ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Gathering and Analyzing the Customers Sentimental Data from Twitter

Bhimaraj Badakundri, DR. SURESH L, DR. CHANDRAKANT NAIKODI

M. Tech Student, Dept. of Computer Science & Engineering, Cambridge Institute of Technology, K. R. Puram, Bangalore, India

Principal, Cambridge Institute of Technology, K. R. Puram, Bangalore, India

Visiting Professor, Cambridge Institute of Technology, K. R. Puram, Bangalore, India

ABSTRACT*:* For opinion mining and sentiment analysis micro blogging web sites are very rich resource. In recent years these web sites become very much interesting and popular communication tool among millions of internet users. There has been a surge of user generated information along with the rise of social networking era. In sites like twitter, facebook and instagram etc, billions of people will share their opinion and thoughts daily because of its extra features, characteristics and also short and simple manner of expression. There is a scope for a large way in a social media and micro blogging websites as exploring the challenges offered by informal and micro blogging sites. In this paper we proposed an architecture that aims at developing a standard model for gathering and analyzing the customers sentimental data from social media, in this we are considering Twitter social network. Twitter currently receives about 190 million tweets a day so it will be useful for us to analyze the sentiment of users. In this paper we are capturing the tweets corresponding to the attributes of a product and voice of the user/customer is separated according to the particular attributes based on attributes list.

KEYWORDS: Customers sentimental data, Twitter social network and twitter.

I. INTRODUCTION

The past few years have witnessed a huge growth in the use of micro blogging platforms. Popular micro blogging websites like Twitter, facebook and instagram etc have evolved to become a source of much more information. Encouraged by the growth of micro blogging platforms, organizations are exploring ways to mine social media for gathering information about how people are responding and giving feedback to their products and services. Many researches has been carried out on how sentiments are expressed in formal text patterns such as product or movie reviews and news articles, but there is no much more inventions and researches on how sentiments are expressed when it comes to informal language and message-length constraints of micro blogging. It is evident that the advent of these real-time information networking sites like Twitter have explored the creation of an unequalled public collection of opinions about every global entity that is of interest. Although Twitter may have access for an excellent channel for opinion creation and presentation, it poses newer and different challenges and the process is incomplete without adept tools for analyzing those opinions to expedite their consumption.

It has become a popular micro blogging service that has a large and rapidly growing user base where users create status messages called tweets. Users use these tweets updating what is on their mind and they express their opinion towards products, services, events and other Twitter users they are interested in. As an extensive source of real-time information the public timeline of twitter displays tweets of all users worldwide. The basic concept behind micro blogging is to provide personal status and opinion updates on several topics and products. But the current scenario surprisingly witnesses tweets covering everything under the world, ranging from current political affairs to personal experiences. Movie reviews, travel experiences, current events etc. add to the list. Tweets are different from others in its basic structure. While reviews are characterized by formal text patterns and are summarized thoughts of authors, tweets are more casual and restricted to 140 characters of text. The main objective of this research is to analyze the effectiveness of various popular classifiers and identify the more suitable classifier(s) for Twitter that could ease the

DOI: 10.15680/IJIRCCE.2016. 0404329

process of classifying sentiments in tweets. A sentiment can be defined as a personal positive or negative feeling. Opinion mining is the computational technique for extracting, classifying, understanding, and assessing the opinions expressed URL various contents. There is a scope for a large way in a social media and micro blogging websites as exploring the challenges offered by informal and micro blogging sites. Sentiment analysis refers to the broad area of natural language processing which deals with the computational study of opinions, sentiments and emotions expressed in text. Sentiment Analysis (SA) or Opinion Mining (OM) aims at learning people's opinions, attitudes and emotions towards an entity. The entity can represent individuals, events or topics. An immense amount of research has been performed in the area of sentiment analysis. But most of them focused on classifying formal and larger pieces of text data like reviews. With the wide popularity of social networking and micro blogging websites and an immense amount of data available from these resources, research projects on sentiment analysis have witnessed a gradual domain shift. In this paper we proposed an architecture that aims at developing a standard model for gathering and analyzing the customers sentimental data from social media, in this we are considering Twitter social network. Twitter currently receives about 190 million tweets a day so it will be useful for us to analyze the sentiment of users. In this paper we are capturing the tweets corresponding to the attributes of a product and voice of the user/customer is separated according to the particular attributes based on attributes list.

II. LITERATURE SURVEY

Geetika Gautam et.al [1] proposed a Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis. They pre-processed the dataset, after that extracted the adjective from the dataset that have some meaning which is called feature vector, then selected the feature vector list and thereafter applied machine learning based classification algorithms namely: Naive Bayes, Maximum entropy and SVM along with the Semantic Orientation based Word Net which extracts synonyms and similarity for the content feature. Finally they measured the performance of classifier in terms of recall, precision and accuracy. Alec Go et.al [2] which has attempted machine learning algorithms like Naive Bayes, Maximum Entropy and SVM on Tweet data the features being Unigrams, bigrams, POS tags etc. The work used emoticon features to build training sets for classifiers and was able to prove the effectiveness of machine learning techniques on Twitter data with their classification accuracy being the best at 82.2% for unigrams. Petrovic et.al [3] contributed to twitter sentiment extraction by creating a huge corpus. They streamed data using Twitter API and segregated and categorized them. Most of the works on extracting twitter sentiment have utilized this dataset until early 2013 when all public Twitter datasets had to be withdrawn as per the notification from Twitter authorities. A two step classification approach, which first classifies messages as subjective and objective, and further distinguishes the subjective tweets as positive or negative attempted to filter the neutral tweets before performing a binary classification. They proposed a polarity based classifier which extracted the Meta information contained in the tweets and the model outperformed unigrams for small training data size. Connor et.al [4] had collected over one billion tweets over 2008-09 on three political topics, analyzed and compared the results to facts. Presence or absence of sentiment bearing words were the features extracted. They used correlation analysis to highlight the potential of text streams in reflecting traditional polling results. The results varied across datasets, with correlations are as high as 80 percent.

III. METHODOLOGY

Figure 1 represents the proposed architecture. Many steps has to be fallow to gather and analyze the sentiment of users, firstly we have to go for Retrieval of tweets. As twitter is the most exaggerated part of social networking site, it consists of various blogs which are related to various topics worldwide, Instead of taking whole blogs; we will rather search on particular topic and download all its web pages then extracted them in the form of text files. Then next start with Pre-processing of extracted data from the twitter, this stage consists of URL elimination, tag elimination POS tagging. After pre-processing will go for Sentiment Analysis in which will consider data from pre-processing as input, in this phase will make use of Naive Bayes and maximum entropy methods for sentimental analysis. These analyzed data is subjected to tweet sentiment scoring model Based on the dictionary assignment of score, the proposed system interprets whether the tweet is positive, negative or neutral.

A. RETRIEVE TWEETS

Interface is provided by the twitter platform, the Twitter API, which connects website or application with the worldwide conversation happening on Twitter. As twitter is the most exaggerated part of social networking site, it consists of various blogs which are related to various topics worldwide, Instead of taking whole blogs; we will rather search on particular topic and download all its web pages then extracted them in the form of text files. We initially stream a large number of tweets based on filter key words which are related to popular voice based conversation in the market using the streaming API. The streaming API serves as an access provider to the immense volume of Twitter data. Streaming connection runs as a separate process apart from the process which handles HTTP requests. The streaming process after successfully establishing a twitter connection retrieves the input tweets which are subjected to filtering if any, and stores the result to a data store. In responding to a customer request, the HTTP handling process queries the data store and gives the results.

B. PRE-PROCESSING

After retrieving tweets from the twitter need to perform pre-processing for those retrieved tweets. We prepare the transaction file that contains opinion indicators, namely the adjective, adverb and verb along with emoticons specially we have taken a sample set of emoticons and manually assigned opinion strength to them. Also we have to identify many emotions explorer, namely, the percentage of the tweet in Caps, the length of repeated sequences & the number of exclamation marks, amongst others. Then, will pre-process all the tweets as below

- Eliminate all URLs (e.g. www.w3school.com), hash tags (e.g. #text), targets (@username), and special Twitter words (“e.g. RT”).
- Percentage of the tweet in Caps has to be calculated.
- Correct spellings; A sequence of repeated characters is tagged by a weight. We do this to differentiate between the regular usage and emphasized usage of a word.
- Replace all the emoticons with their sentiment polarity.
- Remove all punctuations after counting the number of exclamation marks.
- Using a POS tagger, the NL Processor linguistic Parser, we tag the adjectives, verbs and adverbs to those emotions.

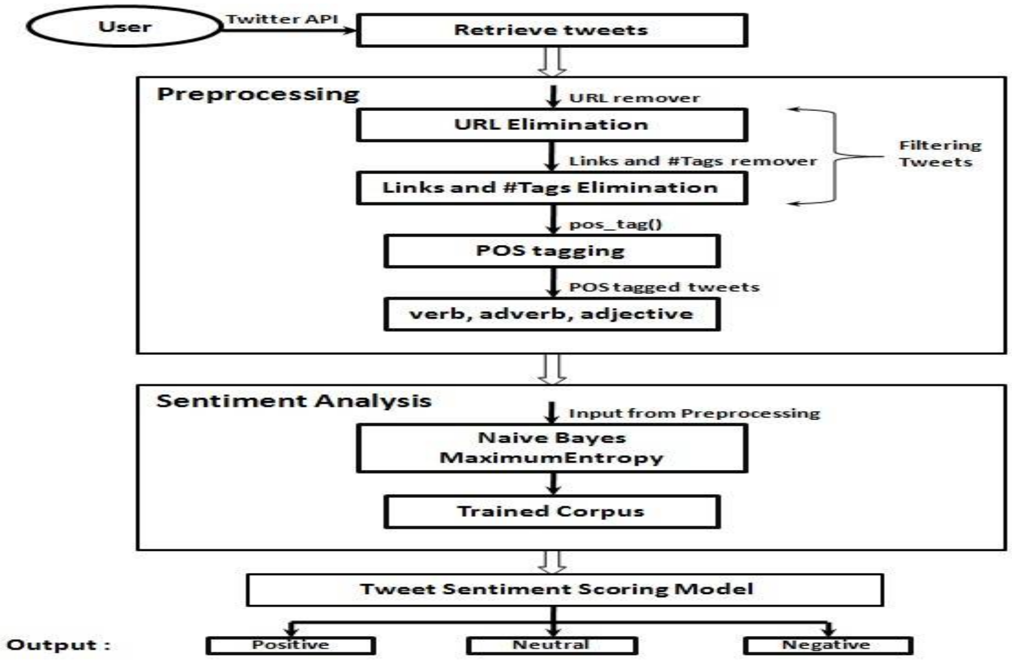


Figure 1: Block Diagram of Proposed System.

C. SENTIMENT ANALYSIS

A sentiment can be defined as a personal positive or negative feeling. In this paper will make use of Naive Bayes and maximum entropy technique for sentimental analysis of customer in a twitter.

I. NAIVE BAYES

In training and classifying stage Naive Bayes is used for its simplicity. It is a probabilistic classifier and can learn the pattern of examining a set of documents that has been categorized. It compares the contents with the list of words to classify the documents to their right category. Nave Bayes Classifier makes use of all the features in the feature vector and analyzes them individually as they are equally independent of each other.

$$C^* = argmac_c P_{NB}(c|d) \qquad (1)$$

$$P_{NB}(c|d) = \frac{\left(P(c)\sum_{i=1}^{m} P\,(f|c)^{n_i^{(d)}}\right)}{P(d)} \qquad (2)$$

Class $C^*$is assigned to tweet d, where, f represents a feature and $n_i^{(d)}$ represents the count of featuref_i found in tweet d. There are a total of m features. Parameters P(c) and P($f|c$) are obtained through maximum likelihood estimates which are incremented by one for smoothing. Pre-processed data along with extracted feature is provided as input for training the classifier using naïve bayes. Once the training is complete, during classification it provides the polarity of the sentiments. For example for the review comment “I am happy’ it provide Positive polarity as result. It does not consider the relationships between features. So it cannot utilize the relationships between part of speech tag, emotional keyword and negation. The Naive Bayes classifier is a probabilistic model based on the Baye’s theorem, which calculates the probability of a tweet belonging to a specific class such as neutral, positive or negative. This assumes that all the features are conditionally independent.

II. MAXIMUM ENTROPY

The idea behind Maximum Entropy models is that one should prefer the most uniform models that satisfy a given constraint. Maximum Entropy models are feature-based models. In a two- class scenario, it is the same as using logistic regression to a distribution over the classes. Maximum Entropy makes no independence assumptions for its features, unlike Naive Bayes. This means we can add features like bigrams and phrases to Maximum Entropy without worrying about features overlapping. The model is represented by the following:

$$P_{ME}(c|d,\lambda) = \frac{(\exp[\sum_i \lambda_i f_i(c,d))]}{\sum_{c'} \exp[\sum_i \lambda_i f_i\,(c,d)]} \qquad (3)$$

In this formula, *c* is the class, *d* is the tweet, and ¸ is a weight vector. The weight vectors decide the significance of a feature in classification. A higher weight means that the feature is a strong indicator for the class. The weight vector is found by numerical optimization of the lambdas so as to maximize the conditional probability. We use the Stanford Classifier to perform Maximum Entropy classification. For training the weights we used conjugate gradient ascent and added smoothing.

III. TWEET SENTIMENT SCORING MODEL

In tweet sentiment scoring model will group adverb and adjective as one group and named as adjective group and similarly verb and corresponding adverb in another group and will name it as verb group The adjective group strength is calculated by the product of adjective score $(adj_i)$and adverb (adv_i)score, and the verb group strength as the product of verb score (vb_i)andadverb score (adv_i)Sometimes, there is no adverb in the opinion group, so the S (adv) is set as a default value 0.5 To calculate the overall sentiment of the tweet, we average the strength of all opinion indicators like emoticons, exclamation marks, capitalization, word emphasis, adjective group and verb group as shown below:

$$S(T) = \frac{(1 + (p_c + \log(N_s) + \log(N_x))\,/3)}{|OI(R)|} * \sum_{i=1}^{|OI(R)|} s(AG_i) + S(VG_i) + N_{ei} * S(E_i) \quad (4)$$

Where,

|OI(R)| denotes the size of the set of opinion groups and emoticons extracted from the tweet, p_c denotes fraction of tweet in caps, Ns denotes the count of repeated letters, N_x denotes the count of exclamation marks, $S(AG_i)$denotes score of the i th adjective group, $S(VG_i)$ denotes the score of the ith verb group, $S(E_i)$ denotes the score of the ith emoticon N_{ei} denotes the count of the ith emoticon.

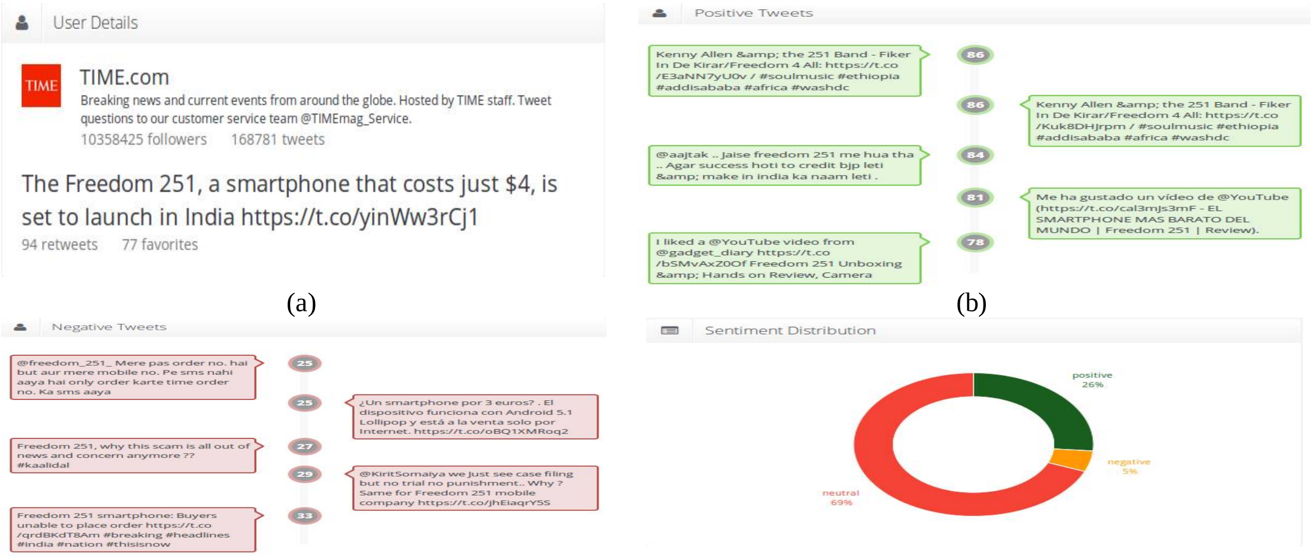
By using equation (4) we can calculate the strength of the tweet as

$$S(T) = \frac{1.33}{5}\sum_{i=1}^{5} s(AG_i) + S(VG_i) + N_{ei} * S(E_i) \quad (5)$$

According to the result obtained from the tweet sentiment score model will classify the sentiment tweet as negative, positive and neutral.

D. EXPERIMENTAL RESULT

Figure 2 represent the experimental result of the proposed system. Firstly will retrieve tweets from tweeter is shown in Figure 2(a) after these tweets are pre-processed, in that URL and tags everything eliminated and applied sentimental analysis methods like naive bayes and maximum entropy technique and result of these is subjected to tweet sentiment score model, based on the value sentiment is calculated and given as negative, positive and neutral, this demonstrated in images shown in Figure 2(b), 2(c) and overall comparison is given in Figure 2(d).



TIME.com

Breaking news and current events from around the globe. Hosted by TIME staff. Tweet questions to our customer service team @TIMEmag_Service.

10358425 followers 168781 tweets

The Freedom 251, a smartphone that costs just $4, is set to launch in India https://t.co/yinWw3rCj1

94 retweets 77 favorites

(a)

Kenny Allen &amp; the 251 Band - Fiker In De Kirar/Freedom 4 All: https://t.co/E3aNN7yU0v / #soulmusic #ethiopia #addisababa #africa #washdc

Kenny Allen &amp; the 251 Band - Fiker In De Kirar/Freedom 4 All: https://t.co/Kuk8DHJrpm / #soulmusic #ethiopia #addisababa #africa #washdc

@aajtak .. Jaise freedom 251 me hua tha .. Agar success hoti to credit bjp leti &amp; make in india ka naam leti .

Me ha gustado un vídeo de @YouTube (https://t.co/cal3mJs3mF - EL SMARTPHONE MAS BARATO DEL MUNDO | Freedom 251 | Review).

I liked a @YouTube video from @gadget_diary https://t.co/bSMvAxZ0Of Freedom 251 Unboxing &amp; Hands on Review, Camera

(b)

@freedom_251_ Mere pas order no. hai but aur mere mobile no. Pe sms nahi aaya hai only order karte time order no. Ka sms aaya

¿Un smartphone por 3 euros? . El dispositivo funciona con Android 5.1 Lollipop y está a la venta solo por Internet. https://t.co/oBQ1XMRoq2

Freedom 251, why this scam is all out of news and concern anymore ?? #kaalidal

@KiritSomaiya we just see case filing but no trial no punishment.. Why ? Same for Freedom 251 mobile company https://t.co/jhEiaqrY5S

Freedom 251 smartphone: Buyers unable to place order https://t.co/qrdBKdT8Am #breaking #headlines #india #nation #thisisnow

(c)

(d)

Figure 2: (a) Retrieving tweets; (b) Positive tweets; (c) Negative tweets; (d) Sentiment distribution Output.

E. CONCLUSION

In this paper, we proposed a set of techniques of machine learning with semantic analysis for classifying the sentence and product reviews based on twitter data. The main aim is to analyze a large amount of reviews by using twitter dataset which are already labeled. Maximum entropy and naive bayes based method was used to find the semantic operations. The naive byes technique which gives us a better result than the maximum entropy, we presented approach for sentiment gathering and analysis on Twitter data. The overall tweet sentiment was then calculated using a linear equation which incorporated emotion intensifiers too. This work is exploratory in nature and gives you precise data on Tweets Statistics, Sentiment Analysis, Topic Analysis and Voice of Customer.

REFERENCES

[1] Altadmri, A. and Ahmed, A, "Automatic Semantic Video Annotation in Wide Domain Videos Based on Similarity and Commonsense Knowledge bases" IEEE International Conference on Signal and Image Processing Applications, UK, Pp. 74 – 79, 2015.

[2] D. Chen, J. Odobez, H. Bourlard, "Text detection and recognition in images and video frames, Pattern Recognition 37" pp. 595 – 608, 2004.

[3] Shih-Wei Sun,Yu-Chiang Frank Wang, "Automatic annotation of web videos" IEEE 2011.

[4] J. W. Jeong, H. K. Hong, D. H. Lee, "Ontology-Based Automatic Video Annotation Technique in Smart TV Environment", IEEE Transactions on Consumer Electronics, Vol. 57, No. 4, pp. 1830-1836, 2011.

[5] N. Magesh, P. Thangaraj, "Semantic Image Retrieval Based on Ontology and SPARQL Query" In proceedings of International Journal of Computer Applications (IJCA) – ICACT, Number 1, pp.12-16, August 2011.

[6] K.Khurana, M.B.Chandak, "Key Frame Extraction Methodology for Video Annotation" International Journal of Computer Engineering and Technology, Volume 4, issue 2, pp.221-228, 2013.

[7] S Zhang, T, "A Generic Framework for Video Annotation via Semi-Supervised Learning" IEEE Transactions on Multimedia, Vol. 14, Issue 4, Pp. 1206 – 1219, 2012.

[8] R Datta, D Joshi, J Li, and J. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age", ACM Computing Surveys, Vol. 40, No. 2, 2008.

[9] Troncy, R, "Integrating Structure and Semantics into Audio-visual Documents" In Proceedings of the 2nd International Semantic Web Conference (ISWC), Sanibel Island, Florida, USA. Lecture Notes in Computer Science, Volume 2870, pp. 566 –58, 2003.