CS 5123 – Cloud and Distributed System

PROJECT REPORT ON:

Analysis of relation between U.S President and social events using twitter data

Team Members:Ateeve, Meghana, Damith, Ippei

//check names

CONTENT:

Contents

[1. INTRODUCTION 3](#_Toc499884716)

[2. BACKGROUND 4](#_Toc499884717)

[3. DESIGN AND SOLUTION: 5](#_Toc499884718)

[4. PROGRAMS AND ANALYSIS 8](#_Toc499884719)

[In fig 4: Emotion Value 1 is the highest. The news reads “NFL Anthem Protests Continue as Owners Dither, Trump Agitates, Fans React”. 10](#_Toc499884720)

[People showed negative impact against NFL. 10](#_Toc499884721)

// will include this after all editing

# 1. INTRODUCTION

As technology is advancing people are using social media as a platform to express themselves. Twitter is one of those platforms where enormous rate of tweets are produced per day. The amount of data generated is very large for a particular topic, such as politics, music, sports, and so on User generated content on twitter provides a rich source of data. This data can be collected to analyse and gain deeper understanding about people’s behaviour, thoughts, actions and reactions.

There are various projects these days on Emotion recognition using the data generated on twitter. In this project we are trying to analyse the relation between U.S President (Donald Trump) and Social events using twitter data. This involves a little of sentiment analysis and emotion recognition.

Data was collected about Donald Trump’s recent activities from twitter. This data was cleaned and staged for further analysis. Sentiment analysis was performed on the data, and the output of this was used as input to Map Reduce Programs to provide the final analysis and graphs

After analysing data, we can come up with diverse types of conclusions and compare with what happens in those dates. Based on those comparisons we can come up with a model that can use in future for any relevant kind of scenarios.

// if u can think of something please include

# 2. BACKGROUND

// idea on what to write …. Cause introduction also has the same

// Is it ok to put the literature review here ?

# 3. DESIGN AND SOLUTION

## Flow chart

TWITTER

Apache Flume Agent

based on keywords

Collecting data

Format: **Date time:: time zone :: tweet**

JSON EXTRACTOR

Cleaning Data

Filter by date 2

Filter by date 2

Filter by date 1

Format: **Date time:: time zone :: Emotion value**

Sentiment Analysis

Emotion value: 0, 1,2,3,4

[**0**- very negative,

**1**-negative**,**

**2**-neutral,

**3**-positive,

**4** – very Positive]

Map reduce programs

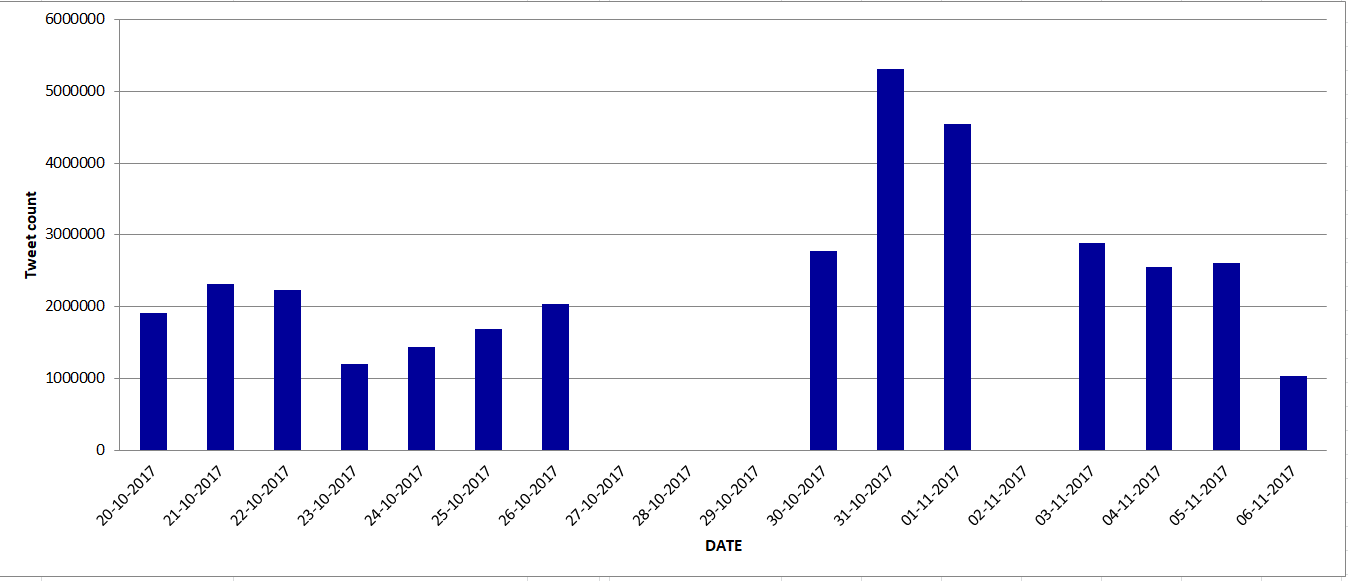
Analysis using the output

## Key words

// Putt all the keywords here

## Data collection –

* Data is collected using flume agent from Twitter
* Size of data collected = 188.4 GB
* Collected from: October 20th, 12:04 am to: November 6th, 10:38 am
* The data collection was stopped by system administrator, since it was connecting to illegal sites. Thereby we have missing data for October 27,28,29 and November 2nd

 Fig 1: Data Collected per day

**Cleaning Data** –

* Data is cleaned using JSON Extractor
* We divided the data based on the dates and cleaned individual files.
* After cleaning we had 4.8GB of data
* Input to JSON extractor : Raw data
* Output : **Data-time:: time zone :: tweet message**

**Sentiment Analysis –**

* For the sentiment analysis, we are using StanfordCoreNLP package and libraries
* StanfordCoreNLP includes the sentiment tool and various programs which support it. The model can be used to analyze text as part of StanfordCoreNLP by adding “sentiment” to the list of annotators [1]
* Sentiment analysis was run on individual date files in-order to reduce the time taken for analysis
* Input: **Data-time:: time zone :: tweet**
* Output: **Data-time:: time zone::emotion value**
* Emotion value: 0, 1,2,3,4

1. very negative
2. negative
3. neutral
4. positive
5. very Positive

//screenshot of output

# 4. PROGRAMS AND ANALYSIS

**Map Reduce Programs –**

Each Map Reduce job was run on individual date files that we got after sentiment analysis, then the output was combined during analysis stage

**Program 1:**

* Input: **Data-time:: time zone::emotion value**
* Output: **Data:: emotion value :: emotion count**

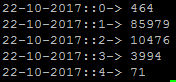


Fig 2: Sample output of Program 1

Graphs:

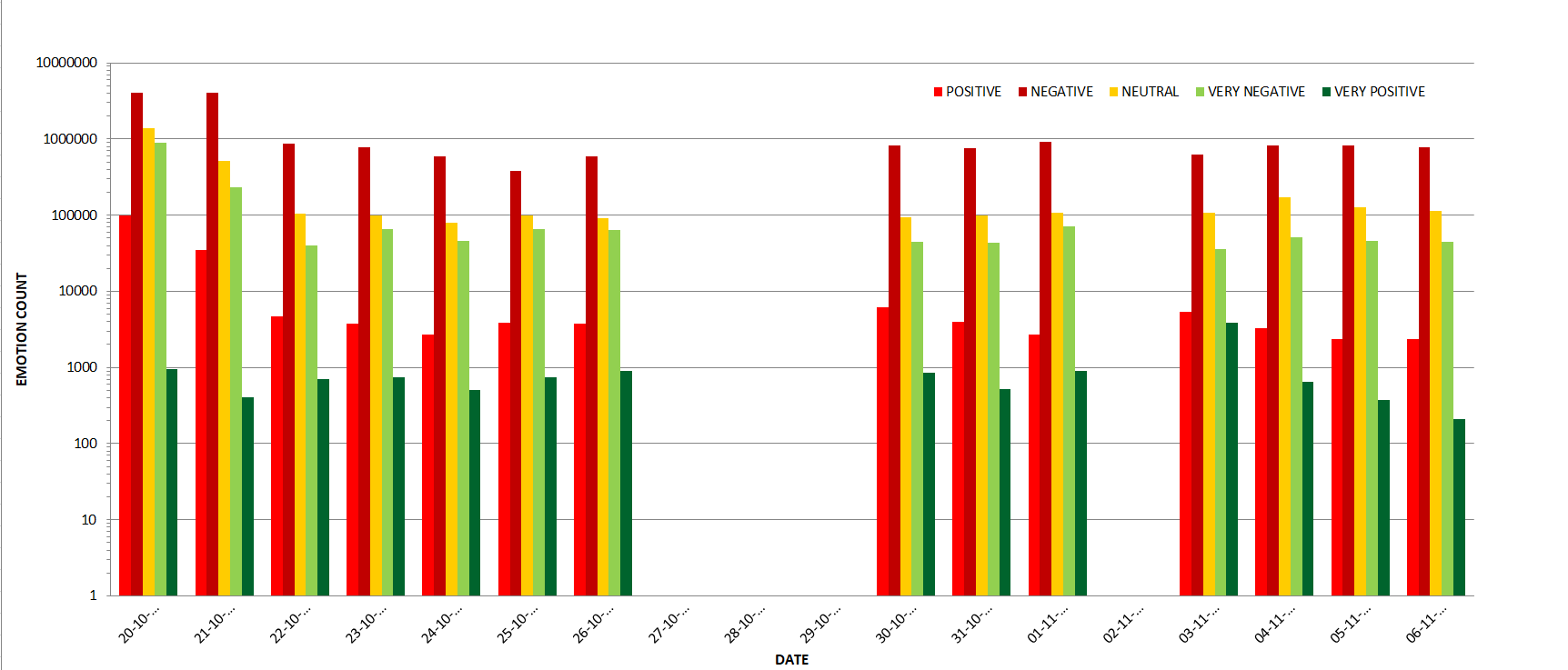


Fig 3: Cumulative graph for output of program 1

Analysis

Based on the Tweets counts

1. Let us consider a single day 20-10-2017

We can see that the value of 1 (which is depicts negative emotion) is high.

When we trace back to the news we found that on this day **both Barack Obama and George W. Bush criticized president Trump.**

We can predict that either people had negative opinion on Obama and Bush or on Trump.

But again when we went through the news on this we found that people had more negative opinion on Barack and Bush for criticizing Trump.

With this we also get to know that there is large number of supporters for trump

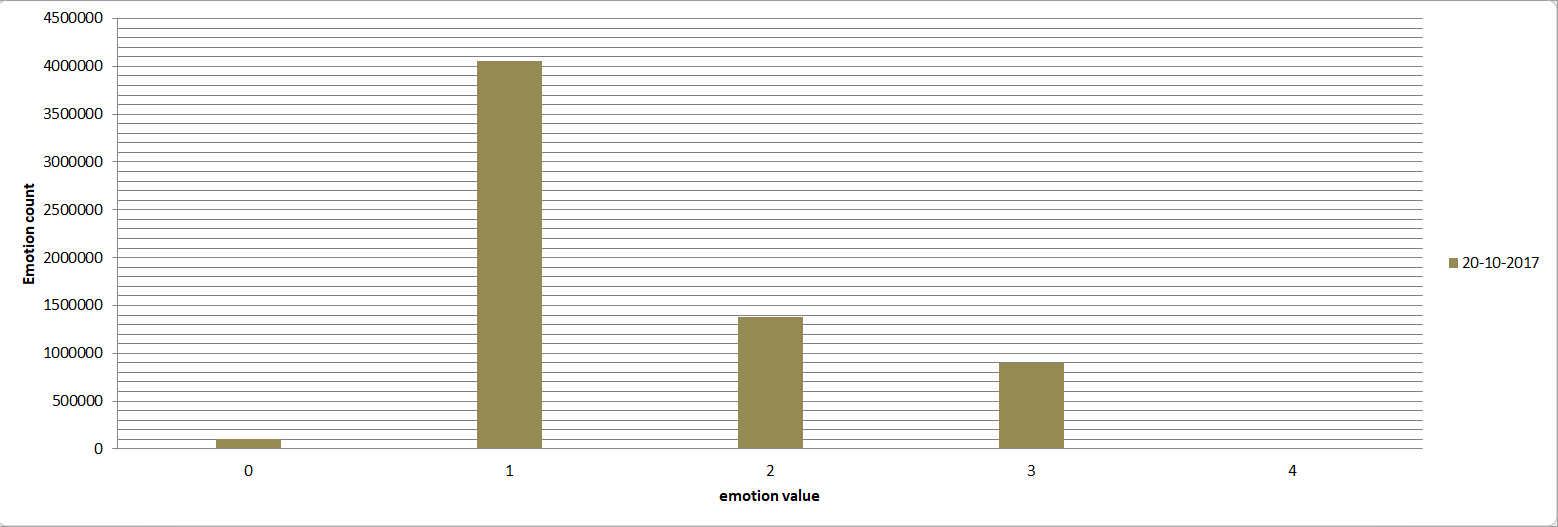


Fig 4: Graph for 20-10-2017

1. Similarly, we generated graphs for other dates and traced back to find the social events

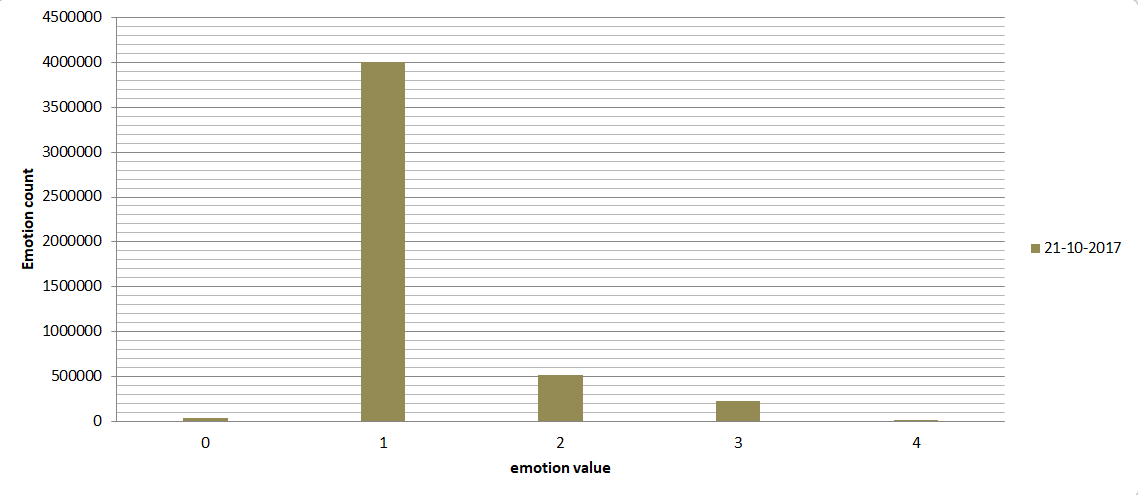


Fig 4: Graph for 21-10-2017

# In fig 4: Emotion Value 1 is the highest. The news reads “NFL Anthem Protests Continue as Owners Dither, Trump Agitates, Fans React”.

# People showed negative impact against NFL.

**Program 2:**

* Input: **Data-time:: time zone::emotion value**
* Output: **time zone:: emotion value :: emotion count**
* **Here we are trying to predict over the entire span of data collection (that is from Oct 21-Nov 6) ,how many people in a particular time zone have been negative or positive towards the social events that happened in Trumps presidency**

**Sample output:**

America/Aruba::2-> 5

America/Aruba::3-> 1

America/Asuncion::1-> 13

America/Asuncion::2-> 2

Analysis :

There were more than 300 time zones; we tried to keep our focus on America and popularity of trump over the period of data collected

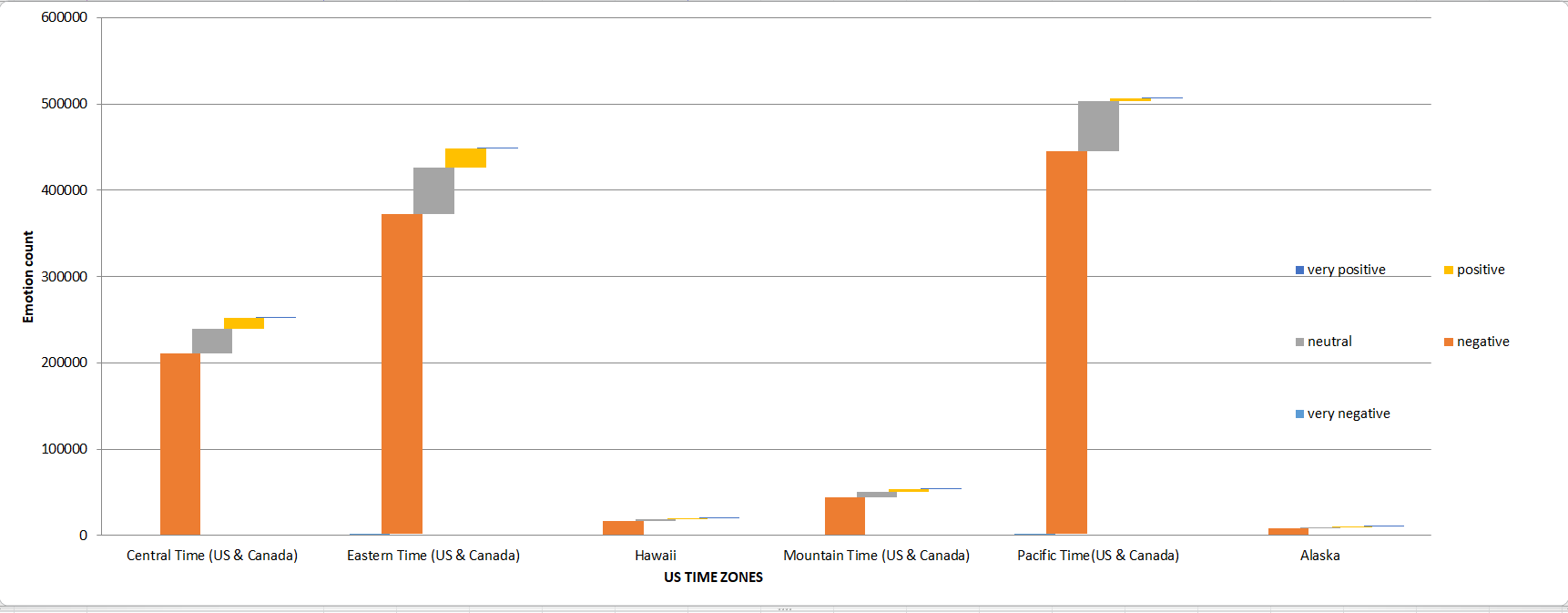
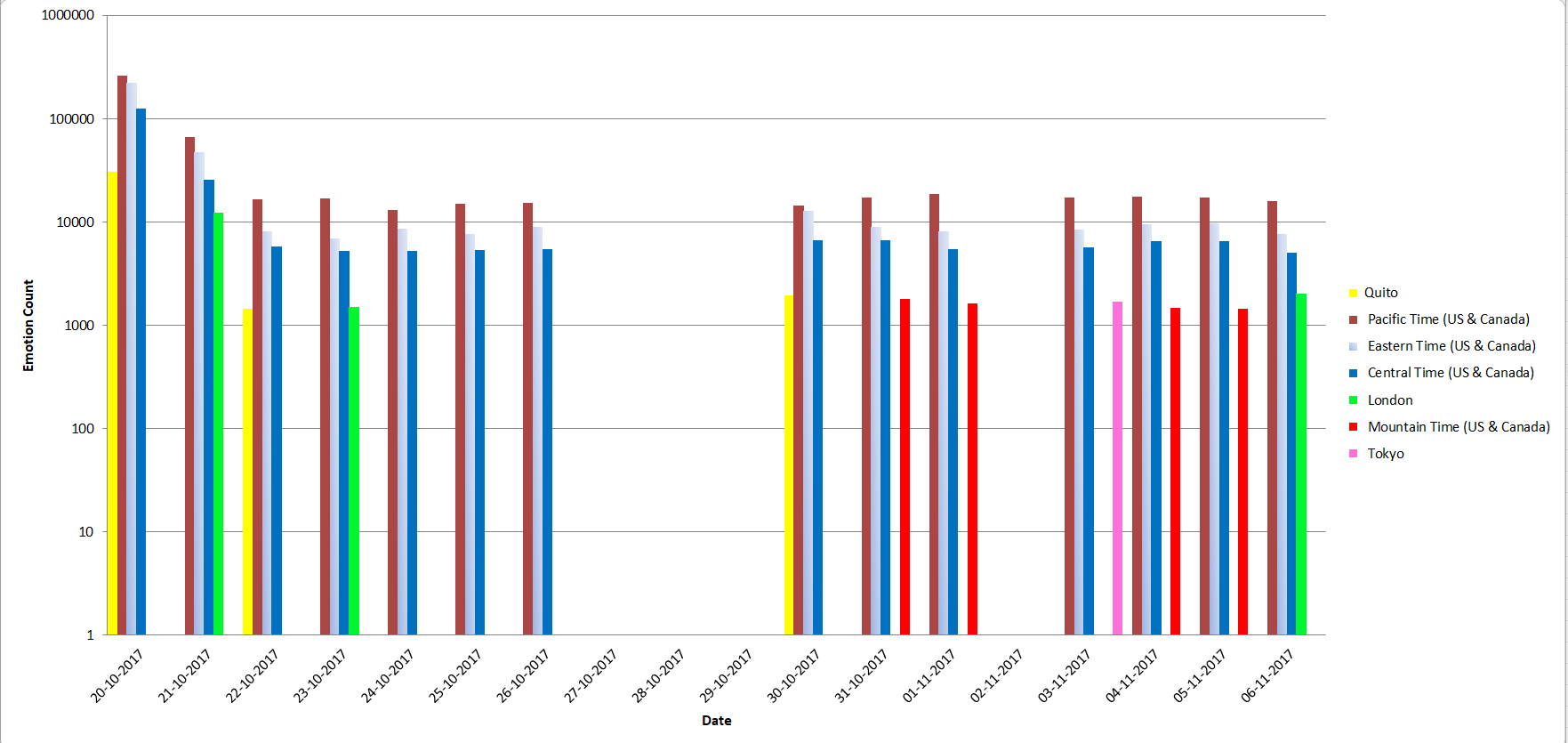


Fig 5: US time zone and emotion value

Program 3:



// https://www.conservativedailynews.com/2017/11/president-donald-trumps-schedule-saturday-november-4-2017/

Trump’s and Melinda trump’s was in Tokyo on 11-04 so we can see number of tweets on the graph.

Also, we can say most of the tweet users are form pacific time.

5. LITERATURE REVIEW:

[1] Xu, Weiai Wayne, Yoonmo Sang, Stacy Blasiola, and Han Woo Park. "Predicting opinion leaders in Twitter activism networks: The case of the Wisconsin recall election." American Behavioral Scientist 58, no. 10 (2014): 1278-1293.

Network and content analysis techniques were used to explore the influence of tweets of individual users in comparison to tweets by organizations. The results showed that organization tweets had greater influence than individual tweets. This paper gives us an idea how to use keywords to collect the data and to do a comparative analysis to know the reaction of people and organizations on various issues during the presidential term.

[2] Kruikemeier, Sanne. "How political candidates use Twitter and the impact on votes." Computers in Human Behavior 34 (2014): 131-139.

This paper investigates how using twitter by political candidates during the course of election campaign had positive impacts on their votes. Through this paper we get an insight how to use the data quantitatively, to analyse how number of followers and popularity of Donald trump changed over the issues.

[3] Golbeck, Jennifer, Justin M. Grimes, and Anthony Rogers. "Twitter use by the US Congress." Journal of the Association for Information Science and Technology 61, no. 8 (2010): 1612-1621.

Here they try to analyse if twitter tends to provide insight or improves transparency into legislature, but the findings show that congressmen rather increase their popularity using twitter. We can gain an insight of how to classify the data based on its contents, and how Donald Trump’s team is using twitter for their and his benefit, this can result in impacting the number of followers he has

[4] Conway, Bethany A., Kate Kenski, and Di Wang. "The rise of Twitter in the political campaign: Searching for intermedia agenda‐setting effects in the presidential primary." Journal of Computer‐Mediated Communication 20, no. 4 (2015): 363-380.

This paper shows the impact of social media on political communication landscape and how media predicts political agenda on twitter by following candidates on certain topics. This paper gives us an idea of how we can use time series analysis and predict future agendas of Donald Trump and the people’s emotional reaction to that.

[5] Hutchison, Ken, and Soundar Kumara. "Big data analytics-sentiment analysis of Twitter data using clustering techniques." In IIE annual conference. Proceedings, p. 2495. Institute of Industrial and Systems Engineers (IISE), 2013.

This paper uses sentiment analysis to process social media data, and explores the use of clustering methods on said sentiment analysis to generate a sparse text corpus in response to a specific subject matter query. We can use cluster analysis technique on the data mined from twitter to do a sentimental analysis of people on presidential talks and issues. Through this w e can get a sentiment score and frequencies of reaction, with which we can conclude Donald Trump’s overall state among people.

[6] Fernández-Gavilanes, Milagros, Tamara Álvarez-López, Jonathan Juncal-Martínez, Enrique Costa-Montenegro, and Francisco Javier González-Castaño. "Unsupervised method for sentiment analysis in online texts." Expert Systems with Applications 58 (2016): 57-75.

The main objective of this research was to predict whether an online text expresses positive, negative or neutral sentiments without the need of supervision. The approach is based on an unsupervised dependency parsing-based text classification method that uses a variety of NLP (natural language processing) techniques and sentiment features derive from sentiment tokens. This paper has used unsupervised learning technique to know sentiments of online text. This paper can be used as reference if we want to apply some unsupervised technique in order to know people sentiments/reactions regarding trump's various decisions or policies. The advantage of using unsupervised learning technique is that it does not need labelled text for prior training, we can apply on in all types of text domain.

[7] Liu, Bing. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5, no. 1 (2012): 1-167.

The book deals with all the issues related to natural language processing, social media analysis, text mining and data mining. The book has various ways and techniques in order to do sentiment analysis of a text in detail manner. Thus, this book can be used as a reference to do our analysis on tweets related to Trump.

[8] Birjali, Marouane, Abderrahim Beni-Hssane, and Mohammed Erritali. "Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks." Procedia Computer Science 113 (2017): 65-72.

This paper tries to find who have the tendency to have suicidal thoughts by analysing their tweets. The paper uses a set of vocabulary related to suicide, and performs sentiment analysis using machine learning algorithms: IB1, J48, CART, SMO, and Naïve Bayes. They used the twitter4j to collect the twitter data. How they used machine learning algorithms can be applied to our project.

[9] Watts, David, K. M. George, TK Ashwin Kumar, and Zenia Arora. "Tweet sentiment as proxy for political campaign momentum." In Big Data (Big Data), 2016 IEEE International Conference on, pp. 2475-2484. IEEE, 2016.

This paper ties to model the momentum of political campaign using Twitter data. It proposes three momentum indicators: sentiment indicator (SI), curvature indicator (CI), and growth indicator (GI). All models were validated with tweets collected during 2014 midterm election and 2016 presidential primaries. Their research is related to our project, but they are more advanced and I will not be able to apply to our project.

Other journals reviewed:

[10] Bora, Prachi. Using machine learning techniques to predict the education level of Twitter users. University of Maryland, Baltimore County, 2015.

[11] Bakliwal, Akshat, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. "Sentiment analysis of political tweets: Towards an accurate classifier." Association for Computational Linguistics, 2013.

6. CONCLUSION AND FUTURE WORK

This model can be used for any political parties or person who involve in politics to keep eye on his popularity among the social media users. Based on the conclusions they can plan their future work that benefits for them.

7. REFERENCES