Social Computing

Scope Document

Problem Statement

The objective of this project is to understand the **target of hate speech and gauge its impact by counter speech analysis**. We would be focusing only on hate speech and their corresponding replies on **twitter**.

Problem Outline , Approach and Justification of the project.

It's important to deepen our understanding of online hate speech by focusing on a largely neglected but crucial aspect of hate speech –*it's target*. We plan to approach this problem from a different direction wherein instead of analysing the tweet content we analyse the counter speech (all the direct replies to a hateful tweet) to gauge who are the people speaking for/against it.

The user characteristics and lexical analysis of the counterspeech will give insights into interesting questions like are people more vocal speaking against one type of hate speech than other , what are the common words/hashtags used by those speaking for versus those against any hateful speech. The user profile would allow us to understand the characteristics of the people who use counterspeech and then draw analysis from it.

Details

Stage 1

- Data Collection

  Hate speech collection would be done using templates, followed by manual removal of false positives. Template examples , I <intensity> <user intent> <hate target>. Additionally, I will look for curated datasets of hate speech and counterspeech pairs.

- Classification of Hate Speech

  Existing black box approaches exist for hate speech classification. Previous literature used Gender, Ethnicity , Physical Traits , Sexual Orientation , Nationality and Religion to classify hate speeches. (keyword based approaches , deep learning etc are some of the existing methods).

  Stage 1 Timeline - 22nd February

  Stage 2

- Identification of Positive and Negative opinion

  Given a hate tweet , we will classify the comments based on whether they are in favour, against or neutral to a hate speech tweet. Blackbox approaches to identify positive and negative comments.

---------- *1st Deliverable - 2nd March* ----------

- User Level Characteristics Analysis

  The counter-speech replies to the hate speech would be analysed on a user level characteristics.

  1) Quantitative analysis of creation dates , likes, followers and followings of users who comment.
  2) Understanding the bystander effect , identifying people who don't actively speak for or against hate speech but instead just like or retweet.

  Stage 2 Timeline - 12th March

- Lexical and Semantic Analysis

  A deeper understanding of the comments to identify other features

  1) Are particular counter speech to hate tweets in one category more personal and vocal compared to others.
  2) The commonly used hashtags, phrases , words.

  ---------- *2nd Deliverable - 2nd April* ----------

Thereafter polishing of the existing work , exploring new possible insights, preparation of the final presentation till *13th April.*

Challenges and Pitfalls

1. The greatest challenge and possible cause of failure is the usage of black-boxes for hate speech classification and identifying counter speech that speaks for or against a particular tweet.

   In case there are errors in the initial results there is a high chance that **later analysis becomes redundant due to propagation of errors.**

   Possible Solution - In case the initial classification is way off then :-

   - Simpler keyword approaches could be tried and then false positives removed manually.
   - Reduce the dataset size significantly (under 10k) so that the tweets may be manually annotated.

2. The project remains mostly surface level analysis.

   Solution - Most of the initial setup and crude level analysis is done by 12th March which provides a month wherein in I can dive deeper into the data and draw more inferences before the final presentation.

Non-Technical Aspects

There are multiple non-technical aspects to consider too. Blanket statements like people are more vocal against gender based hate speech as opposed to ethnicity needs to be critically analysed. In this case the location of the tweets collected in the data set might bias the results in one direction or the other.

Firstly the identification and later the classification of hate speech are both subjective choices and needs to be handled carefully.

However the positives of such a project is enormous , it will give us insights who are the ones speaking against hate speech , how are they doing it and is it effective or not (number of likes can be a crude estimate of the effectiveness).

We can also identify people who wish to speak against hate speech but tend to be passive as opposed to actively expressing their views and maybe delve deeper into why they can’t raise their voice.