ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 10, October 2018

Sentiment Analysis of Twitter tweets through Classification model

Kondru Ravi[1], Naresh Nelaturi [2]

PG Student, Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, India[1]

Research Scholar Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, India[2]

ABSTRACT: With the advent of microblogs and social network sites a huge amount of data generating every day. The data generated by these blogs contains unstructured data which is unable to process by the traditional databases. Lack of handling of this type of data by traditional databases have opens the development of new technologies such as big data, machine learning and artificial intelligence. These technologies providing sophisticated tools for analyzing the data to find the hidden patterns. Everyday Social networking sites like Twitter, Facebook, Instagram generating terabytes of data as they allow people to share and express their views about topics, have discussion with different communities, or post messages across the world. Sentiment analysis is a process of automatically identifying the opinion of users whether it is positive, negative or neutral about an entity in a user-generated text. This paper focuses mainly on sentiment analysis of twitter data which is helpful to analyze the information in the tweets where views are highly unstructured, heterogeneous and are either positive or negative, or neutral in some cases. In this paper, we propose a machine learning approach to enhance the sentiment classification by adding semantics in feature vectors and thereby using support vector machine for classification.

KEYWORDS: Natural language processing (NLP), Sentimental Analysis, Supervised Algorithm, Support vector machine.

I. INTRODUCTION

A Sentiment analysis related to wide area in data mining that covers the people's opinion about particular aspects like politics, product related issues, services and events. Today most of the people share their opinion at social networking sites like twitter, facebook, instagram. Rapid growth of social networks and micro blogging websites has given a richer source of information to perform sentiment analysis. Sentiment analysis allows the people to know the product results in the market and analyze pros and cons of that product. Twitter is a part of social networking site every day millions of tweets are generated. It provides an interface to people to post their opinions and discuss their views across the world in the form of a tweet. Twitter is generating a large volume of sentiment rich data in the form of tweets like status updates, blog posts, comments and reviews.

To perform sentiment analysis on the text that involves application of computational powers in understanding sentiment implied in the text [3]. Natural language processing techniques used to find the semantic meaning of the text and thereby analyzing the text information. For feature vector construction various natural language processing techniques are applied at pre-processing level such as Stemming, Stop Words removing, Parts of Speech Tagging (POS), Named Entity Recognition (NER).The text classified into three categories such as positive, negative and neutral. The sentiment classification can be done in various levels like documentation level, sentence level, and sub-sentence level.

In the documentation level the entire document can be used to classify the sentiment either positive or negative. In sentence level sentiment classification first classify each sentence either subjective or objective then classify the sentence into positive or negative. In sub-sentence level sentiment analysis obtains the sentiment of sub-expressions within a sentence.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 10, October 2018

II. MOTIVATION AND RELATED WORK

In the past decade people using social network sites to communicate with others, such as micro blogging and text messaging have emerged and become ubiquitous. Twitter provides short-range messages to users and there is no limit to share their opinions in the form of tweets. The twitter generates a huge amount of data every day. It is difficult to find the entire opinion about particular entity (person, product, service, issue). By using sentiment analyzing techniques people can easily find the opinion for particular entity. It reduces the user time and gives an accurate opinion. Sentiment analysis and opinion mining were first introduced in the year 2003. Several techniques used for opinion mining in history. The following few works related to this technique;

Pang proposed supervised machine learning techniques to do sentiment classification [3]. In his paper he uses various machine learning approaches such as Naive Bayes, maximum entropy classification, and support vector machines .This methods commonly used for topic classification.

Paper [1], compares two supervised machine learning algorithms of Naive Bayes and Decision tree for sentiment classification of the twitter tweets. The results show that the Decision tree approach outperformed the Naive Bayes approach for datasets of varied sizes and domains.

G.Vinodhini presents a survey covering the techniques and methods in sentiment analysis and challenges appeared in these field. The author gives the summary of where data collected and what are the algorithms mostly used in the machine learning classification [4]. The nature language processing techniques (NLP) used in this area, especially in the document sentiment detection.

Paper[15], compares three supervised machine learning algorithms of SVM, Naive Bayes and KNN for sentiment classification of the movie reviews that contains 1000 positive review and 1000 negative reviews. The results show that the SVM approach outperformed the Naive Bayes and k-NN approaches and the training dataset had a large number of reviews, the SVM approach reached accuracies of more than 80%.

III.APPROACHES OF SENTIMENT ANALYSIS

There are several approaches to perform sentiment analysis some major approaches are

- Lexcon based approach
- Machine learning based approach
- Hybrid based approach

In the lexicon based approach the sentiment analysis can be performed based on some rules, this is restricted to some particular rules only .It can be performed in 2 ways first one is dictionary based approach in this approach performance depends upon the size of dictionary. Second one is corpus based approach, it depend on large corpora for syntactic and semantic patterns of opinion words. The words that are generated are context specific and may require a huge labelled dataset. In machine learning approach automatic system that works based on some machine learning techniques. Various machine learning techniques are there like supervised, un-supervised, semi supervised approaches. Hybrid based approach- in this approach more than one (2 or more) approaches are used in a single combinational approach that gives more accurate results.

IV. PROPOSED SYSTEM

The data present in the twitter is in unstructured format. Before we perform sentiment analysis on twitter data we must convert the data into a proper format for further processing. The following steps are involved in twitter sentiment analysis.

Data collection: Most API's are not allowed collecting data but twitter provides streaming API (Application program interface) to collect last 6-9 days historic data. By using user keys provided by the twitter to collect the required tweets.

Twitter follows its own language conventions, some examples of twitter conventions are listed below

a) Length: Twitter restricted 180 characters per tweet.

b) Web links: Twitter allows URL's to post user account (eg. https://www.google.co.in)

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 10, October 2018

c) Acronym: Twitter allow user to use acronyms in their posts."RT" is an acronym for re-tweet, which means either user repeating or reposting.
d) Special symbols: Some special symbols are used in twitter conventions such as @, #, ~,

Data Pre-processing: Once tweets collected we need to pre-process the tweets. In the twitter data there are many challenges to face before processing the data. the tweets have grammatical mistakes, punctuation marks, stop words, spelling mistakes, etc. We need to do various operations to convert the collected twitter data into a structured format. Some of such operations listed below
a) Remove the re-tweeted symbols present in the data.
b) Removing all the URL links present in the data.
c) Stop word removal: Remove any word that does not help in analyzing the sentiment of a tweet. These words are called stop words
d) case conversions: All the tweets converted into either lower case or upper case to remove the variations like High and high.
e) Punctuation Removal: We need to remove all the punctuation marks like comma or colon often there is no use of it.
f) Word spelling: Correct and convert all the misspelled words into correct words.
g) Stemming: It usually refers to a simple process that chops off the ends of words to remove derivation affixes.
Feature Extraction: We need to extract some features from the pre-processed tweets. Some of the features are listed below
a) Term presence and Frequency: It generally uses to find each term occurrence and their frequency in text.
b) Negations: The text may contain various negative words like 'nor', 'not', 'neither'. It completely change the meaning of sentence e.g. "not good".
c) Parts of speech tagging (POS): Applying POS is very important for a sentiment analysing process.POS is applied according to our sentiment Dictionaries, it works based on language dependent.
d)Emoticons: To find the emoticons present in the text
Training and Testing Machine Learning Classifier: After selecting the features, we have to choose a machine learning classifier for sentiment analysis.The data can be classified into train data and test data.Train data used to train the classifier and its performance is measured using test data.
This paper proposing Support vector machine classifier to perform sentiment analysis. Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification as well as regression challenges. SVM classifier uses a kernel function to map a low dimensional feature space to a higher dimensional space to separate classes.

V. RESULTS

The Support vector machine is trained and tested against the data from twitter. The performance of the classifier is shown below.

Algorithm	Accuracy	Sensitivity	Specificity
Support Vector Machine	95.95 %	98.21	88.89

Table.1. Metric values of SVM classifier

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 10, October 2018

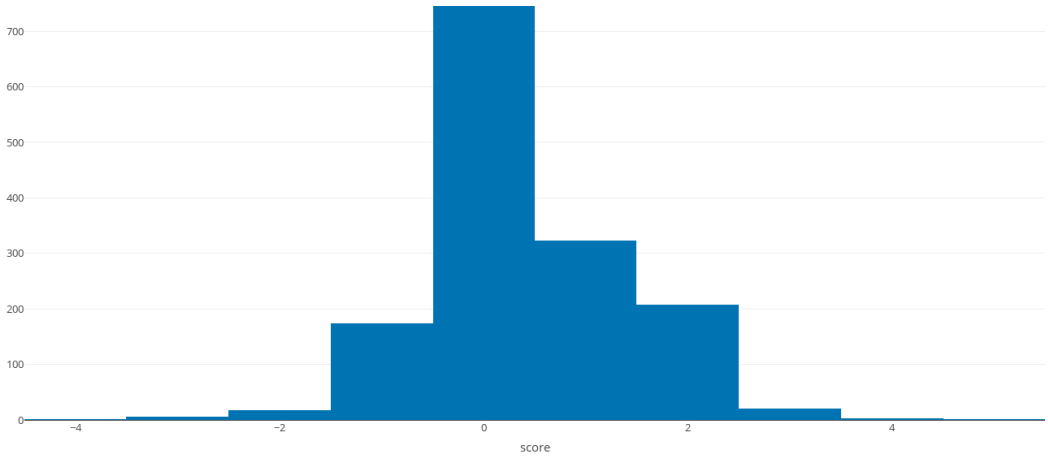

Fig.1 Polarities of the text

Above Figure (Fig.1) shows the polarity of the words and their frequencies. It indicates >1 as positive words and <1 as negative words and 0 as neutral words.

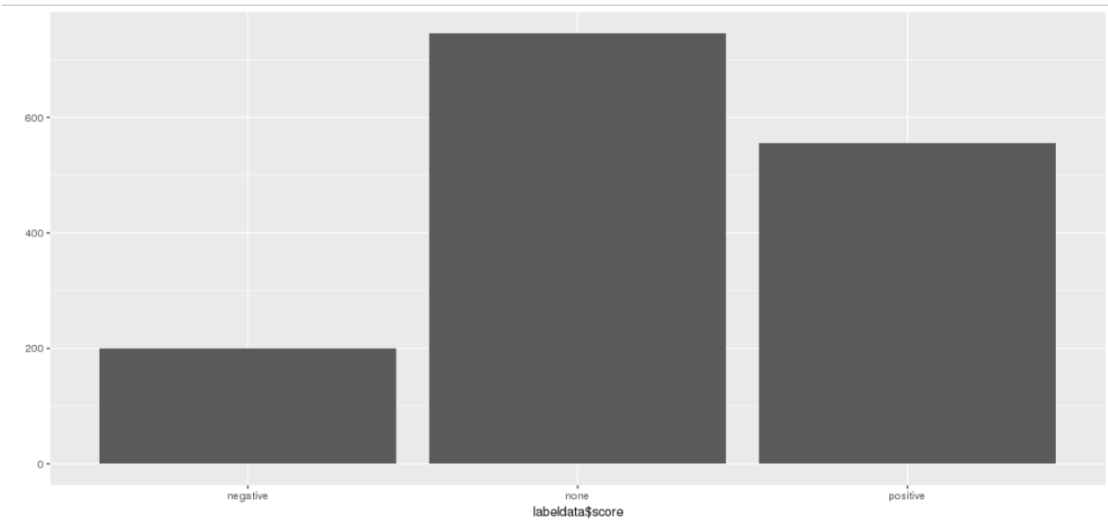

Fig.2. Collecting positive and negative polarities of the text

Above Figure (Fig.2) shows the polarity of the words and their frequencies. To remove the neutral values.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 10, October 2018

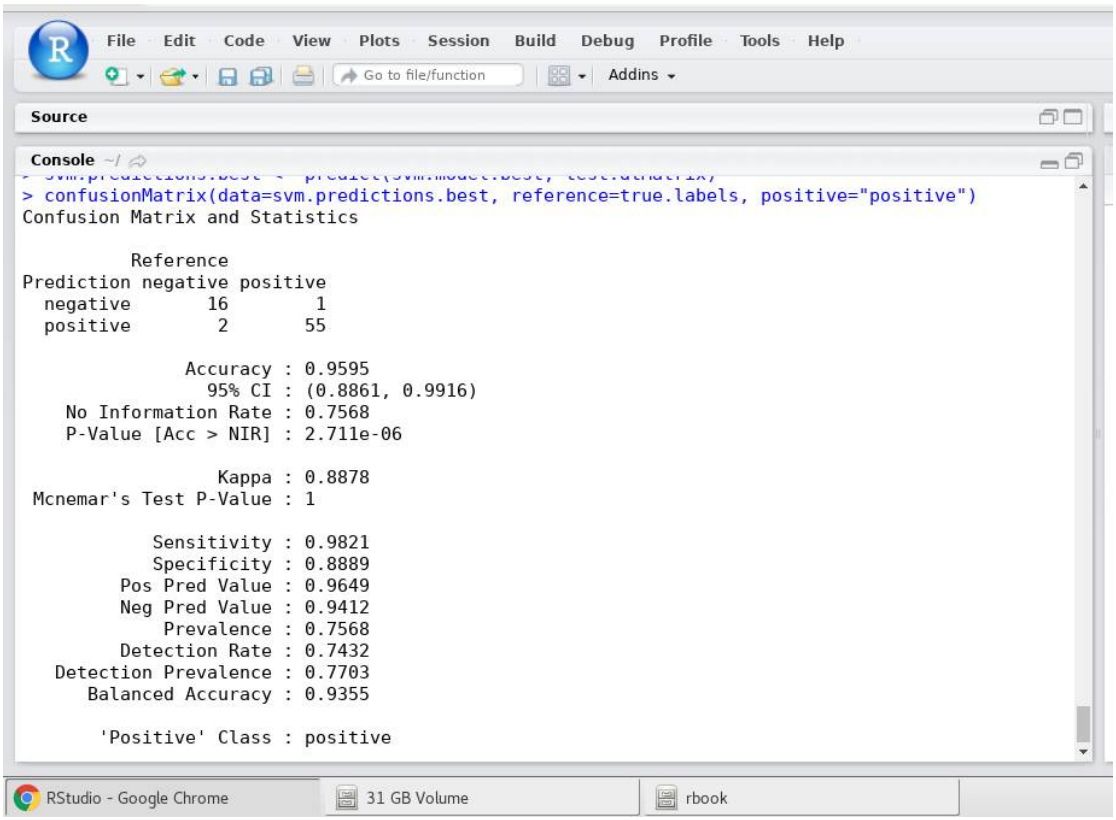

Fig.3. Accuracy of SVM classifier

Above Figure (Fig.3) shows the confusion matrix and accuracy of the SVM Classifier.

VI. CONCLUSION

This paper explains in detail various steps for performing sentiment analysis using machine learning classifier. A machine learning classifier requires a labeled data which is divided into train and test data. Once an appropriate data is collected, the next step is to perform preprocessing on data (tweets) by using NLP based techniques, followed by feature extraction method in order to extort sentiment relevant features. Finally, a model is trained using machine learning classifiers Support vector machine. By using this classification technique it's been concluded that SVM will outperform other classifiers till date and it is best used classifier.

REFERENCES

[1]. Anuja P jain,Padma Dandannavar , "Application of machine learning techniques to sentiment analysis", in 2nd International Conference on Applied and Theoretical Computing and Communication Technology, 2016.
[2]. Suchita V Wawre1, Sachin N Deshmukh"Sentiment Classification using Machine Learning Techniques", International Journal of Science and Research (IJSR), Volume 5 Issue 4, April 2016.
[3]. Monisha Kanakaraj* and Ram Mohana Reddy Guddeti, "NLP Based Sentiment Analysis on Twitter Data Using Ensemble Classifiers", in 3rd International Conference on Signal Processing, Communication and Networking (ICSCN),2015.
[4]. G.Vinodhini,RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012.
[5]. Vishal A. Kharde,S.S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques", International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 10, October 2018

[6]. Saif, Hassan, Yulan He, and Harith Alani. "Semantic sentiment analysis of twitter." International Semantic Web Conference. Springer Berlin Heidelberg, International Semantic Web Conference pp. 508–524, 2012.
[7]. Richa Sharma, Shweta Nigam and Rekha Jain "Opinion mining of movie review at document leve"l, International Journal on Information Theory (IJIT), Vol.3, No.3, July 2014.
[8]. P.Kalaivani, Dr.K.L.Shunmuganathan, "Sentiment classification of movie review by supervise machine learning approach", Indian Journal of Computer Science and Engineering (IJCSE) Vol. 4 No.4 Aug-Sep 2013.
[9]. Hemalatha1, Dr. G. P Saradhi Varma, Dr. A.Govardhan,"Sentiment Analysis Tool using Machine Learning Algorithms ",International Journal of Emerging Trends & Technology in Computer Science Volume 2, Issue 2, March – April 2013.
[10]. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Proceeding of the 18th ACM Conference on Information and Knowledge Management, pp. 375–384. ACM ,2009.
[11]. Yu, B.: An evaluation of text classification methods for literary study. Literary and Linguistic Computing 23(3), 327–343, 2008.
[12]. Andrea Esuli and Fabrizio Sebastiani, "Determining the semantic orientation of terms through gloss classification", Proceedings of 14th ACM International Conference on Information and Knowledge Management, pp. 617-624, Bremen, Germany, 2005.
[13]. Bo Pang and Lillian Lee and Shivakumar Vaithyanathan "Thumbs up? Sentiment Classification using Machine Learning Techniques", Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86.
[14]. Richa Sharma, Shweta Nigam and Rekha Jain "Opinion mining of movie review at document leve"l, International Journal on Information Theory (IJIT), Vol.3, No.3, July 2014.
[15]. P.Kalaivani, Dr.K.L.Shunmuganathan, "Sentiment classification of movie review by supervise machine learning approach", Indian Journal of Computer Science and Engineering (IJCSE) Vol. 4 No.4 Aug-Sep 2013.