International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 7 (2015) pp. 16291-16301

http://www.ripublication.com

Sentiment Analysis of Twitter Data: Based on User-Behavior

Ms.Umaa Ramakrishnan, Ms.Rashmi R? and Mr.Ganesha K

Amrita Vishwa Vidyapeetham, Mysore Campus

Abstract

There are many methodologies and algorithms that are already existing and which are under use by many people. They are broadly categorized as lexicon based approach and statistical approach. We make a comparative study of each of the existing method and propose a new solution for the existing problem. The best methodologies are in statistical approaches. We come up with a statistical approach to solve the problem of BOW(Bag of Word's) problem. The solution promises to solve the sentiment analysis problem that occurs when users are trying to use too many filters and applying all of them at once. This method gives user the full privilege to make use of filters and generate an overall result of the polarity of the Tweets generated on a particular topic.

I. INTRODUCTION

Thanks to social network which has linked a lot of users throughout the globe. The users are able to connect to their long lost friends, express their views about products, movies, their feelings, politicians, book reviews, review of hotels and what not! With the internet there is a lot one can do by just a click of a button. The users express a lot of their opinions about various things that they are going about on blog/social networking websites such as Facebook, Twitter, LinkedIn, Pinterest, Instagram, Flickr, IMDb,Amazon.com etc. When there is such a platform to express ourselves, there is a huge increase in the number of people who are using the web for the same. When the number of users who use such forums are on the rise, it is quite obvious that the information generated on a day to day basis is also on the rise. When data is enormous there is a need to mine the data.

We are particularly talking about the Opinion Mining of the data generated on such forums. The information generated has a lot of sentiment attached to it. Whenever a person review's a movie, his/her comment is usually based on whether the movie was good/bad/ watchable once. If the movie is too good he/she comments about it on a social networking forum and his/her friends promptly go check the

movie and watch it. If they liked it, they might again recommend it to their friends via social network. Be it any kind of information generated, we can always classify the information generated as positive or negative.

Twitter is one such forum where the users can post a Tweet which is 140 characters long. They can share the links, images, videos and make it available to their followers. Twitter follows two Terminologies- Followers (The people who have added me as a friend and who can see my posts apart from myself) & Following (The number of people or public pages that I'm following become my friends until andunless done mutually each of the other can't see all the posts generated).

Lots of Tweets are generated in an hour. The Tweets use the SMS lingo, hyperlinks, images and retweets. It becomes challenging to extract the Twitter data and analyze them. The analysis has to be done, giving the resultant as whether the Tweet is positive or negative. The users are called Tweeples, they make use of a lot of '@' for addressing a particular person and they make use of '#' hashtag to quote hyperlink a particular topic. They make a particular topic trending by making use of the Hashtag. The data has to be analyzed excluding the stop words, the various different forms of the same word and we need to work with shortening and lengthening of various words that are existing.

There are three major steps that are involved in Twitter Sentiment Analysis:

1. Collect the datasets using the OAuth Tool obtain the private and public key from Twitter to access the Tweets and obtain the data sets.
2. Process the data using any of the classifiers which will be discussed in the following section.
3. Generate the result as whether the generated tweet was positive or negative.

II. Objectives of Sentiment Analysis:

1. Companies majorly benefit from the sentiment analysis. The company can track positive and negative reviews of their brands. It also helps them to measure the overall performance, especially on their online presence.
2. It is used across various platforms and hence many are reaping its benefits to develop marketing strategy and improve sales.
3. Identify detractors and promoters of a particular product.

III. Methodologies:

1. Lexicon-Based approach:

We usually start off with a set of words. We then train the data which are similar in meaning or we obtain an online data of a larger lexicon. This feature treats this method as a BOW-Bag Of Words model where the words are often categorized as positive and negative. This actually simplifies the whole process of categorizing the datasets. It is very easily applicable. Anything which is easily will certainly have its own drawback.

2. **Statistical based approach:**

We will have to use a Machine Learning principle to be able to use this.
There are mainly two categories of ML-Machine Learning

a) ***Supervised Learning***:Where we have a trained set of dataset which classifies the given query as positive or negative. The trained set is like a dictionary of words. The classification is based on this dictionary.

b) ***Unsupervised Learning***: The data sets are not trained prior to execution. As and when we run the algorithm of classification, the data gets split and we write into a different document set for classification of positive as 'pos' and negative as 'neg' and neutral set of data as 'neu' respectively. The naming methodology is subject to change for different projects accordingly. The naming nomenclature depends on an individual user and his/her interest.

We have seen the various ways of approaching the problem, but we need to make use of the classifiers to arrive at the solution to the problem. Hence, we can make use of the following classifiers and use them based on our needs.

Various **Classifiers** that can be used are enlisted below:

1. **Naïve-Bayes Classifier:**

It computes the dorsal probability of a class based on the distribution of words. This model works on the

$$P(label|features) = \frac{P(label) * P(f1|label) * \ldots .. * P(fn|label)}{P(features)}$$

BOW (Bag Of Words) Model. It ignores the position of the words that are in the source file. We make use of the Bayes Theorem to envision probability that the given feature belongs to a particular label.

2. **Bayesian Network:**

It assumes that the features are independent of each other. This model makes use of the graph theory to represent the polarity of the data. The nodes are used to represent random variables and edges represent conditional dependencies. One major drawback keeps us from using this particular principle, the computation complexity is very expensive in comparison with any other methodology. Hence we seldom use this algorithm.

3. **Maximum Entropy Classifier:**

It is known as Conditional Exponential Classifier, which converts the labeled features to array or vectors using the principle of encoding. The encoded vector is often used to compute the weights for each feature that can be combined to make use of the labelled set of data. It is often used to detect parallel sentences between any language pairs with small amounts of training data. It is compatible with multiple languages. This is of a higher efficiency when we use multiple different languages.

4. **Linear Classifiers:**
   There are many linear classifiers such as SVM (Support Vector Machines), which is a classifier that determines good linear separation amongst various classes available. It uses the object's features to identify or label it as which class it belongs to.

   a) **SVM (Support Vector Machine):**
      Its principle is to identify the features of each class and classify the accordingly. Text data can be efficiently analyzed using SVM. It is best suited because of the sparse nature of the texts, which can be categorized with many separate categories. SVM's are used in many applications, like apps that are classifying reviews according to the quality of the product.

   b) **Neural Networks:**
      It consists of many neurons. A neuron is a basic unit for the Neural Network. A collection of several neurons make up a Neural Network. There are multiple layers Neural Classifiers. These multilayer classifiers are used to induce multiple linear boundaries, which are used to categorize the regions as belonging to similar class that are in the closer proximity area.

5. **Decision Tree Classifiers:**
   It uses the hierarchical representation of the data and the procedure used to group them into different sub groups is the condition element which groups the data. The condition element always performs check based on the presence or absence of the predicate word given as a condition. The leaf nodes are obtained by recursive checking of the content of words. The decision tree implementations are similar to standard procedure used in Weka tool like ID3 or C4.5.

6. **Rule-based Classifiers:**
   The data space is modeled with a set of rules. The left hand side represents a condition on the feature set assigned while the right hand side represents the class label. There are ample criteria's to generate the rules. We have to train the data based on certain features so that It is able to classify. There are two important factors for the criterion. They are ***support*** and ***confidence***.
   For all of the above methodologies to work, we need to make use of **NLP-Natural Language Processing**.

The main idea behind NLP is that we must analyze the data set by accepting the data set and reading the data. We must read the data compare it with a dictionary of words to check if the word is already defined in the set of dictionary. The principle of NLP works something like a tree: where we break down the paragraph into a set of sentences. The sentences are broken down further into the set of letters and then these letters or words are compared to the existing dictionary of words.We can always work

our way through an NLP because of the ever increasing amounts of online data that is being generated day by day. NLP algorithms get better with more data. Their performance increases by training a huge set of data. The more examples you throw at it, the more examples it can process, the better it gets over time. The main idea behind NLP is that we find words in combination to understand how the words are used in a given sentence, a paragraph and how these words are used. What NLP does is look for a series of words and not just a word. These series of words determine the context and actual intent of the person who created the document and the meaning of the document so that it could help in effective decision making.

NLP has two main approaches:

1. Rule-Based: Built based on dictionary of words and customized rules and it is a supervised learning algorithm that we develop. The data is labelled by human annotators then we make use of various algorithms to annotate the data by the NLP tools.
2. Machine Learning based: Here we need to train the unsupervised data. The training data then becomes the trained data.

NLP can map, encode various concepts from various documentations. We can analyze a text document based on the Pragmatics. Just like how we have various annotations in spoken English, to make a word sound like an expression/ question/ statement, likewise NLP must take into account the grammar. NLP algorithm's put in a lot of emphasis on finding verbs and subjects (in a sentence).

We have to make use of the Phrase-Structure Grammar:

The principle behind the whole process is very simple; there are a set of symbols of sentence parts that are defined as follows:

- DET - Determiner, words such as "the", "in", "at".
- N - Noun, such as "bat", "cup" or "man".
- V - Verb, such as "kick", "kneel or "drink".
- NP - Noun Phrase, combination of a noun and a determiner, such as "the Moon".
- VP - Verb Phrase, combination of a verb and a noun phrase, such as "the ball".
- S - Complete sentence, combination of a noun phrase and a verb phrase, such as "the lady kicked the ball".

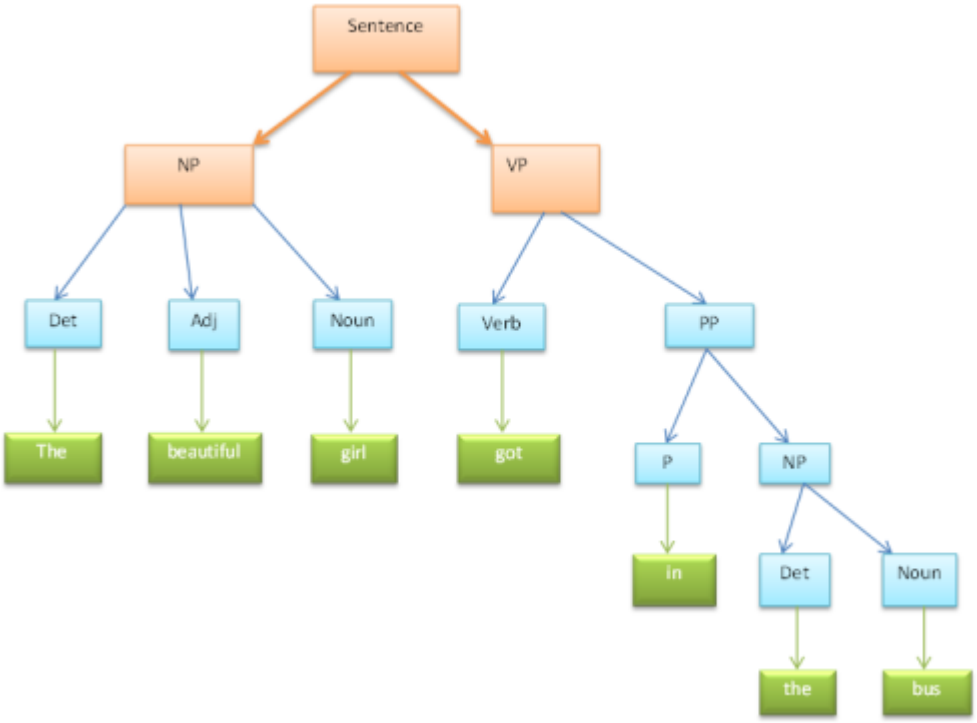
The symbols enlisted above follow a set of rules as given below:

- NP → DET N
- VP → V NP
- S → NP VP

The rules are; noun-phrase when exists a determiner followed by a noun (NP → DET N), verb-phrase when exists a verb followed by a noun-phrase (VP → V NP), and sentence when exists a noun-phrase followed by a verb-phrase (S → NP VP).

The diagram depicts the sentence: “The beautiful girl got on the bus”.

- Det: The
- Adj: Beautiful
- N: girl
- Verb: got
- P: on
- Det: the
- Noun: bus



Drawbacks in the existing system:

- We can’t detect the Sentiments of Tweets because of the polarity it may have in he data. Because of inclusion of data like emoticon, Sarcasm, Hashtags and at the rate , many other symbols that are in use).
- Detecting the Tweets based on specific topic may seem easy, but we have to ensure that the Tweets fetched are of a particular language if we are training a particular dataset.
- The Users may use the account for various personal purposes. So we must know who the authorized official account users of celebrities are.
- Detecting if the user havegenerated authentic or fake information becomes challenging when it comes to searching the tweet information.
- Detecting if the users portray a positive result towards a certain topic or negative in a discussion forum.
- Another Challenge is detecting the links that are shared as Tweets and if we

must figure out if its positive or negative.

- Fetching location-based tweets becomes really challenging especially when we are talking about zillions of Tweets that are generated globally across various platforms.
- When lots of re-tweets are generated about a trending topic, it becomes cumbersome to track the re-tweets through various users because only the originator of the message will be shown in the re-tweets.

Proposed System:

Step 1: Crawl the Twitter data from the Twitter API using the OAuth principle by using the Consumer key and Access Token keys which are both public and private respectively.
Step 2: To obtain the Tweets based on a particular Topic.
Step 3: To obtain the Tweets based on language.
Step 4: Perform region-based search for Tweets, fetch data to know the trending topics from the locality.
Step 5:Determine the user behavior. If there has been a past of fraudulent use by the user, track his activities and identify if the user has been continuing on doing so. If it turns out to be the other way round then, give him a little optimistic approach.
Step 6: Based on the re-tweets being generated train the algorithm to detect if there is a positive response towards the topic or a negative response. Based on this we can have an overall poll of the trending topic as to how much percentile of population considers it positive/ negative/ neutral.
Step 7: Keep track of Famous personalities like Authentic News Reporters on Twitter and monitor their posts so that we know what their say is about the ongoing issues across the globe, Train the data in such a way that it has if these users have a positive approach towards the topic then, train the data in such a way that it is looked at positively.
Step 8: Track the context of search and make sue that the search is very specific and doesn't deviate away from the topic.
Step 9: Collaborating all of these steps to give an optimistic output which is very specific and gives us the result stating the polarity of the Tweet generated i.e., whether the Tweet is positive/ negative or Neutral. Take the help of pie chart to show the trend of the topic.
Step 10: Finish.

Twitter Sentiment Analysis has become an integral part of our lives. When all the data and opinion sharing has gone online, it has become equally important to analyze this to take important decisions on all fronts of our lives. When we Crawl the Twitter data, we must make the search very specific and not generic. Training the data set becomes an integral part of the Sentiment Analysis. We must keep track of the user-behavior by checking the blacklist of users. Once a user has been included in the blacklist, make sure to monitor the user and then keep a close on his/her activities so that we know if the user is continuing to use the account for a fraudulent activity. If the user has changed and is Tweeting about useful information, then train the data set

to treat the user as a normal user and not speculate on all of his posts. This gives the Training data a more optimistic approach. There are various discussion forums on Twitter and everyone's opinion is definitely not the same. Opinions vary from person to person. Hence each user will have a different say about the same topic. Train the data in such a way that it tracks all of the user comments on such data and we frame positive or negative outlook on the topic. Another trending approach is to track some famous personalities on Twitter, by famous personalities we mean people who influence a huge number of people, who have a lot of influence on the public and their say affects the way we look at a problem. For example, News Reporters like Arnab Goswami, RajdeepSardesai, Ravish Kumar, Karan Thapar, Charul Malik etc., These people convey their opinions online and a lot of users follow them. We can automate our data set to fetch and analyze their Tweets on an daily basis and train the data set accordingly. Keep track of the context of the search and make sure the data being fetched from is related to the topic search and that it does not deviate too much away from the topic.

IV. Data Inputs/Sample Data:

The Application was tested for the Data Sets collected from the Twitter API. The data was collected via a Streaming API. The data must be retrieved based on the principle that we need a public access key and a private key in order to fetch the Tweets on some particular topic. This Data Sample shows the data collected for the topic: "Hello". The Tweets are fetched from a Streaming API and the Tweets are rated on a scale of 1-3 where 1-negative 2-neutral and 3-positive.

```
RT @maomaoism: hello. did u need smth http://t.co/lvMM9Ebv9v : 2
@myjobmail_2009 Hello..:) Follow @HotJacquelineF for Hot and Interesting Updates : 2
@Taser_Rani Hello... Follow @HotJacquelineF for latest Updates :) : 1
Feb 02, 2015 1:15:30 PM edu.stanford.nlp.process.PTBLexer next
WARNING: Untokenizable: ? (U+D83D, decimal: 55357)
RT @PoojaSharma_FC: Hello Every1....wid this beautiful edit on #Panchali □□□
Credit : Editor
Hv a great day #PoojaHolics□ http://t.co/Min0f… : 3
RT @InsideScoot: Hello #Dreamstart http://t.co/Q4Ftndl9iJ : 1
@VasantAjmeri Hello... Do Follow @HotSonamKapoor for latest and super Updates :) : 1
@emantahir4 Hello..:) Follow @HotJacquelineF for Hot and Interesting Updates : 2
RT @CR7Fansu: @IbnuYurayama Hello.. Go Follow @TeamRonaldoID for more info, video, pict, and all about Cristiano Ronaldo! : 1
헐. 2차에서 30명탈락했네.....ㄷㄷㄷㄷㄷ : 2
@manal7_7 Hello... Follow @HotJacquelineF for latest Updates :) : 1
RT @CSorgau: Hello Twitter! We are the Australian Clinical Supervision Association - a small group with a big vision #clinicalsupervision #… : 2
@bsps4303 Hello... Do Follow @HotSonamKapoor for latest and super Updates :) : 2
@Vrindaaaa Hello... Follow @HotJacquelineF for latest Updates :) : 1
@Sumukh125 Hello... Do Follow @HotSonamKapoor for latest and super Updates :) : 2
@Shang89 Hello... Follow @HotJacquelineF for latest Updates :) : 1
Feb 02, 2015 1:15:32 PM edu.stanford.nlp.process.PTBLexer next
WARNING: Untokenizable: ? (U+D83D, decimal: 55357)
RT @childhoodisgone: Hello old friend! □ http://t.co/dVThEiFYGY : 1
RT @NikitaBellucciX: Hello! http://t.co/PpYCNK15yy : 2
@ravi_newindia Hello... Do Follow @HotSonamKapoor for latest and super Updates :) : 2
```

V. Experimental Results and Principles:

The basic principle for testing the data set is based on the scores generated. Although we have just used three scores here namely, 1-negative, 2-neutral, 3-positive; we have to follow the following steps to get the best results:

Example: Consider 20 Tweets of the following scores each.

2,3,1,3,2,3,1,3,1,1,3,2,1,2,3,2,3,1,3,1,1

Step 1: Find mean of the Tweet score of the Tweets of a particular famous personality to study his/her traits.
Mean: 42/20=2.1
Step 2: Compute the Standard deviation, for finding an appropriate method to study the variation from the actual scale of work.

SD: $$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Variance: 0.8
SD: 0.9 approx = 1.
Step 3: Remove the noise i.e., remove the data fields which have too much of noise like the rating lies beyond 3 and 1.
Ex: If there were datasets exclusive of the data other than 1/2/3 then we are supposed to eliminate those records.
Step 4: Based on the resultant, train the Classifier to be positive/negative/neutral.
Here, the resultant of Standard deviation is 1, which means Tweets are usually negative in nature and we must check for negativity in the Tweets.

For a given set of data which is as small as the above example, this method may seem inefficient. But, on the longer run, when the data set grows exponentially, we can expect most positive results in this case. With accuracy of 87%, it is very much useful in training a Classifier based on the Tweets generated in the Twitter account.

VI. Application and contribution:

The proposed methodology will solve the problem of mis-classified Tweets, the polarity of the Tweets will be more reliable.

- Application becomes more dynamic and customizable.
- It is very useful in forming a generic opinion on a trending topic.
- Companies mainly benefit from the analysis process.
- Helps in decision making of large firms who will have millions at stake based on a single decision.
- Rebranding a product and launching a new product.
- Studying Customer trends and releasing a product accordingly.

VII. Functional Requirements:

- Sentiment analysis of Twitter Data.
- Sentiment analysis of Dynamic tweets generated based on a topic.
- Twitter user account tracking to categorize white and black list of users.

CONCLUSION

The main idea behind using Standard deviation is that we don't have to evaluate for all the Tweet score. We can consider up to 100 Tweets generated recently and come to a conclusion as to whether the user is likely to post a positive Tweet or negative Tweet and the like. There are so many existing methods for Sentiment analysis; we have enlisted them already and we know that not all of them are very precise. All of the methods have their own advantages and disadvantages. The solution proposed by us solves the problem of ambiguity and it gives more accurate results. The existing systems do not train their systems based on the factors like user behavior, region, language, popularity, trends, re-tweets depicted by RT, formation of opinion based on tweets in authorized portals. These methodologies are useful in deriving a resultant which shows the number of positive, negative and neutral tweets generated for a particular topic that is searched for. The algorithm produced accurate results but takes too long to execute them. In the near future we will work on the time efficiency and get the system running in a fast paced manner.

REFERENCES

[1] Shamanth Kumar, Fred Morstatter, Huan Liu "Twitter Data Analytics" 2013 Springer Journal.

[2] Li Ding, Tim Finin, [Anupam Joshi], "Analyzing Social Networks on Semantic Web, University of Maryland Baltimore Country, Volume-II, 8 pages.

[3] PeterA.Gloor, Jonas Krauss, Stefan Nann, et al., "Web Science 2.0 : Identifying Trends through Semantic Social Network Analysis" In an International Conference on Computational Science and Engineering, 2009, University of MIT.

[4] Vijay Srinivas Agneeswaran, Ph.D "Big Data Analytics Beyond Hadoop" book published by Pearson Education, Inc. April 2014.

[5] Holden Karau, Andy Konwinski, Patrick Wendell, and MateiZaharia "Learning Spark" Book published by O'Reilly on 2014.

[6] Alexandra BalahurBobrescu, Universitatd'Alacant, Univesidad de Alicante did a study on Methods and resources for Sentiment Analysis in Multilingual Documents of Different text types.

[7] M.M. (Mehdi) Aminian,Master Thesis, "Twitter: A Tool for Democratic Measurement", Human Aspects of Information Technology Communication and InformetionSciences,May 2012.

[8] Nada Elgendy and Ahmed Elragal, "Big Data Analytics: A Literature Review Paper", Department of Business Informatics & Operations, German University in Cairo (GUC), Cairo, Egypt

[9] Sivan Alon, Simon Perrigaud, and Meredith Neyrand, "Predicting American Idol with Twitter Sentiment", December 2011.

[10] SergejZerr, Nam Khanh Tran, Kerstin Bischoff, and Claudia Nieder´ee, "Sentiment Analysis and Opinion Mining in Collections of Qualitative Data",

Leibniz Universit¨at Hannover / Forschungszentrum L3S, Hannover, Germany.

[11] Prem, WojciechGryc,Richard D. Lawrence, "Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification", Oxford Univ. Computing Lab ,WolfsonBldg, Parks Rd Oxford OX1 3QD, UK.

[12] Bo Pang1 and Lillian Lee2, "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) 1–135 c 2008 Bo Pang and Lillian Lee.

[13] Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", Universit´e de Paris-Sud, Laboratoire LIMSI-CNRS, Bˆatiment 508, F-91405 OrsayCedex, France.

[14] Sunil B. Mane, YashwantSawant, SaifKazi, VaibhavShinde, "Real Time Sentiment Analysis of Twitter Data Using Hadoop", Sunil B. Mane et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3098 – 3100.