Social Information Networks Review 3 Project Report

Fall Semester 2020-21

Data Extraction and Analysis from "Twitter" Using Web Scrapping & Analysis.

J Component Final Review Report

Submitted by

AMOUGH MITTAL ADITYA MISHRA

(17BCE0210) (17BCE2065)

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

NOVEMBER & YEAR 2020

TABLE OF CONTENTS

Chapter I Abstract

In this project, we attempt to do the sentimental analysis of the 2016 US presidential elections. Sentimental analysis requires the data to be extracted from websites or sources where people present their opinions, views, complaints about the subjects that need to analyzed .Furthermore, it is necessary to ensure that the sample size of the data is large enough to get conclusive results .It is also necessary to ensure that the data is cleaned before it is used to make predictions. Cleaning is done using common techniques like tokenization, spell check, etc. Sentimental Analysis is one of the by-products of Natural Language Processing. This paper includes data collection as well as classification of textual data based on machine learning.

Chapter II Introduction

Web Scrapping is data scraping used for extracting data from the desired or target websites. It is the architecture to access the World Wide Web directly using the HTTP, or through a web browser. The term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying, in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later analysis. Web scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page. Therefore, web crawling is the main component of web scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted and its data copied into a desired format. Sentiment analysis requires data extracted from websites where people share their reviews, opinions or complaints about services, products, movies, music or any other consumer focused offering. Extracting this user generated content would be the first step in any sentiment analysis project and web scraping serves the purpose efficiently. The main objective of our project is extracting the details of various candidates who stood in the 2016 US presidential election from Twitter using web scraping techniques. Then, Sentiment analysis of reviews the candidates will be performed to classify them as positive, negative and neutral. This will help identify the most popular candidates in the election trending on the social media and predict who has the highest possibility of winning the election.

Chapter III Methodology adopted

A. **GATHER DATA:** In this step we made a web crawler targeted at twitter.The web crawler is responsible for collecting all the data from Twitter. The tweets we collected are stored along with a class label for each tweet , based on these the words in the tweets , they are classified under some pre-defined categories (calm,angry, anxious, neutral)

B. **PRE-PROCESS DATA:** In this step, we decided to process the data before we are able to extract the features. The various pre-processing steps we applied are,

1. Tokenization: The tweets are tokenized using the tweettokenizer. A tokenizer divides a string into substrings by splitting on the specified string .These tokens are further used for parsing and data mining.
2. Remove punctuation marks: Unwanted punctuation marks are removed from the data so that the data set remains pure.
3. Remove Stop-words: determiners, prepositions and coordinating conjunctions are removed from the dataset so that it only contains relevant words. Word -removal is a crucial step to supervised learning.
4. Spell check: We perform spell-check on the tweets to ensure that the feature-set being generated has relevant words and not commonly misspelled words, apart from this, spell-check allows for accurate frequency calculation, which is crucial when the basis of the feature-set generation is frequency distribution over the set of processed documents. This is accomplished by using a big.txt file which consists of about a million words. The file is a concatenation of several public domain books from Project Gutenberg and lists of the most frequent words from Wiktionary and the British National Corpus. We then extract the individual words from the file and train a probability model(based on occurrence of each word). The resultant probability distribution is smoothened over the parts that would have been zero(words that have not occurred in the big.txt file) by bumping them up to the smallest possible count. This process of spellchecking is performed two times using an edit-distance of 2, this was done after analyzing that spell-checking twice gives the best result.

C. FEATURE-SET GENERATION: In NLP and information retrieval, bag-of-words is used as a simplified representation. Here, a text is represented as a bag (multiset) of its words, disregarding the word order and the grammar associated with the text. To generate the feature-set two techniques are

considered, tf-idf and term frequency. Upon analysis, it is observed that tf-idf based feature extraction results in removal of words important to the classification of text as positive or negative. Tf-idf ends up penalising words that are crucial for the definition and the words that appear a large number of times in the document.One such instance is with the word “great”, the word great occurred 786 times, whereas the word “of” occurred 745 times. If tf- idf is used , the word “great” is removed , which is key in defining what a user thinks of an application. The alternative to this approach is the frequency distribution method for generating the feature-set. After removal of stop-words, this method gives a feature set that appears to be very similar to a good feature set.

D. CLASSIFICATION:

1. K- Nearest Neighbor: K-nearest neighbors is one of the classification algorithms that trains itself by using similarity measures within its boundary.
2. Naive Bayes: It is a classification technique based on Bayes’ theorem with an assumption of independence among predictors. It defines a class of classifiers, a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.
3. Decision Trees: Decision trees are powerful tools for classification and prediction. They represent rules, which can be understood by humans and used in knowledge system such as database. The following are the key requirements:

• Attribute-value description: Object or case must be expressible in terms of a fixed collection of properties or attributes for e.g., hot, mild, cold

• Predefined classes (target values): The target function has discrete output values for e.g., Boolean or multi class.

• Sufficient data: Enough training cases should be provided to learn the model.

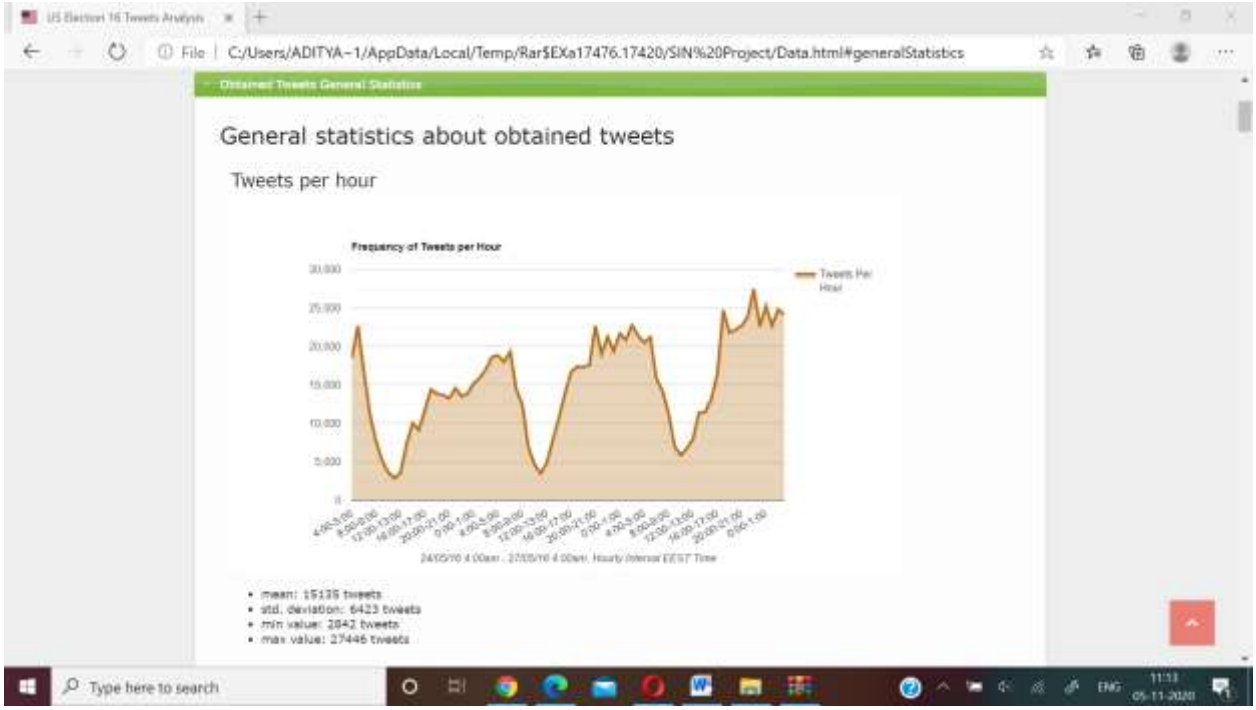
• Parameters: The parameters of this classifier are tuned by varying the minimum depth as threshold.

E. ANALYSIS FOR PICKING GOOD CLASSIFIER: All the results of the evaluation are stored into a csv file and the mean and standard deviations of resubstituting and generalization errors are compared for all the combinations of classifiers with generators. Based on these results it is observed SMO-RBF kernel with C=5,SMO- RBF kernel with C=1, SMO-linear kernel with C=1, Naive Bayes, J48-30 classifiers suited best for the classification of the data. To make sure that the results are true they are compared using the t- values that have been generated using student t-test. The ones whose p value is close to 0 i.e. lesser than 0.05 are picked and the ones whose value is greater than

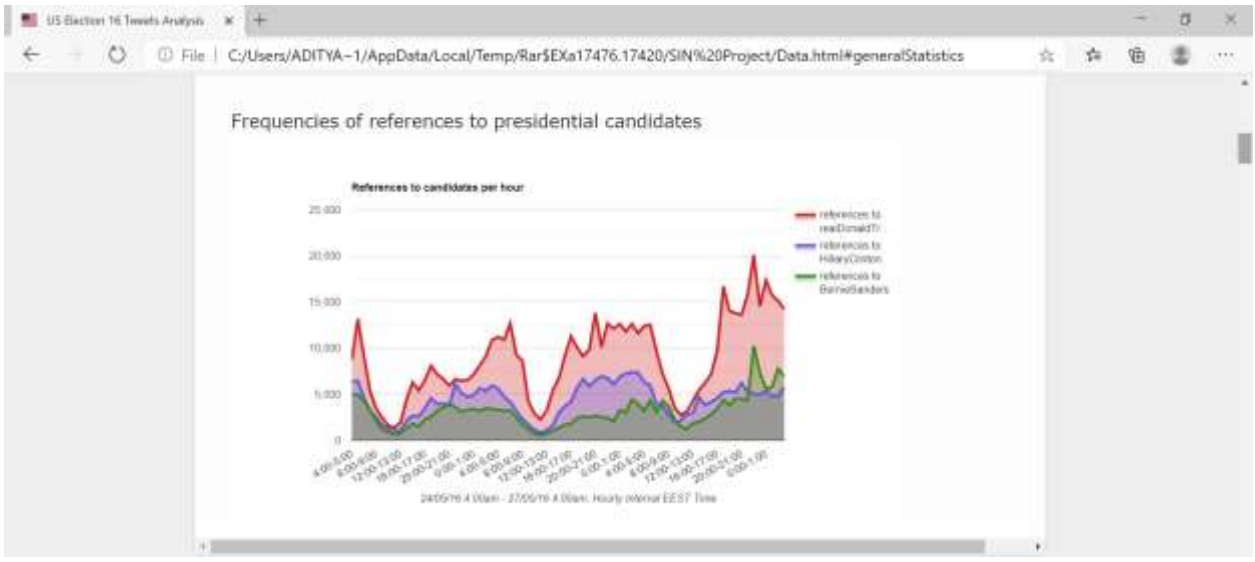
0.05 are ignored. After doing this analysis it is concluded that SMO-RBF kernel with C=1 and Naive Bayes classifiers are best for this data.

Chapter IV Screenshots of Results Obtained

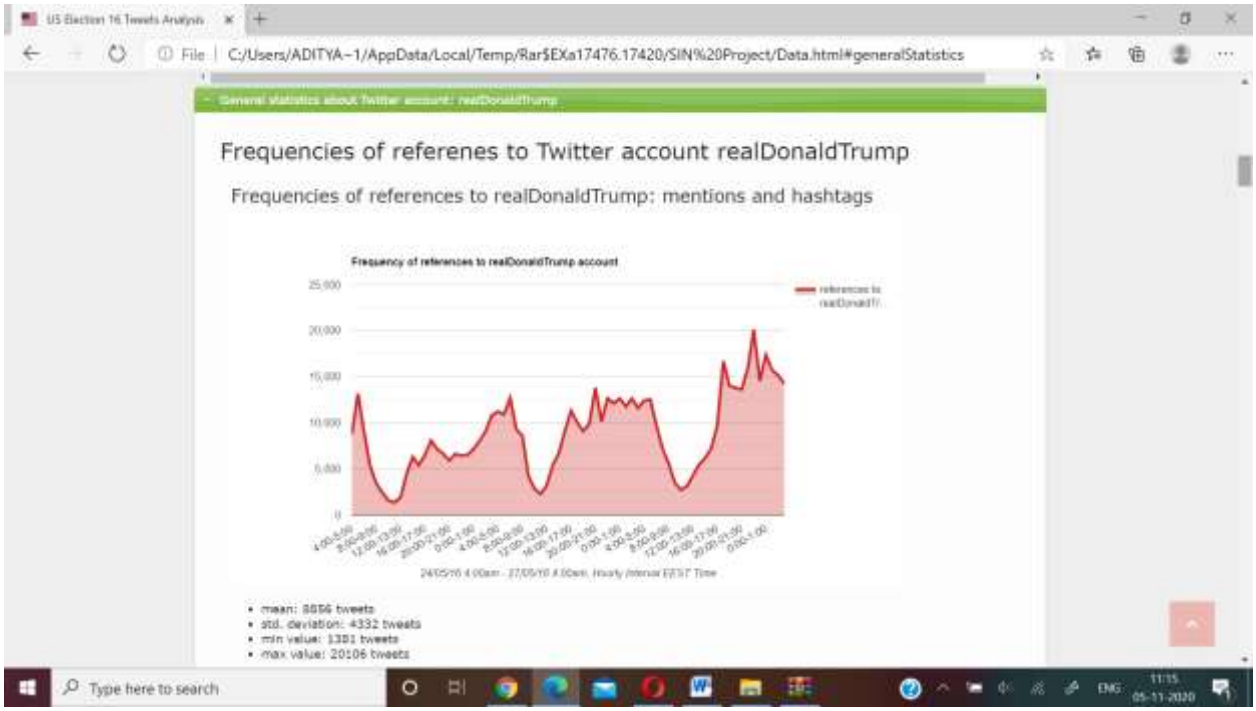
1. Tweets per hour



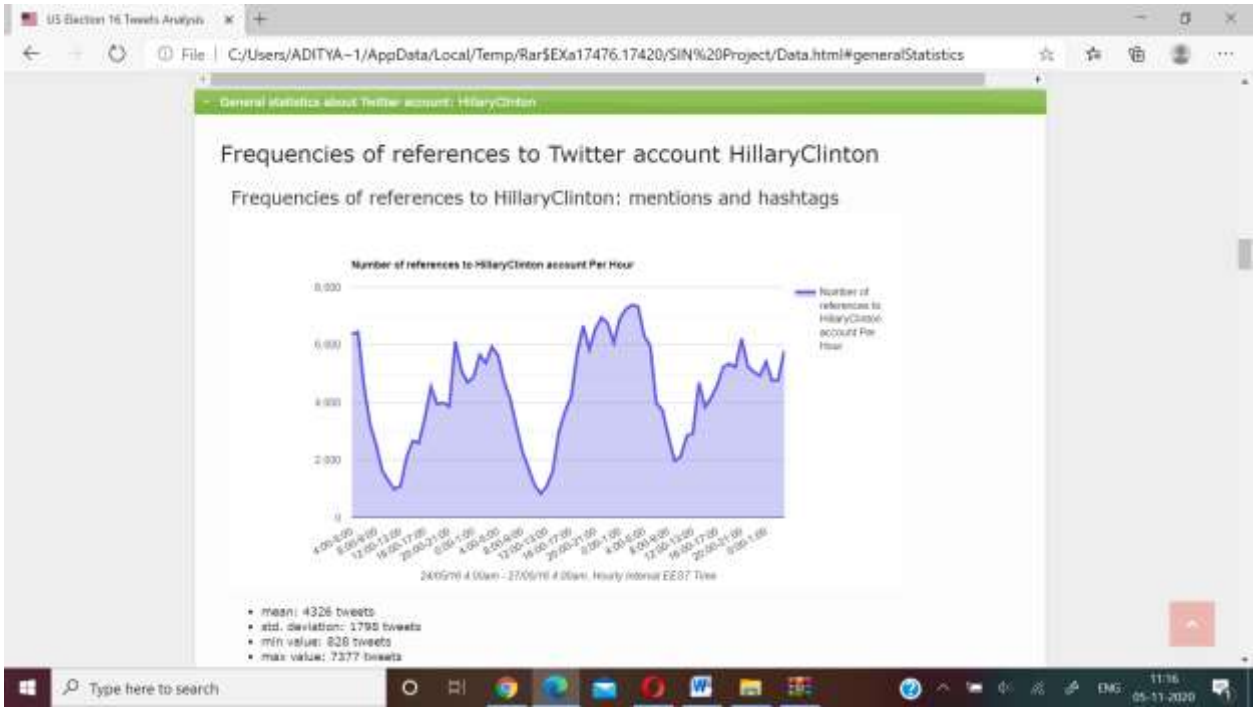
2. Frequencies of references to presidential candidates



2.1 @realdonaldtrump



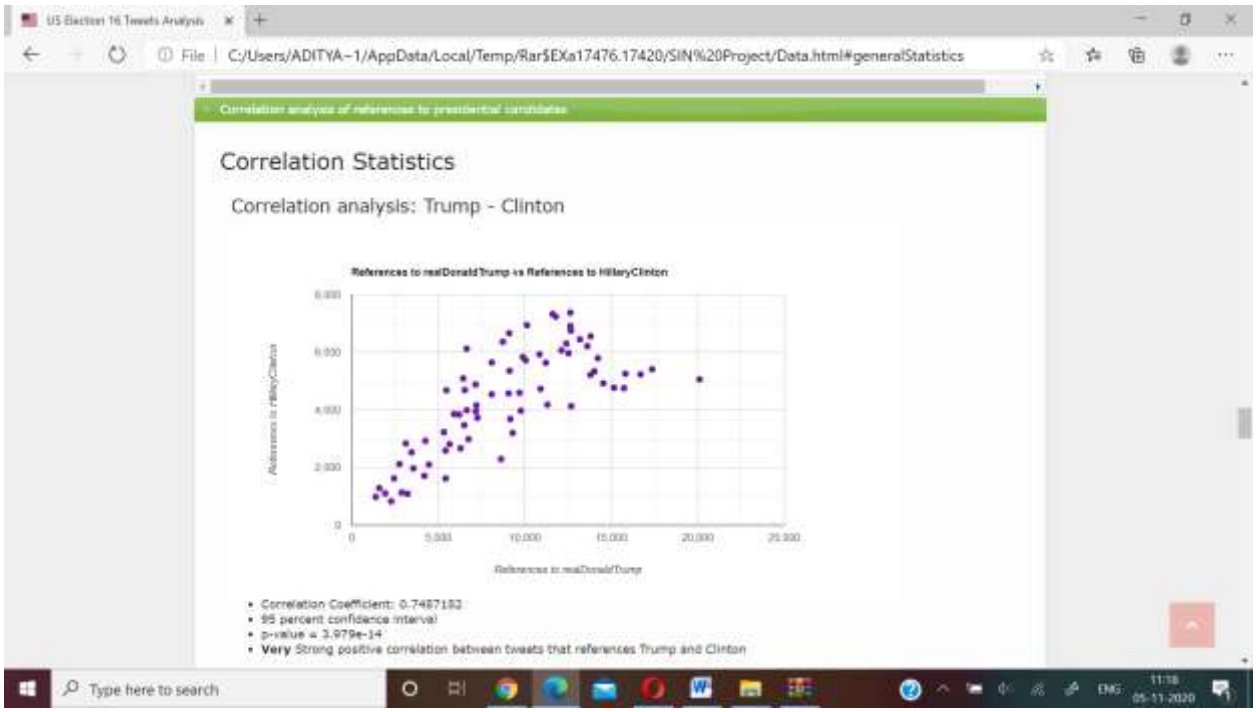
2.2 Hillary Clinton



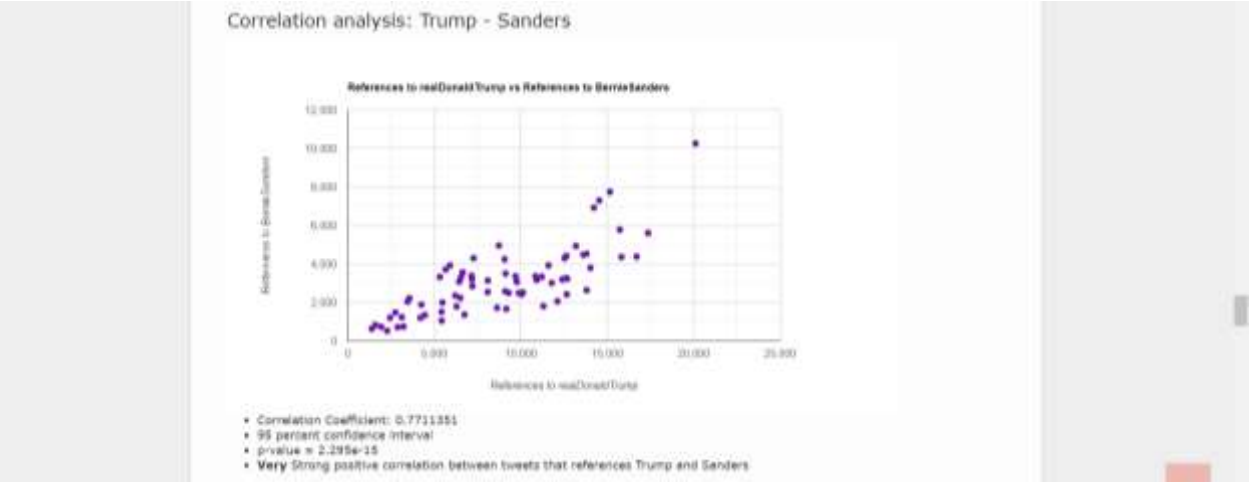
2.3 Bernie Sanders

3. Correlation Statistics

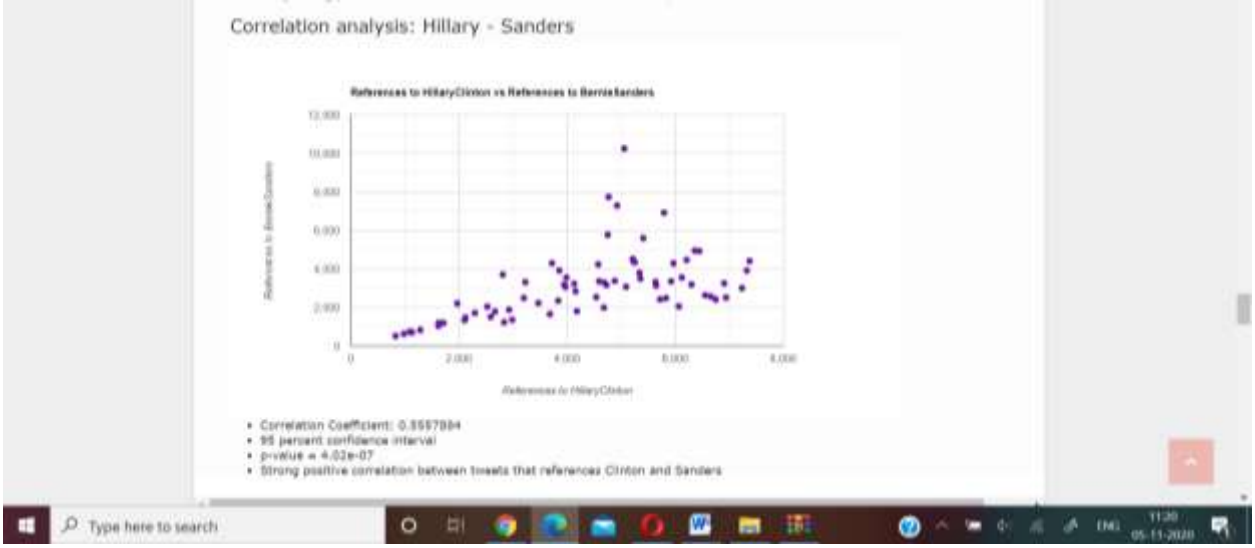
3.1 Trump-Clinton



3.2 Trump- Sanders

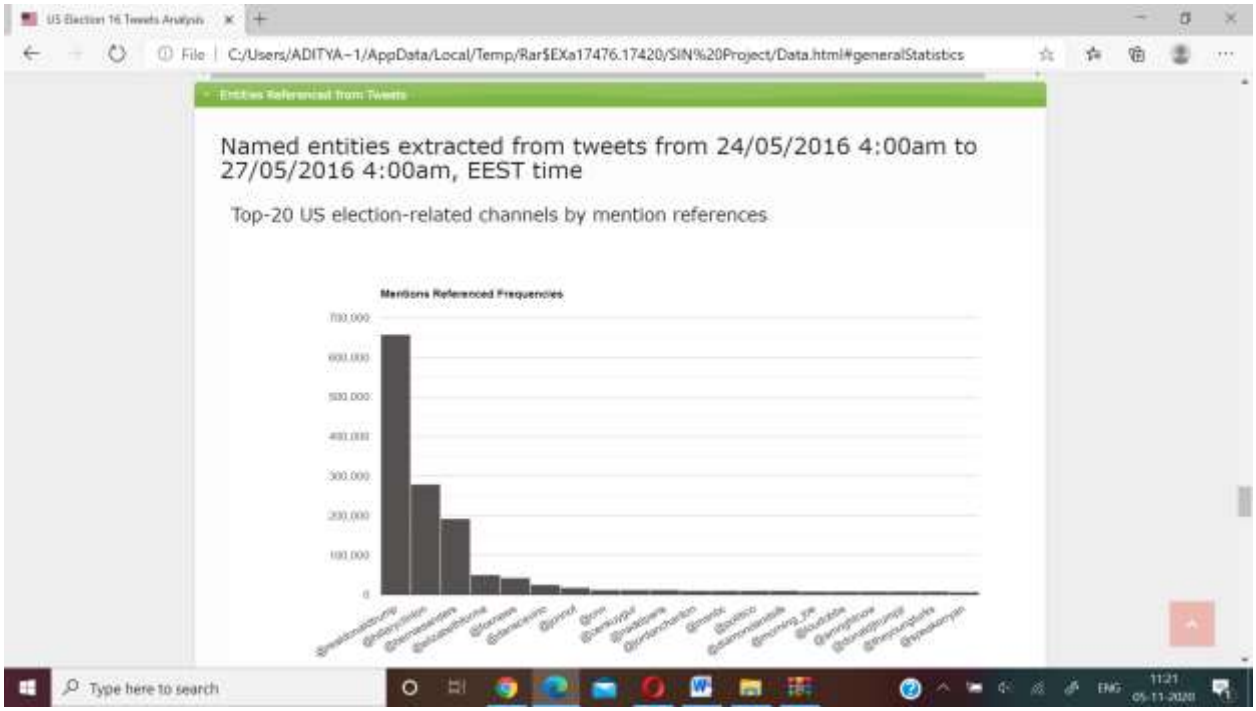


3.3 Hillary- Sanders

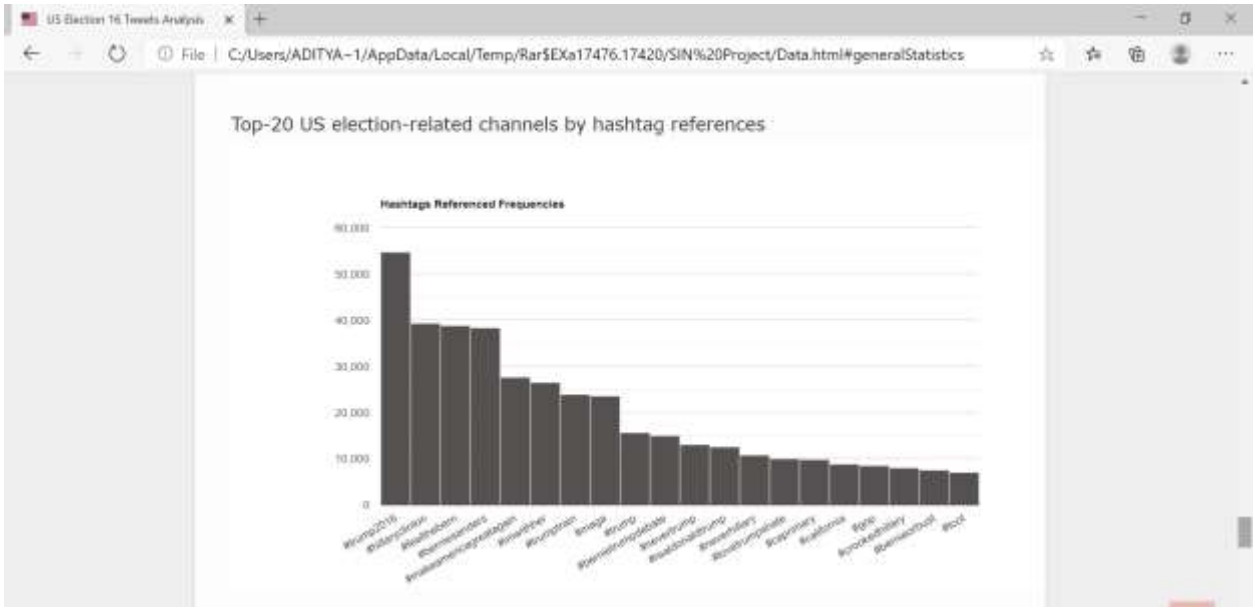


4. Top 20 Statistics

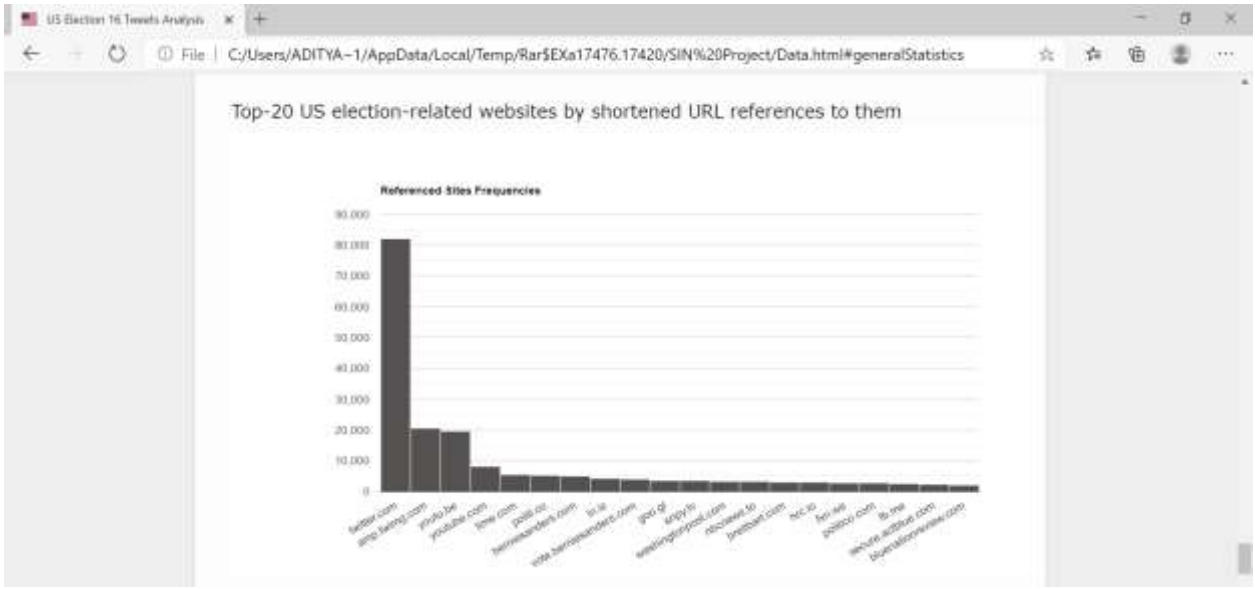
4.1 US election-related channels by mention references



4.2 Election-related channels by hash-tag references

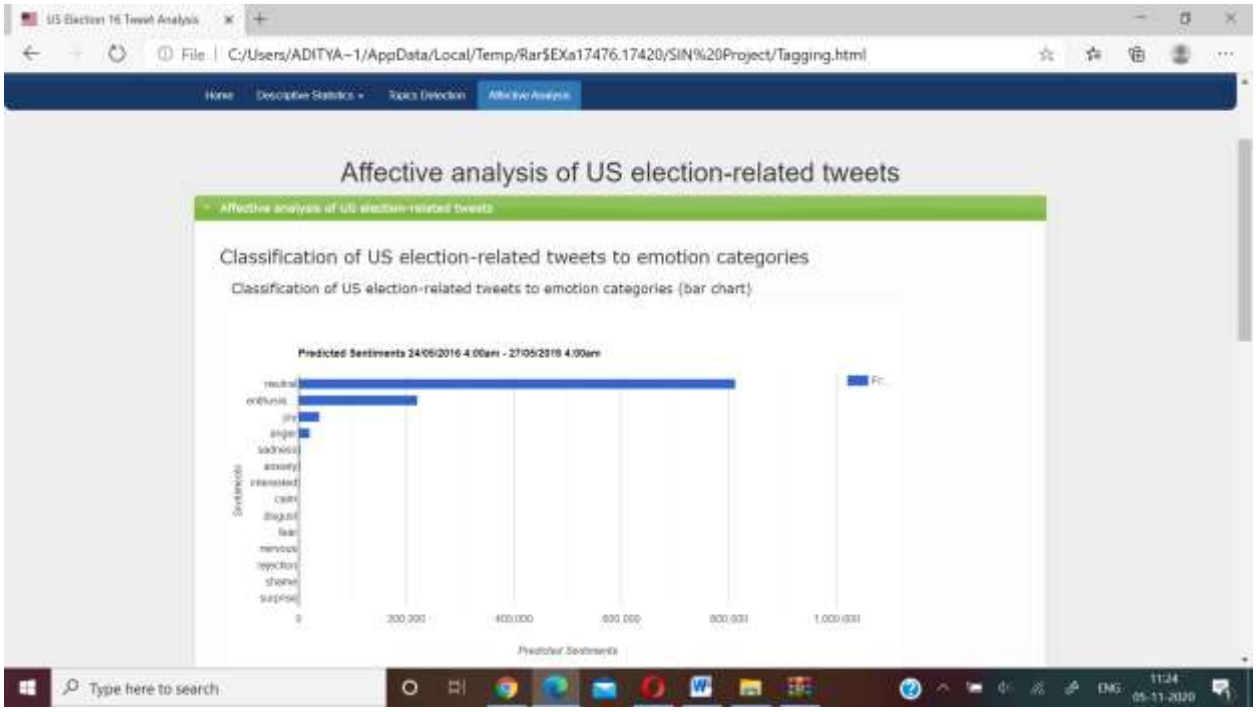


4.3 Top-20 US election-related websites by shortened URL references to them

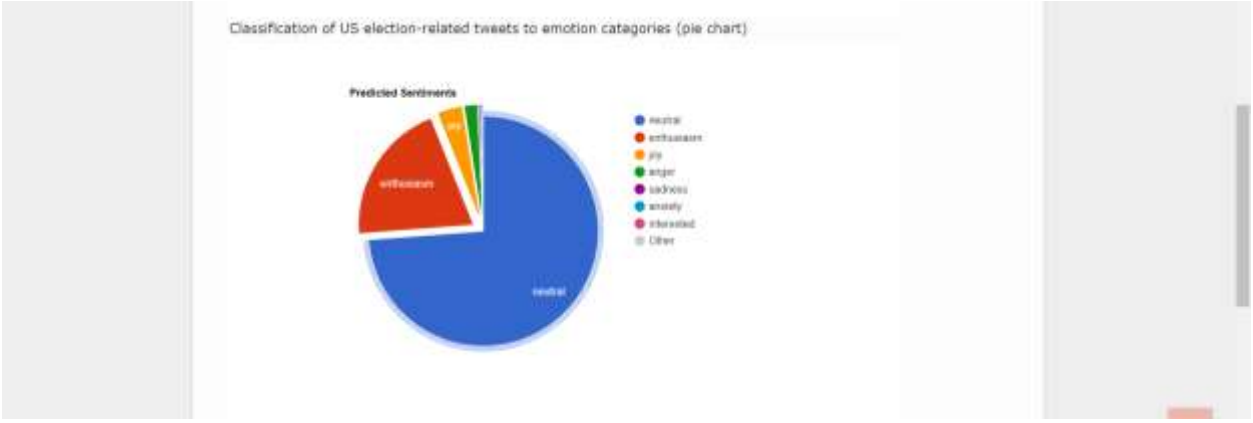


5. Classification of US election-related tweets to emotion categories

5.1 Bar Chart



5.2 Pie Chart


Classification of US election-related tweets to emotion categories (pie chart)
Predicted Sentiments

Conclusion

We can conclude who has the highest possibility of winning the elections by performing sentiment analysis. We, have used name-based classifier and data extraction to pick up references where a particular candidate's name has been mentioned and what kind of review he has been given. Furthermore, the following Statistics are derived using the name-based classifier, these statistics are used to measure how much possibility a candidate has of winning the elections: