IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

Study of Sentiment Classification Techniques

S. Sharma[1*], D. Singh[2]

[1*] Dept. of Computer Science and Engineering, DCRUST, Murthal, Sonepat, India
[2] Dept. of Computer Science and Engineering, DCRUST, Murthal, Sonepat, India

[*]*Corresponding Author: sharma1016.savi@gmail.com, Tel.: 9896742988*

Available online at: www.ijcseonline.org

Accepted: 19/May/2018, Published: 31/May/2018

Abstract- Sentiment analysis is important part of text mining. From the last few years social networking sites like Facebook, Twitter, and Amazon generates large amount of data, such vast amount of data contains lots of useful information which need to be analyzed. As we all know social networking sites generate data in huge amount not only in terabyte but in petabyte so processing of such large amount of data is big challenge, sentiment analysis is the technique which help to analyzie such raw amount of data and extract useful information from it. The reason behind using Sentiment analysis is that it analyze such large amount of data and extract useful information from it that it analyze such large amount of data and extract useful inform Sentiment analysis helps business and organization because it's easy for them to know how people feel about their product or services so that they can make better decision or improve their services. For that purpose we have different sentiment analysis techniques like Naïve bayes, Maximum Entropy, SVM which gives correctness of information or provides us accuracy. For sentiment we use machine learning because it train the computer to recognize the emoticon behind the sentence.

Keywords—: Sentiment analysis, Machine Learning. Sentiment Technique

1. INTRODUCTION

As we all know that social media play a very important role in our day to day life. Data is rush in various social media site like Facebook, twitter, yahoo, YouTube, and many more which has given people new way to express their opinion related to any product, person and place. it allow us to give opinion what we feel what we think about product or our view related to movies .so sentimental analysis is the process which will evaluate whether the give sentence come under positive, negative or neutral. It basically measure the people opinion related to product or anything that happen on social media. The huge amount .of data is stored online here sentimental is use to extract useful information based on some technique. Sentimental analysis is dealing with the people's opinion their attitude and emotions toward a unit. Now a day's data is in large amount for that using traditional technique for mining not give good result. For that we use machine learning technique to handle large amount of data and extract useful information from it.

To sentiment the data here are some methodology like extract data from any site like twitter, amazon etc. Next step is cleaning process where irrelevant data is removed then on the pre-processed data apply feature selection which extract useful data from bag of words then give training and testing to it. Finally apply classifier algorithm which gives accuracy of that classifier.

In sentimental analysis there are three are three level of classification: document level, sentence level, and aspect level. In **document level** classification it determine the overall classification of document.in case of **sentence level** classification it differentiate the subjective information from the objective information. Here it evaluate each sentence and determine whether it come under positive sentence, negative sentence or neutral sentence. Last we have **aspect analysis** here it differentiate what user want and what it doesn't.in this analysis it directly look on the opinion rather than look into any paragraph, document or sentence. For example battery of mi phone is very good, here battery is the feature of phone and "very good" is the opinion. [1]

II. Methodology of sentiment

Fig 1: Methodology of Sentiment Analysis

A. Input data

Testing data we have is in the form of anything like Review from "the times of India" or we can collect the data from any social networking site, like review site or any blogging site.as we all know IMBD is very popular site for movie review while flipkart.com is good source for product review.so that when we select any particular movie from the dataset reviews regarding that movie are displayed on the webpage. [2]

B**. Pre-processing**

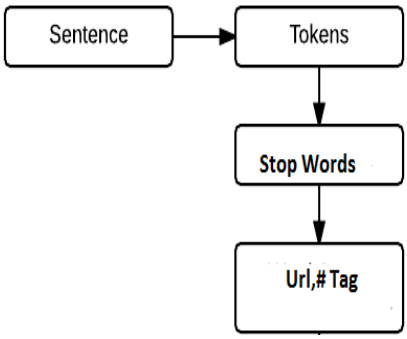

Fig 2: Pre-processing of Data[http://fotiad.is/blog/sentiment-analysis-comparison/]

The pre-processing technique is further divided into sub-categories as shown in [fig2].

1. Tokenization: Data is present in the form of text which contain block of characters called tokens.in this these text document are separated as a token and further use for processing of data.
2. Removal of stop words: A web search tool or any other natural language may contain stop words. Most of the frequently used stop-words in English are "an", "a", "the" or many more which don't carry any meaning so here we need to remove these stop words in pre-processing step.
3. Removal of URL's: sometimes the data may contain URL'S which contains no sentiments so they need to be removed.
4. Case conversion: in this step all the text should be converted into either upper case or lower case.ie there should be no difference between "HELLO" and "hello".
5. Removal of Hash tag: hash tag are generally used in social media for the specification of particular subject which contain no sentiment so need to be removed.

C. Feature extraction: here we have some feature extraction technique

1. Positive sentiment words: These words describes the positive sentiment according to SentiWordNet. For example: pretty, nice, beautiful, and outstanding.
2. Negative sentiment words: These words describes the positive sentiment according to SentiWordNet. For example: bad, ugly, awkward, and pathetic.
3. Unigram model: in the model whole sentence is divide into words [2].

For example:
Unigram example: the dress is so beautiful.
Output set: {the, dress, is, so, beautiful}
Bigram model: Here in this approach we combine two words for creating the feature vector.
Bigram example: It is not best service.
Output Set: {it is, is not, not best, best service}.
N gram Model: In this approach we combine more than two words are together to form feature vector.

D. Training and classify: after feature are extracted we apply classification algorithm. And then evaluate parameter and get the result.

III. Sentimental classification technique

There are basically three type of classification technique in sentimental analysis.

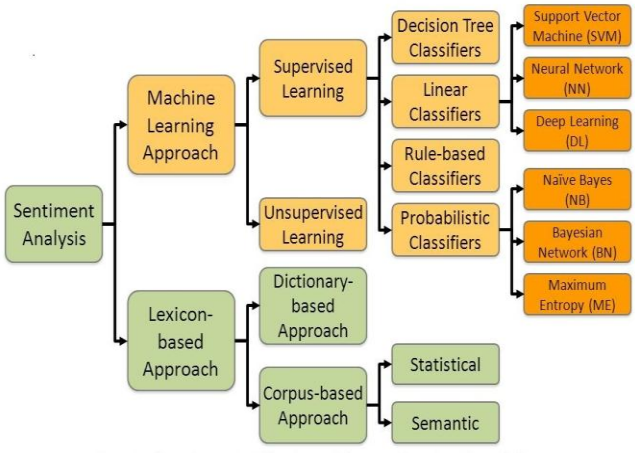

Fig 3: Sentiment ClassificationTechnique [http://slideplayer.com/slide/10538471/]

A. Machine learning

Machine learning is a technique to train a machine that how to learn and react.it allow software application to become more accurate in predicting outcome. Machine learning

classify into supervised learning and unsupervised learning. Both supervised and unsupervised is used to classify the text.

- ❖ Probabilistic Classifier: it is a generative and mixture model where each class we have is component. Probabilistic classifier is further divide into subcategories naïve bays, logistic regression, maximum entropy, Bayesian.

i. Naive bays: Naive bays is a one of the supervised classification technique which are used for classification purpose. Naive bays classifier is frequently used in sentiment categorization [3].Naive bays is a probabilistic classifier based on the Bay's theorem with independence assumption between its features. Naive bays application area are basically text classification, clustering, and target tracking [4].

Mathematical naïve bays is describe in the equation given below:

$$P(H|E1 \dots \dots .. En) = P(E1, \dots \dots . En|H)P(X)/P(E1 \dots .. En)$$

H =is the probability of event.
E =evidence.
P (E1…En |H) is the likelihood.
P (H) is the Prior.
P (E1… En) =Normalization constant
.

KL et al.[4] this paper use naïve bays classifier in which it analyze the emotions behind the review. This paper focus on the amazon product review using naïve bays. Naive bays algorithm proves to be the most efficient algorithm.

Parveen and Pandey [5] proposed naïve bays classifier. Here we apply sentimental on twitter data which help us to provide prediction on business intelligence. Here it use hadoop framework. The analysis of twitter data is categorized into three category positive, negative.

Rana and Singh [6] proposed sentimental analysis on movie review using Naive Bays, Linear SVM and synthetic words. Result shows that Linear SVM has best accuracy. The result also evaluate that people love to watch drama movie than any other movies.

Prabhat and Khullar [7] proposed in this paper twitter data is categorizes as positive, negative. The performance of the algorithm is evaluated on base of accuracy, precision and throughput. Here in this paper they successfully increase the accuracy, precision and throughput using naïve bays

BK et al. [3] this paper focus on many machine learning technique which is used in analyzing the sentimental. Here in this paper we get the 85%accuracy by using supervised technique. This paper give a result is we have small feature set than naïve bays is good technique and if the feature set is in large amount SVM will be the best choice.

ii. Bayesian network: it is the probabilistic classifier that use directed acyclic graph.in this network DAG is use to represent variables and their conditional dependence. Computational cost of this network is so high that's why this Bayesian network is not frequently used.

Trivedi and Tripathi [8] this model presents the sentimentation on Indian movies review using machine learning classifier in this model it uses Bayesian classifier .it has been used for testing the feature selection classifier. Here we use the five feature selection algorithm. The result shows that for maximum number of feature Relief-attribute is good feature selection algorithm and for less number of feature One-R is better.

iii. Maximum entropy: this is the probabilistic classifier that select the model with the highest entropy.in this classifier feature are conditionally dependent on each other. This classifier is basically used to convert feature set into VECTOR and then weight are calculated for each feature using encoded vector. Probability of each label is then computed as

$$P(fs|lbl) = dotprod(wts, encode(fs|lbl))$$

- ❖ Linear classifier: it is one of the most practical classifier.

I. Support vector machine: one of the simplest linear classification approach is the support vector machine.

TK and Shetty [9] this paper proposed various sentimental techniques like naïve bays, maximum entropy and support vector machine. And it shows that support vector machine give high accuracy compared to naïve bays and maximum entropy.it helps to make decision toward any product or any service.

II. Neural network:

Chachra et al. [10] this model capture not only the languages but also the emotion icons.it takes both the syntactic and semantic for better sentiment analysis. For this is uses Deep Convolutional Neural Network which gives 80.69%accuracy.

Chen et al. [11] they proposed their sentimental work not only on the text but also on visual data. Before this model uses either visual or text but it use both of them together. This model used Deep Convolutional Neural Network. And it improves the performance of textual and visual sentimental analysis

Alam and Rahoman [12] designed a framework which is used to analyze the text written in Bangla. For that framework it uses neural network called convolutional

Neural Network.it is used to extract feature for classification automatically. By using this classifier we improve the accuracy by 6.87% and obtains 99.87% accuracy.

B. Lexicon Classifier:

Lexicon Classifier is used to extract sentiment from words. IT is one of the unsupervised technique which is used to define polarity. Lexicon is the important part after cleaning data

Bhoir and Kolte [13] proposed a method which is used to analyze the movies review. This method performs the subjectivity analysis. To extract the opinion it use the SentiWordNet and Naïve bays classifier and method it use Lexicon Based Classifier. Extraction of subjective sentences increase the performance and efficiency both. Naive bays classifier gives more accuracy than SentiWordNet.

Vaitheeswaran and Arockiam [14] in this paper it proposed a Semi_Lexi approach that is part of lexicon classifier.Semi_Lexi approach is used to evaluate the sentimental knowledge on tweets using Lexicon based classifier. In the proposed work by using Semi_Lexi here we increase 8% of accuracy is increase by adding emotion.

Sonawane and Kulkarni [15] use the Lexicon classifier to classify review documents in positive, negative and neutral. Here it use the SentiWordNet to assign polarity to each sentence. By using it increase the accuracy by 3.50%.

- Dictionary classifier

Zhang et al. [16] this model work on not only the normal comments but also the spam comments. For this it used Spam Dictionary. Here we train the classifier to detect the spam comments. This model provides 93.6% accuracy.

- Corpus Based Approach:

This method was proposed by Manohar and Kulkarni [17] .this model is used to find the sarcastic statements by using Corpus based model as well Natural language processing to improve sentimental analysis.

C Hybrid Approach

Sometimes to improve the performance of classifier we need to hybrid the approaches.

Table 1: study of sentiment classification technique

Classifier	Year	Data Source	Feature set	Polarity	Accuracy	Reference
Naïve bays	2016	Twitter	n-gram	P/N	66.66	[4]
	2016	Twitter	SentiWordNet dictionary	P/Ne/N	yes	[5]
	2016	IMDb		P/N	yes	[6]
	2017	Amazon		P/N	yes	[7]
	2017	NA		P/N	85%	[3]
Bayesian	2016	IMDb		P/N	yes	[8]
SVM	2017	Social ware		P/N		[9]
Deep Convolution Neural network	2017	Twitter	POS	P/N	80.69	[10]
Convolution Neural network	2017	VSO	n-gram	NA	yes	[11]
Convolutional Neural network	2017	Facebook	NA	P/Ne/N	99.87	[12]
Lexicon	2015	IMDb	SentiWordNet	P/Ne/N	yes	[13]
Senti_Lexi	2016	Twitter	Uni+bigrams	P/ne/N	73.5	[14]
Lexicon	2017	Amazon	POS	P/Ne/N	78.66	[15]
Dictionary	2017	Chines social site	NA		93.6	[16]
Corpus+NLP	2017	Twitter	POS	P/Ne/N	yes	[17]

IV. Conclusion

Today's world people buy more and more product from Online site and give their reviews regarding that product. Sentiment analysis play an important role to extract the emoticons behind people review.it evaluates the polarity either this sentence comes under positive, negative or neutral.in this paper we discuss overview of sentiment analysis and its various techniques. To provide more accuracy we need to focus on cleaning of data. By using hybrid approach we provide more accuracy.

References

[1]. Kolkur, S., Dantal, G., & Mahe, R. (2015). Study of different levels for sentiment analysis. *International Journal of Current Engineering and Technology*, *5*(2), 768Ű770.

[2]. Wankhede, R., & Thakare, A. N. (2017, April). Design approach for accuracy in movies reviews using sentiment analysis. In *Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of* (Vol. 1, pp. 6-11). IEEE.

[3]. .Bhavitha, B. K., Rodrigues, A. P., & Chiplunkar, N. N. (2017, March). Comparative study of machine learning techniques in sentimental analysis. In *Inventive Communication and*

Computational Technologies (ICICCT), 2017 International Conference on (pp. 216-221). IEEE.

[4]. Kumar, K. S., Desai, J., & Majumdar, J. (2016, December). Opinion mining and sentiment analysis on online customer review. In *Computational Intelligence and Computing Research (ICCIC), 2016 IEEE International Conference on* (pp. 1-4). IEEE.

[5]. Parveen, H., & Pandey, S. (2016, July). Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. In *Applied and Theoretical Computing and Communication Technology (iCATccT), 2016 2nd International Conference on* (pp. 416-419). IEEE.

[6]. Rana, S., & Singh, A. (2016, October). Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques. In *Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on* (pp. 106-111). IEEE.

[7]. Prabhat, A., & Khullar, V. Sentiment classification on Big Data using Naïve Bayes and Logistic Regression

[8]. Tripathi, A., & Trivedi, S. K. (2016, October). Sentiment analysis of Indian movie review with various feature selection techniques. In *Advances in Computer Applications (ICACA), IEEE International Conference on* (pp. 181-185). IEEE.

[9]. Shivaprasad, T. K., & Shetty, J. (2017, March). Sentiment analysis of product reviews: A review. In *Inventive Communication and Computational Technologies (ICICCT), 2017 International Conference on* (pp. 298-301). IEEE.

[10]. Chachra, A., Mehndiratta, P., & Gupta, M. (2017, August). Sentiment analysis of text using deep convolution neural networks. In *Contemporary Computing (IC3), 2017 Tenth International Conference on* (pp. 1-6). IEEE.

[11]. Chen, X., Wang, Y., & Liu, Q. (2017). Visual and Textual Sentiment Analysis Using Deep Fusion Convolutional Neural Networks. *arXiv preprint arXiv: 1711.07798*.

[12]. Alam, M. H., Rahoman, M. M., & Azad, M. A. K. (2017, December). Sentiment analysis for Bangla sentences using convolutional neural network. In *Computer and Information Technology (ICCIT), 2017 20th International Conference of* (pp. 1-6). IEEE.

[13]. Bhoir, P., & Kolte, S. (2015, December). Sentiment analysis of movie reviews using lexicon approach. In *Computational Intelligence and Computing Research (ICCIC), 2015 IEEE International Conference on* (pp. 1-6). IEEE.

[14]. Vaitheeswaran, G., & Arockiam, L. (2017). Lexicon Based Approach to Enhance the Accuracy of Sentiment Analysis on Tweets. *International Journal Of Computer Science And Information Technology & Security (IJCSITS)*, *6*(3), 33-38.

[15]. Sonawane, S. L., & Kulkarni, P. V. (2017, October). Extracting sentiments from reviews: A lexicon-based approach. In *Intelligent Systems and Information Management (ICISIM), 2017 1st International Conference on* (pp. 38-43). IEEE

[16]. Zhang, Q., Liu, C., Zhong, S., & Lei, K. (2017, July). Spam comments detection with self-extensible dictionary and text-based features. In *Computers and Communications (ISCC), 2017 IEEE Symposium on* (pp. 1225-1230). IEEE.

[17]. Manohar, M. Y., & Kulkarni, P. (2017, June). Improvement sarcasm analysis using NLP and corpus based approach. In *Intelligent Computing and Control Systems (ICICCS), 2017 International Conference on* (pp. 618-622). IEEE.

Authors Profile

Dr. Dinesh Singh Assistant Professor at Department of Computer Science and Engineering since 2006. He has total teaching experience of 13 years . His main research area is Signal Processing and Pattern Recognition.

Ms. Savita sharma pursued Bachelors of Technology from BRCM College of Engineering, Behal, India in year 2015 and currently pursuing Masters of Technology in Department of Computer Science from Deenbandhu Chotu Ram university of Science and Technology,Murthal,India. Her main research work focuses on combine the hybrid approach using TF-idf and count vectorizer used at time of feature extraction in sentiment analysis using twitter data.

