IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 2, Ver. V (Mar-Apr. 2016), PP 64-69
www.iosrjournals.org

An Approach to Sentiment Analysis using Artificial Neural Network with Comparative Analysis of Different Techniques

Pranali Borele[1], Dilipkumar A. Borikar [2]
[1] *(M.Tech Student, Department of Computer Science and Engineering*
Shri Ramdeobaba College of Engineering and Management Nagpur, India)
[2]*(Assistant Professor, Department of Computer Science and Engineering*
Shri Ramdeobaba College of Engineering and Management Nagpur, India)

Abstract : *Sentiment Analysis is the process of identifying whether the opinion or reviews expressed in a piece of work is positive, negative or neutral. Sentiment analysis is useful in social media monitoring to automatically characterize the overall feeling or mood of consumers as replicated in social media toward a specific brand or company and determine whether they are viewed positively or negatively on the web Sentiment Analysis has been widely used in classification of review of products and movie review ratings. This paper reviews the machine learning-based approaches to sentiment analysis and brings out the salient features of techniques in place. The prominently used techniques and methods in machine learning-based sentiment analysis include - Naïve Bayes, Maximum Entropy and Support Vector Machine, K-nearest neighbour classification. Naïve Bayes has very simple representation but doesn't allow for rich hypotheses. Also the assumption of independence of attributes is too constraining. Maximum Entropy estimates the probability distribution from data, but it performs well with only dependent features. For SVM may provide the right kernel, but lacks the standardized way for dealing with multi-class problems. For improving the performance regarding correlation and dependencies between variables, an approach combining neural networks and fuzzy logic is often used.*

Keywords - *Machine Learning, Maximum Entropy, Naïve Bayes, Neural Network, Sentiment analysis, Support Vector Machine.*

I. INTRODUCTION

Sentiment mainly refers to feelings, emotions, opinion or attitude. With the rapid increase of World Wide Web, people frequently express their sentiments over internet through social media, blogs, rating and reviews. Due to this increase in the textual data, there is a need to analyze the concept of expressing sentiments and calculate the insights for exploring business. Business owners and advertising companies often employ sentiment analysis to start new business strategies and advertising campaign.

Sentiment analysis can be used in different fields for various purposes. For example in Online Commerce, sentiment analysis is extensively incorporated in e-Commerce activities. Websites allow their users to record their experience about shopping and product qualities. They provide summary for the product and different features of the product by assigning ratings or scores. Customers can easily view opinions and recommendation information on whole product as well as specific product features. Voice-of-the-Market (VOM) is about determining what customers are feeling about products or services of competitors. Voice-of-the-Customer (VOC) is concern about what individual customer is saying about products or services. It means analyzing the reviews and feedback of the customers. Brand Reputation Management (BRM) is concern about managing reputation in market. Opinions from customers or any other parties can damage or strengthen the reputation of business .
Machine learning algorithms are very helpful to classify and predict whether a particular document have positive or negative sentiment. Machine learning is categorized in two types known as supervised and unsupervised machine learning algorithms. Supervised learning algorithm uses a labelled dataset where each document of training set is labelled with appropriate sentiment, whereas, unsupervised learning include unlabelled dataset where text is not labelled with appropriate sentiments.
This paper primarily focuses on applying supervised learning techniques on a labeled dataset. Sentiment analysis is usually implemented on three levels namely sentence level, document level and aspect level. Document Level sentiment classification aims at classifying the entire document or topic as positive or negative. Sentence level sentiment classification considers the polarity of individual sentence of a document whereas aspect level sentiment classification first identifies the different aspects of a corpus and then for each document the polarity is calculated with respect to the obtained aspects for exploring business [18].

Sentiment analysis plays an important role in opinion mining. It is generally used when consumers have to make a decision or a choice regarding a product along with its reputation which is derived from the opinion of others. Sentiment analysis can reveal what other people think about a product. According to the wisdom of the crowd

sentiment analysis gives indication and recommendation for the choice of product. A single global rating could change perspective regarding that product. Another application of sentiment analysis is for companies who want to know the review of customers on their products. Sentiment analysis can also determine which features are more important for the customers. Knowing what people think provides numerous possibilities in the Human/Machine interface domain. Sentiment analysis for determining the opinion of a customer on a product is a non-trivial phase in analyzing the business activities like brand management, product planning, etc. The figure 1 shows the general process flow.

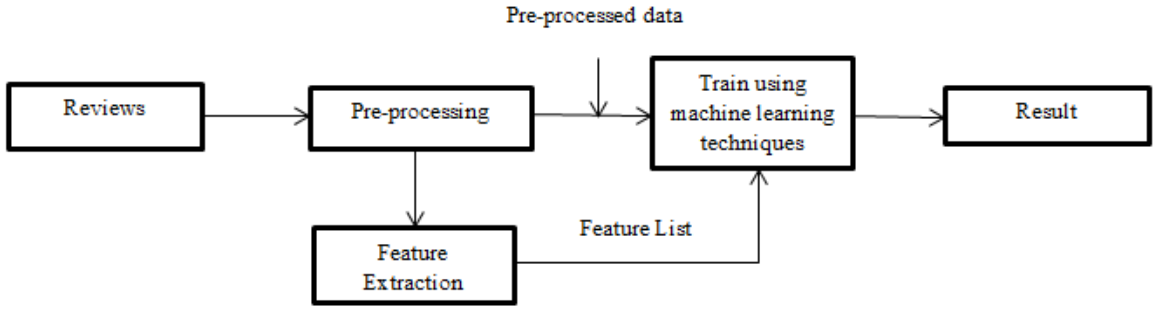

Fig.1 A general process flow using machine learning techniques

II. BACKGROUND

Pang, Lee and Vaithyanathan [1] have done sentiment classification based on categorization feature categorizing sentiments as positive and negative using three different machines learning algorithms i.e., Naïve Bayes classification, Support Vector machine, and Maximum Entropy classification. These techniques are augmented with the use of n-grams. Their experimentation reveals that the SVMs perform better as compared to Naïve Bayes technique.

The structured reviews are used for testing and training and identifying features. This is followed by scoring methods to determine whether the reviews are positive or negative. The classifiers namely NB and SVM are used to classify the sentences obtained from web search through search query using product name as search condition. When operating on individual sentences collected from web searches, performance is limited due to noise and ambiguity. But in the context of a complete web-based tool and helped by a simple method for grouping sentences into attributes, the results are qualitatively quite useful [2]. Among SVM, NB and ME classification techniques for sentiments, Naïve Bayes has been found to achieve better performance over SVM on[5].

K-nearest neighbor classification (kNN) is based on the assumption that the classification of an instance is most similar to classification of other instances that are nearby in the vector space. In comparison to the other text classification methods like Naive Bayes, KNN does not depend on prior probabilities and it is computationally efficient [6].

An approach based on artificial neural networks to divide the document into positive, negative and fuzzy tone has been proposed by Jian,Chen and Han-shi. The said approach uses recursive least squares back propagation training algorithm and in the research, sentiment analysis was performed on a large data set of tweets using Hadoop and the performance was measured in form of speed and accuracy. The results show that the technique shows very good efficiency in handling big sentiment data sets than the small datasets. [7].

Chen, Liu and Chiu have proposed a Neural Network based approach to classify sentiment in blogospheres by combining the advantages of the BPN and SO indexes. Compared with traditional techniques such as BPN and SO indexes, the proposed approach delivers more accurate results. It is found to improve classification accuracy and also reduction in training time [8].

III. COMPARATIVE REVIEW ON APPROACHES TO SENTIMENT ANALYSIS

3.1 The machine learning method

It incorporates machine learning algorithms to deduce the sentiment by training on a known dataset. This approach to sentiment classification is supervised and allows effective text classification. Machine learning classification necessitates two different sets of documents, namely for training and testing. A training set is used

by an automatic classifier to learn and differentiate attributes of documents, and a test set is used to check the performance of the automatic classifier. There are many machine learning techniques adopted to classify the reviews. Machine learning techniques like NB, ME, and SVM have achieved better performances in text categorization.

3.1.1 Naïve Bayes

It is one of the most effective, widely used and a simple approach for text classification. In this approach, first the prior probability of an entity being a class is calculated and the final probability is calculated by multiplying the prior probability with the likelihood. The method is Naïve in the sense that it assumes every word in the text to be independent. This assumption makes it easier to implement but less accurate.

$$P(^{c}/_{d}) = \frac{p(^{d}/_{c})p(c)}{p(d)}$$

3.1.2 Support Vector Machines (SVM)

It is also used for text classification based on a discriminative classifier. The approach is based on the principle of structural risk minimization. First the training data points are separated into two different classes based on a decided decision criteria or surface. The decision is based on the support vectors selected in the training set. Among the different variants of SVM, the multiclass SVM is used for sentiment analysis. The centroid classification algorithm first calculates the centroid vector for every training class. Then the similarities between a document and all the centroids are calculated and the document is assigned a class based on these similarities values.

3.1.3 The K-Nearest Neighbor (KNN)

This approach finds the *K* nearest neighbors of a text document among the training documents. The classification is done on the basis of the similarity score of the class to the neighbor document. Winnow is another commonly used approach. The system first predicts a class for a particular document and then receives feedback. In presence of false classification (i.e., error) the system updates its weight vectors accordingly. This process is repeated over a collection of sufficiently large set of training data.

3.1.4 Maximum Entropy

In Maximum Entropy Classifier, no assumptions are taken regarding the relationship between features. This classifier always tries to maximize the entropy of the system by estimating the conditional distribution of the class label.

3.1.5 Artificial Neural Network

A neural network has emerged as an important tool for classification. During past decade neural network classification has established as a promising alternative to various conventional classification methods. The neural network with appropriate network structure can handle the correlation/dependence between input variables. The advantage of neural networks lies in the following theoretical aspects. First, neural networks are data driven self-adaptive methods in that they can adjust themselves to the data without any explicit specification of functional or distributional form for the underlying model. Second, they are universal functional approximates in that neural networks can approximate any function with arbitrary accuracy. Since any classification procedure seeks a functional relationship between the group membership and the attributes of the object, accurate identification of this underlying function is doubtlessly important.

3.2 The lexicon-based approach

It mostly deals with estimation of sentiment polarity for the input text such as blog, review, comment, etc. using subjectivity and opinion orientation of the text. By using the semantic orientation of words or sentences in the review, the lexicon-based approach evaluates sentiment polarity for the review. Lexicon Based techniques generally work on an assumption that the polarity of a document or any sentence is the sum of polarities of the individual words or phrases.

3.3 The rule-based approach

These approaches involve use of semantic dictionaries. The process creates dictionary for polarity, negation words, booster words, idioms, emoticons, mixed opinions etc. The rule-based approach focuses on opinion words in a document and then classifies the text as positive or negative. The machine-learning based classifier is significantly better than rule based approach. The main advantage of the rule-based approach is that no training phase is required. The rule-based approach fails to ascertain the polarity of the text when the number of positive words and the number of negative words are equal.

3.4 Statistical model

Statistical models treats each review as a mixture of latent aspects and ratings. It is assumed that aspects and their ratings can be represented by multinomial distributions and the head terms may be grouped into aspects and sentiments providing proper ratings. A multiclass sentiment analysis problem can be addressed by combining statistics-based method with sentiment lexicon.

TABLE I. COMPARATIVE STUDY OF SENTIMENT ANALYSIS TECHNIQUE

Sr no	AUTHOR	APPROACH	DATASET	TECHNIQUES	ACCURACY
1	Pane et al.[12]	Supervised	Movie Review	SVM NB ME	82.9% 81.5% 81%
2	Abbasi et al.[13]	Supervised	Movie Review	SVM	95.5%
3	Turney [14]	Unsupervised	Movie Review bank and automobile	PMI	66%
4	Harb et al.[15]	Unsupervised	Movie Review	LEXICON	71%
5	Zhang et al. [16]	Hybrid	Twitter tweets	ML AND LEXICON	85.4%
6	Fang et al. [17]	Hybrid	Multi domain	ML AND LEXICON	66.8%

SVM have been used widely for movie reviews while NB has been applied to reviews and web discourse. In comparisons SVM gives better performance than other classifiers such as NB. [13].
The unsupervised learning algorithm is used for rating a review as thumbs up or down.
The algorithm has three steps:

(1) Extract phrases containing adjectives or adverbs,
(2) Estimate the semantic orientation of each phrase, and
(3) Classify the review based on the average semantic orientation of the phrases.

The algorithm step is important, which uses PMI-IR to calculate semantic orientation. But this technique finds that movie reviews datasets are difficult to classify thus the accuracy on movie reviews is about 66%. On the other hand, for banks and automobiles, the accuracy is 80% to 84% [14]. For the classification, technique used to calculate positive or negative orientation by computing the difference between the number of positive and negative adjectives encountered then count the number of positive adjectives, then the number of negative adjectives, and simply compute the difference. If the result is positive (resp. negative), the document will be classified in the positive (resp. negative) category. Otherwise, the document is considered to be neutral. To improve the classification result, extend the method to consider any adverbs or other words used to invert the polarities (e.g. not, neither, nor, etc.) [15].
The method first adopts a lexicon based approach to perform entity-level sentiment analysis. This method can give high precision, but low recall. To improve recall, additional tweets that are likely to be opinionated are identified automatically by exploiting the information in the result of the lexicon-based method. A classifier is then trained to assign polarities to the entities in the newly identified tweets. Instead of being labelled manually, the training examples are given by the lexicon-based approach. Experimental results show that the proposed method dramatically improves the recall and the F-score, and outperforms the state-of-the-art baselines [16].
Fang and Chen developed a method to incorporate the lexicon knowledge into machine learning algorithms such as SVM to improve sentiment learning [17].

IV. PROPOSED APPROACH

The proposed approach effectively classifies movie review in positive and negative polarities and to increase the accuracy of sentiment analysis. It incorporates neural network and fuzzy logic. Neural network is well trained for handling the correlations an inter dependencies.

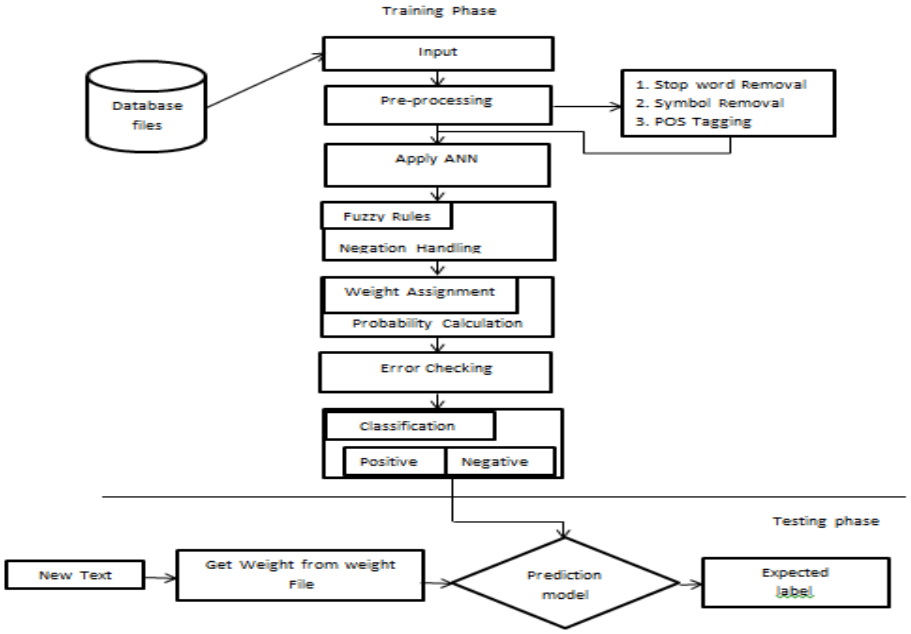

Fig.2 Proposed Approach

4.1 Dataset

The dataset contains movie reviews along with their associated binary sentiment polarity labels. It is intended to serve as a benchmark for sentiment classification. The core dataset contains 50,000 reviews split evenly into 25k train and 25k test sets. The overall distribution of labels is balanced (25k positive and 25k negative)also include an additional 50,000 unlabeled documents for unsupervised learning.

4.2 Pre-Processing

Following are steps in preprocessing.

1) Stop word removal
2) Symbol removal
3) POS tagging (Part Of Speech).
    - Stanford POS tagging is used for our study.
    - This method finds actual parts of speech using the English parser mode.
    - The POS Tagging on the input sentence and uses Verb, Adverb and Adjectives only.
    - It uses the standard Penn Treebank POS tag sets.
      For example: The movie was not quite good. After the Removal of stop word Output is [Movie, not, quite, good] after POS Tagging result is [Movie/NN, not/RB, quite/JJ, good/JJ].

4.3 Feature Extraction

When the input data is too large to be processed then transforming the input data into the set of feature is called feature extraction. If the features extracted properly from the data than it is expected that will perform the needed task. The feature extraction method, extracts the feature (adjective) from the dataset. Then this adjective is used to show the positive and negative polarity in a sentence which is useful for determining the opinion/sentiment of the individuals using unigram model .Unigram model extracts the adjective and separates it. It discards the preceding and successive word occurring with the adjective in the sentences.

4.4 Training And Classification

Supervised learning is an important technique for solving classification problems. In the proposed system we have used Artificial Neural Network for the classification of sentiments. It works in two phases i.e. Training and Testing. The training phase incorporates training number of positive and negative comments using IMDB review dataset. After that assign weights to each comment in the training phase and also apply fuzzy logic to remove the negations like not, never etc. It helps to gain the accuracy in terms of correlations and dependencies. The main purpose of training is to create the dictionary of weights of positive comments. It is need to be trained till the original positive comment will become positive. The next phase is to test the reviews. The reviews will be tested on the basis of the trained weighted dictionary. ANN preforms propagation i.e. back propagation, to train the system, by activation of neurons on hidden layer. This step begins training process of BPN by using

training data set. The back-propagation algorithm includes a forward pass and a backward pass. The purpose of the forward pass is to obtain the activation value and. The backward pass is to adjust weights and biases according to the difference between the desired and actual network outputs. These two passes will go through iteratively until the network converges. The feed-forward network training by back-propagation algorithm can be summarized as follows.

1. For each training pattern (presented in random order):
   - Apply the inputs to the network.
   - Calculate the output for every neuron from the input layer, through the hidden layer(s), to the output layer.
   - Calculate the error at the outputs.
   - Use the output error to compute error signals for pre-output layers.
   - Use the error signals to compute weight adjustments.
   - Apply the weight adjustments.
2. Periodically evaluate the network performance.
3. After the training we test whether the review is positive or negative.

V. CONCLUSION

Applying Sentiment analysis to mine the large amount of unstructured data has become an important research problem. Now business organizations and individuals are putting forward their efforts to find the best system for sentiment analysis. Some of the algorithms have been used in sentiment analysis to gives good results, but no technique can resolve all the challenges. Most of the researchers reported that Support Vector Machines (SVM) has high accuracy than other algorithms, but it also has limitations.To overcome limitation of some techniques, our study focus is on the machine learning approaches and use of artificial neural networks (ANN) in sentiment classification and analysis. Our study suggests that the ANN implementations would result in improved classification, combining the best of artificial neural network with fuzzy logic**.**

References

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002, pp.79–86.

[2] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,"in Proceedings of the 12th international conference on World Wide Web. ACM, 2003, pp. 519–528.

[3] Neethu M S and Rajasree R," Sentiment Analysis in Twitter using Machine Learning Techniques", 4th ICCCNT 2013 July 4 - 6, 2013, Tiruchengode, India

[4] S.B.Kotsianta, I.D. Zaharakis and P.E. Pintelas." Machine Learning: a review of classification and combining technique",Springer,(pp.159-190).-2006

[5] G. Vinodhini and RM. Chandrasekaran,"Sentiment Analysis and Opinion Mining: A survey", International Journal of Advanced Research in Computer Science and Software Engineering, (pp.282 – 291).-2012.

[6] Eui-Hong (Sam) Han, George Karypis and VipinKumar, "Text Categorization Using Weight Adjusted k-Nearest Neighbour Classification", 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), (pp. 53-65).-2001.

[7] ZHU Jian , XU Chen, and WANG Han-shi, " Sentiment classification using the theory of ANNs", The Journal of China Universities of Posts and Telecommunications, July 2010, 17(Suppl.): 58–62

[8] Long-Sheng Chen, Cheng-Hsiang Liu and Hui-Ju Chiu, "A neural network based approach for sentiment classification in the blogosphere", Journal of Informetrics 5 (2011) 313–322.

[9] Vikrant Hole and Mukta Takalikar, "A Survey on Sentiment Analysis And Summarization For Prediction", IJECS Volume 3 Issue 12 December, 2014 Page No.9503-9506

[10] Geetika Gautam and Divakar Yadav, "Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis", 978-1-4799-5173-4/14/$31.00 ©2014 IEEE.

[11] Neha S. Joshi and Suhasini A. Itkat, "A Survey on Feature Level Sentiment Analysis", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5422-5425

[12] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval 2(1-2), 2008, pp. 1–135.

[13] 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.

[14] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", Proceedings of the Association for Computational Linguistics (ACL), 2002, pp. 417–424.

[15] A. Harb, M. Planti, G. Dray, M. Roche, Fran, O. Trousset and P. Poncelet, "Web opinion mining: how to extract opinions from blogs?", presented at the Proceedings of the 5th international conference on Soft computing as trans-disciplinary science and technology, Cergy-Pontoise, France, 2008.

[16] 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.

[17] Ji Fang and Bi Chen, "Incorporating Lexicon Knowledge into SVM Learning to Improve Sentiment Classification", In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), pages 94–100, 2011.

[18] Abinash Tripathy, Ankit Agrawal and Santanu Kumar Rath," Classification of Sentimental Reviews Using Machine Learning Techniques", 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015).