IMPACT OF RACISM ON US ELECTIONS 2020

Shashank Bengaluru Srinivasa
sbengal2@binghamton.edu
State University of New York at Binghamton
Binghamton, New York

Aravind Reddy Yenugula
ayenugu1@binghamton.edu
State University of New York at Binghamton
Binghamton, New York

Darshan Doddaghatta
ddoddag1@binghamton.edu
State University of New York at Binghamton
Binghamton, New York

Abstract

Race and ethnicity have always played a role in politics around the world, in the same order they influencing the politics in a diverse and multicultural country like USA comes as no surprise. In the light of recent events triggered by the death of **George Floyd**, it surely does seem like it is going to have its effect on the upcoming US Elections 2020.

In this project we intend to gain an overview of the general public behind these events, its influence on their political stand and through all this try to predict the result of the US elections 2020. What better option than social media for this ? Therefore, we shall use social media sites such as twitter and reddit to collect the required data. We intend to perform a lexicon based sentiment analysis over the data and try to predict the outcome of the elections.

Keywords: data-sets, sentiment analysis, text polarity

ACM Reference Format:
Shashank Bengaluru Srinivasa, Aravind Reddy Yenugula, and Darshan Doddaghatta. 2018. IMPACT OF RACISM ON US ELECTIONS 2020. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 Introduction

We use twitter as the primary source of data, twitter being one of the most popular social media for political discussions. We chose Reddit as the second source as it had many active political discussion communities also know as subreddits.

We use the twitter's well built **sampled stream** api that gives us 1% of all tweets in real-time and reddit's **search** api.

Following the implementation of code for data collection, we carried out a test run for around 7 days (4 days for reddit) . We use the data collected during this time to understand its qualitative and quantitative measure and how it could aide us in achieving our objective. In this report we try to explain the challenges, projections and expectations in the process of our data collection.

2 Implementation

We made http requests to relevant servers to collect the required data from Twitter and Reddit.

2.1 Twitter Implementation

We use the Twitter's sampled stream api to collect data from twitter. According to the twitters api documentation : "The sampled stream endpoint delivers a roughly1% random sample of publicly available Tweets in real-time. With it, you can identify and track trends, monitor general sentiment, monitor global events, and much more". You can connect one client per session, and can disconnect and reconnect no more than 50 times per 15 minute window. The data that came our way was not filtered and the entire tweet object with all its fields was collected. The code for this is implemented using Node Js.

2.2 Reddit Implementation

We use the Reddit's search api to collect data from reddit. We collect all the subreddits (relevant to our query) returned by the search api and run each of them through a method that fetches all the comments in each subreddit. The code for this is implemented using python and runs every five minutes to collect data.

The data pulled from both the sources is stored directly into our Mongo data base.

3 Modifications

In the proposal, the projected estimation of data was based on the idea that we would be using the twitter's filtered stream api. As of now, we are using the sampled stream api and the proposed estimation is not valid anymore. The sampled stream seems to deliver a handsome amount of data. Since the sampled stream api is being used, we do not use any keyword filtering for queries like we had mentioned earlier in the proposal.

In case of reddit, As proposed initially, we still use the search api. In addition to that we fetch the comment objects by querying the URLs built using the subreddits returned. To make it more efficient we supply the most active politics, ethnicity and race related subreddits along with relevant keywords as queries to the search api.

4 Preliminary investigation of the collected data

4.1 Twitter Data

We collected data from the twitter's sampled stream api from 27th October 2020 through 3rd November 2020. We were able to collect around 20 million tweets. The figure 1 shows the number of tweets collected daily through the week.

Twitter's sampled stream api gives us 1% of tweets from all around the world. Since our study is centered around the United States elections, we understand that most of the

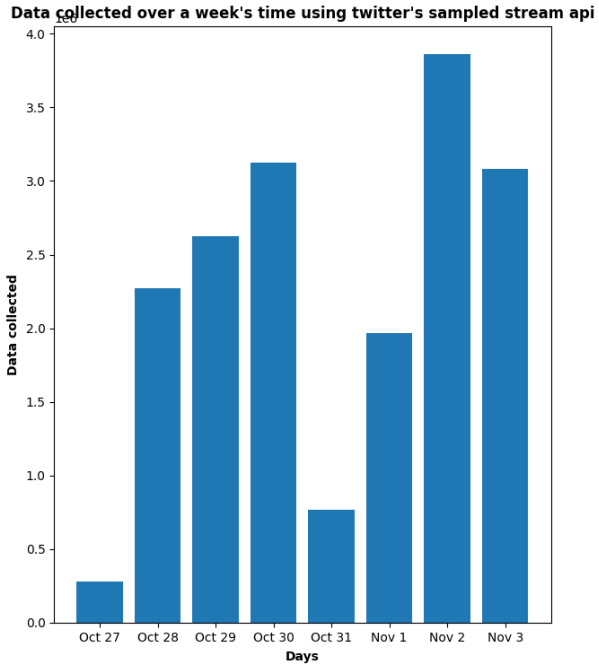

Figure 1. Number of tweets collected over a week daily from **10/27/2020** through **11/3/2020**

collected tweets(coming from outside the USA) is irrelevant. To make some use of it, we ran the data through a set of filters such as twitter hashtags and keywords the are relevant to our area of research.

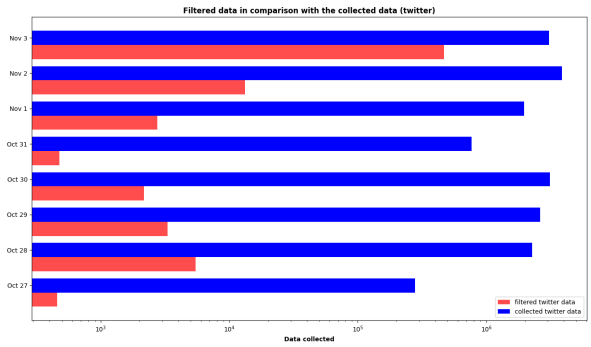

Figure 2. Number of filtered and unfiltered tweets

The 20 million data-set was filtered down to around 500k tweets. The figure 2 shows the number of tweets after filtering the data compared with the unfiltered data taken on logarithmic scale, we call this as 'filtered data'. The table 1 shows the distribution of tweets and comments amongst different topics obtained after filtering.

Number of tweet/comments by topic				
Source	**Trump**	**Biden**	**Elections**	**Racism**
Twitter	22773	17907	161270	26470
Reddit	45106	86	74608	1516

Table 1. Distribution of tweets/comments by topic

4.2 Reddit Data

We collected around half a million comments from reddit, from 2nd November 2020 through 5th November 2020.
The Reddit search api returns everything that matches the query. As we intend to perform a sentiment analysis on the data, we require only the comments. To make the search more efficient we googled a bunch of most active political subreddits on reddit and used them as queries. The subreddits returned by the search api is then used to get all the comments in them. The figure 3 shows the number of comments collected from Reddit through the days mentioned above.

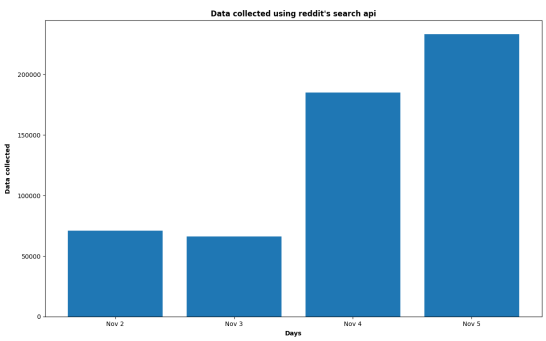

Figure 3. Number of reddit comments collected over a week daily from **11/02/2020** through **11/05/2020**

4.3 Election week

Another interesting development we observed while studying the data was the variation in the number of tweets during the election week. November 3rd which also happened to be the Election Day saw a tremendous increase in the number of tweets related to the USA elections 2020. Using the filtered data, figure 4 shows the rise in the number of election related tweets on November 3rd 2020.

5 Projected Data

Based on the observation made on the collected data, we believe the data we expect to collect over the course of 10 days is not going to exceed the storage limit on the VMs.

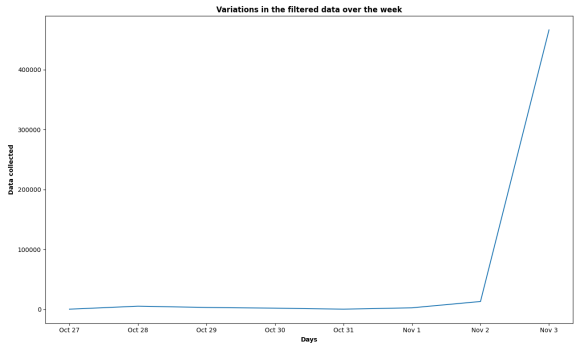

Figure 4. Increase in the number of tweets collected on election day

6 Challenges/Hardships

6.1 Comments/Tweets

- `Sampled stream`: Twitter's sampled stream api provides us with tweets from all over the world. Most of this data is of no relevance to our study. The filtered data-set is quite small and cannot accurately describe the sentiments of people on a huge social network like twitter.
- `Reddit`: Reddit API is vague and not well documented, it took a lot of effort to learn how to retrieve the comments without using libraries such as PRAW.

6.2 Sarcasm

Most of the content on social media is sarcastic. When we studied our data, we found that some of the tweets and comments were sarcastic in nature.For example consider this tweet, **"@realDonaldTrump Politics based on racism, well, that's something new for Donald Trump!"**,it is quite complicated to decipher the sentiment on texts like these.
Therefore, sarcastic content can be quite misleading when it comes to sentiment analysis and detecting it, is quite complex. As of now we believe a solution for this could be beyond the scope of the study.

6.3 Database

We observed that daily querying and dropping of collected data is more efficient than querying it all at once.This also prevents VMs from getting full.

7 Acknowledgments

We thank our Professor, Jeremy Blackburn for all his assistance for this project. We offer our sincere appreciation for all the learning opportunities.