“Are you fucking serious?”
Analysing Offensive Language Detection on Tweets

Nils-Jonathan Schaller
Vrije Universiteit Amsterdam
Humanities Research: Linguistics: Human Language Technologies
schaller.nj@gmail.com

Abstract

With the growing use of social media and the increase of digital verbal abuse, the task of offensive language detection is gaining more importance. In this paper a convolutional neural network, a support vector machine and a multinomial naive bayes classifier are trained and compared to analyse their performance on offensive language detection on the OLID data set of the SemEval 2019 task 6. Additionally, results from other papers and the participants of the SemEval task are included into the analysis. Using the right preprocessing and state of the art machine learning algorithms can result in a good performance, although the general recognition of offensive tweets is still highly improvable. Furthermore, the lack of a universal definition of offensive language is still a problem for a consistent annotation. A main part of the further work in offensive language detection will not rely only on technical and coding solutions but also on the definition of what is or is not offensive language.

1 Introduction

The use and control of offensive language in social media is a topic of growing importance. To be able to get an overview about the actual use and to possibly prevent it with the implementation of a filter, offensive language and hate speech have to be identified and classified. Both tasks are worked upon, but they still remain unsolved. There is still no ultimate answer to the question, what offensive language defines. Still, there are many approaches in natural language processing and computer science to identify and classify texts and comments that contain offensive language.

In this paper, the Offensive Language Identification Dataset (OLID) of the SemEval 2019 Task 6, consisting on 14.100 labeled tweets, is used to train three machine learning classifiers, a convolutional neural network, a support vector machine and a multinomial naive Bayes, on said data set for identifying offensive tweets. The results are examined to identify which problems in the classification occur in terms of annotation, features and use of language. To do so, also previous work is taken into consideration to give an idea of how data for offensive language detection could be annotated and which attempts seem proposing for future experiments.

Furthermore the performances of the classifiers are compared to each other and to approaches from other authors on the OLID data set. The assumption is, that neural networks like the CNN will give the highest performance, followed by the SVM and the NB classifier in that order. Finally, the focus of this paper is on the error analysis and the resulting conclusion.

2 Related Work

Various approaches to solve the offensive language detection task with machine learning have been done. For the SemEval 2019 Task 6 subtask a 104 teams submitted a result. Of the top ten teams, seven used a BERT classifier, the others different unsupervised machine learning classifiers like a CNN, resulting in f1 scores between 0.80 and 0.83.(Zampieri et al., 2019a)

Waseem and Hovy recognised the difficulty in defining hate speech and annotated 16.914 tweets while using an annotation guide based on critical race theory and also the participation of a gender student and a feminist to overcome possible gender bias in the annotation. They received an inter-annotator agreement of 85%. With using a logistic regression classifier and different models they achieved an f1 score of 0.73 and a similar precision and recall.

(Davidson et al., 2017) tried to separate offensive language from hate speech by using a hate speech lexicon created by the hatebase project[1]. They extracted 25.000 tweets containing the terms from this lexicon. For the annotation a definition of hate speech as "*language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group*" was created. Based on this definition and an additional explanation 24.802 tweets were annotated by at least three crowd workers per tweet with an inter-annotator agreement of 92%. The majority of the tweets were labeled as offensive language, only 5 % as hate speech and 16.6% as not offensive, which lead the authors to the assumption, that the definition for hate speech was too strict. For the final test a logistic regression model was used, resulting in a f1 score of 0.90 for offensive language and 0.60 for hate speech. The distinction between hate speech and offensive language was still erroneous, leading to often falsely identify offensive language as hate speech. One of the main problems is also to identify where to differentiate hate speech and offensive language in general.

(Kwok and Wang, 2013) are focussing on detecting tweets containing racism against black people. They used a twitter data set of 24.582 tweets and labeled it for "racist" and "not racist" by creating rules like "contains offensive words", "reference to painful historical context" etc. After preprocessing the tweets they used a naive Bayes classifier with an accuracy of 0.74.

3 Definition of Offensive Language

One of the main problems with detecting types of offensive language is the lack and the subjectivity of a definition. As the previously mentioned work shows, every annotation has to define new, what hate speech or offensive language is and what not, which makes it difficult to compare the results of the different papers. Focussing on specific kinds of hate speech or offensive language seems not to improve the results. Furthermore, as (Waseem and Hovy, 2016) show, even a very specific guide is not necessarily helping with solving the task. The error analysis of those papers demonstrates, that there are always tweets that the authors would have annotated differently than the crowd or chosen experts.

4 Methodology

To investigate, how well different machine learning systems perform, three classifiers are trained on the corpus. A SVM, a naive Bayes classifier and a CNN. The expectation of the experiment is, that the CNN will achieve the best performance. Following that, also the performance of the different labels (OFF/NOT) between all classifiers and experiments from other scientists is analysed. The focus lies on the error analysis, to examine which tweets are difficult to classify and why.

4.1 Data and Annotation

The training and test data is created by Zampieri et al. 2019b and is called the Offensive Language Identification Dataset. OLID is the official data set of the SemEval 2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). It consists of 14.100 tweets, divided into 13.240 tweets for training and 860 for testing. Those tweets are annotated for not offensive (NOT) and offensive (OFF). The tweets were selected by keywords that tend to be used in controversial topics and the focus is on political topics. *50% of the tweets come from political keywords, and the other 50% come from non-political keywords*. The annotation was done with crowd workers, each tweet was given to two annotators, in case they disagreed, a third annotator settled the decision for the label. The agreement of the first two annotators was at 60%.

In table 1 the distribution of the label combinations is displayed. TIN, UNT, IND, OTH and GRP are only relevant for the SemEval 2019 subtasks b and c and will not be further analysed in this paper.(Zampieri et al., 2019b) The labels are defined as follows:

- **Not Offensive (NOT):** Posts that do not contain offense or profanity
- **Offensive (OFF):** Posts containing any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct. This includes insults, threats, and posts containing profane language or swear words.

4.2 Preprocessing

The training set has been divided into a 90%/10% training/development set for programming the classifiers. Afterwards, they were tested with ten

[1] https://hatebase.org/

Table 1: Distribution of label combinations in OLID

A	B	C	Train	Test	Total
OFF	TIN	IND	2.407	100	2.507
OFF	TIN	OTH	395	35	430
OFF	TIN	GRP	1.074	78	1.152
OFF	UNT	-	524	27	551
NOT	-	-	8.840	620	9.460
ALL			**13.240**	**860**	**14.100**

fold cross validation and finally applied to the 860 tweet test set. Since the year 2018 tweets have a maximum length of 280 characters, succeding the 140-character limit. The data was not further preprocessed.

4.3 Multinomial Naive Bayes Classifier

A naive Bayes classifier is a probabilistic classifier based on Bayes Theorem with strong independent, therefore naive, assumptions between features. This means, that each part (word, special character etc.) of a text is taken into account individually without any relation to each other. Based on that a probability for a certain class is computed. For this task, the tweets are used as a bag of words (bow) in which every occurrence of each word is counted (e.g. if the word “house” appears twice in the tweet, the classifier will count both occurrences as input), the bow is vectorised and classified with the python module sklearn and the svm function.

4.4 Support Vector Machine

A support vector machine is a linear model for classification and regression tasks. For the classification, between all different classes of the data set an ideal line or hyperplane is selected for separation. The best possible line does not mean a perfect separation and could still allow to falsely classify individual parts of the data. The data input is the same as in the SVM, the tweets are changed into a bag of words, the bow is vectorised and classified with the python module sklearn and the svm function.

4.5 Convolutional Neural Network

A convolutional neural network is a deep neural network for unsupervised machine learning, which consists of an input and output layer and several hidden layers. The tweets are analysed by a convolutional neural network based on the keras module. As the results vary with every iteration, a series of ten iteration is made and the average of those results are used for the further analysis. Due to computational limitations, the epoch is set to the value 1, the batch size to 16. Additionally, iterations with 2, 3 and 4 epochs are made.

5 Results

The SVM iteration has a f1 macro score of 0.69. It returned a similar performance on precision (0.75) and recall (0.76) on the micro level. The classifier succeeded in predicting offensive labels but had more difficulties identifying not offensive texts.

Table 2: SVM results

Label	prec	recall	f1	supp.	f1 Macro
NOT	0.82	0.85	0.84	620	
OFF	0.58	0.52	0.55	240	
avg	**0.75**	**0.76**	**0.76**	**860**	**0.69**

Table 3: SVM Baseline results Zampieri et al.

SVM	prec	recall	f1	f1 Macro
NOT	0.80	0.92	0.86	
OFF	0.66	0.43	0.52	
Average	**0.76**	**0.78**	**0.76**	**0.69**

The results of the naive Bayes iteration are with a f1 macro score of 0.73 the highest but mirror the results of the SVM, as the classifier does a better performance on the offensive labels but decreases when classifying the not offensive texts.

Table 4: Naive Bayes results

Label	prec	recall	f1	supp.	f1 Macro
NOT	0.84	0.89	0.86	620	
OFF	0.66	0.56	0.61	240	
avg	**0.79**	**0.80**	**0.79**	**860**	**0.73**

For the CNN ten iterations were measured. Displayed in table 4 is the average value. The CNN has the same issues as the other classifiers: it performs well in detecting not offensive tweets but gives a bad performance in detecting offensive tweets. The f1 macro score is in average similar to the SVM with 0.69. The change of the epoch for 2, 3 or 4 gave an f1 macro score of 0.62, 0.62 and 0.68.

Table 5: CNN results (average of ten iterations)

Label	prec	recall	f1	supp.	f1 Macro
NOT	0.82	0.84	0.83	611	
OFF	0.59	0.55	0.56	238	
avg	**0.76**	**0.76**	**0.76**	**849**	**0.69**

Table 6: CNN Baseline results Zampieri et al.

CNN	prec	recall	f1	f1 Macro
NOT	0.87	0.93	0.90	
OFF	0.78	0.63	0.70	
Average	**0.82**	**0.82**	**0.81**	**0.80**

In summary, the results of this experiments do not fulfill the expectation, that the CNN returns the highest performance. On the test set, the naive Bayes classifier gave the highest performance, whereas the SVM and the CNN share a second place. Different settings on the CNNs epoch gave no improvement.

5.1 Error Analysis

In general all three classifiers have difficulties with detecting offensive tweets, whereas they perform well with the detection of not offensive tweets. The naive Bayes classifier classified 174 texts with the wrong label, whereas the SVM classified 204 texts with the wrong label. An analysis shows, that 123 of those texts got the same, wrong, prediction. Taken a final run of the CNN with 2 epochs into account, this reduces to 89 texts, which were falsely classified by all three classifiers. On analysing the texts that got the same prediction, it is clear that most of them consist of political views, which was a focus of the data collection and was expected. Following the annotation guide, tweets containing any form of "non-acceptable language" should have been marked as offensive.

68 of the 89 falsely classified tweets were labeled as offensive. Of those 68, 51 did not contain any profanities, insults, sexual content etc. but mostly political content. 16 tweets are presented here as an example for possible error sources:

1. SierraBurgessIsALoser She is me when my phone dings ,

2. ArrestGeorgeSoros He Funds DomesticTerrorists Antifa URL

3. FF @USER he is an alien from the future URL

4. NoPasaran: Unity demo to oppose the far-right in London antifa Oct13 Enough is Enough! URL

The first tweets is related to a TV show, without context it is not clear, if it is insulting or offensive. The second one is an opinion about a public figure, it therefore relies on political views, if this is considered offensive. The third tweet is similarly to the first one highly contextual but displays no clear insult. The last one is a call to join a demonstration, like the second tweet, it depends again on the political view to find this offensive. This supports the assumption, that the political views of the annotator play an important role in what he or she finds offensive, regardless of the wording.

The other 17 offensive tweets fit the scheme of the definition, but were not recognised by the classifier either:

1. @USER @USER Bull crap. You know she doesn't care. She is trying to get attention for her Presidential run. Do you see any other Senator giving nonsense? Nope.

2. Democrats Liberals you are being threatened by Armed Nazis and you talk about disarmament of yourselves and of Disconnection from the Military Lifestyle ? You should broadcast your Armory and killSkills RightFuckingNOW URL

3. LifeIsButepu when you have sex anytime you want but you are single

4. Emmys Well, most of those ex-crackheads are liberals and Democrats; so ...

All those tweets contain profanities, the third one is related to sexual content, which can be regarded as offensive in some cultures. A possible reason for the misclassification is the choice of words or the spelling. "Bull crap" seems to be an unusual choice of words. "RightFuckingNOW" is containing the profanity "fuck", therefore the classifier could not find it. The third post is assumably considered offensive, as it contains sexual content, though there is no use of profanities or insults. The word "sex" alone is apparently not a strong enough feature to classify this text as offensive. In the last tweet is is to assume that the spelling of "ex-crackheads" prevents the classification as "offensive".

Of the 21 falsely classified tweets that were labeled as not offensive 13 contained profanities etc.:

1. Are you fucking serious? URL
2. An American Tail really is one of the most underrated animations ever ever ever. Fuck I cried in this scene
3. SundayFunday Go check out my girl thecanncierge because she takes bomb photos and drops a hell of a lot of cannabis knowledge! She is the canncierge after all Im just a pot stock URL
4. Conservatives arent afraid of AnitaHill redux: They know power matters more than truth: Republicans arent worried the Kavanaugh hearing will backfire, because in 2018 the facts almost dont matter URL news GOP Trump Liars PostTruth SOCTUS

Tweet one and two should have been clearly labeled as offensive, as they contain "fucking" and "Fuck". Nevertheless, both tweets do not carry an insulting or offensive message, which is probably the reason they have been mislabeled. "Are you fucking serious?" can be interpreted as a simple expression of surprise. The third tweet is drug related, and also uses the word hell. It could happen, that drug related tweets are more often classified as offensive or that the word "hell" is found in more offensive tweets. The last tweet is again political and uses a lot of hash tags. Though the message itself seems not to be offensive it creates the assumption that the annotators consider conservative and right wing tweets more often offensive then liberal or left wing tweets.

The other 8 not offensive tweets did not contain obvious insults, profanity, sexual context etc. but were also misclassified:

1. Kavanaugh John Fund: Kavanaugh and Ford controversy It's 1991 all over again and even some honest liberals are upset URL
2. MAGA BREAKING NEWS: The DNC stole their playbook from George Costanza on Seinfeld. Now at least we know what we're up against. DNC Rule 1: It's not a lie, if you believe it. WATCH: URL
3. ..While reading this, I feel like I am hearing Cardi B's voice rapping. URL
4. @USER Let's just be friends and have a civil discussion." So it's totally fine if he is told to his face that his life is sinful. But criticize him ever so lightly and that is not civil anymore, which deserves a block!"

Tweet number one is again political, it is not clear, why it was classified as offensive. An idea is, that the name "Kavanaugh" and the related incident often occurs in offensive training tweets, a count supports this, by showing that "Kavanaugh" occurs in the training set in 178/8840 not offensive versus 106/4400 offensive tweets, although there may still be other factors at play. The same could be true for the second tweet with the key hashtag MAGA, which often occurs as an abbreviation for "Make america great again" and is related to right wing tweets. As for the two other tweets, there seemed to be no obvious reason for the misclassification.

Furthermore, there are more tweets that used profanities like "fuck" as intensifiers for non insulting statements. In all cases were the word was misspelled, capitalised or embedded in a hasthag or other word, like "fuckin", "FUCK" "RightFuckingNow" it was falsely classified as not offensive. This indicates a heavy weight on profanities for the classifiers, but only if they are displayed unambiguous. The word "fuckin" without the letter "g" in the end and in lower case occurs only once in the test data ("1 son, knockin it out the fuckin park...... URL"), but several times in the training data, where the containing tweets are labeled inconsistently (e.g. "@USER Alex this is so fuckin beautiful NOT NULL NULL") Apparently case is an important feature.

Some tweets, like "3 days before BBC Radio 2s Festival in a Day, they decide to tell us that their headline act isnt coming anymore" were twice in the test set, e.g. number 91430 OFF and 84045 NOT. On further analysis also other tweets occurred twice (88905 and 15180 both OFF). The SVM labeled this tweet false at least once, which could mean that they stuck to their teached pattern, even if the tweet got two different labels. As there was no explanation given, why those tweets were induced into the test set twice, and why one of them got different labels, it is not clear if they are a simple mistake or have the purpose of double checking the classifier against random labeling.

Additionally the ten most important features for the SVM and the NB have been extracted:

Table 7: most important features SVM NB

SVM		**NB**	
NOT	**OFF**	**NOT**	**OFF**
bitch	possive	07405077156	user
coward	repeat	100000	the
fucked	hamas	1001	is
stupid	supergirl	100thmonkey	to
fucker	pedophilia	101	you
fuck	screwing	10am	and
glasses	nowadays	10kg	of
shit	dillinger	10kids	are
idiot	trees	11k	he
1idiots	la	11yrs	that

The words for the SVM under NOT show a great distance from the not offensive class. If they are contained, there is a high chance, that the tweet is classified as not offensive, whereas the words under OFF have a high chance to be related to not offensive tweets. This seems intuitively correct for most of the insults, although the word "glasses" stands out. A quick search showed, that "glasses" are mostly used to demonstrate, that the recipient of the message is not able to see something, mostly as part of an insult or threat.

The words for the NB are less intuitive. The phone number in the first row occurs only once in the training set in an offensive tweet. It is not clear, why those words are considered important for the classifier.

To conclude the error analysis, all classifiers seem to have difficulties with tweets in which the insults are misspelled or use unusual jargon. It also seems, that the annotation has not been done consistently, as tweets containing obvious profanities or insults have not always been labeled OFF, although this is always a given risk when doing a crowd annotation. Though a tweet like "Are you fucking serious? URL" has not to be received as offensive and can also be used in a positive context, it should clearly be marked as OFF, following the guideline, as it contains profanity. The majority of the tweets is political and it is difficult to decide, if they actually contain insults or threats. This is a highly subjective topic and it opens the question, if it is even possible to create annotation rules for this, or if that decision should be left to the intuition of the annotator. A clear difference would be a threat or a racist tweet, but an example like "@USER Holder needed to be impeached" which is labeled as OFF but classified as NOT shows the problematic. Depending on which political side the annotator stands, the request to impeach a politician can be regarded as a necessary demand, a threat or an insult.

5.2 Results compared to other papers

The baseline results of the OLID paper were a few points better for the SVM with a f1 macro score of 0.69, whereas the CNN received a high performance of 0.80. Concerning the different performance between offensive and not offensive tweet recognition, the baseline results had through all classifiers a better performance for the offensive tweets, although the results consistently show lower values for the classification for OFF then for NOT.

Table 8: All NOT/OFF Baseline results Zampieri et al.

All NOT	**prec**	**recall**	**f1**	**f1 Macro**
NOT	-	0.00	0.00	
OFF	0.72	1.00	0.84	
Average	**0.52**	**0.72**	**0.00**	**0.42**
All OFF	**prec**	**recall**	**f1**	**f1 Macro**
NOT	0.28	1.00	0.44	
OFF	-	0.00	0.00	
Average	**0.08**	**0.28**	**0.12**	**0.22**

From all participants in the SemEval 2019 Task 6 subtask a, the best performance was given by a Bidirectional Encoder Representations from Transformers (BERT) classifier by Liu, Lu and Zou ,followed by Radivchev and Nikolov who also used a BERT. In total, seven of the top ten teams used BERT. BERT is a 2018 state of the art deep learning classifier, developed by google with a focus on NLP tasks (Devlin et al., 2019).

Liu, Lu and Zou did several preprocessing steps. Besides splitting hashtags into space separated words they also used a module to interpret emojis as English phrases and exchanged the emojis in the tweets for those phrases, furthermore all tweets were lower cased. After experimenting with different classifiers, they selected BERT with 3 epochs for the submission as it gave the best performance. They received an f1 macro score of 0.82 with an f1 micro score of 0.75 for OFF and 0.90 for NOT.

Table 9: BERT-Large by (Liu et al., 2019)

Label	prec	recall	f1	supp.	f1 Macro
NOT	0.90	0.90	0.90	620	
OFF	0.74	0.75	0.75	240	
				860	**0.82**

Radivchev and Nikolov compared several classifiers after preprocessing the data, e.g. deleting symbols like “@” or “#”, split and tokenise words marked with a hashtag (e.g. *#HelloThere* becomes *hello* and *there*) etc. Furthermore they used pretrained word embeddings from GloVe. This lead to a f1 macro score of 0.815. The data in the paper still indicates a divergence in the performance of recognising offensive and not offensive with an f1 micro score of 0.73 (OFF) and 0.90 (NOT) (table 9).(Radivchev and Nikolov, 2019)

Table 10: BERT-Large by (Radivchev and Nikolov, 2019)

Label	prec	recall	f1	supp.	f1 Macro
NOT	0.91	0.88	0.90	620	
OFF	0.70	0.75	0.73	240	
				860	**0.81**

Contrary to the results of the three classifiers used for this paper, the results of the participants of the SemEval 2019 Task 6 support the expectation, that neural networks, especially the BERT, but also a CNN, are able to produce a higher performance then SVM or NB. It would be of value to test, how large the performance gain from the various preprocessing step is. It is also notable, that both leading teams used word embeddings and splitted hashtags etc. into single space words.

6 Conclusion

In this paper three machine learning classifiers, a convolutional neural network, a support vector machine and a multinomial nave Bayes were trained on the OLID data set of the SemEval 2019 task 6 subtask a for detecting offensive language. Besides the actual performance, also the annotation and the general problematic of defining offensive language should be considered.

All three classifiers were giving an underperformance, compared to similar experiments and the results of the SemEval 2019 Task 6 participants. Analysing the top results, it is necessary to add advanced preprocessing steps on the data set, e.g. splitting hashtags into words with separated space, lower case all tweets and delete redundant information like the symbols or . A further step could be the adding of word embeddings. The analysis shows also, that a state of the art classifier, like BERT could improve the performance. Nevertheless, it would be important to test how much the preprocessing influences the final result. The top ten papers have a variation of performance from 0.79 to 0.82, indicating only a small influence due to preprocessing but rather a high correlation with the choice of the classifier itself, as seven of them used BERT. In a follow up work, different steps of preprocessing can be added to evaluate if and how large an increase in the performance can be gained.

Besides the general performance, all systems performed better in the recognition of not offensive tweets then offensive tweets. Even the leading system has a difference of 0.15 in the f1 micro score of OFF and NOT. This could be due to the comparable small training set and the discrepancy in the separation between OFF and NOT tweets of 8840 and 4400. An evenly distributed data set might approximate the results.

Furthermore, not all annotations followed the general description for offensive tweets. Many tweets that were annotated as NOT contained obvious profanities or insults and should therefore have been labeled as OFF. This is a known risk of crowd annotation and could only be solved by trained annotators, resolving in a greater financial and time consuming effort for future annotations. Or alternatively an active learning approach, which allows to reduce manual annotation effort for supervised machine learning.

The part of the offensive language definition that relies on the use of profanities, opens the question, if the addition of a lexicon of profanities could have improved the results, at least when using this definition.

It has to be said, that the main problem lies in the general problematic of the subjectivity in experiencing offensive language. Many tweets that have no obvious use of profanities or insults were annotated as offensive. As the data has a focus on political topics, it is to assume, that the political views of the annotator play an important part in the annotation and overshadow the actual annotation guideline. This can also be read into the comparable low inter-annotator agreement of 60%. An optimal data set should be annotated by a larger group, preferably from mixed backgrounds to diminish bias in the inter-annotation agreement. As the views of the annotator can not be ignored, it has to be investigated, how they can be included in further tasks.

To conclude, for a better performance in detecting offensive language using machine algorithm the right choice of preprocessing and classifier can enhance the results. More importantly, a clear definition or annotation guide is needed. Therefore, improvements for this task lie not only in better algorithms and computational power but in a decision, what type of offensive language should be filtered and following a clear annotation guideline. It is also questionable to include profanities as part of the definition, as those can be partly filtered with a lexicon, another interesting examination would be the annotation on offensive tweets without the use of profanities.

References

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Irene Kwok and Yuzhou Wang. 2013. Locate the Hate: Detecting Tweets against Blacks .

Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 Task 6: Transfer Learning for Offensive Language Detection using Bidirectional Transformers .

Victor Radivchev and Alex Nikolov. 2019. Offensive Tweet Classification with BERT and Ensembles.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval).

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Predicting the Type and Target of Offensive Posts in Social Media.

A Supplemental Material

Python files and github links (follows)