Journal of Engineering Sciences

Vol11,Issue2,FEB/2020
ISSN NO:0377-9254

SOCIAL MEDIA SENTIMENT ANALYSIS

[1]DEEPA A, [2]Dr. CHANDRAMOULI H, [3]V. CHANDRASEKHAR, [4]ABHIMAN J R

[1]MTech Scholor, Department of Computer Science and Engineering East Point College of Engineering and Technology, Bidharahalli, Bangalore-560067,deepaacse@gmail.com,

[2]Professor, Department of Computer Science and Engineering East Point College of Engineering and Technology Bidharahalli, Bangalore-560067,Hcmcool123@gmail.com,

[3]Asst. Professor, Department of Computer Science and Engineering, Samskruti College Of Engineering & Technology, Ghatkesar, Hyderabad, **chandrasekhar_bly@yahoo.co.in**,

[4]Professor, Department of Computer Science and Engineering East Point College of Engineering and ,Technology Bidharahalli, Bangalore-560067, abhimankrishna@gmail.com

Abstract-Social network has gained great attention in the last decade. Using social network sites such as Twitter through the internet and the web 2.0 technologies has become more affordable. The heavy reliance on social networks causes them to generate massive data characterized by some computational issues as: size, noise and reliability. These issues make social network data complex to analyze manually, resulting in the adoption of computational tools. In this paper we discuss a recent software architecture, named lambda-architecture, modified with the introduction of machine learning components, in order to perform sentiment analysis on big data streams, as the one provided by the Twitter social network.

Keywords Twitter, Sentiment analysis (SA), Opinion mining, Machine learning, Naive Bayes (NB), Maximum Entropy, Support Vector Machine (SVM).

I. INTRODUCTION

Sentiment essentially relates to feelings; attitudes, emotions and opinions. Sentiment Analysis refers to the practice of applying Natural Language Processing and Text Analysis techniques to identify and extract subjective information from a piece of text. A person's opinion or feelings are for the most part subjective and not facts. Which means to accurately analyze anindividual's opinion or mood from a piece of text can be extremely difficult. With Sentiment Analysis from a text analytics point of view, we are essentially looking to get an understanding of the attitude of a writer with respect to a topic in a piece of textand its polarity; whether it's positive, negative or neutral.

- **Business:** In marketing field companies use it to develop their strategies, to understand customers' feelings towards products or brand, how people respond to their campaigns or product launches and

  why consumers don't buy some products.
- **Politics:** In political field, it is usedto keep track of political view, to detect consistency and inconsistency between statements and actions at the government level. It can be used to predict election results as well!
- **Public Actions:** Sentiment analysis also is used to monitor and analyse social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.

Twitter is an online microblogging tool that disseminates more than 400 million messages per day, including vast amounts of information about almost all industries from entertainment to sports, health to business etc. One of the best things about Twitter—indeed, perhaps its greatest appeal—is in its accessibility. It's easy to use both for sharing information and for collecting it. Twitter provides unprecedented access to our lawmakers and

to our celebrities, as well as to news as it's happening. Twitter represents an important data source for the business models ofhuge companies as well.

When working with text mining applications, we often hear of the term "stop words" or "stop word list" or even "stop list". Stop words are basically a set of commonly used words in any language, not just English. The reason why stop words are critical to

many applications is that, if we remove the words that are very commonly used in a given language, we can focus on the important words instead.

Stop words are generally thought to be a **"single set of words"**. It really can mean different things to different applications. For example, in some applications removing all stop words right from determiners (e.g. the, a, an) to prepositions (e.g. above, across, before) to some adjectives (e.g. good, nice) can be an appropriate stop word list. To some applications however, this can be detrimental. For instance, in sentiment analysis removing adjective terms such as 'good' and 'nice' as well as negations such as 'not' can throw algorithms off their tracks. In such cases, one can choose to use a minimal stop list consisting of just determiners or determiners with prepositions or just coordinating conjunctions depending on the needs of the application. Formally, given a training sample of tweets and labels, where **label '1'** denotes the tweet is **racist/sexist** and **label '0'** denotes the tweet is **not racist**, our objective is to predict the labels on the giventest dataset.

- id : The id associated with the tweets in the given dataset.
- tweets : The tweets collected from various sources and having either positive or negative sentiments associated with it.
- label: A tweet with **label '0'** is of **positive sentiment** while a tweet with **label '1'** is of **negative sentiment.**

II.BACKGROUND WORK

Sentiment analysis deals with identifying and classifying opinions or sentiments which are present in source text. Social media is generating a huge amount of sentiment rich data in the form of tweets, status updates, reviews and blog posts etc. Sentiment analysis of this user generated data is very useful in knowing the opinion of the crowd. Twitter sentiment analysis is arduous as Literature Survey on Sentiment Analysis of Twitter Data using Machine Learning Approaches (IJIRST/ Volume 3 / Issue 10/ 004) All rights reserved by www.ijirst.org 20 compared to basic sentiment analysis due to the presence of slang words and misspellings. The maximum limit of characters that are allowed in Twitter is 140. Machine learning approach can be used for analyzing sentiments from the text. Some sentiment analysis are performed by analyzing the twitter posts about electronic products like cell phones, computers etc. using Machine Learning approach. By performing sentiment analysis in a specific domain, it is possible to identify the effect of domain information in sentiment classification. They presented a new feature vector for classifying the tweets as positive, negative or neutral and extract people's opinion about products [1]. Another research tried to pre-processed the dataset, after that extracted the adjective from the dataset that have substantial meaning which is called feature vector, then selected the feature vector list and thereafter applied machine learning algorithms such as Naïve-Bayes, Maximum Entropy and SVM along with the Semantic Orientation based Word-Net which extracts synonyms and relation for the content feature. At the end, they measured the performance of classifier in terms of recall, precision and accuracy [2]. Some researchers had an approach where posted tweets from the Twitter micro-blogging site are subjected to pre- processing and classified based on their emotional content as positive, negative and neutral or irrelevant; and compares the performance of various classifying algorithms based on their precision and recall in such cases. Further, the paper also discusses the applications of this research and its limitations [3]. A number of machine learning like Naïve Bayes and Random Forest models performed sentiment analysis on product review data [8]. Some work in this field included experiments with mood classification on blog posts. One of the researches also deals with review of aspect-based opinion polling from unlabelled free-form textual customer reviews without requiring customers to answer any questions [10]. The tweet retrieval process needs access tokens from the twitter developer site and a piece of code which perform the operation of retrieving those tweets. As the base language used will be java, we choose to implement the Java library called Twitter4J. This library is developed for the twitter API. With Twitter4J library, you can easily integrate your Java application with the Twitter service. Twitter4J has the following

features like: 100% Pure Java - works on any Java Platform v5 or later, Android platform and Google AppEngine

ready, Zero dependency: Additional jars are not required, Built-in OAuth support, already built gzip support. There are some system requirements that needs to be followed for the Twitter 4J java library to successfully operate. The library supports Windows and Unix Operating systems with Java 1.5 or higher versions installed on it. A java document is also provided in case a user needs to find the method name, syntax, or the root package while implementing the code. To use the java library, the user just needs to add the .jar file to the java application class path.[4]

III. EXISTING SYSTEM

The existing system works only on the dataset which is constrained to a particular topic. The existing systems also do not determine the measure of impact the results determined can have on the field taken into consideration and it does not allow retrieval of data based on the query entered by the user i.e. it has constrained scope. In simple words, it works on static data rather than dynamic data. Unsupervised algorithms like Vector Quantization, are used for data compression, pattern recognition, facial and speech recognition, etc and therefore cannot be used in determining sentiment in twitter data. Apriori algorithm fails to handle large datasets and as a result can generate faulty results.

IV. PROPOSED SYSTEM

In the proposed system, we will retrieve tweets from twitter using twitter API based on the query. The collected tweets will be subjected to pre-processing. We will then apply the supervised algorithm on the stored data. The supervised algorithm used in our system is Support Vector Machine (SVM). The results of the algorithms i.e. the sentiment will be represented in graphical manner (pie charts/bar charts). The proposed system is more effective than the existing one. This is because we will be able to know how the statistics determined from the representation of the result can have an impact in a particular field.

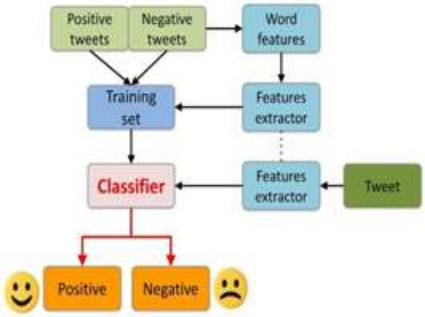

Fig 4.1 Sentiment analysis Architecture

4.1 Pre-Processing

Raw tweets scraped from twitter generally result in a noisy dataset. This is due to the casual nature of people's usage of social media. Tweets have certain special characteristics such as retweets, emoticons, user mentions, etc. which have to be suitably extracted. Therefore, raw twitter data has to be normalized to create a dataset which can be easily learned by various classifiers. We have applied an extensive number of pre-processing steps to standardize the dataset and reduce its size. We first do some general pre-processing on tweets which is as follows.

- Convert the tweet to lower case.
- Replace 2 or more dots (.) with space.
- Strip spaces and quotes (" and ') from the ends of tweet.
- Replace 2 or more spaces with a single space.

Special twitter features as follows.

URL

Users often share hyperlinks to other webpages in their tweets. Any particular URL is not important for text classification as it would lead to very sparse features. Therefore, we replace all the URLs in tweets with the word URL. The regular expression used to match URLs is ((www\.[\S]+)|(https?://[\S]+)).

4.2 User Mention

Every twitter user has a handle associated with them. Users often mention other users in their tweets by @handle. It replaces all user mentions with the word USER_MENTION. The regular expression used to match user mention is @[\S]+.

4.3 Emoticon

Users often use a number of different emoticons in their tweet to convey different

emotions. It is impossible to exhaustively match all the different emoticons used on social media as the number is ever increasing. However, we match some common emoticons which are used very frequently. the matched emoticons is replaced with either EMO_POS or EMO_NEG depending on whether it is conveying a positive or a negative emotion. A list of all emoticons matched by our method is given in table 3.

4.4 Retweet

Retweets are tweets which have already been sent by someone else and are shared by other users. Retweets begin with the letters RT. We remove RT from the tweets as it is not an important feature for text classification. The regular expression used to match retweets is \brt\b. After applying tweet level pre-processing, we processed individual words of tweets as follows.

- Strip any punctuation ['"?!, ():] from the word.

Convert 2 or more letter repetitions to 2 letters. Some people send tweets like I am sooooo happpppy adding multiple characters to emphasize on certain words. This is done to handle such tweets by converting them to I am soo happy.

- Remove - and '. This is done to handle words like t-shirt and theirs's by converting them to the more general form t-shirt and theirs.
- Check if the word is valid and accept it only if it is. We define a valid word as a word which begins with an alphabet with successive characters being alphabets, numbers or one of dot(.) and underscore (_). Some example tweets from the training dataset and their normalized versions are shown in table4.

4.5 Feature Extraction

We extract two types of features from our dataset, namely unigrams and bigrams. We create a frequency distribution of the unigrams and bigrams present in the dataset and choose top N unigrams and bigrams for our analysis.

4.6 Unigrams

Probably the simplest and the most commonly used features for text classification is the presence of single words or tokens in the the text. We extract single words from the training dataset and create a frequency distribution of these words.

4.7 Bigrams

Bigrams are word pairs in the dataset which occur in succession in the corpus. These features are a good way to model negation in natural language like in the phrase – This is not good.

4.8 Naive Bayes

Naive Bayes is a simple model which can be used for text classification. In this model, the class ĉ is assigned to a tweet t, where

$$\hat{c} = \underset{c}{argmax}\ \mathrm{P(c|t)}$$

$$\mathrm{P(c|t)} \propto \mathrm{P(c)} \prod_{i=1}^{n} \mathrm{P}(f_i|\mathrm{c})$$

In the formula above, f i represents the i-th feature of total n features. P(c) and P(f i |c) can be obtained through maximum likelihood estimates.

4.9 Maximum Entropy

Maximum Entropy Classifier model is based on the Principle of Maximum Entropy. The main idea behind it is to choose the most uniform probabilistic model that maximizes the entropy, with given constraints. Unlike Naive Bayes, it does not assume that features are conditionally independent of each other. So, we can add features like bigrams without worrying about feature overlap. In a binary classification problem like the one we are addressing; it is the same as using Logistic Regression to find a distribution over the classes. The model is represented by

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

Here, c is the class, d is the tweet and λ is the weight vector. The weight vector is found by numerical optimization of the lambdas so as to maximize the conditional probability.

4.10 Decision Tree

Decision trees are a classifier model in which each node of the tree represents a test on the attribute of the data set, and its children represent the outcomes. The leaf nodes represent the final classes of the data points. It is a supervised classifier model which uses data with known labels to form the decision tree and then the model is applied on the test data. For each node in the tree the best test condition or decision has

P to be taken. We use the GINI factor to decide the best split. For a given node t, where p(j|t) is the relative frequency of class j at node t.

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

4.11 Random Forest

Random Forest is an ensemble learning algorithm for classification and regression. Random Forest generates a multitude of decision trees classifies based on the aggregated decision of those trees. For a set of tweets x 1 , x 2 , . . . x n and their respective sentiment labels y 1 , y 2 , .
. . n bagging repeatedly selects a random sample (X b , Y b) with replacement. Each classification tree f b is trained using a different random sample (X b , Y b) where b ranges from 1 . . . B. Finally, a majority vote is taken of predictions of these B trees.

4.12 XGBoost

Xgboost is a form of gradient boosting algorithm which produces a prediction model that is an ensemble of weak prediction decision trees. We use the ensemble of K models by adding their outputs in the following manner

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F$$

where F is the space of trees, x i is the input and y ˆ i is the final output. We attempt to minimize the following loss function

$$L(\Phi) = \sum_i l(\hat{y}_i, y_i) + \sum \Omega(f_k)$$

$$where\ \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

4.13 SVM

SVM, also known as support vector machines, is a non-probabilistic binary linear classifier. For a training set of points (x i , y i) where x is the feature vector and y is the class, we want to find the maximum-margin hyperplane that divides the points with y i = 1 and y i = −1.
The equation of the hyperplane is as follow

$$w \cdot x - b = 0$$

We want to maximize the margin, denoted by γ, as follows:

$$\max_{w,\gamma} \gamma, s.t. \forall i, \gamma \leq y_i(w \cdot x_i + b)$$

4.14 Positive, negative and neutral tweets The difference between these types of tweets might seem easy, but it is a bit more complex but easy to understand. A positive tweet is a tweet that throws a positive sentiment after having analyzed all its words. Sentiment analytics in Twitter is an amazing science because the system that analyzes Twitter information has to be so accurate in just a few words and in a network where irony plays an important role. Each word of the tweet has its own score that can vary depending on the context. For example: the word “amazing” has a high score, but if it is around the words “piece of sh**” like “amazing piece of sh**” then it is not as good. You can understand why this is so exciting.To have positive tweets in your campaign is great, but it is more important to have more positive impacts. This is what twitter sentiment analysis is all about. This will be explained in the next section, but imagine this: you’re running a campaign, what would you prefer: 10 positive tweets from 1 person with 5 followers or 1 positive tweet sent by @LadyGaga about your campaign. The Twitter impressions generated byLady Gaga will be huge. In this posts we will explain how we have played around with all this Twitter sentiment analytics and we hope you feel proud of our work.The aim of the *Sentiment Score* is to calculate how positive or negative a report is in general. This is calculated by taking into account a greater number of variables than just the number of positive and negative tweets. For Tweet Binder it is important to considerthe number of Twitter impacts (or impressions) and users that have tweeted in a positive, negative or neutral way. Therefore, the *Sentiment Score* takes into consideration, in general lines, the following variables as we saw before:

- Number of positive, negative or neutral **tweets**
- Amount of **users** that have participated in the report
- Number of positive, negative or neutral **Twitter impressions (impacts)**

Why have we chosen these variables? Because Tweet Binder Twitter sentiment analytics consider "more positive" that more users tweet in a positive way rather than just one. Meaning that if 500 users have tweeted positively about our hashtag, it is better than if just 1 user sends 500 tweets. For example:

- 500 positive tweets sent by 1 user = bad (probably spammer or someone associated to the Twitter campaign or event)
- 500 positive tweets sent by 100 users = better (it means that a lot of people agree that hashtag is great)

Same thing happens with **negative tweets**. Many times, the commonly named *trolls* will try to sabotage an event or campaign. These users usually send tons of tweets but they are not a big group of users. Meaning that a little amount of users will send a big amount of tweets (and probably they have a little number of followers). It is not very negative for the report that this little amount of users sends tons of tweets rather than if they were sent from a huge amount of accounts. For example:

- 500 negative tweets sent by 500 users = bad
- 500 negative tweets sent by 20 users = not that bad, probably someone trying to undermine the hashtag

Fig 4.2 Sentiment score

Conclusion

Thus, the basic knowledge required to do sentiment analysis of Twitter is well stated in this review paper. What is Sentiment Analysis with respect to levels of sentiment analysis, what are the approaches to do sentiment analysis, methodologies for sentiment analysis, features to be extracted from text and the applications where it can be utilized is mentioned hierarchically. If we want to do Twitter's sentiment analysis we need to know about the twitter, about extracting the tweets, its structure, their meaning. This paper gives brief notion of tweets. When one wants to do sentiment analysis of tweets, he has to do it in a specialized aspect of sentiment analysis. So, the brief knowledge about Twitter Sentiment Analysis is given in this paper. Different methods and techniques are discussed in a comparative manner. The accuracy/ result of each method enables us to imagine the efficiency of applied technique in respective circumstances. ACKNOELEDGEMENT This review has been partially supported by Department

REFERENCES

[1] A.Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320- 1326

[2] R. Parikh and M. Movassate, "Sentiment Analysis of User- GeneratedTwitter Updates using Various Classi_cation Techniques",CS224N Final Report, 2009

[3] Go, R. Bhayani, L.Huang. "Twitter Sentiment ClassificationUsing Distant Supervision". Stanford University,
Technical Paper,2009

[4] L. Barbosa, J. Feng. "Robust Sentiment Detection on Twitterfrom Biased and Noisy Data". COLING 2010: Poster Volume,pp. 36-44.

[5] Bifet and E. Frank, "Sentiment Knowledge Discovery inTwitter Streaming Data", In Proceedings of the 13th InternationalConference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.

[6] Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011Workshop on Languages in Social Media,2011 , pp. 30-38

[7] Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". Coling 2010: Poster Volumepages 241{249,
Beijing, August 2010

[8] Po-Wei Liang, Bi-Ru Dai, "Opinion Mining on Social MediaData", IEEE 14th International Conference on Mobile Data Management,Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN: 978-1-494673-6068-5,
http://doi.ieeecomputersociety.org/10.1109
/MDM.2013.

[9] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, "Using wordnet to measure semantic orientations of adjectives," 2004.

[10] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

[11] Liu, S., Li, F., Li, F., Cheng, X., &Shen, H.. Adaptive cotraining SVM for sentiment classification on tweets. In Proceedings of the 22nd ACMinternational conference on Conference on information & knowledge management (pp. 2079-2088). ACM,2013.

[12]Saha, S.; Yadav, J.; Ranjan, P. Proposed approach for sarcasm detectionin twitter. Indian J. Sci. Technol. 2017, 10. [CrossRef]

[13]Turney, P.D. Thumbs up or thumbs down: Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meetingon Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002;
pp. 417–424. 55.
Sentiment Classifier. Available online: https://github.com/kevincobain2000/senti
ment_classifier (accessed on 26 February 2018).