Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 03, March -2018

A SENTIMENT MINING APPROACH TO BIG DATA ANALYTICS – COMPARISON STUDY

Mrs.Raviya.K[1] , Dr. Mary Vennila.S[2]

[1]*Assistant Professor & Research Scholar, Department of Computer Applications,*
Guru Nanak College, Chennai.
[2]*Associate Professor & Research Supervisor, PG & Research Department of Computer Science,*
Presidency College, Chennai.

ABSTRACT - *Sentiment Analysis is an unending and enduring research in the Text mining field. It is a technology to extract the sentiment or opinion from a message to obtain the valuable information. Social media big data and machine learning technology are widely used in sentiment classification because of their ability to "learn" from the trained dataset to predict decision with relatively high accuracy. This paper deals with the importance of Big Data Analytics and how it is used in sentiment analysis to track sentiment in social media big data. This paper intends to give a whole image of sentiment analysis techniques and their comparisons.*

Keywords - *Big Data; Big Data Analytics; Sentiment analysis; Sentiment classification; opinion mining; Predictive Analytics, SVM*

I. INTRODUCTION

This Big data is a technique which comprises of a huge volume of both the unstructured and structured data which can be manipulated only by new software and framework. It cannot be performed by classical software and database techniques within a bounded time frame. Big data in reference to the size shows to the data of size larger than a Gigabyte. But, even the smaller size of data will also refer to the 'Big data' as it depends on the context on which it is being used. The previous techniques that existed before 'Big Data' like relational databases are purely different from it. The main difference is that in olden days the single processor is used to perform and analyze the whole or infinite number of information. As in the case of 'Big Data', Infinite number of processors is used to manipulate infinite number of information.[1] When experimenting with machine learning and big data, we may establish data set that contains text content which consists of customer reviews, or social media posts where customers (or potential customers) are telling about a product, trade mark or service that that they offer. Classifying such data to prove how the people think about the trade mark, product, or service, is called Sentiment Analysis. Our Social Media Sentiment Analysis and Big Data Predictive models are used in many scenarios to analyze and predict the data. Data technologies such as Data mining, Social Media Sentiment Analysis, Predictive Modeling, etc. have opened new avenues for businesses across industries. By using these technological innovations, businesses can improve their decision making for both present and future events. In today's age, various sentiment analysis modules and enhancement has been anticipated for finding opinion at various stages.

II. BIG DATA AND BIG ANALYTICS

Big data is gaining importance in various fields such as health care, business, science, research etc. The main aspect of Big Data is to process huge amount of data in parallel to the infinite number of processors. It manipulates, compute, predict, analyze, compare and provide result to any number of information that is fed as input in a big data tool within a short span of time. Basically there are four main characteristics of 'Big Data' which is often mentioned as four V's of Big Data. They are Volume, Velocity, Variety and Value. But, there are also other characteristics such as Veracity, Validity and Volatility. The Big Data analytics was broadly classified into three types. They are Descriptive, Predictive, Prescriptive analytics.

2.1 Descriptive Analytics.

Business intelligence is the broad area where descriptive analysis technique was carried out. It is the beginning stage of data processing that provides some suggestion to make use of historical data for prediction. It uses both data mining and data aggregation methods.[2] In BI, the traditional applications include scoreboard, dashboard, data screening and visualization which are the primary applications. In recent days, the Descriptive Analytics uses the major application to identify and analyze what had happened in the past data and what can be done to improve the decision / prediction. For this, purpose a new analytical technique emerged known as Predictive Analytics.

2.2 Predictive Analytics.

Predictive Analytics is a technique which utilizes the past or the historical data to provide the future prediction with reasonable accuracy in prediction. It can be used in various fields such as weather forecasting, stock prediction, economy variation prediction, etc. The major tool used for Predictive Analytics is R and Hadoop which is a combination of R and Hadoop to provide a better result with more accuracy. The combination of both the predictive and descriptive techniques constitute to a new analytics called Prescriptive Analytics described next.

2.3 Prescriptive Analytics.

Prescriptive Analytics refers to the process of analyzing the abstraction of an exact data related to a particular field to enhance the classification result. It is the combination of both the predictive and descriptive analytics.[3] The major application of prescriptive analytics is Business Intelligence. Prescriptive analysis will always produce a better result as it combines both the trends in market and social network, and unknown statistical relations" a new technique was proposed in accordance with Big Data analytics is referred to as the sentiment analysis.

III. SENTIMENT ANALYSIS

The process of establishing and dividing opinions which are expressed in series of words, especially in order to identify whether a person's attitude on a particular product or topic is either positive, negative or neutral. The main aim of Sentiment Analysis is to bring out the subjective information from the scripts in existing language, in order to form a structured and knowledge which can be put in action either by the decision support system or a person who takes decision.

3.1 Generic sentiment analysis system framework.

The frame work of generic sentiment analysis from Twitter data is shown in Figure-1.

3.1.1 Data extraction- The relevant data are extracted from data source such as twitter, facebook Amazon on which the sentiment analysis will be performed in future.

3.1.2 Preprocessing by tokenization – The extracted words are tokenized to remove the duplications and other noisy data and the normalized text is provided as the input to the next stage.

3.1.3Training the classifier- A part of the obtained words serve as the training data, where the classifier is trained based on specific features.

3.1.4 Sentiment classification- mainly emphasis on the positive, negative and neutral polarities in the test data, with the support of trained classification algorithms.

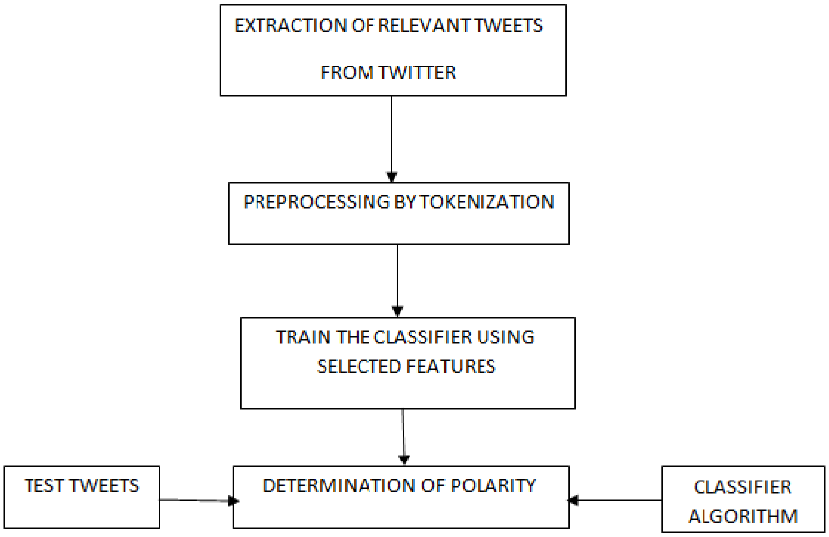
EXTRACTION OF RELEVANT TWEETS FROM TWITTER
PREPROCESSING BY TOKENIZATION
TRAIN THE CLASSIFIER USING SELECTED FEATURES
TEST TWEETS
DETERMINATION OF POLARITY
CLASSIFIER ALGORITHM

Figure-2 presents a detailed classification of sentiment analysis. From the technical aspect this can be classified into four approaches.

4.1 Machine Learning Approaches- Determination of trained dataset is done by this machine learning using the learning algorithms.

4.2 Lexicon-based approach- Sentiment polarity using the semantic orientation of words or sentences in the text is extracted using the lexicon-based approach.

.

4.3 Statistical Approaches- Mixture of hidden aspects and ratings is represented by the statistical models. It is implicated that aspects and their ratings are showcased by polynomial distributions and clustered in terms of aspects and sentiments into ratings.

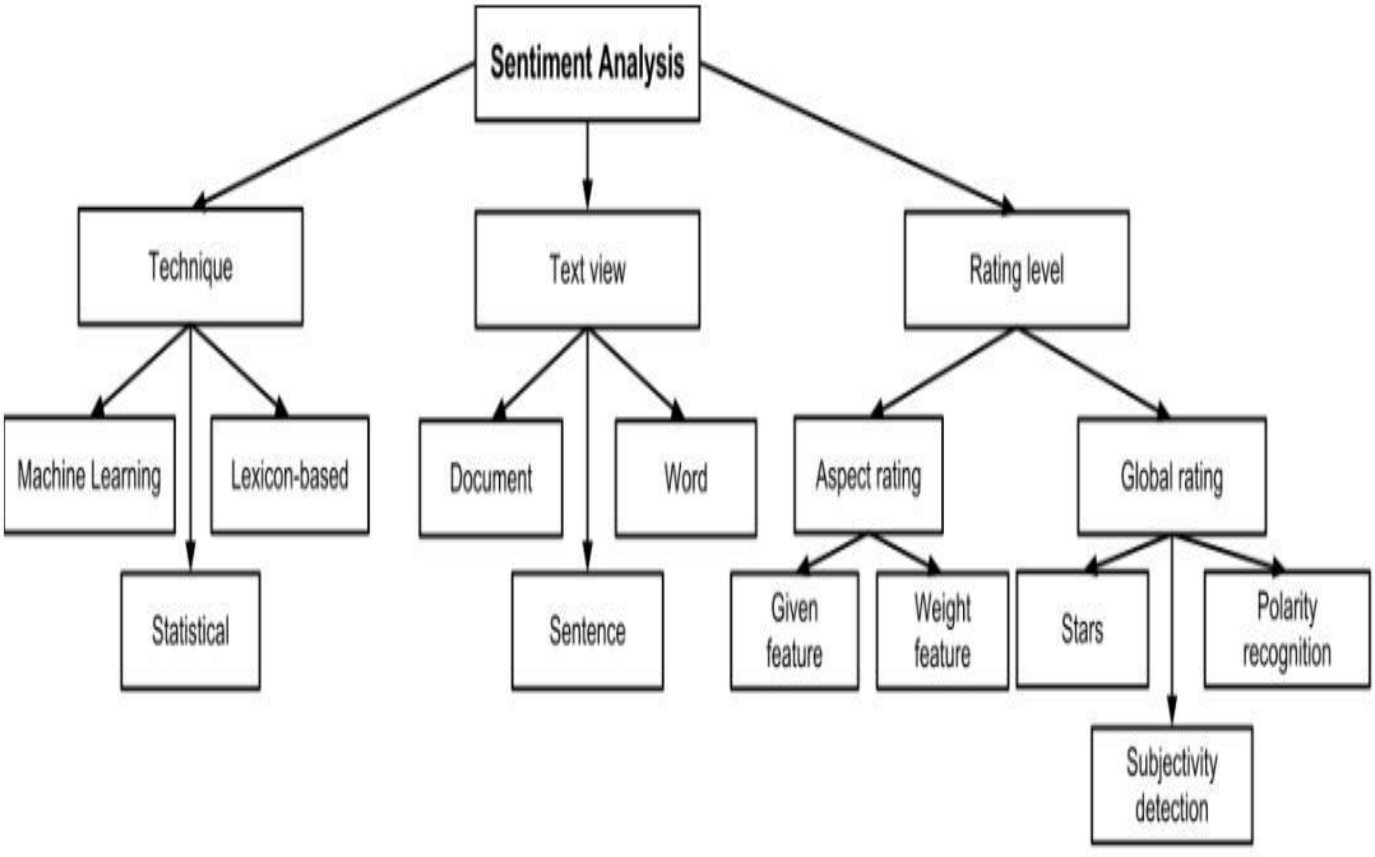

Figure2. Sentiment analysis classification

In text view, Classification is based on the structure of the text such as document level, sentence level or word/feature level classification. Sentiment polarity for the whole review, is mainly done by the Document level classification, sentence level is well expressed for each sentence and also word by word.

Rating level is distinguished by measuring the sentiment strength of a product and attempts to rate to review on a global level.. Solutions that aim a more detailed classification of reviews (e.g., three or five star ratings) use more linguistic features including, negation, possibility and language.

V. SENTIMENT ANALYSIS COMPARISON

Mainly two approaches are focused in this paper such as machine learning based approach and lexicon based approach. In Figure-3 These two approaches further classified into various sentiment analysis methods

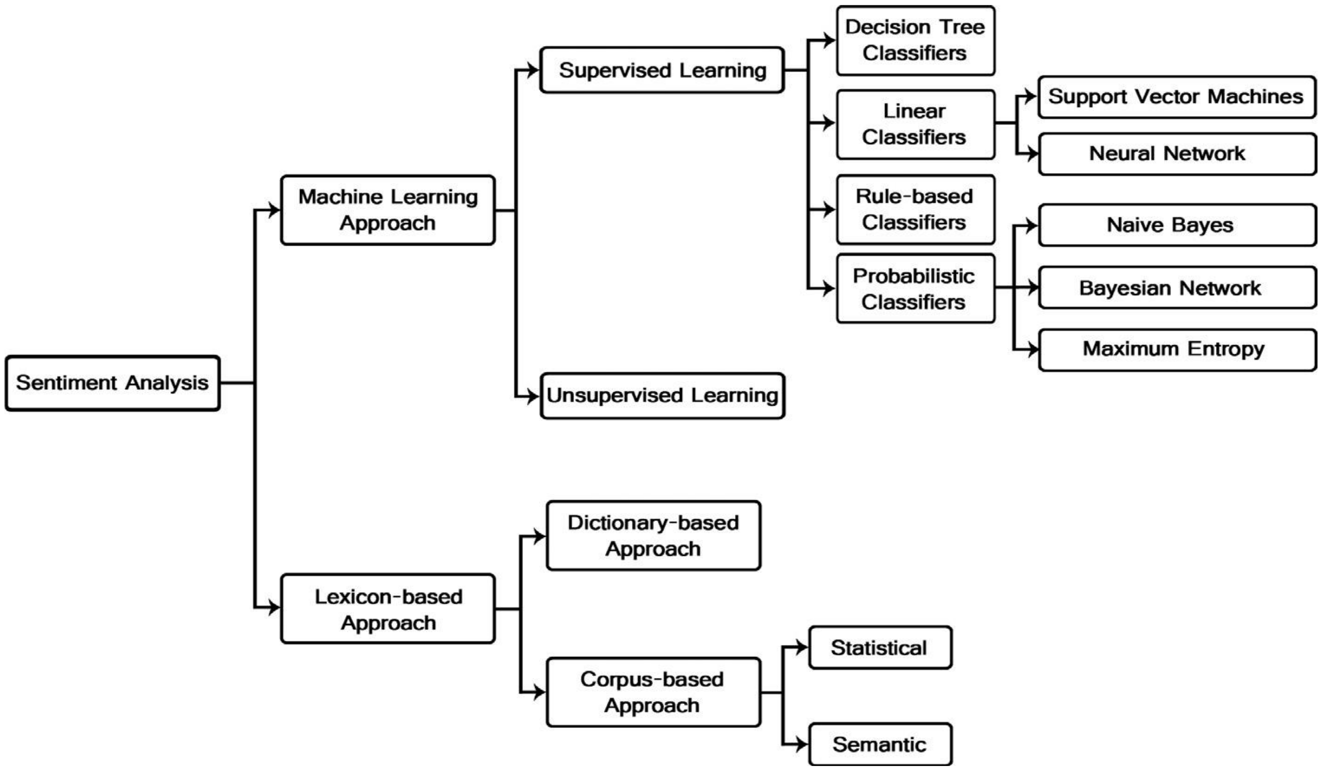

Figure3. Sentiment analysis methods

5.1 The Machine learning approach.

The machine learning approach is a fully automatic tool to analyze the sentiment in large collections of Web data. The supervised and unsupervised learning are the two categories of this approach. Supervised classification algorithms are linear classifier, decision tree, rule based classifier and probabilistic classifier. Supervised learning technique is based on labeled dataset which is provided as input to train the model using trained data to generate output. Sentiment classification in machine learning consists of two steps. First step is to extract feature and store it in feature vector and second step is to train feature vector by using classification algorithms.

5.1.1 Supervised learning.

Supervised learning is the traditional classification algorithm and has been adopted for investigation of opinion detection with satisfactory results. Important supervised classification algorithms are: Naïve Bayes, a generative classifier that estimates prior probabilities of P(a|b) and P(b) from the given data and creates the posterior probability of P(b|a), based on earlier probabilities. SVM is based on the training data and estimate P(b/a) directly and lazy learning algorithm, which does not need prior construction of classification model. Naïve Bayes and SVM are the widely used most popular and effective supervised learning algorithms.

Naïve Bayes is highly suitable for large dataset. Usually the accuracy of NBC is increased when the data size increase. The biggest restriction associated with supervised learning is the sensitivity to the quantity and quality of the training data and this may fail only in the training data. Additional challenge is raised by the opinion detection at the sub-document level.

5.1.2 Unsupervised learning.

It is found that it is sometimes complex to create labeled training documents, in text classification. But the unlabeled documents are collected easily. The difficulties are surmounted by the unsupervised learning methods. The hidden opinion in the text documents are brought out by the traditional models.
The constraint of the unsupervised approach is that a large volume of data is needed to trained system perfectly. Logical topics are produced because of unsupervised models for they do not always match well with human judgments. Because of this disadvantage, unsupervised learning still offers us a way to gain knowledge about the data without any explanatory comment.

5.2 Lexicon-based approach.

Finding the opinion lexicon that is used to analyze the text is totally depended on the lexicon based approach.

This is mainly classified in two categories namely dictionary based and corpus based approach. Wordnet helps in identifying sentiments using synonym and antonym dictionary approach. On the other hand, corpus based approach, identifies opinion words using word list. This can be divided into statistical and semantic approach. Sentiment is identified by the co-occurrences in the statistical approach, whereas the word semantic represents the semantic space to discover the correlation between the terms.

Table 1. Comparison between machine learning and lexicon-based approaches

CRITERIA	MACHINE LEARNING	LEXICON BASED
Domain	Dependent	Independent
Classification Approach	Supervised	Unsupervised
Require Prior Training	Yes	No
Adaptive Learning	Yes	No
Time of result generation	Slow	Fast
Maintenance	Do not require maintenance	Require maintenance of corpus
Accuracy	Higher	Low

VI. PERFORMANCE EVALUATION ANALYSIS

The performance of frequently used sentiment analysis technique on the social media data shows good and promising result. Analyzed papers includes various dataset those are namely movie review dataset, Twitter dataset, Customer dataset (amazon.com, flipkart, cnet.com),. The movie review mining is more challenging than other dataset review because real life word and ironic terms are mixed in movie review. For example unpredictable terms indicate negative opinion but it gives positive opinion for movie review. The performance of sentiment analysis is measured by using the confusion matrix in table-3 which is generated when algorithm is implemented on dataset. Various performance measures are used that are Precision, Recall, F-measure and Accuracy.

Referred papers denotes various classification techniques for sentiment analysis which produces result with vary accuracy. The accuracy can be improved by using such combination of classification techniques. Support vector machine is most widely used classification algorithm for sentiment analysis to high accuracy.

Table 3. Confusion Matrix

	Correct Labels	
	Positive	**Negative**
Positive	True Positive	False Positive
Negative	False Negative	True Negative

Accuracy is one of the performance evaluation parameter and it is calculated by number of correctly predicted reviews divide by total number of reviews present in the corpus.

Table 4. Comparison of Evaluated Result

REFERENCE	DATASET	TECHNIQUE	ACCURACY
[3]	Movie Review	SVM NB	82.9% 81.5%
[4]	Movie Review	Supervised Unsupervised	83.54%
[5]	Movie Review	SVM NB	94% 89.5%
[6]	Customer Review	NB	92.4%
[7]	Movie Review	SVM	83%
[8]	Movie Review	Fuzzy classifier	good
[9]	Customer Review	SVM	78%
[10]	Twitter	SVM NB	85.5% 88.2%
[11]	Unsupervised	Twitter	80%
[12]	Ensemble Classifier (RF,NB,LR,SVM)	Sanders Twitter Stanford Twitter OMD Twitter HC Twitter	84.89% 87.20% 76.81% 78.35%
[13]	Twitter	NB SVM	82.7% 82.2%

SVM performs very well among the given classifier for the given dataset.

VII. THE MAJOR BENEFITS OF SENTIMENT ANALYSIS

Most companies generally obtain a lot of advantages from sentiment analysis today. The direct beneficiaries of sentiment analysis are marketing persons, campaign managers, politicians famous personalities and online shoppers. The important point shows that the companies can individually track positive and negative reviews of their brands, to measure their overall performance, mainly on the online sentiment analysis. This acts as a major component in measuring sales and improving their marketing strategies. To popularize more, some firms develop their own modules while some rely on outsourcing.

Popular personalities and every human being reap the benefit of the idea of sentiment analysis. They can be aware of the public reaction towards them. Every human being benefit out of this sentiment analysis.

VIII. CONCLUSION

This survey portrays the study of all techniques related to Big Data Analytics and sentiment analysis. It is also clear from the study that the sentiment analysis will be the best Analytical technique to predict the actionable insights well in advance. The comparison and performance evaluation of various sentiment analysis techniques result in SVM is the most promising approach to give high accuracy for all type of data. Sentiment analysis offers organizations to analyze various social media data in real time and work systemically.

REFERENCES

[1] Paul C. Zikopoulos, Chris Eaton, Dirk deRoos [2013], “Understanding Big Data.”

[2] Chun□Wei Tsai1, Chin□Feng Lai, HanChieh Chao and Athanasios V. Vasilakos [2015], “Big data analytics: A survey

[3] Liu, B., 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), pp.1-167.

[4] Thien Hai Nguyen, Kiyoaki Shirai, Julien Velcin [2015], “Sentiment analysis on social media for stock movement prediction.”

[5] Pang, B., Lee, L. and Vaithyanathan, S., 2002, July. sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

[6] Chaovalit, P. and Zhou, L., 2005, January. Movie review mining: A comparison between supervised and unsupervised classification approaches, Hawaii International Conference IEEE

[7] Tripathy, A., Agrawal, A. and Rath, S.K., 2015. Classification of Sentiment Reviews using Machine Learning Techniques.Procedia Computer Science, 57, pp.821-829

[8] Shahana, P.H. and Omman, B., 2015. Evaluation of Features on Sentimental Analysis. Procedia Computer Science, 46, pp.1585-1592.

[9] Jeyapriya, A. and Selvi, K., 2015, February. Extracting aspects and mining opinions in product reviews using supervised learning algorithm. In Electronics and Communication Systems (ICECS), 2015 2nd International Conference on (pp. 548-552). IEEE.

[10] Mouthami, K., Devi, K.N. and Bhaskaran, V.M., 2013, February. Sentiment analysis and classification based on textual reviews. In Information Communication and Embedded Systems (ICICES), 2013 International Conference on (pp. 271-276). IEEE.

[11] Bhadane, C., Dalal, H. and Doshi, H., 2015. Sentiment analysis: Measuring opinions. Procedia Computer Science, 45, pp.808-814.

[12] Gautam, G. and Yadav, D., 2014, August. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In Contemporary Computing (IC3), 2014 Seventh International Conference on (pp. 437-442). IEEE.

[13] Khan, F.H., Qamar, U. and Javed, M.Y., 2014, November. Sentiview: Avisual Sentiment analysis Frame Work. In Information Society, 2014 International Conference on (pp. 291-296). IEEE.

[14] da Silva, N.F., Hruschka, E.R. and Hruschka, E.R., 2014.Tweet sentiment analysis with classifier ensembles. Decision Support Systems, 66, pp.170-179.

[15] Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1, p.12.

[16] Analysis of Various Sentiment Classification Techniques. Bhumika M. Jadav, Vimalkimar B. Vaghela. International Journal of Computer Applications (0975 – 8887) Volume 140 – No.3, April 2016.