieve relatively high precision and recall in sentence detection.

5 Methodology

For these three tasks, we are facing three classification problems on top of each other. There are various network designs for solving classification problems such as Bidirectional LSTM, Bert, ELMO. Most of the new methods based on transformers and attention. In this part, we discuss what a transformer and Bert is. And How we can use Bert for classification problems. In Bert(Bidirectional Encoder Representations from Transformers.), they first pre-trained their model on mask language modeling, and then they fine-tune their model on 11 Natural language processing tasks. They showed they could get state of the art results on these tasks. Bert has two sets of input one for the first sentence and one for the other sentence. During training, Bert will learn weather the second sentence could be the next sentence or not. This structure helps Bert to be useful for some language processing tasks such as Question-Answering.In this task, the question could be the first input, and a paragraph is the second input, and Bert tries to understand where is the answer of the question in the paragraph by detecting where is the starting point and where is the ending point for the question. Bert Transformer uses bidirectional self-attention.In Bert-Base, the number of transformer blocks is 12. The hidden size 768. And the amount of the self-attention head is 12. The total number of parameters in BertBase is 110 M, which is comparable with Open AI GPT.).[4] The Transformer, its self, consists of encoder and decoder, where encoder and decoder consist of N self-attention layers, and N feed layers stacks forwards on top of each other. Specifically, the inputs to each layer are projected into keys K, queries Q, and values V. Scaled dot product attention is then used to compute a weighted sum of values for each query vector: Attention(Q, K, V) = softmax($QK^T/\sqrt{d_k})V(1)$

dk is the dimension of the keys. [9]

6 Evaluation and Results (1 point)

In the dataset, there was "@user" in each sentence. I remove all during preprocessing since "@user" does not have any useful information for the classification task. After tokenization with Bert specialized tokenizer and setting the maximum sentence length 256, and pertaining Bert on a large

corpus, I used pre-trained Bert for the first task. Here are the results.

epoch	Training Loss	Valid. Loss	Valid. Accur.	Training Time	Validation Time
1	0.48	0.43	0.81	0:12:11	0:00:24
2	0.36	0.46	0.81	0:12:15	0:00:24
3	0.26	0.79	0.80	0:12:14	0:00:25
4	0.17	1.02	0.79	0:12:12	0:00:24

Figure 1: Training and validation loss,time , and Accuracy.

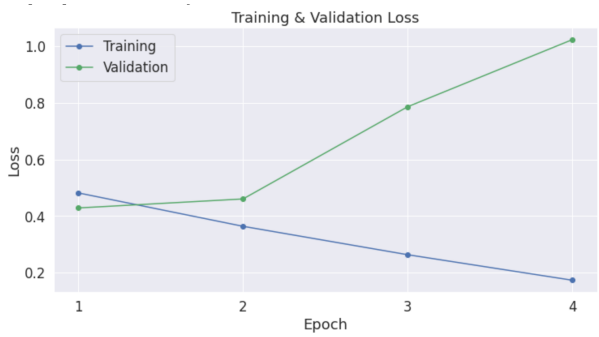

Figure 2: Trainig and validation loss per epoch.

Percision	Recall	F1-score	avg_test_accuracy
0.85	0.9	0.88	0.82

Figure 3: Evaluation of the trained network on test data with various metrics

7 Discussion

Based on the plot, we can understand that 2 Epochs were enough for training since validation loss increases after two Epochs. This experiment Accuracy was imperfect measurement due to the imbalance number of offensive and non-offensive tweets. Therefore, we had to compute precision and recall and F-1 score. Our F-1 score is comparable to the previous results on this dataset and relatively high.

8 Conclusion

In recent years several papers studied the identification and classification of offensive and abusive content in social media such as Facebook and Twitter. In this paper, we investigated the state of the art models for Identifying and Categorizing Offensive Language in Social Media. We used a new dataset, the Offensive Language Identification Dataset (OLID). We examined the results of the trained network on Test data with various metrics such as Precision, Recall, Accuracy, and F1-score.

9 Other Things we tried

Here I investigate the results of SUBTASK-A. I worked on SubtaskB and SubtaskC. However, As It takes a long time to Train network, I could not run several times to get an excellent hyperparameter to get high Accuracy and F-1 scores on that Task,yet.

10 Work Plan

If any time is remaining, I will do some extra preprocessing steps such as tweets such as normalizing the tokens, hashtags, URLs, or converting emojis to text. I will study which feature the trained network learned too.

References

[1] Segun Taofeek Aroyehun and Alexander Gelbukh. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 90–97, 2018.

[2] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE, 2012.

[3] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*, 2017.

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

[5] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30, 2015.

[6] Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90, 2017.

[7] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, 2018.

[8] Andrei-Bogdan Puiu and Andrei-Octavian Brabete. Semeval-2019 task 6: Identifying and categorizing offensive language in social media. *arXiv preprint arXiv:1903.00665*, 2019.

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

[10] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*, 2017.