International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 06 | June -2017 www.irjet.net p-ISSN: 2395-0072

SENTIMENT ANALYSIS ON TWITTER DATA

Chandan Arora[1], Dr. Rachna[2]

[1]*Research Scholar, Department of Computer Science, Global Institutes of Management &Emerging Technologies, Amritsar, India*

[2]*Associate Professor, Department of Computer Science, Global Institutes of Management & Emerging Technologies, Amritsar, India*

Abstract - *Twitter is one of the most commonly used platforms for sharing opinions, expressing views. Sentiment Analysis on twitter can allow users to understand the opinions expressed in tweets and classifying them in positive or negative categories. The organizations can use sentiment analysis to get an idea of the customer reviews of their products, and subsequently try and improve their services based on the reviews.*

Keywords: Data Mining, Sentiment Analysis, Twitter, Classifiers.

1. INTRODUCTION

Data Mining refers to extracting knowledge and discovering patterns from large data-sets. Almost all organizations collect and store data, and extract useful information, while discarding unnecessary data. The useful data is then analyzed in order to discover meaningful patterns [1]. Computing large data sets is an integral component of almost all organizations. The goal is to review the data sets and transform them into usable patterns. Data mining is sometimes also referred to as "knowledge discovery from data", or KDD.Earlier techniques that were used to identify data patterns were Bayes' theorem and Regression Analysis. As the times have gone by, the size of data sets has increased remarkably. As such, the discoveries in computer sciences like neural networks, clustering, etc. have made it easier to manage these data sets better [2].

The traditional techniques such as database can handle just a limited amount of data. In order to analyze millions of records, data mining has to be used. Specific computer algorithms such as neural networks, decision trees are applied to extract patterns from the given data sets.

2. SENTIMENT ANALYSIS

Sentiment Analysis (SA) elucidates users whether information or opinion regarding a certain product is positive, negative or neutral. Sentiment basically refers to any opinion or a feeling expressed by someone. Various organizations use this analysis to understand users' opinion for their products. For example, a particular e-commerce website can utilize sentiment analysis to discern if their products are being liked by the customers or not. The reviews for the products can be generalized into either positive or negative as well as neutral categories[3]. SA can be simply put as "What other people think?"

The terms views, belief, sentiment and opinion can be defined as follows:

- Opinion- A conclusion open to dispute
- View- A subjective opinion
- Belief- Deliberate acceptance and intellectual assent
- Sentiment- opinion representing someone's feelings[4].

2.1 SENTIMENT ANALYSIS TECHNIQUES

There are basically three techniques to perform Sentiment Analysis.

1. SA using machine learning.

2. SA using lexicon based techniques

3. SA using the above two techniques combined together.

1. Machine learning technique involves both supervised and unsupervised learning.

1.1 **Unsupervised Learning** is based on just inputs, without any mention of targets. It just relies on clustering.

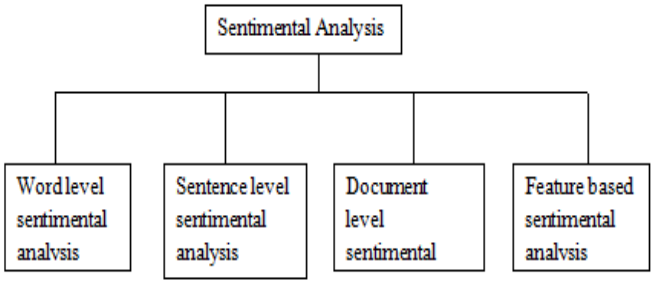
1.2 **Supervised Learning** defines pre-specified targets which should be achieved, along with the inputs. Data set are trained to achieve significant outputs when encountered during decision-making.

2. **Lexicon-Based Approaches:** Lexicon based method assigns positive or negative polarity based on the sentiment of each word and then a dictionary is created.

We can use a combining function, for example, sum or average to find out the general sentiment of a document.

3. Hybrid techniques combine both machine learning and lexicon based approaches to get better classification results. The dominant features of both these methods can be used to obtain steady results[5].

2.2 LEVELS OF SENTIMENT ANALYSIS



There are 4 levels of sentiment analysis:

1) **Document level**: In this, whole document is classified as either positive or negative[6]. Respective words and sentences are checked for sentiments and are subsequently combined to find the sentiment polarity of the complete document.

2) **Word level:** Word level sentiment analysis utilizes adjectives and adverbs that define the sentiment of each word. Two methods that comment on sentiment at word level are :

   1. Dictionary-based approaches

   2. Corpus-based approaches

3) Sentence or phrase level: In this sentiment analysis, each sentence is categorized as either positive, or negative and may be neutral as well. If a sentence turns out to be neutral, it means there is no opinion. All the sentences can then be combined to find polarity of a paragraph or even complete document.

4) Feature-level or aspect-level: It helps to analyze what people are trying to suggest. Feature level tries to extract sentiment from the opinion directly.

Various tasks involved in this are:

i. Identify and extract object features on which opinion holder has commented.

ii. Check the polarity of each opinion whether it is positive or negative. It can be even neutral.

iii. Find feature synonym[17].

3. CLASSIFIERS

3.1 Naive Bayes: It is a probabilistic classifier based on Bayes' Theorem. It can learn the pattern of examining a set of documents. A comparison is done between the subject matter of the document and a given set of words, so that a correct category of classification can be found. Assume 'd' is a chance to be the tweet and c* is a class assigned to d, where

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

$$P_{NB}(c|d) = \frac{(P(c))\sum_{i=1}^{m} p(f|c)^{n_{i(d)}}}{P(d)}$$

In the above equation, 'f' refers to feature, count of feature (fi) is denoted with ni(d) and is available in d representing tweet. Here, number of features are denoted by m. Parameters P(c) and P(f|c) are computed through maximum likelihood estimates.

For training and classification of Naïve Bayes Machine Learning technique, Python NLTK library[8] is used. NLTK comes with all resources to get started on sentiment analysis like feature extraction. NLTK classifiers work with dictionaries that map a feature name to a feature value.

3.2 Maximum Entropy[ME]: This is an exponential model. In ME Classifier, we don't make any assumptions in relation to conditional independence between features. Maximum Entropy Classifier needs more time to be trained as compared to Naive Bayes because of

optimization problems.Maximum entropy can handle overlap feature as well, and then chooses the model with maximum entropy.

3.3 Support Vector Machine: Support vector machine is mostly used for pattern recognition and analysis of data. This was invented by Vladimir Vapnik. In Support Vector Machine, classification is performed by construction of an N-Dimensional hyperplane, which can separate data into separate categories. Two vectors of a particular size are fed as inputs, and classification is performed.

4. Literature Review

In 2015, Rincy Jose, et.al used a Natural Language (NLP) approach to enhance sentiment classification by adding semantics in feature vectors and thereby using ensemble methods for classification.

Generally, bag-of-words approach has been used for mining sentiments online. In this approach, individual words are considered instead of complete sentences. Traditional machine learning algorithms such as Support vector Machines, Naive Bayes' and Maximum entropy etc. are commonly used to solve the classification problems[10].

There is a certain level of bias toward a particular class using above techniques. Therefore, Natural Language (NLP) based approach has been used to enhance the sentiment classification. Conducted experiments have shown that semantics based feature vector gives 3-5% better results than the above mentioned bag of words approach.

In 2016, Aldo Hernández, et.al presented a paper on sentiment analysis method on Twitter content to predict future attacks on the web [12]. The method is based on the daily gathering of tweets from two sets of users; the individuals who utilize the platform as a method for expression for views on relevant issues, and the individuals who utilize it to present contents identified with security attacks in the web.

Predicting attacks is an imperative task that considers what actions ought to be taken if the assault is latent. The daily Daily information is converted into data that can be broke down statistically to predict whether there is a plausibility of an assault. The last is finished by investigating the aggregate sentiment of users and groups of hacking activists in response to a global event. The goal is to predict the response of specific groups involved in hacking activism when the sentiment is sufficiently negative among various Twitter users. For two contextual analyses, it is demonstrated that having coefficients of determination greater than 44.34% and 99.2% can figure out whether a significant increase in the percentage of negative opinions is identified with attacks.

In 2015, Anurag P. Jain, et.al presented an approach for examining the sentiments of users utilizing data mining classifiers [13]. It additionally compares the performance of single classifiers for sentiments analysis over ensemble of classifier.

With quick growth in client of Social Media as of late, the researcher get attracted towards the utilization of social media data for sentiment analysis of individuals or particular product or person or event. Twitter is one of the broadly utilized social media platforms to express the considerations. Experimental results acquired demonstrate that k-nearest neighbor classifier gives high predictive accuracy. experiments have shown that single classifiers give better results than ensemble of classifier approach. It can be seen from the test results that data mining classifiers is a decent decision for sentiments prediction utilizing twitter data. In experimentation, k-nearest neighbor (IBK) outperforms over every one of the three classifiers in particular RandomForest, baysNet, Naive Baysein. RandomForest additionally gives great prediction accuracy. There is a no compelling reason to utilization of ensemble of classifier for sentiments predictions of tweets as single classifier (i.e k-nearest neighbor) gives a better accuracy over all combinations of ensemble of classifier.

In 2011, Ming Hao, et.al used novel techniques three novel time based visual sentiment analysis techniques to explore high volume of Twitter data. These techniques are: (1) topic-based sentiment analysis that extracts, maps, and measures customer opinions; (2) stream analysis identifying interesting tweets depending on density, negativity, and impact attributes; and (3) pixel cell-based sentiment timetables and high density geo maps that visualize substantial volumes of data in a single view. These techniques were connected to a variety of twitter data, (e.g., movies, amusement parks, and hotels) to demonstrate their distribution and patterns, and to recognize influential opinions. A visual analysis of Twitter time series was displayed, to explore equivalent Twitter data streams.

In 2015, Manju Venugopalan, et.al proposed building up a half and half model for sentiment classification that explores the tweet specific features and uses domain independent and domain specific lexicons to offer a domain oriented approach to analyze sentiment of shoppers regarding different smart phone brands [15]. The analyses have demonstrated that the results enhance by around 2 points on an average over the unigram baseline. The SVM accuracy has improved in the range 1.5 to 3.5 and J48 could provide an accuracy improvement ranging from 1.5 to 4 points across domains. The improved lexicon which have adapted polarities learning the domain and the tweet specific features extracted have added to the improvement in classification accuracies.

In 2015, Gaurav D Rajurkar, et.al. proposed an approach of consolidating the Apache Open Source platform which solves the issues of Real Time Analytics utilizing HADOOP. It additionally provides scalability and reduced cost over analytics by utilizing open Source Software. The work proposes to combine the Apache Open Source Modules and configure them to get the required result [16]. Data can be downloaded at a faster rate on HDFS by utilizing source and sink mechanism. The Hadoop is flexible and scalable architecture. The proposed work is based upon the phenomenon of combination of open source software alongside commodity hardware that will increase the profit of IT Industry. So the proposed system utilizes an efficient Apache Open Source Product which presents the model that can have Twitter Trend Analysis utilizing HADOOP where no additional work like scraping, cleansing and data protection required. The proposed work concludes with the phenomenon of Open Source Software alongside Commodity Hardware that will increase IT Industry Profit.

TABLE 1

Author	Year	Description	Outcome
Rincy Jose, et.al," Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation"	2015	Natural Language (NLP) based approach to enhance the sentiment classification by adding semantics in feature vectors using ensemble methods for classification	Ensemble method outperforms the traditional classification methods by about 3- 5%.
Aldo Hernández, et.al," Security Attack Prediction Based on User Sentiment Analysis of Twitter Data"	2016	Sentiment analysis method on Twitter content to predict future security attacks on the web	Coefficients of determination greater than 44.34% & 99.2% can figure out whether a significant increase in the percentage of negative opinions is identified with attacks.
Anurag P. Jain, et.al," Sentiments Analysis Of Twitter Data Using Data Mining"	2015	Examining the sentiments of users utilizing data mining classifiers; and comparison between performance of single classifiers for sentiments analysis over ensemble of classifier.	k-nearest neighbor classifier gives high predictive accuracy. Single classifiers outperforms ensemble of classifier approach
Ming Hao, et.al," Visual Sentiment Analysis on Twitter Data Streams"	2011	(1) topic-based sentiment analysis; (2) stream analysis that identifies interesting tweets based on their density, negativity, and impact attributes; and (3) pixel cell-based sentiment timetables to visualize substantial volumes of data in a single view.	A visual analysis of Twitter time series, which combines sentiment and stream analysis with geo and time-based interactive visualizations for the exploration of genuine Twitter data streams.
Manju Venugopalan, et.al," Exploring Sentiment Analysis on Twitter Data"	2015	Half and half model for sentiment classification that explores the tweet specific features	The results enhance by around 2 points on an average over the unigram baseline.

Gaurav D Rajurkar, et.al," A speedy data uploading approach for Twitter Trend and Sentiment Analysis using HADOOP"	2015	Consolidating the Apache Open Source platform which solves the issues of Real Time Analytics utilizing HADOOP.	Quick data downloading approach for efficient Twitter Trend Analysis.
Akshi Kumar, Teeja Mary Sebastian," Sentiment Analysis on Twitter"	2012	Both Corpus-based method and dictionary-based methods are used to extract opinion words.	Tweets with a negative score are classified as negative, while tweets with positive value are classified as positive.
Kushal dave, et.al,"Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews"	2003	Traversing through product reviews and and automatically extricating between positive and negative reviews using corpus of of self-tagged reviews available from major websites.	Mixed reviews offer serious noise to the problem of scoring words. Also shorter reviews are difficult to categorize. Thresholding of shorter reviews can help in classification

CONCLUSION

This section concludes the survey provided in the Literature Review section. Various techniques such as corpus-based, dictionary based methods and Natural Language processing technique have been used for sentiment analysis on Twitter. It can be concluded that sentiment analysis can be further improved and more accurate results can be achieved in future using more efficient algorithms.

REFERENCES:

[1] Nikita Jain , Vishal Srivastava, "Data Mining Techniques: A Survey Paper", IJRET, eISSN: 2319-1163 | pISSN: 2321-7308.

[2] Yihao Li, "Data Mining: Concepts, Background and Methods of Integratign Uncertainty in Data Mining"

[3] Vishal A. Kharde, S.S. Sonawane," Sentiment Analysis of Twitter Data: A Survey of Techniques", 2016, International Journal of Computer Applications, Volume 139 – No.11

[4] Deepali Arora, Kin Fun Li and Stephen W. Neville," Consumers' sentiment analysis of popular phone brands and operating system preference using Twitter data: A feasibility study", 2015, IEEE, 1550-445X

[5]Haseena Rahmath P , Tanvir Ahmad, "Sentiment Analysis Techniques - A Comparative Study", IJCEM International Journal of Computational Engineering & Management, Vol. 17 Issue 4, July 2014

[6] P. Grandin and J. M. Adán," Piegas: A System for Sentiment Analysis of Tweets in Portuguese", 2016, IEEE LATIN AMERICA TRANSACTIONS, VOL. 14, NO. 7

[7] Alexander Porshnev, Ilya Redkin, Alexey Shevchenko," Machine learning in prediction of stock market indicators based on historical data and data from Twitter sentiment analysis," 2013, IEEE, 879234-645-345

[8] LI Bing, Keith C.C. Chan, Carol OU," Public Sentiment Analysis in Twitter Data for Prediction of A Company's Stock Price Movements", 2014, IEEE, 978-1-4799-6563-2

[9] Ryan M. Eshleman and Hui Yang," A Spatio-temporal Sentiment Analysis of Twitter Data and 311 Civil Complaints", 2014, IEEE, 978-1-4799-6719-3

[10] Rincy Jose, Varghese S Chooralil," Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation", 2015, IEEE, 978-1-4673-7349-4

[11] Nehal Mamgain, Ekta Mehta, Ankush Mittal, Gaurav Bhatt," Sentiment Analysis of Top Colleges in India Using Twitter Data", 2016, IEEE, 978-1-5090-0082-1

[12] Aldo Hernández, Victor Sanchez, Gabriel Sánchez, Héctor Pérez, Jesús Olivares, Karina Toscano, Mariko Nakano and Victor Martinez," Security Attack Prediction Based on User Sentiment Analysis of Twitter Data", 2016, IEEE, vol. 56, pp.45

[13] Anurag P. Jain, Mr. Vijay D. Katkar," Sentiments Analysis Of Twitter Data Using Data Mining", 2015 International Conference on Information Processing (ICIP), 978-1-4673-7758-4

[14] Ming Hao, Christian Rohrdantz, Halldór Janetzko, Umeshwar Dayal, Daniel A. Keim, Lars-Erik Haug, Mei-Chun Hsu," Visual Sentiment Analysis on Twitter Data Streams", 2011, IEEE, 3927504-365-4-54

[15] Manju Venugopalan, Deepa Gupta," Exploring Sentiment Analysis on Twitter Data", 2015, IEEE, 978-1-4673-7948-9

[16] Gaurav D Rajurkar, Rajeshwari M Goudar," A speedy data uploading approach for Twitter Trend and Sentiment Analysis using HADOOP", 2015, IEEE, 978-1-4799-6892-3

[17] Changbo Wang, Zhao Xiao, Yuhua Liu, Yanru Xu, Aoying Zhou, and Kang Zhang," SentiView: Sentiment Analysis and Visualization for Internet Popular Topics", 2013, IEEE Transactions on Human-Machine Systems, VOL. 43, NO. 6