Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter

Zeerak Waseem
University of Copenhagen
Copenhagen, Denmark
csp265@alumni.ku.dk

Dirk Hovy
University of Copenhagen
Copenhagen, Denmark
dirk.hovy@hum.ku.dk

Abstract

Hate speech in the form of racist and sexist remarks are a common occurrence on social media. For that reason, many social media services address the problem of identifying hate speech, but the definition of hate speech varies markedly and is largely a manual effort (BBC, 2015; Lomas, 2015).

We provide a list of criteria founded in critical race theory, and use them to annotate a publicly available corpus of more than 16k tweets. We analyze the impact of various extra-linguistic features in conjunction with character n-grams for hate-speech detection. We also present a dictionary based the most indicative words in our data.

1 Introduction

Hate speech is an unfortunately common occurrence on the Internet (Eadicicco, 2014; Kettrey and Laster, 2014) and in some cases culminates in severe threats to individuals. Social media sites therefore face the problem of identifying and censoring problematic posts (Moulson, 2016) while weighing the right to freedom of speech.

The importance of detecting and moderating hate speech is evident from the strong connection between hate speech and actual hate crimes (Watch, 2014). Early identification of users promoting hate speech could enable outreach programs that attempt to prevent an escalation from speech to action.

Sites such as Twitter and Facebook have been seeking to actively combat hate speech (Lomas, 2015). Most recently, Facebook announced that they would seek to combat racism and xenophobia aimed at refugees (Moulson, 2016). Currently, much of this moderation requires manual review of questionable documents, which not only limits how much a human annotator can be reviewed, but also introduces subjective notions of what constitutes hate speech. A reaction to the "Black Lives Matter" movement, a campaign to highlight the devaluation of lives of African-American citizens sparked by extrajudicial killings of black men and women (Matter, 2012), at the Facebook campus shows how individual biases manifest in evaluating hate speech (Wong, 2016).

In spite of these reasons, NLP research on hate speech has been very limited, primarily 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.

While online hate speech is a growing phenomenon (Sood et al., 2012a), its distribution is not uniform across all demographics. Neither is the awareness of what constitutes hate speech (Ma, 2015). Considering that hate speech is not evenly distributed in the United States of America (Zook, 2012) and perpetrators of hate speech should be a small minority from a limited demographic group. Including available demographic information as features should thus help identification accuracy.

Our contribution We provide a data set of 16k tweets annotated for hate speech. We also investigate which of the features we use provide the best identification performance. We analyze the features that improve detection of hate speech in our corpus, and find that despite presumed differences in the geographic and word-length distribution, they have little to no positive effect on performance, and rarely improve over character-level features. The exception to this rule is gender.

2 Data

Our data set consists of tweets collected over the course of 2 months. In total, we retrieved 136,052

Proceedings of NAACL-HLT 2016, pages 88–93,
San Diego, California, June 12-17, 2016.

tweets and annotated 16,914 tweets, 3,383 of that for sexist content sent by 613 users, 1,972 for racist content sent by 9 users, and 11,559 for neither sexist or racist and is sent by 614 users.

Since hate speech is a real, but limited phenomenon, we do not balance the data, to provide as realistic a data set as possible.

Our data set will be made available as tweet IDs and labels at Github[1].

Corpus collection We bootstrapped our corpus collection, by performing an initial manual search of common slurs and terms used pertaining to religious, sexual, gender, and ethnic minorities. In the results, we identified frequently occurring terms in tweets that contain hate speech and references to specific entities, such as the term "#MKR", the hashtag for the Australian TV show *My Kitchen Rules*, which often prompts sexist tweets directed at the female participants[2]. In addition, we identified a small number of prolific users from these searches.

Based on this sample, we used the public Twitter search API to collect the entire corpus, filtering for tweets not written in English. This particular corpus construction ensures that we obtain non-offensive tweets that contain both clearly offensive words and potentially offensive words, but remain non-offensive in their use and treatment of the words. For example, even though "muslims" is one of the most frequent words in racist tweets, it also occurs in perfectly innocuous tweets, such as "`you are right there are issues but banning Muslims from entering doesn't solve anything.`"

We manually annotated our data set, after which we had the help of an outside annotator (a 25 year old woman studying gender studies and a non-activist feminist) to review our annotations, in order to mitigate annotator bias introduced by any parties.

Identification and annotation While it is easy to identify racist and sexist slurs, hate speech is often expressed without any such terms. Furthermore, it is not trivial for humans to identify hate speech due to differences of exposure to and knowledge of hate speech. Similarly to identifying privileges, a critical thought process is required to identify hate speech (McIntosh, 2003; DeAngelis, 2009). In order to reliably identify hate speech, we need a clear decision list to ensure that problematic tweets are identified.

We propose the following list to identify hate speech. The criteria are partially derived by negating the privileges observed in McIntosh (2003), where they occur as ways to highlight importance, ensure an audience, and ensure safety for white people, and partially derived from applying common sense.

A tweet is offensive if it

1. uses a sexist or racial slur.
2. attacks a minority.
3. seeks to silence a minority.
4. criticizes a minority (without a well founded argument).
5. promotes, but does not directly use, hate speech or violent crime.
6. criticizes a minority and uses a straw man argument.
7. blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims.
8. shows support of problematic hash tags. E.g. "#BanIslam", "#whoriental", "#whitegenocide"
9. negatively stereotypes a minority.
10. defends xenophobia or sexism.
11. contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria.

As McIntosh (2003) highlights the way that they are privileged by being white. Many of these observations underline apparent safety and visibility granted by skin color. As such, our list highlights ways in which minorities are undercut and silenced as these occur as methods of oppression of minorities (DeAngelis, 2009).

While most of the criteria are easily identified, others such as identifying problematic hash tags is far more unclear. We define problematic hash tags as terms which fulfill the remaining one or several of other criteria.

Annotator agreement The inter-annotator agreement is $\kappa = 0.84$. 85% of all disagreements occur in annotations of sexism, with 98% of all reviewer changes being set to neither sexist nor

[1] http://github.com/zeerakw/hatespeech

[2] All terms queried for: "MKR", "asian drive", "feminazi", "immigrant", "nigger", "sjw", "WomenAgainstFeminism", "blameonenotall", "islam terrorism", "notallmen", "victimcard", "victim card", "arab terror", "gamergate", "jsil", "racecard", "race card"

racist, the remaining set to racist. In most of these cases we find that the disagreement is reliant on context or the lack thereof. Where our outside annotator would tend to annotate such cases lacking apparent context as not being sexist, we preferred to annotate as sexist for many of these cases. For instance, our outside annotator did not find "There just horrible #lemontarts #MKR" to be a case of sexist language whereas we had annotated it as such. Another common case of disagreement was the difference of opinion in what constitutes sexism. Where we found tweets such as ""Everyone else, despite our commentary, has fought hard too. It's not just you, Kat. #mkr"" to be singling out a single woman, our annotator found that such a comment was not coined on the gender but in fact an (assumed) expression hard work from the competitor.

3 Demographic distribution

Twitter does not directly provide fields for demographic information beyond location, so we collect this information by proxy. We extract gender by looking up names in the users profile text, the name, or the user name provided and compare them to known male and female names (Kantrowitz, 1994) as well as other indicators of gender, such as pronouns, honorifics, and gender specific nouns.

We find that the gender distributions in our hate speech are heavily skewed towards men (see Table 1).

	All	Racism	Sexism	Neither
Men	50.08%	33.33%	50.24%	50.92%
Women	02.26%	0.00 %	02.28%	01.74%
Unidentified	47.64%	66.66%	47.47%	47.32%

Table 1: Distribution of genders in hate-speech documents.

While men are over represented in our data set for all categories, the majority of users cannot be identified with our method, which heavily impairs use of gender information as features. For instance, in our racist subset, we only identify 3 out of 9, all of them men. Furthermore, (Roberts et al., 2013) find that 75% and 87% of perpetrators of hate crimes against African Caribbeans and Asians respectively, were men. Considering that hate speech is a precursor to hate crime (Watch, 2014), we find it unsurprising that such a large part of the perpetrators of hate speech in our data set are men.

And while we manage to identify 52.56% of the users in our annotated database, we find that the vast majority are users associated with sexist tweets and tweets that do not contain hate speech. Given that both have nearly the same distribution (see Table 1), we do not expect this feature to yield a substantial increase in F1 score.

4 Lexical distribution

We normalize the data by removing stop words, with the exception of "not", special markers such as "RT" (Retweet) and screen names, and punctuation.

We construct the ten most frequently occurring words by selecting the ten words with the most frequent occurrence for each class. We find that the terms frequently occurring in each class differ significantly (see Table 2). The most frequent tokens for racism are necessary in order to discuss Islam, while discussing women's issues does not require the use of most of the terms that occur most frequently.

We also see a sampling effect of the data set, as many of the tweets flagged as sexist are generated by viewers of *My Kitchen Rules*. Similarly, and more obviously, many of the tweets flagged as racist pertain to Judaism and Islam.

Lengths Drawing inspiration from Tulkens et al. (2015), we add average and total the lengths of the tweets and the lengths of the user descriptions. We expect lengths to discriminate between tweets that contain hate speech and those that do not (see Table 3).

5 Geographic distribution

We find that using location as a feature negatively impacts the F1-score attained. In order to identify the geographical origin of a tweet, we need to consider more than just the tags Twitter provides, given that only 2% of Twitter users disclose their location (Abbas, 2015).

We therefore identify whether any location or their proxy is given in the tweet or user meta data (name given and user name). In each of these fields we extract markers indicating geographical location or time zone. Time zone is also used as a proxy for location by (Gouws et al., 2011).

If a time zone or location is identified, we map it to longitude and latitude and add to the set of tweets originating from that time zone. If a location name, such as "Sydney" is given, it is also used as a feature for classification.

Sexism	Distribution	Racism	Distribution
not	1.83%	islam	1.44%
sexist	1.68%	muslims	1.01%
#mkr	1.57%	muslim	0.65%
women	0.83%	not	0.53%
kat	0.57%	mohammed	0.52%
girls	0.48%	religion	0.40%
like	0.42%	isis	0.38%
call	0.36%	jews	0.37%
#notsexist	0.36%	prophet	0.36%
female	0.34%	#islam	0.35%

Table 2: Distribution of ten most frequently occurring terms

	Racism	Sexism	None
Mean	60.47	52.93	47.95
Std.	17.44	21.16	23.43
Min.	11.00	2.00	2.00
Max.	115.00	118.00	129.00

Table 3: Overview of lengths in characters, subtracting spaces.

6 Evaluation

We evaluate the influence of different features on prediction in a classification task. We use a logistic regression classifier and 10-fold cross validation to test the influence of various features on prediction performance, and to quantify their expressiveness.

Model Selection In order to pick the most suitable features, we perform a grid search over all possible feature set combinations, finding that using character n-grams outperforms using word n-grams by at least 5 F1-points (60.42 vs. 69.86) using similar features. For that reason, we do not consider word n-grams.

To determine whether a difference between two feature sets is statistically significant (at $p < 0.05$), we run a bootstrap sampling test on the predictions of the two systems. The test takes 10,000 samples and compares whether the better system is the same as the better system on the entire data set. The resulting (p-) value of the bootstrap test is thus the fraction of samples where the winner differs from the entire data set, giving the p-value a very intuitive interpretation.

Results We find that using character n-grams of lengths up to 4, along with gender as an additional feature provides the best results. We further find that using location or length is detrimental to our scores. By using our n-gram features we achieve the results shown in Table 4.

We find that across our features only adding gender information improves our F1-score. All other features and feature combinations are detrimental to the performance of the system. We find that gender, the only additional feature that provides an improvement, is not statistically significant, whereas the addition of location as well as gender is significant, at p = 0.0355.

Features We collect unigrams, bigrams, trigrams, and fourgrams for each tweet and the user description. To assess the informativeness of the features we sum the model coefficients for each feature over the 10 folds of cross validation. This allows for a more robust estimate.

We find that the most influential features for the logistic regression (see Table 5) largely correspond with the most frequent terms in Table 2. We see, for instance different n-gram lengths of the word "Islam" and "sexist".

Intuitively, it makes sense that not only will the most frequent terms be indicative, but also that character n-grams would outperform word n-grams, due to character n-gram matrices being far less sparse than the word n-gram matrices.

One of the notable differences between the n-grams for our two categories is the occurrence of a gender-based slur, and normal words pertaining to women. On the other hand, all of the racist features are n-grams of normal terms, which are re-appropriated for building a negative discourse. One such example is: "`@BYRONFBERRY Good. Time to confront the cult of hatred and murder #Islam`".

	char n-grams	+gender	+gender +loc	word n-grams
F1	73.89	73.93	73.62*	64.58
Precision	72.87%	72.93%	72.58%	64.39%
Recall	77.75%	77.74%	77.43%	71.93%

Table 4: F1 achieved by using different features sets.

Feature (sexism)	Feature (racism)
'xist'	'sl'
'sexi'	'sla'
'ka'	'slam'
'sex'	'isla'
'kat'	'l'
'exis'	'a'
'xis'	'isl'
'exi'	'lam'
'xi'	'i'
'bitc'	'e'
'ist'	'mu'
'bit'	's'
'itch'	'am'
'itc'	'm'
'fem'	'la'
'ex'	'is'
'bi'	'slim'
'irl'	'musl'
'wom'	'usli'
'girl'	'lim'

Table 5: Most indicative character n-gram features for hate-speech detection

Gender (*F1* 73.89) We train our model on character bi- to fourgrams and the gender information for each use obtained as described in section 3. We find that this combination yields the highest score (see Table 4), though the score only increases slightly.

Length (*F1* 73.66) This feature set contains the total of each tweet and description and average lengths of the words occurring along with the n-grams of lengths 1 to 4.

Gender + location (*F1* 73.62) In this feature set contains the locations obtained in 5 along with our 1 to 4-grams, and the gender for each user. Adding locations occurs to be slightly detrimental to the performance of the classifier.

Gender + location + length (*F1* 73.47) For completeness we train on gender, geographic information, and length features along with 1 to 4-grams. Our score decreases by the use of all features, as we expected given the results of using location in combination with gender, and length.

7 Related Work

Most related work focused on detecting profanity, using list-based methods to identify offensive words (Sood et al., 2012b; Chen et al., 2012a). While studies suggest that these are good, robust ways to identify abusive language (Sood et al., 2012b); this approach is limited by its reliance on lists. Chen et al. (2012b) addresses this by the use of character n-grams among other features, in order to identify various forms of bullying.

Sood et al. (2012b) extend their system from static lists to incorporating edit distances to find variants of slurs. This allows for finding a better recall, but does not address the core issue of detecting offensive sentences, which do not use terms that occur in the list. Chen et al. (2012a) address this by using lexical and syntactical features along with automatically generated black lists.

Warner and Hirschberg (2012) perform a similar task of detecting hate speech using a support vector machine classifier, trained on word n-grams, brown clusters, and "the occurrence of words in a 10 word window" (Warner and Hirschberg, 2012). They find that their best model produces unigrams as most indicative features, and obtains an F1 score of 63, which is similar to the F1 score we achieve using word n-grams.

8 Conclusion

We presented a list of criteria based in critical race theory to identify racist and sexist slurs. These can be used to gather more data and address the problem of a small, but highly prolific number of hateful users. While the problem is far from solved, we find that using a character n-gram based approach provides a solid foundation. Demographic information, apart from gender, brings little improvement, but this could be due to the lack of coverage. We plan to improve location and gender classification to update future data and experiments.

References

Diana Abbas. 2015. What's in a location. https://www.youtube.com/watch?v=GNlDO9Lt8J8, October. Talk at Twitter Flight 2015. Seen on Jan 17th 2016.

BBC. 2015. Facebook, google and twitter agree german hate speech deal. http://www.bbc.com/news/world-europe-35105003. Accessed on 26/11/2016.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012a. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*, pages 71–80. IEEE, September.

Yunfei Chen, Lanbo Zhang, Aaron Michelony, and Yi Zhang. 2012b. 4is of social bully filtering: Identity, inference, influence, and intervention. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2677–2679, New York, NY, USA. ACM.

Tori DeAngelis. 2009. Unmasking 'racial micro aggressions'. *Monitor on Psychology*, 40(2):42.

Lisa Eadicicco. 2014. This female game developer was harassed so severely on twitter she had to leave her home. http://www.businessinsider.com/brianna-wu-harassed-twitter-2014-10?IR=T, Oct. Seen on Jan. 25th, 2016.

Stephan Gouws, Donald Metzler, Congxing Cai, and Eduard Hovy. 2011. Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 20–29, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mark Kantrowitz. 1994. Name corpus: List of male, female, and pet names. http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/names/0.html. Last accessed on 29th February 2016.

Heather Hensman Kettrey and Whitney Nicole Laster. 2014. Staking territory in the world white web: An exploration of the roles of overt and color-blind racism in maintaining racial boundaries on a popular web site. *Social Currents*, 1(3):257–274.

Natasha Lomas. 2015. Facebook, google, twitter commit to hate speech action in germany. http://techcrunch.com/2015/12/16/germany-fights-hate-speech-on-social-media/, Dec. Seen on 23rd Jan. 2016.

Alexandra Ma. 2015. Global survey finds nordic countries have the most feminists. http://www.huffingtonpost.com/entry/global-gender-equality-study-yougov_us_564604cce4b045bf3deeb96d, November. Seen on Jan 19th.

Black Lives Matter. 2012. Guiding principles. http://blacklivesmatter.com/guiding-principles/. Accessed on 26/11/2016.

Peggy McIntosh, 2003. *Understanding prejudice and discrimination.*, chapter White privilege: Unpacking the invisible knapsack, pages 191–196. McGraw-Hill.

Geir Moulson. 2016. Zuckerberg in germany: No place for hate speech on facebook. http://abcnews.go.com/Technology/wireStory/zuckerberg-place-hate-speech-facebook-37217309. Accessed 10/03/2016.

Colin Roberts, Martin Innes, Matthew Williams, Jasmin Tregidga, and David Gadd. 2013. Understanding who commits hate crime and why they do it.

Sara Sood, Judd Antin, and Elizabeth Churchill. 2012a. Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1481–1490. ACM.

Sara Owsley Sood, Judd Antin, and Elizabeth F. Churchill. 2012b. Using crowdsourcing to improve profanity detection. In *AAAI Spring Symposium: Wisdom of the Crowd*, volume SS-12-06 of *AAAI Technical Report*. AAAI.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2015. Detecting racism in dutch social media posts, 2015/12/18.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 19–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hate Speech Watch. 2014. Hate crimes: Consequences of hate speech. http://www.nohatespeechmovement.org/hate-speech-watch/focus/consequences-of-hate-speech, June. Seen on on 23rd Jan. 2016.

Julia Carrie Wong. 2016. Mark Zuckerberg tells Facebook staff to stop defacing Black Lives Matter slogans. http://www.theguardian.com/technology/2016/feb/25/mark-zuckerberg-facebook-defacing-black-lives-matter-signs. Accessed on 10/03/2016.

Matthew Zook. 2012. Mapping racist tweets in response to president obama's re-election. http://www.floatingsheep.org/2012/11/mapping-racist-tweets-in-response-to.html. Accessed on 11/03/2016.