Measurements and Analysis

Bhuvan Chadha
Computer Science
SUNY Binghamton
bchadha1@binghamton.edu

Daniel Paul
Computer Science
SUNY Binghamton
dpaul12@binghamton.edu

Naveen C. Poda
Computer Science
SUNY Binghamton
npoda1@binghamton.edu

Naresh K. Anantharam
Computer Science
SUNY Binghamton
nananth1@binghamton.edu

ABSTRACT

Social media has emerged as a medium to raise users' opinions and influence the way businesses, political interests, public activities exist and operate. Opinion of people matters a lot to analyze how the propagation of information impacts the lives in a large-scale network like Twitter. Sentiment analysis of the tweets determine the polarity and inclination of vast population towards specific topic, item, or entity. These days, the applications of such analysis can be easily observed during public elections, movie promotions, brand endorsements and many other fields. The primary aim is to provide a method for analyzing sentiment score in noisy twitter and 4Chan streams and other public forums. This project reports on the design of a sentiment analysis, extracting vast number of tweets. Results classify users' perception via tweets into positive and negative. Secondly, we discuss various techniques to carryout sentiment analysis on twitter data in detail.

KEYWORDS

Social Media, Twitter, 4Chan, Tweets, politics, textblob, sentiment analysis

RESEARCH OBJECTIVES

For this project, we have thought about and shall try to work upon three research questions/areas based on our datasets and the type of data we have gathered. Our research objectives shall revolve around the sentiments and behavior of common public on the social media and public discussion forums like 4/Chan and Twitter.

- Sentiments of people on US Presidential Elections 2020 with measurements such as anxiety, frustration, and happiness
  This objective gives an idea on how people reacted to the ongoing US Elections which has been followed by people all around the world.
- Measurement of toxicity and extreme radical language on Twitter and 4Chan in comments, posts, and Tweets
  This is done by using Perspective API from google which helps to cross check keywords from a dataset and enumerates the toxicity count.
- Retweet frequency based on historian style, viral news style or troll slang style and the impact they have created in reaching to the wider audience.
  Historian style is a type of twitter post where a post is mentioned along with a historical correlation to the existing post. Usually a previous incident or a milestone is mentioned in a post along with its similarity to the new incident and in this case US elections.
  Viral news style is another type where a link is provided in the tweet with a reference to a fact.
  Troll slang is the most common type of post where a troll or a sarcastic slang is mentioned in the post with reference to a recent situation.

With these research areas, we try to analyze the political mood of various types of users.

DATASET COLLECTION

As part of the previous subproject we have managed to collect relevant data from both Twitter and 4/Chan platforms.
For the sentimental analysis part of our project we shall focus majorly on political posts.
This was relatively more feasible in 4chan as there was a category linked to politically related posts called as pol.
From Twitter we had collected data everyday as suggested and filtered the necessary data to make a suitable data frame to perform analysis on.

METHODOLOGY

Sentiment analysis is done by using the algorithms that find polarity as below.

Finding polarity: For discovering the polarity, we will use an algorithm of counting strong positive and negative words in tweets, in comments and posts. For both, positive and negative words, different lists were made. Next step is to compare every word in a tweet against both these lists. If the current word matches a word in positive list, then a score of 1 is incremented and if a negative word is found then it is decremented. More positive words lead to higher sentiment score and more negative words lead to lesser sentiment score.

Stop words: Tweets and 4/Chan posts and comments contain stop words which are common in nature. These can be “is”, “am”, “are” which hold no additional information. These words serve no purpose, and this feature is implemented using a list stored in stopwords.dat file. We will then compare each word in a tweet with this list and delete the words matching the stop list to focus on strong words/language.

For the first part of our sentiment **analysis, we** have used TextBlob library as a **classifier. Instead** of training a classifier using Naïve-Bayes or any other technique we had used this existing library in python to classify our data. We have trimmed down our data to be suitable for analysis by converting all the alphabets to lower case using lambda() function and getting rid of any punctuation marks using regular expression and replacing them with a blank space.

From this data, we have sorted out only the necessary content for our analysis by choosing specific attributes like id, timestamp, content of the post instead of all the attributes.

Now we have used sentiment attribute of the TextBlob library to provide a polarity and subjectivity score. Polarity in this case tells us how whether a post is negative or positive. We had used a scale of -1 to +1 to measure positivity with -1 being the most negative and +1 being most positive. Subjectivity is also measured on the same scale and it tells us how relevant a post is to our existing analysis.

Based on the inputs of polarity and subjectivity, we have plotted a graph with these values over time.

For the second part of our analysis i.e., measuring the toxicity in a post we have used the perspective API from google which provides us with a toxicity score based on the words in a post.

How **this function** is the algorithm crosschecks every word in the post with a set of predefined words considered toxic and increments the toxicity score when a match is found. We have reached out to google to receive the Perspective API key. Once we received **it,** we used one of the attributes TOXICITY of the perspective API for our analysis. Perspective API provides us with two outputs namely **‘**span score**’** and **‘**summary score**’**.

‘Span score**’** gives us with a score of **parts** in a text whereas **‘**summary score**’** gives an aggregate of the spanscore combining all the **‘**span score**’**.

For our third research question, we need to provide a comparison between different types of posts based on the style in which they were posted i.e., Historian, Viral **news,** or Troll slang type of post.

Among these we need to check the retweet frequency.

Below are the results of the measures of ‘Toxicity’ and ‘Severe Toxicity’, i.e. the span scores and summary scores.


```
"attributeScores": {
  "SEVERE_TOXICITY": {
    "spanScores": [
      {
        "begin": 0,
        "end": 92,
        "score": {
          "value": 0.3336587,
          "type": "PROBABILITY"
        }
      }
    ],
    "summaryScore": {
      "value": 0.3336587,
      "type": "PROBABILITY"
    }
  },
  "TOXICITY": {
    "spanScores": [
      {
        "begin": 0,
        "end": 92,
        "score": {
          "value": 0.8785311,
          "type": "PROBABILITY"
        }
      }
    ],
    "summaryScore": {
      "value": 0.8785311,
      "type": "PROBABILITY"
    }
  }
}

{
  "attributeScores": {
    "TOXICITY": {
      "spanScores": [
        {
          "begin": 0,
          "end": 145,
          "score": {
            "value": 0.44006696,
            "type": "PROBABILITY"
          }
        }
      ],
      "summaryScore": {
        "value": 0.44006696,
        "type": "PROBABILITY"
      }
    },
    "SEVERE_TOXICITY": {
      "spanScores": [
        {
          "begin": 0,
          "end": 145,
          "score": {
            "value": 0.28657368,
            "type": "PROBABILITY"
          }
        }
      ],
```

DATA VISUALISATION

Below is a sample snapshot of the data we have used in a tabular format. This contains the id of the post, its **timestamp,** and the content

_id	time_stamp	com
so-whats-going-to-happen	2020-11-05T00:52:31.570Z	To these people when joe is officially president? Are they worried about viewership plummeting? The news has been dominated by trump for 4 years without him t
did-biden-fuck-up-on-election-night-when-he	2020-11-05T00:52:31.604Z	did Biden fuck up on election night when he jumped the gun as he was the first one on screen, WHILE the ballots were still being counted, and he smiled with confi
activate-them	2020-11-05T00:52:31.607Z	ACTIVATE THEM ALL
say-it-with-me	2020-11-08T15:58:53.051Z	>I will own nothing >I will live in a pod >I will eat the bugs</span
when-will-they-stop	2020-11-05T00:52:31.612Z	I thought the raids would end post-election or at least during it. Half the posts now are bait the other half sound like Joe himself typed them
i-love-seeing-trump-lose-its-my-daily-medicine-my	2020-11-05T00:52:31.615Z	I love seeing TRUMP lose, it's my daily medicine, my weekly energy, my monthly inspiration and my yearly motivation. His loss is the only reason i&#039
lets-be-honest-here-trump-was-a-joke-that-went	2020-11-05T00:52:31.618Z	Let's be honest here, Trump was a joke that went too far.
is-it-too-late-to-just-redo-the-election-with	2020-11-05T00:52:31.620Z	Is it too late to just re-do the election? With better counters. Perhaps there could be a coin toss to determine who gets the President gig? Maybe just mak
accelerate	2020-11-05T00:52:31.622Z	THIS IS THE CORRECT FUCKING TIMELINE BUCKLE UP AND HIT THIS SHIT
is-trump-the-last-republican-ever	2020-11-05T00:52:31.624Z	Well /pol/
maga-bros-shilling-for-pa	2020-11-05T00:52:31.626Z	Please remember Biden only needs 6 more AC votes not 56
literally-no-record-to-say-you-voted-not-even	2020-11-05T00:52:31.628Z	>Literally no record to say you voted (not even getting to the issue of secret ballot retardation) unless you mail-in vote and only in certain
az-mi-and-pa-bros	2020-11-05T00:52:31.630Z	Cernovich will be in Phoenix at the AZ State Capitol building at 8pm tonight ALSO Tomorrow at 9am FreedomWorks will be at City Hall Let's yo
why-dont-you-americans-count-till-its-done	2020-11-05T00:52:31.632Z	Why do you stop counting and take breaks and shit? Faggots.
election-stealing-get-in-here-quick	2020-11-05T01:10:49.757Z	https://www.youtube.com/watch?v=iviT=wbn2s5Ug#ho&feature=emb_title - 12:27 Just look at those nigresses walking around with bags and coole
how-long-has-it-been	2020-11-05T00:52:31.636Z	How long has it been?
why-dont-they-just-revote-but-only-have-trump-and	2020-11-05T00:52:31.638Z	Why don't they just revote but only have Trump and Biden?

Table 1. Dataset with columns ID, Date/Time and Comments

Below are the plots modelled after our sentiment analysis was performed.

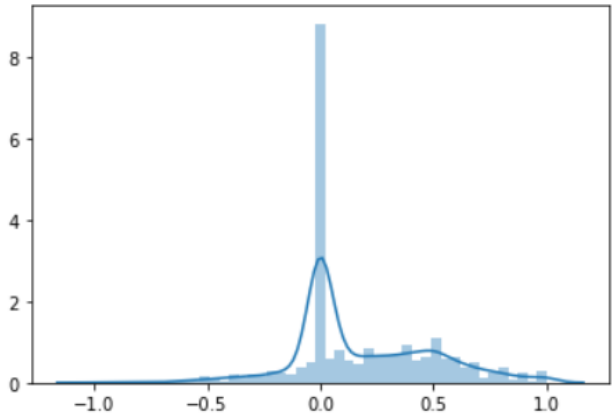

Figure 1. 4/Chan Data Analysis plotting

The above plot is plotted using POL data. Looking at the above graph we notice most of the posts were **neutral (**close to 0).

Below is a similar graph we have received from twitter **data analysis.**

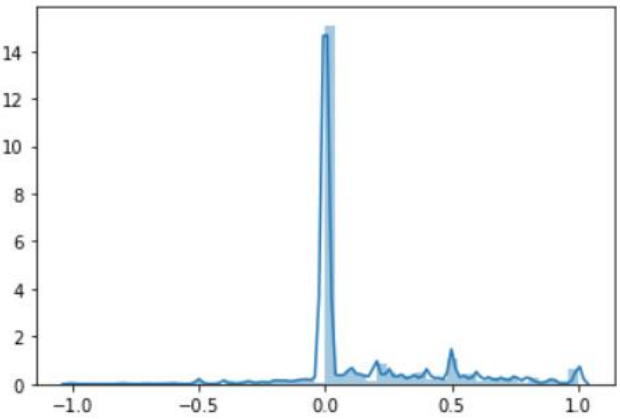

Figure 2. Twitter data Analysis Plotting

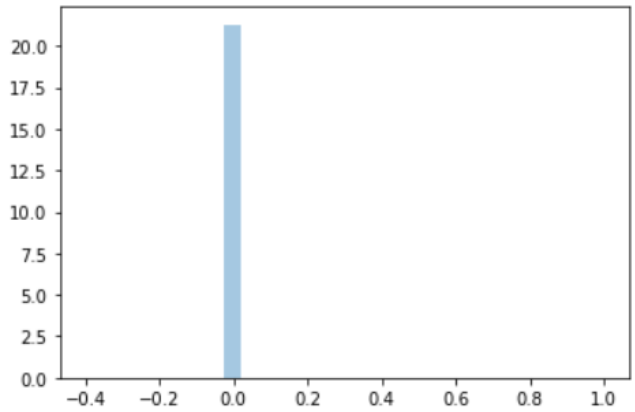

Figure 3. The Plot post-election results

Looking at the graphs we can say most of the posts were neutral in nature (close to 0).

We also noticed that although the dataset used for twitter was huge, there was a significant increase in the negative posts in POL data comparatively.

	4/Chan Data Collection	Twitter Data Collection
Size	36,000	141,000

Table 2. Size of data for 4/Chan and Twitter on which our sentiment analysis is performed

From this we can say that pol data had more negative sentiments over twitter.

Analyzing our data for retweet frequency we noticed that highest number of posts which were retweeted were following the Viral News style. Next in line would be posts which were of the Troll Slang style.

CONCLUSION

In this project, we tried to analyze the political mood and sentiments of public on Twitter and 4/Chan platforms based on their comments and the words/language they used. Many of the comments had extreme language in both, positive and negative sense. We analyzed these sentiments based on two major measures – polarity and subjectivity. Further, for our second research question, we used two different measures – span score and summary score. We used TextBlob library for analysis which is a great tool for getting the scores. We used *matplotlib* library for plotting the graphs based on various scores and outputs. We tried to answer our third research question regarding the type/category of the comments, which shall be worked upon further.

REFERENCES

[1] Twitter Developer Docs
https://developer.twitter.com/en/docs/apps/overview

[**2**] Yehia Khoja. 2019. *Twitter data collection tutorial using Python* https://towardsdatascience.com/twitter-data-collection-tutorial-using-python-3267d7cfa93e

[**3**] Deepesh Khaneja. 2017. *SENTIMENT ANALYSIS ON TWITTER USING APACHE SPARK* https://www.researchgate.net/publication/320625064_PROJECT_REPORT_SENTIMENT_ANALYSIS_ON_TWITTER_USING_APACHE_SPARK

[4] Perspective API https://www.perspectiveapi.com/#/start

[5] https://github.com/conversationai/perspectiveapi/tree/master/1-get-started

ACKNOWLEDGEMENT

It gives us immense pleasure to thank Prof. Jeremey H Blackburn for guiding us through this project and helping us improve in the subject with all his knowledge, expertise, and experience.