Hate Speech Detection in Twitter: A Selectively Trained Ensemble Method

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Jackson Houston

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Richard Maclin

May 2020

Acknowledgements

We would like to thank all who helped us complete this project whether from the beginning or stepping in at short notice. Especially our committee members, Catherine Reich and Peter Peterson. We are grateful for your support.

Dedication

To people targeted by hate speech whether online or in person. There is no excuse for that form of speak.

Abstract

This thesis tests classification models from Natural Language Processing and Machine learning in the task of identifying hate speech. We tested on multiple annotated data sets (Davidson et al. 2017) of tweet data labeled as hate speech, offensive speech, both, or neither. Hate speech has become an unavoidable topic in the current social media environment due to poorly monitored comment sections and news feeds. With that, studies showing the negative affects that it brings to people's well-being have also begun to surface (Gelber and McNamara 2015). Therefore, being able to identify hate speech accurately and precisely has grown in importance. Hate speech is often contextual, subjective, and a matter of opinion which makes creating an accurate model of such speech all the more difficult. We have found that using an ensemble method of a classic Naive Bayes classifier (Pedregosa et al. 2019c), Random Forest (Pedregosa et al. 2019b), K-Means (Pedregosa et al. 2019d), and Bernoulli (Pedregosa et al. 2019a) performed better than similar studies in precision, accuracy, recall, and f-score (Malmasi and Zampieri 2018). The ensemble performed better than using the strongest of the individual models, Random Forest, by a small but useful margin. We believe this to be due to the nuanced nature and context behind hate speech being more than one model can fully encompass. In addition to the ensemble strategy, training on data which was labeled as 'clean' (not hate speech or offensive) or labeled 'dirty' (hate speech) with higher confidence ratings increased the precision of our model by around 10% in some cases when compared to training on the complete data set including the tweets which have a blurred sentiment such as offensive but not hate speech tweets. Having an accurate and precise model such as this will allow organizations to protect their users from such language to prevent the negative effects of hate speech. Additionally, it will allow us to identify more hate speech tweets or statements to have more data to research in the future and find deeper trends than simply the tweet text, such as replies, retweets, and user biographies.

Contents

List of Tables

List of Figures

1 Introduction

This thesis focuses on the problem of accurately classifying hate speech in text, specifically tweets which are short posts on the social media platform Twitter. There are many nuances to language that makes this a difficult task, even more so than simple sentiment analysis which we will explain in this thesis. Identifying hate speech is interesting especially now as the internet has become a common ground for people to express their opinions, good or bad. The latter will sometimes come in the form of aggressive hate speech towards and individual or group of people, causing real emotional, psychological, and social harm to the target. Therefore, being able to identify hate speech automatically could bring positive and immediate benefits to targets of hate speech and society in general by granting people the option to shield themselves from such content.

Although we do believe protecting free speech is important, an argument against the study of hate speech detection, we find it imperative to give a company or individual the tools to discourage or even hide hate speech from their online experience. The internet is an incredible source of information that everyone should feel welcome to access and use. As developers and members of the computer science community, we feel it is our responsibility to help make that so.

We can see the presence of hate speech nearly everywhere we look online. Comment sections of YouTube videos or news articles, Reddit posts and threads, and the news feeds of social media sites like Facebook and Twitter, which this study focuses on. To put it simply hate speech is a form of bullying against an individual or group of people, and it has been shown what negative effects bullying can have. According to Kann et al. n.d. study in 2009, youth that did not identify as heterosexual were much more likely to attempt to die by suicide. Taking that further, a study by Meyer et al. 2019 showed that states that had enacted anti-bullying laws (hate speech is a type of bullying) had much lower suicide rates. Therefore, having a system to accurately protect users from hate speech could have immensely important impacts by possibly reducing the amount of suicide attempts targets of hate speech. We can see this backed up by a study looking at the reported increase in "stress expression" of college students exposed to hate speech on reddit (Saha, Chandrasekharan, and De Choudhury 2019). Looking further into hate speech and You Tube, there has been a recent scenario of a right-wing talk show host and comedian, Steven Crowder, had his entire channel "Louder With Crowder" demonetized because of charges of hate speech against the homosexual, Latino, VOX reporter, Carlos Maza (Nett 2019). From this, and a few other scenarios, YouTube made a sweep over their content to attempt to remove hate speech as well as Nazi rhetoric. In doing so, however, many accounts had content wrongly removed or were even completely demonetized because the algorithm used incorrectly classified data (Dwoskin 2019). Improving our techniques in identifying hate speech will also aid in preventing innocent accounts and users from getting flagged.

Our approach to improve on current hate speech detection is to use various different Machine Learning and Natural Language Processing models together as an ensemble in order to catch as many different nuances and types of hate speech as possible. Additionally, we train the model on widely agreed upon data, avoiding tweets in the middle ground, or in the gray area, to improve the model's understanding of hate speech. These ideas seem to work rather well as we got much more promising results than traditional methods. Training on the "extreme" data increased accuracy alone by a significant margin which is certainly contrary to the notion of “more data, the better.” Tests at different confidence intervals were done as well, all of which led to improvements for the different scores of the models.

2 Background

This section will give an overview of the definition of hate speech and sufficient evidence as to it being a legitimate category of speech. There will also be an overview on relevant laws, where it is commonly found, how it affects people, and why it should or should not be allowed openly in public or online domains. The basic ideas behind machine learning processes used for classification, natural language processing, and sentiment analysis will also be discussed. Additionally, we will discuss how hate speech detection specifically differs from other classification practices such as movie reviews and why it may be more difficult.

2.1 Hate Speech

With the openness of the internet and the wide use of social media platforms hate speech has become a prominent issue online. This has caused a increasing need and experimenting with automatically detecting hate speech (Schmidt and Wiegand 2017). There is, however, no globally accepted definition of hate speech. Different countries and groups will hold varied definitions, or none at all. The best example of a widely accepted definition is "the public incitement to violence or hatred directed against a group of person or member of such a group on the ground of race, color, religion, descent and national or ethnic origin" (Wenguang 2018) which is defined by European Union law. This could of course be extended to include other forms of identity such as age, disability, sexual orientation, and gender.

Various countries such as France, Germany, the United Kingdom, and Canada have made are attempting to combat this issue by imposing civil and criminal laws on hate speech. Germany, for example, has fairly strict regulations on Nazi rhetoric. It is against the law there (and in France) to say the Holocaust did not happen or express other anti-Semitic ideas. Germany has even recently enacted the Network Enforcement Act which aims to fight online hate speech (Wenguang 2018). International agreements exist which denounce hate speech, including the International Covenant on Civil and Political Rights and the International Convention on the Elimination of All Forms of Racial Discrimination (Gelber and McNamara 2015). The United States, however, hate speech is often classified as freedom of expression under the first amendment (Gelber and McNamara 2015).

Speech is not limitless in the United States, though. There have been cases that do not fall under free speech, such as the court case Schenck v. United States (*Schenck v. United States* 2019). In this hearing Oliver Wendell Holmes Jr used the analogy of shouting "fire" in a crowded theater to describe what he stated as words which are normally under the First Amendment may be deemed unprotected if they cause a "clear and present danger." The actual case was to determine whether the acts of Charles T. Schenck and the US Socialist Party were protected by the First Amendment when they were encouraging men to ignore the military draft in wartime. The acts of encouraging men to dodge the draft were deemed as speech which incited a *clear and present danger*, therefore unprotected and Schenck lost (Britannica 2019). This ruling has not stood the test of time, however, and has been ignored and overruled in various court cases since then. The idea that speech could pose *clear and present danger* was made more strict and changed to *directed at inciting or producing imminent lawless action*, for example, trying to start a riot (*Schenck v. United States* 2019). This can be seen in a ruling by the District Court of Northern California in a case against Yahoo! for displaying Nazi paraphernalia on its auction site (Banks 2010). Although hate speech was denounced as terrible, it would still be protected under the First Amendment as having those items available for purchase did not produce any immediate danger or dangerous action.

Why should we care about hate speech? A study funded by the Polish Ministry of Science and Higher Education shows that exposure to hate speech has negative affects on people who view it, whether they were part of the intended target group or not (Soral, Bilewicz, and Winiewski 2018). When repeatedly exposed to this type of language we are desensitized to it, which causes us to become more indifferent in future interactions with hate speech and also subconsciously increases prejudice towards the groups of targeted people (Soral, Bilewicz, and Winiewski 2018). Arguably more importantly, hate speech has been found to cause a multitude of psychological and emotional harm to the recipients of hate speech (those belonging to an specific race, religion or other group) (Gelber and McNamara 2015). This is not limited to damaging self-esteem, feeling unsafe to go to certain places or locations, and feeling excluded from a sense community. Additionally, at a higher scale, hate speech can prolong entrenchment of prejudice and worsen racial divisions by dehumanizing the targets and instilling an "us" vs. "them" mentality. Not to mention that the recipients are often already marginalized in society, indigenous groups, racial minorities, and LGBTQ members for example (Gelber and McNamara 2016).

2.2 Previous Work

This section will summarize previous studies done in the broad area of sentiment analysis as well as hate or offensive speech detection. These summaries are meant to give insight on how sentiment analysis and classification works as well as highlight practices and methods that have proven to be quite accurate and those that seem to rank poorly compared to others.

2.2.1 Sentiment Analysis and Classification

Sentiment analysis and classification is the practice of determining the overall sentiment (whether or not a statement or article was positive or negative, for example) through an algorithm. This has been attempted in many ways from learning methods to using a simple “bad word” list. Traditionally, sentiment analysis is a difficult task as humans all have a unique way of explaining their emotions and feelings on a given topic and these nuances are difficult to formulate.

Pang, Lee, and Vaithyanathan 2002 described, in their paper about the classification of movie reviews, “Thumbs Up?: Sentiment Classification Using Machine Learning Techniques” that people tend to use an “anti-narrative.” Although a movie contained purely positive qualities and should have received positive reviews, in the end the authors will state that all of those positives did not matter as they still did not like the movie in general. For example, one could state about Star Wars episode 8,“I thought that the Last Jedi had absolutely stunning visuals throughout with the beautiful landscapes and creative creature designs. Additionally, I loved the relationship growth between Rey and Kylo Ren and am excited to see where that goes! However, I ultimately found the movie lacking in the end apart from those pieces”. This example, or the opposite approach where most of the review was critical but still rated the movie as positive overall, would be easy for a person to identify if the reviewer enjoyed the movie or not. A computer, however, finds this task much more difficult as it would identify an overwhelming amount of positive or negative words in the statement and would not know which elements to give more weight in the final analysis. The same goes for cases of sarcasm (Pang, Lee, and Vaithyanathan 2002). This is only a small example for movie reviews, but sentiment classification has seeped into many areas of our lives. This is often seen in open discussion forms and social websites which allow people to talk freely and openly, for good or bad, and broadcast their opinions to the world.

Twitter and other social media sites have been a growing platform for users to promote their stance on a wide variety of topics. The ease of access to online posts allows researchers to analyze tweets to understand and forecast important events such as stock market events, political influence, and consumer trends. However, the so-called state-of-the-art tool used by Twitter to do this, Twitter sentiment analysis, is not as accurate as it once was (Zimbra et al. 2018). Compared to current classification analysis practices Twitter sentiment analysis is sometimes less than 70% accurate, which will negatively effect research relying on that data (Zimbra et al. 2018).

Twitter sentiment analysis uses techniques similar to basic sentiment analysis for forums and reviews. This is done with a combination of a lexicon, or sets of related terms in an unsupervised application and a supervised machine learning algorithm to acquire the relationship between values. One weakness is that the lexicon approach does not take context and other social indicators into account (Zimbra et al. 2018). Also, in order to train the machine learning side requires a large amount of data.

One study by Koumpouri, Mporas, and Megalooikonomou 2015 compared four practices for determining sentiment classification on their performance on correctly classifying movie reviews. The analyzed approaches were statistical-based, bag-of-words-based, synonyms-and-antonyms-based, and lexicon-based. The results of this experiment found the bag-of-words approach was the most accurate, but by a rather small margin. It was followed extremely closely by synonyms-and-antonyms-based and lexicon-based classification, all 3 of which were within 2% of each other. The least effective of these was statistical-based which was 10% less accurate than bag-of-words (Koumpouri, Mporas, and Megalooikonomou 2015).

2.2.2 Hate Speech Detection

Detecting hate speech automatically is a unique task. Many basic techniques for filtering online data such as simple“bad word” lists do not work as they would for things like profanity. Hate speech is significantly contextual, two people tweeting the same phrase can be taken completely different. This could be due to the people’s backgrounds, timing of the statement, media connected to the tweet, along with countless other factors (Schmidt and Wiegand 2017).

Pelle, Alcântara, and Moreira 2018 used a technique of which we saw used much less than expected while reading hate speech classification articles. They used multiple classification techniques in an ensemble in order to predict whether a phrase was offensive or not. The ensemble was more accurate than any of its individual parts; in some cases more than 6 percentage points better in F measure and ROC-AUC. This particular ensemble was made from three techniques – Support-Vector-Machine (SVM), HateWord2Vec, and HateDoc2Vec. SVM approach was a traditional bag-of-words approach using uni-gram features which is usually a reliable system in sentiment classification. HateWord2Vec is a process built by the authors using lexicons and word embedding. HateWord2Vec references a list of dirty words that are generally considered offensive and then uses word embedding above to catch anything which could be a dirty word but slightly modified to avoid any explicit detectors, such as misspellings or using symbols in place of letters. If a dirty word is detected it classifies that comment as offensive. Lastly, HateDoc2vec is a multi-step process with Bag-of-Words training and a vector based from the document to create doc vectors. These are then put through a logistic regression classifier. This experiment also used three different data sets to evaluate its ensemble: 1) Tweets-EN which contains 16 thousand tweets labeled racist, sexist, or not, 2) Kaggle with six thousand tweets labeled offensive or not, and 3) OffComBR which contains over one thousand Portuguese tweets. The ensemble performed better than its parts for all of these data sets except that HateDov2Vec tied with the ensemble for the Kaggle data. The ensemble also excelled at preventing false positives and false negatives, with false negatives being slightly more common (Pelle, Alcântara, and Moreira 2018).

Another study was on a large collection of tweets which contained either hate speech, offensive speech, both hate speech and offensive speech, or neither. In fact, this is the data set used in the experiments in this thesis. In the referenced study they used four different groups of information from the tweets to attempt and classify and predict hate speech. First was the detection of sentiment in the tweet, assuming that hate speech usually carries a negative sentiment. Next, semantics were examined, such as use of punctuation or capitalization to emphasize or suggest hate which may not be explicitly stated. Third was to employ uni-grams (single word/feature sequence) to analyze and score occurrences of words and their general likelihood of being used in a hate speech tweet or not. Lastly, patterns were studied to see how closely a tweet resembles the tweets which the model was trained on. The model was trained on all data types (hate speech, offensive, neither, and both) and achieved an accuracy of 87.4% when only predicting offensive or non-offensive tweets and a lower accuracy of 78.4% when predicting a tweet as hateful, offensive, or clean speech (Watanabe, Bouazizi, and Ohtsuki 2018). The fact that their model's accuracy would have a 10% swing based on simply adding or removing data that is a large motivator of the research in this thesis. We aimed to test that theory and see how that could be applied to improve accuracy instead to aid in creating a model which can detect hate speech at a percent greater than what was given in this experiment. In addition to that study we found a paper by Malmasi and Zampieri 2017 which scored a 78% accuracy when detecting hate speech using a character 4-gram model. They suggest applying a classification ensemble to the issue of hate speech detection in hopes of differentiating between profanity and hate speech.

2.2.3 Dependency

Identifying and handling negation in phrases and sentences (such as “that is *not* good”) is a must with sentiment analysis and classification. When the negation dependency is lacking in an analysis approach, such as the basic bag-of-words approach, where only the occurrences of words are considered and not the order, the results are significantly less accurate. For example, if a sentence contains a “no” or “not” at the beginning, the polarity of the following words should be changed for the rest of the phrase until another word comes along to break up the negation, such as but or a comma. This has its downsides as some words cannot be flipped, such as what Long et. al describe as intensifiers, and sentences all have different structures and possibly multiple conjunctions or dependent clauses, double negatives or *however*s for example. The authors negation-handling algorithm used a dependency-based parse tree to break apart the structure of the sentence and assign sentiment scores to sub-trees (Diamantini, Mircoli, and Potena 2016).

To simplify further, dependency in the realm of natural language processing is the idea of breaking down the words in a sentence by their type, such as if they are a verb, noun, conjunction, or adjective, to name a few. By forming relationships between these types of words we are able to decide whether the emotion of a word is dependent on a descriptor or needs to be negated due to a negative word such as *not* or *no*. This idea is then carried to a larger scale such as phrases on either side of a comma or conjunction such as *and*, and then to calculate the sentiment of the entire sentence or statement (Quan, Wei, and Ren 2013).

In addition to the basic strategies for sentiment classification of lexicon scoring and supervised learning, adding a factor which determines the syntax of the sentence seems to improve the accuracy of the classification. The two forms of sentence syntax representations are constituency grammar, which breaks down a sentence into a series of constituents made of smaller constituents until the constituent is simply a word and dependency grammar. Deng, Sinha, and Zhao 2017 research implemented the latter of the two forms of sentence syntax, which is done by making an association between two non-consecutive words in the sentence and making them part of a triplet. A triplet is the combination of the two associated words, the head word and dependent word (which is usually the subject, or modifier) and a description of their relationship. "I love pasta" is an example where "I" and "pasta" are the head and dependent respectively and "love" is the descriptor. Despite a few constraints such as the time and quality of the parsing strategy this study showed that adding the dependency strategy to sentiment classification was successful with better results in multi-gram and part-of-speech features. This was accomplished by the dependency grammar's ability to give context to words and relationships while using the supervised sentiment classification strategy (Deng, Sinha, and Zhao 2017).

It is clear that there are plenty of the nuances in sentiment analysis as described above and even more complexity when adding the contextual difficulties of hate speech in particular. Because of this we plan to train our model on very clear and agreed upon data. In doing this we would have a model which will be extremely precise and while probably not catching all tweets with hate speech, will likely be correct on the ones that it does classify as such.

3 Implementation

3.1 The Model

In order to tackle the many nuances and versions of hate speech we are using an ensemble method to combine the power of multiple machine learning and natural language processing classification models. Those being:

- Random Forests
- K Means Clustering
- Multinomial Naive Bayes
- Bernoulli Naive Bayes

The tweet is labeled as hate speech if any of the ensemble models classify the tweet as such. Different combinations of the models that are the components of our ensemble were tested, for example, if Random Forest and at least one of the other three models labeled a tweet hate speech or at least two models did. The best results, however, were using the approach of labeling a tweet as hate speech if any of he component models did so. The program is written in Python using the sklearn library (Pedregosa et al. 2011) for the different machine learning and natural language processing models.

3.1.1 Random Forests

The Random Forest model (Pedregosa et al. 2019d) was used due to its reputation of being able to handle most machine learning problems with decent accuracy. This was true in our ensemble as well as it is without a doubt the best scoring model when used by itself. We feared that the ensemble would only be as strong as the random forest model, however adding the other models to the mix did increase the score. After running multiple calibration tests to find the ideal parameters, this Random Forest model uses ten estimators, or forests, which are decision trees that determine classification.

3.1.2 K Means Clustering

For our K Means Clustering model (Pedregosa et al. 2019b) we group the training data tweets into 100 different clusters with a max iteration size of 500 iterations (these parameters were decided based on preliminary testing results). This is done without the model taking into account the tweet's hate speech classification, what is called unsupervised machine learning. Once grouped, each of the cluster's purity is calculated. The clusters which have 100% hate speech tweets are stored in a list. When the model then processes the test data, only the tweets which were categorized into one of the pure hate speech clusters is marked as hate speech. This is done with the intention to increase precision and avoid mislabeling a tweet as hate speech when it is not.

3.1.3 Naive Bayes

Both the Multinomial (Pedregosa et al. 2019c) and Bernoulli (Pedregosa et al. 2019a) models were used as a very basic baseline of the ensemble being simple and standard NLP models. Naive Bayes is a fairly rudimentary classification technique so we thought it important to have to see how useful and impact it could be in hate speech detection. During the origins of this study and participation in the International Workshop on Semantic Evaluation's Semeval Task Five: "Multilingual detection of hate speech against immigrants and women in Twitter" we found that combining Multinomial and Bernoulli Naive Bayes as an or ensemble improved the results so included them both here for that reason.

3.2 Testing Data

The data used in this analysis was the biggest challenge and actually crucial to our findings. We started with a data set of tweets labeled as hate speech, aggressive, targeted, or any mixture of the three. This was acquired while participating in the International Workshop on Semantic Evaluation's Semeval Task Five: "Multilingual detection of hate speech against immigrants and women in Twitter." (Basile et al. 2019). Although it provided a good platform for this thesis, we found many of the labeled tweets to be inconsistent and began to doubt the accuracy of their original classification process. We ended up using a second data set that had each tweet classified by multiple people and gave them a confidence score based on the amount of people who viewed it and rated it hate speech or not. This process seemed to have much more consistent data with fewer questionable classifications, making training the model much easier and more accurate. A more thorough explanation of how that data set was constructed is included below.

3.2.1 Preprocessing

In order to prepare the training and testing tweets we used sklearn's feature extraction library, specifically TfidfVectorizer. TfidfVectorizer converts all of the tweets into a feature matrix weighted by tfidf term weighting (a weight calculated by term frequency). Tf-idf term weighting was used due to it being found to be more accurate than the other common practice called BM25. This was determined in a study comparing the F1-score of models using data processed by the two methods (Kadhim 2019). Additionally, stop words such as *the* or *and* are being removed in order to reduce clutter and likely non-influential features. Links, punctuation, and capitalization was all left in, however.

3.2.2 Semeval Data Set

This data set is the training data used in SemEval 2019 Task 5 - Shared Task on Multilingual Detection of Hate: Multilingual detection of hate speech against immigrants and women in Twitter (hatEval) (Basile et al. 2019). Therefore, the tweets consist of non hate speech tweets and tweets with hate speech directed at women and immigrants. The data was collected through a variety of means. One route was to identify accounts that were or could be targets of hate speech and inspect tweets directed at them (public immigrant figures or women's advocates for example). From there, the tweet history of accounts which were discovered to use hate speech were stored. Lastly, tweets collected based on offensive, crude, or clearly targeted language or words. There are 10,000 total tweets in this data set split evenly between women and immigrant related content. It should be noted that there are many more hate speech tweets present in this data set in attempts to have a more even split between hateful and clean tweets (Basile et al. 2019). This means that hate speech tweets make up about 50% of the total tweets where as in reality that percentage is much lower. As stated above this data set appeared to have a good amount of inconsistent data, preventing the models from being trained accurately.

3.2.3 Data.World Data Set

The data set “Hate Speech Identification” is the main data set used in this experiment. It was put together by Davidson et al. 2017 during their study of issues related to offensive language and automated detection of hate speech. In order to construct this data they first went to Hatebase.org in order to acquire a list of language labeled as hate speech from the internet community. Then, tweets containing these words or phrases were identified from Twitter using Twitter API. By doing this there is a broader collection of hate speech than only speech directed at women or immigrants as in the semeval data set. The users of those tweets were identified and all of their tweet history was saved resulting in over 33,000 twitter users and over 85 million tweets. These tweets were shuffled and set of 25,000 tweets which also contained language from the original collection were uploaded to CrowdFlower, a site which provides crowd sourced annotations to data sets for machine learning. This is where this data set took a step further than the semeval data set. While on CrowdFlower the tweets were marked by the employees there as either hate speech (dirty), offensive, or neither (clean) based on the definitions of the categories as well as “to think not just about the words appearing in a given tweet but about the context in which they were used.” Of the provided tweets, most were labeled as offensive but not hate speech, around 75% of the time. Clean tweets were next at around 15% and lastly hate speech tweets around 5%. The remaining 5% were tweets that did not come to a conclusive classification (Davidson et al. 2017).An example tweet from this data set which was labeled hate speech is:

> "*LMFAOOOO I HATE BLACK PEOPLE https://t.co/RNvD2nLCDR This is why there's black people and [REDACTED]*"

We chose to remove the last word in the tweet for sensitivity and keep unnecessary language out of the thesis. Two examples from this data set which were not labeled hate speech are:

> "*Hardcore way to eat Mac and cheese*"

and

> "*My nigga was drinking Gatorade like it's not Dumass hot lol bitch drink water*"

These are good examples of a fairly innocent tweet but also one that could blur the lines of hate speech or not depending on who was the person saying it and its context.

Due to this data set providing us with more than a simple classification but also counts of the amount of people who labeled the tweet as either hate speech, offensive, or neither, we were able to run the model with different training sets. First, we ran the model by training and testing on all the tweets. Then, a 66% threshold (due to there being a minimum of three annotators from Crowdflower) was tested along with a 100% threshold after that. All three of these tests yielded varying results and scores.

3.2.4 Scoring

We measured our ensemble's performance with four scores also from an sklearn library (Pedregosa et al. 2011).

- Accuracy: Tweets with correct classifications / All Tweets

- Precision: Tweets correctly labeled hate speech / All Tweets labeled hate speech

- Recall: Tweets correctly labeled hate speech / All hate speech Tweets

- F-Score: A weighted average of precision and recall. More specifically, 2 x ((Precision x Recall) / (Precision + Recall))

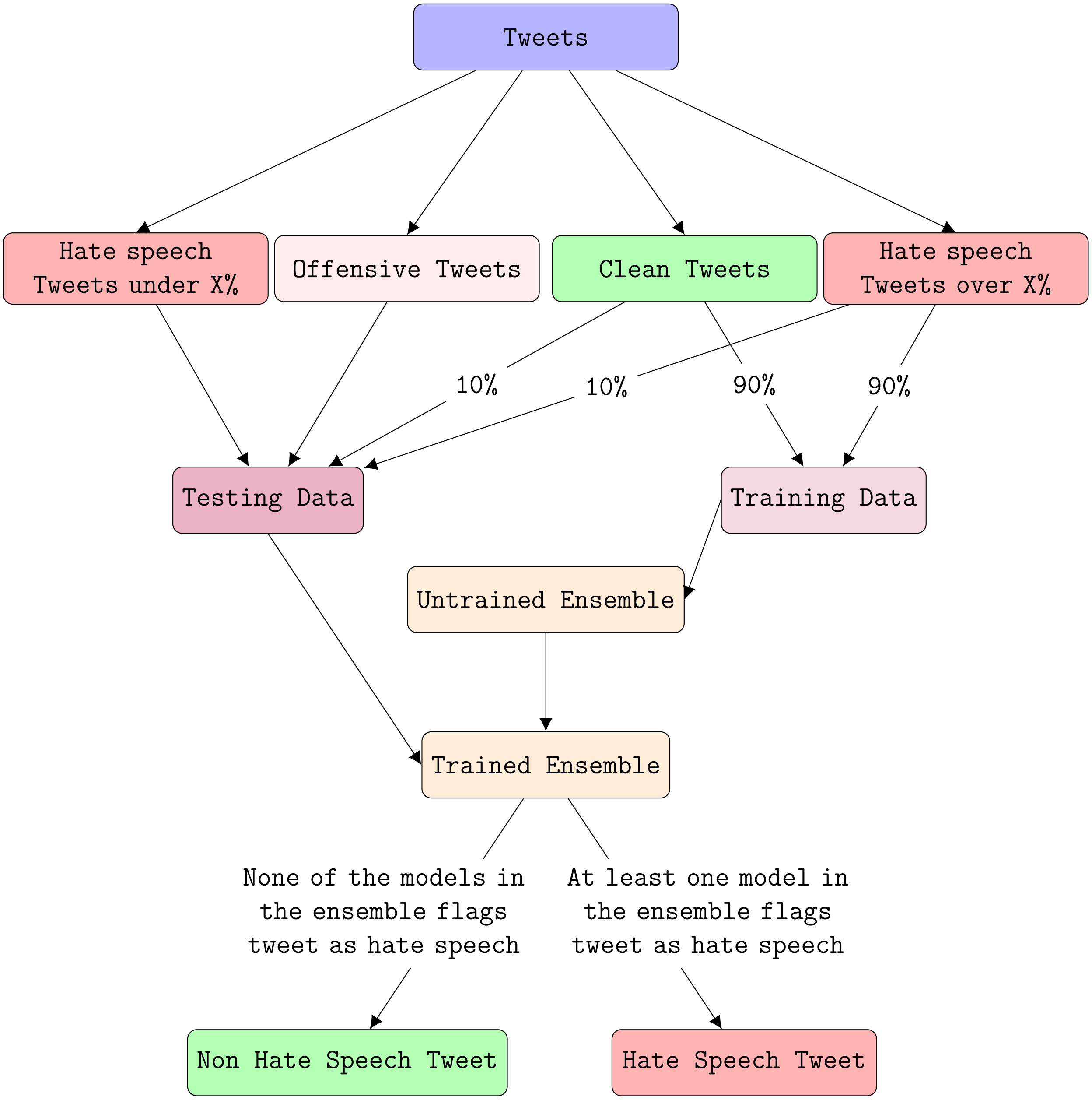
Tweets
Hate speech Tweets under X%
Offensive Tweets
Clean Tweets
Hate speech Tweets over X%
10%
10%
90%
90%
Testing Data
Training Data
Untrained Ensemble
Trained Ensemble
None of the models in the ensemble flags tweet as hate speech
At least one model in the ensemble flags tweet as hate speech
Non Hate Speech Tweet
Hate Speech Tweet

Figure 3.1: Flow chart of how our ensemble works. Breaking the tweets into categories based on whether they are hate speech or not, training on different sets of that data, then testing on all data. Tweets that any of the ensemble methods found to be hate speech were labeled as such, all others were labeled not hate speech.

4 Results

We collected four sets of data from our model. First, Table 4.1 shows the scores of each model in the ensemble, the ensemble score, and a base line of labeling all tweets as not hate speech in terms of F-Score, precision, recall and accuracy. This can be seen in the graph in Figure 4.1. Likewise, Table 4.2 and Figure 4.2, and Table 4.3 and Figure 4.3 show the scores when training the ensemble on 100% and 66% confident data, respectively. We are able to see how all of these figures compare in Figure 4.4. When looking at training with all data, accuracy was the highest score in this test, all others at or below 75% with an f-score of just over 50. training on 100% excelled in precision and accuracy about about 95% each, however, had a lower recall and overall f-score than the 66% trained model which had 83% f-score and 80% recall. This is helpful as it shows our ensemble was more specific in it's classification when trained with extreme data, however, it was able to cover a larger scope when using the slightly less strict threshold of 66%. In both cases performance was better using a threshold than training on all data apart from accuracy, where training on all data did slightly out perform the 66% threshold by about five percentage points.

Lastly, Table 4.4 and Figure 4.5 show the ensemble scores when using data from the semeval competition. Here we can see the how different models in the ensemble excel in different areas like precision and accuracy. This shows that there is not one model that is best for every situation, at least not with the methods we used here. For example, k-means clustering had higher precision than all other models, bringing up the precision of the ensemble, but it had low recall and accuracy comparatively

Test	F-Score	Precision	Recall	Accuracy
$AllModels$	0.559	0.748	0.54	0.943
$RandomForest$	0.559	0.749	0.54	0.943
$NaiveBayes$	0.485	0.471	0.5	0.942
$BernoulliNaiveBayes$	0.486	0.546	0.501	0.942
$KMeansClustering$	0.485	0.471	0.5	0.942
$AllNo$	0.478	0.458	0.5	0.916

Table 4.1: Average scores from training on all hate speech tweets from data.world data set.

Test	F-Score	Precision	Recall	Accuracy
$AllModels$	0.768	0.947	0.703	0.948
$RandomForest$	0.768	0.947	0.703	0.948
$NaiveBayes$	0.494	0.659	0.508	0.917
$BernoulliNaiveBayes$	0.478	0.458	0.5	0.916
$KMeansClustering$	0.484	0.558	0.503	0.918
$AllNo$	0.478	0.458	0.5	0.916

Table 4.2: Average scores from training on non hate speech and hate speech with 100% confidence from data.world data set.

Test	F-Score	Precision	Recall	Accuracy
$AllModels$	0.827	0.885	0.796	0.884
$RandomForest$	0.815	0.883	0.782	0.878
$NaiveBayes$	0.687	0.899	0.658	0.826
$BernoulliNaiveBayes$	0.724	0.899	0.688	0.84
$KMeansClustering$	0.464	0.81	0.518	0.756
$AllNo$	0.478	0.458	0.5	0.916

Table 4.3: Average scores from training on non hate speech and hate speech with 66% confidence from data.world data set.

Test	F-Score	Precision	Recall	Accuracy
AllModels	0.769	0.768	0.774	0.773
RandomForest	0.747	0.763	0.743	0.762
NaiveBayes	0.737	0.756	0.733	0.754
BernoulliNaiveBayes	0.753	0.765	0.749	0.765
KMeansClustering	0.441	0.799	0.536	0.609
AllNo	0.478	0.458	0.5	0.916

Table 4.4: Average scores from training on all hate speech tweets from Semeval data set.

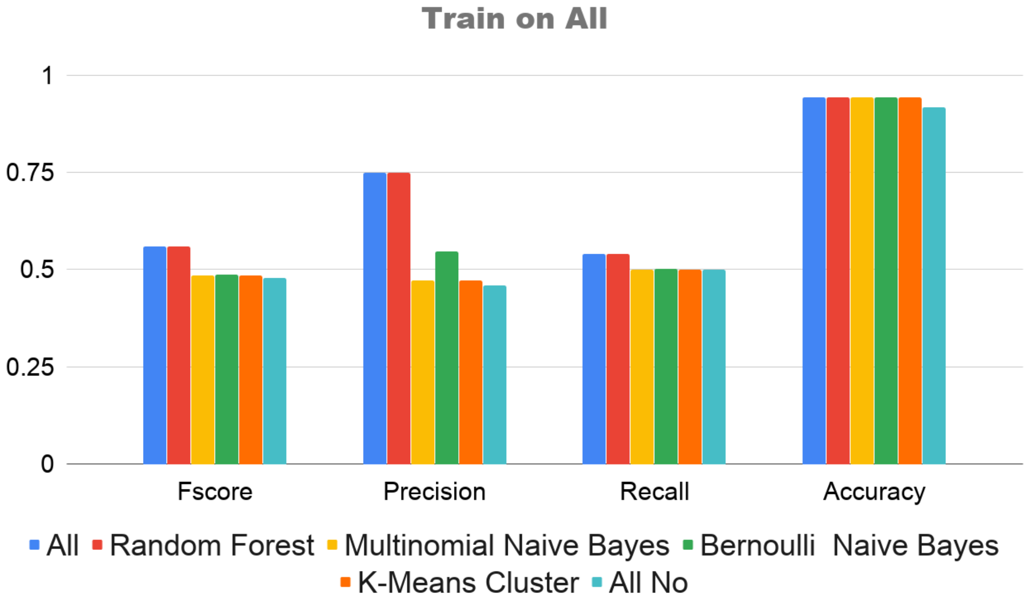

Figure 4.1: Average scores from training on all hate speech tweets from data.world data set.

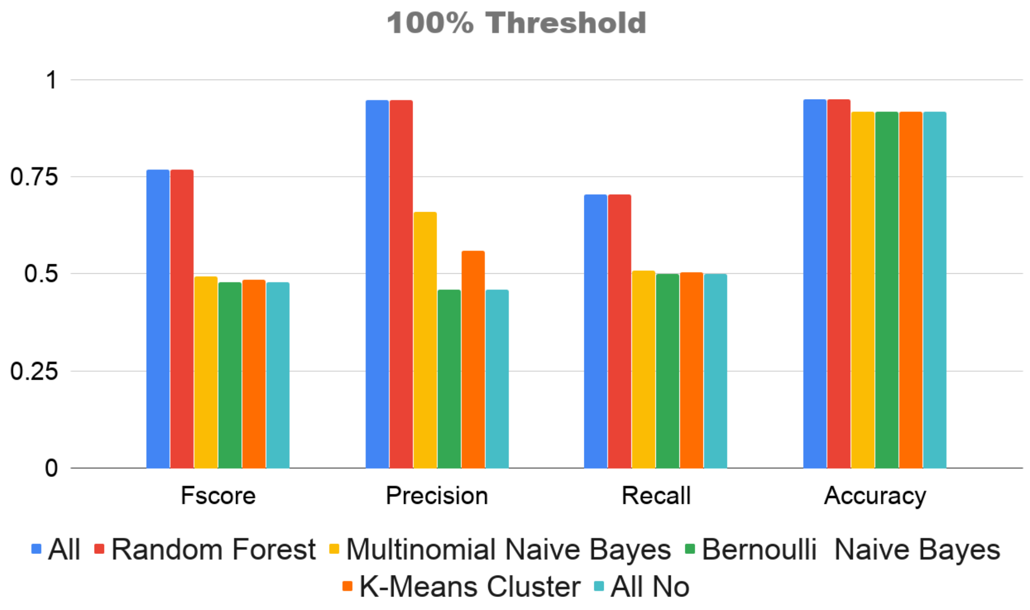

Figure 4.2: Average scores from training on hate speech and clean tweets with 100% confidence from data.world data set.

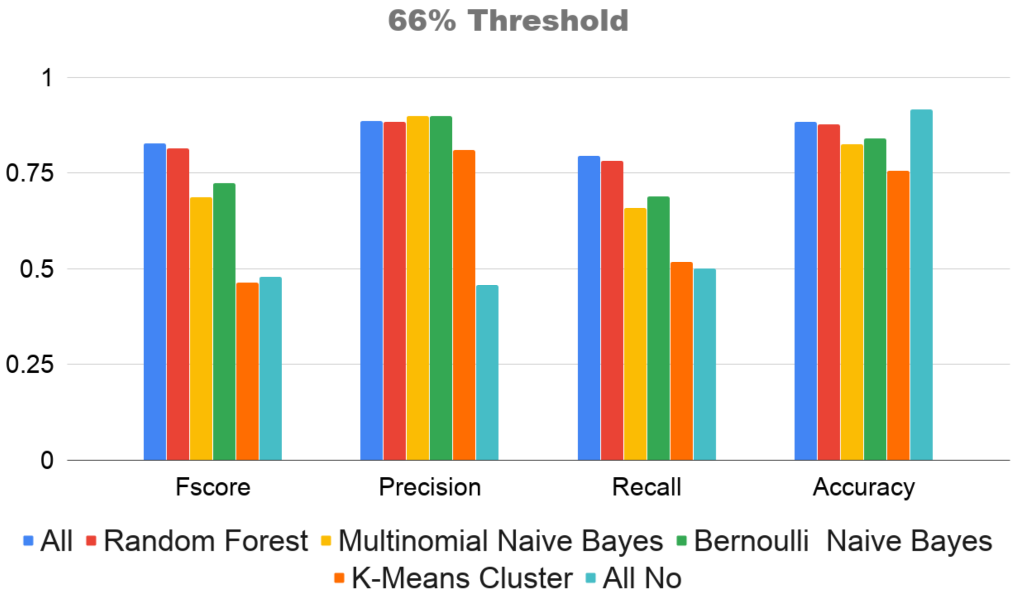

Figure 4.3: Average scores from training on hate speech and clean tweets with 66% confidence from data.world data set.

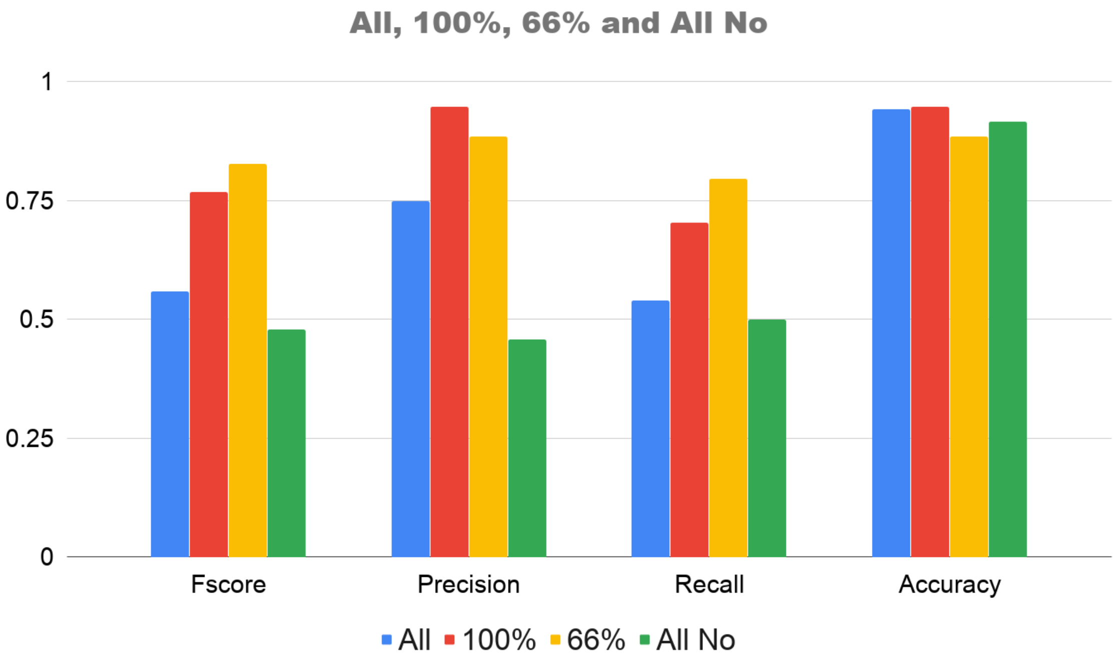

Figure 4.4: Average scores of the ensemble at all thresholds.

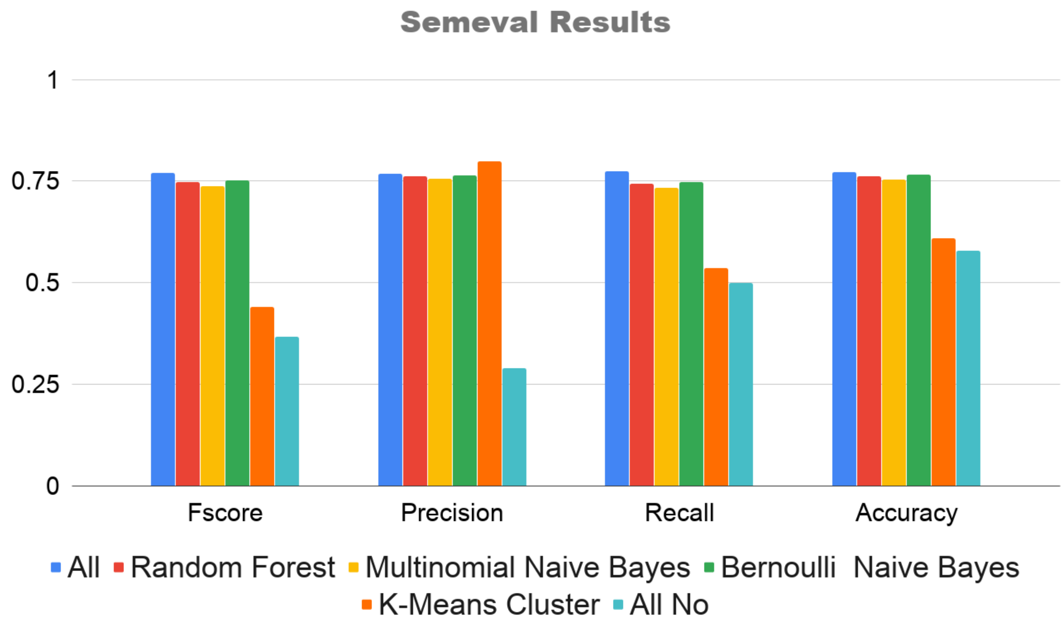

Figure 4.5: Average scores from training on all hate speech tweets from Semeval data set.

5 Conclusions

The results from these tests present two findings. First, and less prominent is that the ensemble will not provide worse results than that of any of its parts. If anything, the results will be slightly improved. Therefore, if there are resources available to tackle a problem with multiple models it will absolutely not hurt to do so when doing hate speech detection. There are many nuances to hate speech and varying models can pick up on different patterns making some more accurate but less precise where others precise but misses a good deal of cases. More importantly, however, we found that selectively training the model provided better results than simply training the model off of a random subset of the data. The selective "extreme" data (non-hate speech or offensive tweets and hate speech tweets over 66% confidence and 100% confidence in our two tests) increased the model's scores by the significant margin of nearly 30%, suggestion that the human element of labeling the test data increased the model's ability to recognize context. Having more people than three review and classify tweets as hate speech or not would allow for a more specific threshold to optimize the training of the model and perhaps improve the recognition of context even more, as there were still falsely labeled hate speech tweets from people quoting hateful content in an informative way as well as tweets labeled hate speech that simply used profanity. From here, future research would be able to better scrape tweets from twitter to be able to find more data and context to the tweets beyond text, such as replies, likes, retweets, and user bio information. This information could then be used to find patterns to hate speech that are beyond simply text. As we know words and phrases will change and they are different depending on where you are in the world. If we are able to pick up some other pattern to hate speech our online filters would be able to stay up to speed with lingo changes and we would be able to look at what common correlations there are that bring people to speak in such a way.

References

Banks, James (2010). “Regulating hate speech online.” In: *International Review of Law, Computers Technology* 24.3, pp. 233–239. ISSN: 13600869. URL: https://login.libpdb.d.umn.edu:2443/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=keh&AN=54862967&site=ehost-live (cit. on p. 6).

Basile, Valerio et al. (2019). “SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter”. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 54–63. DOI: 10.18653/v1/S19-2007. URL: https://www.aclweb.org/anthology/S19-2007 (cit. on pp. 15, 16).

Britannica, Encyclopedia (2019). *Schenck v. United States*. Website. Encyclopedia. URL: https://www.britannica.com/event/Schenck-v-United-States (cit. on p. 5).

Davidson, Thomas et al. (2017). “Automated Hate Speech Detection and the Problem of Offensive Language”. In: *CoRR* abs/1703.04009. arXiv: 1703.04009. URL: http://arxiv.org/abs/1703.04009 (cit. on pp. iii, 17).

Deng, Shuyuan, Atish P. Sinha, and Huimin Zhao (2017). “Resolving Ambiguity in Sentiment Classification: The Role of Dependency Features”. In: *ACM Trans.*

Manage. Inf. Syst. 8.2-3, 4:1–4:13. ISSN: 2158-656X. DOI: 10.1145/3046684. URL: http://doi.acm.org/10.1145/3046684 (cit. on p. 12).

Diamantini, C., A. Mircoli, and D. Potena (2016). "A Negation Handling Technique for Sentiment Analysis". In: *2016 International Conference on Collaboration Technologies and Systems (CTS)*, pp. 188–195. DOI: 10.1109/CTS.2016.0048 (cit. on p. 11).

Dwoskin, Elizabeth (2019). "How YouTube erased history in its battle against white supremacy". In: URL: https://www.washingtonpost.com/technology/2019/06/13/how-youtube-erased-history-its-battle-against-white-supremacy/ (cit. on p. 2).

Gelber, Katharine and Luke McNamara (2015). "The Effects of Civil Hate Speech Laws: Lessons from Australia". English. In: *Law Society Review* 49.3. Date revised - 2015-12-01; Last updated - 2016-09-28; CODEN - LWSRAA; SubjectsTermNotLitGenreText - Speech; Hate; Law; Australia; Grievances; Prejudice; Civil Law; Advocacy; Risk, pp. 631–664. URL: https://login.libpdb.d.umn.edu:2443/login?url=https://search-proquest-com.libpdb.d.umn.edu:2443/docview/1735651594?accountid=8111 (cit. on pp. iii, 5, 6).

— (2016). "Evidencing the harms of hate speech." In: *Social Identities* 22.3, pp. 324–341. ISSN: 13504630. URL: https://login.libpdb.d.umn.edu:2443/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=keh&AN=114016476&site=ehost-live (cit. on p. 6).

Kadhim, A. I. (2019). "Term Weighting for Feature Extraction on Twitter: A Comparison Between BM25 and TF-IDF". In: *2019 International Conference on Advanced Science and Engineering (ICOASE)*, pp. 124–128. DOI: 10.1109/ICOASE.2019.8723825 (cit. on p. 16).

Kann, Laura et al. (n.d.). "Sexual Identity, Sex of Sexual Contacts, and Health-Risk Behaviors Among Students in Grades 9–12 — Youth Risk Behavior Surveillance, Selected Sites, United States, 2001–2009". In: URL: https://www.cdc.gov/mmwr/preview/mmwrhtml/ss6007a1.htm (cit. on p. 2).

Koumpouri, Athanasia, Iosif Mporas, and Vasileios Megalooikonomou (2015). "Evaluation of Four Approaches for "Sentiment Analysis on Movie Reviews": The Kaggle Competition". In: *Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS)*. EANN '15. Rhodes, Island, Greece: ACM, 23:1–23:5. ISBN: 978-1-4503-3580-5. DOI: 10.1145/2797143.2797182. URL: http://doi.acm.org/10.1145/2797143.2797182 (cit. on pp. 8, 9).

Malmasi, Shervin and Marcos Zampieri (2017). "Detecting Hate Speech in Social Media". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. Varna, Bulgaria: INCOMA Ltd., pp. 467–472. DOI: 10.26615/978-954-452-049-6_062. URL: https://doi.org/10.26615/978-954-452-049-6_062 (cit. on p. 11).

— (2018). "Challenges in Discriminating Profanity from Hate Speech". In: *CoRR* abs/1803.05495. arXiv: 1803.05495. URL: http://arxiv.org/abs/1803.05495 (cit. on p. iii).

Meyer, Ilan H. et al. (2019). "Sexual Orientation Enumeration in State Antibullying Statutes in the United States: Associations with Bullying, Suicidal Ideation, and Suicide Attempts Among Youth". In: *LGBT Health* 6.1. PMID: 30638436, pp. 9–14. DOI: 10.1089/lgbt.2018.0194. eprint: https://doi.org/10.1089/lgbt.2018.0194. URL: https://doi.org/10.1089/lgbt.2018.0194 (cit. on p. 2).

Nett, Danny (2019). "Is YouTube Doing Enough To Stop Harassment Of LGBTQ Content Creators?" In: URL: https://www.npr.org/2019/06/08/730608664/

is-youtube-doing-enough-to-stop-harassment-of-lgbtq-content-creators (cit. on p. 2).

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan (2002). "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques". In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*. EMNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 79–86. DOI: 10.3115/1118693.1118704. URL: https://doi.org/10.3115/1118693.1118704 (cit. on pp. 7, 8).

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830 (cit. on pp. 13, 18).

— (2019a). *BernoulliNB*. Website. URL: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html?highlight=nb#sklearn.naive_bayes.BernoulliNB (cit. on pp. iii, 14).

— (2019b). *KMeans*. Website. URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html?highlight=k%5C%20means#sklearn.cluster.KMeans (cit. on pp. iii, 14).

— (2019c). *MultinomialNB*. Website. URL: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html (cit. on pp. iii, 14).

— (2019d). *RandomForestClassifier*. Website. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html?highlight=random%5C%20forest#sklearn.ensemble.RandomForestClassifier (cit. on pp. iii, 14).

Pelle, Rogers, Cleber Alcântara, and Viviane P. Moreira (2018). "A Classifier Ensemble for Offensive Text Detection". In: *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*. WebMedia '18. Salvador, BA, Brazil: ACM, pp. 237–243. ISBN: 978-1-4503-5867-5. DOI: 10.1145/3243082.3243111. URL: http://doi.acm.org/10.1145/3243082.3243111 (cit. on pp. 9, 10).

Quan, C., X. Wei, and F. Ren (2013). “Combine sentiment lexicon and dependency parsing for sentiment classification”. In: *Proceedings of the 2013 IEEE/SICE International Symposium on System Integration*, pp. 100–104. DOI: 10.1109/SII.2013.6776652 (cit. on p. 12).

Saha, Koustuv, Eshwar Chandrasekharan, and Munmun De Choudhury (2019). “Prevalence and Psychological Effects of Hateful Speech in Online College Communities”. In: *Proceedings of the 10th ACM Conference on Web Science*. WebSci fffdfffdfffd19. Boston, Massachusetts, USA: Association for Computing Machinery, 255fffdfffdfffd264. ISBN: 9781450362023. DOI: 10.1145/3292522.3326032. URL: https://doi.org/10.1145/3292522.3326032 (cit. on p. 2).

Schenck v. United States (2019). URL: http://academic.eb.com/levels/collegiate/article/473962 (cit. on p. 5).

Schmidt, Anna and Michael Wiegand (2017). “A Survey on Hate Speech Detection using Natural Language Processing”. In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Valencia, Spain: Association for Computational Linguistics, pp. 1–10. DOI: 10.18653/v1/W17-1101. URL: https://www.aclweb.org/anthology/W17-1101 (cit. on pp. 4, 9).

Soral, Wiktor, Michafffdfffd Bilewicz, and Mikofffdfffdaj Winiewski (2018). “Exposure to hate speech increases prejudice through desensitization”. In: *Aggressive Behavior* 44.2, pp. 136–146. DOI: 10.1002/ab.21737. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ab.21737. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ab.21737 (cit. on p. 6).

Watanabe, H., M. Bouazizi, and T. Ohtsuki (2018). “Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform

Hate Speech Detection". In: *IEEE Access* 6, pp. 13825–13835. DOI: 10.1109/ACCESS.2018.2806394 (cit. on p. 10).

Wenguang, Yu (2018). "INTERNET INTERMEDIARIES' LIABILITY FOR ONLINE ILLEGAL HATE SPEECH". eng. In: *Frontiers of Law in China* 13.3, p. 342. ISSN: 1673-3428 (cit. on pp. 4, 5).

Zimbra, David et al. (2018). "The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation". In: *ACM Trans. Manage. Inf. Syst.* 9.2, 5:1–5:29. ISSN: 2158-656X. DOI: 10.1145/3185045. URL: http://doi.acm.org/10.1145/3185045 (cit. on p. 8).