NLP

PROJECT SYNOPSIS

Detection Of Offensive Text using Sentiment Analysis

Manish Shetty M (01FB16ECS192)

Neelesh C.A (01FB16ECS223)

Pallavi Mishra (01FB16ECS243)

Abstract

Hateful, offensive comments are becoming a huge problem with the rise in popularity of the internet. There are places like internet forums, twitter, facebook etc where anyone can comment. Generally, human moderators are used to filter comments. But a user has to first see the comment, and then submit the report by which time damage would already be done. Hence, a solution that uses algorithms to accomplish the same is preferable.

Introduction

Hate speech is commonly defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic. Examples :

1. The Jew Faggot was Behind The Financial Collapse
2. Hope one of those idiots falls over and breaks her leg ..

Insult/hate speech detection is a process of predicting whether a text is insulting or not. This kind of text classification can be solved manually or automatically. Manual detection can be done by classifying each text individually by means of reading the texts. This process is very time consuming and infeasible. Automatic detection of insults may use several machine learning and natural language processing techniques.

Natural language processing focusing specifically on this phenomenon is required since basic word filters do not provide a sufficient remedy.

What is considered as hate speech message might be influenced by aspects such as the domain of an utterance, its discourse context, as well as context consisting of co-occurring media objects (e.g. images, videos, audio), the exact time of posting and world events at this moment, identity of author and targeted recipient.

Current software packages that perform filtering are based on a simple keyword search i.e. comments, pages or other forms of content, are blocked if an offensive or vulgar word is found. While this technique can capture some of the offensive texts that are found online, research has shown that only 12% of insults contain vulgar words while over 33% of texts containing vulgar words are not insults.

There are cases where a system may classify the text as offensive even though it is not, an example follows. " John said that Mary called his friend an idiot." The comment is not offensive, merely giving information on what someone else did.

Despite a large amount of work, it remains difficult to compare performance of models, largely due to the use of different datasets by each work and the lack of comparative evaluations.

Sentiment Analysis

At a high level, sentiment analysis attempts to discern text with a positive disposition from text with a negative disposition. Previously classifiers would generally ignore texts that would conventionally be classified as "neutral" as they are considered to lie close to the boundary condition in a binary classifier, and are thus subject to noisy effects that are beyond the predictive capabilities of the classifier.

Hate speech and sentiment analysis are closely related, and it is safe to assume that usually negative sentiment pertains to a hate speech message. Because of this, several approaches acknowledge the relatedness of hate speech and sentiment analysis by incorporating the latter as an auxiliary classification.

Consider the following paragraph "Actually, John told that because he usually says that nonsense. Lisa said he is an idiot. But, that idiot said Lisa is a good girl. And she still prays that God heals his heart from all of his meanness. Get that socialist out of my pocket!".

The last sentence is an insult when we interpret is as somebody wants to get a socialist(human) from their pocket.. This interpretation needs some incorporation of world knowledge for capturing the demeaning of a human being's personal status. Only semantic analysis wouldn't necessarily help.

Scope of the Project

The aim of this project is to use sentiment analysis as the major portion of a larger pipeline used to detect hate speech within tweets. Some preprocessing may need to be done as people can use alternative forms as a word such as, "He is such an i****". A list of words considered bad can also be used to aid the model is deciding what is considered offensive. The project will encompass the use of hybrid models to apply sentiment analysis.

Models to consider

1. Hybrid SVMs which combine unigram-style feature-based SVMs with those based on real-valued favorability measure.
2. Hybrid system using n-gram analysis and dynamic artificial neural network
3. Multi level classifier (level 1 - lexicon based, level 2 - n gram SVM classifier)
4. IDEA : Unsupervised sentiment classification method that allows for the topical context of words in documents to be accounted for when classifying sentiment. That is given a set of positive words and negative words , some form of clustering to predict sentiment.
5. Naive Bayes Classifier
6. Logistic Regression
7. Random Forest
8. Deep Learning Model based on a few features (RNNs / LSTMs)

Literature Survey

S.no	Reference
1	https://github.com/eandrews597/Automatic-Detection-of-Insulting-Text
2	https://www.researchgate.net/publication/49242911_Detecting_flames_and_insults_in_text
3	https://aclweb.org/anthology/W18-5104
4	https://www.researchgate.net/publication/273381302_Multi-level_classifier_for_the_detection_of_insults_in_social_media
5	https://arxiv.org/abs/1712.06427
6	https://cs224d.stanford.edu/reports/Sax.pdf
7	https://beta.vu.nl/nl/Images/werkstuk-biere_tcm235-893877.pdf
8	https://www.academia.edu/38464940/Word_Embeddings_for_Sentiment_Analysis_A_Comprehensive_Empirical_Survey