© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 09 | October – December 2016
Paper is available at www.jrps.in | Email : info@jrps.in

A Comparative Study on Sentiment Analysis Techniques

Nikhil Kalraiya
Infinity Management & Engineering College Sagar (M.P.)
RGPV,Bhopal
Sagar (M.P.), India
kalraiya.nikhil@gmail.com

Asst.Prof Sarvesh Rai
Infinity Management & Engineering College
RGPV, Bhopal
Sagar (M.P.), India
sarvesh.s51@gmail.com

Abstract

The fastest growing popularity of E-commerce website, blogs, social Medias, forums, etc. created a new platform where everyone can explore and exchange their views, suggestions, ideas and events about any product or services. This new moment assembled a huge amount of data generated by user on the web. If this data can be draw out and examine properly then it can act as a key factor in decision making. But human extraction of data and examine of this content is an impossible task, because the data is unstructured in nature and it is written in natural language. This condition opened a new area of research called Sentiment Analysis or Opinion Mining. Data mining have extensions as Opinion Mining and Sentiment Analysis; it extracts and examines the unstructured data automatically. The main purpose of this paper is to compare the main concept used in Opinion Mining and Sentiment Analysis, with proposed work.

03
9 772278 684800
© iJRPS International Journal for Research Publication & Seminar

Keywords: Sentiment Analysis, Opinion Mining, Natural Language Processing (NLP), Sentiment Lexicon, support vector machine (SVM).

1. Introduction

Textual data on the internet is growing day by day. To mine and analysis the textual information is becoming more difficult day by day. The text is usual data format on the web, since it is easy to produce and publish. Buyer and producers both are advantageous of this content: Buyer can consider others opinion and experience while taking decision about any particular product or services and producers can get clear idea about their product from the Buyer point of view and thereby buyer can increase the grade of the product.morover the main challenging task is extracting data and analyzing the useful things from the content of data..The unstructured data and the natural language used to write these content added up the complexity more and it opened a new area of research for researcher is called Opinion Mining and Sentiment Analysis. Sentiment Analysis involves determining the evaluative nature of a piece of text data. For example, a product review can express a positive, negative, or neutral sentiment. By sentiment Analysis improves the customer relation models, detect the happiness and well-being, and improve the automatic dialogue systems. Twitter is a micro blogging services which is used in wed as well mobile also. There is huge interest in sentiment analysis of short informal texts message, such as tweets and SMS messages, across a variety of domains (e.g., e-commerce, health service, military intelligence, and disaster management). Sentiment Analysis has gained popularity in recent years due to its on the spot applicable in business environment, such as summarize the feedback from the product reviews, discover the collaborative recommendations, or assist in election campaigns. Sentiment Analysis and Opinion Mining is a Natural Language Processing (NLP) technique that automatically extracts the sentiments, emotions, opinion, attitude, views etc. in proper context and then categories it into different categories like positive, negative , neutral etc. The two important steps include in Opinion Mining and Sentiment Analysis are 1) Opinion Extraction: extracting the generated phrases, in proper context, from free text and (2) Sentiment classification: classifying generated phrases based on sentiment orientation. It utilizes various machine learning techniques such as Naïve Bayes, SVM, character Based N-gram model etc. for sentiment classification.

The main three levels of sentiment Analysis are the document level, aspect level and the sentence level. This classification depends on the different levels of analysis. The first one document level is known as document-level sentiment classification because the main task is to determine the document, if the document as a whole opinion has a negative or a positive sentiment. This method is not the most suitable one for texts with entities comparison or evaluating more than one entity. The other sentence level is very similar to the document level, but with the main difference that in this case each sentence is analyzed individually to see if it expresses a negative, neutral or positive opinion. In the aspect level, most fine-grained analysis as compare to the other level unlike the sentence and document levels, the aspect level discovers what each opinion is about. The main difference in this is that analysis finds a target for each opinion, instead of focusing on language units, like sentences, documents or paragraphs. The goal of this level is to identify the opinion or sentiment on entities and their different aspects. The majority of real-time sentiment analysis systems are based on this level.

Sentiment Classification is mainly divided into two different approaches: the machine learning approach and lexicon- based approach. below figure of types of SA

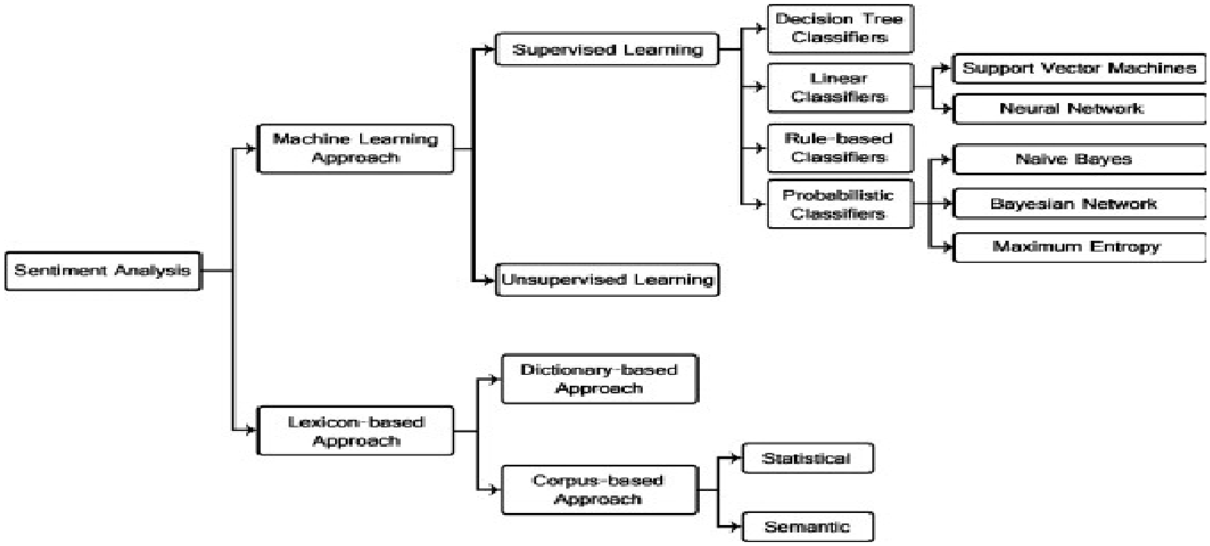

Fig: 1 Types of Sentiment Analysis

2 Machine learning based Approach

Machine learning based approaches classification can be done in two ways: 1) Sentiment Analysis by using supervised machine learning techniques and 2) Sentiment Analysis by using unsupervised machine learning techniques. In approaches, two types of data sets are required: training dataset and test data set. An automatic classifier learns the classification factors of the document from the training data set and the accuracy in classification can be finding using the test set. The machine learning algorithms like maximum entropy (ME), Naive Bayes (NB) and Support Vector Machine (SVM).Supervised learning problems can be further grouped into regression and classification problems.

- **Classification**: A classification problem is when the output variable is a category, such as “black” or “white” or “rain” and “no rain”.
- **Regression**: A regression problem is when the output variable is a real value, such as “money” or “measure”.

Unsupervised learning is a type of machine learning algorithm used to draw conclusion from content consisting of input data without labeled responses. Because unlike supervised learning above there is no right answers and there is no teacher. Algorithms are left to their own tool to discover and present the interesting structure in the data set.

Unsupervised learning problems can be further grouped into clustering and association problems.

- **Clustering**: A clustering issue where you want to discover the inherent groupings in the data, such as grouping consumer by purchasing action.
- **Association**: An association where learning problem want to discover rules that describe big portions of your data

3 Lexicon Based Approach

This approach compared to supervised learning, lexicon-based unsupervised learning uses a sentiment dictionary, which doesn’t require storing a large data corpus and training - which makes the whole process much faster Lexicon Based Method is an Unsupervised Learning approach since it does not require prior training data sets. Sentiment lexicon can be constructed in three ways: 1) manual lexicon establishes, 2) corpus-based lexicon construction and 3) dictionary-based lexicon construction.

4 Comparative Analysis of different Technique

Comparison table below shows most of the cases the supervised machine learning methods outperformed the unsupervised learning based methods. But, the demand of big labeled training data set for supervised machine learning approaches; pressurize the researchers to adopt the unsupervised methods, as it is very easy to collect unlabelled dataset.

To obtain the results using Accuracy of the classification, confusion matrix was used. The sentiment value of the tweets is classified into ‘0’ – showing positive sentiment and ‘1’ – showing negative sentiment.

The accuracy is used as a statistical measure of how well a binary classification test rightly discovers the sentiment of the tweets as positive or negative. In other words, accuracy is the proportion of true results (true positives and true negatives) between the total numbers of cases.

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}$$

The accuracy of the proposed ‘Extremely Randomized Tree’ classifier is found to be the highest, as shown in the figure below:

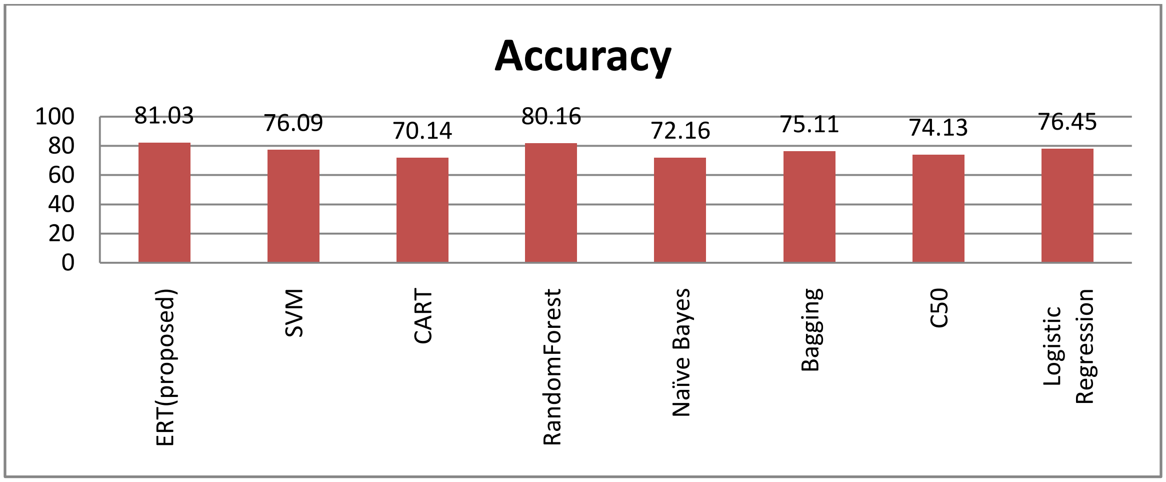

Table 1: show accuracy of sentiment analysis on proposed work

The accuracy of classification of the sentiment for proposed approach is highest. It means that using the ‘Extremely Randomized Tree’ approach, 81.03% of the tweets in test data are correctly classified as showing the actual sentiment of the users towards the event ‘Demonetization in India 2016’.

Conclusion

The classifiey model built using ‘Extremely Randomized Tree (ERT)” near was applied to the test data. The classifiey sentiment values had been compared with the actual sentiment values to do result analysis. After executing the machine learning process for twitter sentiment analysis, the result was evaluated on Accuracy of the Classification.It is shown in this research the approach of supervised machine learning classifier ‘Extremely Randomized Tree’ has a major effect on the overall accuracy of the analysis. This approach has an accuracy of around 81% for classification.The simulation of the given technique was presented in ‘R’ language.

References

[1] A. Abbasi, H. Chen, and A. Salem, “Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums,” In ACM Transactions on Information Systems, vol. 26 Issue 3, pp. 1-34, 2008.

[2] A. Khan, B. Baharudin, K. Khan; “Sentiment Classification from Online Customer Reviews Using Lexical Contextual Sentence Structure” ICSECS 2011: 2nd International Conference on Software Engineering and Computer Systems, Springer, pp.317-331, 2011.

[3] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B.Liu, “Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis”, Technical report, HP Laboratories, 2011.

[4] W. Zhang, H. Xu, W. Wan, “Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis,” Expert Systems with Applications, Elsevier, vol. 39, 2012,pp. 10283-10291.

[5] Neha Raghuvanshi, Prof. J.M. Patil “A Brief Review on Sentiment Analysis” International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016 978-1-4673-9939-5/16/$31.00 ©2016 IEEE

[6] Mr. S. M. Vohra, Prof. J. B. Teraiya “A Comparative Study Of Sentiment Analysis Techniques” Journal Of Information, Knowledge And Research In Computer Engineering Issn: 0975 – 6760| Nov 12 To Oct 13 | Volume – 02, Issue – 02

[7] Zhao jianqiang, Cao xueliang” Combining Semantic and Prior Polarity for Boosting Twitter Sentiment Analysis” 2015 IEEE International Conference on Smart City/SocialCom/SustainCom together with DataCom 2015 and SC2 2015.

[8] ChetanKaushik, AtulMishra”Comparative Analysis of Sentiment Analysis Techniques” ISSN (PRINT) : 2320 – 8945, Volume -2, Issue -1,2014.

[9] Chiyu Cai1, Linjing Li1 Daniel Zeng “ New Words Enlightened Sentiment Analysis in Social Media” 978-1-5090-3865-7/16/$31.00 ©2016 IEEE.

[10] Pankaj Gupta, Ritu Tiwari and Nirmal Robert “Sentiment Analysis and Text Summarization of Online Reviews: A Survey” International Conference on Communication and Signal Processing, April 6-8, 2016, India.