Hate speech on Twitter and other online platforms

Meera Whitson

1 Introduction

As the internet becomes more and more ubiquitous every day, new avenues for hateful language appear. The drastic increase in the number of users has made the need to monitor hateful content on online platforms exceed the capabilities of the companies that operate them. In this analysis, I explore how linguistic features characteristic of hate speech appear in Tweets and examine if previous linguistic findings for longer texts of hate speech are consistent with online posts, which are much shorter in length.

The main linguistic source and inspiration for this study is the work of Alexandria Marsters (2019) on linguistic threat assessment of hate speech. This study will use her comprehensive definition of hate speech, which is as follows:

> "a problematic expressive speech act which conveys derogatory sentiment towards a person or persons based on the perceived possession of a socially defined group characteristic, which is made to the detriment of the target, and is addressed to the ingroup with the intention of inciting animus and/or violence or is addressed to the target with the effect of instilling a fear of violence or harm" (Marsters, 2019, : 65--66).

Other research consulted focused on the applications of computer algorithms to automate hate speech detection. Due to considerable obstacles in distinguishing hateful Tweets from harmless Tweets, data in this paper is manually selected. Marsters's methodologies are then applied in order to evaluate their relevance since the medium of hate speech being examined here is different from that of Marsters's original work.

Overall, hate speech detection, particularly automatic hate speech detection, is found to be an incredibly challenging task that requires a lot of computer science research. The fact that Marsters' findings do appear to be consistent with hate speech on Twitter suggests that future natural language processing (NLP) research into automating hate speech detection should be informed by forensic linguistic research. Online hate speech has been shown to have concrete consequences, including ethnic cleansing in Myanmar and a siege on the United States Capitol, so being able to identify and monitor it is essential to prevent unnecessary violence.

2 Literature Review

Detection of hate speech online has been the focus of a lot of natural language processing research, specifically the work of Zeerak Waseem, who has been studying hate speech for over five years and has published numerous papers. In his work with Dirk Hovy, Waseem trained various language models on annotated data, sixteen thousand tweets that were manually labeled with tags including "racism", "sexism", and "none". They then trained multiple models and evaluated their performances against each other, concluding that using character n-grams of lengths up to 4 produce the best results (Waseem & Hovy, 2016 : 91). Using character n-grams refers to determining probabilities of a potential next character given the n-1 previous characters, so a bigram would depend on one preceding character, a trigram would depend on the two preceding characters, and so on.

In this paper, Waseem emphasizes that there is an observable connection between hate speech and hate crimes and describes how sites such as Twitter and Facebook have been looking for ways to actively combat certain forms of hate speech on their platforms (Waseem & Hovy, 2016 : 88). Here, the authors also mention the inconsistency of manual labels due to the subjectivity of what truly constitutes hate speech, which is built upon in Waseem's paper on annotator influence. The authors also point out that there is very limited NLP research on hate speech due to "the lack of a general definition of hate speech, an analysis of its demographic influences, and an investigation of the most effective features" (Waseem & Hovy, 2016 : 88).

This paper also includes a link to the dataset of tweet IDs and labels (Waseem & Hovy, 2016 : 89). Upon examining the dataset which contains sixteen thousand annotated tweets, I have found that it does not directly include the contacts of the tweets, just corresponding ID numbers, so their model must use the IDs to access the API (application product interface – the way for developers to retrieve data from Twitter's database) and then correlate the retrieved text of the tweet to the labels, including "racism", "sexism", and "none". They make the distinction between racism and sexism likely because these two subtypes of hate speech will not always contain similar language features, so it is easier for the model to view them as two independent categories instead of simply trying to binarily classify hate speech and not hate speech. Waseem and Hovy explain the importance of their research by citing research which claims that "hate speech is a precursor to hate crime" (Waseem & Hovy, 2016 : 90), and that proactive measures are necessary to prevent more tragedies resulting from hateful beliefs. Catching concerning posts before they turn into action can save lives.

In another source by Waseem, he examines the effects of different people annotating data to train models. The way automated identifiers work is they study large datasets of Tweets labeled as "hateful" or "not hateful", but where these labels come from matters. Whoever labeled the data may not be using the same definition from Marsters shown above, which can result in making different decisions about what does and does not count as hate speech. Any NLP model will pick up on these distinctions and repeat assumptions or patterns in the dataset. This paper discusses the issue of inconsistency in labeling hate speech depending on who is annotating it. Most labeled datasets are manually labeled by a student on a research team, and models are trained based on those "true" labels. However, as discussed before, "hate speech" is hard to define, therefore it is challenging to definitively say whether something qualifies as hate speech, especially when individual pieces of data are single sentences, and whether or not something is hateful can be very context dependent. Additionally, different people will label things differently, so Waseem examined the "influence of annotator knowledge" on classification by comparing "expert and amateur annotations" (Waseem, 2016 : 138).

For the expert annotations, Waseem selected "feminist and anti-racism activists" to annotate the data set (Waseem, 2016 : 139). Waseem compared the labels assigned by these activists and the labels assigned by members of the general population ("amateurs") and used both of these as the true labels to train language models to detect hate speech. The paper includes a lot of statistical analysis, including the observation that the main cause of error is false positives in both amateur and expert annotations (Waseem, 2016 : 141). This is significant because even advanced models created by researchers who have spent years on this topic still struggle with flagging Tweets that are not hate speech and should not actually be flagged. If manually assigned labels are supposed to be "true" hate speech, they should be consistently

assigned, and they're not, so creating a reliable computational model is quite challenging. People do not always agree with each other on what is or is not hate speech, so even an ideal computational model will not always be agreed with by everyone.

Another source for this study is the work of Gambäck and Isaksen (2020), a computer science paper which focuses on specific architectures of machine learning models. Gambäck and Isaksen analyzed the success of different types of language models at identifying hate speech online. One important observation they make is that hate speech does not always include inherently offensive words like slurs, and offensive language does not always constitute hate speech (Gambäck & Isaksen, 2020 : 16). On social media in particular, offensive words and swears are used casually and frequently, often without any malice. This mode of communication simply lends itself to foul language more so than other mediums, so searching for hate speech by specifically filtering for sensitive words like slurs will not only miss a considerable quantity of more veiled, polished hate speech, it will also falsely flag posts that are actually harmless. According to Gambäck and Isaksen, systems which can track and automatically detect "abusive language" are of great importance to companies which need to monitor content, and a lot of capital is invested into these technologies. The inability to distinguish between "offensive" language and legitimate hate speech is a common issue in these tools (Gambäck & Isaksen, 2020 : 16).

One last NLP paper discussing automating online hate speech detection is the work of Saleh, Alhothali, and Moria (2020) which explores more advanced machine learning models. Saleh, Alhothali, and Moria argue that the pervasiveness of social media has made it "an essential element of our society by which people communicate and exchange information on a daily basis" (Saleh, Alhothali, & Moria, 2020 : 2). Thus, not studying social media would be ignoring a medium through which a huge portion of speech events occur. Additionally, use of social media offers a level of anonymity which allows users to feel more comfortable spreading harmful messaging. The authors focus on white supremacist hate speech specifically, as it is easily observable in hateful content online. The authors point out studies have shown "links between hate speech and hate crimes against others" (Saleh, Alhothali, & Moria, 2020 : 3). Thus, tracking hate speech via linguistic markers ought to be used to predict and prevent hate crimes and violent events. The authors do not go into more detail here about what they consider "linguistic markers or signs," but they do make some important observations about the speech of white supremacist extremists. According to the authors, white supremacists often use rhetoric, specialized vocabularies, abbreviations, and coded words to convey their stance while avoiding "being detected by traditional detection methods" (Saleh, Alhothali, & Moria, 2020 : 4). This implies that these extremists will generally not use slurs or profanity, but instead will speak in an apparently civil manner which is hateful in the meaning. The authors go into a little more detail about the types of speech expected from white supremacists, like claims that other races are undermining them or that white people are the victim in modern society (Saleh, Alhothali, & Moria, 2020 : 5). This paper concludes that "content of tweets is a good indicator for hateful accounts," which confirms that looking through the language used in posts is still worthwhile for identifying hate speech online (Saleh, Alhothali, & Moria, 2020 : 11), even though authors may use coded language, as mentioned before . In their own data analysis, they describe how they collected their own corpus of white supremacist posts using hashtags like #white_privilege and #it_is_ok_to_be_white. Using hashtags could be a useful idea for hate speech detection, which will later be shown to be a significant hurdle for this kind of research. From the corpus that the

authors constructed, they identified most used terms. These terms are notably not slurs or inherently offensive, but their presence can be an indicator of white supremacist ideologies. The authors provided a word cloud of these most used terms, which is shown in Figure 1.

Figure 1: Word cloud of most used terms of white supremacist corpus (Saleh, Alhothali, & Moria, 2020 : 14)

My main linguistic resource and inspiration for this work is Alexandria Marsters's work. Her research informs two important elements of this paper. First, she has a thorough discussion of the definition of hate speech and dedicates a chapter to analyzing competing definitions and ultimately constructs her own encompassing definition. This is useful because to combat hate speech, there needs to be an agreed upon definition of what exactly is being combatted. Second, she identifies and explains linguistic features that identify hate speech and classify it as either "Hunter" (author is violent) and "Howler" (author incites others to violence). Both classes are incredibly harmful and need to be addressed, but being able to identify Hunter tendencies can allow appropriate authorities to identify individuals that may pose an immediate threat to the wellbeing and safety of others.

Most of the other sources discussed here emphasize the computer science and NLP side of this topic, but this paper is a true linguistics paper. Marsters studies hate speech from a sociolinguistic perspective using discourse analysis. She emphasizes that, while hate speech is widely talked about, there is no clear agreed upon definition. Through combining previous linguistics research, legal definitions, and other interdisciplinary perspectives, she develops a working definition of hate speech off of which she bases her research. Marsters dedicates an entire chapter (57 pages) to defining hate speech. There are a lot of competing factors that seem to affect whether or not a given speech act constitutes hate speech, and the need to balance them all poses a considerable challenge for creating a single definition. Hate speech is a speech act which "declares feelings of hatred" and "adopts an attitude" with some notion of harm towards a targeted group (Marsters, 2019 : 64--65). The definition is tricky because it relies on words like "insult", "harm", and "hate" which all have their own connotations and can be interpreted differently. Marsters's final, complete definition is restated again here for convenience:

> "a problematic expressive speech act which conveys derogatory sentiment towards a person or persons based on the perceived possession of a socially defined group characteristic, which is made to the detriment of the target, and is addressed to the ingroup with the intention of inciting animus and/or violence or is addressed to the target with the effect of instilling a fear of violence or harm" (Marsters, 2019, : 65--66).

There is a lot of nuance to this definition, which Marsters explains throughout the chapter, delving into what is meant by harm and the idea that hate speech has an "ingroup addressee" and "outgroup addressee" (Marsters, 2019 : 67).

Additionally, Marsters discussed a contrast of two genres of hate speech, referred to as "Hunters" and "Howlers" (Marsters, 2019 : 71). This distinction is made between the latter, who "simply engage in hateful rhetoric", and the former, who exhibit the potential to act violently (Marsters, 2019 : 71--72). This is a notable distinction because if resources are needed to track down potentially violent individuals, those should be allocated to Hunters instead of Howlers to address more urgent threats. Marsters cites Tammy Gales's research on stance in the context of realized versus unrealized threats.

Marsters also identifies many linguistic features of hate speech, which will inform the methodology of this study. If patterns she observes in her data, which mainly concerns longer instances of hate speech like articles or manifestos, are also apparent in much briefer speech instances such as Tweets, then her findings could be applicable to improving hate speech detection models described by the other sources reviewed in this paper.

3 Data and Methodology

Because this is a linguistics focused study, I am following the methodology of a linguistics paper, the work by Marsters, in this case. Unlike Waseem and Hovy, who focus on determining what architectures of language models are most useful for automated hate speech detection, Marsters focuses on identifying language patterns in hate speech which distinguish speakers who may eventually act on their words from those who are all talk (Marsters, 2019 : 3). For Marsters's research, she used two corpora of texts that were divided into speech by people who did act violently against the targets of their hate speech, and speech by people who did not act violently. These corpora include fewer, larger pieces of text than any Tweet dataset (as Tweets are limited to 140 characters).

Marsters's methodology is to look for linguistic features and manually identify patterns within each corpus. Due to the size of both Waseem & Hovy's dataset and the dataset I originally compiled using flagged words and scraping directly from the Twitter API, this seems unrealistic to apply on this scale. The scraped dataset is close to 100,000 Tweets, many of which are definitely false positives and not actually hateful, and Waseem & Hovy's dataset is 16,000 Tweets. Neither of these would be particularly easy to manually sift through and try to apply that methodology to everything.

Instead, hateful Tweets will be identified manually by going through the Twitter feeds of inflammatory public figures. Once these Tweets are collected, Marsters's findings will be applied to look for similar patterns in the Tweets. Marsters specifically focuses on the way "the

authors position themselves in the texts, how both authors use rhetorical questions as a positioning tool, the positioning of the author's ingroup in the role of victim, and how the authors demonstrate commitment to their hateful claims through evidentiality and epistemics (Marsters, 2019 : 13).

Trying to combine these two methodologies is challenging because Waseem and Hovy use advanced language models that are unrealistic for this project, and the data that Marsters uses is inherently different from Tweets. She uses long articles and manifestos, such as a four-page statement by Dylann Roof, the shooter in the 2015 Charleston church attack, and an article published by The Aryan Alternative (Marsters, 2019 : 263--269). Thus, it is unclear how transferrable work on this format will be to Tweets, which are obviously much shorter.

4 Analysis

This study applies Marsters' methodology to data by looking for indicators of epistemic stance, because "epistemic stance suggests an author's certainty and commitment to the propositions s/he expresses (Marsters, 2019 : 153). Additionally, Marsters discusses positioning, "an act by which an interlocutor situates him/herself along an affective or epistemic scale and through which s/he claims or is assigned responsibilities of stancetaking" (Marsters, 2019 : 153).

Marsters identifies the following features as indicative of the "Hunter" class (meaning hate speech by individuals who did act violently at some point): "first person singular pronouns, justification of violence via predictive modals, and expressions of inevitability, to position himself with a storyline of a personal journey to enlightenment whereby he has gained knowledge, and to present himself as the only person who can act as a reluctant but inevitable savior to address the issues he mentions," along with "inductive modes of knowing based on sensory and inferential evidence and rhetorical questions" (Marsters, 2019 : 154). In contrast, a member of the "Howler" class (those who do not actually act on their hateful ideologies) is identified by "first person plural pronouns, other plural referents, and constructed dialogue, to position himself as the spokesperson for a larger group who share his ideology" (Marsters, 2019 : 154). According to Marsters, Howlers will also use rhetorical questions (so this is not an important distinctive feature, and "explicit and repeated constructions in which an outgroup is the Agent of violence against an ingroup to position his ingroup as the victims in a storyline in which the outgroup are the violent victimizers" (Marsters, 2019 : 154). Marsters summarizes the distinction between Howlers and Hunters as follows: "the Howler projects a more detached and less emotionally involved stance than the Hunter while he seeks to incite the reader to violence by exploiting strategies which have been noted in 'dangerous speech' associated with incitement to genocide" (Marsters, 2019 : 154).

Using Marsters's work as a basis, I am looking for the following features in Tweets that could be considered hateful: whether first person pronouns are singular or plural, predictive modals, expressions of inevitability, and how the author positions himself. Tweet (1), shown below, was collected from the previously mentioned archive of President Trump's tweets.

(1) "States want to correct their votes, which they now know were based on irregularities and fraud, plus corrupt process never received legislative approval. All Mike Pence has to do is send them back to the States, AND WE WIN. Do

it Mike, this is a time for extreme courage!" @realDonaldTrump

This may not necessarily be classified as hate speech on its own as it arguably does not contain anything explicitly derogatory towards a certain group. However, it contains multiple features discussed by Marsters. First, the author, Trump, clearly positions himself as a victim by decrying "irregularities and fraud". He also uses first person plural pronouns in "WE WIN", a feature that is characteristic of the "Howler" group. This is consistent with general knowledge about Trump's relationship with his supporters and the events leading up to the January 6th, 2021 mob at the United States Capitol. Trump's account was eventually suspended due to accusations that he incited violence, which is the definition of Howlers that Marsters provides.

Tucker Carlson, the author of Tweet (2), is a conservative cable television personality who has been applauded by many white supremacists as beneficial to their movement and spreading their message to a wider audience. This Tweet, describing the riots and violence in Kenosha, Wisconsin in August 2020 following the shooting of Jacob Blake by a police officer.

(2) "Kenosha devolved into anarchy because the authorities abandoned the people. Those in charge, from the governor on down, refused to enforce the law. They've stood back and watched Kenosha burn. Are we really surprised that looting and arson accelerated to murder?" @TuckerCarlson

In this Tweet, Carlson uses a rhetorical question, something Marsters identified as common among instances of hate speech. Additionally, like Trump in Tweet (1), he uses a first person plural pronoun ("are we really surprised…"), indicating Howler tendencies. He also makes his stance clear through words like "anarchy" and "looting", both of which have clear negative connotations and frame protestors as the inciters of violence. His final sentence, "Are we really surprised that looting and arson accelerated to murder?" is particularly interesting because it implies that the protestors, the people he alleges are looters and arsonists, committed a murder, when the deaths that actually occurred in Kenosha were caused by a police officer (which led to the protests), and then a vigilante who believed he was assisting law enforcement. Carlson's positioning here is misleading and perpetuates a division between his ingroup (law enforcement and "law abiding" citizens) and the outgroup (those protesting the shooting of Jacob Blake). Additionally, this sentence contains an expression of inevitability, another feature that Marsters identified as characteristic of hate speech. She attributed this feature to the Hunter class. Tweet (3) is also written by Tucker Carlson.

(3) "The larger lesson of the past 2 yrs is the left will not abide losing power, even temporarily. For liberals, political power is personal power. W/out it, they're exposed & terrified. Some become vicious. They believe they're meant to run this country, our government & our culture" @TuckerCarlson

This is a clear example of in-group out-group positioning, as shown through the use of pronouns "they" and "our". Carlson positions some undefined (but understood to be white, conservative Americans) "us" as a victim of the controlling "liberals", an antagonistic out-group.

Carlson also uses the predictive modal "will", another feature identified by Marsters as a common characteristic of hateful speech. Through this modal, Carlson expresses the inevitability of some action because "the left" will never concede some perceived "power".

Tweet (4) is authored by Kaitlin Bennett, a right-wing public figure. She initially gained recognition for her outspoken stance on gun rights, and now has a considerable presence on social media, particularly Twitter.

> (4) "*black people caught on camera killing Asians at an alarming rate*
>
> Big brain liberals: 'Actually even that's white people's fault.'
>
> If you want to see what real racism looks like, look no further." @KaitMarieox

In this Tweet, Bennett clearly conveys her position by the disparaging reference to "Big brain liberals". Additionally, her post very blatantly frames black people as violent by asserting that there exists evidence of "an alarming rate" of black people killing Asians. Marsters's work does not explicitly discusses whether false statistics should be construed as hate speech, but I certainly think this "conveys a derogatory sentiment" (Marsters, 2019 : 65--66).

Tweet (5) was posted by Todd Starnes, a right-wing conservative author and radio host, in response to Black Lives Matter protests in Brooklyn Center, Minnesota.

> (5) "Praying for the small business owners in Brooklyn Center tonight as they guard their stores and prepare to defend their property from the pillaging and plundering mob." @toddstarnes

Because (5) does not appear to incite fear or violence, it could be argued that it does not fit all the criteria Marsters's definition of hate speech and thus does not qualify as hateful. However, framing the out-group, the protestors (a group with a higher proportion of black and minority individuals than the presumed in-group), as a "pillaging and plundering mob" very blatantly attributes violence to the out-group. Language such as this is not profane, but Starnes dehumanizes protestors and positions himself as a bystander that only has the best interests of small business owners in mind. (5) does not contain Marsters's defined linguistic features of hate speech other than framing, so I will argue that it does not constitute hate speech, but still exhibits some hateful questions which invites the question of whether an ideal hate speech detection model should flag Tweets such as this. Like discussed by Waseem and other sources, disagreement between human evaluators over what does and does not constitute hate speech further complicates designing a successful hate speech detection algorithm. Another example of a Tweet that could be argued to be hate speech or not hate speech is Tweet (6), authored by Tomi Lahren, a political commentator and television personality.

> (6) "How many tragedies could be prevented if [resisting] arrest didn't become glorified martyrdom?" @TomiLahren

Tweet (6) is a rhetorical question, a sentence type that Marsters classified as frequent in hate speech. Like Starnes in (5), Lahren is vilifying the Black Lives Matter movement. She claims that at least some of the black Americans unjustly killed by police would have never been hurt had they not "resisted arrest," and implies that victims of police violence deliberately try to be martyrs. I consider this assertion incredibly offensive and believe that this Tweet qualifies as hateful based on Marsters's definition. However, the context of police brutality is needed to understand the true meaning of the Tweet, and automatic hate speech detection models cannot take context into account. Thus, this Tweet would likely not be flagged as the subject matter is difficult to discern without any contextual information because the only word even referencing police is "arrest", and nothing explicitly mentions Black Lives Matter. This illustrates that even the best possible algorithm may miss information necessary to identify hateful language, so it will likely take years and much more research to develop an effective solution for hate speech detection.

As explained in each of the above examples, Marsters's findings do seem to generally hold up when compared to Tweets, a different medium of speech than what she studied. Thus, using linguistic research to inform hate speech detection seems like it could have useful applications, especially because current models focus on patterns of vocabulary and not epistemic stance and positioning, which are more significant markers of hate.

5 Discussions and Conclusion

The goal of this paper is to discuss occurrences of hate speech online, particularly on Twitter, identify methods for identifying problematic posts, and apply linguistic research to classify example Tweets into the Hunter and Howler groups described by Marsters. The findings of this study are as follows. First, identifying hate speech is very challenging, especially using algorithms and automated systems. Originally, an objective of this study was to collect useful data by seeking out certain terms, but that yielded a lot of false positives (Tweets which are definitely not actually hate speech). Certainly, there were some true positives there, but not enough to justify that method of data collection. For the Tweets that were selected by hand through looking up individual profiles, there were many instances of patterns described in Marsters's research. Because bias is a factor when manually selecting data, I tried to avoid bias by simply looking for Tweets that could be considered hateful and not look for patterns such as rhetorical questions. Generally, it seems that my findings conform with Marsters's work, especially because her methods suggest that Donald Trump and Tucker Carlson fall into the Howler category, which seems accurate given their roles in society. This could suggest that these Tweets do indeed qualify as hate speech. In all of the Tweets studied, at least one of Marsters's linguistic features of hate speech is observed.

Research on hate speech detection is prone to limitations, as was discussed in the literature review. NLP researchers have been working on this problem for years, and algorithms currently in use by Twitter, Facebook, and other media platforms are still works in progress. Additionally, even if linguistic research could be applied, how to turn findings such as Marsters's into instructions for an algorithm is still unclear. A model can learn to recognize rhetorical questions, modals, and expressions of inevitability, but markers of stance are inarguably more difficult for a computer to understand, especially when contextual information is missing, as illustrated by Tweet (6). Additionally, both linguistic markers and frequently appearing terms like those shown in Figure 1 may be characteristic of hate speech, but they also appear frequently

in normal speech, so both are imperfect features for distinguishing between hate and not hate. Though Marsters did identify rhetorical questions as common in hateful language, it is misguided to assume that the use of a rhetorical question is indicative of hate speech as rhetorical questions are a normal persuasive rhetorical device. Thus, features such as these should be examined in conjunction with existing methods and manual review of flagged posts.

Despite all the challenges, computer mediated communication is increasingly becoming the default mode of communication, thus it should be one of the main focuses of current linguistic research. The need for this research only grows as time passes, especially because online speech can have very real, very dramatic effects on people's lives. In order to protect users, online platforms need to be aware of hate speech and have mechanisms in place to handle it. Linguistic research can help identify the most problematic posts before they reach a wide audience and spread extremist, hateful ideologies.

As discussed by Waseem & Hovy, an area of further research would be developing better language models that can catch hate speech, because content is being posted on the internet far faster than humans can monitor it. This is outside the scope of this project but would be interesting to study further. Other areas of further research could be investigating other platforms to investigate hate speech in computer mediated communication as a whole. The world is moving more and more online, and the linguistic community ought to keep up with that change.

6 References

Brown, B. (2021). Trump twitter archive. Retrieved April 01, 2021, from https://www.thetrumparchive.com/faq

Gambäck, B., and Isaksen, V. (2020). Using transfer-based language models to detect hateful and offensive language online. *Proceedings of the Fourth Workshop on Online Abuse and Harms* :16--27

Marsters, A. (2019). When hate speech leads to hateful actions: a corpus and discourse analytic approach to linguistic threat assessment of hate speech. *DigitalGeorgetown*. Retrieved 1 April 2021, from https://repository.library.georgetown.edu/bitstream/handle/10822/1056009/Marsters_georgetown_0076D_14371.pdf?sequence=1&isAllowed=y.

Saleh, H., Alhothali, A., and Moria, K. (2020). Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT. *ArXiv*. Retrieved 1 Apr 2021 from https://arxiv.org/pdf/2010.00357.pdf.

Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. *Proceedings of the First Workshop on NLP and Computational Social Science,* 138--142

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. *Proceedings of the NAACL Student Research Workshop:* 88—93