Volume 4, Issue 10, October 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Approaches of Opinion Mining and Performance Analysis: A Survey

Ms. Vaishali Mehta
PG Student
Computer Science & Engg Dept.
SIMS, Indore, M.P., India

Prof. Ritesh K Shah
Professor
Computer Science & Engg Dept.
SIMS, Indore, M.P., India

Abstract- ***In the present scenario, Opinion mining or Sentiment analysis is used to identify user's view. It is a process of judgment and evaluation to extract subjective information from a text file or a text document. The important task of opinion mining is to extract the emotions according to the polarity and neutrality of text. Social networking websites like Twitter, Linkdin and many more provide the facility to user to share their thoughts via text messages i.e. updating post and status. In order to analyze the above scenario, Text classification is used and therefore is an important part of opinion mining. Various techniques such as Naïve Bayes theorem for classification of emotions into different classes based on hierarchical classification of text, CBTA (machine learning), Smooth filtering, Distributed clustering are used for text mining. Hybrid techniques like Mining of text using Naïve Bayes with Genetic optimization has been also introduced which have several benefits. After surveyed the different methods of opinion mining and conclude that Hybrid technique gives better results as compare to other techniques.***

Keywords: Opinion Mining, Text Classification, Naïve Bayes theorem, Machine Learning, Genetic Algorithm..

I. INTRODUCTION

Opinion mining or Sentiment analysis is a growing area in this age of information technology to evaluate the opinions, emotions, and sentiments of the text used in text classification. Emotions are extracted using classification methodology like Flat classification and Hierarchical classification techniques into different classes like sadness, fear, happy, anger etc. according to polarity and neutrality of words.

In opinion mining the major task is to evaluate the emotions of text. Naïve Bayes theorem categorize data into 3-level of hierarchy by using hierarchical classification on the basis of polarity and neutrality of words and it gives the advancement over Flat classification approach which is traditional method.

Now, Support vector machine is used for automatic classification of text message using a vector space model arising competition between the similar type of data belonging to different classes of emotion. In Competition, Support and Confidence factor is calculated between the similar words. For fully sentiment analysis of text classification several steps are required which are as follows:

(a) Extracting the text information from already define data set.
(b) Tokenized the text into specific manner.
(c) Filtering our spam and irrelevant terms associate with data.
(d) Calculate, support and confidence using vector space model.

Apply the above process on predefined data set e.g. Twitter data set. After the above process we classify the data into six classes of emotion using Hierarchical classification approach. After that machine learning gives many approaches in the area of text mining. After analyzing the machine learning process genetic optimization approach increases the system performance in terms of precision, recall and F-measure because it is fully optimized and evaluate the result by using its to important operator i.e. Mutation and Crossover. A hybrid approach Naïve Bayes with Genetic Optimization technique is used to generalize the result and comparatively give better result as compare to Naïve Bayes approach and SVM based approach.

II. METHODOLOGY

1) Naïve Bayes classifier:

First technology is Naive Bayes classifier algorithm based on the Bayes classification theory. This technique classifies the text according to particular feature of text. The value of particular feature is depends on the probability of class variables.

Naïve Bayes theorem trained the system efficiently follow supervised learning strategy according to probability reasoning. Naive Bayes classifiers have worked to solve many complex real-world conditions. The most important and effective benefit of this algorithm is requires a small amount of training data to evaluate the parameters like means and variances for text classification.

To predict the future events Bayesian reasoning is applied to decision making and inferential statistics which is deals with probability inference rule. Probability Rule according to Naïve Bayes theorem which are as follows:

The Naïve bayes Theorem:

$$P(h/D) = \frac{\{P(D/h)\ P(h)\}}{P(D)}$$

P(h) : Prior probability of hypothesis h P(D) : Prior probability of training data D

P(h/D) : Probability of h given D

P(D/h) : Probability of D given h

2) Machine Learning:

Machine learning is a process of automatic learning, the system provides the techniques how to train and test data. Machine learning is used in web search, to remove bag of words, recommended systems, audio extracting, image processing etc.

There are the three main components of Learning which are as follows:

1. Representation.
2. Evaluation.
3. Optimization.

In the present scenario, learning problem used to solve or predict the properties of unknown data using n samples of data. If the value of sample is not unique and having several attributes or features e.g. Perceptron learning model is used to separate the samples of a plane into the different classes.

Support vector machine is one of the best example of machine learning process. SVM uses vector space model (VSM) to separate the sample into different classes which is done by learning process of SVM. Three types of learning process used in SVM which is supervised, unsupervised and semi-supervised learning.

Statistical learning theory can identify rather precisely the factors that need to be taken into account to learn successfully certain simple types of algorithms, however, real-world applications usually need more complex models and algorithms. SVM can be seen as lying at the intersection of learning theory and practice. They construct models that are complex enough a large and that are simple enough to be analyzed mathematically. This is because an SVM can be seen as a linear algorithm in a high-dimensional space.

The below figure 2.1 shows that how to separate data into different categories by SVM:

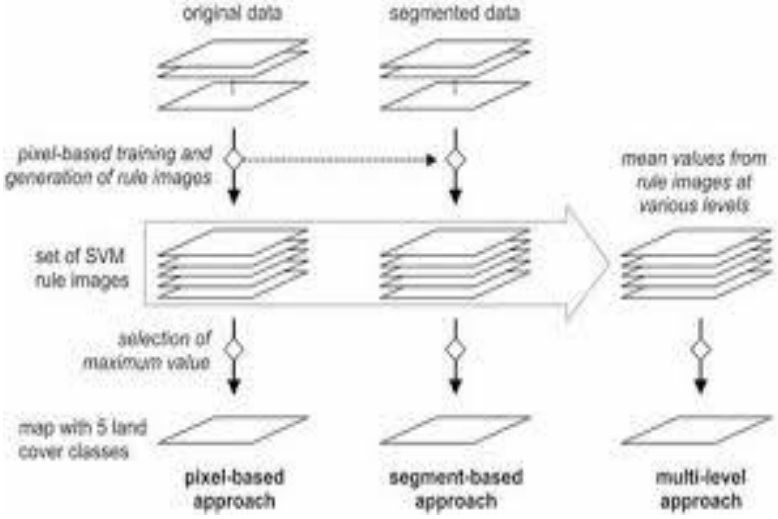

Fig 2.1: Working of SVM

1) Genetic Algorithm:

Genetic algorithm (GA) is an optimized technique derived from Darwin's principle. It is an adaptive procedure of survival of the first natural genetics. GA maintains a population of potential solution of the candidate problem termed as individuals. By manipulation of these individuals through genetic operators such as Selection, Crossover and mutation, GA gives better solutions over a number of generations.

GA are characterized by the five basic components as follows:

1. Chromosome representation for the feasible solutions to the optimization problem.
2. Initial population of the feasible solutions.
3. A fitness function that evaluates each solution.
4. Genetic operators that generate a new population from the existing population.
5. Control parameters such as population size, probability of genetic operators, number of generation etc.

The Following figure 2.2 shows that the functionality of genetic optimization technique:

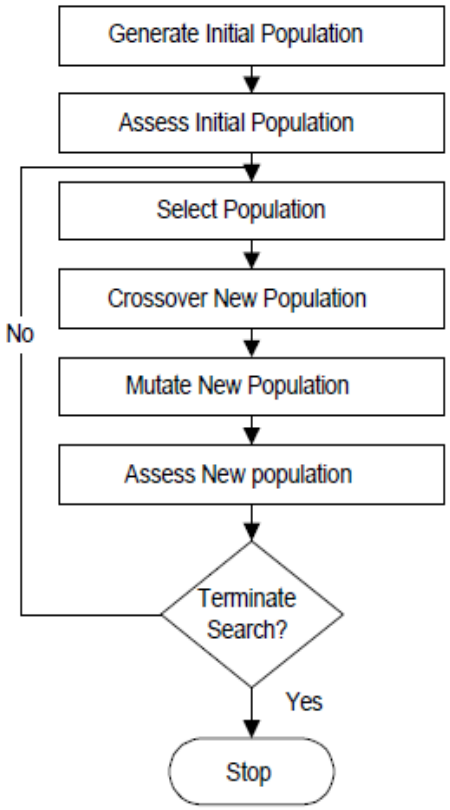

Fig 2.2: Flow chart of genetic optimization technique

III. LITERATURE SURVEY

Text classification is a major task in this age of computer science and it is an important part of sentiment analysis. Several techniques have been introduced to classify the text into different classes of emotions. Firstly, analyze that classification of text is possible by two approaches which are Flat classification and Hierarchical classification. flat classification approach uses traditional approach to mine the text into different classes and on the other hand hierarchical classification approach converts the emotion into different levels of hierarchy[1]. On comparing these two methods, Hierarchical classification approach gives better results as compare to flat classification approach because flat classification is not suitable for large amount of predefined data set. In recent internet applications, text classification methods is aimed to identify the emotions according to polarity of text, but Naïve Bayes approach is suitable for classifying text into various classes of emotion focusing on reviews given by the user on different social networking web sites[3].

Naïve Bayes theorem has several advantages. It is a suitable algorithm which mines the data into emotions. It uses Bayesian probability, Bayes classifier to filter the spam and irrelevant items from the text. It is quite useful for small data set and gives the accurate results. We analyze that the naïve Bayes theorem has many benefits but it is not compatible for large amount of data set and is not generalize for future.

In this analysis[8] the concept of machine learning when applied on a large amount of predefined data set provides automatically learning and training to the system (using SVM) and also gives comparatively good results against Naïve Bayes. Several other approaches like CBTA [2] method overcomes the drawbacks of keyword based classification of text and gives experimentally good results. This approach is used to address the problem of keyword based method. Term association is originally proposed as an information retrieval method for query expansion. It also works well when applied to text classification. It is based on the hypothesis that terms have relationships if they co-occur often in the document.

Another method called Smooth Filtering [4] for sentiment analysis was applied on Wikipedia articles that exists in training documents and further more are categorized and redirect to those articles as topic signatures. Wikipedia based semantic smoothing approach is also applied since it exploits significant amount of semantic information encoded in the relation between article titles, categories and redirects. Smooth filtering overcomes the problems of synonymy, polysemy and hyponymy.

Naïve Bayes theorem has several benefits but it does not specifying how to use unlabeled data and how to extract implicit domain knowledge. Mutual beneficial learning (MBL) algorithm [6] is an another advancement in machine learning algorithm. This algorithm is used for on-line news classification task.MBL algorithm has three major advantages over naïve bayes theorem. It uses of unlabeled data, remove noisy data and extracts implicit domain knowledge. The output of MBL consists of two components, first is a common classifier and a set of rules corresponding to local structures. A set sample first matches with the discovered rules. If a matched rule is found, the label of the rule is assigned to sample.

Another advancement in machine learning algorithm is word clustering algorithm used in hierarchical classification task [7]. This algorithm uses information theoretic approach to hard word clustering for text classification. CBTA and smooth filtering approach does not work on distributional clustering. This approach introduces the concept of clustering and achieves optimality by the use of objective function. The NB-GA approach[9] is combination of Naïve Bayes and Genetic algorithm, Naïve Bayes approach is good in filtering the text and converts text into different level of hierarchy and genetic approach gives the optimized solution using property of feature extraction. Genetic algorithm improves the system performance by using its two important operator i.e. mutation and crossover.

IV. PERFORMANCE ANALYSIS

The performance analysis of Naïve Bayes theorem, Genetic algorithm and Hybrid method(NB-GA) in terms of accuracy applied on movie review data set as shows in below fig 4.1[9]

Dataset	Classifiers	Accuracy
Movie-Review Data	Naïve Bayes (NB)	91.15 %
	Genetic Algorithm (GA)	91.25 %
	Proposed Hybrid NB-GA Method	93.80 %

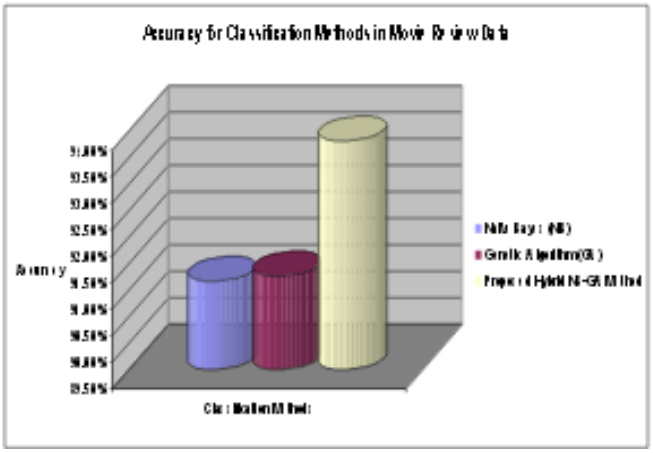

Fig. 4.1 Performance analysis of Naïve Bayes theorem, Genetic algorithm and Hybrid NB-GA method.

V. CONCLUSION

This paper presented a survey on different approaches which are used to extract emotions and also analyze the comparison between these approaches. Naïve Bayes theorem using Hierarchical classification approach is classifying the emotions into different hierarchy of class but Naïve Bayes approach is not apply on generalize data set. Text mining is an important part of opinion mining, The CBTA, Smooth Filtering, MBL and Distributed clustering approaches for text classification based on machine learning have been introduced for text mining. These methods have several benefits like find out the relation between terms, remove bag of words, increase efficiency and reduce complexity during text classification. These approaches gives several advancement in text classification task. Now in present scenario hybrid approaches are introduced for classifying the emotions into the text. The Naïve Bayes approach with the concept of Genetic optimization is used to improve the result during classification of emotions. In Genetic optimization technique based on feature extraction and feature selection and increase the performance of system in terms of accuracy by using its two important operator i.e. Mutation and Crossover.

REFERENCES

[1] Ahmed A. A. Esmin, Roberto L. de Oliveira Jr. & Stan "Hierarchical Classification Approach to Emotion Recognition in Twitter" Year 2012, IEEE conference on Machine learning.

[2] YunFei Yi, Lijun Liu, Cheng Hua Li & Wei Song "Machine Learning Algorithms With Co-occurrence based Term Association for Text Mining" Year 2012, IEEE Conference on Computational Intelligence and Communication Network.

[3] Diman Ghazi, Diana Inkpen & Stan Szpakowicz"Hierarchical Approach to Emotion Recognition and Classification in Texts" Year 2010 Springer.

[4] Dilara Totunoglu,Gurkan Telseren,Ozgun Sagturk & Murat C.Ganiz "Wikipedia Based Semantic Smoothing For Twitter Sentiment Classification" Year-2013 IEEE.

[5] Anand Mahendran, Anjali Duraiswany, Amulya Reddy & Clayton Gonsalves "Opinion Mining For Text Classification" 1 June Year-2013, IJSET Vol 2.

[6] Lei Wu, Zhiwei Li, Mingjing Li, Wei-Yang Ma & Nenghai Yu "Mutual Benificial Learning with Application to On-line News Classification" Year 2007, Sixteenth ACM international conference on iformation and knowledge management.

[7] Inderjit S. Dhillon, Subramanyam Mallela & Rahul Kumar "Enhanced Word Clustering for Hierarchical Text Classification" Year 2002, Eighth ACM international conference on Knowledge discovery and data mining.

[8] Lei Shi, Rada Mihalcea & Mingjun Tian " Cross Language Text Classification by Translation And Semi-Supervised Learning" 9-11 October Year-2010 Conference on Empirical Methods in Natural Language Processing IEEE.

[9] M.Govindarajan " Sentiment Analysis of Movie Reviews using Hybrid Method of Naïve Bayes and Genetic Algorithm" 13 December Year-2013 International Journal of Advanced Computer Research Vol-3.

[10] A. Nisha Jebaseeli, Dr. E. Kirubakaran " Genetic Optimized Neural Network Algorithm to Improve Classification Accuracy for Opinion Mining of M-Learning Reviews" 3 May-June Year 2013, International Journal of Emerging Trends Of Technology in Computer Science Vol 2.

[11] S. Chandrakala & C. Sindhu "Opinion Mining and SENTIMENT Classification: ASurvey" IJSC Vol-3.

