HATE SPEECH DETECTION

BY YASHI JAIN

ABSTRACT

A key challenge for automatic hate-speech detection on social media is the separation of hate speech from other instances of offensive language. Lexical detection methods tend to have low precision because they classify all messages containing particular terms as hate speech and previous work using supervised learning has failed to distinguish between the two categories. We used a crowd-sourced hate speech lexicon to collect tweets containing hate speech keywords. We use crowd-sourcing to label a sample of these tweets into three categories: those containing hate speech, only offensive language, and those with neither. We train a multi-class classifier to distinguish between these different categories. Close analysis of the predictions and the errors shows when we can reliably separate hate speech from other offensive language and when this differentiation is more difficult. We find that racist and homophobic tweets are more likely to be classified as hate speech but that sexist tweets are generally classified as offensive. Tweets without explicit hate keywords are also more difficult to classify.

INTRODUCTION

What constitutes hate speech and when does it differ from offensive language? No formal definition exists but there is a consensus that it is speech that targets disadvantaged social groups in a manner that is potentially harmful to them (Jacobs and Potter 2000; Walker 1994). In the United States, hate speech is protected under the free speech provisions of the First Amendment, but it has been extensively debated in the legal sphere and with regards to speech codes on college campuses. In many countries, including the United Kingdom, Canada, and France, there are laws prohibiting hate speech, which tends to be defined as speech that targets minority groups in a way that could promote violence or social disorder. People convicted of using hate speech can often face large fines and even imprisonment. These laws extend to the internet and social media, leading many sites to create their own provisions against hate speech.

In extreme cases this may also be language that threatens or incites violence, but limiting our definition only to such cases would exclude a large proportion of hate speech. Importantly, our definition does not include all instances of offensive language because people often use terms that are highly offensive to certain groups but in a qualitatively different manner. For example some African Americans often use the term n*gga2 in everyday language online (Warner and Hirschberg 2012), people use terms like h*e and b*tch when quoting rap lyrics, and teenagers use homophobic slurs like f*g as they play video games. Such language is prevalent on social media (Wang et al. 2014), making this boundary condition crucial for any usable hate speech detection system . Previous work on hate speech detection has identified this problem but many studies still tend to conflate hate speech and offensive language. In this paper we label tweets into three categories: hate speech, offensive language, or neither. We train a model to differentiate between these categories and then analyze the results in order to better understand how we can distinguish between them. Our results show that fine-grained labels can help in the task of hate speech detection and highlights some of the key challenges to accurate classification. We conclude that future work must better account for context and the heterogeneity in hate speech usage.

OBJECTIVE

A key flaw in much previous work is that offensive language is mislabeled as hate speech due to an overly broad definition. Our multi-class framework allows us to minimize these errors; only 5% of our true offensive language was labeled as hate. The tweets correctly labeled as offensive tend to contain curse words and often sexist language, e.g. Why you worried bout that other h*e? Cuz that other h*e aint worried bout another h*e and I knew Kendrick Lamar was onto something when he said “I call a b*tch a b*tch, a h*e a h*e, a woman a woman”. Many of these tweets contain sexist terms like b*tch, p*ssy, and h*e. Human coders appear to consider racists or homophobic terms to be hateful but consider words that are sexist and derogatory towards women to be only offensive, consistent prior findings (Waseem and Hovy 2016). Looking at the tweets misclassified as hate speech we see that many contain multiple slurs, e.g. @SmogBaby: These h*es be lyin to all of us n*ggas and My n*gga mister meaner just hope back in the b*tch. While these tweets contain terms that can be considered racist and sexist it is apparent than many Twitter users use this type of language in their everyday communications. When they do contain racist language they tend to contain the term n*gga rather than n*gger, in line with the findings of Kwok and Wang (2013). We also found a few recurring phrases such as these h*es ain’t loyal that were actually lyrics from rap songs that users were quoting. Classification of such tweets as hate speech leads us to overestimate the prevalence of the phenomenon. While our model still misclassifies some offensive language as hate speech we are able to avoid the vast majority of these errors by differentiating between the two.

METHODOLOGY

MODEL USED

We first use a logistic regression with L1 regularization to reduce the dimensionality of the data. We then test a variety of models that have been used in prior work: logistic regression, na¨ıve Bayes, decision trees, random forests, and linear SVMs. We tested each model using 5-fold cross validation, holding out 10% of the sample for evaluation to help prevent over-fitting. After using a grid-search to iterate over the models and parameters we find that the Logistic Regression and Linear SVM tended to perform significantly better than other models. We decided to use a logistic regression with L2 regularization for the final model as it more readily allows us to examine the predicted probabilities of class membership and has performed well in previous papers (Burnap and Williams 2015; Waseem and Hovy 2016). We trained the final model using the entire dataset and used it to predict the label for each tweet. We use a one-versus-rest framework where a separate classifier is trained for each class and the class label with the highest predicted probability across all classifiers is assigned to each tweet. All modeling was performing using scikit-learn (Pedregosa and others 2011).

DATA USED

We begin with a hate speech lexicon containing words and phrases identified by internet users as hate speech, compiled by Hatebase.org. Using the Twitter API we searched for tweets containing terms from the lexicon, resulting in a sample of tweets from 33,458 Twitter users. We extracted the time-line for each user, resulting in a set of 85.4 million tweets. From this corpus we then took a random sample of 25k tweets containing terms from the lexicon and had them manually coded by CrowdFlower (CF) workers. Workers were asked to label each tweet as one of three categories: hate speech, offensive but not hate speech, or neither offensive nor hate speech. They were provided with our definition along with a paragraph explaining it in further detail. Users were asked to think not just about the words appearing in a given tweet but about the context in which they were used. They were instructed that the presence of a particular word, however offensive, did not necessarily indicate a tweet is hate speech. Each tweet was coded by three or more people. The intercoder-agreement score provided by CF is 92%. We use the majority decision for each tweet to assign a label. Some tweets were not assigned labels as there was no majority class. This results in a sample of 24,802 labeled tweets. Only 5% of tweets were coded as hate speech by the majority of coders and only 1.3% were coded unanimously, demonstrating the imprecision of the Hatebase lexicon.

CODE

C:\Users\yashi\Downloads\Hate speech detection on twitter data_live.py

CONCLUSION

If we conflate hate speech and offensive language then we erroneously consider many people to be hate speakers (errors in the lower triangle of Figure 1) and fail differentiate between commonplace offensive language and serious hate speech (errors in the upper triangle of Figure 1). Given the legal and moral implications of hate speech it is important that we are able to accurately distinguish between the two. Lexical methods are effective ways to identify potentially offensive terms but are inaccurate at identifying hate speech; only a small percentage of tweets flagged by the Hatebase lexicon were considered hate speech by human coders.4 While automated classification methods can achieve relatively high accuracy at differentiating between these different classes, close analysis of the results shows that the presence or absence of particular offensive or hateful terms can both help and hinder accurate classification. Consistent with previous work, we find that certain terms are particularly useful for distinguishing between hate speech and offensive language. While f*g, b*tch, and n*gga are used in both hate speech and offensive language, the terms f*ggot and n*gger are generally associated with hate speech. Many of the tweets considered most hateful contain multiple racial and homophobic slurs. While this allows us to easily identify some of the more egregious instances of hate speech it means that we are more likely to misclassify hate speech if it doesn't contain any curse words or offensive terms. To more accurately classify such cases we should find sources of training data that are hateful without necessarily using particular keywords or offensive language.

REFERENCES

Bird, S.; Loper, E.; and Klein, E. 2009. Natural Language Processing with Python. O'Reilly Media Inc. 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. Policy & Internet 7(2):223– 242. Djuric, N.; Zhou, J.; Morris, R.; Grbovic, M.; Radosavljevic, V.; and Bhamidipati, N. 2015. Hate speech detection with comment embeddings. In WWW, 29–30. Gitari, N. D.; Zuping, Z.; Damien, H.; and Long, J. 2015. A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering 10:215–230. Hutto, C. J., and Gilbert, E. 2014. VADER: A parsimonious rulebased model for sentiment analysis of social media text. In ICWSM. Jacobs, J. B., and Potter, K. 2000. Hate crimes: Criminal Law and Identity Politics. Oxford University Press. Kwok, I., and Wang, Y. 2013. Locate the hate: Detecting tweets against blacks. In AAAI. Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; and Chang, Y. 2016. Abusive language detection in online user content. In WWW, 145–153. Pedregosa, F., et al. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12:2825–2830. Silva, L. A.; Mondal, M.; Correa, D.; Benevenuto, F.; and Weber, I. 2016. Analyzing the targets of hate in online social media. In ICWSM, 687–690. Walker, S. 1994. Hate Speech: The History of an American Controversy. U of Nebraska Press. Wang, W.; Chen, L.; Thirunarayan, K.; and Sheth, A. P. 2014. Cursing in english on twitter. In CSCW, 415–425. Warner, W., and Hirschberg, J. 2012. Detecting hate speech on the world wide web. In LSM, 19–26. Waseem, Z., and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In SRW@HLT-NAACL, 88–93. Waseem, Z. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In Proceedings of the First Workshop on NLP and CSS, 138–142.