IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 4, Ver. I (Jul - Aug 2017), PP 09-13
www.iosrjournals.org

SentiT: A Semi Real Time System for Interpreting Sentiment in Twitter

Dr.G.Krishna Kishore[1], Kambhampati Dheeraj[2]
[1](Professor, Computer Science Department, VR Siddhartha Engineering College, Vijayawada, India)
[2](Student, Computer Science Department, VR Siddhartha Engineering College, Vijayawada, India)

Abstract: *SentiT is an opinion analysis application for Twitter. Based on the keyword searched, SentiT collects tweets having to do with it , separates and labels them into the different polarity classes neutral, negative and positive , simultaneously we also categorize them into emotions which are anger, disgust, fear, joy, sadness, surprise .Our main objective is to prepare a system that takes real time data from the twitter and come to a conclusion about the opinion on particular product/keyword*

Keywords: *Public opinion mining, Social media, Analysis Introduction*

I. Introduction

Twitter, has 313 million monthly users, 1 billion Unique visits [9] and 6000 tweets tweeted per second has become a huge database for organizations to monitor their products and brands by extracting and analysing the sentiment of the tweets posted by the public about them, so that later based on public opinion necessary changes or modifications can be done on their next product. Sentiment analysis is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, and their attributes" [10]

Twitter is a micro blogging administration with a social viewpoint. It permits its clients to express their perspectives/assumptions through an Internet SMS, called "tweets" with context to Twitter. These tweets are data framed of most extreme of 140 characters. The data of these tweets can be anything extending from a individual's temperament to individual's area to a person's interest. The stage on which these tweets are posted is known as a Timeline

Example of tweet for Dummies:
Had fun learning the big data at #BIGEVENT @pilio you should have joined us

Handle : @pilio
Hastags/Keyword: #BIGEVENT

Sentiment analysis is essentially executed by programming that can naturally extricate suppositions and states of mind in content records, be that as it may, as the main part of the feelings and mentalities that are transferred to web-based social networking sites are unstructured kind of information, it is a challenging undertaking for computers to process the data and mine noteworthy and important data from the information. The rise of social media such as blogs and social networks has driven interest in sentiment analysis. For Example : Companies like apple release there IOS beta initially to limited customers so that they can make changes necessary prior to release of original version this is part of opinion mining based on users sentiment, similarly if a new flagship mobile is released into market people will tweet about it and company can capture the public opinion and can make necessary changes in their next mobile to favour the customer .

II. Related Work

Online journals are a medium of articulation of individuals' perspectives, suppositions and so in a vernacular style. Micro blogging has turned into an indispensable part in a man's day to day life. Particularly in Twitter, there is a colossal number of tweets from individuals on different subjects each day. An effective technique is required to infer asignificant information out of it and further determine suppositions and conclusions from the important informational indexes. This is accomplished through Opinion mining and Sentiment analysis. [8] There are few approaches to concentrate tweets from twitter like by utilizing Twitter Search API, Twitter Streaming API and the firehose. This is clarified in detail in [4]

Already number of systems such as C-Feel-It[1], & TwiSent[2] has been developed which were an inspiration to our work . Although those systems were good they don't consider real time data for the analysis.
In [7], has studied Tweets are entered on the micro blogging network. The main problems associated with the large volume of data and the lack of official language so two main issues arose during the review of the tweets:

Misspellings and slang in tweets lead to a new culture of the vocabularies. There are tweets in different areas and issues that unlike other areas, such as blogs, news coverage, and text that is discussed about specific issues. Here they have used group method to classify tweets and unlike the classic method that data enter into a set of algorithms, data are imported on some collections. In general, it is done by using a set of algorithms and combines them together and there are three reasons for using the group method [3] that includes:

• Statistics: using of multiple classifications and combining them in order to get the best accuracy.
• Calculation: making a model from a lot of points to improve and classify model in order to minimize the error function.
• Representation: if a model can't achieve alone a special boundary to make decision, it makes a group with ability of set boundaries.

So we developed a system which uses Twitter API to connect to Internet and Twitter account, retrieve the tweets, Preprocess those retrieved tweets and then give it to system for analysis. The system processes the preprocessed tweets and produces 5 outcomes namely: Emotion, Polarity, Emotion cloud, Polarity Cloud, Frequenc

III. System Architecture

In this section, we give an outline of the system and explain the functionality of each module. Figure 1 presents the architecture of the system.

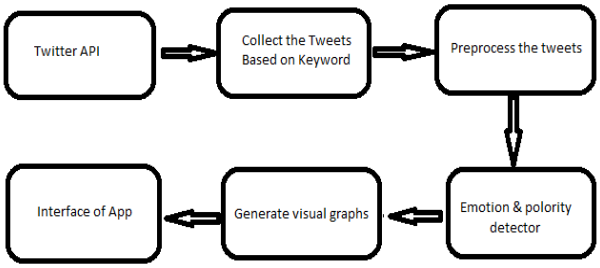

Fig 1. SentiT application Architecture

3.1 API Creation

As the basic idea of our system is to obtain real time analysis we need to create an API(Application programming Interface) for twitter which is readily available with twitter . For this API creation first we need to have a user account for twitter which can be created easily with an email and other simple steps also we need to provide full access for developer options

3.2 Tweets Collector

A tweet is simple message conveyed by a user from his/her personal account in general a tweet is limited to 140 characters per tweet, which may include special characters, symbols, images , Http Links , numbers and emoticons. A user if in interest may retweet others tweet which is indicated by RT. Also a hash tag written with a # symbol—is used to index keywords or topics on Twitter. This function was created on Twitter, and allows people to easily follow topics they are interested in. So by accessing the twitter API using a specific keyword for which we want to know the current sentiment about and retrieve the tweets based on the keyword , it is advisable to use #keyword as in general users hashtag it based on a topic or discussion

3.3 Preprocessing Tweets

As a tweet consists of all the multimedia including Http links ,images , numbers etc we need to preprocess the retrieved tweets prior feeding them to the system so we start by
Processing techniques include:

- Remove Retweets
  Some tweets in twitter can be tricky as people recently started using them for business purposes
  Ex:*We'r giving away 5 packs of Hair Dizzle as seen in this months Cosmopolitan-to get you Sparling for Summer . Just RT and FOLLOW to win*
- Remove punctuation
  Removal of punctuation include removal of symbols such as @,#,$,%,&,*(,),! Etc from the retrieved tweets
- Remove http links
  Users may include several types of https links which turn out to be part of images or random websites they want to tweet about

- Remove Tabs
  Clear the unnecessary tab spaces from the tweets
- Remove blank spaces
  Clear the unnecessary blank spaces from the tweets
- Remove next line tweets

As we retrieve tweets from the tweeter DB "/n" are included in the tweets to indicate that new line is starting from the point. As that won't help us in this context we have removed /n from the corpus. Including all the above function we need to consider the case of text so we need to convert upper case words to lower case for easy analysis of the tweets. However as tweets are tweeted by different people with different slangs and meanings although even after preprocessing the tweets still may be having some blunt english sentences which can be ignored as our most key area of concentration is extracting only the polarity and emotion for the keyword. This is also one of the disadvantages of our current system

3.4 Emotion and Polarity Detector

The preprocessed tweets are taken as an input into the system. For polarity 3 categories are considered which are positive, negative and neutral respectively. For emotion 6 categories are considered which are anger, disgust, fear, joy, sadness and surprise. All the preprocessed set of tweets are run against a simple naïve bayse logic algorithm with an inbuilt sample dataset of tweets which are predefined. Example: *a lovely film . . . elegant , witty and beneath a prim exterior unabashedly romantic* is a positive response to a tweet and *bella is the picture of health with boundless energy until a few days before she dies . this is absolutely and completely ridiculous and an insult to every family whose mother has suffered through the horrible pains of a death by cancer* is considered negative tweet For all the preprocessed tweets in the object primarily a document – term matrix is created with language as english and minimum document frequency of 1 and minimum word length of 3 and the matrix is given to system after removal of sparse terms in the corpus. Also stop words are removed from the corpus such as the, are ,of etc

3.4.1 Algorithm
For each word in words
For each category
log(score*prior/count)
scores[[category]] <- scores[[category]]+score
For each key
log(count/total)
scores[[key]] <- scores[[key]]+score
End

This applies for both polarity and emotion classifier, However we have used a separate affin list for both emotion [5] and polarity [6] respectively

3.5 Results

This systems unique feature is that once you decided on keyword for which you want to know the public opinion, on just click of button it will generate the following graphs
Polarity
Emotion
Polarity Cloud
Emotion Cloud
Frequency

On comparison we can conclude whether opinion on the specific keyword is Positive or Negative .Polarity is simply divided in to Positive, Negative and Neutral where as Emotion is classified into anger, disgust, fear, joy, sadness, surprise and unknown respectively. We may observe unknown ratio for some specific keywords such as #ps4, #xboxone etc, Here our main motive is only to capture the response from the tweets rather than analyzing the whole set of tweets

3.5.1. Sample Results

A sample output for the keyword #happy is shown below for both emotion and polarity which is captured for 100 tweets. The numbers of tweets taken from twitter data base can manually adjusted we can capture tweets ranging from 20 to 5000 tweets. There is trade off between the number of tweets user requested and run time of the system

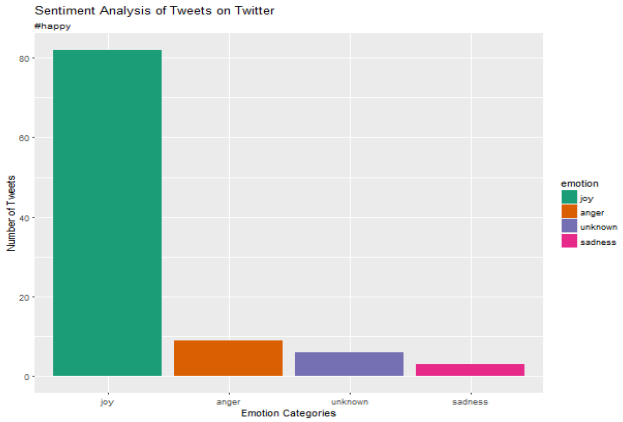

Fig 2. Emotion captured on keyword

In Fig 2 we can observe that for keyword #happy more tweets were classified as joy (i.e upto 80+) and below 10 were classified as anger and some were classified as unknown and sadness respectively

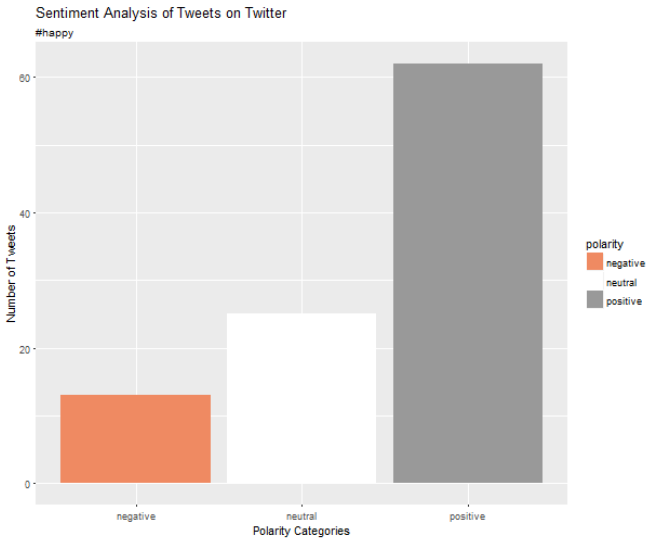

Fig 3. Polarity captured on keyword

In Fig 3 for keyword #happy up to 60+ tweets are classified as positive and less than 15 are classified as negative and rest are neutral which tend to have no polarity One of the simplest and most frequently used visualization libraries is the *simple word cloud*. The basic intent to using word cloud is to visualize the weights of the words present. The weights are proportional to the size and color of the word you see in the plot. In fig 4 shows the emotion cloud which merely gives the user a glimpse of more repeated words from each category In fig 5 shows the polarity cloud which gives the user a glimpse of more repeated words for each category

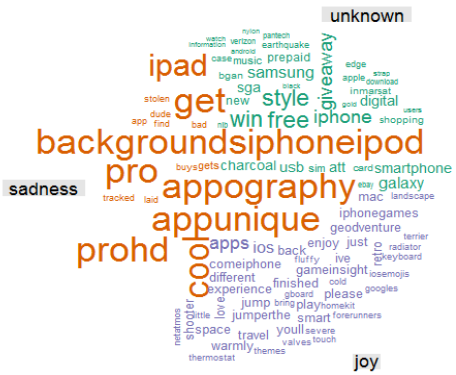

Fig 4. Emotion cloud for the corpus

Fig 5. Polarity cloud for corpus

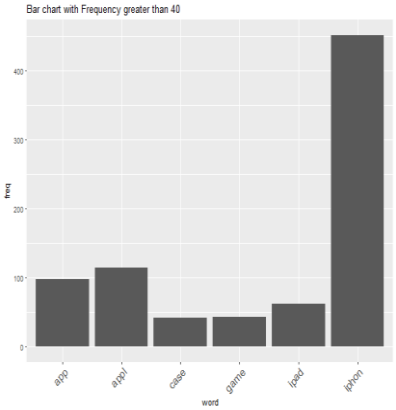

Fig 6. Frequency graph for the corpus

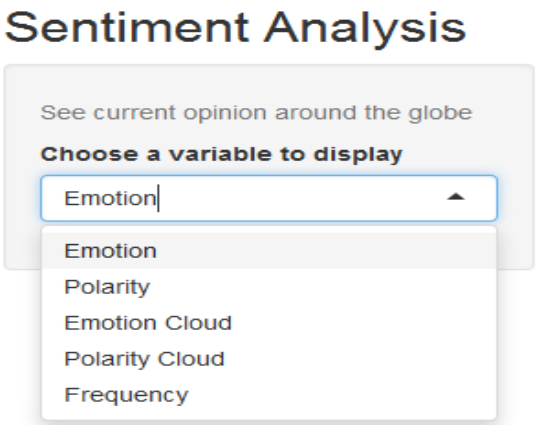

Fig 7. Glimpse of user interface

In Fig 6 we merely capture the words which appear frequently in the entire corpus of the pre-processed data So that user can come to know that what the people are currently interested in twitter. Fig 7 is the interface for accessing the outputs of sentiT application Once the system is run direct authentication is done with twitter API so that we can connect to twitter DB. Currently twitter freely supports limited use of API .After the connection is established we need to manually enter the keyword for which we need to know the sentiment about, after feeding the system with keyword (For best results usage of keyword with #hastag is recommended) , the system runs and we can select the required output . Emotion Cloud, Polarity Cloud and Frequency are merely add on features of our system which displays the word cloud for emotion , polarity and frequency respectively In the word cloud we can capture the glimpse of the most repeated words and less repeated words categorized into same classes of polarity and emotion respectively. The order of arrangement in the word cloud is merely in incremental order of occurrence of the word .So the most repeated words appear larger in the word cloud and less repeated appear smaller

IV. Conclusion and Future Work

The proposed system depicts the sentiment on particular domain based on simple keyword on just click of button with hassle free User Interface , As the dataset input to the system are live tweets from different users around the globe from twitter platform we can capture different opinions of people on the go. Currently there is tradeoff between the number of tweets retrieved and run time of the system this can be further improved up on, also handling of pragmatics in the tweets can be taken up as the future work

References

[1] Joshi, A.; Balamurali, A. R.; Bhattacharyya, P.; and Mohanty, R. 2011. C-feel-it: a sentiment analyzer for microblogs. In Proceedings of ACL Demo Papers, HLT '11, 127–132
[2] Mukherjee.; Balamurali, A. R.; Bhattacharyya,P.;2012. TwiSent:A Multistage System for Analyzing Sentiment in Twitter
[3] Dietterich, T.G., Ensemble methods in machine learning, in Multiple classifier systems. 2000, Springer. p. 1-15.
[4] F. Morstatter, J. Pfeer, H. Liu, and K. M. Carley. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In Proc. ICWSM, 2013.
[5] Carlo Strapparava and Alessandro Valitutti, "WordNet-Affect: an affective extension of WordNet". In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, May 2004, pp. 1083-1086
[6] Riloff and Wiebe (2003). Learning extraction patterns for subjective expressions. EMNLP-2003
[7] da Silva, N.F.F., E.R. Hruschka, and E.R. Hruschka Jr, Tweet sentiment analysis with classifier ensembles. Decision Support Systems, 2014. 66(0): p. 170-179.
[8] B. Pang and L. Lee, "Opinion mining and sentiment analysis,"Found.Trends Inf. Retr., vol. 2, no. 1-2, pp. 1–135, Jan 2008.
[9] https://about.twitter.com/company
[10] Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.