ANALYZING THE SENTIMENTS OF TWITTERATI BASED ON THE CURRENT PROCEEDINGS OF U.S. PRESIDENCY

Suchina Parihar
Dept. of Electrical and Computer Engineering
University of Calgary
Calgary, Canada
suchina.parihar2@ucalgay.ca

Abstract— This paper describes a system for analysis of public sentiment expressed on Twitter, a microblogging service. Twitter has become a central site where people express their views on emerging events or news, providing a unique opportunity to predict the relation between expressed public sentiment and Presidential events. Researchers have carried out multiple techniques to extract underlying sentiments from the text. In this paper, I focused on exploratory analysis for figuring out the sentiments of the people in United States post Elections. For this purpose, twitter corpus is collected and linguistic analysis is performed on the collected corpus. Using the training corpus, a Sentiment Classifier is trained, which is able to detect positive, negative, neutral and sarcastic sentiments. A comparative study is also performed using multiple machine learning algorithms for sentiment analysis on tweets. This comparative analysis also shows the best way for selecting the classifier on the basis of accuracy. Different algorithms are used for the confidence level of the result. After labeling the sentiments, an attempt was made to visualize different trends in the sentiments based on all of the findings.

Keywords—sentiment analysis; Naïve Bayes classifier; MNB Classifier; machine learning algorithms

1. INTRODUCTION

In the past few years, there has been a huge growth in the use of microblogging platforms such as Twitter [1]. This is due to nature of microblogs on which people post real time messages about their opinions on variety of topics, discuss current issues, complain, and express positive sentiment [3]. Millions of messages are appearing daily in popular web-sites that provide services for microblogging such as Twitter, Tumblr, Facebook etc. Authors of those messages write about their life, share opinions on variety of topics. Because of a free format of messages and an easy accessibility of microblogging platforms, Internet users tend to shift from traditional communication to microblogging services. As more and more users post about products and services they use, or express their political and religious views, microblogging websites become valuable sources of people's opinions and sentiments. Such data can be efficiently used for marketing or social studies [4].

For my Project, I have used a dataset of messages from Twitter. Twitter contains a very large number of very short length messages created by the users. The contents of the messages vary from personal thoughts to public statements. The dataset I collected, consists of tweets of people on the current proceedings of U.S. Presidency. Then, Sentiment Analysis is performed to extract the underlying sentiments from those tweets.

1.1 SENTIMENT ANALYSIS:

Sentiment analysis refers to the use of natural language processing to systematically identify, extract and quantify subjective information. Sentiment analysis is widely applied to the reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine [3]. Sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction. A basic task in sentiment analysis is classifying the *polarity* of a given text in the document [5].

In this Project, a Corpus based approach is followed where a classifier is trained on a corpus of negative or positive tweets. I have used 4 sets of corpus, Negative, Positive, Neutral and Sarcastic. Various different algorithms are used for Comparative analysis like Naïve Bayes Classifier, MNB Classifier, Linear SVC Classifier and Logistic Regression Classifier.

2. RELATED WORK

With the increase in population of blogs and social networks, opinion mining and sentiment analysis became a field of interest for many researches. Sentiment analysis is a growing area of Natural Language Processing with research ranging from document level classification (Pang and Lee 2008) to learning the polarity of words (e.g., (Hatzivassiloglou and McKeown 1997; Esuli and Sebastiani 2006)) [1]. In (Yang et al., 2007), the authors used web-blogs to construct a corpora for sentiment analysis and emotion icons assigned to blog posts were used as indicators of users' mood. SVM and CRF were applied to learners to classify sentiments at the sentence level and then investigated several strategies to determine the overall sentiment of the document [2].

Some of the recent results on sentiment analysis of Twitter data are by Go et al. (2009), (Bermingham and Smeaton, 2010) and Pak and Paroubek (2010). Go et al. (2009) use distant learning to acquire sentiment data. They use tweets ending in positive emoticons like ":)" ":-)" as positive and negative emoticons like ":(" ":-(" as negative. They build models using Naive Bayes, MaxEnt and Support Vector Machines (SVM), and they report SVM outperforms other classifiers. Pak and Paroubek (2010) collected data following a similar distant learning paradigm. They performed a different classification task though: subjective versus objective. For subjective data they collected the tweets ending with emoticons in the same manner as Go et al. (2009). For objective data they crawled twitter accounts of popular newspapers like "New York Times", "Washington Posts" etc. They reported that POS and bigrams both help. Both these approaches, however, are primarily based on ngram models. Another significant effort for sentiment classification on Twitter data is by Barbosa and Feng (2010). They used polarity predictions from three websites as noisy labels to train a model. Barbosa and Feng used 1000 manually labeled tweets for tuning and another 1000 manually labeled tweets for testing. However, they did not mention how they collected their test data. They proposed the use of syntax features of tweets like retweet, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words [3].

Past studies of political sentiment on social networks have been either post-hoc and/or carried out on small and static samples. To address these issues, I built a unique infrastructure and sentiment model to analyze in real-time (during scraping of data) public sentiment on Twitter about the current proceedings. My effort to gauge political sentiment is based on combining data processing and statistical sentiment modeling.

3. DATA DESCRIPTION

I have collected a data set of approximately 1000 tweets from Twitter using Twitter API. The tweets are stored in json file format. In order to process the data file, I had to create a text corpus. I have even collected some predefined Positive, Negative, Neutral and Sarcastic Twitter training datasets from Github for the purpose of training my classifiers. Table 1 shows the examples of twitter posts.

USERNAME	TWEETS
@kumailn	Kumail Nanjiani: As someonewho was born in Pakistan I can tell you coming into America is VERY difficult. A #Muslimban accomplishes nothing but hate.
@Mikekeen	@realDonalTrump Yes it will! Thank you for fighting for the people!!!!!
@BrandanJR	Welp. Some honesty. Trump thinks anything that disagrees with him is fake news
@NawtyLisanne	AbsolutelyLOVING@TrumpVoterRegrets! I often muse if anyone thinks "Oh shit. I fucked up." and now I know!

Table 1. Examples of Twitter posts

4. DATA PREPROCESSING

The data when scraped from Twitter is raw and we cannot perform any sort of computation on the dataset unless we convert it into machine readable form. Data pre processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent and lacking in certain behaviours or trends, and is likely to contain many errors. Data pre processing is a proven method of resolving such issues [6].

Data preprocessing is one of the important parts of my project because it really helps in increasing the accuracy of my results. Moreover, we know that the text in tweets differ from the text in articles, books or even spoken language as it includes many idiosyncratic uses, such as emoticons, URLs, RT for re-tweet, @ for user mentions, # for hashtags and repetitions. It is necessary to preprocess and normalize the text. There are various preprocessing techniques used in my project.

4.1 TOKENIZATION:

Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing [7]. Tokenization separates the text by counting the white spaces. Whenever a white space occurs, the word gets tokenized. Figure 1 shows an example of how Tokenization works.

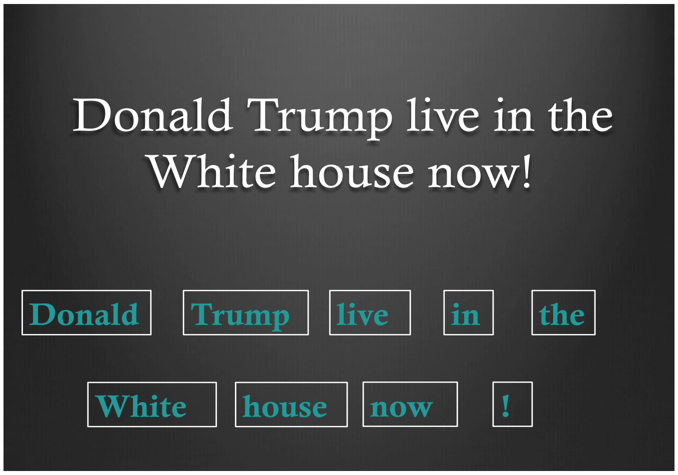

Figure 1. Example of how sentence is Tokenized

4.2 REMOVE STOPWORDS:

Stop words are some of the most common, short function words, such as *the*, *is*, *at*, *which*, and *on*. These words have little meaning in my dataset. It is necessary to remove them as they take up about 10% of the dataset and are just considered noise. Along with stop words, other special characters and symbols are also removed from the text corpus.

4.3 STEMMING:

For grammatical reasons, documents are going to use different forms of a word, such as *organize*, *organizes*, and *organizing*. Additionally, there are families of derivationally related words with similar meanings, such as *final*, finalize and *finally*. In many situations, it seems as if it would be useful for a search for one of these words to return documents that contain another word in the set. *Stemming* usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly, most of the time and often includes the removal of derivational affixes [8].

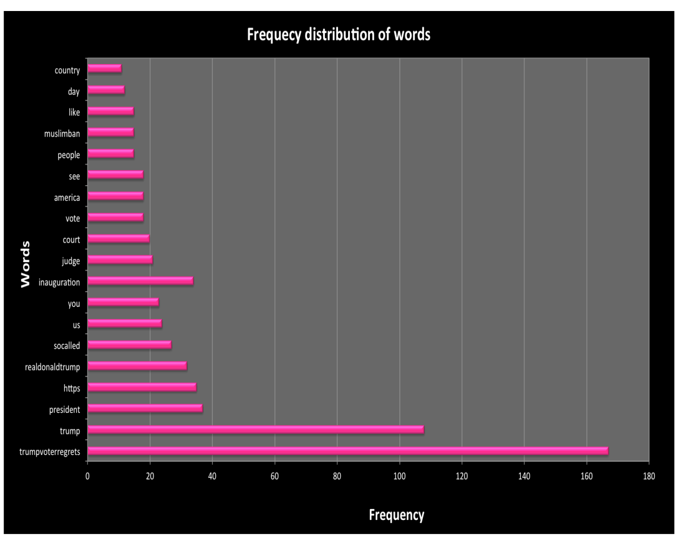

Figure 2. Frequency count of 20 most frequent words

After performing all the above mentioned pre processing steps, now we get a list of words and we can find the most frequent words in the list of words as shown in the Figure 2. Figure 3 and Figure 4 shows the word cloud of frequent words and the clusters of frequent hashtagged words, respectively.

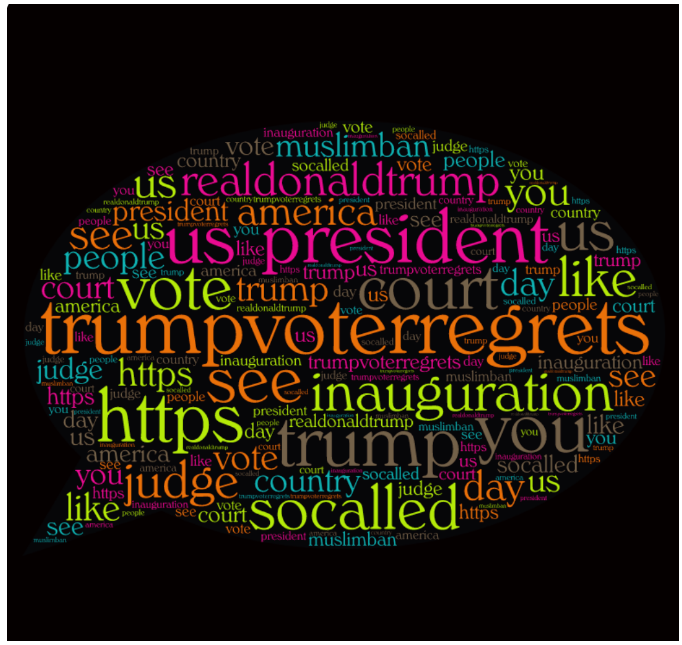

Figure 3. Word Cloud of most frequent words

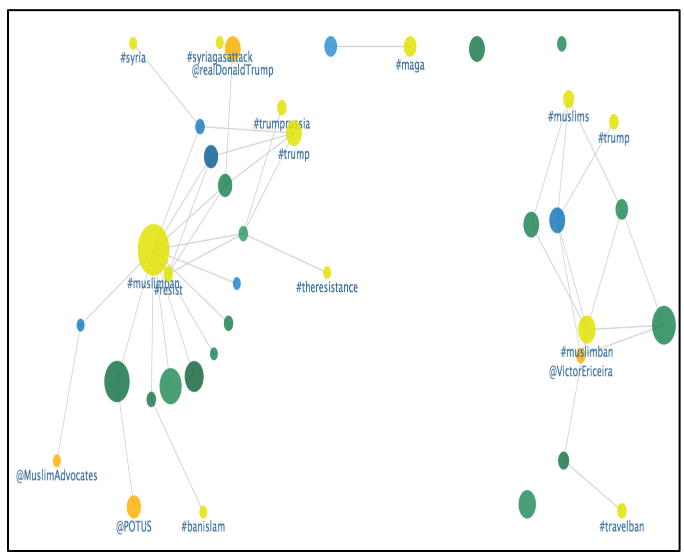

Figure 4. Affinity Graph

4.4 POS TAGGING:

A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word such as noun, verb, adjective etc. It is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context—i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school children, in the identification of words as nouns, verbs, adjectives, adverbs etc [9].

```
[('reality', 'NN')]
[('Redditors', 'NNS')]
[('collecting', 'VBG')]
[('minorities', 'NNS')]
[('camps', 'NNS')]
[('collages', 'NNS')]
[('female', 'NN')]
[('Clinton', 'NN')]
[('voters', 'NNS')]
[('crying', 'VBG')]
[('election', 'NN')]
[('results', 'NNS')]
[('I', 'PRP')]
[('miss', 'NN')]
[('Trump', 'NN')]
[('said', 'VBD')]
[('minorities', 'NNS')]
[('camps', 'NNS')]
```

Figure 5. Screenshot of POS tagged words from my dataset

4.5 CHUNKING:

Chunking is a strategy that breaks down difficult text into more manageable pieces. Dividing content into smaller parts help in identifying key words and ideas and makes it easier to organize and synthesize information [15].

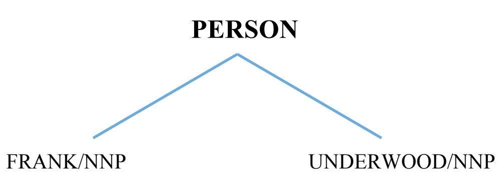

Figure 6.1. Example of Chunking

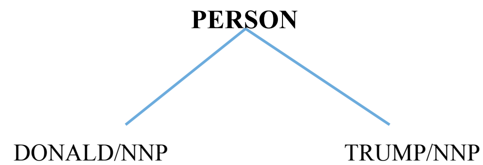

Figure 6.2. Example of Chunking

5. FEATURES

I used a variety of features for the Project. For the baseline, unigrams and bigrams were used. I also included features typically used in sentiment analysis, namely features representing information from a sentiment lexicon and POS features. Finally, features to capture some of the more domain-specific language of microblogging were also included [4].

5.1 FEATURE CATEGORIZATION:

Different types of features, identified from literature review on sentiment analysis are categorized as under:

1 Frequent Features

2 Implicit Features

5.2 FEATURE SELECTION:

Proper feature selection techniques in sentiment analysis have got significant role for identifying relevant attributes and increasing classification accuracy [10]. Feature selection methods are grouped into four main categories NLP or heuristic based, Statistical, Clustering based and Hybrid. In this Project I have used NLP based Feature selection techniques.

Natural language processing (NLP) based techniques mainly operate on two principles:

1. Noun, nounphrases, adjectives, adverbs

2. Terms occurring near subjective expressions can act as features [11].

5.3 FEATURE CLEANSING:

Large numbers of unnecessary features are produced during frequent feature set generation and they need to be removed. Feature cleansing process removes such extra features by applying feature pruning algorithms. Irrelevant features are eliminated by [13, 14] using compactness pruning method. Figure 7 shows the most informative features based on the training datasets.

```
Most Informative Features
                 not = True           sarcas : neutra =    543.8 : 1.0
                more = True           sarcas : pos    =    295.5 : 1.0
               right = True           sarcas : neg    =    178.8 : 1.0
               great = True           sarcas : neutra =    172.1 : 1.0
                many = True           sarcas : pos    =    158.1 : 1.0
                back = True           sarcas : neg    =    135.9 : 1.0
                 old = True           sarcas : neg    =    121.6 : 1.0
               first = True           sarcas : neutra =    117.0 : 1.0
          understand = True           sarcas : pos    =    116.8 : 1.0
               using = True           sarcas : neg    =    107.3 : 1.0
                 new = True           sarcas : neg    =    107.3 : 1.0
              entire = True           sarcas : neutra =    103.2 : 1.0
                used = True           sarcas : pos    =    103.1 : 1.0
                such = True           sarcas : neg    =     93.0 : 1.0
                down = True           sarcas : neutra =     91.4 : 1.0
                mean = True           sarcas : pos    =     89.3 : 1.0
                save = True           sarcas : pos    =     89.3 : 1.0
            terrible = True           sarcas : pos    =     89.3 : 1.0
                 buy = True           sarcas : neg    =     81.5 : 1.0
                last = True           sarcas : neg    =     78.7 : 1.0
                kill = True           sarcas : neutra =     75.7 : 1.0
              enough = True           sarcas : neutra =     75.7 : 1.0
                 bad = True           sarcas : neutra =     73.7 : 1.0
                next = True           sarcas : neg    =     64.4 : 1.0
               quick = True           sarcas : neg    =     64.4 : 1.0
              notice = True           sarcas : neg    =     64.4 : 1.0
                read = True           sarcas : neg    =     64.4 : 1.0
              middle = True           sarcas : neg    =     64.4 : 1.0
                sure = True           sarcas : neg    =     64.4 : 1.0
               enjoy = True           sarcas : neg    =     64.4 : 1.0
```

Figure 7. Screenshot of 30 Most-Informative-Features

6. EXPERIMENTS & RESULT ANALYSIS

The goal of the experiments performed, is to find out the sentiment of the words and then label the sentences with relevant sentiment polarity and finally extract the polarity of the whole dataset. To perform Sentiment Analysis, firstly a Classifier is trained. For this project, I have started with the basic Classifier, Naïve Bayes (NB) Classifier.

6.1 NEGATIVE VS POSITIVE:

For the first part of the Research, the NB Classifier was trained against the Positive and Negative tweets corpus. The testing set was the scraped twitter dataset. Based on the dataset, classifier performs at 69% accuracy on the 2 category classification of negative and positive.

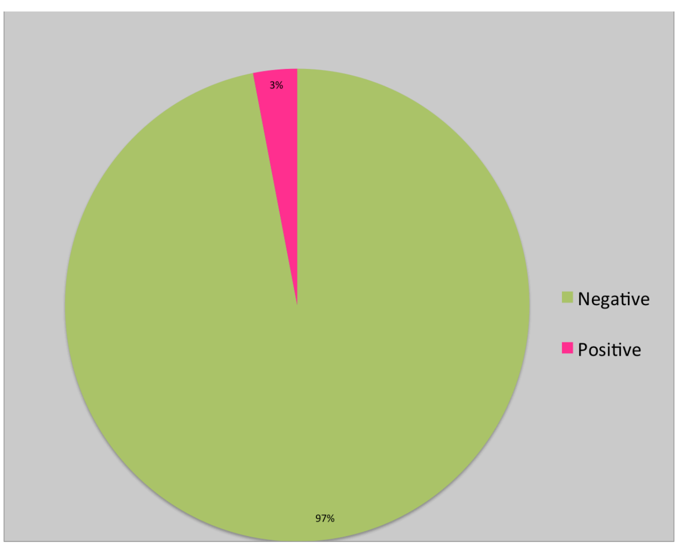

Figure 8. Pie Chart of the Sentiment Score

Figure 8 shows the final Sentiment score of the whole document based on the 2 categories. It clearly shows that there are a large number of negative tweets and thus Trump's Presidency is not successful till now. The results are tested by using 3 other classifiers and their accuracy values are shown in Figure 9.

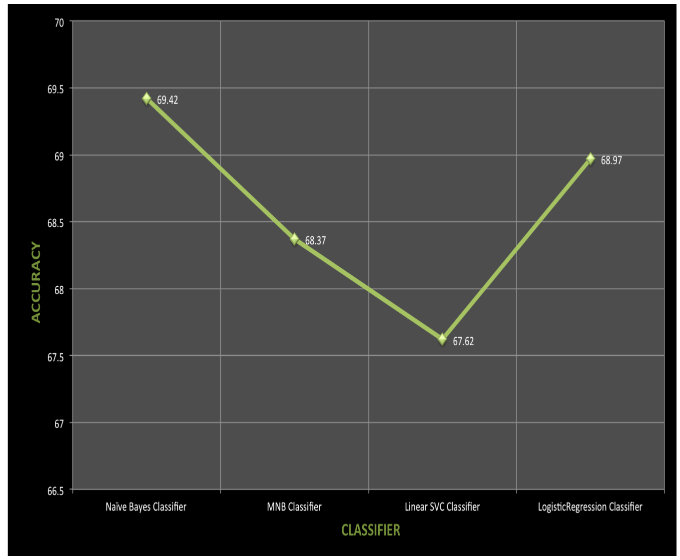

Figure 9. Graph showing the accuracy values of different classifiers

Figure 10 shows the result values for particular tweets. Only a few sentence scores are shown in the picture. The *Sentiment* column refers to the negative or positive sentiment polarity value. The *Confidence* column refers to the confidence percent of the result based on different classifiers.

Tweets	Sentiment	Confidence
https://twitter.com/Sifill_LDF/status/825487287819567104 ƒ" "Donald Trump is an unwell, evil human being.	-0.5	0.55
https:///" " A ""so-called"" President is calling a real President and true leader: bad and sick guy.	-0.22	0.62
can't think of a more fitting way to ring in this day #bye If it's 'America First'...then don't knock 20 million Americans off of Health Insurance.	0.5	0.5
_#trump #trumpvoterregrets #fail" Oh this is perfect!	0.25	0.65
_ https://twitter.com/womensmarch/status/825477468748410880 ƒ I feel sick.	-0.71	0.86
You're all dangerous to us.""	-0.6	0.9
You guys...why did my email count turn to this number this minute right as he is being sworn in?!?!	0.45	0.54
You breed evil here & everywhere."	-1	1
Wonderful speech!	1	1
Will b sending lots of #love&light 2 poor #trumpvoterregrets u will need it ___ politicususa.com/2016/12/13/left-roars-bernie-sanders-mobilizes-opposition-trump-pick-secretary-state.htmlæƒ We are in a lot of trouble with this pick.	-0.3	0.4
What is Twitter is making us all automatically follow Scott Baio and 3 Doors Down where does it end I think maybe Trump did write his own speech.	0.22	0.64
What is Easy D?	0.43	0.83
WTF?	-0.5	1
WTF is this about?	-0.5	1
WORLD shocked!!	-1	0.8
Trump voters - Feeling cheated much?	0.2	0.2
Trump screwed his own base w harmful trade policies.	-0.1	1
Today is a sad day for America.	-0.5	1

Figure 10. Picture showing the Sentiment and Confidence values

Below are Figures 11 and 12, which represent the Sentiment trees of 2 different sentences. The sentences are:

1. @realDonaldTrump Americans are all colors, faiths, cultures & genders. We have voices. We refuse fear. We believe in the Dream. #WeAreHere

2. I have never seen my country on an inauguration day so divided, so anxious, so fearful, so uncertain of its course.

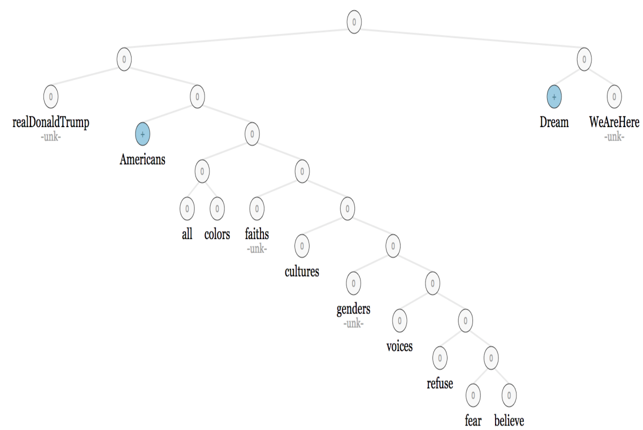

Figure 11. Sentiment Tree for Sentence 1

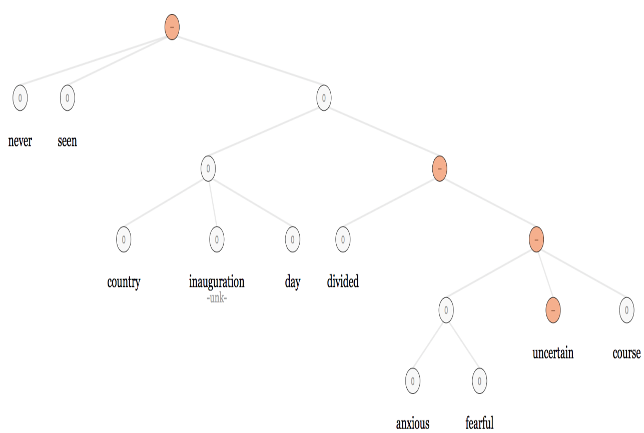

Figure 12. Sentiment tree for Sentence 2

6.2 POSITIVE vs NEGATIVE vs NEUTRAL vs SARCASM:

For the second part of the Research, the Classifiers are trained against 4 different training datasets, negative, positive, sarcastic and neutral. The result of this research is shown in Figure 13. Figure 14 shows the accuracy values of different classifiers.

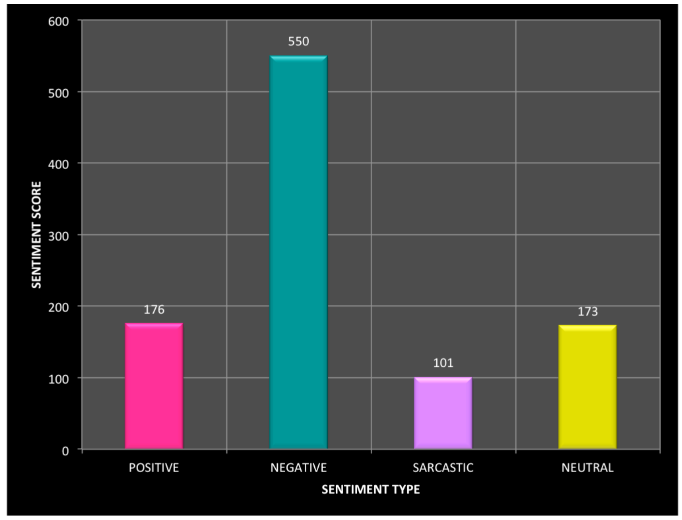

Figure 13. Bar Graph showing the Sentiment Score

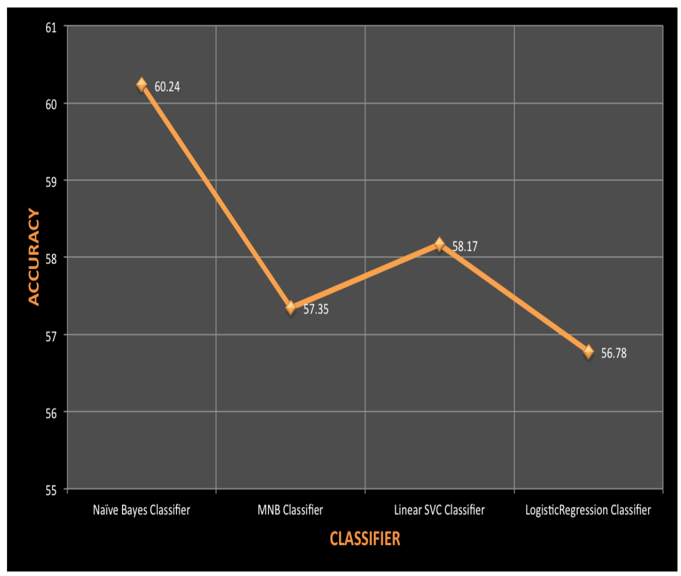

Figure 14. Graph showing the Accuracy values of different Classifiers

7. THREATS TO VALIDITY

1. I trained my classifier on text corpuses downloaded from Github. There might be a slight possibility that those corpuses cannot be trusted. Thus this imposes a threat to the accuracy of the results.

2. The dataset is not so big. This might lead to false positive results and for more accurate results we need a very large dataset.

3. The accuracy percentage of Naïve Bayes Classifier is 60%, which is average and thus can be improved.

8. CONCLUSION

1. The results clearly show that the Negative tweets are larger in number as compared to other categories which means at current stage, people are not happy with the U.S. Presidency.

2. From the experiments performed, it is found that sentiment classifiers are severely dependent on domains and it becomes evident that neither classification model consistently outperforms the other. Thus, even when trained against different training datasets, the results came out different for the same testing set.

3. Experiments also show that part-of-speech features may not be that useful for sentiment analysis in the microblogging domain. For my project, I have used only Adjectives as other POS features did not prove worth enough.

4. It is also found that different types of features and classification algorithms are combined in an efficient way to enhance the sentiment classification performance as more the confidence value, more accurate the result is.

5. In terms of relative performance, Naive Bayes tends to do the best, although the differences are not large. In terms of the overall performance, when the categories were only 2 i.e. when trained against 2 training datasets, the classifiers show higher accuracy as compared to the accuracy when trained against 4 datasets. Therefore, best performance is achieved by using a basic classifier trained to detect just two categories.

10. FUTURE WORK

In future, more work is needed on further improving the performance of the Classifiers. Sentiment analysis can be applied for new applications. Although the techniques and algorithms used for sentiment analysis are advancing fast, however, a lot of problems in this field of study remain unsolved. The main challenging aspects exist in use of other languages, produce a summary of opinions based on product features/attributes, complexity of sentence/ document, handling of implicit product features etc. More future research could be dedicated to these challenges [16]. Moreover, there is a need to research more into detecting Sarcasm. In this Project, I tried to detect sarcasm but the accuracy values were really low and thus proving the need for investing more time and research.

9. REFERENCES

[1] Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. "Twitter sentiment analysis: The good the bad and the omg!." *Icwsm* 11.538-541 (2011): 164.

[2] Wang, Hao, et al. "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle." *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, 2012.

[3] Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." *Proceedings of the workshop on languages in social media*. Association for Computational Linguistics, 2011.

[4] Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." *LREc*. Vol. 10. No. 2010. 2010.

[5] https://en.wikipedia.org/wiki/Sentiment_analysis

[6] https://www.techopedia.com/definition/14650/data-preprocessing

[7] https://en.wikipedia.org/wiki/Tokenization_(lexical_analysis)

[8] https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

[9] https://en.wikipedia.org/wiki/Part-of-speech_tagging

[10] Peter Koncz and Jan Paralic, An approach to feature selection for sentiment analysis,, INES 2011 • 15th International Conference on Intelligent Engineering Systems • June 23–25, 2011, Poprad, Slovakia

[11] S.-M. Kim and E. Hovy, "Automatic identification of pro and con reasons in online reviews," in Proceedings of the COLING/ACL Main Conference Poster Sessions, pp. 483–490, 2006.

[12] Popescu, A.-M. And Etzioni, O. 2005. Extracting product features and opinions from reviews. In HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Morristown, NJ, USA, 339–346

[13] Hu, M., and Liu, B. 2004. Mining Opinion Features in Customer Reviews. AAAI'04, 2004.

[14] D. K. Ly, K. Sugiyama, L. Ziheng, and M.-Y. Kan. Product Review Summarization from a Deeper Perspective. In Proc. of the 11th CM/IEEE Joint Conference on Digital Libraries(JCDL 2011), pages 311–314, 2011.

[15] https://www.facinghistory.org/resource-library/teaching-strategies/chunking

[16] Vinodhini, G., and R. M. Chandrasekaran. "Sentiment analysis and opinion mining: a survey." *International Journal* 2.6 (2012): 282-292.