Insight Report - Stream A

Nebojsa Ajdarevic

Queensland University of Technology

Student number: N10434348

Course Code: IFN619

Summary

Hateful text is becoming more and more prominent in our online interactions. Companies like Twitter and Facebook, whose business models rely on pleasant online experiences may be at risk of being left behind due to their negative user experience. Twitter data was used to determine how hateful text can be policed. Hashtags were extracted to allow us to target commonly negative subject areas, this can be used to reduce the amount of data that my machine learning algorithms have to go through. This is especially useful due to the immense amount of text data that is posted onto twitter daily. After implementation the most negatively used hashtags should be updated regularly to ensure currency in the targeted hashtags. Count Vectorizers were created for the test and train datasets and various machine learning techniques were run. Due to the lack of computing power the machine learning algorithms produced shouldn't be the one that is used in final implementation, but they do serve as an example of what can be done. With more features in the count vectorizer the machine learning techniques would prove very accurate and allow Twitter to identify the problematic users. There are many ethical considerations when banning users from platforms and many issues that may arise from implementing automation with the power to ban users proactively. As a result, the action recommended for Twitter executives is that users with multiple instances of specifically targeting a range of other users should be dealt with bans. Thus reducing false positives while effectively cleaning up the confrontational behaviour seen on the site.

How can hateful text be policed more effectively?

Due to polarising figures like Donald Trump having more influence on social media, especially on twitter in this case, hate and hate filled comments seem more and more commonplace on the platform. Furthermore, many have put the blame onto social media companies like Facebook and Twitter. The rise of negativity on platforms can cause users to move to platforms less likely to upset them. Social media companies have been known to rise and fall overnight with the likes of Myspace being a prominent example. Detecting the users who constantly hurl abuse would prove to be a large challenge for any company. Furthermore, reporting functions still require a lot of manpower to ensure the targeted users are justly reported and then banned. Therefore, creating a machine learning algorithm that can proactively keep the negative users in check will prove an essential business advantage.

Description of the Data Analytics Process

The data was collected from Kaggle where a test and train dataset were already established. Due to this fact it is difficult to judge the accuracy and quality of the data. Sentiments were already determined in the train dataset to test the accuracy of my algorithm. The dataset had mostly been pre-processed already with no missing values and all usernames replaced with "@user". Hashtags were extracted as they offer an easy way to see the topics of the tweets. Positive hashtags and words were visualised separately from negative hashtags and words used for context. Count vectorizers were created for the test and train datasets. However, due to the limited computing power available for the count vectorizers, the machine learning models produced lacked accuracy in predicting the negative tweets. With more computing power these models would be useful for identifying positive and negative tweets, thus allowing Twitter to sanction the common perpetrators accordingly.

It should be noted that how the uploader of the dataset determined the sentiment in the train dataset is unknown. This may pose a threat to the validity of any models that use this data. Furthermore, what is and isn't hateful text is hard to determine, even for humans.

Insight Gained

Hashtags were extracted for the negative and positive sentimental tweets. There seems to be a clear difference in topic for positive and negative hashtags. Most of the negative hashtags focus on political or religious topics. This aligns with observations made in the word clouds. Users posting with these hashtags could potentially be targeted with advertisements or twitter campaigns that promote having an open dialogue with people whom we disagree with to help quell the hate Twitter is often known to contain on its website. Furthermore, users who commonly spread hateful text can be banned accordingly. This would obviously have a big impact on the Twitter user experience. Targeting specific users would obviously have an impact on them with the content they see(or if they are allowed to use Twitter at all), however, this may also have an impact on the content they post. Thus, making the twitter browsing experience a more positive one for all users involved. The most popular negative hashtags can be viewed in figure 1 below.

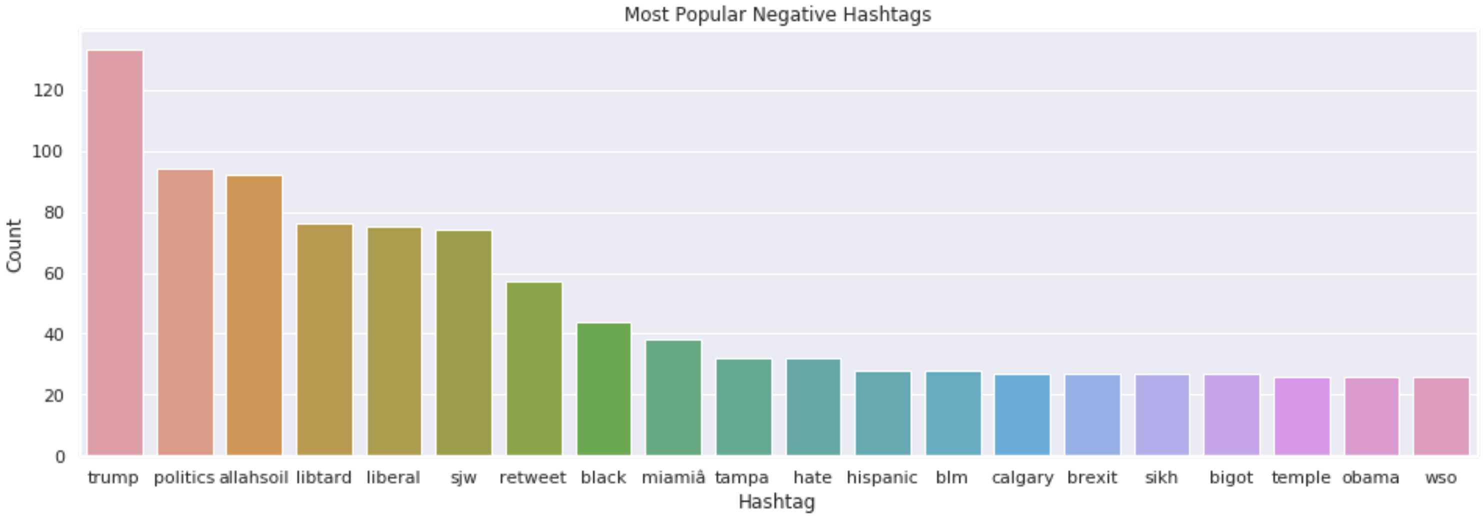

Figure 1

Additionally, a model was constructed to test the sentiment of tweets. The model had a 90% accuracy score. Due to the accuracy here, this model would misclassify 10% of tweets. This may not seem like a lot, but with the potential of censoring people, and the sheer number of tweets tweeted everyday there could be serious repercussions for twitter here. If people feel they are being unfairly censored they may boycott the site. Furthermore, as many of the topics are either religion or politics they could feel their opinions and core beliefs are being unjustly censored. This could lead to mass boycotts of the site. All this should be considered along with the fact that only 53% of hateful tweets (the actual target of this algorithm) would be correctly identified. A new model should be developed with more computing power and more features for our Count Vectorizer. Once this model has been tested Twitter must consider how impactful negative tweets are to their website. Then they must decide how much collateral they would be willing to accept depending on the models misclassification rate.

Ethical Considerations

There are many ethical considerations that come up when the censoring of people on social media is concerned. Is it the same for someone to call out a Nazi on twitter and for someone to slur racial abuse to an unsuspecting bystander? How would an algorithm differentiate someone going on the attack or someone simply biting back? In fact, should these things even be considered when it comes to policing hate speech? In the public hate is often met with hate. People victimised by hate could also be affected by this algorithm if they fight fire with fire. Just one high profile case of this could lead to a PR nightmare suggesting that twitter may be trying to silence minorities, or people at risk. Additionally, due to the opacity of the machine learning algorithm used here, we have no real way of telling why the algorithm classifies each tweet as hateful or not. The effects of opaque algorithms need to be considered appropriately before they are implemented. Furthermore, if a model of this like is to be used it should be incredibly accurate and staff response to misclassification should be quick and efficient.

How the banning system works needs to be refined. Users should not be banned without being given a warning that lets them know that their tweets have been flagged as inciting or spreading hatred on multiple occasions. Once they have been banned users should be shown which tweets were flagged so that they can report false positives of any arise. This would also (hopefully) offer the banned users a chance for insight into the way they have been acting allowing them time to change.

In a broader context, is banning people from social media be something that we as a society consider acceptable. This text classification could be applied to any variety of Social media companies that have text as part of their platform, including; Facebook, Instagram, LinkedIn and others. With the increasing prominence these companies play in our everyday lives, how would someone be affected if they were banned from all of these. Would they be able to find work in society? Social media companies have as much a role to play in enabling free speech as countries do. Silencing people that we disagree with is a slippery slope to removing our freedom of speech as human beings, especially if governments start to get involved.

Governments are potential stakeholders in this situation as well. They may start to mandate what can and can't be said on social media sites that have users within their country. China is an example of this. Political parties could also be considered stakeholders in the situation as they may suggest policing speech in this way is unconstitutional and impersonal. This is especially true as they are the main topic for the negative tweets.

Bias

Machine learning algorithms encode bias and errors dominant in the training dataset. due to the high skew of non-hateful tweets in the training dataset the ML model is biased to assume tweets are non-hateful. This explains its lack of effectiveness in correctly identifying hateful tweets. Furthermore, due to the training dataset the machine learning algorithm is likely to assume the "negative words" in the wordcloud are bad to use despite context, or it may be that it fails to identify hateful tweets about new topics with new contexts to make any real impact on the Twitter browsing experience.

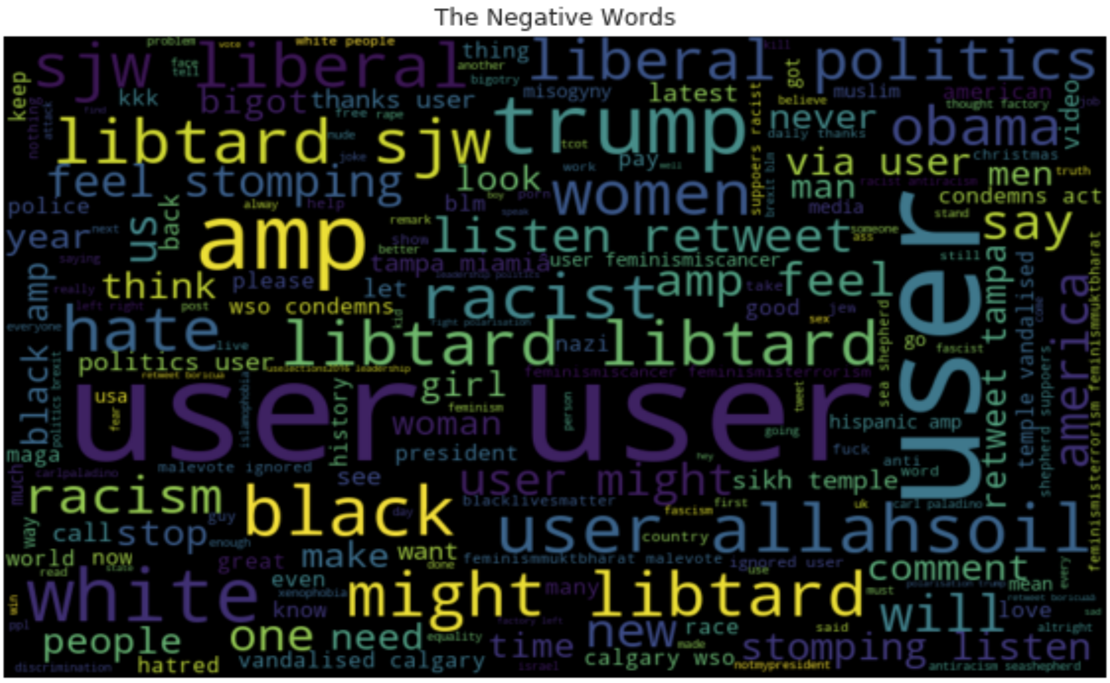

Figure 2 Wordcloud of commonly used words in negatively labelled tweets

Conclusion

It is clear that the divisiveness of politics and religion are not specific to Twitter, or any online platform. This is a problem ingrained much deeper into society. With hate fuelled crimes seemingly becoming the norm more and more. We as a big player in how people interact online owe it to society to do something about it. The vocal minority play a large role in inciting hatred between religions or political parties which in turn leads to groupthink and ingroup vs outgroup mentalities (Brewer, 1999). Decreasing the amount of people inciting groupthink has the potential to have great benefits in both societies at large and on Twitter itself. Making it is clear that something needs to be done to clean up twitter and make it a more pleasant experience for those involved. However, banning users from your platform could cause backlash amongst the active user base, especially if something goes wrong with the algorithm and the innocent are being affected. Therefore, this algorithm should only target users with multiple instances of specifically targeting a range of other users. This would avoid the false positives while still allowing for users who actively use twitter to express their opinions, even if they have negative connotation, to do so. Despite the fact that nothing may solve the negativity we see an experience both offline and online, this would serve an important role in cleaning up the visceral hate on the website without impacting the majority of the user base in any negative way, making browsing Twitter a more pleasant experience for all those involved.

Referneces

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate?. Journal of social issues, 55(3), 429-444.