

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

Machine Learning Techniques for Sentiment Analysis and Opinion Mining

Priyanka Bhadoria¹, Gajanand Sariyam²

Lecturer, Dept. of Computer Science, Kalaniketan Polytechnic College, Jabalpur, India¹

Lecturer, Dept. of Electronics and Telecommunication, Government Women's Polytechnic College, Jabalpur, India²

ABSTRACT: Analysis of sentiments or views is a leading way for evaluating text messages, and this provides optimal findings on opinions or feelings by drawing and analysing opinion-oriented text and recognising positive and negative opinions. The key power of human operations is opinion. We decide on the basis of others' feedback. It's also true for organisations, not merely for individuals. This study aims mainly to develop the model and to consider the viewpoints related to the enormous amount of data from film reviews. Film reviews with labelled opinions as good as 1 and negative as 0. This trained the system, using machine learning methods such as Random Forest, Naive Bayes and SVM, utilising the 25,000-labeled film review data. Once trained, the machine will more accurately predict the opinion on unmarked test data. This model helps uncover the customer's views on the digital web in connection with unstructured film review. The model is built entirely on NLP, text analysis, machine education and statistics.

KEYWORDS: Radom Forest, Naive Bayes, SVM (support vector machine), NLP(Natural Language Process) I.

I. INTRODUCTION

The field of research employs people's opinions, opinions, assessments, ratings, perspective and feelings for such things as products, services, corporations, people, complications, opportunities, issues and its credits as well as sentiment analysis. This is also known as Sentiment analysis. The analysis of sentiment concentrates mostly on opinions expressing feelings favourable and negative. The key values of human operations are all human activities based on opinions and opinions. We wish to know other emotions when we want to make a resolution. A company always wants to find customers or public feedback on its products and services in the actual business environment. If companies or enterprises want public or consumer views, surveys, opinion polls and market surveys shall be carried out.

Acquiring public and consumer opinions for marketing, public relations and policy propaganda organisations has been a significant business in itself for many years. Every consumer also wants to know the views of current users of a product to buy it and others like the public wants to collect views on political candidates before deciding to vote. Social networks (e.g. review, forum discussion, recordings, microblogs, Twitter, comments, and posts on social network sites) are currently growing more and more the Internet, with individuals and companies increasingly using the contents of such media for decision-making purposes. If you want to buy a consumer product nowadays, it is because there are many user reviews and analyses in public websites about the product that you would consult your family and friends for thoughts.

Nowadays, surveys, opinion polls, and market research are not necessary at all for an organisation's purpose to make the information available to the public because there is a wealth of this data. However it is still an incredible work to find and analyse the opinions on the websites and refine the information they contain due to the escalation of different websites. Typically, each site offers a huge list of opinion texts which cannot be always decoded in long blogs and blog posts. In recognition of the allowable locations for segregating and encapsulating the opinions within them, the average individual peruser will have difficulty. So technologies of sentiment analysis are necessary to overcome this programmed

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

A. Different Levels of Analysis

Document level: At this level, whole file contains group opinion and verify that whole file conveys a positive or negative sentiment. For example, given a commodity review, the system governs whether the review convey an overall positive or negative thoughts about the product. This task is commonly known as document-level sentiment classification.

Sentence level: In Sentence level, task goes to the take decision and determines whether each sentence conveyed a positive, negative, or neutral opinion. Neutral usually means no thoughts. Sentence level is intently matched to emotional classification and peculiar sentences (called objective sentences) that express real information and conveys anomalous views and opinions.

Entity and Aspect level: Aspect and entity level is not performing actually on the people likes and dislikes. This level performs like a fine tuner analysis. This aspect level also called as featured level. This always performs on the emotions (positive and negative) which are present at the each opinion.

Instead of looking at language constructs (report-log, segment text, sentences, clauses or phrases),

B. Different Types of Opinions

Regular opinion: A regular opinion is frequently referred to simple opinion type and it has two main sub-types,

Direct opinion: A direct opinion mention to thoughts expressed straight away on an entity or an entity aspect, e.g., “The image quality is awesome.”

Indirect opinion: An Unintended opinion is an opinion that is conveyed indirectly on an entity or aspect of an entity based on its outcome on some other entities. This sub-type frequently occurs in the medical territory. For example, the sentence “After getting new laptop, my operating system not working properly” describes an undesirable effect of the laptop on “operating system”, which indirectly gives a negative opinion or sentiment to the laptop. In this case, the entity is the laptop and the aspect is the effect on operating system.

Comparative opinion: A comparative opinion conveys a relation of homogeneous or variations between two or more entities and/or a preference of the opinion holder based on some apportion characteristics of the entities. For example, the sentences, “Pepsi tastes better than Maza” and “Pepsi tastes the best” express two relative opinions.

C. Data crawling

Data Crawling refers to the operation of dealing with datasets of enormous compendium, for which tools called crawlers (spiders or bots) are used to crawl into the deepest of web pages and converge data. This resemble of information extraction is predominantly used in the web and hence the name Web Crawlers. The crawling agents/tools are designed to be competent of reaching the maximum depth of a website and crawl the data constantly from each level. This ability of a crawler plays an vital role in the creation of real-time datasets. The basic web crawling process is outlined in Fig.1

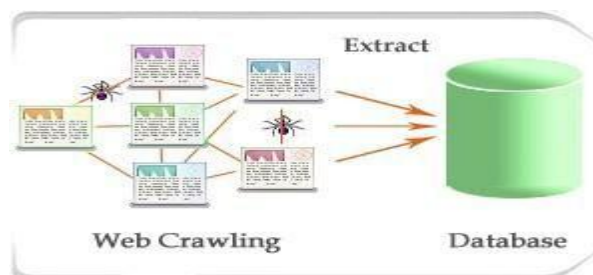


Fig: 1 A Basic web crawling process

If the requirement is to gather a particular type of data for a specific duration, then the crawler follows a scheduled

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

execution, so that the data collection is systematic and at regular intervals of time.

D. Data Pre-Processing

Once the data is crawled, the next step is to process and transform the data into a suitable structure. Data pre-processing is an information searching technique that necessitates the stated transformations to clarify the data into an understandable format. The data has to go through a sequence of steps during pre-processing like cleansing, amendment, cutting and discontinued. Data cleaning is responsible for flattening the noisy data, or resolving inconsistencies like missing values. Once the data is cleaned, dissimilar forms or presentation of data can be combined jointly which forms the integration step during pre-processing.

E. Supervised Classification

Supervised classification is nothing but choosing proper class label for a given input. In basic classification, each input is considered as a fragment from all other inputs, and the set of labels should be priority defined. Some examples of classification tasks are:

- Determining whether an email is spam or not.
- Identifying what the topic of a news article is, from a fixed list of topics like "sports," "technology," and "politics."

The basic classification task has a number of attractive variants. For example, in multi-class classification, each case taken with multiple labels; the set of labels is not priority defined that is in case of open classification and in sequence classification, list of inputs are arranged mutually.

Supervised classifier is built based on the training corpus containing the correct label for each input. The structure used by supervised classification is shown in below fig 2.

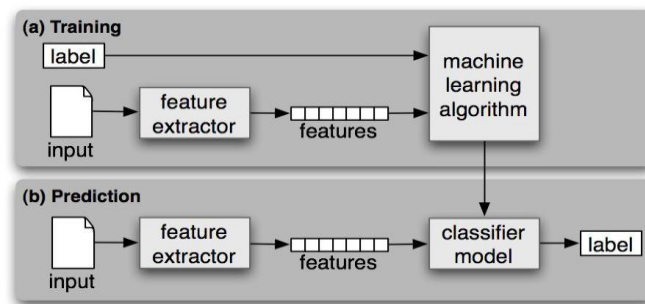


Fig: 2 Frameworks for Supervised Classification

In the process of supervised classifier, the feature extractor extracts the words those comprise either positive or negative sentiment. Then, Machine learning algorithms classify precisely and allocate the label as positive for positive sentiments and negative for negative sentiments. Once machine trained with huge compendium of data, then it will envisage precisely test data based on their knowledge. During prediction, the feature extractor can be used to convert unidentified inputs to feature sets. These inputs sets are then forwarding to the model, which generates expected labels.

II. RELATED WORK

Because of popularity of social media, field of Big Data is hot topic in today scenario. Data stored in the form of reviews, blogs, comments etc..., sentiment analysis and opinion mining became interesting field of research. A complete overview of sentiment analysis is presented by Pang and Lee in 2008. According to the survey Pang and Lee

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

mentioned various methodology of opinion mining on opinion oriented data. Only few of the researchers in opinion mining considered micro blogs, web data, e business web reviews like Amazon, flipkart. W. Wang and Y.Zhou, 2009 explains about e -business websites using popularity of internet. Yang et. al., constructed corpus using web-blogs, Corpus contains emotion icons represents users' mood. The researchers used SVM for sentence level sentiment analysis. J.Read, 2005 used emoticons such as “;-)” and “;-(“ for development of training set for sentiment classification. For this, researcher collected text contains emoticons from various newsgroups. Dataset classified as “positive” and “negative” samples. SVM and Naive Bayes had given up to 70% accuracy on this test data. In (Go et al., 2009), used Twitter to collect training data and then perform sentiment analysis on constructed corpora with “positive”, “negative” emoticons. The author obtained 81% accuracy using Naive Bayes Classifier. But he got less performance on three classes “negative”, “positive” and “neutral”.

III. DATA AND METHODOLOGY

Around 50,000 IMDB movie reviews are available in labelled data set, by using this analyse the sentiments of the reviews. The sentiment reviews are representing with binary numbers, those IMDB reviews contain < 5 ratings for those assign binary 0, and rating >= 7 assign a binary value 1. Every single movie not contains more than 30 reviews. The unique 25,000 reviews are labeled trained set using for analysis. In addition to this 50,000 reviews are unclassified labeled data set are present

A. File Description

- (i) Labelled Train Data: The labelled trained data set contains 25,000 rows containing an id, sentiment, and text for each review.
- (ii) Test Data: Here the test data is nothing but test set. Task is to predict the sentiment based on each reviews

```
<-----Sample Training Data with 25000 rows----->
id sentiment review
0 "5814_8" 1 "With all this stuff going down at the moment ...
1 "2381_9" 1 "\"The Classic War of the Worlds\" by Timothy ...
2 "7759_3" 0 "The film starts with a manager (Nicholas Bell...
3 "3630_4" 0 "It must be assumed that those who praised thi...
4 "9495_8" 1 "Superbly trashy and wondrously unpretentious ...
5 "8196_8" 1 "I dont know why people think this is such a b...
6 "7166_2" 0 "This movie could have been very good, but com...
7 "10633_1" 0 "I watched this video at a friend's house. I'm...
8 "319_1" 0 "A friend of mine bought this film for £1, and...
9 "8713_10" 1 "<br /><br />This movie is full of references....
```

Fig: 3 Samples of Labelled Train Data

```
<-----Sample of Test Reviews----->
0 "Naturally in a film who's main themes are of ...
1 "This movie is a disaster within a disaster fi...
2 "All in all, this is a movie for kids. We saw ...
3 "Afraid of the Dark left me with the impressio...
4 "A very accurate depiction of small time mob l...
5 "...as valuable as King Tut's tomb! (OK, maybe...
6 "This has to be one of the biggest misfires ev...
7 "This is one of those movies I watched, and wo...
8 "The worst movie i've seen in years (and i've ...
9 "Five medical students (Kevin Bacon, David Lab...
Name: review, dtype: object
```

Fig: 4 Samples of Test Data

B. Data Pre-Processing

Data pre-processing steps as follows:

- (i) Read tab delimited csv file LabeledTrainData.csv into python.
- (ii) Remove html tags using python package Beautiful Soup

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

- (iii) Remove numbers, punctuations, stop words.
- (iv) Convert reviews into lower case letters and words also called tokens using nltk

C. Data Fields

- (i) ID: Each row in the review contain a unique identification number
- (ii) Sentiment; for each sentiments in the review assigning binary numbers like 1 for positive reviews and 0 for negative reviews
- (iii) Review – Message of the review.

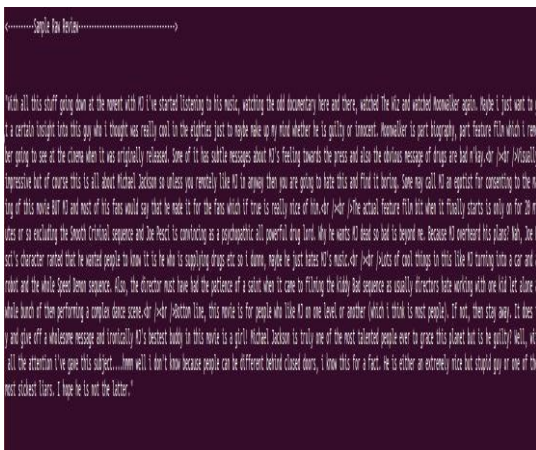


Fig.5: Sample Raw Review before processing



Fig 6: Sample processed review

D. FEATURE SELECTION FROM BAG OF WORDS

Feature selection by definition is the process of selecting a subset of attributes or features during model construction. Generally, a dataset may consist of many attributes, all of which may not have significant contributions or in other words, may not be as informative as the others. Hence, it is sensible to choose only a handful of features from the set of available features. Different approaches can be adopted for feature selection. Large set of vocabulary for IMDB data set generated by Bag of Words using feature extraction module from scikit-learn. Bag of words holds the set of features selected from Labelled Train Data.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

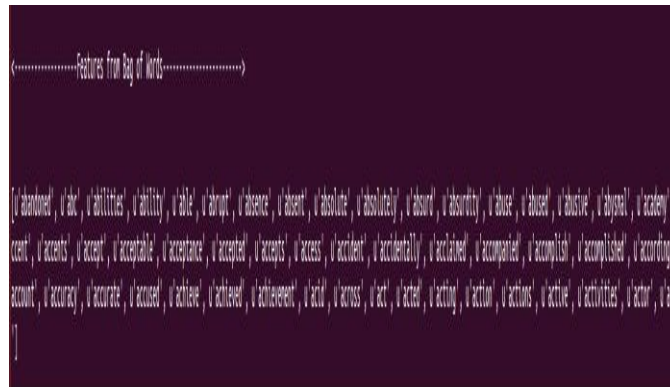


Fig.7: Features from Bag of Words

E. Classification Algorithms used for Analysis

Based on feature selection three prominent classifiers used for review classification as positive and negative .

- (i) Random Forest
- (ii) Naive Bayes
- (iii) Support Vector machine

Train the machine using above 3 classifiers with feature selected from Bag of words. For each algorithm need to evaluate accuracy on trained data, predicted sentiment on test Data, Number of positive reviews and negative reviews and time taken by each classifier.

(i) Support Vector Machines

Support Vector Machines (SVMs) are among the best off-the-shelf supervised learning algorithms available. Basically SVM is a supervised classification technique, which can be extended for regression as well, known as Support Vector Regression (SVR). The relevant concepts associated with the technique are discussed below. Consider the following figure, where 'x' and 'o' denote the two different classes of training examples, separated by a separating hyper plane known as the decision boundary. Figure 3 represents a binary classification problem mentioned above.

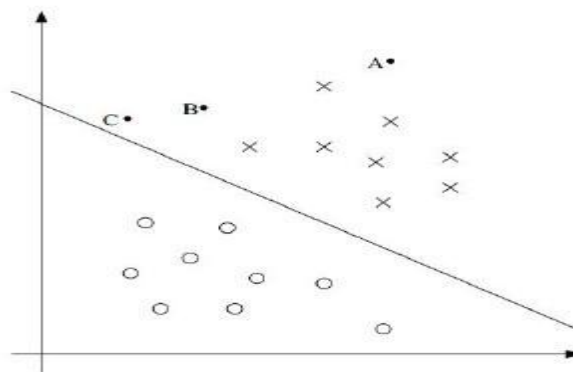


Fig.8: A simple linear classifier for binary classification

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

The following notation is adopted for the remainder of this discussion, $y \in \{-1, 1\}$ denotes the class labels, $b(x) = g(w^T x + b)$ represents the classifier, where $g(z) = 1$ if $z \geq 0$ and $g(z) = -1$ otherwise. The parameter b can be treated as the intercept term, and the vector w is associated with the input feature vector.

(ii) Naive Bayes Classifier

A Language Model is a statistical approach of modelling the words from a text by means of a probability distribution. The model assigns probabilities to the sequence of words. This approach is commonly used in a wide range of domains which include Information Retrieval, Speech Recognition, Parts-of-speech tagging and more. A Language Model is very similar in terms of approach to the Naive Bayes classifier that it behaves essentially as an extension of a Naive Bayes classifier. Generally, a Naive Bayes classifier is known for its simplicity and efficiency and this happens to be the reason why Naive Bayes is commonly used in Text Analytics, despite the presence of many other NLP approaches using k-Nearest Neighbour algorithm or Support Vector Machines. Language Models build on these traits of a Naive Bayes classifier and makes further improvements. In text classification, a document is tokenized into an unordered collection of words,

Where each word is treated as a token. A Naive Bayes classifier associates a probability measure to each token (word) that appears in the document. This estimation forms the basis for the construction of a Language Model. The model may vary from a simple bag of words model to what is known as an n-gram model. A bag of words model is a simple, unordered representation of words in a text. The frequency of occurrence of each word is used as a feature for training a classifier. However, when considers the occurrence of previous words in the text, the model is said to be n-gram. In addition, the approach employs smoothing techniques to tackle the problem of zero probability. A Naive Bayes classifier employs Laplace smoothing to assign non-zero probabilities to unseen words as well. Smoothing is defined as follows: Consider, a set of observations following multinomial distribution:

$X = (x_1, \dots, x_d)$ from N trials, and parameter vector $\theta = (\theta_1, \dots, \theta_d)$ The smoothed version of the estimator is given by $\hat{\theta}_i = \frac{x_i + \alpha}{N + \alpha d}$, ($i = 1, \dots, d$) where $\alpha > 0$ is the smoothing parameter.

In text classification, the Naive Bayes classifier is founded on the assumption of independence of words/tokens given the category/class. Given a document d and a predefined set of classes $\{\dots, c_i, \dots\}$. The Naive Bayes classifier starts off with the calculation of the posterior probability associated with the document that it belongs to a particular class c_i , using Bayes rule. $P(c_i | d) = P(d | c_i)P(c_i) / P(d)$.

(iii) Random Forest

For classification, aggregating the data and for other tasks, the best learning method is Random Forest method. In this method it forms a multilevel decision tree at training time and gives the result and that is in the mode of classification or aggregated of the individual tree. Decision based trees are very popular in machine learning tasks. The tree learning concept is very close to meet the requirements for data mining aspects. Because tree learning method or decision tree is never changing process under any circumstances of scaling and transformation features values and also this is very robust in nature to insert any irregular features produces identifiable models and those are rarely accurate in nature.

In certain examples, tree grows very deep to learn highly irregular patterns like those out of the training set data. Because those data has low inclination and very high variation. Random forest is the way of aggregating the multiple deep decision trees, on trained different parts of the training sets with the aim of reducing the variations. This can be achieved by small inclination and small loss in interoperability but in general it boosts the performance of final model.

IV. SIMULATION RESULTS

The simulation performed with three different machine learning techniques gives the best accuracy and the prediction rates on the revived data to select best opinion on product, organization etc. Here using random forest, naive

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

bayes and support vector machines. Figure 9 gives result on random forest, figure 10 gives on Naive bayes and figure 11 gives on support vector machine. A. **Random Forest Classifier** Number of Positive Reviews: 12222

Number of Negative Reviews: 12778

Mean Accuracy on Trained Data: 0.91

Time Taken by Random Forest for prediction: 513.60 seconds

Accuracy on Test Data: 0.88

Confusion Matrix: [4290 791]
[839 4080]

Precision= number of accurate positives / (sum of accurate positives and false positives) = 0.83

Recall= number of accurate positives / (the sum of accurate positives and false negatives) = 0.82

B. Naive Bayes

Number of Positive Reviews: 12020

Number of Negative Reviews: 12980

Mean Accuracy on Trained Data: 0.94

Time Taken by Naive Bayes for prediction: 47.31 seconds

Accuracy on Test Data: 0.89

Confusion Matrix: [4357 724]
[756 4163]

Precision= number of accurate positives / (sum of accurate positives and false positives) = 0.85

Recall= number of accurate positives / (the sum of accurate positives and false negatives) = 0.84

C. Support Vector Machine

Number of Positive Reviews: 12020

Number of Negative Reviews: 12980

Mean Accuracy on Trained Data: 0.96

Time Taken by SVM for prediction: 50.44 seconds

Accuracy on Test Data: 0.87

Confusion Matrix: [4495 586]
[1128 3791]

Precision= number of accurate positives / (sum of accurate positives and false positives) = 0.86

Recall= number of accurate positives / (the sum of accurate positives and false negatives) = 0.77

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

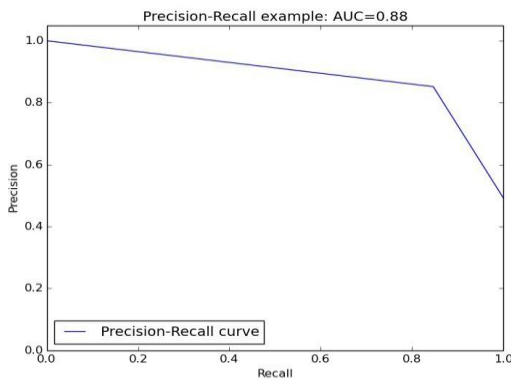


Fig.9 Precision-Recall plot of Random Forest

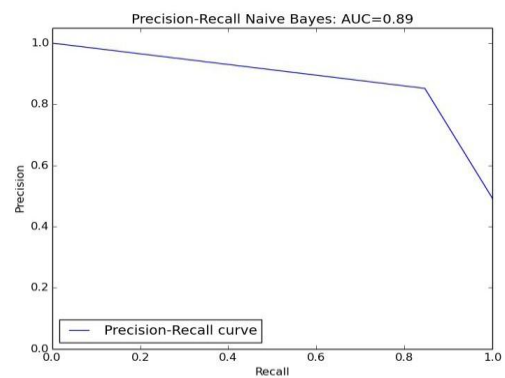


Fig 10: Precision-Recall plot of Naive Bayes

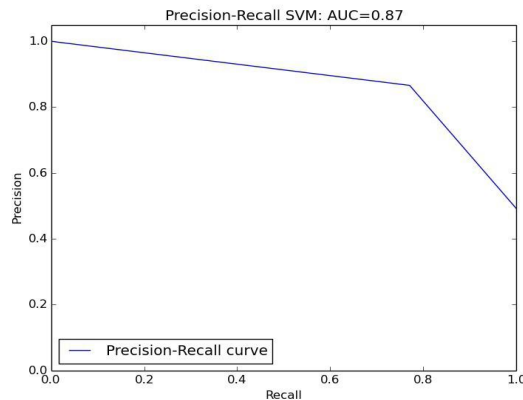


Fig 11: Precision-Recall plot of SVM

V. CONCLUSION

Sentiment analysis is a multidisciplinary field covering NLP, text mining and machine education. This technique can be used to analyse blogs, news item information, social network data and movie review data for various applications. The Sentiment Analysis here is used to analyse the positive or negative opinions related to the IMDB film review data. More than 25,000 etiquette film review data are trained using various machine learning techniques, such as Random Forest, Naiva Bayes and Vector Support machine (SVM). Once trained, the Machine will reliably predict the view of unmarked test data. This technique helps to discover customer opinions in digital format related with unstructured film review data. Model also delivers the best accuracy, accuracy and the test data retrieval. The model is built entirely on NLP, text analysis, machine education and statistics.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

REFERENCES

1. Parkhe, V.; Biswas, B. "Aspect Based Sentiment Analysis of Movie Reviews: Finding the Polarity Directing Aspects", Soft Computing and Machine Intelligence (ISCM), 2014
2. Tayal, D.K.; Yadav, S.; Gupta, K.; Rajput, B.; Kumari, K. "Polarity detection of sarcastic political tweets", Computing for Sustainable Global Development (INDIACom), 2014
3. Khurana, S.; Relan, M.; Singh, V.K. "A text analytics-based approach to compute coverage, readability and comprehensibility of eBooks", Contemporary Computing (IC3), 2013
4. Sheibani, A.A. "Opinion mining and opinion spam: A literature review focusing on product reviews", Telecommunications (IST), 2012 Sixth International Symposium
5. W. Wang and Y. Zhou, "E-business Websites Evaluation Based on Opinion Mining" in 2009 International Conference on Electronic Commerce and Business Intelligence, 2009
6. K. Khan, B. B. Baharudin, a. Khan, and F. e-Malik, "Mining opinion from text documents: A survey" 2009 3rd IEEE International Conference on Digital Ecosystems and Technologies
7. Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.
8. Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. Found. Trends Inf. Retr., 2(1-2):1-135