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, 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. It gets more complicated with 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.

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.

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:

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…)
2. Tweets with specific words were almost always classified as hate speech, regardless of context, which was mostly agreed with.
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.
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 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.

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 and great reads.

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*