Identifying Offensive Content in Social Media Posts

Ashwin Singh and Rudraroop Ray

Indraprastha Institute of Information Technology, Delhi
ashwin17222@iiitd.ac.in , rudraroop17311@iiitd.ac.in

Abstract. The identification of offensive language on social media has been a widely studied problem in recent years owing to the volume of data generated by these platforms and its consequences. In this paper, we present the results of our experiments on the OLID dataset from the OffensEval shared from SemEval 2019. We use both traditional machine learning methods and state of the art transformer models like BERT to set a baseline for our experiments. Following this, we propose the use of fine-tuning Distilled Bert using both OLID and an additional hate speech and offensive language dataset. Then, we evaluate our model on the test set, yielding a macro f1 score of 78.8%.

Keywords: Offensive Language · Social Media · Machine Learning.

1 Introduction

The wide reach of social media comes with the curse of people using it to indulge in problematic behaviour. Offensive speech can make people averse to using social media, or worse yet, embolden those who wish to participate in similar behaviour. While it is easier for human moderators to judge if the contents of a post are offensive, it is neither feasible nor efficient to hire a gigantic number of people that would be required to check if each post going up on a social media platform contains offensive content. Automation, therefore, is the logical next step.

In our work, we perform various experiments on sub-task A of OffensEval 2019 where the task is a binary classification task to distinguish offensive tweets from non-offensive ones. Broadly, it can be said that we use three approaches - (i) Extraction of statistical, sentiment-based, TF-IDF and offense-based features, following which traditional machine learning methods such as SVM, logistic regression, Naive Bayes are used for classification, (ii) the use of sentence embedding as features and the same models as (i), and (iii) fine-tuning state of the art transfomer language models such as BERT and its lighter counterpart DistilBERT, on two datasets - OLID (Zampieri et al., 2019) [10] and an additional offensive language and hate speech dataset (Davidson et al., 2017) [1]. We present the results of our experiments in this paper.

2 Related Work

Much work has been done in this area in the past decade. Two important works published in recent years include (Davidson et al., 2017) [1] and (Malmasi and Zampieri, 2017) [2]. The former introduced the Hate Speech Detection dataset with the use of machine learning techniques for classification, such as Logistic Regression, Naive Bayes, Random Forests and Linear SVMs. The latter experiments further with the same Hate Speech Detection dataset, making use of n-grams and skip-grams. Two other works, (Gamback and Sikdar, 2017) [3] and (Zhang et al., 2018) [4], make a comparison between the performance of Neural Networks and traditional machine learning methods. Another work, (Doostmohammadi et al., 2019) [8], uses a Support Vector Machine (SVM) with BERT (Devlin et al., 2018) [13] encoded sentences as input and compares the performance to that obtained by using RNNs and CNNs. Traditional ML methods along with Deep Learning methods are also used in (Emad et al., 2019) [9]. This work then uses a combination of traditional ML and Deep Learning methods along with Data Augmentation to make the model more robust.

A few surveys have been published that cover the work done in this field. One such survey (Fortuna and Nunes, 2018) [6] asserts that most of the work done in this area is limited to the English language and machine learning is used by almost all the researchers. Further, the survey suggests that most of the previous works have modeled the problem as a binary classification task. According to another survey (Schmidt and Wiegand, 2017) [7], these works have chosen to use features such as surface features, sentence embeddings, sentiment-based features, lexical features, linguistic features, knowledge-based features and multimodal information features. The survey also suggested that supervised learning methods such as SVM, Random Forest, Naive Bayes alongside deep learning approaches are commonly used by the previous works.

3 Dataset

The primary dataset we have used is the OLID dataset from the OffensEval 2019 Shared Task (Zampieri et al., 2019) [10]. This dataset contains 14,100 tweets annotated using a heirarchical three layer annotation model and is divided into training and testing sets of 13,240 and 860 tweets respectively.

Table 1. Description of OLID dataset

Label	Train	Test
OFF	4,400	240
NOT	8,840	620
TOTAL	13,240	960

An additional Hate Speech and Offensive Language dataset (Davidson et. al., 2017) [1] is also used in the latter section which contains 24,802 labeled tweets manually annotated by CrowdFlower workers. Owing to the ambiguity between hate speech and offensive language, we discard instances labelled as 'hate speech' from this dataset prior to using it for our task.

Table 2. Description of the Offensive Language dataset

Label	# of Instances
OFF	19,190
NOT	4,163
TOTAL	23,353

4 Methodology

4.1 Feature Extraction and Machine Learning

We describe the feature extraction process of our first approach in detail, before moving on to the experiments, which involve the use of four traditional machine learning techniques, namely Naive Bayes, Logistic Regression, Random Forest and SVM. In this approach, we make use of four types of features which can be described as statistical, sentiment-based, TF-IDF and offense-based features. We discuss these below:

Content-based Features Based on our literature survey, we extracted various statistical features from the content of the tweet which included the number of mentions, the number of hashtags, the number of links, the number of words, the number of uppercase words, the average word length, the average sentence length, the number of punctuation marks and the number of emoticons in each tweet.

Sentiment-based Features We used a sentiment lexicon designed for social media to assign sentiment scores to each instance [12] . Along with this, we also used the TextBlob lexicon to assign subjectivity scores to each instance.

TF-IDF Features The tweets which constituted the OLID dataset had many special characters associated with mentions, hashtags, emoticons etc. which were removed along with commonly used stop-words in the cleaning process. These cleaned tweets were stemmed together to form the corpus from which the TF-IDF features were extracted.

Offense-based Features We used the Offensive/Profane word dictionary [14] to identify offensive language within our dataset. We considered the number of offensive words as a feature for each instance.

Experiments We make use of four traditional machine learning models (Logistic Regression, Naive Bayes, Random Forest and Support Vector Machines) along with a Multi-Layer Perceptron. We evaluate the first four models using 5-fold cross validation on the training set as shown in Table 3. Finally, we evaluate these methods on the test set, with the obtained results given in Table 4.

Table 3. Results for all methods on the training set using 5-fold cross validation.

Model	F1 (macro)	Accuracy
Logistic Regression	69.67	74.19
Naive Bayes	59.44	60.41
Random Forest	**70.15**	**76.51**
SVM	66.70	75.13

Table 4. Results for all methods on the test set

Model	F1 (macro)	Accuracy
Logistic Regression	73.12	79.31
Naive Bayes	65.23	68.61
Random Forest	**75.41**	**83.01**
SVM	70.91	80.17
Multi-Layer Perceptron	71.08	76.54

4.2 Sentence Embedding and Machine Learning

In this approach, we initially perform cleaning on the tweets to remove special characters associated with mentions, hashtags, emoticons etc. along with commonly used stop-words. Then, we use the sentence embedding corresponding to the CLS token in a pretrained BERT (Reimers et. al., 2019) [5] as the set of features for every tweet. Following this, we make use of four traditional machine learning models (Logistic Regression, Naive Bayes, Random Forest and Support Vector Machines) along with a Multi-Layer Perceptron. We evaluate these on the test set.

Table 5. Results for all methods on the test set.

Model	F1 (macro)	Accuracy
Naive Bayes	67.02	69.44
Logistic Regression	**77.12**	**83.05**
Random Forest	70.34	80.21
SVM	75.49	82.36
Multi-Layer Perceptron	71.53	77.42

4.3 Fine-Tuning BERT and DistilBERT

BERT (Bidirectional Encoder Representations from Transformers) [13] is a state of the art natural language model that has proven to be extremely useful for NLP tasks. The principle that a bidirectionally trained model can develop a better understanding of context in language than single-direction language models serves as its foundation. DistilBERT is a lighter variant of BERT that preserves 95 percent of its performance while being 40 percent lighter.

In this approach, we initially perform cleaning on the tweets to remove special characters associated with mentions, hashtags, emoticons etc. along with commonly used stop-words. Then, we perform fine-tuning on both BERT and DistilBERT using the cleaned corpus from both our datasets. We do it for four combinations, namely - (i) OLID dataset with BERT, (ii) OLID dataset with DistilBERT, (iii) OLID + Offensive Language dataset with BERT and (iv) OLID + Offensive Language dataset with DistilBERT.

5 Results and Evaluation

Upon evaluating our fine-tuned BERT and DistilBERT models on the test set, we obtained the following results:

Table 6. Transformer Models and Performances

Model	Dataset	F1 (macro)	Accuracy	Training Time
BERT	OLID	78.08	83.13	38m
DistilBERT	OLID	77.24	82.44	2h 13m
BERT	OLID + Offensive Language	77.84	82.90	6h 56m
DistilBERT	**OLID + Offensive Language**	**78.80**	**83.25**	4h 23m

The evaluation metric for the OffensEval task (Zampieri et al., 2019) [10] was chosen as the macro-averaged F1 score due to the high class imbalance. The highest score on the task leaderboard (82.9%) was recorded by team NULI, who used BERT-base-uncased with a maximum sentence length of 64. However, we were not able to replicate their results in our experiments.

6 Conclusion

In this work, we have presented the results of various experiments to the problem of Offensive Language Identification in Social Media posts. Among the traditional machine learning methods, a Random Forest classifier outperformed the rest. However, upon using BERT-generated sentence embeddings with CLS tokens as input, Logistic Regression produced the best results. On fine-tuning

BERT and DistilBERT with OLID and OLID + Offensive Language dataset, our results were better than those produced by traditional machine learning models. DistilBERT fine-tuned on OLID + Offensive Language performed best. However, the addition of the Offensive Language dataset did not lead to a statistically significant jump in our performance despite the increased complexity. We will also be releasing our code on Github for all the experiments in the near future.

References

1. Thomas Davidson, Dana Warmsley, Michael Macy and Ingmar Weber: Automated Hate Speech Detection and the Problem of Offensive Language. In: Proceedings of ICWSM (2017)
2. Shervin Malmasi and Marcos Zampieri: Detecting Hate Speech in Social Media. In: Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP) (2017).
3. Bjorn Gamback and Utpal Kumar Sikdar: Using Convolutional Neural Networks to Classify Hate Speech. In: Proceedings of the First Workshop on Abusive Language Online (2017)
4. Ziqi Zhang, David Robinson, and Jonathan Tepper: Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In: Lecture Notes in Computer Science. Springer Verlag (2018)
5. Nils Reimers and Iryna Gurevych: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2019)
6. Paula Fortuna and Sergio Nunes: A Survey on Automatic Detection of Hate Speech in Text. In: ACM Computing Surveys (CSUR) (2018)
7. Anna Schmidt and Michael Wiegand: A Survey on Hate Speech Detection Using Natural Language Processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Valencia, Spain (2017)
8. Ehsan Doostmohammadi, Hossein Sameti, Ali Saffar: Ghmerti at SemEval-2019 Task 6: A Deep Word and Character-based Approach to Offensive Language Identification In: Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), Minneapolis, Minnesota, USA, (2019)
9. Emad Kebriaei, Samaneh Karimi, Nazanin Sabri, Azadeh Shakery: Emad at SemEval-2019 Task 6: Offensive Language Identification using Traditional Machine Learning and Deep Learning approaches In: Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), Minneapolis, Minnesota, USA, (2019)
10. Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, Ritesh Kumar: SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval) In: The 13th International Workshop on Semantic Evaluation (SemEval-2019), Minneapolis, Minnesota, USA, (2019)
11. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13). Curran Associates Inc., Red Hook, NY, USA, 3111–3119

12. Hutto, C.J. and Gilbert, Eric. (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.
13. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers forLanguage Understanding In: Google AI Language (2018)
14. Useful Resources: Luis Von Ahn Research Group, Carnegie Mellon University, https://www.cs.cmu.edu/ biglou/resources/.