ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

Analysis of Machine Learning Techniques for Opinion Mining

***Hima Suresh[1] and Dr. Gladston Raj.S[2]**

1. Research Scholar**,** School of Computer Science, M.G University, Kerala.

2. Assistant Professor, Department of Computer Science, Govt. College, Nedumangadu, Kerala.

Manuscript Info

Manuscript History:

Received: 18 October 2015
Final Accepted: 22 November 2015
Published Online: December 2015

Key words:
Opinion Mining (OL), Machine Learning Techniques (MLT), Sentiment Analysis (SA), Support Vector Machine (SVM)

****Corresponding Author***

Hima Suresh

Abstract

With the onset of the exponential growth in the field of web resources and the emergence of micro-blogging websites such as Twitter in particular, which allows for the dissemination of user opinion into the social networks and the World Wide Web in general; many companies and organizations have identified these resources as a rich mine of marketing and capturing knowledge.

Opinion mining, also called sentiment analysis is a process of detecting user opinion, in reference to any particular topic, problem or product. A topic can be anything like an event, news, product, movie, location etc. Finding an efficient machine learning technique for opinion mining is a challenging problem today, due to the sheer volume and uninterrupted influx of unsorted and unprocessed information available for analysis. This paper discusses about sentiment analysis, with reference to a specific product using twitter information with a brief overview of some of the machine learning techniques and a comparison between those techniques are comprised.

INTRODUCTION

The social media has emerged as a rich and veritable source of primary data. Owing to the continuous and connected nature of periodic information flow from the source (users) to the social networks, the social media could logically be used to predict, model and analyze social behavior and trends. Social media is already being used by online retail mega stores like Amazon.com, to predict buyer behavior; Google Analytics uses information derived from social media to deliver targeted advertisement to web users. The creation of algorithms to analyze social media information, and the mining, segregation and the substituent analysis of social media information to derive and forecast behaviors, patters and trends with the aid of a supervised learning technique (Wikipedia, 2015) shall be analyzed as part of this research paper.

Sentiment Analysis (SA) is the process of detecting user's opinion about a specific topic or product. It can be considered as classification analysis. SA can be done on a document level or sentence level. In document level SA, the entire document is evaluated to determine the opinion polarity, where, the features describing the products or services should be extracted first. In sentence level SA, the document is divided into sentences where each one is evaluated separately to determine the opinion polarity. SA has been recently applied to many other areas to analyze and to predict the public behavior and feelings towards various products, services, social and political events. SA can be performed using two methods. They are Lexicon based and Machine Learning Techniques (MLT).

Lexicon based techniques works on the assumption that the contextual sentiment orientation is the sum of the sentiment orientation of each word or phrase available (Kaushik et al., 2014).

Machine Learning is the process to enhance machine performance using artificial intelligence techniques. It involves developing and learning of algorithm in order to design a model from the data's received. This helps in taking predictions and decisions based on model developed from the inputs. Real time SA of twitter information using MLT has become increasingly critical for organizations to master in order to predict market trends, analyze consumer opinions, and remain competitive. This technical report provides a survey on the existing methods to analyze sentiments of a particular product using twitter information.

The remainder of this paper is organized as follows: Section II represents related works. Section III presents proposed architecture. Section IV describes experimental analysis. Section V presents results and finally conclusion and future work is discussed in Section VI.

II. Related works

In this section some related concepts and techniques are discussed such as: Sentiment Analysis, Machine Learning Techniques etc.

A. Sentiment Analysis

Sentiment analysis, also called *opinion mining*, is the field of study that analyzes opinions, sentiments, evaluations, appraisals, attitudes, and emotions of people towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes (Bing Liu, 2012).

(Vinodhini et al., 2002), as part of the research paper entitled 'Sentiment analysis and Opinion mining: survey' undertook a study covering the techniques and sentiment analysis and the challenges that might appear in the field. It shows that neither classification model consistently outperforms the other; different types of features have distinct distributions. It is also observed that different types of features and classification algorithms are combined in an efficient way in order to overcome their individual drawbacks and benefits from each other's merits, and finally enhance the sentiment classification performance. Pravesh Kumar Singh and Mohd Shahid Husain analyzed various techniques used for opinion mining and sentiment analysis. As per the analysis, SVM produces results with better accuracy than other three methods, when N-gram feature was used. The four classification methods used in this work are applied in different areas like clustering of movie reviews and SVM technique is applied in biological reviews & analysis. They also found that Naïve Bayes is best suitable for classification such as text classification, clustering for consumer services and SVM is good for biological reading and rendition as part of the research paper entitled 'Methodological study of opinion mining and sentiment analysis techniques' (Singh et al., 2014). (Ansarul et al., 2014), analyzed the sentiments on a topic which are extracted from the Twitter and concluded a remark (positive/negative) of the defined topics. They have implemented an easier procedure to analyze sentiments on any interested field or topic.

B. Machine Learning Techniques

Machine learning techniques simulate the way humans learn from their past experiences to acquire knowledge and apply it in making future decisions. These learning techniques are widely used in artificial intelligence and document classification. The classification using machine learning can be summed up in two .One of it is learning the model using the training dataset and the other is applying the trained model to the test dataset. Sentiment analysis is a classification problem and thus any existing supervised classification method can be applied (Tripathi et al., 2015). This work uses the Naive Bayes classifier, Support Vector Machines, Max Entropy and J48 for classifying a particular product obtained from social media, twitter and compares the results obtained using these approaches to come up with an efficient MLT with accurate prediction.

(i) Naïve Bayes

It's a probabilistic and supervised classifier named after reverend Thomas Bayes. The algorithm itself is derived from Bayes theorem:

$$P(c/d) = (p(c)\sum \sum_{i=1}^{m} P\left(\frac{f}{c}\right)^{n(d)} \quad (1)$$

where P(c) and P(f/c) are obtained by calculating the relative frequency of a feature 'f', n(d) is the number of corresponding features. The total number of features are represented by 'm' and the document containing input data is represented by'd'.

One of the major works in the area of sentiment analysis using Naïve Bayes was done by (Pak et al., 2010). Using the assumption of emoticons contained in the text; large training data set was automatically collected. Ensembles of two Naïve Bayes classifiers were used for the study. First classifier was trained using the presence of unigrams while the other one was used with part of speech (POS) tagging. After combining both the classifiers they obtained an accuracy of 74%.

As part of the research conducted by (Pang *et* al., 2004), used a single Naïve Bayes classifier on a movie review corpus. Different Naïve Bayes classifiers were trained using various features such as POS, unigrams and bigrams. The resulting model achieved an accuracy of 77.3% which was considered as a good performance of the corresponding algorithm on the particular domain.

Much of the algorithm's popularity is due to its simplicity, low computational cost but it suffers multi-co linearity

(ii) J48 classifiers

J48 method is an open source java implementation of C4.5 algorithm. One of the major works in the field of sentiment analysis using J48 was carried out by (Castillo et al., 2011).The analysis mainly focused on accessing the credibility of tweets posted on twitter but they also focused on sentiment analysis. A decision tree using J48 algorithm was implemented to classify sentiment in twitter dataset. The algorithm produced an accuracy of 70%.

From the related work it can be concluded that the algorithm can be very effective for text base classification.

(iii) Support Vector Machine

The support vector machine is a non-probabilistic binary linear classifier. It works by plotting the input data in multidimensional space and separate the classes with a hyperplane. When the classes are not immediately linearly separable in the multidimensional space, the method will add a new dimension in an attempt to separate the classes. The size of the feature space increases exponentially by adding extra dimensions; this is one of the major problems with the SVM. SVM algorithm counteracts this by using dot products in the original space. Since all the calculations are performed in the original space and mapped to the feature space, this method hugely reduces processing. This increase in the size of feature space has a negative effect on the model's ability to accurately classify data, known as Hughes effect (Hughes et al., 1968). As the feature space increases, the training data becomes extremely sparse in that space and has a strong negative effect of classification. To overcome this phenomenon the training data would need to be increased exponentially with all dimensions added, that is not really practical in real world applications.

As per the studies carried out by (Pang et al., 2002), the SVM was used to extract sentiment from a movie review database. Here multiple SVM were trained using various features like POS, unigrams and bigrams. The model achieved a classification accuracy of 82.9%. The same principles as pang et al was followed in a later study were the training data was substantially increased. Increase of the training data was actually due to the researcher using the assumption that emoticons contained in text represented the overall sentiment in that text. Using this assumption large quantities of training data were automatically collected. Thus the model achieved classification accuracy of 70% on a movie review data (McCallum et al., 1998).

From the studies, it is observed that the algorithm has the ability to produce high classification accuracy. One of the main drawbacks of this algorithm is its complexity that makes it difficult to obtain a solid understanding of how exactly it works when compared to some of the simpler algorithms.

(iv) Max Entropy

The Max Entropy is a probabilistic classifier. Max Entropy is considered amongst the class of exponential models. The features set for Max Entropy classifier does not assume that the features are conditionally independent of each other. The Max Entropy is based on the Principle of Maximum Entropy that selects the model which has largest entropy from the entire models that fit the training data. Maximum Entropy is used when it is impossible to consider the conditional independence of the features. This is particularly true in Text Classification problems where the features are usually words which obviously are not independent. Time required to train the Max Entropy classifier would be more compared to Naive Bayes. This is mainly due to the optimization problems that need to be resolved in order to estimate the parameters of the model. However, after computing the parameters, the method yielded robust results and it is competitive in terms of CPU and memory consumption. As per the paper titled 'Dynamic feature sub sumption based multiclass sentiment analyzer using machine learning tool' (Jayanag et al., 2014), have proposed a new methodology to get multi-class sentiment labels regarding each feature of the product. Tagged features are extracted from the text to identify the opinion of a user with the help of Max Entropy parts of speech tagger. Results are evaluated on training and testing data and the resulting model achieved high percentage of accuracy when compared to other existing works in the particular domain.

III. Proposed Architecture

The following architecture is a depiction of methodology being followed as part of this research.

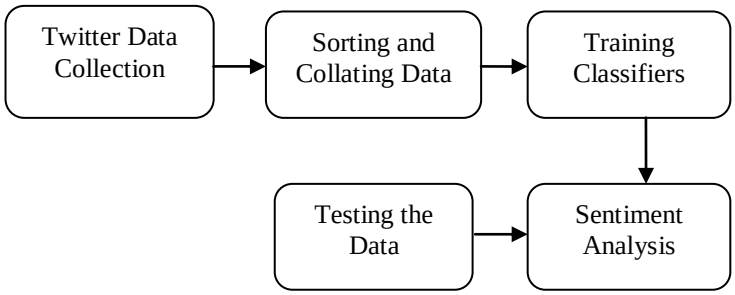

Fig1: Architecture of Sentiment Analysis

The first phase in the process involves collecting social media information from a select target group using twitter feeds. Twitter allows access to a range of streaming APIs which offer low latency access of twitter data flow. Streaming API provides access to a global stream of twitter data that could be filtered as required. Second phase of the system is sorting and collating data which involves pre-processing of data collected. In this phase, data is passed through a series of processes such as the following:

(a) Replacing emoticons: This process involves replacement of emoticons with a keyword
(b) URL extraction : This process replaces URL with an equivalent word <URL>
(c) Detection of pointers: This process abstracts username and hash tags with symbols <USER> and <HASHTAG>
(d) Identification of punctuation: This process removes irrelevant punctuations with <PUNCT>
(e) Removal of stop words: Irrelevant stop words are removed in this process
(f) Compression of words: In this process, elongated words that are used to express strong emotions are compressed. for eg: " happyyyyyyyy"

Third phase is training the classifiers. In this phase a supervised learning algorithm is trained. In order to train such a supervised classifier, a training dataset is collected. This dataset consists of training examples and the corresponding expected output for each sample. Fourth phase is testing data. Test data is the documented data that is basically used to test the corresponding software program. Final phase is the sentiment analysis task. It deals with classification of opinions expressed in twitter feeds. This is classified into multiple categories such as positive, negative or neutral based on the valence of the opinion expressed in the text units.

IV. Experimental Analysis

This section discusses about the comparison and performance analysis of different machine learning techniques on the training and test datasets with the purpose of choosing the best model for the task of twitter sentiment analysis. This part covers different processes which are used to evaluate the model. The classification algorithm with highest performance will then be applied to the twitter feeds gathered on the brand "Samsung galaxy S6" to produce the overall sentiment in the corresponding data set.

V. Results

The experiment was carried out using an open data mining tool WEKA on Pentium R processor with memory of 2GB RAM. Cross validation is the process used to evaluate the model. This is the method of estimating accuracy and validity of a statistical model. One of the standard methods used in many machine learning applications is Tenfold cross validation. The metrics used here for evaluating the performance of the classifiers are accuracy, precision and recall. The accuracy is defined as the percentage of correctly classified instances. It is defined by the following formula:

$$Accuracy = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \quad (2)$$

where *Tp* is 'true positive', *Fp* is 'false positive', *Tn* is 'true negative' and Fn is 'false negative'.

Precision is the number of instances correctly classified as its true class out of all the instances classified as that class. It is calculated for each class. Precision is defined by the following formula:

$$Precision = \frac{Tp}{Tp+Fp} \quad (3)$$

Recall represents the number of correctly classified instances of a class out of all instances of that class. The formula for calculating recall is defined below:

$$Recall = \frac{Tp}{Tp+Fn} \quad (4)$$

The following table 1 shows the results obtained after analyzing different supervised learning models:

Table1: Comparative analysis of models

Model	Accuracy (%)	Recall	Precision
Naïve Bayes	76	0.76	0.578
SVM	88	0.88	0.816
Max Entropy	89.2	0.89	0.82
J48	92	0.92	0.88

From the above result it is clear that J48 classifier outperforms the other classification models with an accuracy of 92%. The fig 2 represented below shows the accuracy of four different machine learning techniques.

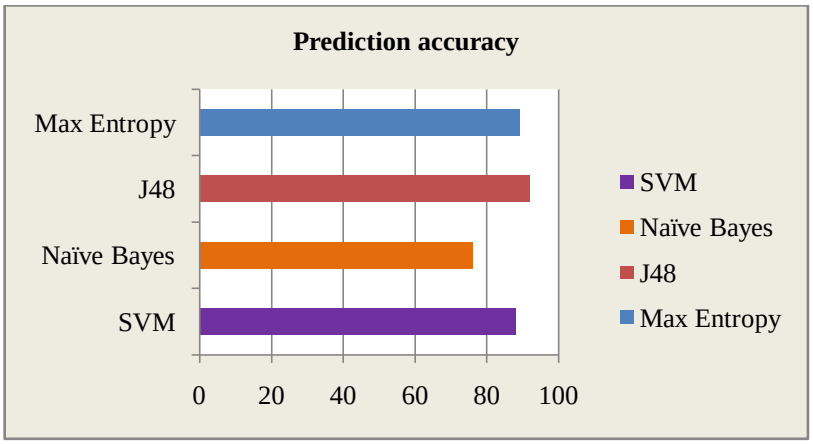

Fig2: Prediction accuracy of classification models

Following pictorial representations in fig 3 and fig 4 portraits the precision and recall of classification models for sentiment analysis.

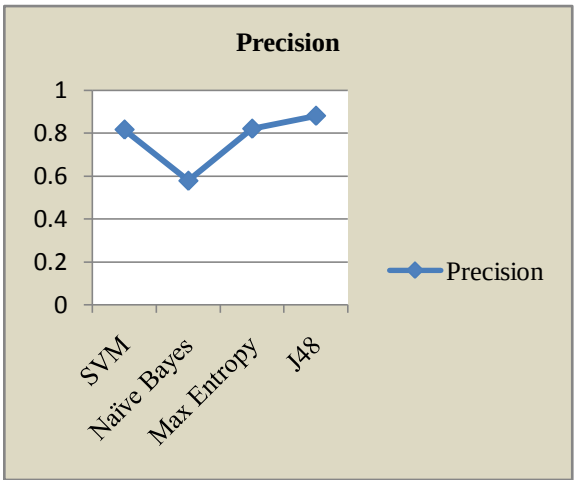

Fig3: Precision of models for SA

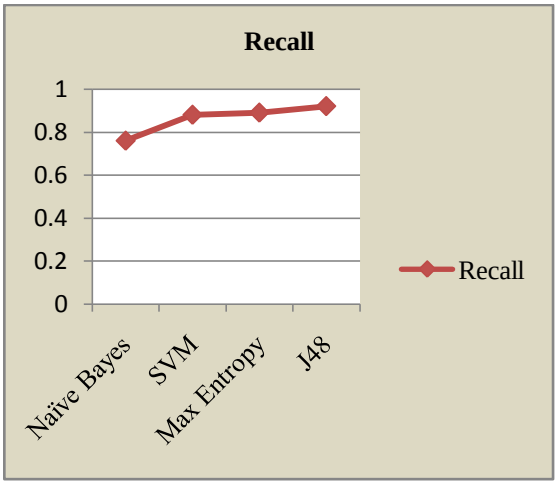

Fig4. Recall of models for SA

The experimental analysis shows that J48 predicts 80% positive sentiments and 20% negative sentiments over the other three models where SVM makes a prediction of 70% positive and 30% negative sentiments, Max Entropy gives a prediction of 75% positive and 25% negative sentiments and Naïve Bayes with 50% positive and negative sentiments. The results are depicted in the figures fig 5 and fig 6 shown below.

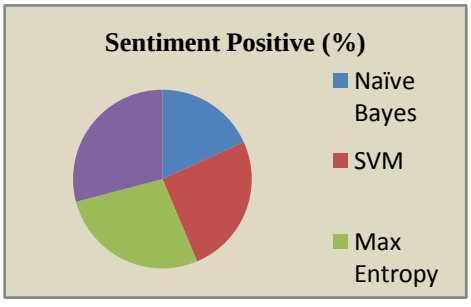

Fig5.Predicted positive sentiment

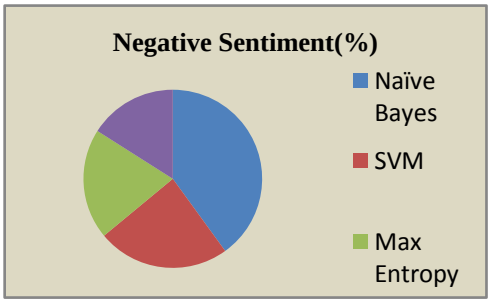

Fig 6: Predicted negative sentiment

VI. Conclusion and Future work

Many techniques have already been proposed for the classification of data. The main aim of this analysis is to explore highly efficient machine learning technique for twitter sentiment analysis. In this work we have presented a novel application of twitter sentiment analysis of a particular product and have examined the four commonly used supervised classification models. Among all those models J48 classifier turned out to be the most efficient machine learning technique compared to the other three techniques with an overall accuracy of 92% in the domain of sentiment analysis.

The J48 classifier predicted approximately 80% of positive sentiments and 20% of negative sentiments from the tweets extracted for sentiment analysis of a particular product. Recall and precision measure obtained for the model were approximately 0.92 and 0.88. In future we intend to come up with a new model which would yield results of still higher percentage of accuracy in the particular domain of twitter sentiment analysis.

VII. References

Ansarul Haque M.D, Rahman.T, (2014). Sentiment analysis by using fuzzy logic. International Journal of Computer Science Engineering and Informational Technology, 4: 33-48.

Bing Liu, (2012). Sentiment Analysis and Opinion Mining, Morgan and Clay pool publishers.

Castillo, Carlos, M.M., P. B., (2011). Information credibility on twitter.AMC, proceedings of the 20 th international conference on World Wide Web.

Hughes, G.,(1968). On the mean accuracy of statistical pattern recognizers. IEEE Transactions on Information Theory. 14(1): 53-63.

Jayanag.B, D.r Rao K.V.S, (2014). Dynamic Feature sub-sumption based Multi class Sentiment Analyzer using Machine learning Techniques, International Journal of Advanced research in computer and Communication Engineering .

Kaushik.C, Mishra.A, (2014). A Scalable, Lexicon based techniques for Sentiment Analysis, International Journal in Foundationa of Computer Science and Technology.

McCallum, Andrew, K.Nigam. (1998). A comparison of event models for Naïve Bayes text classification. AAA 198 workshop on learning for text categorization, 725: 41-48.

Pak.A, P.P, (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In proceedings of the seventh conference on International Language Resources and Evaluation.1320-275.

Pang.B, Lee.L, (2004). A Sentimental education: Sentiment nalysis using subjectivity summarization based on minimum cuts. In proceedings of the ACL. 271-278.

Pang.B, Lee.L,V.S., (2002). Thums up? Sentiment classification using Machine Learning Techniques. Association for Computer Linguistics.

Singh.P.K, H.M.S., (2014). Methodological study of Opinion Mining and Sentiment Analysis Techniques.5:11-21.

Supervised-learning. In Wikipedia. Retrived October 28, 2015, from https://en.wikipedia.org/wiki/supervised-learning.

Tripathi.G, Naganna.S, (2015). Feature selection and classification approach for Sentiment Analysis, Machine Learning and Applications: An International Journal (MLAIJ). 12: 1-16.

Vinodhini.G, Chandrasekaran.R.M, (2002). Sentiment Analysis and Opinion Mining: A Survey, International Journal of Advanced Research in Computer Science and Software Engineering, 2: 282-292.