Analysis of Hate Speech on Twitter: What qualifies, what slides, and what it means

Kareem Aboueich
Computer Science
Portland State University
Portland, Oregon, USA
aboueich@Pdx.edu

ABSTRACT

The most difficult aspect for differentiating between hate speech and offensive language is recognizing the context behind the speaker. Intuitively, it makes sense that censoring particular words and racial slurs would help improve this, but learning how to identify the speech and separate it from offensive language is a gray area we still have not figured out how to navigate between the two. In order to gauge some of the reasons why, I will extract a small sample of an over 25,000-line sample of Twitter data to try and identify what qualified as hate speech, why it was, and what should have been and was not. Contextualizing the data is crucial, as some tweets that were labeled hate speech were not and vice versa. Intriguingly, the data I analyzed shows that speech that is considered racist or homophobic is generally seen as hate speech, a good starting point, however, speech that is sexually explicit or outright sexist is just seen as offensive. Also interestingly, some tweets that do not contain slurs or vulgar language are significantly more difficult to analyze, as context is often misunderstood or is not clear.

KEYWORDS
Twitter, hate speech, offensive language, political speech, censorship, algorithm, content moderation, racism, sexism, homophobia

1 Introduction

As technological advancements figure out more and more ways to move our lives behind a screen, content moderation has been an area that has not quite seen the advancement that it needs in order to keep up with the times. And this is something that, unfortunately, has gotten much worse as time passed. Treading the fine line between what is defined as hate speech and what is defined as offensive speech is difficult, and even the definition itself leaves a few loopholes. Per Google, hate speech is defined as "*abusive or threatening speech or writing that expresses prejudice against a particular group, especially on the basis of race, religion, or sexual orientation*" (Google Dictionary definition). Finding a definition for offensive speech is even more difficult, as the dictionary does not have an exact definition for it. This does not take into account threats of violence that are not catered to a particular group – should that be qualified as hate speech? Or just offensive speech? And when should content moderators take action?

Before internet was as prominent as it was, most hateful ideologies were not as public, as society typically frowned upon it. Social media, in particular, has changed this significantly. This is partially due to the fact that an increasing number of the global population has access to the internet, and that creating social media accounts is free, but primarily, the anonymity that the internet gives its users enables ideas that are not socially acceptable to be shared, and instead, will amplify them, not because it is a flaw, but because that is how the algorithms are designed. There are a few issues when it comes to this.

First, because of the way that social media algorithms work, they are designed to share content that the site

thinks you will enjoy based on your previous searches, your friends, or the general fields you post in. It becomes apparent how this can be incredibly dangerous, as the algorithm will unknowingly connect white supremacists, neo-Nazis, and other hate groups together because that is what the algorithm is designed to do. And hate groups having a voice and being able to talk to other hate groups will bring legitimacy to their voice. There is also the issue of more moderate views being exposed to extremist, hateful rhetoric. This is more of a problem on YouTube, and many have addressed the danger of the YouTube recommended algorithm work the way it does – specifically, it does not take very long to get from an informative news report to incredibly crude speech about immigrants. This is not what this paper will focus on; instead, it will focus on what speech is detected, either has hateful or offensive, and how it was distinguished.

Second, often times, many minority personalities will often get their posts marked and removed as hate speech due to the biggest problem when it comes to identifying hate speech – context. To a machine learning algorithm, understanding context is not the primary goal, but rather to filter between hate speech and offensive speech respectively, and to remove content that is considered hate speech. However, it can completely miss important context as to why certain language or slurs are used. For example, a lot of music lyrics use vulgar terms in regards to women, and across races or cultures, the use of certain terms or slurs is permissible. Distinguishing between that, and knowing what is acceptable in a certain context and what is not, is a very key factor in effective content moderation.

In order to help distinguish between them, the data that I am using contains a breakdown of tweets that were labeled as hate speech, offensive speech, or neither. The goal is to break down why each tweet was labeled as it was, and what tweets possibly slipped through the cracks even though they shouldn't have.

2 About the Data

This data was taken from a previous study on the same topic. The paper is titled *Automated Hate Speech Detection and the Problem of Offensive Language* and it was written by Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber (referenced below). The data was gathered by obtaining words and phrases deemed hate speech compiled by a site called *hatebase.org*, and searching Twitter for tweets containing that language. The data was then analyzed by humans, who were asked to rate each tweet as hate speech, offensive speech, or neither.

2.1 Breaking down the data

Their data is beneficial for my analysis in that it is random and was analyzed by humans working at CloudFlower and not an AI algorithm, which makes it more ideal for understanding things like context, and made their analysis differentiating hate speech and offensive speech different from how an AI or a Twitter algorithm would recognize it. The data is sorted in 5 columns, with a numeric code for each tweet, identified as follows:

1. Count: the number of people at CrowdFlower who analyzed each tweet. a minimum of 3 people analyzed each tweet with no given maximum.
2. Number of people (from total) who classified the tweet as *hate speech*
3. Number of people (from total) who classified the tweet as *offensive speech*
4. Number of people (from total) who classified the tweet as *neither hate speech nor offensive speech*
5. A label given by the majority of analyzers:
    a. 0: hate speech classification
    b. 1: offensive speech classification
    c. 2: neither hate speech nor offensive speech classification

2.2 Objectives from using this data

The goal here is to try and analyze a few things from this data:

1. What primarily contributes to a hate speech classification?
2. What primarily contributes to an offensive speech classification?
3. Which hate speech tweets were not classified as hate speech, and why?

For the last question, there were some tweets that were classified as both hate speech and offensive speech, but were labeled as offensive speech because a majority of the reviewers saw it as offensive speech and not as hate speech, and the goal is to figure out why.

Seeing as the data is over 25,000 tweets, I will not be analyzing all of it for this class, but rather analyzing a subsect of it. The goal is to include an even number of tweets that were found to be hate speech and ones that were found to be offensive speech.

3 General Data Analysis

The overwhelming majority of the tweets in this data sample were classified as offensive speech, and almost all of the ones that were classified as offensive speech had some sort of curse word or vulgar reference. Tweets that were music lyrics with vulgar language were also considered hate speech. The data gets more interesting when discussing crude language against women specifically.

Hate speech, on the other hand, usually contained outright slurs regardless of context, and most were classified correctly. The slurs that commonly resulted in being classified as hate speech were primarily against the African American community, the Hispanic community, and the Asian American community.

Tweets that were classified as neither often times had no swear words or crude humor, and often, the tweets were often in a multitude of different fields (popular culture, news, sports, food, etc.) as opposed to offensive speech, which was predominantly sexist, and hate speech, which was predominantly racist or homophobic.

There were some tweets that were written near identically, however, one would be considered offensive and one would be considered hate speech. Indicating reasons why was tricky, but a common factor observed is that if the tweet contained the word ‘racist’ (not even ‘racism’), it was more likely to be considered hate speech rather than offensive speech.

The context that greatly resulted in tweets being classified as hate speech was in regards to politics in particular. The analysts were typically good at distinguishing between offensive banter between two sports team fans arguing over their respective teams, for example, and hateful speech going too far and attacking the person’s ethnicity, religion, race, or sexual orientation to name a few examples.

4 Offensive Speech Analysis

One of the gray areas that lead to more tweets being defined as offensive speech rather than hate speech was violence. Threats were dominantly considered to be offensive, so long as they weren’t followed up with a slur of sorts. While violence against minorities fit the hate speech classification, it gets more complicated with what arguably is hate speech against women, however, which is where the disparity between gender and the other categories exists.

4.1 Offensive tweets regarding women

Many of the tweets refer to women in vulgar terms (b----, h--) and end up getting classified as offensive even though it could easily be considered hate speech. While use of racial slurs, as an example, is considered by society to be hateful speech, referring to women in derogatory terms is not, and that is difficult to determine if that is a societal effect or a fault on the part of the readers for not understanding the context behind the tweet, especially when many of them are crudely sexual tweets about women. Some of this may be to the socially acceptable parts of our society that convey such sentiment, such as music or movies, despite the fact that often times, the people part taking in such a production are vehemently against sexism.

Sexism is not seen as hate speech in the same context that racism is, implying that the human bias plays more of a role in how sexist speech is identified. This could partly be a reason why female harassment online is as prominent as it is – if the speech is not classified as hate speech, it becomes increasingly difficult to take it down. In Twitter's hateful conduct policy, it mentions that you cannot promote violence or directly attack others on a multitude of different bases; however, it does not make any specific mention on harassment, allowing people to tread the fine line there and get away with speech that, if put into a different context, would get flagged.

Additionally, often times sexist speech would get at least one person who said that it was hate speech, but the majority said it was not, so therefore, it was not classified as hate speech. This did not nearly happen as much as it should have, and is not necessarily a good indicator of how content moderation views sexism.

This is one of the most concerning conclusions I could have come to from researching this data – and much like many societal problems, this is also a systemic one, and there could certainly be a case made that an AI would be better at being fair at detecting sexist speech as hate speech than what the people analyzing would find if that is indeed a systemic bias they have.

4.2 Tweets that crude sexual comments

There were posts that were very sexual in nature but were not sexist, which were also flagged as offensive speech, though that is also treading another gray area, seeing as it is not necessary offensive so much as it is meant for a more mature audience. This is a flaw in how the classification system works – and while the system should remove sexual images and videos that are made public off the site, per the terms and conditions, having it in the same category as sexism and threats of violence is not a fair classification. Similar to Google's "Safe Search" feature, something similar could be implemented on Twitter, for example, to allow these sorts of comments to those who want to view them but not put them in the same class as racism or other forms of prejudice.

5 Hate Speech Analysis

What generally differentiated hate speech from offensive speech was the use of racial slurs, though there were some that slid through the cracks, and some uses that were considered culturally acceptable in context that were mostly classified correctly. Hate speech was predominantly political, and was mostly against minority communities. Some interesting points that made hate speech interesting to study from the data that I used (Important to note: These tweets were judged exclusively by people, not by AI/other machine learning algorithms):

1. Tweets from the same account with two varying spellings of a slur resulted in both being classified as hate speech. In this case, the people learned the context behind the person's previous tweets and understood context and intention. This is not always accurate, but generally can be a safe assumption to make (if it walks like a duck, and looks like a duck, and quacks like a duck…). The effectiveness of this method will be discussed later on in the paper.
2. Tweets with specific words were almost always classified as hate speech, regardless of context, which was mostly agreed with. Often times, these were slurs against a particular group.
3. Tweets that were 3 words or less and contained a slur were almost always recognized as hate speech
4. Sometimes, tweets with the term 'racist' were considered to be hate speech in some contexts. There is not really a clear reason for this, as often times the tweets did not contain any slurs or implications of hate speech.
5. Tweets that discussed violence were considered offensive unless they referenced rape or referenced a minority group. Violence against a minority group was almost always

considered to be hate speech, and mentions of rape were likewise also considered to be hate speech. The latter is an interesting point because it would lead one to think that speech is not hate speech against women unless they are threatened with sex crimes.

6. Sometimes responses to hate speech was considered hate speech, especially if that speech contained a slur. This was usually justified.

6 An Engineered Solution

The interesting thing about how people judged is the frequency that tweets in a certain category would slip through the cracks, mainly sexism. This is an evident problem.

To get a rough idea of how a machine could respond to this without building something that uses AI, I took a section of the data set and decided to manually search for certain language to see if the simplest of methods would be adequate in that regard – Any tweets that use a certain word would immediately be flagged as hate speech. In order to do this, I compiled a short list of words and put them into categories of racism, sexism, homophobia, and miscellaneous, which consisted of other hateful language that did not necessarily fit in one of the previous categories.

To keep it simple, I picked out some common words in the tweets that met each classification, but the difference here is that I placed words that were derogatory against women in the hate speech category. The idea here was to try and get some understanding of how we can

6.1 Test Filters

In order to conduct this test, I took a sample of 1,250 tweets from the data and decided to run a few search tests on them based on a few conditions. Important disclaimer here: this method did not use any AI or machine learning algorithm, so this does not really solve the contextualizing the speech problem so much as it tries to find a new way to solve the sexism is classified as offensive speech problem.

For the setup, here is how the categories were split:

1. Hate Speech: This category mostly contained slurs that were used in throughout the list of tweets; however, the difference here is that they also included some commonly used derogatory terms against women.
2. Offensive Speech: This category mostly contained swear words and words that could be used to convey violence (e.g. kill, murder, etc.). This did not include the term rape, which, for the purpose of this study, I decided to place in the hate speech category.

Lastly, if a tweet both qualified as hate speech and as offensive speech, it was placed in the hate speech category. If a tweet was neither classified as hate speech nor offensive speech, it would go in the neither category.

6.2 Tests Conducted

To get somewhat of a variety of how the simple search algorithm would do with finding hate speech, I devised 4 different conditions of which to determine if speech is hate speech or offensive speech:

1. Banning the use of certain words altogether: The only rule here is that if a tweet has any of the words in the hate speech category, it would automatically be classified as hate speech.
2. Banning the use of certain words if preceded immediately by ‘you’: The rule here is that if a tweet has any of the words in the hate speech category and is preceded by ‘you’, it would be classified as hate speech. If the term was not preceded by ‘you’, it would not count. Tweets that were less than 3 words or less and contained a slur would get a hate

speech classification even if they were not preceded by ‘you’.

3. Banning the use of certain words if preceded immediately by ‘you’ and an adjective: The rule here is that if a tweet has any of the words in the hate speech category and is preceded by ‘you’ or ‘you’ and an adjective, it would be classified as hate speech. Likewise, to the previous condition, tweets that were less than 3 words or less and contained a slur would get a hate speech classification even if they were not preceded by ‘you’ or ‘you’ and an adjective.
4. Banning the use of certain words if preceded immediately by ‘you’ and an adjective, as well as banning tweets from users who previously tweeted something that was classified as hate speech: For this one, one of any of these conditions had to be met:
    a. Tweet contains a term in the hate speech category, preceded by either ‘you’ or ‘you’ and an adjective
    b. Tweet comes from a user who previously tweeted something that was classified as hate speech AND whose new tweet contains a slur, be it misspelled or not.

Each of these tests adds a filter to the previous one. The idea is to do two things:

1. Get some sort of idea of how well the find bot will detect hate speech.
2. Observe how it compares to how the humans judged it.

6.3 Tests Data

The sample of 1,250 tweets extracted from the larger sample of test data consisted of the following, according to the human reviewers:

1. 894 tweets were considered offensive speech
2. 201 tweets were considered hate speech
3. 155 tweets were considered neither hate speech nor offensive speech.

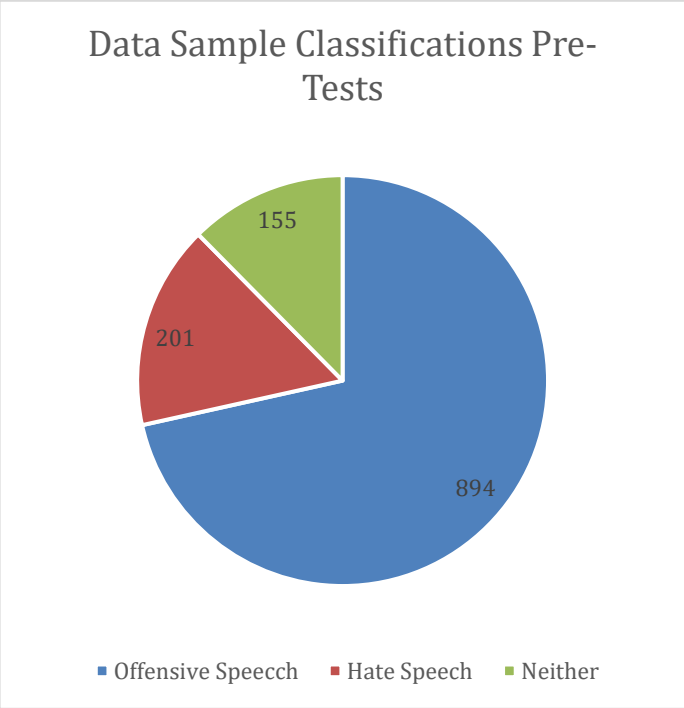

Figure I: Classification of pre-test data in chart form, for visual clarification.

The sample was extracted from by using a random number generator between 1 and the list maximum minus 1,250 to make it fair. The number that was chosen would be the starting point, and would pick the next 1,250 tweets from there.

It is important to note here that a different number could have yielded different results.

7 Electronic Provisions

Aside from the first test, the results of the other tests were pretty close in their results. All tests had more tweets classified as hate speech versus what the humans decided.

7.1 Test I Results

The first test yielded the most hate speech of all the tests, as such:

1. 876 tweets were classified as hate speech
2. 176 tweets were classified as offensive speech
3. 196 tweets were classified as neither hate speech nor offensive speech

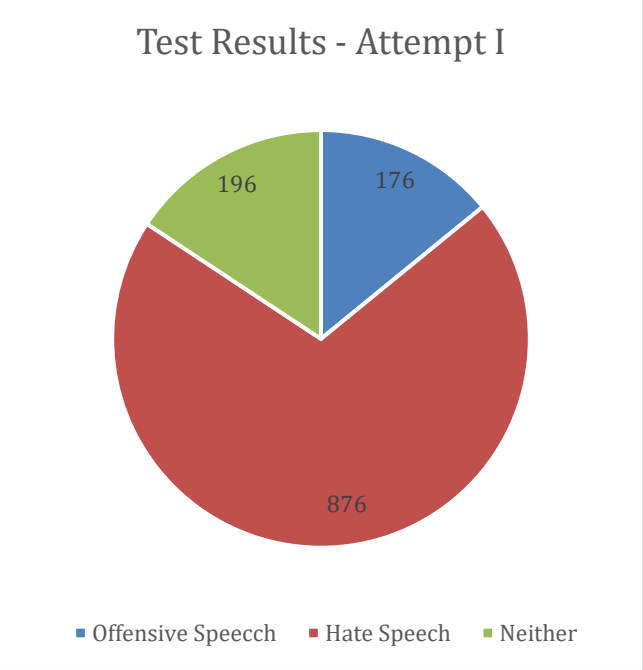

Figure II: Test results for Attempt I, visualized.

Due to the lack of context for each tweet, as well as a very general classification, it is not much of a surprise that this result yielded the most tweets in the hate speech category. This is not a very effective method of filtering hate speech.

Part of the results being as skewed as they are is due to the inclusion of derogatory terms against women to the hate speech category.

7.2 Test II Results

The second test yielded much closer results to how the humans rated the tweets:

1. 301 tweets were classified as hate speech
2. 773 tweets were classified as offensive speech
3. 176 tweets were classified as neither hate speech nor offensive speech

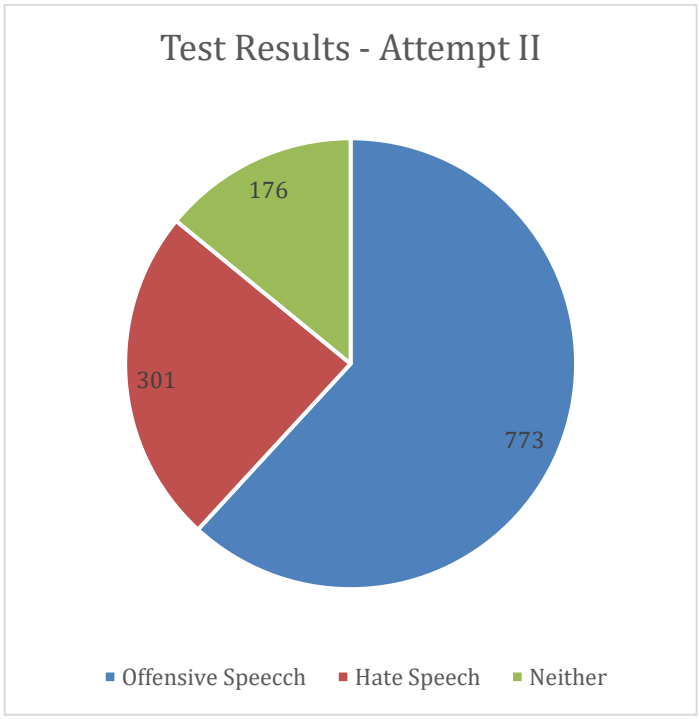

Figure III: Test results for Attempt II, visualized.

The advantage of having the 'you' precede the word in the hate speech category is that it adds a filter for words that might not be used to convey a person. Obviously, certain racist, homophobic, xenophobic, etc. slurs would be classified as hate speech regardless should something like this actually be implemented, but it does a better job at distinguishing a reason why a certain term is used.

7.3 Test III Results

The third tests yielded slightly more hate speech results than its predecessor, particularly because it added to the previous rule:

1. 377 tweets were classified as hate speech
2. 697 tweets were classified as offensive speech
3. 176 tweets were classified as neither hate speech nor offensive speech

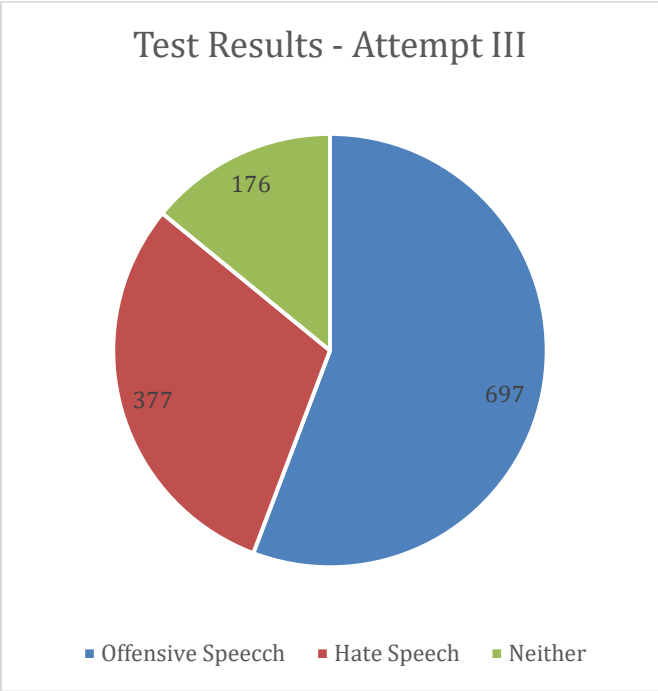

Figure IV: Test results for Attempt III, visualized.

Adding the condition that an adjective can precede 'you' helps to add to context a little bit, as often times, people will use words for emphasis. It is rare that you will find tweets with more than two adjectives before a potential slur, but for the sake of this study, I decided to focus exclusively on one adjective to get a rough idea of the impact it would have. The result was about a 20% increase in tweets classified as hate speech.

7.4 Test IV Results

The final test had a negligible impact on the results as a whole, only having a single digit impact on results compared to the previous test:

1. 380 tweets were classified as hate speech
2. 694 tweets were classified as offensive speech
3. 176 tweets were classified as neither hate speech nor offensive speech

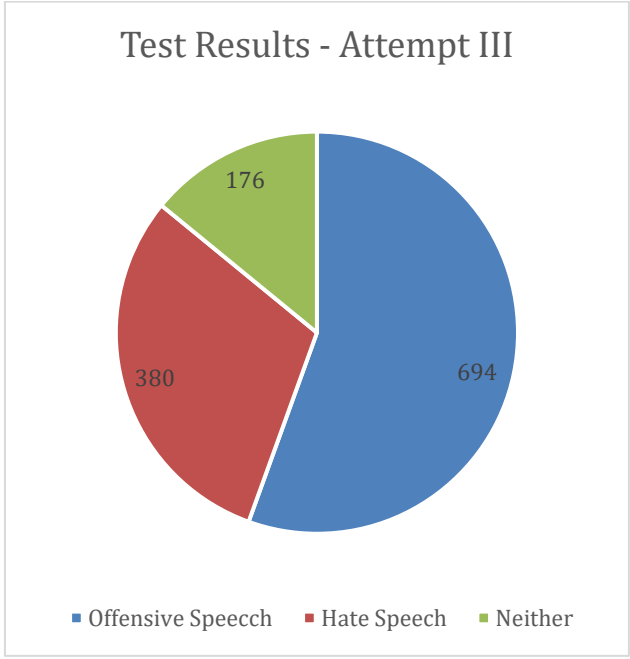

Figure V: Test results for Attempt VI, visualized.

While humans sometimes were suspicious of tweets that contained slurs that were tweeted by the same account, and rightfully so, the machines did not really seem to be phased by it, only having a 3 tweet difference compared to the previous test. However, this could differ for a different random sample, which could yield different results.

It is safe to assume that, from the same account, a previous tweet containing a slur that was classified as hate speech likely will mean their next tweet(s) containing a slur, even if misspelled, should also fall under the hate speech label. However, it does not mean that this should always be how they are classified (e.g. a minority tweets a tweet with a slur that gets falsely flagged, if they were to tweet that same slur, having it automatically flagged would not be effective at its job)

8 Analyzing the Results

While it is impossible to determine if the search algorithm did a better job from the humans searching through the data, the primary issue in content moderation is not that too many tweets are being

taken down for hate speech, but that not enough tweets are being taken down for hate speech, and given that 3 of the tests yielded results that were close to what the humans came up with, it is safe to say that they could have done a better job detecting certain prejudices (e.g. sexism primarily) than the human reviewers could.

None of the methods tested were perfect or anywhere near though; many tweets that were considered sexist still slipped through the cracks and were classified as offensive speech even though they should have been classified as hate speech. Running a similar test where tweets by category (e.g. sexist tweets, racist tweets, xenophobic tweets, etc.) and seeing how the human reviewers rated them would be the next step in figuring out how to better content moderation.

9 Conclusion

The way in which hate speech is processed is still flawed. While intrinsic human bias will always exist, the fact that sexism is not nearly picked up as hate speech as much as racism and homophobia is a concerning factor in how we process and analyze hate speech, and is something that needs to be improved on. Humans generally understand context well, however, at the volume that posts on Twitter are made, it is easy for a view vicious posts to slip through and not be removed, even if flagged by multiple people. Detecting accounts that have posted hate speech in the past can make it easier to determine if their future tweets will also be hate speech. Violent speech generally should be considered hate speech, as it very much fits the "abusive or threatening" part of the definition of hate speech. And lastly, it is clear that training the humans overseeing this system is incredibly important to ensuring the right things get removed and that context is understood.

The study shows us that it is possible for machines to do effective content moderation so long as they are trained. In the real world, classifying tweets by just searching for words is effective for yielding results on a small sample, but many still slipped through the cracks, and the machines needs to be able to keep up with societal and social trends that introduce new terms and memes into the mainstream. Often times, those forms of hate speech are the hardest to detect, as there are no visual queues to give the tweets away unless you understand the context it is in.

Both humans and machines can easily detect the obvious tweets, for example, ones containing blatant racial slurs that are societally unacceptable, but they get more complicated for terms that are not yet in the mainstream. If society does not look down on the use of a particular term, it complicates its classification as hate speech, and often times, many of these terms blend in mainstream speech to make criticism of them absurd.

Ultimately, the largest problems in content moderation lie in understanding context, the constant revolution of social, political, and societal acceptability of various terms, and the need for more content moderators, as billions of posts are made worldwide, and monitoring all of that by people is next to impossible unless they make a significant percentage of your company's employee base, which most companies will choose not to do. What it will come down to is how effectively can we train our software to do this work for us, and how we as a society can better ourselves in terms of condemning the use of this content so that way it becomes easier for the future content moderators and possibly their inhuman sidekicks.

ACKNOWLEDGMENTS

I would like to thank Professor Kristin Tufte, the instructor for this course, for providing advice, ideas, and resources for this project.

Special thanks also goes to Molly Shove who helped me obtain the data I studied, as well as miscellaneous articles that were excellent sources for research and great reads on the subject matter.

REFERENCES

[1] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated Hate Speech Detection and the Problem of Offensive Language,* 4 pages

[2] Warner, W., and Hirschberg, J. 2012. *Detecting hate speech on the world wide web.* 8 pages

[3] Waseem, Z., and Hovy, D. 2016. *Hateful symbols or hateful people? predictive features for hate speech detection on Twitter,* 6 pages

[4] Waseem, Z. 2016. *Are you a racist or am i seeing things? annotator influence on hate speech detection on Twitter,* 5 pages

[5] Burnap, P., and Williams, M. L. 2015. *Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making,* 10 pages

[6] Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; and Chang, Y. 2016. *Abusive language detection in online user content,* 9 pages

[7] Walker, S. 1994. *Hate Speech: The History of an American Controversy. U of Nebraska Press*