Twitter Abuse Detection

"English Football announces social media boycott" Premier League (April 2021)

#TheProblem

Twitter is a vastly popular social media site seeing constant use by users globally. With an expected 353 million monthly users and a barrage of nearly 6000 tweets every second Smith (2020). With such a large volumes of data being created it has become a rich environment for academics to perform research. Ranging from Sentiment analysis to trend predictions etc. This project looks to tackle the very imprudent issue of abuse.

Abuse has become a significant issue within the realm of social media. With large cooperation's like the English football federation and Cricket clubs coming out recently and declaring a total social media blackout. In order to raise awareness of the abuse suffered by players.

Solving abuse has become an astronomical task with many issues surrounding the execution of removing it from the platform. Twitter currently has a very high reliance on the likes of manual reporting with the aid of public users. With senior director of twitter, Gasca (2019) mentioning how only 38% of abusive content is flagged without the aid of public reporting. This paper looks to discuss and experiment on methods of creating an automatic detection abuse method. It does so by looking at two of the popular methodologies explored , while also presenting an experimental concept.

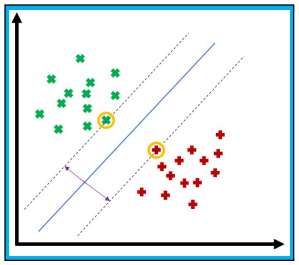
#Automatic?

By being able to create a suitable automatic detection system will drastically help the current reliance on manual reporting. Manual reporting is open to interpretation as well as many miss reportings. Furthermore automatic detection can be implemented within a higher system of methods, as an additional step, to help reduce the provenance of abusive content. In order to create abuse detection that can be executed automatically there are two main ways that have been explored in this paper. The lexicon methodology and the machine learning methodology. This paper aims to evaluate and compare the two methodologies and also explore a supervised lexicon model.

#Machine Learning

Machine learning utilises mathematical models that allow for Artificial intelligence to be able to learn how to differentiate between data. This project uses an SVM model to create a machine learning method that is able to distinguish between non abusive and abusive labels.

SVM model uses hyperplanes and marginal classifiers to create AI, that can conclusively differentiate between types of data, based on datasets used in training.



SVM Model Diagram

#Lexicon

The lexicon methodology entails the use of a large dictionary corpus of pre labelled words. These words will have associated scores that can be added up to create a total score. Once a tweet is broken down, the tweets words are isolated and examined to see if they exist within the pre labelled corpus. If they do their scores are tallied up across the entire sentence and then a total score is printed which gives an insight in to the nature of the tweet. This project uses the popular AFINN Lexicon and also explores a supervised learning lexicon concept with the use of TFIDF feature extraction.

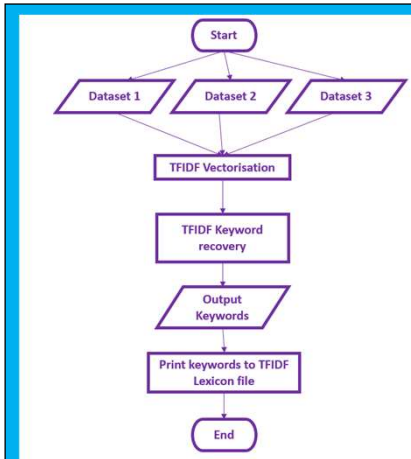
#Evaluation

To create a thorough evaluation and complete comparison of the methods presented. This project adopts a double testing method utilising validation testing and Survey testing. From the results observed, machine learning models provide a much more robust solution to issues like context and abuse definition.

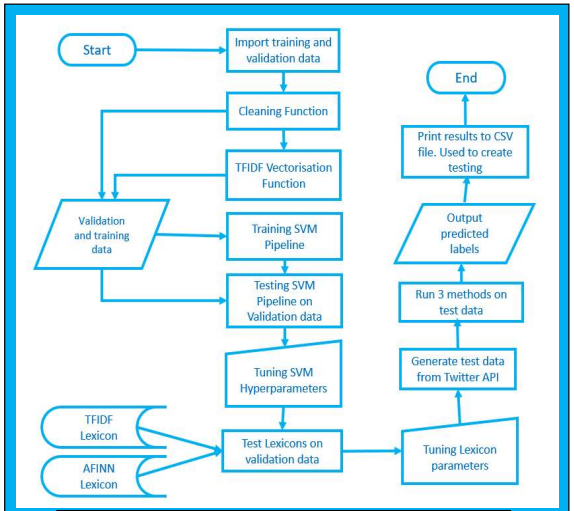
#TFIDF Concept

TFIDF is often used in documents as a keyword recovery tool. This concept was loosely applied in this project, with 3 different abuse datasets being selected and TFIDF extraction was executed. Retrieving the most important words within the datasets. Assigning these words in to a lexicon system to attempt to detect abuse in unseen tweets. This lexicon showed poor results however the concept could be explored further

#FlowCharts



TFIDF Lexicon Flow chart



Entire System flow chart

#References

Gasca, D. (2019), 'A healthier twitter: Progress and more to do'.URL:https://blog.twitter.com/enus/topics/company/2019/health−update.htmlSmith,

K. (2020), '60 incredible and interesting twitter stats and statistics

" English football announces social media boycott" (2021) URL: https://www.premierleague.com/news/2116111