ISSN:0975-9646

Location Based Opinion Mining of Real Time Twitter Data

Gargi Mishra
M.Tech. (CT) Scholar,
Department of CSE
RCET, Bhilai,Chattisgarh

Shivani Varshney,
Assistant Professor,
Department of CS
RCET, Bhilai,Chattisgarh

***Abstract* – Social Networking sites provides tremendous impetus for giant data in mining people's opinion. To recognize people's opinion, tweets are labeled into positive, negative or neutral indicators. This paper provides an effective mechanism to perform opinion mining from real time knowledge from Twitter. In this competitive world of business, politics and media we need to seek out what is best and fit to use and is entertaining, to search out we've designed a process that extracts data from twitter (which is growing and widespread social networking sites with has over 310 million monthly active users as on March 2016) which is crawl real time data from twitter using ASP.net. Initial part of the research concerns creeping knowledge from Twitter and store them into database. Second a part of the research is classifying the text to three major classification of positive negative or neutral. We'll apply our sentiment analysis and opinion mining on tweets based on them. The additional to the sentiment analysis are going to be classification of the location in line with the user tweet or user's current location showing the number of tweets from the location and percentage of sentiments in positive, negative and neutral. We have a database of words with more than 10k words and phrases to classify data according to their sentiment.**

***Index Terms* – sentiment; opinion mining; Web mining; web mining software; big data analytics; social networking**

I. INTRODUCTION

In this competitive world where we want to keep updated with latest and best available products , news , entertainment etc. we want to seek out what's best and suitable for use, is trending , is entertaining. To search out we've designed a process that extracts knowledge from twitter (which is growing and widespread social networking sites with has over 310 million monthly active users as on March 2016) which is used to crawl real time knowledge from twitter using ASP.net and so we'll apply our sentiment analysis and opinion mining on tweets. The additional to the sentiment analysis we are going to classify tweets according to the tweet location or user's current location, also showing the number of tweets for the location and percentage of sentiments in positive, negative and neutral. We have a database of words with more than 10k words and phrases to classify data according to their sentiment.

The execution of opinion examination strategies over Twitter is a small arrangement of assessment data sets .The paper evaluating assumption investigation at target (element) level, is the absence of particular slant comments among the tweets and the substances contained in them. This paper likewise gives a relative investigation of the different datasets along a few measurements including: downright number of tweets, vocabulary size.[1]

A framework for continuous investigation of open estimation toward presidential applicants in the 2012 U.S. race as communicated on Twitter, a smaller scale blogging administration. It offers general society, the media, government officials and researchers another and opportune viewpoint on the flow of the constituent procedure and popular conclusion. [2].

The paper connects measures of popular opinion measured from polls with sentiment measured from text. The paper analyse many surveys on consumer confidence and political opinion over the 2008 to 2009 period, and notice they correlate to sentiment word frequencies in contemporaneous Twitter messages. Whereas our results vary across datasets, in many cases the correlations are as high as 80% and capture necessary large-scale trends. The results highlight the potential of text streams as a substitute and supplement for ancient polling. The results recommend that additional advanced natural language processing techniques to enhance opinion estimation may be very helpful. During this paper the polls are treated as a gold customary .Many techniques from ancient survey methodology can even be used once more for automatic opinion measuring. Eventually, we have a tendency to see this analysis attending to align with the additional general goal of query-driven sentiment analysis wherever one will raise a lot of varied queries of what individuals are thinking based on text they are already writing. [3]

In this paper, we present a novel methodology of including semantics as extra components into the preparation set for opinion examination. This paper applies this way to deal with foresee opinion for three distinctive Twitter datasets. This paper found that best results are accomplished while adding the generative model of words given semantic ideas into the unigram dialect model of the NB classifier. Additionally directed broad tests on three Twitter datasets and contrasted the semantic components and the Unigrams and POS arrangement highlights and also with the notion subject elements. The outcomes demonstrate that the semantic component model beats the Unigram and POS benchmark for recognizing both negative and positive assessment. The outcomes demonstrates that the semantic methodology is more fitting when the datasets being broke down are extensive and spread an extensive variety of subjects, while the slant theme methodology was most suitable for moderately little datasets with particular topical foci. [4].

This paper assesses the value of existing lexical assets and also includes that catch data about the casual and imaginative dialect utilized as a part of micro blogging. We take a managed way to deal with the issue, yet influence existing hash tags in the Twitter information for building preparing information. This investigation on twitter opinion examination demonstrates that grammatical feature elements may not be valuable for estimation investigation in the micro blogging space. Highlights from a current estimation dictionary were to some degree helpful in conjunction with micro blogging highlights; however, the micro blogging highlights unmistakably the most valuable. Utilizing hash tags to gather preparing information proved valuable, as did utilizing information gathered in view of positive and negative emotions. [5].

In this paper, we demonstrate to naturally gather a corpus for assumption investigation and sentiment mining purposes. The exploration performs etymological investigation of the gathered

corpus and clarifies found marvels. Utilizing the corpus, we fabricate an opinion classifier that can decide positive, negative and unbiased suppositions for a record. Trial assessments demonstrate that our proposed systems are proficient and performs superior to anything beforehand proposed strategies. In this exploration, we worked with English; be that as it may, the proposed system can be utilized with some other dialect. [6].

Also two primary issues about big data: (1) big data as a sample, and (2) a particular type of big data— that is, social media data. [7]

We discuss the current and future trends of mining evolving data streams and the challenges that the field will have to overcome during the next years. [8].

We also see that large amount of information on web platforms make them viable for use as data sources, in applications based on opinion mining and sentiment analysis. The algorithm for detecting sentiments on movie user reviews, based on naive Bayes classifier. This make an analysis of the opinion mining domain, techniques used in sentiment analysis and its applicability. [9]

Earlier studies suggest that target-subordinate Twitter opinion arrangement; in particular, we group the suppositions of the tweets as positive, negative or nonpartisan as indicated by whether they contain positive, negative or impartial slants about that inquiry. [11].

We looked at two techniques for time arrangement determining. The main strategy utilized a Nonlinear Autoregressive Network with exogenous inputs (NARX) and RCDESIGN. By the by RCDESIGN had yielded optimizing and astounding execution in some benchmark issues.[13].

Discussing about different strategies for preparing a repetitive neural network [14].

The paper is to bring in light the value of Web Content Mining. The paper gives us knowledge into its techniques and processes. Web content mining is an integral role by getting rich set of contents and uses those contents in the decision making in the industry, education sector and research field.[15]

With millions of users tweeting around the world, real time search systems and different types of mining tools are emerging to allow people tracking the repercussion of events and news on Twitter. However, although appealing as mechanisms to ease the spread of news and allow users to discuss events and post their status, these services open opportunities for new forms of spam. Trending topics, the most talked about items on Twitter at a given point in time, have been seen as an opportunity to generate traffic and revenue. Spammers post tweets containing typical words of a trending topic and URLs, usually obfuscated by URL shortness, that lead users to completely unrelated websites. This kind of spam can contribute to de-value real time search services unless mechanisms to fight and stop spammers can be found. In this paper we consider the problem of detecting spammers on Twitter. We first collected a large dataset of Twitter that includes more than 54 million users, 1.9 billion links, and almost 1.8 billion tweets. Using tweets related to three famous trending topics from 2009, we construct a large labelled collection of users, manually classified into spammers and non-spammers. [16]

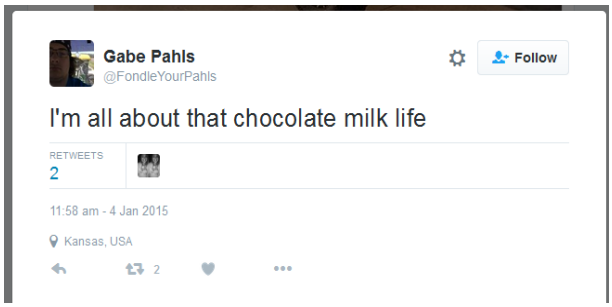
II. PROBLEM STATEMENT

To recognize people's opinion, tweets are labeled into positive, negative or neutral indicators. This paper provides an effective mechanism to perform opinion mining from real time knowledge from Twitter. In this competitive world of business, politics and media we need to seek out what is best and fit to use and is entertaining, to search out we've designed a process that extracts data from twitter (which is growing and widespread social networking sites with has over 310 million monthly active users as on March 2016) which is crawl real time data from twitter using ASP.net. Initial part of the research concerns creeping knowledge from Twitter and store them into database. Second a part of the research is classifying the text to three major classification of positive negative or neutral. We'll apply our sentiment analysis and opinion mining on tweets based on them. The additional to the sentiment analysis are going to be classification of the location in line with the user tweet or user's current location showing the number of tweets from the location and percentage of sentiments in positive, negative and neutral. We have a database of words with more than 10k words and phrases to classify data according to their sentiment.

III. BACKGROUND

CRAWLING

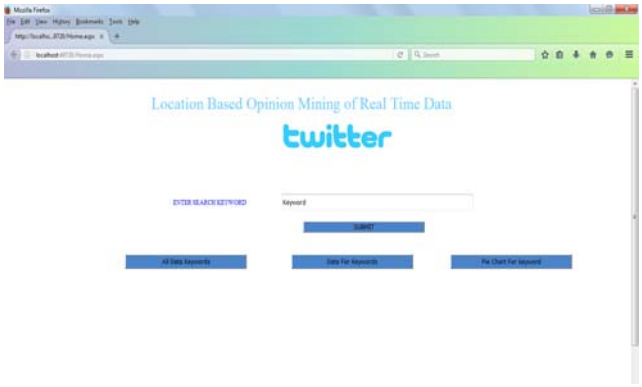
A web crawler is a system that consistently browses on the World Wide Web, generally is used for the purpose of web classification. Web search engines and a few alternative sites use internet crawl to update their web content or indexes of varied alternative available from obtainable sites web page. Web crawlers extract information from all the pages they visit for later utilized by a search engine that indexes and relates the extracted pages therefore the users will search rather more with efficiency and precisely. Twitter is that the most generally and popularly used social media platform for expressing emotions for applications starting from whole observation to complete quality, sentiment analysis, and client pattern researches to guide generation. If you're a Twitter user, this offers helps you gather all the Twitter data you wish to go looking in close to period of time. Twitter is mine of information. We are able to gather enough data from tweet.



Twitter let us share our views in 165 characters. Also we can add smileys, pictures and location to the tweet. Also People can share your tweets and comment on your tweet. If people like your tweets they can also like your tweet. In the above picture we see that tweet has following information

1) Tweet Text
2) Location
3) Date
4) No of People Favourite tweet
5) No of People Re-tweeted Tweet
6) Username
7) User-id

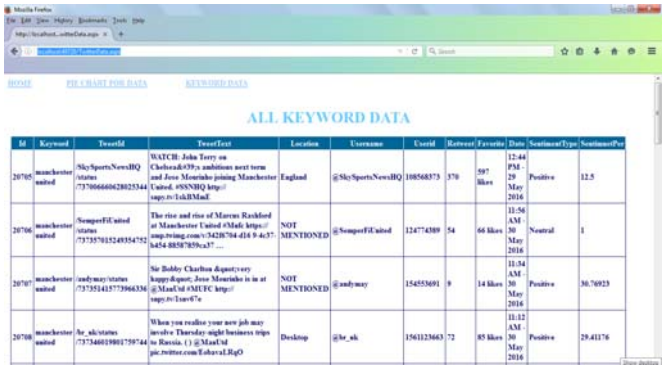
The user enters a keyword and clicks submit. As we click submit the crawler starts working in the backend without hanging the GUI of the website. The crawler then continuously crawls for tweets with keyword in the textbox and stores the detail in the database.



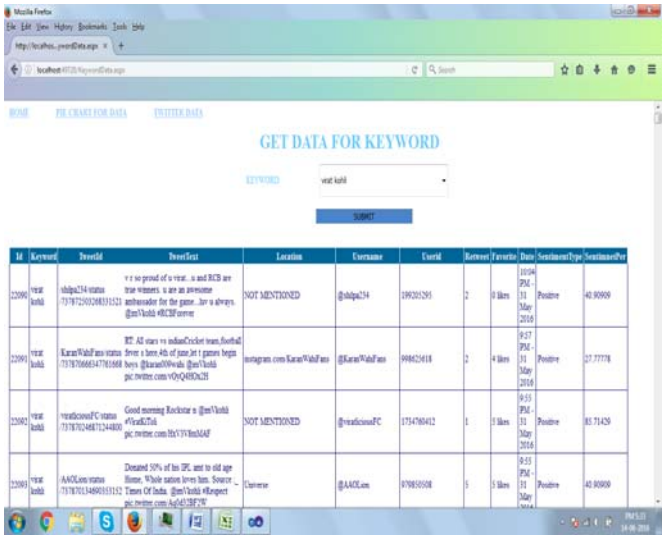
To Stop GUI from hanging we have added multithreading code so that the crawling keeps going on at the backend and we can check the crawled data from the "All Data Keyword" and "Data for Keywords".

```
new Thread(() =>
                {
                    GetData();
                }).Start();
```

We can view all the crawled data for all entered keywords from "All Data Keyword".



We can also view data for keywords individually also by selecting the data for keywords from "Data for Keywords".



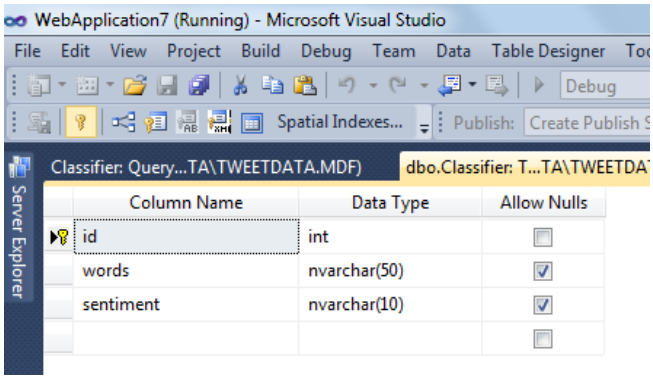
Opinion Mining:-

Opinion mining, which is also known as sentiment analysis, involves building an automated software to collect and categorize opinions about a product, political opinion , research projects , entertainment etc. Automated opinion mining often uses machine learning, a type of artificial intelligence (AI), to analyze the sentiment of data.

Opinion mining can be useful in several ways. It can help marketers evaluate the effectiveness of an ad campaign or launching the new product, determine which type of a product or service are popular and analyzing which part of the product is like or dislike particular product features. For example, a review on a website might be broadly positive about a digital camera, but be specifically negative about how heavy it is. Being able to identify this kind of information in a systematic way gives the vendor a much clearer picture of public opinion than surveys or focus groups do, because the data is created by the customer.

The Steps to be followed in Opinion Mining Phase will be

i. **Dataset: -** Loading the data set of words with classified positive, negative and neutral keywords to be used to perform sentiment analysis.



ii. **Pre-processing:-** Pre-processing of data is the process of preparing and cleaning the data of dataset for classification to reduce the noise in the text should help improve the classification.

1) **Stop Words Removal: -** A stop list is the name commonly given to a set or list of stop words. It is typically language specific, although it may contain words. Some of the more frequently used stop words for English include "a " , "of " , " the" , " I" , "it", "you" , and "and" these are generally regarded as functional words" which do not carry meaning. When assessing the contents of natural language, the meaning can be conveyed more clearly by ignoring the functional words.

2) **Stemmer: -** It is the process for reducing derived words to their stem, or root from. Stemming program is commonly referred to as stemmer or stemming algorithm. The stemming is the process for finding the root words or is the procedure of describing relevant tokens into a single type. For example "He teach us in an interesting manner" This sentence after stemming is converted into "teach interest manner" thus, by using stem (root) word, the comparison of sentence word with number of positive/negative words becomes easy. Also we will use the porter stemming algorithm: the porter stemming algorithm is a process for removing the commoner morphological and inflexional endings from words in English. Its main use is as part of

term normalization process that is usually done when setting up Information Retrieval System. The algorithm was originally described in Porter, M.F, 1980.

iii. **Classifier:** - Sentiment polarity is vague with regard to its conceptual extension. Sentiment polarity does not gives a clear boundary between the concepts of "positive", "neutral" and "negative". Classification is done on the feature extracted for every sentiment. There is classifier for each entity to be evaluated for a set of faculty. For classification we have chosen two classifiers:

Support Vector Machines: - Support Vector Machine is a supervised learning technique, which is basically used for classification of binary data. SVM classifiers are more widely used in sentiment classification. For classifying document we are using separate SVM classifier for each and different type of aspect. As we are computing sentiment of tweet from aspect level SA, we choose a single aspect from the set and classify that tweet using SVM classifier for that particular aspect. Support vector machines (SVMs) are used for classification and regression as a set of related supervised learning methods. In simple words, given a set of training examples, each marked as belonging to one of two categories, a SVM training algorithm builds a model that predicts whether a new example falls into one category or the other. Intuitively, an SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

A linear support vector machine is composed of a set of given support vectors **z** and a set of weights **w**. The computation for the output of a given SVM with N support vectors $z_1, z_2, \ldots, z_N$ and weights $w_1, w_2, \ldots, w_N$ is then given by:

$$F(x) = \sum_{i=1}^{N} w_i \langle z_i . x \rangle - b$$

Given some training data $\mathcal{D}$, a set of n points of the form

Where the

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\}_{i=1}^{n}$$

y_i is either 1 or −1, indicating the class to which the point $\mathbf{x}_i$ belongs. Each $\mathbf{x}_i$ is a p-dimensional real vector. We want to find the maximum-margin hyper plane that divides the points having

$$y_i = 1$$

From those having

$$y_i = -1$$

where $\cdot$ denotes the dot product and $\mathbf{w}$ the (not necessarily normalized) normal vector to the hyper plane. The parameter $\frac{b}{\|\mathbf{w}\|}$ determines the offset of the hyper plane from the origin along the normal vector $\mathbf{w}$.

We can put this together to get the optimization problem:

Subject to (for any $i = 1, \ldots, n$)

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1.$$

This problem can now be solved by standard quadratic programming techniques and programs. The "stationary" Karush–Kuhn–Tucker condition implies that the solution can be expressed as a linear combination of the training vectors

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i.$$

Only a few α_i will be greater than zero. The corresponding $\mathbf{x}_i$ are exactly the *support vectors*, which lie on the margin and satisfy

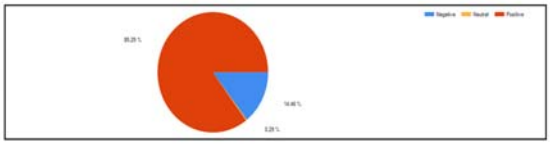
$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) = 1$$

which allows one to define the offset b. This estimate of b, the center point of the division, depends only on the single pair y_i and x_i. We may get a more robust estimate of the center by averaging over all of the N_{SV} support vectors, if we believe the population mean is a good estimate of the midpoint, so in practice, b is often computed as:

$$b = \frac{1}{N_{SV}} \sum_{i=1}^{N_{SV}} (\mathbf{w} \cdot \mathbf{x}_i - y_i)$$

4.5 Display and Visualization Classified According to Location

We represent data on a pie chart to show that which tweets from which location show which amount of positive tweets, negative tweets and neutral tweets. If the Tweets are posted on twitter without location then these tweets will be grouped into one pie chart.



When we go to page, we see two drop downs. From the first drop down we select keyword for which we want to know the sentiments of the people and click submit. After this we see that the location is reloaded with the Location from where tweets have been gathered. Then we select a single location out of all the listed and we get the data of that particular location in form of pie chart. Also it shows number of tweets that have been used to get the percentage of each positive, negative and neutral.

REFERENCES

[1] Hassan Saif Miriam Fernandez, Yulan He and Harith Alani , "Evaluation Datasets for Twitter Sentiment Analysis" ,in Knowledge Media Institute, The Open University, United Kingdom, School of Engineering and Applied Science, Aston University, UK , 2012

[2] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar and Shrikanth Narayanan , "A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle " in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Republic of Korea, 8-14 July 2012. c 2012 Association for Computational Linguistics

[3] Brendan O'Connor ,Ramnath Balasubramanyan, Bryan R. Routledge, Noah A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series" in Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media 2008-2009.

[4] Hassan Saif, Yulan He and Harith Alani , "Semantic Sentiment Analysis of Twitter" in Knowledge Media Institute, The Open University, United Kingdom

[5] Efthymios Kouloumpis , TheresaWilson,Johanna Moore , "Twitter Sentiment Analysis: The Good the Bad and the OMG!" in Work performed while at the University of Edinburgh Copyright _c 2011, Association for the Advancement of Artificial Intelligence (www.aaai.org)

[6] Alexander Pak, Patrick Paroubek , " Twitter as a Corpus for Sentiment Analysis and Opinion Mining" , Annual IEEE India Conference (INDICON) , 2013

[7] Jonathan Nagler , Joshua A. Tucker , "Drawing Inferences and Testing Theories with Big Data" in APSA Annual Meeting in Washington, DC. Supported by the INSPIRE program of the National Science Foundation (Award #SES-1248077) and Dean Thomas Carew and the Research Presented at the August 2014

[8] Albert Bifet , "Mining Big Data in Real Time" in CiteSeer x Publication 2012

[9] Ion SMEUREANU, Cristian BUCUR , "Applying Supervised Opinion Mining Techniques on Online User Reviews" in *Informatica Economică* vol. 16, no. 2/2012

[10] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow ,Rebecca Passonneau , "Sentiment Analysis of Twitter Data" in Department of Computer Science Columbia University New York, NY 10027 USA [7] Alexander Pak, Patrick Paroubek , " Twitter as a Corpus for Sentiment Analysis and Opinion Mining" , Annual IEEE India Conference (INDICON) , 2013

[11] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, Tiejun Zhao , "Target-dependent Twitter Sentiment Classification" in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 151–160, Portland, Oregon, June 19-24, 2011. c 2011 Association for Computational Linguistics

[12] Theresa Wilson, Janyce Wiebe, Paul Hoffmann , "Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis" in Submission received: 14 November 2006; revised submission received: 8March 2008; accepted for publication: 16 April 2008.

[13] Bo Pang and Lillian Lee , "Opinion mining and sentiment analysis" in Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) 1–135

[14] Basant Agarwal, Namita Mittal, Pooja Bansal, and Sonal Garg , "Sentiment Analysis Using Common-Sense and Context Information" in Hindawi Publishing Corporation , Computational Intelligence and Neuroscience , Volume 2015, Article ID 715730, Accepted 23 February 2015

[15] Fabr´ıcio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virg´ılio Almeida , "Detecting Spammers on Twitter**" in** Seventh annual Collaboration, Electronic messaging, AntiAbuse and Spam Conference July 1314,2010, Redmond, Washington, US

[16] Jiliang Tang , Chikashi Nobata, Anlei Dong, Yi Chang† and Huan Liu , "Propagation-based Sentiment Analysis for Microblogging Data" in Computer Science and Engineering, Arizona State University, Tempe, AZ