A Survey of Non-Neural Machine Learning Techniques in Hate Speech Classification

Noah R. Carver

Abstract

Social media censorship drives current research on natural language processing of hate speech, focusing it merely on detecting bigotry accurately, occasionally providing a "badness" score. While this metric may be useful in tracking the spread of harmful ideas, further metrics may supply further insights that could spark more influential systemic and cultural change. This project aims to move away from that censorship and content moderation and toward analysis and understanding of the use of damaging and hateful language. As such, I utilized five non-neural techniques from baseline papers on hate speech detection in order to solve the joint task of both recognizing *and* classifying hate speech. While most performed decently when classifying definite hate speech, none were able to properly classify non-hate as such. However, this is likely due to the inaccuracy of the database used.

Introduction

In the past few years, the problem of hateful and offensive speech has exploded in the public eye. Aided by the anonymity of social media, bigoted elements of our society seem emboldened to express and spread their bigotry. In attempts to counter this, Social Media Companies have begun to incorporate content moderation in their services. However, the push is merely to detect and hide hate speech, not to analyze it. Much has been written in social sciences fields regarding this (Noble and Tynes 2016), but there is little technical research to enable potential solutions, only discriminators to serve in automatic censorship machines.

Related Work

There are a variety of methods used when dealing with Hate speech in short text. For simplicity's sake, this project focuses only on classical methods, as defined by (Zhang, Robinson, and Tepper 2018) as Methods that require manual designing of feature encoders. The most common classical methods in literature are:

1. Linear Support Vector Machine
2. Logistic Regression
3. Naive Bayes
4. Decision Tree
5. Random Forest

(Davidson et al. 2017; MacAvaney et al. 2019; Zhang, Robinson, and Tepper 2018)

Defining Hate Speech

Hate Speech is a tricky thing to define. It seems each organization, research group, even every database annotator has a different definition of what is hateful, what is merely offensive and what constitutes acceptable speech. Hate Speech is often conflated with offensive, abusive or profane language, each of which may exist in tandem with hate speech, but are separate concepts.(Zhang, Robinson, and Tepper 2018) For example, although they are correlated the usage of slurs does not necessarily imply hate speech, nor does all hate speech include slurs. Generally, Hateful speech is characterized by:

1. Expression of intent or desire to harm, incite harm, or spread hatred.
2. Targeting of groups based on Characteristic (ie: racial or ethnic groups, religious groups, disabled and neurodivergent, Women, non-cis-heteronormative folk, etc)

(MacAvaney et al. 2019) Note that by this definition, hate speech is held separately from bigotry, as *"abolish white people"* would technically qualify as hate speech, despite white people not being at any disadvantage as a group whatsoever.(DiAngelo 2018)

It follows from the general definition of Hate speech that any system capable of concretely identifying hate speech should trivially be able to determine

Approach

Dataset

The dataset used was collected by (Ousidhoum et al. 2019) and annotated under their instruction using Amazon's Mechanical Turk. It contains a corpus of 5646 tweets annotated as *Abusive, Hateful, Offensive, Disrespectful, Fearful* or *Normal* as well as a separate classification over whether the tweet discriminates by *Origin, Gender, Sexual Orientation, Religion, Disability* or *Other*. There are a number of additional labels applied to each tweet, but those proved either irrelevant or too inconsistent to be useful.

This classification scheme was simplified into *Inoffensive, Origin, Gender, Sexual Orientation, Religion, Disability* and *Other*, as it makes little sense to classify inoffensive text by the group it offends.

the 5646 tweets in (Ousidhoum et al. 2019) were split randomly into 3 groups using `sklearn.model_selection.train_test_split` 56% of the data (3161 tweets) was allocated to training, 14% (791 tweets) to development and 30% (1694 tweets) for the final test data.

Stop-words (ie: *the, is, at, which...*) and superfluous, twitter-specific, artefacts (ie: #ff,USR or rt) and excessive white space were removed.

Data was vectorized using tf-idf over word 1-grams, 2-grams and 3-grams, with a max vector size of 12000. This size was chosen via heuristic, as there are 3161 tweets in the training set. It was originally intended to use a more sophisticated vectorization scheme, such as google's `word-2-vec`, but that proved unwieldy over so little data.

Models

Linear Support Vector Machines, Logistic Regression, Naive Bayes, Decision Trees and Random Forests, were selected to learn the dataset as mentioned earlier. These five were chosen due to their prevalence in literature as well as their relative ease of implementation. Some of these algorithms do not natively handle multi-class classification, so, When applicable, a one-versus-all multi-classing scheme was used. Algorithms were implemented using the `scikit-learn` library. Code for this project can be found at https://github.com/ncarver1/HateSpeechClassification

Experimental results

Results are presented as precision, recall and F1 scores calculated with weighted averages in table 1, as well as a multi-class confusion matrix for each classifier in figure 1 and an accuracy by class score. Precision, Recall and F1 scores give a decent understanding of the general efficacy of each each method, Accuracy by class is meant to compare the inbuilt bias of each method, and Confusion Matrices provide pairwise intersectional breakdown of error.

	Precision	Recall	F1
Linear SVM	0.543	0.444	0.411
Logistic Regression	0.572	0.539	0.553
Naive Bayes	0.374	0.382	0.357
Decision Tree	0.596	0.556	0.573
Random Forest	0.708	0.610	0.651

Table 1: Classifier Scores

Discussion

It is clear from Table 1 that these models did not perform very well. This stands in contrast to their performance in literature. In fact, the highest performing algorithm was Random Forests, an algorithm barely mentioned in baseline literature, and even that didn't perform too well, with only a 61% Recall. By Class (table 2), we can see that algorithms performed best in identifying speech that is harmful to people based on Gender and Religion. Furthermore, No model was effective in separating Inoffensive speech from offensive speech. The plurality of inoffensive text was mischaracterized as targeting by *origin*, which is consistent with the experiences of many black Anti-racist activists (Sankin 2017).

Dataset

Every data point in (Ousidhoum et al. 2019) was labeled by multiple anonymous crowd sourced annotators with no definite training in socio-linguistics.

The need for a comprehensive hate speech dataset that accurately represents public understanding of what is and is not acceptable speech *and why it is or is not acceptable* has been thoroughly documented in literature. This means that *all* available labeled datasets on this topic have a large amount of noise.

Furthermore, due to funding, each datapoint was only ever labeled by three annotators. This clearly cannot reflect full societal understanding, and will obviously lead to lack of representation for certain minorities.

Critical Theory

Critical theory examines the concepts of race, gender, class, disability, etc as *systems of power* and posits that all efforts considering those power structures should aid the underprivileged. This stands in contradiction to the most common definitions of Hate Speech.(MacAvaney et al. 2019), as abusive text that targets an empowered group would not require intervention. This is only confirmed by measures in literature, In a 2019 paper analyzing Racial bias in commonly used datasets, (Davidson, Bhattacharya, and Weber 2019) found significant racial bias against black people in every single dataset they tested. An attempt to mitigate this was made by (Sap et al. 2019) using what they call *dialect* and *race priming*. Essentially, annotators are asked to consider the dialect (as measured using a linguistic tool) or a presumed race. Each tweet (Sap et al. 2019) analysed was annotated by at least five annotators, however, there was significantly more agreement among annotators when compared to (Ousidhoum et al. 2019). When used to train a model, that same linguistic tool was used to add an additional "dialect" feature to the dataset that served as a proxy for the race of the speaker.

Intersectionality

Discrimination of all sorts is generally **intersectional**. This means that the discrimination faced by those who lie at the intersections of disadvantages can take on additional characteristics from any of those individual disadvantages alone. Put simply: the whole is greater than the sum of it's parts.(Crenshaw 1990)

Unfortunately, the only database available that even bothered to classify by offended party did not take this into account, a decision that likely contributed a large amount of error. How should an annotator or classifier label hate directed

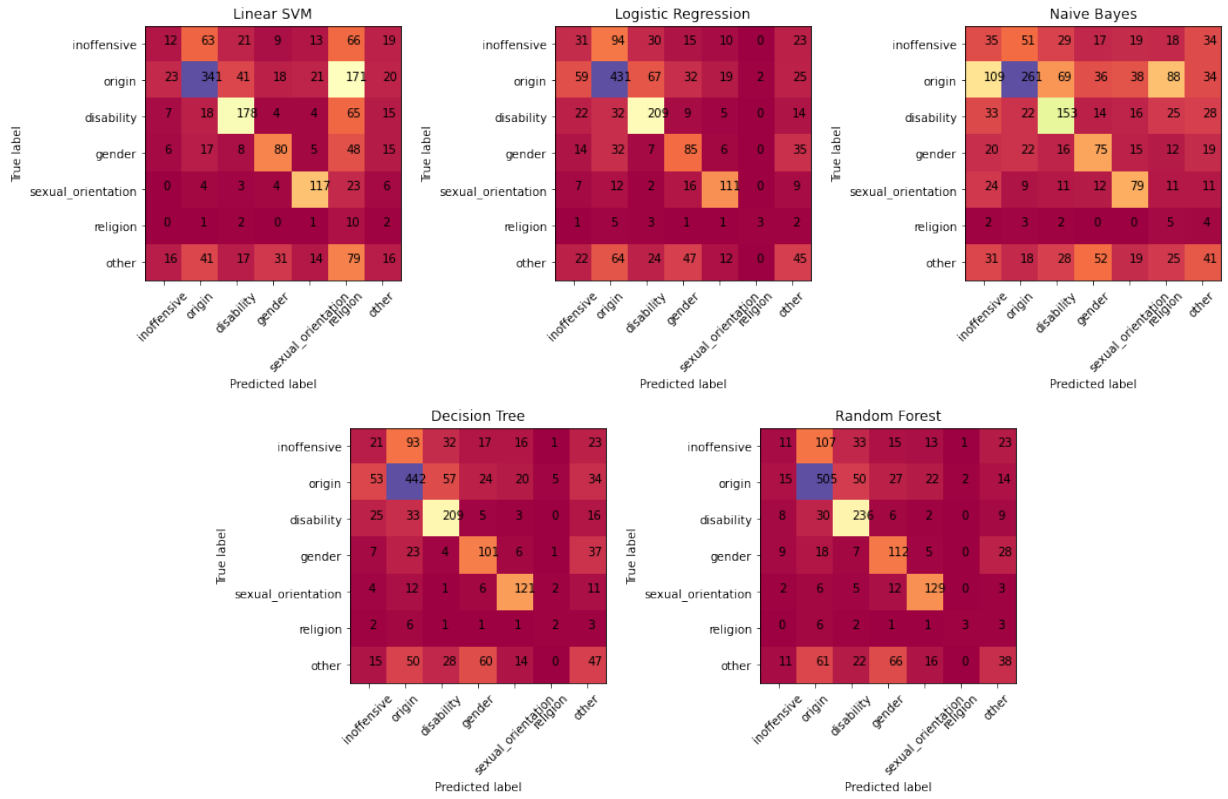

Figure 1: Confusion Matrices

	Inofffensive	Origin	Gender	Sexual Orientation	Religion	Disability	Other
Linear SVM	0.059	0.537	0.611	0.447	0.745	0.625	0.074
Logistic Regression	0.152	0.678	0.718	0.474	0.707	0.1875	0.210
Naive Bayes	0.172	0.411	0.526	0.419	0.503	0.313	0.192
Decision Tree	0.103	0.696	0.718	0.564	0.771	0.125	0.220
Random Forest	0.054	0.795	0.811	0.626	0.822	0.188	0.178

Table 2: Accuracy by Class

toward Black Women, *origin* or *gender*? Furthermore, Judaism manifests somewhere between ethnicity and religion in the harmful stereotype that all Jews are pro-Israel, . Ideally a classifier would use a multi-class multi-output system, but the data to facilitate such a classifier is not available right now.

Returning to (Sap et al. 2019), it is clear that this technique, while promising, cannot account for other disadvantaged groups, nor for the intersections of those groups. Furthermore, the few, seemingly oxymoronic, Black Racists, such as Candace Owens, who holds that systemic racism is a myth, would be a potential blind-spot for their technique.

Conclusion and Future Work

From these experiments, it is reasonable to question the validity of these techniques

It would be interesting to collect a corpus of tweets, along with some direct demographic information about the author of said tweet, and annotate them intersectionally, either by use of significantly more Mechanical Turks than either (Ousidhoum et al. 2019) or (Sap et al. 2019), or through a council of anti-discrimination experts.

Future Work: a Design Fiction

Imagine, if you will, a tool used, not to censor, but to seek out and identify potential hateful speech.

This tool would be open source and iterative, using a platform like Github to manage versions and continuously self improve. It would be free to use, on the one condition that it is never used to censor. It would provide accurate and precise probabilities that a piece of text constituted Bigoted or hateful speech, what group it though the text offended, and would be able to back-trace its internal logic in order to explain its findings. Such a tool would likely be ”NLP Complete” - capable of understanding text at a human level of comprehension - in order to make sense of context and lex-

ical distinctions (ie: "the nazi's organization was great" (the nazis were adept at organizing) versus "the nazi organization was great").

This tool would allow activist groups to strike at the heart of the issue, finding and confronting bigotry wherever it was. It could be deployed at large scale, using web crawlers to root out even the deepest of deep-net forums.

Such a tool *could* be the end of overt racism as we know it. Paradoxically, Not by stifling bigotry, but by amplifying it.

References

Crenshaw, K. 1990. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.* 43: 1241.

Davidson, T.; Bhattacharya, D.; and Weber, I. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516* .

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009* .

DiAngelo, R. 2018. *White fragility: Why it's so hard for white people to talk about racism*. Beacon Press.

MacAvaney, S.; Yao, H.-R.; Yang, E.; Russell, K.; Goharian, N.; and Frieder, O. 2019. Hate speech detection: Challenges and solutions. *PloS one* 14(8): e0221152.

Noble, S. U.; and Tynes, B. M. 2016. *The intersectional internet: Race, sex, class, and culture online*. Peter Lang International Academic Publishers.

Ousidhoum, N.; Lin, Z.; Zhang, H.; Song, Y.; and Yeung, D.-Y. 2019. Multilingual and Multi-Aspect Hate Speech Analysis. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Sankin, A. 2017. How activists of color lose battles against Facebook's moderator army .

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678.

Zhang, Z.; Robinson, D.; and Tepper, J. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, 745–760. Springer.