iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VIII **Month of publication:** August 2017

DOI: http://doi.org/10.22214/ijraset.2017.8188

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VIII, August 2017- Available at www.ijraset.com

A Study on Sentiment Analysis Algorithms and its Application on Movie Reviews-A Review

Mr. Sourav De[1], Prof. Samir Kumar Bandyopadhyay[2]
[1]*Student of M. Tech. 4th Semester, Department of Computer Science, University of Calcutta*
[2] *Department of Computer Science, University of Calcutta*

Abstract: Most of the time reviews on movies carry sentiment which indicates whether review is positive or negative. The aim of this paper is to predict the sentiments of reviews using basic algorithms and compare the results.
Keywords: Sentiment Analysis, Movie Reviews, Classification Techniques

I. INTRODUCTION

Sentiment Analysis is the method of extracting subjective information from any written content. It is being widely used in product benchmarking, market intelligence and advertisement placement. Sentiment Analysis reveals the emotions, beliefs and feelings of the author on a particular topic. It uses natural language processing and machine learning techniques to effectively apply general patterns and determine the attitude expressed in the written text. Sentiment Analysis has gained popularity in recent years due to its immediate applicability in business environment, such as summarizing feedback from the product reviews, discovering collaborative recommendations, or assisting in election campaigns.

Sentiment analysis, also opinion mining is the field of computational study that analyses people's opinions expressed in written language, where focus of research is on the processing of text in order to identify opinionated information. This differs from mining and retrieval of factual information which is the target of much of the existing research in natural language processing and text analysis.

There are three main classification levels in sentiment analysis: document-level, sentence-level, and aspect-level sentiment analysis. Document-level sentiment analysis aims to classify an opinion document as expressing a positive or negative opinion or sentiment. It considers the whole document a basic information unit (talking about one topic). Sentence-level sentiment analysis aims to classify sentiment expressed in each sentence. The first step is to identify whether the sentence is subjective or objective. If the sentence is subjective, Sentence-level sentiment analysis will determine whether the sentence expresses positive or negative opinions. Classifying text at the document level or at the sentence level does not provide the necessary detail needed opinions on all aspects of the entity which is needed in many applications, to obtain these details; we need to go to the aspect level. Aspect-level sentiment analysis aims to classify the sentiment with respect to the specific aspects of entities. The first step is to identify the entities and their aspects. The opinion holders can give different opinions for different aspects of the same entity.

In this paper we discuss about sentiment analysis process and two different sentiment classification techniques i.e. Naïve Bayes classifier and support vector machine and their comparison.

II. REVIEW WORKS

Current research focuses on sentiment analysis of information gathered from social networking websites like Twitter, Facebook to conclude viewers' response to a particular social event or issue. Sentiment analysis has endless applications like forecasting market movement based on news, blogs and social media. Currently, sentiment analysis is a very lucrative approach for hefty applications like 'Smart Cities'. These applications use methods based on document level and sentence level classification which use purely supervised or unsupervised classification algorithms. These algorithms are advanced by Fuzzy Formal Concept, Genetic Algorithms or Neural Networks by making them semi-supervised. Research also focused on sentiment analysis with networking to give a degree of parallelism. It focused on online accrued utility scheduling algorithm which gave them high speed on multiple processors. But this made the system much more complex. Research was also focused on Twitter sentiment analysis for security-related information gathering using normalized lexicon based sentiment analysis [1]. While it provided a positive outcome, a universal dataset was not used. Two long and detailed surveys were presented [2-6]. They focused on the applications and challenges in sentiment analysis. They mentioned the techniques used to solve each problem in sentiment analysis.

Some researchers have given short surveys illustrating the new trends in sentiment analysis [7-10]. Others have presented a survey which discussed the main topics of sentiment analysis in details. For each topic they have illustrated its definition, problems and development and categorized the articles with the aid of tables and graphs [11].

III. SENTIMENT CLASSIFICATION TECHNIQUES

Sentiment Classification techniques can be roughly divided into machine learning approach, lexicon based approach and hybrid approach. The Machine Learning Approach applies the famous Machine Learning algorithms and uses linguistic features. The Lexicon-based Approach relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. The hybrid Approach combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods. The various approaches and the most popular algorithms of sentiment Classification are illustrated in Figure 1.

The text classification methods using machine learning approach can be roughly divided into supervised and unsupervised learning methods. The supervised methods make use of a large number of labelled training documents. The unsupervised methods are used when it is difficult to find these labelled training documents [12-13].

The lexicon- based approach depends on finding the opinion lexicon which is used to analyze the text. There are two methods in this approach. The dictionary-based approach which depends on finding opinion seed words, and then searches the dictionary of their synonyms and antonyms. The corpus based approach begins with a seed list of opinion words, and then finds other opinion words in a large corpus to help in finding opinion words with context specific orientations. This could be done by using statistical or semantic methods.

Machine learning approach relies on the famous Machine Learning algorithms to solve the sentiment analysis as a regular text classification problem that makes use of syntactic or linguistic features.

Text Classification Problem Definition: We have a set of training records D = {X_1, X_2, . . . ,X_n} where each record is labelled to a class. The classification model is related to the features in the underlying record to one of the class labels.

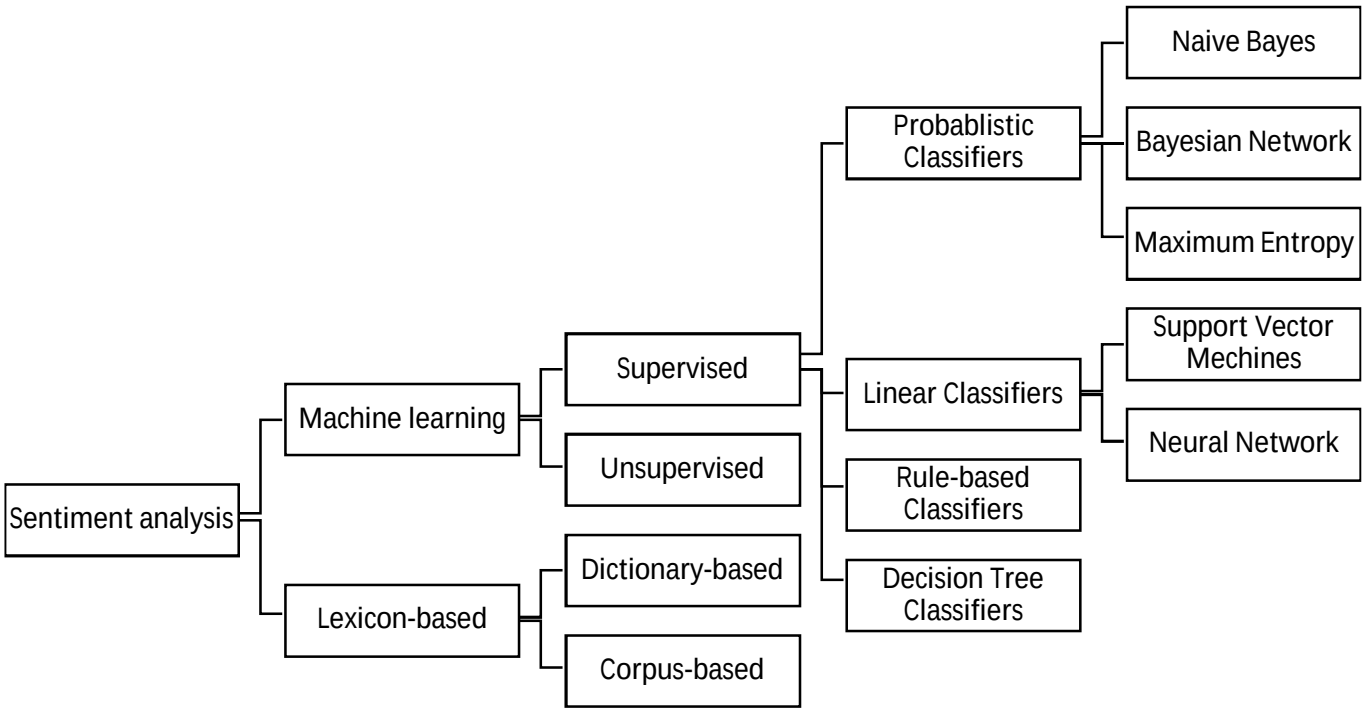


Figure 1 Sentiment Classification techniques

The lexicon- based approach depends on finding the opinion lexicon which is used to analyze the text. There are two methods in this approach. The dictionary-based approach which depends on finding opinion seed words, and then searches the dictionary of their

synonyms and antonyms. The corpus based approach begins with a seed list of opinion words, and then finds other opinion words in a large corpus to help in finding opinion words with context specific orientations. This could be done by using statistical or semantic methods.

Machine learning approach relies on the famous Machine Learning algorithms to solve the sentiment analysis as a regular text classification problem that makes use of syntactic or linguistic features.

Text Classification Problem Definition: We have a set of training records D = {X_1, X_2, . . . X_n} where each record is labelled to a class. The classification model is related to the features in the underlying record to one of the class labels.

Then for a given instance of unknown class, the model is used to predict a class label for it. The hard classification problem is when only one label is assigned to an instance. The soft classification problem is when a probabilistic value of labels is assigned to an instance.

The supervised learning methods depend on the existence of labelled training documents. There are many kinds of supervised classifiers in literature. Probabilistic classifiers use mixture models for classification. The mixture model assumes that each class is a component of the mixture. Each mixture component is a generative model that provides the probability of sampling a particular term for that component. These kinds of classifiers are also called generative classifiers.

The Naive Bayes classifier is the simplest and most commonly used classifier. Naive Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. The model works with the BOWs feature extraction which ignores the position of the word in the document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label.

$$P(lable|features) = \frac{P(label) * P(features|lable)}{P(features)}$$

P (label) is the prior probability of a label or the likelihood that a random feature set the label. P (features| label) is the prior probability that a given feature set is being classified as a label. P (features) is the prior probability that a given feature set is occurred. Given the Naive assumption which states that all features are independent, the equation could be rewritten as follows:

$$P(lable|features) = \frac{P(label) * P(f1|lable) * \ldots * P(fn|lable)}{P(features)}$$

The main assumption of the Naive Bayes classifier is the independence of the features. The other extreme assumption is to assume that all the features are fully dependent. This leads to the Bayesian Network model which is a directed acyclic graph whose nodes represent random variables, and edges represent conditional dependencies. Bayesian Network is considered a complete model for the variables and their relationships. Therefore, a complete joint probability distribution (JPD) over all the variables, is specified for a model. In Text mining, the computation complexity of Bayesian Network is very expensive; that is why, it is not frequently used.

The Maxent Classifier (known as a conditional exponential classifier) converts labelled feature sets to vectors using encoding [17]. This encoded vector is then used to calculate weights for each feature that can then be combined to determine the most likely label for a feature set. This classifier is parameterized by a set of X {weights}, which is used to combine the joint features that are generated from a feature-set by an X {encoding}. In particular, the encoding maps each C {(feature set, label)} pair to a vector. The probability of each label is then computed using the following equation:

$$P(fs|lable) = \frac{dotprod(weights, encode(fs, lable))}{sum(dotpod(weights, encode(fs, l))forlinlables)}$$

Maximum entropy classifier detect parallel sentences between any language pairs with small amounts of training data. The other tools that were developed to automatically extract parallel data from non-parallel corpora use language specific techniques or require large amounts of training data. Their results showed that Maximum entropy classifiers can produce useful results for almost any language pair. This can allow the creation of parallel corpora for many new languages [14-17].

Given $\bar{X} = \{x1, \ldots, xn\}$ is the normalized document word frequency, vector $\bar{A} = \{a1, \ldots, an\}$ is a vector of linear coefficients with the same dimensionality as the feature space, and b is a scalar; the output of the linear predictor is defined as $p = \bar{A}.\bar{X} + b$, which is the output of the linear classifier. The predictor p is a separating hyper plane between different classes.
The main principle of SVMs is to determine linear separators in the search space which can best separate the different classes. In Figure 2 there are 2 classes x, o and there are 3 hyperplanes A, B and C.

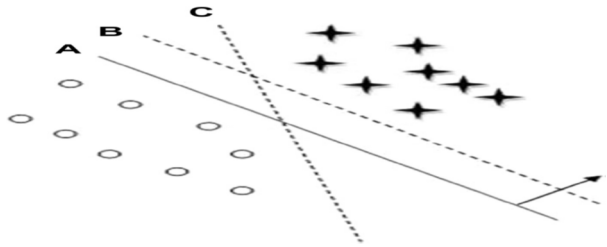


Figure 2 Using Support Vector machine on a classification problem

Hyperplane A provides the best separation between the classes, because the normal distance of any of the data points is the largest, so it represents the maximum margin of separation.
Text data are ideally suited for SVM classification because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories. SVM can construct a nonlinear decision surface in the original feature space by mapping the data instances non-linearly to an inner product space where the classes can be separated linearly with a hyperplane.
SVMs are used in many applications, among these applications are classifying reviews according to their quality. Researchers used two multiclass SVM-based approaches: One-versus-All SVM and Single-Machine Multiclass SVM to categorize reviews. They proposed a method for evaluating the quality of information in product reviews considering it as a classification problem. They also adopted an information quality (IQ) framework to find information oriented feature set. They worked on digital cameras and MP3 reviews. Their results showed that their method can accurately classify reviews in terms of their quality. It significantly outperforms state-of-the-art methods [18].
SVMs were used for sentiment polarity classifier. Unlike the binary classification problem, they argued that opinion subjectivity and expresser credibility should also be taken into consideration. They proposed a framework that provides a compact numeric summarization of opinions on micro-blogs platforms. They identified and extracted the topics mentioned in the opinions associated with the queries of users, and then classified the opinions using SVM. They worked on twitter posts for their experiment. They found out that the consideration of user credibility and opinion subjectivity is essential for aggregating micro-blog opinions. They proved that their mechanism can effectively discover market intelligence (MI) for supporting decision-makers by establishing a monitoring system to track external opinions on different aspects of a business in real time [16].
Neural Network consists of many neurons where the neuron is its basic unit. The inputs to the neurons are denoted by the vector X_i which is the word frequencies in the *i*th document. There are a set of weights A which are associated with each neuron used in order to compute a function of its inputs $f(\blacksquare)$.The linear function of the neural network is: $P_i = A \cdot \bar{X}_i$. In a binary classification problem, it is assumed that the class label of X_i is denoted by y_i and the sign of the predicted function p_i yields the class label [15-16].
Multilayer neural networks are used for non-linear boundaries. These multiple layers are used to induce multiple piecewise linear boundaries, which are used to approximate enclosed regions belonging to a particular class. The outputs of the neurons in the earlier layers feed into the neurons in the later layers. The training process is more complex because the errors need to be back-propagated over different layers.

Decision tree classifier provides a hierarchical decomposition of the training data space in which a condition on the attribute value is used to divide the data. The condition or predicate is the presence or absence of one or more words. The division of the data space is done recursively until the leaf nodes contain certain minimum numbers of records which are used for the purpose of classification [19].
Opinion words are employed in many sentiment classification tasks. Positive opinion words are used to express some desired states, while negative opinion words are used to express some undesired states. There are also opinion phrases and idioms which together

are called opinion lexicon. There are three main approaches in order to compile or collect the opinion word list. Manual approach is very time consuming and it is not used alone. It is usually combined with the other two automated approaches as a final check to avoid the mistakes that resulted from automated methods.

In dictionary based approach a small set of opinion words is collected manually with known orientations. Then, this set is grown by searching in the well-known corpora WordNet or thesaurus for their synonyms and antonyms. The newly found words are added to the seed list then the next iteration starts. The iterative process stops when no new words are found. After the process is completed, manual inspection can be carried out to remove or correct errors. The dictionary based approach has a major disadvantage which is the inability to find opinion words with domain and context specific orientations [14].

The Corpus-based approach helps to solve the problem of finding opinion words with context specific orientations. Its methods depend on syntactic patterns or patterns that occur together along with a seed list of opinion words to find other opinion words in a large corpus [15].

IV. PROPOSED METHOD

A proposed test model is discussed and it will give direct result of sentiments. We have used movie reviews dataset for training and testing of text. Sentiment Analysis can be considered a classification process as illustrated in Figure 3. In this paper we used movie review dataset for sentiment analysis (IMDB–movie-review polarity dataset V 2.0) [20].

This set is a collection of movie-review documents labelled with respect to their overall sentiment polarity (positive or negative). The set was released in June 2004 and it contains 1000 positive and 1000 negative processed reviews. The reviews were pre-processed by the dataset editors so that each review is formatted as a plain tokenized text, containing no structured information that can imply on the polarity of the text. The average review size (in words) is 746.3. Although this set is considered as a user-generated content, the language is relatively grammatically correct in most cases, probably because users are not restricted with the size of the text.

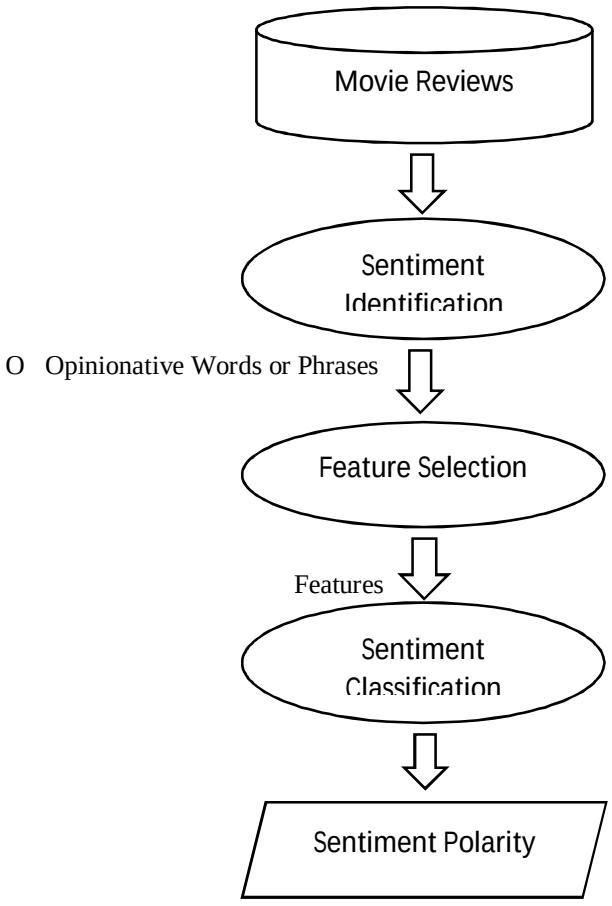


Figure 3 Process of Sentiment Analysis

Data preprocessing is a technique that involves transforming raw data into understandable format by eliminating incomplete, noisy and inconsistent data. Online informal text requires more sophisticated methods to clean noise in raw text to perform sentiment analysis. Therefore equal importance should be given to preprocessing along with classification. 'Bag of words' is required to identify opinion targets (features) from this pure textual information. This is also known as feature extraction.

It includes the following steps:

A. Lower case of all words

B. Delete all the stop word such as “the”, “a”, “to”, etc. and delete special characters

C. Part of speech (POS) tagging: Parts of speech or POS tagging is a linguistic technique used since 1960 and has recently got particular attention of Natural language processing researchers for feature extraction. POS tagging assigns a tag to each word in a text and classifies a word to a specific morphological category such as norm, verb, adjective, etc.

D. Stemming: Stemming is essential morphological processes of pre-processing module during feature extraction. The stemming process converts all the inflected words present in the text into a root form called a stem. For example, ‘automatic,’ ‘automate,’ and ‘automation’ are each converted into the stem ‘automat.’

E. Count the frequency of all words

F. Discard all words that occur less than or equal to 15 times because there is little change they relate to the sentiment of the review

Overall, there are left only 49928 unique words. The following most and least popular words were found after stemming, converting to lower case and “cleaning” the reviews from stop words, punctuations, etc. Each recorded word appears in the train set more than 15 times (Table 1). The reason to keep only words of this frequency and higher is because completely rare words are often just mistyped, non-existing or just meaningless words that would not contribute much as features to our models.

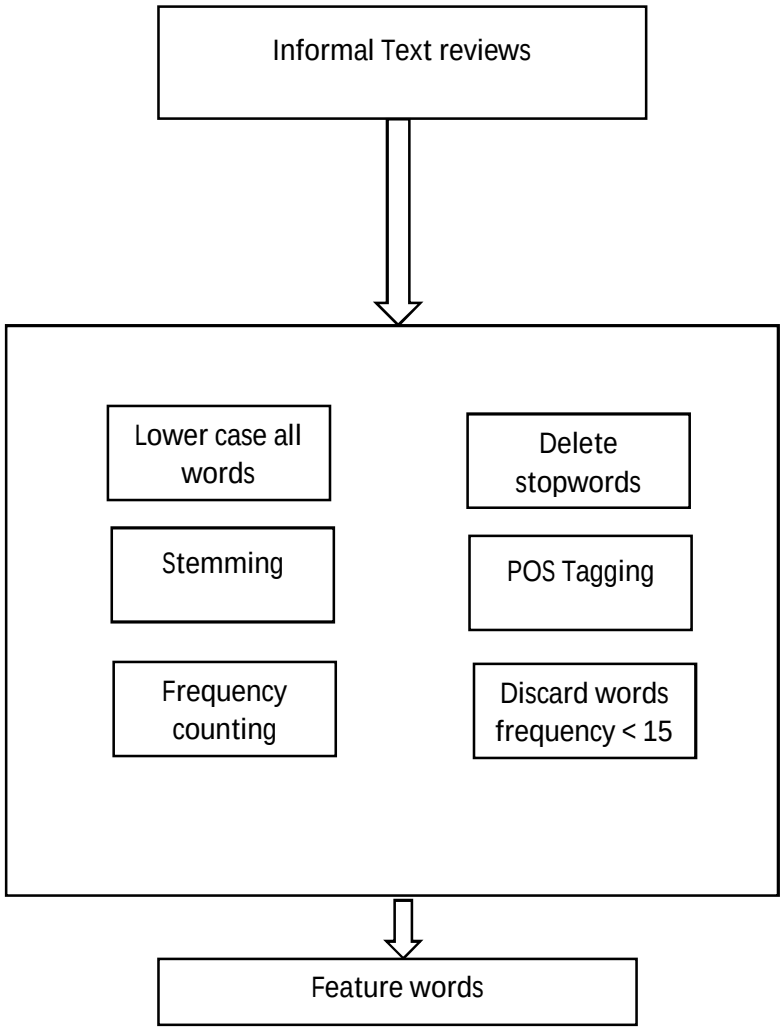
Most Popular Words	
Words	Frequency
movie	103613
film	97770
one	56253
like	45718
time	31882

Least Popular words	
Words	Frequency
Abnormal	16
Advani	16
Afar	16
Airhead	16
Alaska	16

Table 1 Most and least popular words

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VIII, August 2017- Available at www.ijraset.com

So, after preprocessing we have got the target words i.e. the feature words. Figure 4 describes the steps of data preprocessing.



After the data analysis that is done in the previous section, our goal is to make a prediction of the review sentiment based only on review text. Our output will be either positive or negative i.e. 0 or 1.

$$f(review\ text\ features) \rightarrow sentiment[0\ or\ 1]$$

We will use Naive Bayes Model and support vector machine classifier to predict the sentiment. We will test how Bayes model and svm will work if the features are selected more carefully. By choosing only those words that appear the most in positive or the most in negative, it is natural to assume that our prediction accuracy will improve.

As mentioned in the dataset portion there are 1000 positive and 1000 negative review files. For training and testing we have shuffled all files so that we can choose all types of files i.e. positive and negative files for training and testing of our classification algorithm. For training we have chosen first 1900 shuffled files and for testing we have chosen rest files.

In machine learning terms, classification is the problem of identifying to which of a set of categories a new observation belongs. This is decided on the basis of a training set of data containing observations whose category membership is known. We have used two classifiers i.e. Naive Bayes classifier and support vector machine classifier. Figure 5 describes the sentiment classification technique.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VIII, August 2017- Available at www.ijraset.com

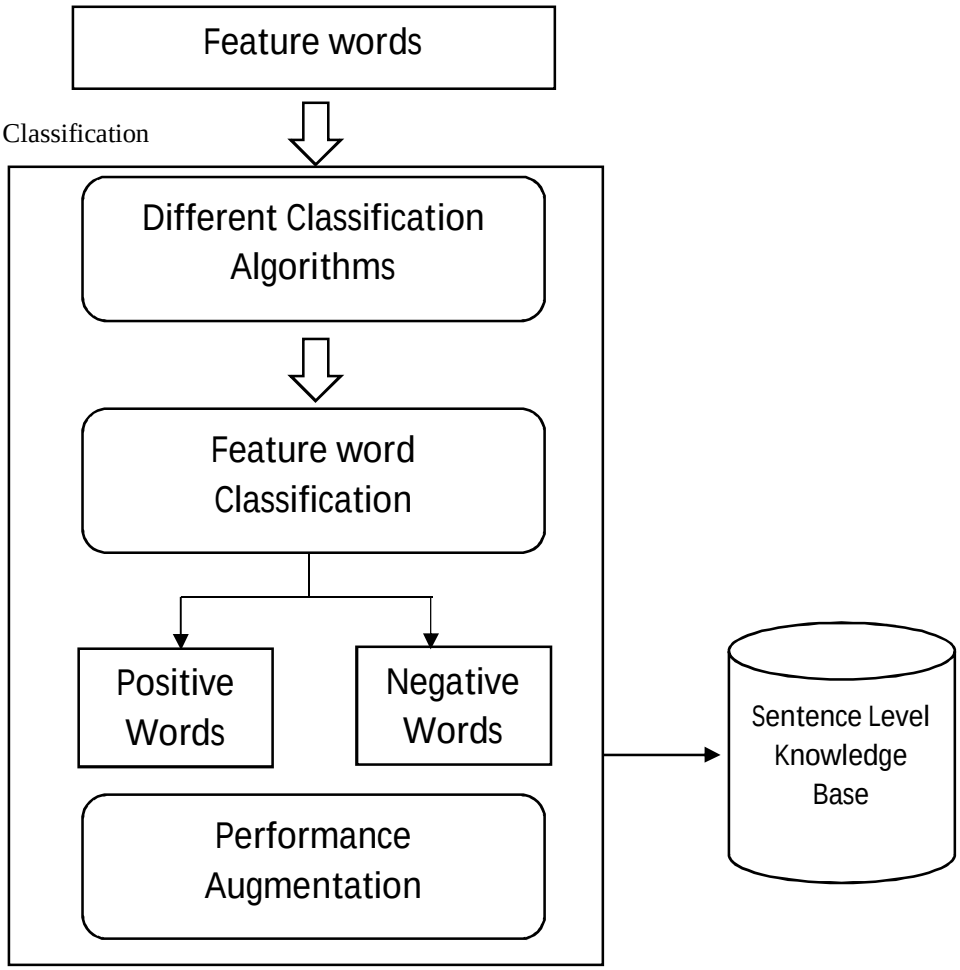


Figure 2 Classification of words

The comparison of two classifiers are shown in Table 2.

Features	Naïve Bayes	Support Vector Machine
Based on	Bayes theorem	Distance vector
Simplicity	Very simple	Moderate
Performance	Good	Good
Accuracy	Good	Better than Naïve Bayes
Memory Requirement	Low	High

Table 2 Comparison of classifiers used

A movie review comprises of a number of sentences. To calculate the polarity of the review, the polarity of each individual sentence needs to be calculated. Aggregation is finding out the polarity of each review to conclude if it falls in the positive class or negative class.

1) *Positive Words Count:* We calculated the number of positive words in the sentence and added it as a feature. This is a very important feature because if there are more positive words then the sentence tends to be a positive sentence. For example, "It's intelligent, thought provoking, emotional, and damn well entertaining" has four positive words so it is a positive sentence.
2) *Negative Words Count:* We also calculated the number of negative words present in the sentence and added it as a feature. For example, "The only problem with this film is the acting" has one negative word.

If the number of positive sentences was greater than the number of negative sentences, then we considered the overall sentiment of the review to be positive with probability number of positive sentences by total subjective sentences and vice versa. Figure 6 shows the sentiment aggregation process.

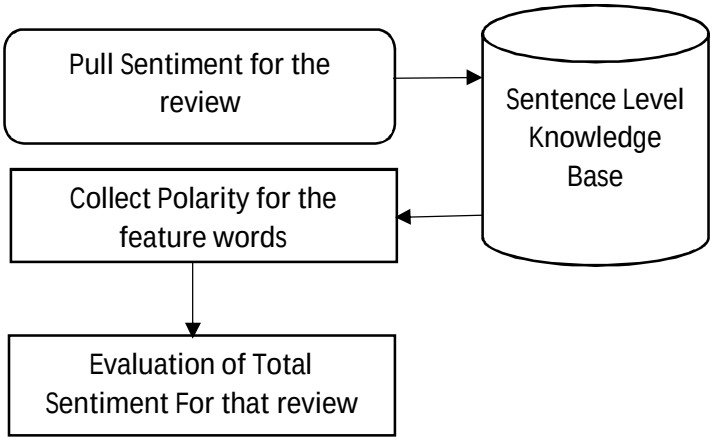


Figure 6 Sentiment Aggregation

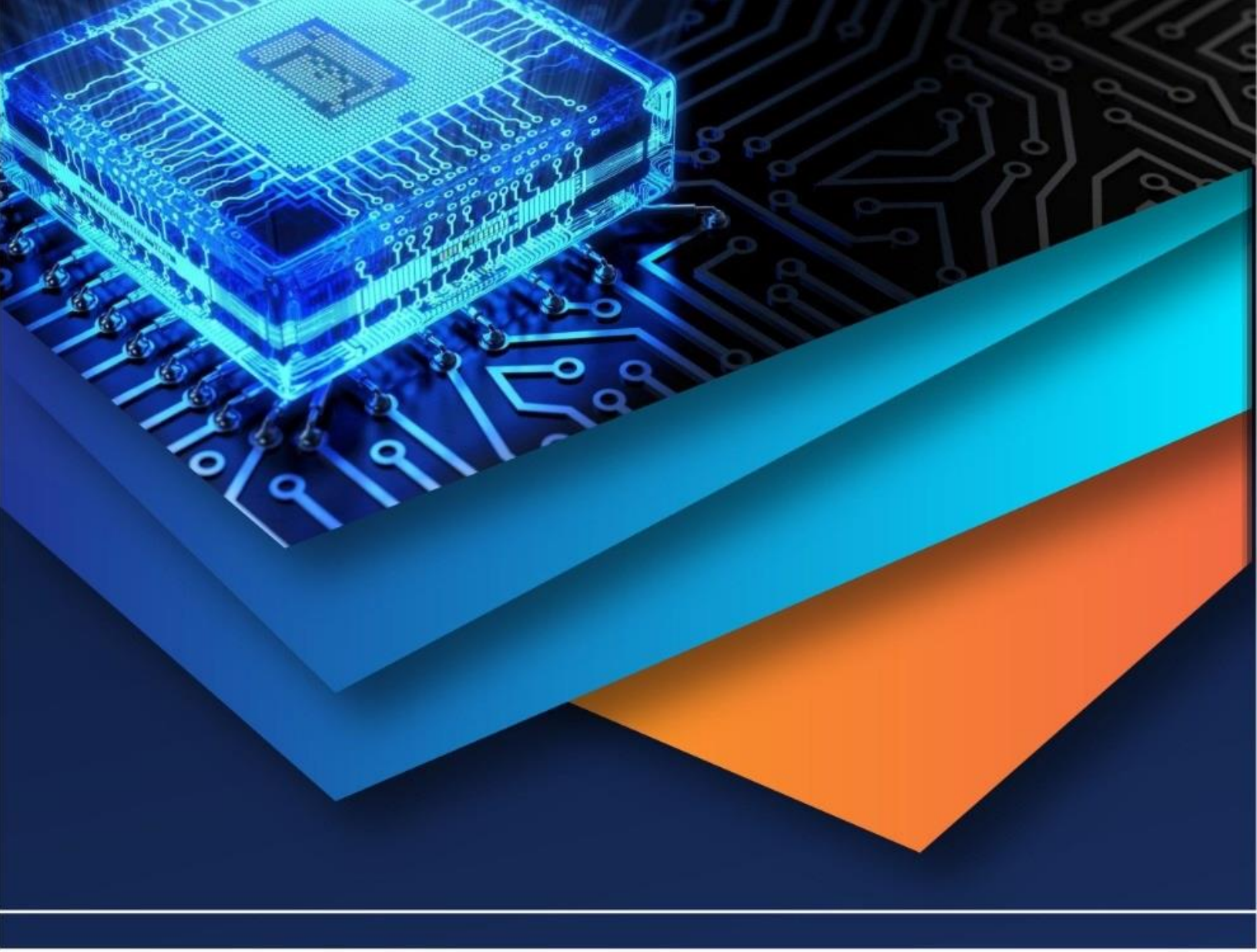
V. CONCLUSIONS

In this paper we review different sentiment classification techniques, steps of sentiment classification and use basic machine learning techniques and explore how useful they can be in predicting sentiment of movie reviews. Even with small amount data and using simple approaches to train our model we can make a quite accurate prediction of the text's sentiment.

REFERENCES

[1] Pang B, Lee L. "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval, 2008.
[2] Barbosa, Luciano and Junlan Feng. "Robust sentiment detection on twitter from biased and noisy data", In Proceedings of the International Conference on Computational Linguistics (COLING-2010), 2010.
[3] Bespalov, Dmitriy, Bing Bai, Yanjun Qi, and Ali Shokoufandeh. "Sentiment classification based on supervised latent n-gram analysis" In Proceeding of the ACM conference on Information and knowledge management (CIKM- 2011). 2011.
[4] Boiy, Erik and Marie-Francine Moens. "A machine learning approach to sentiment analysis in multilingual Web texts. Information retrieval", 2009.
[5] Dellarocas, C., X.M. Zhang, and N.F. Awad. "Exploring the value of online product reviews in forecasting sales: The case of motion pictures", Journal of Interactive Marketing, 2007.
[6] Ganesan, Kavita, ChengXiang Zhai, and Jiawei Han. Opinosis. "A graph based approach to abstractive summarization of highly redundant opinions", In Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010). 2010.
[7] Ghahramani, Zoubin and Katherine A. Heller. "Bayesian sets", Advances in Neural Information Processing Systems, 2006.
[8] He, Yulan, Chenghua Lin, and Harith Alani. "Automatically extracting polarity-bearing topics for cross-domain sentiment classification", In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011). 2011.
[9] Joachims, Thorsten. "Making large-Scale SVM Learning Practical in Advances in Kernel Methods - Support Vector Learning", MIT press, 1999.
[10] Kennedy, Alistair and Diana Inkpen. "Sentiment classification of movie reviews using contextual valence shifters", Computational Intelligence, 2006.
[11] Moghaddam, Samaneh, Mohsen Jamali, and Martin Ester. "ETF: extended tensor factorization model for personalizing prediction of review helpfulness", In Proceedings of ACM International Conference on Web Search and Data Mining (WSDM-2012). 2012.
[12] Sumathi T, Karthik S, Marikannan M. "Performance Analysis of Classification Methods for Opinion Mining", International Journal of Innovations in Engineering and Technology, 2013.
[13] Lina L. Dhande and Dr. Prof. Girish K. Patnaik, "Analyzing Sentiment of Movie Review Data using Naïve Bayes Neural Classifier", in International Journal of Emerging Trends and Technology in Computer Science August 2014.
[14] Sudipto Shankar Dasgupta, Swaminathan Natarajan,Kiran Kumar Kaipa, Sujay Kumar Bhattacherjee, Arun Viswanathan,"Sentiment Analysis of Facebook Data using Hadoop based Open Source Technologies", Proceedings of Data Science and Advanced Analytics(DSAA),2015.
[15] Akaichi, J. "Sentiment classification at the time of the Tunisian uprising," IEEE International European Conference on Network Intelligence, 2014.
[16] Mahyoub, F., Siddiqui, M., and Dahab, M. "Building an arabic sentiment lexicon using semi-supervised learning," Journal of King Saud University–Computer and Information Sciences, vol. 26, p. 417-424, 2014

[17] Al-Radaideh, Q., and Twaiq, L. “Rough set theory for arabic sentiment classification,” IEEE International Conference on Future Internet of Things and Cloud, 2014.
[18] Shrote, Khushboo R., and A. V. Deorankar. "Review based service recommendation for big data." Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2016 2nd International Conference on. IEEE, 2016.
[19] Raghuvanshi, Neha, and J. M. Patil. "A Brief Review of Sentiment Analysis." in International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016.
[20] http://www.cs.cornell.edu/people/pabo/movie-review-data/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY