International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017),e-ISSN:2455-2585
International Conference on
Recent Explorations in Science, Engineering And Technology (ICRESET'19)
Volume-5, Special Issue-March, 2019.

SENTIMENT ANALYSIS ALGORITHMS COMPARISION IN TWITTER DATASET - REVIEW

N.SRIRAM

Assistant Professor, KG College of Arts and Science, Saravanampatti, Coimbatore – 35.

Abstract: ***Social networking sites like twitter have millions of people share their thoughts day by day as tweets. As tweet is characteristic short and basic way of expression. So in this review paper we focused on sentiment analysis of Twitter data. The Sentiment Analysis sees as area of text data mining and NLP. The research of sentiment analysis of Twitter data can be performed in different aspects. This paper shows sentiment analysis techniques used to perform extraction of sentiment from tweets. In this survey paper, we have taken comparative study of different techniques and approaches of sentiment analysis having twitter as a data.***

Keywords— Sentiment Analysis, Opinion Mining, Social Media, Twitter Data

I. INTRODUCTION

The social networking sites like Twitter, Facebook, and YouTube have obtained so much popularity now days. [1] The area of sentiment analysis is known as opinion mining, it is under umbrella of computational linguistics and data mining. Its main aim is to detect the person's mood, behavior and opinion from text documents. With the expanded use of social networking sites, sentiment analysis techniques have started to use these sites' public data to do sentiment analysis studies in different sociological areas, such as politics, sociology, economy and finance. [2] Most of the data that available in social networks is unstructured [3]. Such unstructured data is almost 80% of the data all over the world. This makes it difficult to analyze and gain valuable judgment from such data. Sentiment analysis or opinion mining is the important technique, which help in detecting opinions of people on social media data. [4] Opinions of others can be important when it is need to make a decision. When those decisions involve valuable resources people think about their companions' past experiences. Now a day's social media gives new tools to conveniently share ideas with peoples linked to the World Wide Web. Though sentiment analysis concentrate on polarity detection (positive, negative or neutral). Twitter is a micro blogging site which contains large number of short length utilizes for marketing, social networking. For example, political parties might be eager to know whether people support their curriculum or not. In present scenario the need to gather opinions from social networking sites and draw conclusions that what people like or dislike, has been the most important perspective. The objective of this review paper is to discuss concept of sentiment analysis of twitter tweet. [5]

II. SENTIMENT ANALYSIS

The area of study that interprets people's opinions, against any particular topic, about any event etc. in text mining it is known as opinion mining or sentiment analysis. It produces a vast problem zone. There are also various names and having different tasks, e.g., sentiment analysis, opinion extraction, opinion mining, sentiment mining, affect analysis, subjectivity analysis, review mining, etc. [6]

Levels of Analysis: In general, sentiment analysis is categorized into mainly three different levels:

A. Document Level Analysis:

This level classifies that whether the complete document gives a positive sentiment or negative sentiment. The document is on single topic is considered. Thus texts which comprise comparative learning cannot be considered under document level.

B. Sentence Level Analysis:

The task of this level is sentence by sentence and decides if each sentence represents opinion into negative, positive, or neutral. Neutral, if sentence does not give any opinion means it is neutral. Sentence level analysis is related to subjectivity classification. That expresses factual information from sentences that gives subjective aspect and opinions. i.e. good-bad terms.

C. Entity/Aspect Level Analysis:

Both the document and the sentence level analysis don't find peoples like and dislikes. Entity/Aspect level gives throughout analysis. Entity/Aspect level was earlier called feature level. The core task of entity level is to identification constructs, aspect level straightforwardly gives attention at the opinion or sentiment. It is based on the concept that an opinion resides of an attitude and a destination of opinion.

III. RELATED WORK

There are various text mining approaches used to mine the data. Prabhsimran Singh, Ravindra Singh and Karanjeeet Singh Kalhon, [7] they have examined this government policy the demonetization from the ordinary person's viewpoint withthe use of the approach of sentiment analysis and using Twitters data, Tweets are collected using certain hashtag (#demonetization). Analysis based on geo-location (State wise tweets are collected). The sentiment analysis API used from meaning cloud and classified the states into six categories, they are happy, sad, very sad, very happy, neutral, and no data.

Xing Fang, Justin Zhan, [8] they have solved the issue of sentiment polarity categorization, and it is one of the basic problems of sentiment analysis. Online product reviews data is used in this study, collected from Amazon.com. In this paper, Investigation for both sentence-level categorization and review-level categorization are achieved. Scikit-learn software is used for this study. Scikit-learn is an open source machine learning software package in Python. Naïve Bayesian, Random Forest, and SVM: These classification techniques selected for categorization.

Geetika Gautam, Divakar yadav, [9] they contribute to the sentiment analysis for customers' review classification.Already labeled twitters data is used in this task. They have used three supervised techniques in this paper: naïve-Bayes, Max-entropy and SVM followed by the semantic analysis which was used along with all three methods to calculate the similarity. They have used Python and NLTK to train and classify the: naïve-Bayes, Max-entropy and SVM. Naïve-Byes approach gives a better result than the Max-entropy and SVM with unigram model gives a better result than using SVM alone. Then the correctness is then increased when the WordNet of semantic analysis is Applied after the above procedure.

Neethu M S, Rajasree R, [10] in this paper, they analyze the twitter data related to Electronic products using Machine Learning approach. They existent a new Feature-Vector for classification of the tweets and extricate peoples' opinion about Electronic products. Thus Feature-Vector is created from 8 relevant features. The 8 features used are specialkeyword, presence of negation, pos tag, and number of positive keywords, emoticon, and number of negative keywords, number of negative hash tags and number of positive hash tags. Naïve-Bayes and SVM classifiers are implemented using built in functions of Matlab. Max-Entropy classifier is implemented using Maximum-Entropy software. All the used classifiers have almost equal performance.

Akshay Amolik, Mahavir Bhandari, Niketan Jivane, Dr.M.Venkatesan, [11] in this paper they proposed a more correct model of sentiment analysis of twitter data about reviews of coming Hollywood and Bollywood movies. With the help of classifiers and Feature-Vector such as SVM and Naïve-Bayes we are accurately classifying these tweets. For sentiment of each tweet. Naïve-Bayes has better precision than to SVM, but slightly lower accuracy and recall. SVM has better accuracy than Naïve Bayes. The Feature-Vector gives more good sentiment analysis than of selected classifier. The accuracy of classification will increase with the increase of training data.

IV. TWITTER

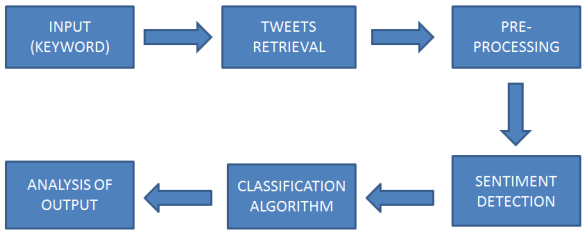
The aim while performing twitter sentiment analysis is classifies the tweets in different sentiment classes accurately. In this field of research, various techniques have evolved, which come up with methods to train a model and then test it o check its effectiveness. Performing sentiment analysis is challenging on twitter tweets. Here we define some reasons for this: [12]

1) Limited tweet size: with just 280 characters in hand, reduced statements are generated, which results sparse set of features.
2) Use of slang: these words are not quite the same as English words and it can make an approach out dated because of the evolutionary use of slangs.
3) Twitter features: it permits the use of hash tags, user reference and URLs. These require different processing in comparison to other words.
4) User variety: the users express their opinions in a different ways, some using different language in between, while others using repeated words or symbols to express their emotion.

All the above problems are required to be faced in the pre-processing section.

IV. SENTIMENT ANALYSIS ON TWITTER DATA

The work flow for sentiment analysis is shown in Figure 2. The system consists of the four main modules: data collection module, data processing module, classification module and analysis of output.



1) Input (Keyword):

First start by choosing a subject then we will collect the tweets with that keyword and perform sentiment analysis on those tweets.

2) Tweets Retrieval:

Tweets can be a structured, semi-structured and unstructured type. Sentiment Analysis research, we can collect tweets using different programming languages like R or python.

3) Pre-Processing:

Data pre-processing is nothing but filtering the data to remove the incomplete noisy and inconsistent data.

Following tasks are involved in pre-processing task:

- Removal of Retweets (in case of twitter dataset)
- Removing URLs, Special characters, Punctuations, Numbers etc.
- Removing Stopwords
- Stemming
- Tokenization

4) Sentiment Detection:

Sentiment word identification is important work in many applications of sentiment analysis and opinion mining, such as tweets mining, opinion holder finding, and tweet classification. Sentiment words can be classified into Positive, Negative and Neutral words.

The fundamental task in Sentiment Analysis is classifying the polarity of a given tweets feature. The polarity is in three classes i.e. Positive, Negative and Neutral. Polarity identification is done by using different lexicons e.g. Bing Lui sentiment lexicon, Senti Word Net etc. which help to calculate sentiment strength, sentiment score, etc. [13]

5) Classification Algorithm:

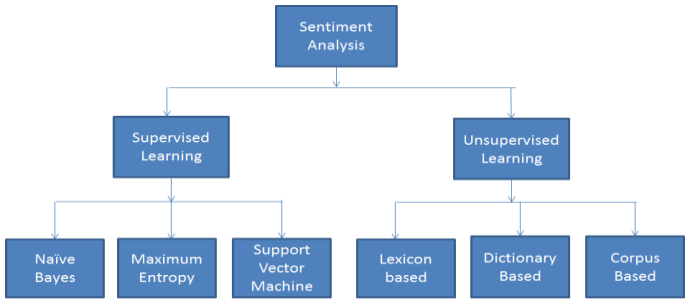


Fig2: Sentiment Analysis Algorithms

Two fundamental approaches are there in sentiment analysis i.e. Supervised learning Approach and unsupervised learning Approach. Sentiment classification of twitter data is done using supervised machine learning approaches like Naïve-Bayes, SVM, and Maximum-Entropy etc. Efficiency of classifier is built upon which dataset is used for which classification methods. In the case of Supervised machine learning approaches to train the classification model Training dataset is used which then help for classification of test data. [13]

6) Analysis of Output:

The fundamental thought of sentiment analysis is to change unstructured data into the significant or meaningful data. After the completion of analysis, the results are displayed on graph like pie chart, bar chart and line graphs.

VI. COMPARATIVE STUDY OF TECHNIQUES OF SENTIMENT ANALYSIS HAVING TWITTER DATASET

The Following table shows the works of various authors on Sentiment Analysis having Twitter dataset.

Author & Year	Dataset	Techniques	Accuracy
Geetika Gautam (2014) [9]	Customer Review Twitter Dataset	Naive Bayes Maximum Entropy SVM Semantic Analysis (WordNet)	88.2% 83.8% 85.5% 89.9%
Neethu M. S. (2013) [10]	Twitter posts about electronic products	Navive Bayes SVM Maximum Entropy Essembled	89.5% 90% 90% 90%
Seyed-Ali Bahrainian (2013) [14]	Twitter data on Smartphones	Unigram feature, SVM, NB, MaxEnt Hybrid Approach	89.78 %
Dhiraj Gurkhe (2014) [15]	Twitter Data	Unigram Bigram Uni+Bigram	81.2% 15% 67.5%
Apoorv Agarwal (2011) [16]	11,875 manually annotated Tweets	Unigram Senti-features Kernel Unigram + Senti-features Kernel + Senti-features	71.35 71.27 73.93 75.39 74.61

Table1: Summary of Research Articles having Twitter Dataset

The analysis of Twitter data is being done in different points of view to mine the opinion or sentiment. This paper defined the concept of sentiment analysis and opinion mining with respect to various levels of sentiment analysis. This survey paper discussed different techniques of sentiment analysis and methodology for sentiment analysis. If we are doing twitter sentiment analysis, it is necessary to know about the twitter, about extricating the tweets, its structure, and their significance.

This paper gives brief idea of tweets. Thus the essential information required to do sentiment analysis of Twitter is well discussed in this review paper. The study of literature shows that, the accuracy is improved when the semantic analysis WordNet is followed up by the machine learning techniques, like SVM, Naïve-Bayes and Maximum entropy. Also the accuracy can be increased up to 4-5% using the Hybrid approach.

REFERENCES

[1] Surnar, Avinash, and Sunil Sonawane. "Review for Twitter Sentiment Analysis Using Various Methods."

[2] Eliacik, Alpaslan Burak, and Erdogan Erdogan. "User-weighted sentiment analysis for financial community on Twitter." *Innovations in Information Technology (IIT), 2015 11th International Conference on*. IEEE, 2015.

[3] Ahmed, Khaled, Neamat El Tazi, and Ahmad Hany Hossny. "Sentiment Analysis over Social Networks: An

Overview." *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*. IEEE, 2015.

[4] Ko, Youngjoong, and Jungyun Seo. "Automatic text categorization by unsupervised learning." *Proceedings of the 18th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 2000.

[5] Kharche, S. R., and Lokesh Bijole. "Review on SentimentAnalysis of Twitter Data." International Journal of Computer Science and Applications 8 (2015).

[6] Liu, Bing. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5.1 (2012): 1-167.

[7] Singh, Prabhsimran, Ravinder Singh Sawhney, and Karanjeet Singh Kahlon. "Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government." ICT Express (2017).

[8] Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data." Journal of Big Data 2.1 (2015): 5.

[9] Gautam, Geetika, and Divakar Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." Contemporary computing (IC3), 2014 seventh international conference on. IEEE, 2014.

[10] Neethu, M. S., and R. Rajasree. "Sentiment analysis in twitter using machine learning techniques." Computing,

Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, 2013.

[11] Amolik, Akshay, et al. "Twitter sentiment analysis of movie reviews using machine learning techniques." International Journal of Engineering and Technology 7.6 (2016): 1-7.

[12] Gupta, Bhumika, et al. "Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python." International Journal of Computer Applications 165.9 (2017).

[13] Jagdale, Rajkumar S., Vishal S. Shirsat, and Sachin N.Deshmukh. "Sentiment Analysis of Events from Twitter Using Open Source Tool." (2016).

[14] Bahrainian, Seyed-Ali, and Andreas Dengel. "Sentiment analysis and summarization of twitter data." Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on. IEEE, 2013.

[15] Gurkhe, Dhiraj, and Rishit Bhatia. "Effective Sentiment Analysis of Social Media Datasets using Naive Bayesian Classification." (2014).

[16] Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." *Proceedings of the workshop on languages in social media*. Association for Computational Linguistics, 2011