Social Computing

Deliverable - 1

The code for the project is available at the private repo. I have provided access to Projit. (Will grant access to others who want it) https://github.com/arnavkapoor/CounterSpeech_Analysis

Introduction

For the first phase I have worked on the data collection aspect :-

- The quality and value of the quantitative and qualitative analysis depends upon the initial dataset that needs to be sufficiently large , well labelled and diverse enough so as to make relevant analysis on it.
- There isn't a dataset with the above qualities for hate tweets and it's counterspeech which means that using pre-existing datasets isn't going to work.

DataSets

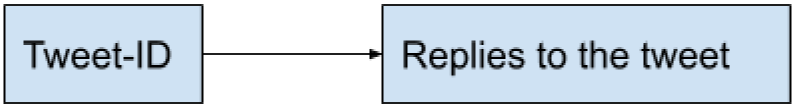
- One of the research papers[1] does provide a curated list of hate tweets and counter speech twitter-id tuples . However these aren't labelled (in the data made public). More importantly **these only contain template generated hate speech** of the form I < intensity >< userintent >< hatetarget > , thus missing out on more nuanced hate and since this 'nuanced hate' makes up the majority of the tweets , this dataset doesn't provide the complete picture. (Also the dataset contains only about 500 unique hate-tweets)

- To get a more representative and broader range of tweets , existing **annotated dataset of tweet id and category of speech** needed to be considered. However this brought about technological/implementation challenges that needed to be overcome. One of the most trusted dataset , for tweet-id and hate speech seems to be NAACL_SRW_2016.csv , thus this was the focus for the time-being.

[1] "Analyzing the hate and counter speech accounts on Twitter." 6 Dec. 2018, https://arxiv.org/abs/1812.02712.

Challenges

- Thus our objective is to get all the replies , given a particular tweet id to get the replies to it.



- However the problem isn't as trivial as it seems because the twitter api doesn't provide a direct way to query for the replies from an id.
- Thus the steps to achieve the same are as follows:
  1. Get information about the tweet from the tweet-id , when and who posted it.
  2. Do a Search Query with the relevant parameters (to: 'the hate tweet author' , between dates: 'hate tweet date' to 'hate tweet date + 21 days'.) to identify possible replies.
  3. Parse the json of the possible replies to match the conversation id of the result with the conversation id of the hate tweet.

We cannot query the twitter search API and simply use the 'in_reply_to_status_id' field which contains the original Tweet's ID for replies.

This is because of the restriction of the search API that only returns results of the last week while the actual hate speech data set contains tweets from 2016. Thus modifications in this repo[2] was done to make it possible.

[2] "Jefferson-Henrique/GetOldTweets-python ...." https://github.com/Jefferson-Henrique/GetOldTweets-python.

Current Status Summary

At present I have a system in place that allows me given a tweet id to retrieve all reply tweets from it. I have run it for the above data set and have some hate-tweet reply pairs (less than 1000) because a lot of the original tweets don't exist any more. (About 50% of the tweets are no longer available) and most of the remaining have zero replies to them.

However I am more hopeful as all I require are tweet-id and labels and other resources (eg- https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection) for the same are available.

However instead of just focusing on hate speech , I might broaden the scope to abusive language too as the objective is to focus more on the replies and draw inferences from it.

Future Plans

I plan to manually go through the data entries first to understand it better. Thereafter would see what interesting quantitative analysis can be drawn from it. Parallely I am also finding more tweet_id , counter_id pairs so the overall analysis can be on more substantial data. (at least 2k unique hate tweet id)