Depicting the Public Sentiment Variations on Twitter

Shubham Pacharne, Vaibhav Sonawane, Sahil Rajeshirke, Pranav Kolhatkar

***Abstract*— Twitter platform is valuable to follow the public sentiments. Knowing users point of views and reasons behind them at various point is an important study to take certain decisions. Categorization of positive and negative opinions is a process of sentiment analysis. It is very useful for people to find sentiment about the person, product etc. before they actually make opinion about them. In this project, we stream tweets based on a topic and then plot a pie chart, which represents the percentage of positive and negative sentiments in a convenient manner. We also display the number of tweets on which the result is based. Instead of analysing individual sentiment, we stream multiple sentiments and represent percentage of positive and negative sentiments about that topic.**

***Index Terms*— Sentiment analysis, Tokenization, hashtags, Stop words, Part-of-speech tagging, NLTK, Naïve-Bayes classifier**

I. INTRODUCTION

Twitter today has become a very popular communication tool among web savvies. Millions of tweets are appearing daily. Authors of these messages write about their life, share opinions on variety of topics and discuss current issues. As more and more users post about products and services they use, or express their political and religious views, micro-blogging web-sites become valuable sources of people's opinions and sentiments. Such data can be efficiently used for marketing, social studies or improving services. Twitter contains a very large number of very short messages (up to 140 characters) created by the users. The contents of the messages vary from personal thoughts to public statements. Table 1 shows examples of typical posts from Twitter. As the audience of micro-blogging platforms and services grows every day, data from these sources can be used in sentiment analysis tasks. For example, restaurants may be interested in the following questions:

- What do people think about us (food, service etc.)?
- How positive (or negative) are people about our food-items?

Political parties may be interested to know if people support their program or not. News channels may ask people's opinion on current debates (News hour).

Shubham Pacharne, Department of Computer Engineering, P.E.S's Modern College of Engineering, Shivaji Nagar, Pune 411005, India.

Vaibhav Sonawane, Department of Computer Engineering, P.E.S's Modern College of Engineering, Shivaji Nagar, Pune 411005, India.

Sahil Rajeshirke, Department of Computer Engineering, P.E.S's Modern College of Engineering, Shivaji Nagar, Pune 411005, India.

Pranav Kolhatkar, Department of Computer Engineering, P.E.S's Modern College of Engineering, Shivaji Nagar, Pune 411005, India

Hashtags are words prefixed with "#" and are used to indicate the topics of tweets. For example, "#Election2014" can be used in tweets related to India's General Election of

2014. Hashtags play an important role in Twitter. Popular hashtags can become trending topics in the home page of Twitter.

TABLE I

EXAMPLES OF TWITTER POSTS WITH USER'S OPINIONS

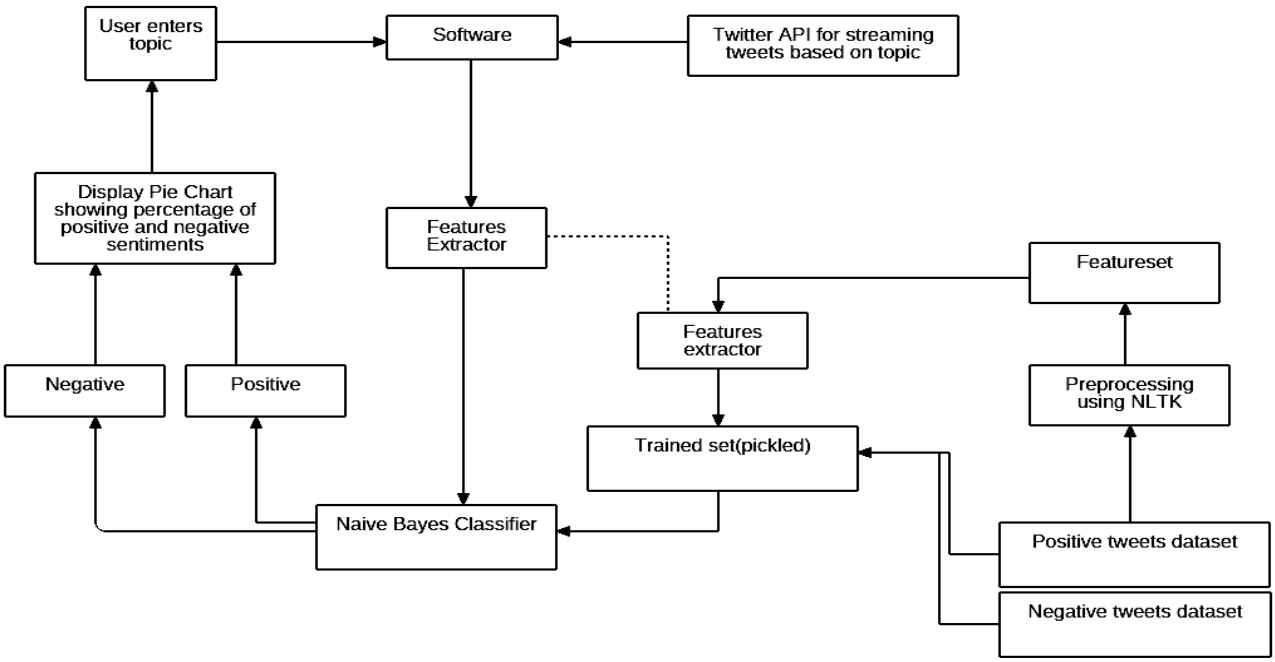
Shubham Pacharne: Python is the best programming language. http://bit.ly/1Ro76HT
Sundar Pichai: Tremendous excitement among Googlers for PM Modi's visit. #ModiInUSA
Darklightnjh: I hate how when im watching a YouTube video the ad is all hq but when the real video starts its DSI camera quality.
Semantria: "I love the summer in New York, but I hate the winter."
Curiosity Rover: Namaste, @MarsOrbiter! Congratulations to @ISRO and India's first interplanetary mission upon achieving Mars orbit.

II. LITERATURE SURVEY

Sentiment analysis is the process of analysing the opinions which are extracted from different sources like the comments given on forums, reviews about products, various policies and the topics mostly associated with social networking sites and tweets. A very broad overview of the existing work was presented in (Pang and Lee, 2008). In their survey, the authors describe existing techniques and approaches for an opinion oriented information retrieval. However, not many researches in opinion mining considered blogs and even much less addressed micro-blogging. In (Yang et al., 2007), the authors use web-blogs to construct corpora for sentiment analysis. The authors applied SVM and CRF learners to classify sentiments at the sentence level and then investigated several strategies to determine the overall sentiment of the document. As the result, the winning strategy is defined by considering the sentiment of the last sentence of the document as the sentiment at the document level.

In (Go et al., 2009), authors used Twitter to collect training data and then to perform a sentiment search. The approach is similar to (Read, 2005). The authors construct corpora by using emoticons to obtain "positive" and "negative" samples, and then use various classifiers. The best result was obtained by the Naive Bayes classifier with a mutual information measure for feature selection. The authors were able to obtain up to 81% of accuracy on their test set. However, the method showed a bad performance with three classes ("negative", "positive" and "neutral").

III. ARCHITECTURE DIAGRAM



IV. SENTIMENT TRACKING

- We have 2 datasets – one containing 5000+ positive sentences and other containing 5000+ negative sentences

- Pre-processing tasks (tokenization, stop word removal, part-of-speech tagging) are done using nltk on each sentence of the datasets.

(1) Tokenization – We segment text by splitting it by spaces and punctuation marks, and form a bag of words. (2) Stop word removal – Stop words are natural language words which have very little meaning, such as "and", "the", "a", "an", and similar words. (3) Part-of-speech tagging – It is a sentence-based process and given a sentence formed of a sequence of words, part-of-speech tagging tries to label (tag) each word with its correct part-of-speech. Therefore, pre-processing techniques on tweets are necessary for obtaining satisfactory results on sentiment analysis

- We create a word_features list from which only the features which appear most frequently are extracted.

- This list is used to create feature set list which contains all the words from the two datasets and a Boolean value indicating whether or not that particular word is from the most commonly occurring words list.

- This feature set randomly shuffled and split into 2 sets- majority of it in training set and remaining in testing set.

- The training set is trained against by the Naïve Bayes classifier and the testing set is tested against for accuracy. When user enters topic, streaming tweets are fetched from the Twitter API about that topic.

- Each tweet is tokenized and a list is created which contains the words followed by the Boolean value indicating whether or not the words are present in the word_features list. Using the Naïve Bayes classifier, the tweet is classified as positive or negative, depending on whether the words in the tweet appear more frequently in the positive category or the negative category.

- A Pie Chart is generated showing the percentage of positive and negative tweets about that topic out of the total number of tweets.

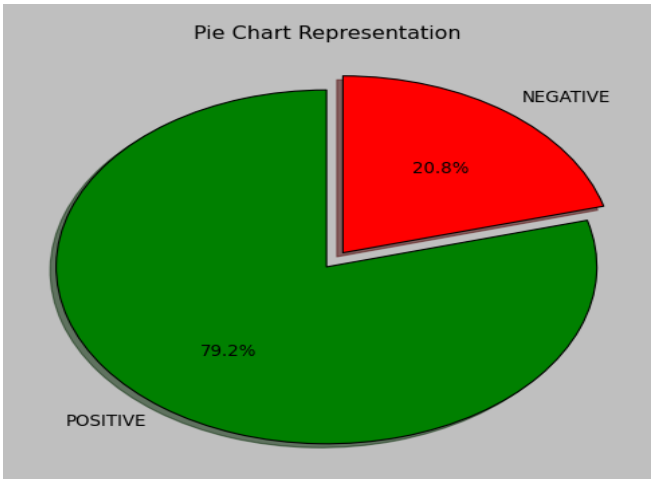
- The number of tweets based on which the pie chart is generated is also displayed.

V. RESULTS

In this section, we can segregate the reviews based on different categories and generate graphical representation of the sentiments in the form of pie-charts. The user will enter the topic and press the button “Get Streaming Twitter Data”. The software will start getting the streaming Twitter data based on topic from the Twitter API. Then user will click on “Generate Pie Chart”.

1. Pie-Chart representation

Software will generate a pie-chart showing the percentage of positive and negative sentiments based on the user's entered topic.



2. Number of tweets based on which Pie-Chart is plotted

```
/usr/bin/python3.4 /root/main_proj/pie.py
This is based on 53 tweets
```

VI. Conclusions

Nowadays, the opinions and reviews of people on social networking sites like Twitter are hugely influential on other people and their decisions. This system can help such people find about sentiments about a topic on Twitter and make necessary decisions conveniently.

Despite all the challenges and potential problems that threatens Sentiment analysis, one cannot ignore the value that it adds to the industry. Because Sentiment analysis bases its results on factors that are so inherently humane, it is bound to become one the major drivers of many business decisions in future. Improved accuracy and consistency in text mining techniques can help overcome some current problems faced in Sentiment analysis.

Looking ahead, what we can see is a true social democracy that will be created using Sentiment analysis, where we can harness the wisdom of the crowd rather than a select few experts. A democracy where every opinion counts and every sentiment affects decision making.

Acknowledgment

We take this opportunity to express our profound gratitude and deep regards to our mentor Mr.Pradeep Pattayat (QE Lead, Persistent Systems) for his exemplary guidance, monitoring and constant encouragement throughout the course of this thesis.

We also take this opportunity to express a deep sense of gratitude to Dr. B. D. Phulpagar, Mr. Vijeth Rao for their cordial support, valuable information and guidance, which helped us in completing this task through various stages.

References

[1] Alexander Pak, Patrick Paroubek, *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*, F-91405 Orsay Cedex, France.

[2] D. Chakrabarti and K. Punera. Event summarization using tweets. In *Proc. of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Catalonia, Spain, 2011.

[3] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Found. Trends Inform. Retrieval, vol. 2, no. (12), pp. 1135, 2008.

[4] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proc. 10th ACM SIGKDD, Washington, DC, USA, 2004.

[5] D. Tao, X. Tang, X. Li, and X. Wu,"Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," IEEE Trans. Patt.Anal. Mach. Intell., vol. 28, no. 7, pp.10881099, Jul. 2006

[6] O. Tsur and A. Rappoport, "What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities," in *WSDM*, 2012,
pp. 643–652.

[7] A. Cui, M. Zhang, Y. Liu, S. Ma, and K. Zhang, "Discover breaking events with popular hashtags in twitter," in *CIKM*, 2012,
pp.1794-1798.

[8] H. Kwak, C. Lee, H. Park and S. B. Moon, "What is twitter, a social network or a news media?" in *WWW*, 2010, pp 591-600