International Journal of Engineering & Technology, 7 (2.8) (2018) 284-290

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Comparative analysis of machine learning algorithms on social media test

R. Ragupathy*, Lakshmana Phaneendra Maguluri

Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Chidambaram, Tamil Nadu 608002

Abstract

Sentiment analysis deals with identifying and classifying opinions or sentiments expressed in main text. It mainly refers to a text classification. Social media is generating a vast amount of sentiment rich data in the form of tweets, blog posts, comments, status updates, news etc. Sentiment analysis of this user generated data is very useful in knowing the opinion of the public. Knowledge base approach and Machine learning approach are the two strategies used for analyzing sentiments from the text. In this paper, Machine learning approach has been used for the sentiment analysis of movie review dataset and is analysed by Naïve Bayes, Decision tree, KNN, and SVM classifiers. Commencing the most efficient classification technique is the moto of the paper. Efficiency of the classifier is decided based on some regular parameters that are outputs of the classification techniques.

Keywords: *Sentimental Analysis, Social Reviews, Text Pre-processing, Sentiment Score, Machine Learning Techniques, Comparative Study.*

1. Introduction

As the popularity of social media, e-commerce, forums, blogs etc. is being increased in recent times gives rise in huge storage of user data on the web in the form of opinions, reviews and comments on different products, events and services and this is continually evolving day by day. Both producers and customers are benefit holders in this context, consumers can consider opinions of different people and experience that while taking decision about any product or services and producers thereby knowing the opinion of customers on the product, will increase their product or service quality. But extracting and analyzing the useful data from this content is a major task. The unstructured nature of the data and human(natural) language being used by customers to write these content increases the complexity and this results in new research area called Opinion Mining and Sentiment Analysis.

Sentiment Analysis is the analysis of feelings (i.e. emotions, opinions and attitudes) behind the words using Natural Language Processing tools. For example: “I am very happy today, good morning to everyone”, is a general positive text. This sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the document. It is also known as opinion mining. Basically, Sentiment Analysis is the task of identifying whether the opinion expressed in a text is Positive or Negative. Natural language processing (NLP) is a field of computer science which deals with the actual text element helps to transform it into a format that the machine can use., artificial intelligence, It uses the information given by the NLP and uses a lot of statistics to determine whether something is negative or positive used for clustering. Sentiment analysis is in demand because of its efficiency. Machine learning helps for effectively computing of sentiment analysis. Thousands of text documents can be processed for sentiment in seconds, where a team of people take hours manually to complete. Because as it is efficient. This method of analysis looks beyond the number of likes, comments, shares you get on an product release, blog post, ad campaign or a video to know how people are responding to it. Was the review positive? Negative? Ideologically biased? This has several dimensions such as how does a machine define sentiment? How does machine analyze polarity (positive/negative)? Is this customer email satisfied or dissatisfied? Based on a sample of reviews, how are people responding to this ad campaign/product release/news item? How have bloggers attitudes about the movie? Related tasks such as information extracting, question-answering and summarization. Sentiment analysis can be performed in three levels they are document level, sentence level and aspect level. In the document level it takes multiple opinions which are closely related. Here, entire document is taken as a single opinion score best used in surveys. Sentence level, in this each sentence is calculated and decides whether the sentence is positive, negative or neutral. Aspect level, it calculates what aspect or feature people like or dislike and calculate the sentimental score accordingly.

In this paper, we compare the popular machine learning approaches (Naïve Bayes, decision trees, SVM (support vector machine) and KNN (K nearest neighbor) in the context of sentence-level sentiment classification. It reduces the cost of the finding insight in the customer reviews by computing the sentimental score rather than calculating and analyzing each review. It gives the customer opinion on the product weather it is negative or positive. It is used in spam detection in emails. It is used to know the mood of the person by analyzing the text.

2. Literature Review

Many researchers have been working on the problem of sentiment classification on textual reviews. Some datasets are available and have already been used by many researchers in order to compare the results. Since the focus of our study is on overall opinion (positive or negative) expressed in the review, we have oriented our literature survey towards sentence level sentiment classification.

Liu B [8] when a text is taken for the classification. The algorithm classifies the text into any of the polarity it may be positive or negative, good or bad, like or dislike, happy or angry etc., based upon the text. There are three levels of sentence polarization they are sentence level, document level and aspect level. The sentence and document level states weather the text is positive or negative whereas aspect level states which aspect is positive and which aspect is negative based on the text. [3]Opinion mining helps to detect the subjective information about emotions, opinions and feelings. According to work conducted and analysed on online paper reviews which were based on reading time and save paper costs. Sentiments were predicted on those reviews. They proposed a new technique which was called Sentiment analysis of online papers (SAOOP). This was good in understanding the reviews in different languages. This work also provides ranking and judging parameters of the whole reviews. At last this technique was compared with other existing techniques. It gave good predictions in area of topic domain dependency of sentiment analysis. [6] As there are several applications for sentiment analysis it is very important to find out the technique that must be used for classification of sentiments. A comparison was done among different classification algorithms that compare some of the machine learning algorithms like Naive Bayes, maximum entropy and balanced winnow on the tweets collected from social media. Results were like naive Bayes though generally gives less accuracy but was better than maximum entropy where maximum entropy was better than balanced winnow.

3. Machine Learning Approaches

Sentiment Analysis can be implemented with the help of two approaches. They are Machine Learning Approach and Natural Language Processing Approach. Machine learning approach needs a dataset, a classier to training set., basic idea behind this approach is that first we collect the data set which can be movie review dataset, twitter dataset etc. These data sets are available for free on internet. The processes involved in the machine learning are predictive modelling and data mining. There are three types of machine leaning algorithms they are supervised learning, Unsupervised learning, Semi-supervised learning.

3.1 Naive Bayes Classification:

It is primarily used for text classification, which involves high-dimensional training data sets. A few examples are spam filtration, sentimental analysis and classifying news articles. It is not only known for its simplicity but also for its effectiveness. Using naive Bayes algorithm you can build models fast and make quick productions. This algorithm learns the probability of an object with certain features belonging to a particular group and class. It's a probabilistic classifier. The Naïve Bayes algorithm is called "naive" because it makes the assumption that the occurrence of a certain feature is independent of the occurrence of other features. Even these features depend on each other or in the presence of each other all these properties individually contribute the probability that this fruit is an orange. In machine learning classification problem, there are multiple features and classes. The main aim of the Naive Bayes algorithm is to calculate the conditional probability of an object with a feature vector which belongs to a particular class. Posterior probability is calculated for all the feature classes and results are predicted by comparing the posterior probabilities.

$$\text{Posterior probability} = \frac{\text{Likelihood}*\text{Proposition prior probability}}{\text{Evidence prior probability}}$$

3.2 SVM (Support Vector Machine):

SVM is suited for extreme cases. It looks at extremes of datasets and draws a decision boundary also known as Hyperplane near the extreme points in the datasets. Hyperplane is the widest margin that separates the two classes. Extreme data points are known as support vectors. Support vectors are most difficult to classify. The distance between the support vectors and the hyper plane are as far as possible which is another way of saying we maximized the margin. Due to maximum margin both sides, it is less likely to result into over fitting. SVM is a frontier which best segregates the two classes (Hyper-plane/ Line). Optimized decision boundary could result in greater misclassifications on new data. SVM implies that only support vectors are important whereas training examples are ignorable. SVM can be made even more powerful by adding some additional complexities like higher dimensions, C Parameter, Multiple Classes, and Kernel Trick. SVM is also known as Large-Margin Classification. SVM is similar to LDA, logistic regression, classification tree. SVM is popular among many classification techniques because it can be used to separate linear as well as non-linear space. There are two types of SVM. Linear Support Vector Machine (LSVM) uses datasets are linearly separable, classification is based on two Dimensional datasets. Non-Linear Support Vector Machine (NLSVM) are datasets are linearly not separable. SVM is effective in high dimensional spaces like dimensions more than 2D. They are much effective in cases where number dimensions are greater than number of samples. It also works well on small datasets. But if number of features are greater than number of samples then the method is likely to give poor performance. SVM's do not provide probability estimates. They are calculated using expensive k-fold cross validation. It is not very suitable for non-linear separation of humongous datasets in its original format - It's complexity with number of records 'n' in the dataset is in n^3. SVM works as an alternative to artificial neural network. Medical Imaging i.e. classification of brain MRI, Surgical planning, Simulation and Therapy. SVM based regression models to study the air quality in urban areas. Image interpolation as well as medical classification tasks. In financial industry support vectors are used for times series production as well as financial analysis. SVM in support with neural networks in coding theory.

3.3 Decision Trees:

A decision is a type of supervised learning algorithm that is mostly used in classification problems. A tree has many similarities in real life, and it has impact on a wide area of machine learning, for both classification and regression trees also known as CART. So it is a flowchart like model, where every internal node resembles a test on an attribute, each branch resembles the outcome of a test, and each terminal node resembles a class label. The top most node of a tree is known as the root node. In decision analysis, a decision tree can be visually represent decision and as well as decision making. Similar to the name decision tree, it makes a tree like structure of decisions. So, the advantages of the classification and regression trees are easy to understand and visualize, internally perform variable reading or feature selection. They can deal with both numerical and categorical data and require relatively little effort for the data preparation and it even handles non-linearity. Decision trees generally give assurance to a user that it will work on the new datasets. It also has few disadvantages are they can be unstable because small variance in the data results in a huge difference called variance which can be lowered by bagging and boosting. The data with different number of levels, where the information gain in decision trees is biased in support of those attributes with more levels and the calculations get very tough to resolve, mainly if many values are uncertain and/or if many outcomes are attached, greedy algorithms cannot give any assurance to return the globally optimal decision tree and can be decreased by training the multiple trees. They are mainly used in predicting an email as spam or not, in prediction of tumor whether is cancerous or predicting a loan as a bad or good credit based on the factors in the problem.

3.4 K Nearest Neighbour:

It is also simply known as KNN algorithm is one of the most simply understandable algorithm in the machine learning algorithms. It is simple to understand and works incredibly well in application. It is a non-parametric learning algorithm that is it does not makes any assumption on data distribution. Its training phase is minimal and very fast. It stores all the training data in its memory as it does not have any type of generalization. In other words, all of the data that is used for training is used for the testing of the data. It is unlike the SVM in which all the non-support vectors are discarded. The KNN works fine with the data being either scalar or multidimensional. The k value is given by the user. K means the number of nearest neighbours that are to be taken into consideration. The distance between the new data and each of the data that is in the training data is calculated the distance can be hamming distance, Manhattan distance or city block distance, Euclidean distance. The k nearest neighbours are selected based upon their distances from the new data in the data set. The classes of the k nearest neighbours are taken into consideration. The class in which the majority of the data fall is taken as the class of the new data. K should always be taken as an odd number to eliminate the ties. The main advantage of the KNN algorithm is its simplicity, robustness, and its performance in classification of the data. Draw back in the KNN algorithm is the complexity the KNN has to find the each and every distance between the new data and trained data to compute the data it is easily done in simple data as the data size increase the time complexity increases. So KNN is best used for the simple data. Another drawback in KNN is utilization of memory unlike other algorithms it does not predict the data. So it stores the entire trained data in the memory and used it at the time of testing. The KNN algorithm is mostly used in the text mining to mine the data that is required or the data that is nearest to the required data. KNN is also used in agriculture for weather forecasting, and to find the soil and water parameters. It is used in the finance as the stock market forecasting i.e. finding market trends, bank customer profiling, currency exchange rate, loan management etc. Identifying the chances of occurrence of cancer based on the clinical variables. KNN search can be used as the recommender system to show the user the similar items to that of what he purchased.

4. Proposed Model

Sentiment analysis in this paper is compared among various classification techniques. A model is built to perform sentiment analysis. In this work datasets are reviews collected from social sites which are user provided reviews. As sentiments can be calculated on text which has some emotion and user given reviews are best in expressing emotions these datasets are selected for this model. According to the model next after collecting datasets preprocessing of text must be done. Preprocessing includes removal of numbers, symbols, stop words, whitespaces. Stemming is replacing words with root words. Next step to proceed is Feature selection that is extracting the features based on our output requirement. Frequency is also done in term matrix which is one of the features. Based on the features selected classification techniques are applied. Evaluating results among Naïve Bayes, SVM, Decision tree and KNN classifications. Sentiments are calculated on the classified texts. This model results in sentiment classification and comparative results of the above-mentioned classifiers.

In our work movie review are selected as datasets. In R programming, there are many packages provide functions for classification of text and also to provide sentiment analysis. Some of the packages used are tm, caret, snowballC, wordcloud, dplyr, syuzhet, e1071. Text mining applications use tm package in R. For classification and regression training caret package is used. Stemming words library is available in snowballC package. Wordcloud package have functions that are used to build pretty wordclouds. Data manipulation can be done with dplyr package. Sentiment dictionaries, sentiment derived plots and sentiments can be extracted from syuzhet package. Datasets are imported from local library. Dataset collected will be allocated to a corpus variable which can be used for preprocessing in R.

Any text must be preprocessed before further experimenting the text. Some text preprocessing methods are used to for removing unused text form input datasets. As already discussed there are many methods available in text preprocessing, only some of them which are necessary for this work are used in this model. Special characters like @, #, / ... have no value adding to the sentiments of the review. Term Document Matrix is a datatype that is frequently used in R programming. This is mainly used to obtain word frequencies in the input. Corpus variable can be type casted to a Document term matrix. For better visualization, we are also producing word cloud for the dataset selected. Not all the words are displayed in the cloud but words with more frequencies are listed out and displayed as a word cloud. Now that words are arranged in the document matrix, it's time for classification.

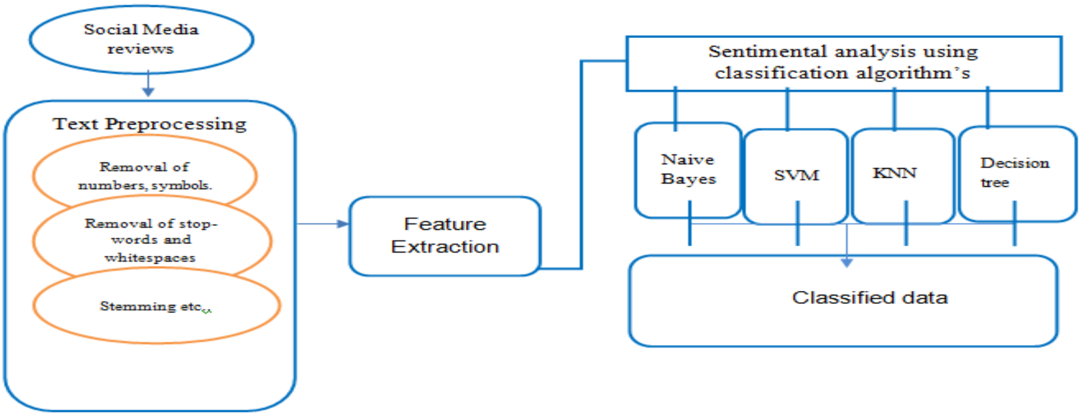
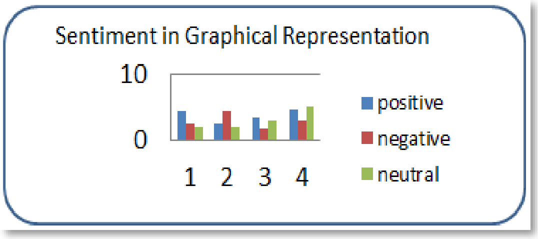

Social Media reviews
Text Preprocessing
Removal of numbers, symbols.
Removal of stop-words and whitespaces
Stemming etc,,
Feature Extraction
Sentimental analysis using classification algorithm's
Naive Bayes
SVM
KNN
Decision tree
Classified data

Fig. 1: Outline of the comparative model for sentimental analysis



For classification, initially we need to take datasets into two sets they are training set and test set. Training sets are considered to be the actual data which is used for further classification. Whereas test set is used for prediction using any of the classification techniques. As already chosen in the flow model Naïve Bayes, SVM, KNN, Decision tree classifiers are used and compared with each other for better results. For Naïve Bayes classifier, a convert function is defined which is practiced in addition during classification of text. This function also labels text as per their belonging classes. A classifier variable is trained as per the convert function and the training data. With this classifier, more predictions can be done. Now that as we have to predict the classes of test set. Predict functions are also available to do that process. This results in the count whether all the predictions are classified properly with their respective classes or not. That gives a confusion matrix as output which has true positive, true negative, false positive, false negative values. Moving to SVM classifier, package 'e1071' in R programming have required functions that are used to train the classification algorithms. Training set of data is used to train svm classifier along with the factors of classification. Classifier along with the test set of data now results in predictions and confusion matrix. Implementing KNN is so similar to implementing of SVM. Train method does work of training the classifier when method attribute is included to the function. Predictions are made again based on the KNN trained classifier and the test set. These predictions are used to display confusion matrix of KNN predictions compared to the actual values. Decision tree classifier will also be trained additional to the training set. A training control function is defined that cross folds the training set repeatedly for limited number of times while training the Decision tree classifier. Trained classifier can now be represented as Decision tree. Confusion matrix can be formed using trained classifier of Decision tree and the test set.

Overall details of confusion matrix provide details like accuracy, sensitivity, specificity, prevalence and balanced accuracy. Sentiments are extracted from classified data as outputs from functions of syuzhet package. Binding of corpus variable with sentiments extracted outcomes text results of sentiments. Sentiment tools classify the text with their respective sentiment value. A bar chart is drawn that indicates the sentiments analyzed from the classified texts. On basis of results from confusion matrix attributes are compared for different classification techniques.

5. Results

Results obtained from the analysis of classification techniques are provided in this section. Sentiment classification, Overall attributes of classification, word cloud are the results obtained from the experiment. Wordcloud gives the words with highest frequency in the document in arranged in a format provided by wordcloud package. This display varies based on the attributes given to the function. Wordcloud helps us find the most used words.

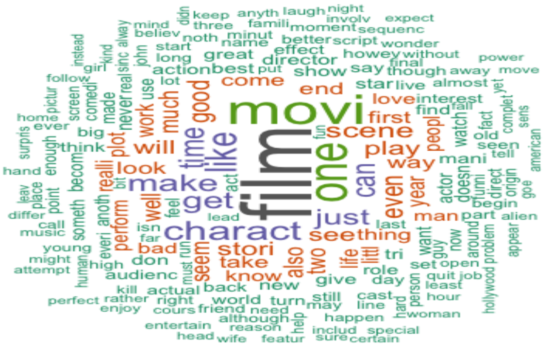


Fig. 2(a): Word cloud of movie review dataset

Table 1: Confusion matrix of Naïve Bayes

	Actual	
Negative	254	225
Positive	11	10

Table 2: Confusion matric of SVM

	Actual	
Prediction	Negative	Positive
Negative	0	0
Positive	150	401

Table 3: Confusion matric of KNN

	Actual	
Prediction	Negative	Positive
Negative	0	0
Positive	50	51

Table 4: Confusion matric of Decision tree classifier

	Reference			
Prediction	More Negative	More Positive	Negative	Positive
More Negative	0	0	0	0
More Positive	10	25	6	7
Negative	0	0	0	0
Positive	0	0	0	0

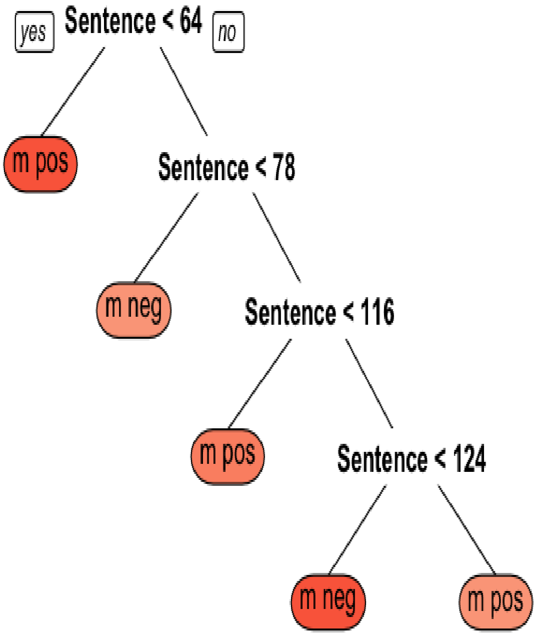
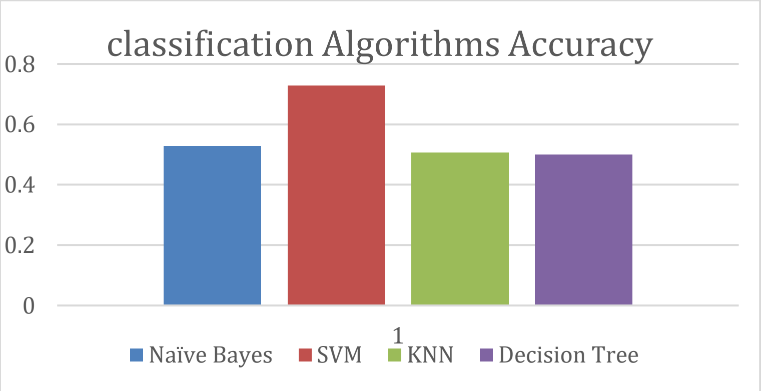


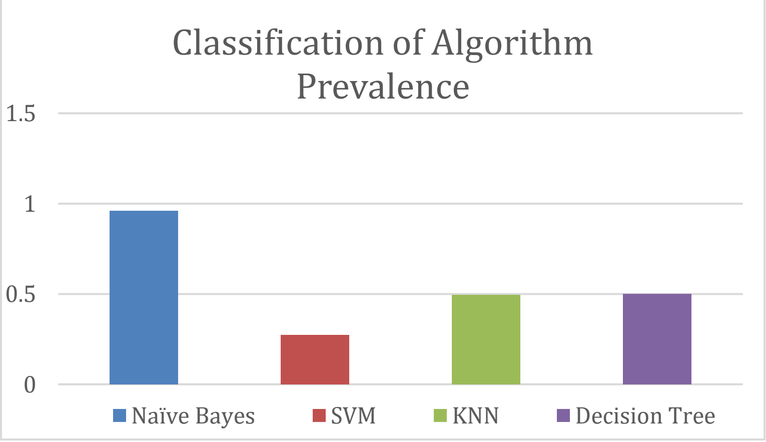
Fig. 2(b): Decision tree plotted for movie review dataset

Classification techniques came up with results of measures from confusion matrix. They accuracy, kappa, specificity, sensitivity etc. These results are compared among the classifiers. Naïve Bayes, SVM, KNN, Decision tree have accuracies 52%, 72%, 50.5%, 50% respectively.

Table 5: Comparative results of various classification techniques

Classifiers	Naïve Bayes	SVM	KNN	Decision Tree			
				M -ve	M +ve	negative	positive
Accuracy	0.528	0.7278	0.505	0.5	0.5	0.5	0.5
Kappa	0.0011	0	0	0	0	0	0
McNemar’s test p-value	<2e-16	<2e-16	4.219e-12	NA	NA	NA	NA
Sensitivity	0.95849	0.000	0.0	0	0.0	1.0	0.0
Specificity	0.04255	1.000	1.00	1	1.0	0.0	1.0
Prevalence	0.95800	0.2722	.495	0.04	0.2	0.5	0.0
Detection Prevalence	0.95800	0	0	0.00	0.0	1.0	0.0
Balanced Accuracy	0.50052	0.50	0.50	0.50	0.5	0.5	0.50





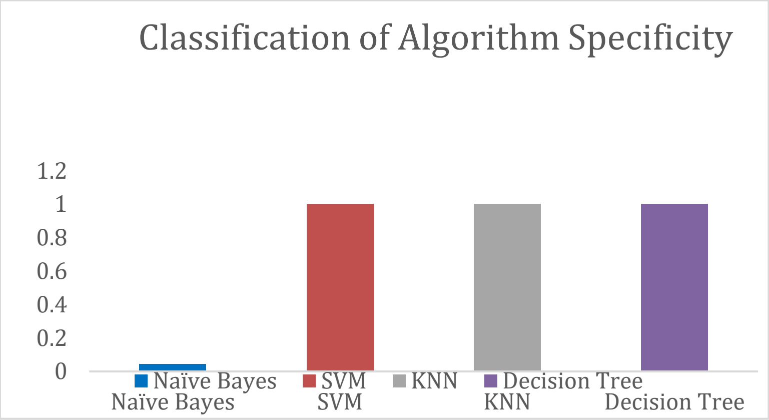


Fig. 2(c): Accuracy, Prevalence, Specificity plots of various classifiers

Sentiments analyzed from the classified data are distributed into ten emotions provided by the packages of R programming. The results are plotted on a bar chart.

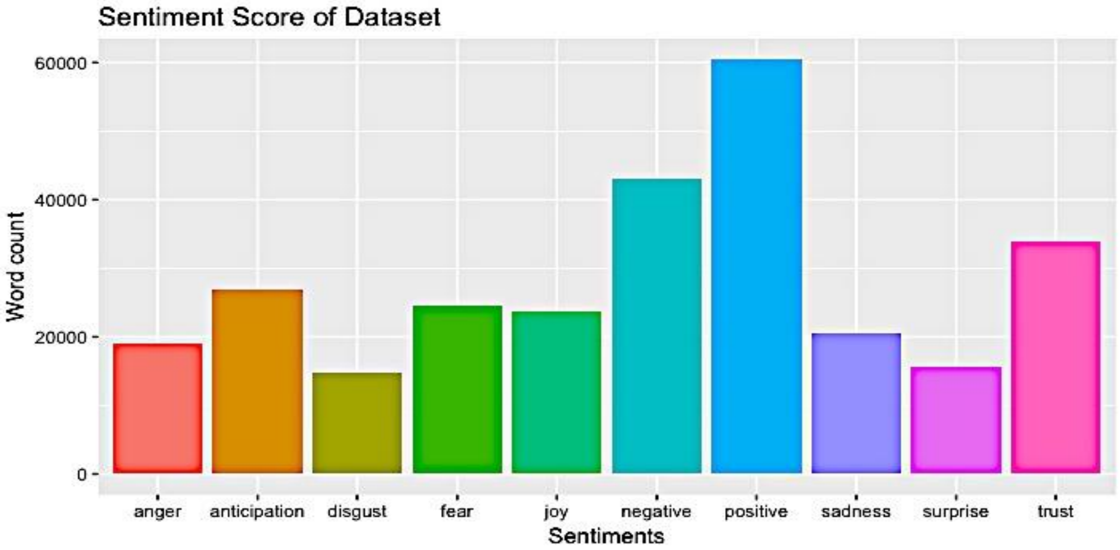


Fig. 3: Sentiment scores of movie reviews

7. Conclusion

In this paper, we have conducted experimental study on sentiment classification over a corpus of movie reviews. Considering different classifiers like Naïve Bayes, SVM, Decision tree and KNN text is classified. We have constructed an architecture that performs required analysis on those classification techniques and sentimental results are calculated. Accuracies of test set predictions are listed out. Among which SVM came out with best results of accuracy and specificity. Accuracy gained using SVM is 72.7%. There are some limitations found in other classifiers like KNN classifier cannot handle more amounts of data with ease. Decision tree classifier is not applicable for linear data which has only one factor to be processed. Sentiment is also classified among ten basic sentiments provided by syuzhet package in R. We came up with plot that represents sentiments available in the text. Wordcloud which represents terms of the documents as per their frequencies are also plotted. Accuracy, Prevalence and Specificity gives reports of the efficiency of the classification techniques used. This model helps in finding the best classification technique and user emotion for that particular aspect. This work can be extended by selecting different features for classification. Different other classification techniques might give better accuracy.

References

[1] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093-1113, 2014.

[2] A. Deshwal and S. Sharma, "Twitter sentiment analysis using various classification algorithms", 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2016.

[3] Doaa Mohey El-Din , Hoda M.O. Mokhtar ,Osama Ismael, "Online Paper Review Analysis", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 9, 2015.

[4] B. Liu, "Sentiment Analysis and Opinion Mining", Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1-167, 2012.

[5] A. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis", 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (ICATCCT), 2016.

[6] A. Pak, and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining, " Special Issue of International Journal of Computer Application, France: Universida Paris-Sud, 2010.

[7] DauméIII and D. Marcu , "Domain adaptation for statistical classifiers", Journal of Artificial Intelligence Research, 26:101–126, 2006.

[8] Jaap Kamps, Robert J. Mokken, Maarten Marx, and Maarten de Rijke, ‘Using WordNet to measure semantic orientation of adjectives’, in Pro- ceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), pp. 1115–1118. European Language Re- sources Association, Paris, (2004).

[9] S. Argamon, C. Whitelaw, P. Chase, S. R. Hota, N. Garg, and S. Levitan. Stylistic text classification using functional lexical features: Research articles. J. Am. Soc. Inf. Sci. Technol., 58(6):802–822, Apr. 2007.

[10] Dr. Seetaiah Kilaru, Hari Kishore K, Sravani T, Anvesh Chowdary L, Balaji T “Review and Analysis of Promising Technologies with Respect to fifth Generation Networks”, 2014 First International Conference on Networks & Soft Computing, ISSN:978-1-4799-3486-7/14,pp.270-273,August 2014.

[11] S.V.Manikanthan and T.Padmapriya “Recent Trends In M2m Communications In 4g Networks And Evolution Towards 5g”, International Journal of Pure and Applied Mathematics, ISSN NO: 1314-3395, Vol-115, Issue -8, Sep 2017.

[12] T. Padmapriya and V.Saminadan, “Improving Performance of Downlink LTE-Advanced Networks Using Advanced Networks Using Advanced feedback Mechanisms and SINR Model”, International Conference on Emerging Technology (ICET), vol.7, no.1, pp: 93, March 2014.

[13] Dr. Seetaiah Kilaru, Hari Kishore K, Sravani T, Anvesh Chowdary L, Balaji T “Review and Analysis of Promising Technologies with Respect to fifth Generation Networks”, 2014 First International Conference on Networks & Soft Computing, ISSN:978-1-4799-3486-7/14,pp.270-273,August2014.