IMPACT OF RACISM ON US ELECTIONS 2020

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

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

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

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 police brutality (death of **George Floyd**) and the persistent racial inequality, it surely does seem like it has had its effect on the 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 ? We use social media sites such as twitter[10] and reddit[4] to gather the required data. We intend to perform a lexicon based sentiment analysis over the data using VADER, provide a binary classification of tweets / comments and find out how well it agrees with the results of the US Elections 2020 .

Keywords: Racism, Sentiment Analysis, US Elections, Black Lives Matter, Donald trump, Joe Biden

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, 5 pages. https://doi.org/10.1145/1122445.1122456

1 Introduction

The 2020 US Presidential Elections are done. Joe Biden is the 46th president of the United States, Kamala Harris, has become the first woman, first Black American, and first Asian-American to be elected vice president of the United States.

After a year of nationwide protests over systemic racism and police brutality, the results of the now concluded elections were not so surprising. Racism and related issues has always bothered the United States [11].

But did racism really have a key role in determining the results of this year's presidential elections?

What is the view of people on the the two presidential candidates? After almost a decade long of Donald Trump's administration, were people tired of his discriminatory politics? What is their opinion on Joe Biden?

Did Joe Biden really win the elections? Is he peoples' favorite or is he just an alternative to Donald Trump that people wanted.

Biden's eight years as Barack Obama's vice president and their apparently effective professional and affectionate personal relationship – might have helped inspire African American voters. Biden-Harris victory also marked a new high in the representation of African American women in federal politics. A big question is whether the mobilisation that occurred through BLM rallies translate into high participation by minority and young people in the election [17]. In an election year marked by pandemic,unemployment, and widespread protest against racial inequality in the United States, we turn to social media for answers. We try to provide insight and analysis to make sense of the impact of racism on the elections.

We collected the data for around 10 days, from 11th November 2020 through 22nd November 2020. Twitter is the primary source of data, tweets are collected using twitter's sampled stream API. Alongside Twitter, Reddit is used as a secondary source, reddit comments are collected using its search API.

The data collected is run through a set of keywords related to our topic of research. This is our **'Race'** data-set. The race data-set is further filtered into two data-sets, **'Trump'** and **'Biden'** data-sets, again using a different set of keywords for each. The tweets/comments in the three data-sets are then pre-processed to make it suitable for lexicon based analysis.

A well compiled twitter trainer data-set containing 1.6 million tweets[8] was used to determine which of the two libraries (VADER/TextBlob) was more suitable to obtain the sentiment of the tweets/comments. The VADER library had a better accuracy rate than TextBlob.

The pre-processed tweets and comments are then assigned a compound score for their sentiment using the VADER

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

library. Based on the compound score each tweet/comment is tagged as 'Positive' or 'Negative'.

We calculate the percentage of postive and negative tweets / comments for each dataset. The Biden dataset received a better percentage of positive tweets/comments than the trump dataset corroborating very well with results of the US Elections 2020.

2 Background and Related Work

2.1 Twitter as a data source

Twitter has long been a rich source of data for researchers to track, collect and analyze peoples' opinion on events happening around the world. In the paper "#Hashtagging hate: Using Twitter to track racism online"[14] the author provides an insight into the ability of twitter to track racism online. Quoting the works of Dhiraj Murthy[16] and R.Rogers[18] he explains how twitter makes a great source of data for researchers. He also reviews the sampled stream API of twitter and the method to collect tweets from it.

2.2 VADER

"VADER performed as well as (and in most cases, better than) eleven other highly regarded sentiment analysis tools" , say the authors in "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text"[15] explaining how well equipped and remarkable VADER is for sentiment analysis, especially in microblog-like contexts .

3 Methodology

Our study consists of three parts. Part one concerns the collection of data from twitter and reddit. Part two deals with data preprocessing work used to clean and remove irrelevant information from the tweets and comments. Part three deals with the use of the [2] NLTK's VADER analyzer for scoring each tweet and comment and consequent polarity tagging method applied .

4 Data Acquisition

4.1 Twitter

Wikipedia describes Twitter[10] as an American microblogging and social networking service on which users post and interact with messages known as "tweets". Tweets were originally restricted to 140 characters, but was doubled to 280 for non-CJK languages in November 2017.

" Twitter has become a powerhouse in sharing news information (both locally and globally). Users are able to search for key words or hashtags related to an event, moment, or experience, allowing them to feel like they are experiencing the event in real time" says author Irfan Chaudhry in his paper, "Hashtagging hate: Using Twitter to track racism"[14]. The author cites work of R.Rogers from "Debanalising Twitter: The Transformation of an Object of Study"[18] to explain how twitter can be a great source of data for researchers as it allows to track,capture and analyze users' responses and activities, calling it Twitter III. The author also explains how although twitter's streaming API is great for collecting data on a current event it can still be useless in collecting historical data. All of this makes twitter a great source to gather people's opinion on the 2020 US elections.

We use the Twitter's sampled stream API[7] to collect data from twitter. This method allowed real-time access to publicly available raw tweets. 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 and data was gathered using the http requests.

We collected data from the twitter's sampled stream API from 11th November 2020 through 22nd November 2020. We were able to collect around 38 million tweets. The figure 1 shows the number of tweets collected hourly everyday through a period of 10 days (from 11th November 2020 through 22nd November 2020)

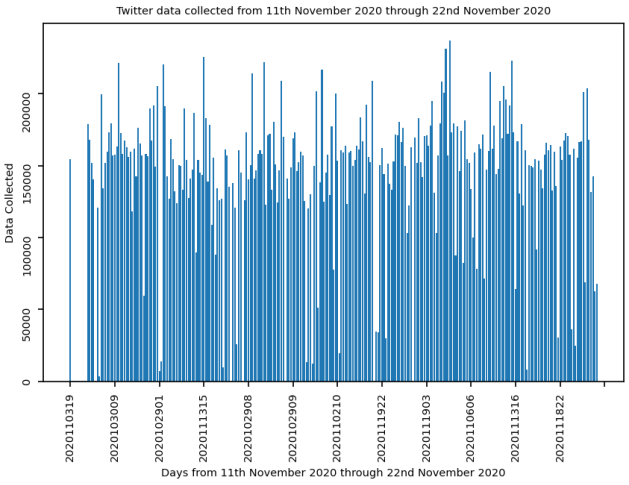

Figure 1. Twitter data collected over a period of 10 days, from **11/11/2020** through **11/22/2020** shown hourly

4.2 Reddit

Wikipedia describes Reddit as an American social news aggregation, web content rating, and discussion website. Registered members submit content to the site such as links, text posts, and images, which are then voted up or down by other members. Posts are organized by subject into user-created boards called "subreddits", which cover a variety of topics

such as news, politics, science, movies, video games, music, books, sports, fitness, cooking, pets, and image-sharing.

According to The website 'Statista' [6] more than 50 percent of users of reddit are from United States, making reddit a perfect source of data for our study.

The Reddit search API [5] 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.A subreddit called "What are the biggest political subs on Reddit?"[13] helped us get a list of most politically active subreddits. The subreddits returned by the search API is then used to get all the comments in them.

We collected around 2.5 million comments from reddit, from 11th November 2020 through 22nd November 2020. The figure 2 shows the number of reddit comments collected hourly everyday through a period of 10 days.

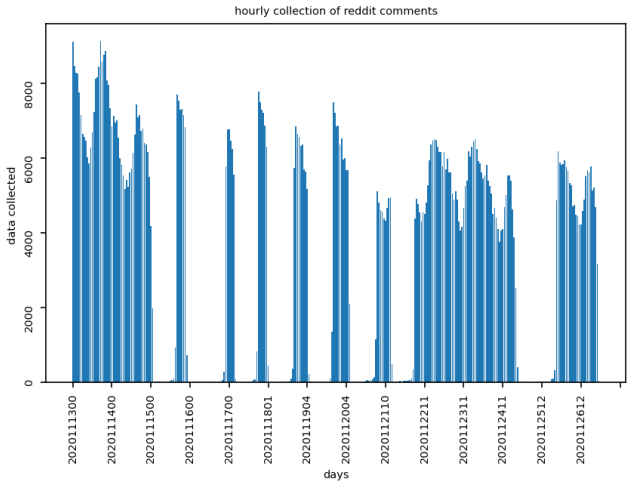

Figure 2. Reddit data collected over a period of 10 days shown hourly

5 Data Filtering

As mentioned earlier twitters sampled stream API provides us with tweets from all over the world, therefore most the tweets are irrelevant to our study. To make the collected data more specific we filter it. We apply the same filtering techniques to both twitter and reddit data.

5.1 Relevant Datasets

The raw data collected is first filtered to create a broad, working dataset. We call it 'Race' dataset. This dataset is created using a set of 95 racism related keywords. We searched the web for a wide range of racism related words commonly used on social media sites. We sourced most of the keywords from the twitter handle "Racism Watchdog" (@RacismDog)[3] and website "racial equity tools"[1].

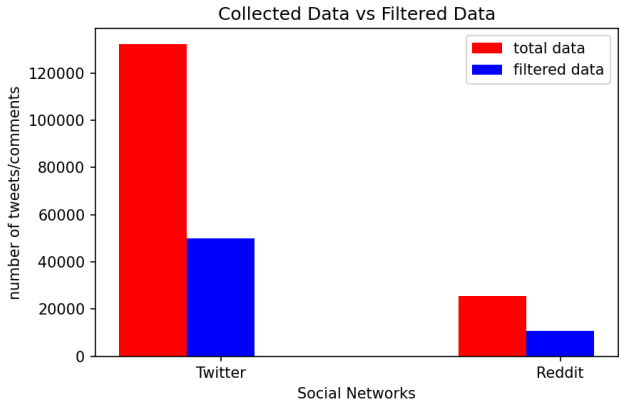

Figure 3. collected data vs filtered data over a period of 10 days, from **11/11/2020** through **11/22/2020**

The Race dataset was further filtered into Trump dataset (Using keywords related Donald Trump) and Biden dataset (Using keywords related to Joe Biden and Kamala Harris). The table 1 shows the distribution of tweets and comments amongst different topics obtained after filtering.

Number of tweets / comments by data set				
Source	**Overall**	**Race**	**Trump**	**Biden**
Twitter	38213834	332790	34922	43583
Reddit	2509653	132235	25397	29378

Table 1. Distribution of tweets / comments by data set

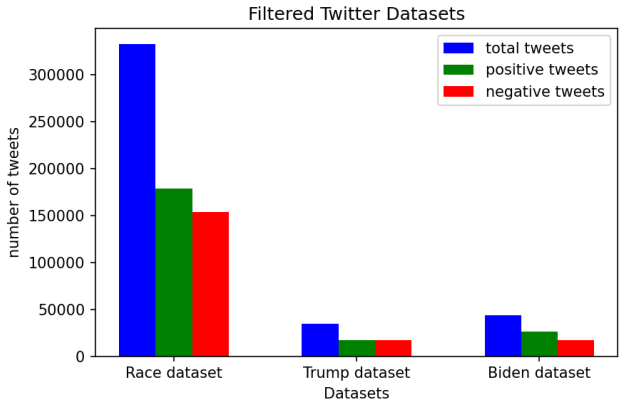

Figure 4. Distribution of tweets with their polarity amongst each dataset **11/11/2020** through **11/22/2020**

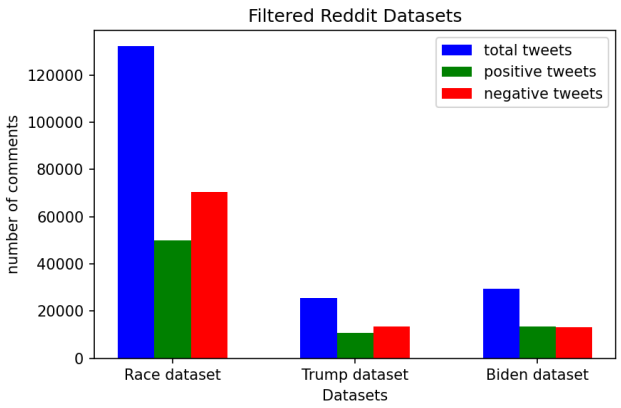

Figure 5. Distribution of comments with their polarity amongst each dataset.

6 Sentiment Analysis

6.1 VADER vs TextBlob

When we decided to perform a sentiment analysis on tweet / comments we had two well developed python libraries at our disposal, VADER[12], a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media and TextBlob[9], a python library for processing textual data for common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

We measured the accuracy rate of the two libraries using a twitter trainer dataset[8] containing 1.6 million tweets with their polarity scores (0 for Negative and 4 for Positive).The scores were converted into tags, 'Positive if 4 and 'Negative' if 0. The tweets in the trainer dataset were preprocessed before running them through the two tools. The tweets were then assigned a compound score using VADER and polarity score using TextBlob. The compound score and the polarity scores were converted into vader_polarity and textblob_polarity tags respectively ('Positive' if score was greater than 0, 'Negative' if score was lesser than 0). The tweets without any compound score or polarity score were removed. The vader_polarity and textblob_polarity tags were compared to the original tags to determine their accuracy. Testing over 473987 tweets (reduced from 1.6 million tweets) VADER scored 294699 and TextBlob scored 273657. With 621% accuracy VADER was chosen for the sentiment analysis.

6.2 Data Preprocessing

Most tweets / comments contain text and embed URLs, pictures, usernames, and emoticons. They also contain misspellings. Hence, we perform a series of preprocessing steps to remove irrelevant information from them. The clean data would provide a better accuracy for text based analysis. Duplicate tweets / comments were removed from the dataset.

Natural Language Toolkit (NLTK)[2] library was used to preprocess the data further. First, a regular expression (Regex) in Python was run to detect and discard special characters, such as URLs ("http://url"), retweet (RT), user mention (@), and unwanted punctuation. Since Hashtags (#) often explain the subject of the tweet and contain useful information related to the topic of the study, they are added as a part of the tweet, but the "#" symbol was removed.

Further, NLTK were used to convert the tweets to lowercase, remove stop words (i.e., words that do not express any meaning, such as is, a, the, he, them, etc.), tokenize the tweets into individual words or tokens, and stem the tweets using the Porter stemmer. When the preprocessing steps are complete, the dataset was ready for sentiment classification.

6.3 Sentiment Classification

VADER Sentiment Analyzer was applied to the dataset. Using VADER each tweet/comment was classified as positive, negative or neutral. VADER is a rule-based sentiment analysis tool and a lexicon that is used to express sentiments in social media. Each tweet is assigned a positve, negative, neutral and compound score. If the compound score is lesser than 0, the tweet or comment is tagged as 'Negative', if the compound score is greater than 0, it is tagged 'Positive'.

7 Limitations

7.1 Sarcasm and Context

When it comes to a any language the context plays a significant role in determining the meaning and intent. Analyzing a few tweets manually we found a good number of tweets were sarcastic in nature and some tweets were mere banters, perhaps amongst friends. 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.

"Language classifiers are useful tools to filter and manage large data sets, however, also need to be considered within the contextual nature of the tweet itself. This calls for the researcher to not only consider the initial tweet but also understand its context, content, and construct. What this means is the tweet cannot be taken only at face value"[14].

7.2 VADER

Although VADER seems to be a good choice for lexicon based sentiment analysis for social media texts, with an accuracy of 62%(as per our test), using it to project people's opinion on a large social network like twitter seems quite inefficient.

Machine learning and Deep Learning would be a better choice in this aspect.

8 Results And Discussions

The results of this study clearly corroborate with the results of the US elections 2020.

8.1 Race data set

Race related twitter dataset has 332790 tweets, out of which 53% of tweets are positive and 46% of tweets are negative and the race related reddit dataset has 132235 comments with 37% of comments being positive and 54% being negative. This actually shows the mixed opinions people had about racism and elections on the two large social media networks.

8.2 Trump data set

Donald trump related twitter data set has 34922 tweets, 48% of tweets are positive and 50% are negative. Coming to the reddit data set with 25397 comments, 42% are positive and 53%are negative.

This shows the negative emotion of people towards trump on the two social networks.

8.3 Biden data set

Joe Biden twitter data set has 43583 tweets, 60% of tweets are positive and 40% are negative. The reddit data set has 29378 comments, 46% of comments are positive and 44% of comments are negative

This shows the positive emotion of people towards biden on the two social networks.

9 Conclusion

On both, Twitter and Reddit, Joe Biden had a larger percentage of positive tweets / comments and smaller percentage of negative tweets / comments as compared to Donald Trump. Both these social networks are have a large number of users and with most of the users belonging to United States, we believe, it is safe to assume that the user opinions on these sites truly reflect the opinions of majority of people in United States

With 14% of all positive tweets on twitter and 27% of all positive comments on reddit we declare Joe Biden as the winner of the US elections 2020.

10 Acknowledgments

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

References

[1] 2019. *Racial Equity Tools.* Retrieved November 26, 2020 from https://www.racialequitytools.org/glossary

[2] 2020. *Natural Language Toolkit.* Retrieved November 26, 2020 from http://www.nltk.org/

[3] 2020. *Racism Watchdog.* Retrieved November 26, 2020 from https://twitter.com/RacismDog?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor

[4] 2020. *Reddit.* Retrieved November 22, 2020 from https://reddit.com

[5] 2020. *Reddit API documentation.* Retrieved November 26, 2020 from https://www.reddit.com/dev/api/

[6] 2020. *Reddit Users By Country.* Retrieved November 26, 2020 from https://www.statista.com/forecasts/1174696/reddit-user-by-country

[7] 2020. *Sampled Stream.* Retrieved November 26, 2020 from https://developer.twitter.com/en/docs/twitter-api/tweets/sampled-stream/introduction

[8] 2020. *Sentiment140 dataset with 1.6 million tweets.* Retrieved November 26, 2018 from https://www.kaggle.com/kazanova/sentiment140

[9] 2020. *TextBlob: Simplified Text Processing.* Retrieved November 26, 2020 from https://textblob.readthedocs.io/en/dev/

[10] 2020. *Twitter.* Retrieved October 21, 2020 from https://twitter.com

[11] 2020. *US election 2020: Why racism is still a problem for the world's most powerful country.* Retrieved November 23, 2020 from https://www.bbc.com/news/election-us-2020-54738922

[12] 2020. *vaderSentiment.* Retrieved November 26, 2020 from https://pypi.org/project/vaderSentiment/

[13] 2020. *What are the biggest political subs on Reddit?* Retrieved November 22, 2020 from https://www.reddit.com/r/neoliberal/comments/9195w2/what_are_the_biggest_political_subs_on_reddit/

[14] Irfan Chaudhry. 2015. Hashtagging hate: Using Twitter to track racism online. *First Monday* 20, 2 (Feb. 2015). https://doi.org/10.5210/fm.v20i2.5450

[15] C.J. Hutto and Eric Gilbert. 2015. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.*

[16] D. Murthy. 2018. *Twitter.* Wiley. https://books.google.com/books?id=LJ1PDwAAQBAJ

[17] Racism 2020. *Racism has long shaped US presidential elections. Here's how it might play out in 2020.* Retrieved October 29, 2020 from https://theconversation.com/racism-has-long-shaped-us-presidential-elections-heres-how-it-might-play-out-in-2020-147556

[18] R. Rogers. 2014. Foreword: Debanalising Twitter: The Transformation of an Object of Study.