Correctly Identifying Anti-Racist Comments from Hate Speech in Social Media using Bag of Words and BERT

By Benjamin K. Schmidt (bschmidt66@gwu.edu)
and Armand Heydarian (aheydarian94@gwu.edu)
George Washington University, Data Science Department

Abstract: Detection of hate speech and offensive language is a major concern for scholars, social media companies, and society at large due to the large scale ripple effects of such dialogue on discourse and potential acts of violence due to this speech. Automated detection must be able to distinguish hate speech from anti-racist or counter hate speech, which can often include shared language. By comparing Bag of Words models and a BERT model, detection of anti-racist speech does seem to be systematically miscategorized; however a more diverse collection of anti-racist speech taggers and multi-label datasets will increase confidence in such a system and reduce the risk of dataset bias for training models.

Keywords: Hate Speech, Algorithmic Bias, Social Media, Automated Detection, BERT

I. Introduction to Subject[1]

With the prevalence of hate based ideologies, there is a growing effort to detect hateful content on social media, one platform in particular is Twitter. Twitter has become a breeding ground for hateful content based on racial and identity based bigotry. Hate speech refers to a kind of speech that denigrates a person or multiple persons based on their membership to a group, usually defined by race, ethnicity, sexual orientation, gender identity, disability, religion, political affiliation, or views. According to the United Nations, the distinction between hate speech and free speech should fall into three different categories of: "expression that constitutes a criminal offence; expression that is not criminally punishable, but may justify a civil suit or administrative sanctions; expression that does not give rise to criminal, civil or administrative sanctions, but still raises concern in terms of tolerance, civility and respect for the rights of others." All these platforms fundamentally do the same thing of connecting everybody throughout the world, but Twitter is a leading platform for hate-based content and has garnered the attention of researchers. Twitter serves as a real time blogging network that allows for the dissemination of information on a scale that allows it to report news prior to it reaching major media outlets and it's limitation of characters to short and concise messages lends itself to mass usage, especially during the surge of events like the global protests after the death of George Floyd.

II. Problem Statement

Twitter has been increasingly utilized for the diffusion of hate-based ideologies and world-views. This becomes prolonged due to the anonymous environment and mobility of the platform. There is a lack of an efficient automatic hate speech detection model based on natural language processing and machine learning, particularly with regards to differentiating between racist and anti-racist speech. One of the key challenges of hate speech detection in the realm of social media is the need to separate offensive language from true hate speech. The lexical detection methods that are used tend to have low precision as they classify all messages containing particular terms as hate speech; with previous work using supervised learning continuing to fail to distinguish between the two categories. This project aims to analyze and compare posts on the Twitter platform to improve the accuracy rate of the model, particularly related to false positives (offensive language being flagged as hate speech). Prior attempts to do so utilizing other NLP methods have shown to have high false-positive rates. In order to accomplish this comparative analysis, this project will be utilizing a Twitter dataset containing hate speech labeled from a 2017 paper (Davidson, 2017) which used older NLP methods. This project will seek to improve accuracy and speed with iterations of BERT. BERT's advantage in this effort is it's capability of working across a framework and with different pre-trained models, which can either be: contextual or context free; and unidirectional or bidirectional.

III. Literature Review

As automated moderating and hate speech detection is of interest in a cross section of society and scholarship, the ability to distinguish between hate speech and non-hate speech is increasingly pressing during periods of increased conversation and discussion. It was also found that automated classification methods can produce relatively high accuracy at differentiating between these different classes, close analysis shows that the presence or absence of particular offensive or hateful terms can both help and hinder accurate classification, leading to false positive identification. Though slurs and explicit words help to easily identify some of the most blatant instances of hate speech, but this approach also leaves open the possibility of miscategorization of counter-speech or anti-racist speech that is critizing these kind of slurs through explanation, often with examples.

[1] Sample code support this paper can be found at:
https://github.com/benjaminkschmidt/Hate-Speech-Detection-With-BERT

There are many different journals and publications regarding hate speech detection on social media, these publications approach hate speech from different avenues but there's a common selection of models that have been applied in their respective cases. In particular, the SVM and Linear Regression models appear to be most common, along with naive Bayes, random forest and decision trees (Kohatsu & Carlos, 2019). An influential paper that examined new methods of detecting hate speech was a paper by Marzieh Mozafari, Reza Farahbakhsh, Noel Crespi, called A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media as it was one of the first publications on using a BERT - based model in the realm of hate speech detection in social media. The project aimed to identify text that would be deemed as racist, sexist, hateful or sexist in the context of social media content. Their findings also alluded to the possibility of biases in the collection and annotation process of data. Through all these different methods, there is a common theme of false positives that erroneously inflate the percentage of hate speech found and the clear presence of bias in data collection, annotations and the classification algorithm itself. This bias can attribute to not only the inflation of hate speech detection, but also conflate explicitly anti-racist speech from what is truly hate speech (Davidson & Thomas, 2017). Through all the publications that were examined on hate speech detection, a gap in research that was found was that other researchers did not test for possible anti-racist speech that could erroneously be classified as hateful content. Of the many common models that were used in these different publications, it wasn't found that any researcher was trying to test the possible efficacy of BERT together with the Bag of Words model. Unlike the previous project, which used TF-IDF and bigrams/trigrams for word representation this project is employing word embeddings which is basically one level deep. In addition to bigrams/trigrams features, this project employed bi-directional relationships between the words of a particular Tweet to learn the semantics. Another difference in terms of models, they are using Naive Bayes, random forest and linear SVMs, but this project is employing a multi-layer convolution neural network to gather insights at a deeper level than previous attempts.

IV. Dataset

The dataset for this project will be the dataset gathered by the authors of "Automated Hate Speech Detection and the Problem of Offensive Language." (2017). This dataset contains ~25000 mutliclassed labeled offensive comments and Tweets from social media. This data is labeled as offensive but not racist, racist, and not offensive. These Tweets were searched on the basis of the lexicon, yielding Tweets from over 33,000 unique Twitter users and the timeline of each user was extracted, allowing us to create a sample of over 25,000 Tweets. This project does not seek to address issues related to gathering hate speech, which is heavily influenced by bias of data collectors. This issue would need to be addressed in future work. The distribution of this dataset is shown in Fig 1.

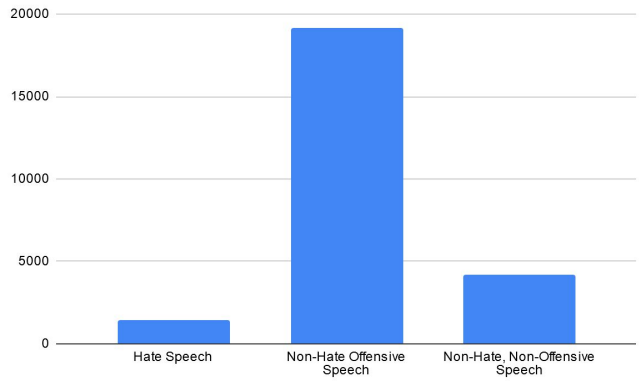

Fig 1. Distribution of Twitter Training Dataset

As this dataset has particular emphasis on offensive, non-hate speech Tweets, there is an expectation that the balance of the dataset will label most Tweets as Category 1.
For a dataset of Tweets that can be considered anti-racist, an additional set of tweets were gathered related to the Black Lives Matter protests following the death of George Floyd in the summer of 2020. Using Social Feed Manager, we collected approximately 200K Tweets regarding the events surrounding the George Floyd from June 1st to 12th using hashtags: ***#blm***, ***#bluelivesmatter, #blacklivesmatter,#georgefloyd, #defundthepolice, and #police***. Through manual inspection of each Tweet, approximately 2,000 anti-racist Tweets were identified to test against the experimental models.These Tweets were selected on the basis of being explicitly anti-racist and critical of systemic racism in media, culture and government. These Tweets were determined based on the guidance below. For purposes of evaluation, each of these Tweets would be considered Category 2, non-offensive, non-hate speech.

V. Anti-Racist Definitions and Rubric

For purposes of this evaluation and research, anti-racist were selected using the following broad criteria:

A. Anti Racist Speech:

- Invokes calls for equality, dismantling of systematic racism, police reform, defunding the police. Example: “Heard this at a protest: "Our tax money-our money pays for the police department! So when they kill, they are killing in OUR name." You're a part of this whether you wanna be or not. #blm #defundthepolice #accountability #wantbetter”
- Calling out criticism of hateful rhetoric from public figures. Example: “Just made it official, I have #crossfitiscanceled thanks to the #crossfitceo. No more @Reebok either. I'm now a @Brooksrunning guy. #EndRacism don't support racists or their sponsors.”
- Acknowledging biases in our culture and government. Example: “Even the algorithms are compromised: IBM exits facial recognition business, calls for police reform | #BLM #Defund”
- Dedication or affirmation of the Black Lives Movement. Example: "Premier League Captains Plan Show of Support for Black Lives Matter"

B. Racist Speech:

- Comparisons of protestors to thugs or terrorists: “George Floyd was used by the looters and shameless rioters.. The demo-zombies used him to advance their greed hate for Trump.”
- Denial or victim blaming for George Floyd’s death. Example: Why didn’t you in the #MSM describe the people attacking the police and damaging property at the weekend during the #BLM protest as left wing violent thugs? I mean that actually happened this hasn’t happened and these are thugs already!!??”
- Delegitimizing Black Lives Matter by associations with crime/wrong doing. Example: “BLM's push to defund police & abolish law enforcement is not about fighting racism or protecting liberty, it's a shameless excuse for anarchy without impunity. They want to destroy the fabric of our society & watch crime rates soar as ordinary citizens suffer.”

VI. Methodology

For effective comparison this project uses the benchmark dataset as described above. After gathering the data, the researchers need to preprocess the data by tokenizing the Tweets. The Tweets will then be

converted into a list of words, with the researchers removing unwanted characters and words as well as words that are not part of the English dictionary. This list will be the corpus of words. The researchers will utilize BERT embeddings to convert a particular word in the Tweet to its numerical vector representation. Next, the Researchers will employ the usage of batching and padding to convert each word to a vector representation of fixed/uniform size. This enables a fixed size of the input data, which is a precondition for any machine learning model. Following tokenization of the Tweets, the training and testing sets need to be formed. To form the training and testing data, the researchers have used a 70-30 split respectively. In this BERT based neural network, the structure consists of several layers. The first input layer is the embedding layer, in simpler terms this means the embeddings which the researchers generated from preprocessing act as an input to the neural network. The next three layers are convolution layers using the ReLu activation function with max pooling layers in between for dimensionality reduction, after this the model produces a dense layer to generate a scalar representation of the data. Finally, the model produces the output layer with three output units using soft max activation for generating the output class. The researchers then move on to testing the model and once the model is trained for a predefined number of epochs, it is tested on the testing data in terms of performance or evaluation metrics such as accuracy, precision and recall. Additionally, this project seeks to use additional data gathered from #BlackLivesMatter to compare against the benchmark data to study impact on the model against anti-racist speech.

VII. Modeling

The Bag of Words model is one of the most common methods of representation used in natural language processing for object categorization, it utilizes text as a representation of a bag of words along with disposing and ignoring the word order and grammar while maintaining multiplicity. The sentence is represented as a string of numbers though the process of vectorization. Vectorization is done through either finding the frequency that each word appears in a file out of all the words or counting the number of times each word appears in a document. The vector representation increases as the vocabulary size increases. In order to preprocess first, all text must be converted to lowercase letters for consistency, then all non-word characters and punctuation must be removed as it is considered unnecessary noise. The next step is to find and identify the most frequently used words in the text. This is then followed by declaring a dictionary to hold the bag of words, each sentence is then tokenized to words and for each word in the sentence, it's checked to see if the word exists in the dictionary. If it does, then the increment is increased by 1 and if it doesn't, it's added to the dictionary and set its count as 1. Finally, the Bag of Words model is then built by constructing a vector, this will indicate whether a word in each sentence is a frequent word or not, and if a word in a sentence is a frequent word, it is set as 1, otherwise it is set as 0. Implementation of Bag of Words algorithms can be seen in Fig 2.

As Bag of Words algorithms are based on word choice similarity, the similarity of language used to counter hate speech and racist messages (Mozafari, 2019) poses a possibility of communications being flagged as hate speech when they are intended to counter hate speech by automated systems. As an alternative, an algorithm that is focused on context using bi-grams may be better able to differentiate between the three categories of speech being studied in this paper. Using an implementation of Bidirectional Encoder Representations from Transformers or BERT (Devlin, 2018), the context of an individual Tweet could be taken into account. As a newer technology trained on a large corpus of text, BERT has the potential to be applied to a variety of text classification tasks, including hate speech classification. Using the pre-trained model and additional subject data of this data set as a key portion of the classification of the process.

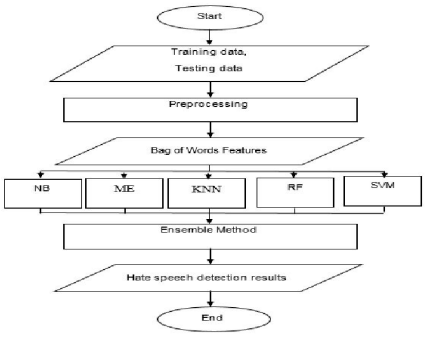

Fig 2. Block Diagram of Bag of Words Algorithms
Courtesy of Fauzi, et Al. 2018

As BERT is an algorithm that utilizes bigrams as a major component to it's predictions and training over a form of cluster analysis of the relative density of specific non-stop words as Bag of Words would use, this type of classification has potential to provide much deeper analysis and examination of the text to be analysed. The large corpus of data that BERT is pre-trained with allows it to be adapted to smaller classification tasks such as social media analysis (Fig3.). As other scholars have found, BERT is a major contribution to the Natural Language Processing domain and provides opportunities for more accurate predictive results. BERT has not yet been used in applications to hate speech, and this paper seeks to explore the opportunities here. By using additional labeled data, the flexibility of BERT allows scholars to advance their studies quickly and find accurate results.

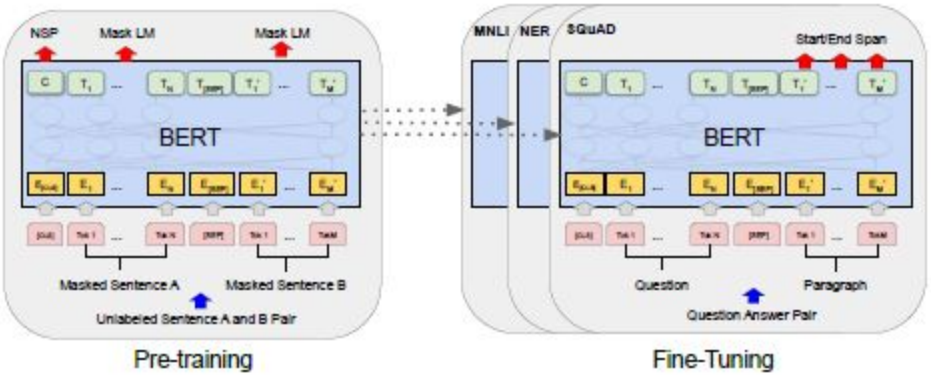

Fig 3. Pre-training and Fine-tuning of BERT.
Courtesy of Devlin et al., 2019.

VIII. Bag of Words Analysis

One of the models that were utilized for testing was the Gaussian Naive Bayes mode. The Naive Bayes classifier is known to be effective for the classification of texts as it works as a linear classifier and also a probabilistic classifier being a supervised machine learning method. When in cases where the dataset's features are all continuous and categorical, the Gaussian Naive Bayes classifier is deemed to be fit for use. An advantage of this model is that it's easier to train on data, particular smaller datasets, whereas other models need to utilize more data. Since it performs a simple linear function, it prevents it from overfitting its training data with low variance, causing there to be less overfitting. The classifier is smaller in size with the ability to process faster. There are issues however with the model as it is very brittle as there could be over-fitting without the regularization assumption. It should be noted that the Naive Bayes

classifier produces competitive classification accuracy, but the data point-class label association probability estimates could be inaccurate.

Upon examining the results for each of the models that were tested, it has been found that the Gaussian Naive Bayes Model would be best as the results have the highest precision and recall. The model provides the best results as it is learning at a relatively better rate on the training data, this is not surprising as Gaussian Naive Bayes is known for promising results with textual data. As mentioned before, the classifier produces competitive classification accuracy, but the data point-class label association probability estimates could be inaccurate and this could be due to when a data point is being classified, the Naive Bayes classifier calculates the probabilities first along with the data points that belong to each possible class label. The classification is created with the selection of the class label that comes with the largest probability. Even though the largest probability is associated with the correct class label, the close and important correlation between the correct class label and the largest probability are not correlated with classification confidence. The following diagrams show the results of the individual Bag of Words models. Full comparison of these models is offered at the end of this document. For individual model's results, please see Fig. 4 and Fig. 5.

Running the GaussianNB Model on BOW data

Accuracy	Precision	Recall	F1	Confusion Matrix
74.7226	52.1692	61.4151	0.5331	[[35 65 190] [246 2853 733] [1 18 816]]

Fig 4. BOW Gaussian Naive-Bayes Metrics

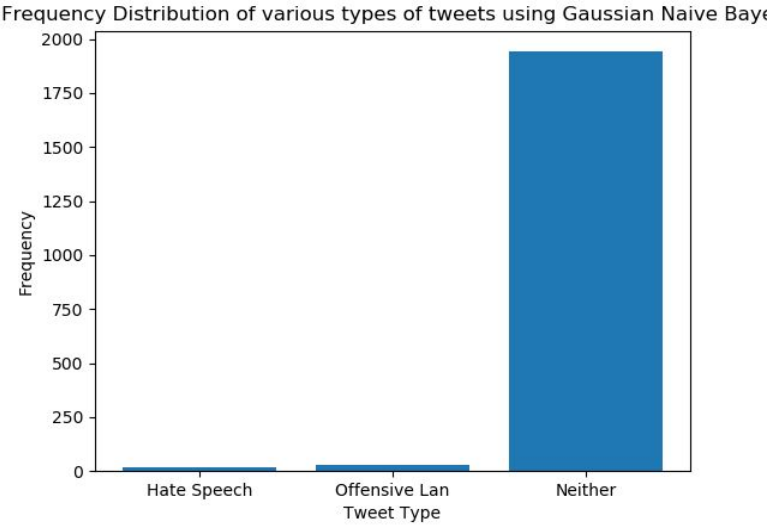

Fig 5. Gaussian Naive-Bayes Anti-Racist Distribution

IX. BERT Model and Analysis

As a BERT has multiple options for implementation, the model being evaluated for this task uses the iteration of BERT as prepared by the transformers library, extending the original . The hyper parameters of this model that are utilized in this evaluation are covered in Figure 10, any hyperparameter that is not in this table is the standard value as defined by the transformers library. The training loss of the BERT Model is shown in Figure 11.

Batch Size	Epochs	Learning Rate	Optimizer	Max Sequence Length
16	20	2e-5	Adam	128

Fig 6. BERT Hyperparamenters

Using the model trained with the above parameters, the model was evaluated with the anti-racists Tweets as used above. As with the Bag Of Words models, the expected result is that the majority of Tweets will be predicted as neither hate-speech nor offensive speech. Using the evaluation functions of pytorch for transfer learning, the anti-racist Tweets were labeled as category 2, non-offensive, non-hate speech, the anti-racist Tweets were used to validate the model prior to being tested with the model. The resulting correct labels was 92 percent, with the majority of the incorrect labels as offensive, non-hate speech. This suggests that the model correctly predicted the majority of cases where anti-racist speech correctly. The histogram in Figure 12 shows the distribution of these results.

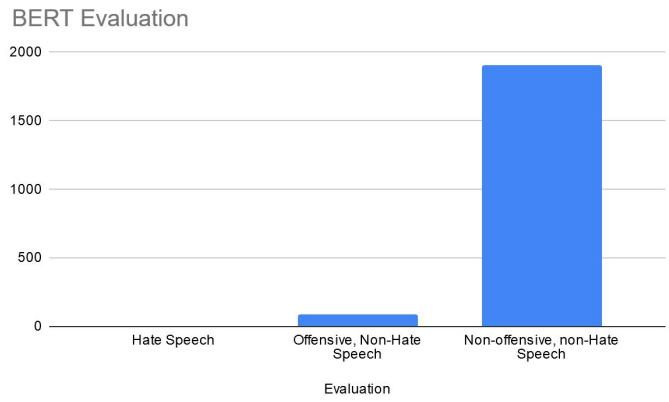

Fig 7. BERT Labeling of Anti-Racist Tweets

X. Comparison

In reviewing this evaluation of the multiple models prepared for this project, there are a few major takeaways. First, each model is largely accurate in labeling anti-racist speech, with the majority of errors occurring with a miscategorization being offensive, non-hate speech. This error is most likely for using “curse” words, though further analysis may explain the cause of incorrect labels. The second, is there is significant variability of training time between the various models, which may impact the use of these models. Lastly, there is only a minor difference between these models and do not suggest underlying bias in the training data. Further study that utilizes a broader range of anti-racist Tweets or ambigious Tweets annotated with reference to a wider range of labelers will enable a more robust model that can handle edge cases. However, the Bag of Words models all consistently have poor Precision, Recall, and F1 scores compared to BERT model, raising some questions on the possibility of general application. However, the ultimate accuracy with the anti-racist Tweets are comparable, though the Gaussian Naive

Bayes Model is the only one that labeled anti-racist tweets as hate speech, despite being the most accurate of the Bag of Words models. The results of the four models explored in this paper are summarized in Figure 13 (raw numbers and Figure 14 (model metrics) below.

Model	Hate Speech	Offensive Speech	Neither
GaussianNB	16	30	1941
BERT	0	92	1908

Fig 8. Anti-racist Labeling by Model (raw numbers)

Model	Accuracy	Precision	Recall	F1	Confusion Matrix
GaussianNB	74.7226	52.1692	61.4151	0.5331	[[35 65 190] [246 2853 733] [1 18 816]]
BERT	92.95	90.85	95.51	.9220	[204, 95, 14], [105, 3009, 64], [20, 84, 1208]]

Fig 9. Combined Model Metrics

XI. Conclusions and Recommendations

As has been noted by scholars working on Automated Hate Speech detection and social media, the automatically identifying hate speech is very dependent on the underlying training data. However, as recent events in the United States have continued to show, the perception of racism, hate speech, and other violent language is deeply variable based on the life experience of the person reading it. As has been discussed in multiple areas of research, larger training datasets that have been labeled by a large number of respondents who have also had demographic data gathered may allow increased effectiveness of an automated system. Resolving these questions represent questions about user privacy and data gathering that is outside the purview of this research. As has been shown by this paper, the type of model used by hate speech detection is critical and may even suggest that a deep learning model may not offer the same degree of accuracy and precision as non-deep learning solutions. Far more important is a balanced dataset that identifies not just hate speech or offensive language but in equal measure non-hate speech. Having sufficiently large datasets will allow for models to avoid over-labeling anti-racist speech, which still presents a challenge to automated solutions as well as non-automated detection. To avoid extending the biases further on the underlying datasets, an effort to expand the size of datasets being used to train automated detection systems along with an expansion of the backgrounds of data gathering efforts will create greater confidence in detection systems. More broadly, this research suggests that for short text such as Tweets, there is clear and consistent benefit from the additional modeling of the additional computational requirements of a BERT Model. Further, each of these models should be tested with additional withheld data specifically for each of the categorical variables to ensure that overfitting is minimal, as the confusion matrix results suggest there is a possibility of over fitting, which would perpetuate hate speech.

Works Cited

1. Davidson, Thomas et al. "Automated Hate Speech Detection and the Problem of Offensive Language." Proceedings of the 11th International Conference on Web and Social Media (ICWSM) (2017). Web
2. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.
3. Fauzi, Muhammad & Yuniarti, Anny. (2018). Ensemble method for indonesian twitter hate speech detection. Indonesian Journal of Electrical Engineering and Computer Science. 11. 294-299. 10.11591/ijeecs.v11.i1.pp294-299.
4. Florio, Komal et al. “Time of Your Hate: The Challenge of Time in Hate Speech Detection on Social Media.” Applied sciences 10.12 (2020): 4180–. Web.
5. Fortuna, Paula, and Sérgio Nunes. “A Survey on Automatic Detection of Hate Speech in Text.” ACM Computing Surveys (CSUR) 51.4 (2018): 1–30. Web.
6. Klubička, Filip, and Raquel Fernández. “Examining a Hate Speech Corpus for Hate Speech Detection and Popularity Prediction.” (2018): n. pag. Print.
7. MacAvaney, Sean et al. “Hate Speech Detection: Challenges and Solutions.” PloS one 14.8 (2019): e0221152–. Web.
8. Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media.
9. Oriol Sàbat, Benet. “Multimodal Hate Speech Detection in Memes.” Universitat Politècnica de Catalunya, 2019. Print.
10. Pereira-Kohatsu, Juan Carlos et al. “Detecting and Monitoring Hate Speech in Twitter.” Sensors (Basel, Switzerland) 19.21 (2019): 4654–. Web.
11. Pitsilis, Georgios K, Heri Ramampiaro, and Helge Langseth. “Effective Hate-Speech Detection in Twitter Data Using Recurrent Neural Networks.” Applied intelligence (Dordrecht, Netherlands) 48.12 (2018): 4730–4742. Web.
12. Ribeiro, M., Calais, P., Santos, Y., Almeida, V., & Meira Jr, W. (2018). Characterizing and Detecting Hateful Users on Twitter.
13. Sajjad, Muhammad et al. “Hate Speech Detection Using Fusion Approach.” IEEE, 2019. 251–255. Web.
14. Watanabe, Hajime, Mondher Bouazizi, and Tomoaki Ohtsuki. “Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection.” IEEE access 6 (2018): 13825–13835. Web.
15. Xia, Mengzhou, Anjalie Field, and Yulia Tsvetkov. “Demoting Racial Bias in Hate Speech Detection.” (2020): n. pag. Print.
16. Zhou, Yanling et al. “Deep Learning Based Fusion Approach for Hate Speech Detection.” IEEE access 8 (2020): 1–1. Web.