Sentiment Analysis of IMDb Movie Reviews

Karthik Sivarama Krishnan[1] and Aayush Kumar Chaudhary[2]

Abstract— **Sentiment Analysis is the process of determining and categorizing the opinions expressed in the given text into positive and negative sentiments. The Sentiment Analysis method here is performed on the movie reviews available in the IMDb website. There are a large amount of reviews to every movie written by professionals and amateur reviewers. In every review, there are few keywords which would contribute to particular sentiments. These keywords are filtered out and classified into positive and negative review. The agents can be made to do the sentiment analysis on its own. The proposed idea helps the agent learn on its own, the different reviews on the movies and classify them into positive sentiment and negative sentiment. The data set is initially obtained, Pre-processing is done so as to remove all unwanted data and classifiers are used to classify the sentiments of the reviews. The data is validated and the efficiency of the classifiers are compared. The Support Vector Machine (SVM), Random Forest and Logistic Regression is used in this proposed method and their efficiency is compared and tabulated.**

I. INTRODUCTION

Movie reviews impact the future success of the movie in the present modern world. After having huge hype and expectations about every movie, the critics review the movie on the IMDb website. This review acts as a decider for every other audience. While selecting a perfect movie that would suit their current situation, people just go through the IMDb rating and reviews. The IMDb ratings for every movie is calculated by the overall average ratings provided by various reviewers. The reviewers also write a detailed review about each and every part and plot of the movie. The first impression about the movie comes through the reviews posted in IMDb website. The IMDb website is open to public and the reviews can be written by any critic. There may be amateur reviewers and professional reviewers too.

Every reader cannot read through thousands of review about a movie, this is where Artificial Intelligence comes into play. The Artificial Intelligence that is coded to the agent, analyze the samples of data and examines the words or sentences that give the positive sentiment and negative sentiment about the movie. The classifier is used to classify the given data and predicts the sentiment of the movie according to the sampled texts. The classifier also classifies the fake positive and fake negative results. These results are filtered out at the end after the classification is done. Here we propose the use the most common classifier Support Vector Machine (SVM). The support vector machine is used for the reduced data set and the complete data set, in-order to determine the efficiency throughout. The support vector machine algorithm is used in this proposal as it is the most common and the most efficient algorithm for classifying data set with only 2 classes. Another approach is by using the Random Forest classification algorithm. Random Forest algorithm is an Ensemble learning algorithm which predicts the data by aggregating the votes obtained with the old data. Baseline classification is done by the logistic regression algorithm. All three results are compared. The efficiency obtained in the random forest algorithm is compared with the support vector machine and the best classification algorithm for sentiment analysis is determined.

II. RELATED WORKS

The data must undergo the pre-processing techniques to remove all those unwanted data and noise in the given data set. The pre-processing techniques are used to convert the categorical attributes to numerical data. The authors in [5] used various pre-processing techniques and proved that the efficiency of claasifier was improved when the pre-processing techniques like Porter Stemming and Stop word removal were used in their data set.

The authors in [6] used a similar kind of data set available in the Kaggle website. The authors of this paper exhibited the usage of TF-IDF computation techniques which would help in improving the efficiency of the classifier by giving more weights or tokens to the particular words which contribute towards a particular sentiment.

Also the authors approach with the use of Support Vector Machine and Logistic Regression model helped in obtaining a greater accuracy with the proposed model. The authors in [7] considered the same Kaggle data set for IMDb sentiment analysis. The approach used by the authors of this paper focussed on using the Random Forest algorithm which gave them a greater accuracy. The author also used the Support Vector Machine and Logistic Regression along with the random forest.

The authors in [8] considered the data set provided by the stanford university and used the same Term frequency and Inverse Document Frequency techniques as a pre-processing algorithm. On considering the various pre-processing techniques used previously for sentiment analysis tasks per-

*Rochester Institute of Technology

[1]Graduate Student of Electrical and Microelectronics Department, Rochester Institute of Technology, Rochester, New York 14623, USA ks7585@rit.edu

[2] Ph.D Student of Center for Imaging Science Rochester Institute of Technology, Rochester, New York 14623, USA akc5959@rit.edu

Both authors [1] and [2] contribute equally formed by various authors, the most common pre-processing techniques performed to the data set are tokenization, stemming, stop words removal, and tf-idf computation. The positive and negative sentiments classification are done with the help of various classification algorithms such as linear classifier, Linear perceptron, decision trees, random forest, Bayes classifier or deep learning. The efficiency of various models are quite considerable and the algorithms works fine with the data set. The work Proposed in this sentiment analysis will be the use of classification algorithms like Support Vector Machine and Random Forest Algorithm with Logistic Regression as Baseline Algorithm. The pre processing techniques such as Stemming Algorithm, Stop Words removal, Tf-Idf compute and tokenization will be performed to clean the data set and set it for the classifier to classify with maximum accuracy

III. Dataset

The data set for Sentiment Analysis is provided by Stanford University Website. The system will be trained initially for about 25000 examples referred to as training examples in the dataset and later test will be compared with 25000 examples of the testing dataset. The overall data set does not contain more than 30 reviews for every movie. The ratings with 5 above are generally considered as positive review movies and the ratings less than 5 are generally considered as negative sentiment movies. The data set consists of corresponding ratings for every review. The data set also comes with a bag of words (BOW) which consists of 89,527 words which are compared with the given reviews. This could be used to generate the positive or negative sentiment given by movie based on the reviews. So the outputs can be the rating of the movie or the sentiment of the movie with the possible accuracy of the prediction. The data set consists of sentiment polarity labels. These BOW are available in .feat format which can be opened in LIBSVM packge.

IV. Simulation

The data is simulated using the MATLAB software. The MATLAB software has a wide application in dealing with the Machine Learning Algorithms. The raw data is imported to the MATLAB workspace. This raw data is filtered and the unwanted signals and pre-processing is done to the signals. This pre-processed data is then sent to the classifiers. The classifier uses particular Classification Algorithms to classify the given cleaned data set into two classes which corresponds to Positive and Negative sentiments respectively. The classification learner toolbox in MATLAB is used to perform various classification algorithms and determine the efficiencies of every algorithm.

V. Methods

The data is initially loaded to the MATLAB workspace. This Raw data consists of various noise and other unwanted keywords which does not contribute to sentiment analysis. In order to remove all these unwanted data from the given data set, it must undergo a series of pre-processing techniques. This pre-processing techniques helps in rendering a clean data set which could improve the performance and efficiency of the classifiers. The pre-processing techniques are used to convert the categorical attributes to numerical data. when the data set consists of words and not numbers, it has a specific and a different types of pre-processing techniques. The Various pre-processing techniques used in this experiment are Tokenization, Stemming, Stop-words removal and TF-IDF compute.

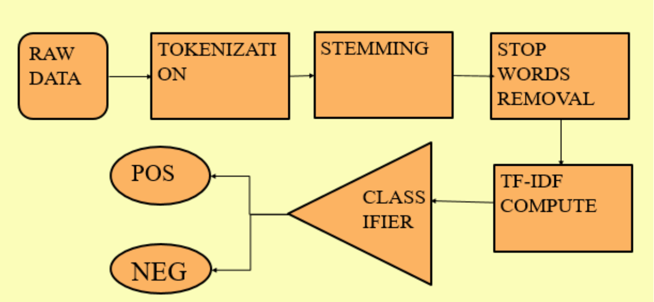

Fig. 1. Block Diagram of the Proposed Method

A. Tokenization

Tokenization is the process of removing the unwanted token and punctuations in the given data set. Tokenization is the first step to be performed to reduce the data set. Every review in the IMDb website consists of paragraphs of data in general English grammatical order. These sentence review about every movie consists of some basic punctuations like a Question mark (?), Comma (,) etc. These punctuations does not contribute to any sentiments. The data should be cleaned by removing all these punctuations and other HTML tags. The process of tokenization helps in performing the removal of these tokens. This is done to reduce the data set and helps in extracting the features which gives a strong hard-hit to the sentiments. This can in-turn increase the efficiency of the classifier.

B. Stemming

Stemming is the process of converting a word to its root word. IMDb reviews consists of various words which are used in various grammatical forms. The verbal form, the adjective form of the words will be present in every review paragraphs. The verbal form of the word in the paragraph should be replaced with its root word (or) parent word. For example, the word books should be replaced with the parent word book. The word ran should be replaced with the parent word run. The stemming process has a predefined algorithm which could be used so as to convert the given word to its root. The algorithm should run through each and every word on the reviews and then automatically convert every word to its root. This is the second step towards sentiment analysis. This does not reduce the data set but would increase the classification efficiency. This would help the classifier identify that the words ”run” converted from ”ran” is same as the word ”run” found in all other reviews.

C. Stop-Words Removal

The stop words are the words that are available in the reviews which does not contribute to any sentiment. These are generally the words which do not contribute towards the classification of the data. The stop words are the word list available in the English language. The examples of few stop words are a, the,to,too etc. These stop words are not removed from the reviews and they are available so has to write grammatically by the critics. But the agent that is reading the reviews here isn't a human and does not follow any grammar english. The agent here is an Intelligent system which could recognize words like ”bad”, ”horrible” as negative sentiments and words like ”good”, ”awesome” as positive sentiments.

The removal of the stop words became our first step towards pre-processing the data set. The data set available on the Stanford website is already tokenized and stemmed to the parent words. The removal of stop words was not done previously to the data set. The list of stop words available in the english language is collected and then these words are compared with the given Bag of Words (BOW). The data set provided consists of repetitions of words provided in Bag of Words (BOW) as features. These features will be reduced by this process. On comparing these stop words with the given Bag of Words (BOW), the Stop-words available in the given Bag of Words (BOW) can be removed. This will also help in removing the respective columns from the data set.

Removing the unwanted columns (or) the columns that does not contribute towards sentiments, helps in reducing the data set and also improve the classification efficiency of the classifier.

D. TF-IDF Compute

TF stands for time frequency and IDF stands for inverse document frequency. This process is used to set up a weight for the data. This process symbolizes the use of ranks for the important set of data or the keywords. This is the most useful technique for the ranking of keywords in the sentiment analysis process. For analyzing the sentiment of the reviews, there are few words which give a strongly positive review and few words giving strong negative reviews. The words which provide strong positive sentiments are given more weightage than the other words in the sentiments. This process is done in Matlab so as to obtain the words with greater influence on sentiments. This process does not reduce the data set but would help the classifier to increase its efficiency of classifying the data.

TF(t) = (Total occurrences of term t)/(Total number of terms in the document).

IDF(t) = [(Total number of documents)/(Number of documents with term t)] log_e

After pre-processing the features reduce to 5635 from 89,527

E. Baseline Algorithm

Baseline algorithm is the algorithm with which the classifier's efficiency is compared. Here we use the Logistic Regression as a Baseline algorithm. Logistic Regression is actually a classification algorithm which is used to classify the classes. There are two classes in this proposed model. class 0 and class 1. where class 0 represents the negative sentiment and class 1 represents the positive sentiment. Logistic Regression gives values between 0 and 1.

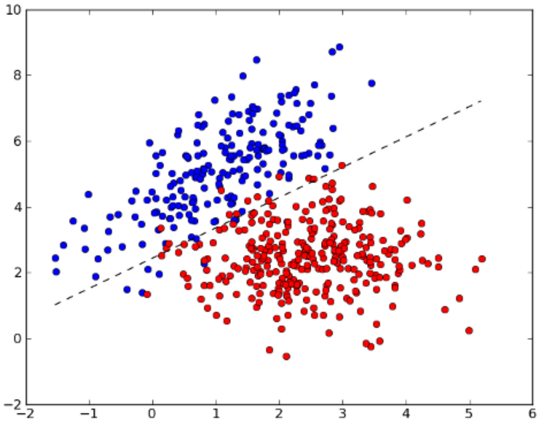

Fig. 2. Logistic regression

Logistic regression is available in the classification learner toolbox of Matlab 2017a. The train dataset is used first to train and get the accuracy. The training data reduces to train label and train dataset. The train label consists of class 0 and class 1. The train data set consists of the filtered data set. These variables are imported to the Classfication learner toolbox in Matlab. The class label is taken as the response and the train data set is taken as the predicators. The holdout validation is performed as the data set is huge for sentiment analysis. Now in the classification drop down box, logistic regression is selected and the model is made to train on the data set. Now the efficiency of the algorithm is obtained. The precision and Recall is obtained by continuous training the algorithm. This efficiency is considered as the baseline and the other classification algorithms are compared with the efficiency of the Logistic Regression algorithm.

Logistic Regression is used in place of Linear regression as the linear regression gave varying results with output classification decimal values greater than 0 and greater than 1. The classes available in this proposed method is 0 class and 1 class. But Linear regression did not help with classification. Here, Logistic regression was useful as it always has values between 0 and 1 and classifies the group of classes according to the values taken in correspondance with the vector dividing the classes into two. Once the classification is done, the model is exported to the workspace and this model is used to classify the test data.

$$\sigma(t) = 1/(1 + e^-t)) \quad (1)$$

$$t = linearcombinationofclasses \quad (2)$$

$$\sigma(t) = logisticfunction \quad (3)$$

F. Support Vector Machine (SVM)

Support Vector Machine is the most common and the most efficient classification algorithm. Support Vector Machine uses a hyper-plane to differentiate the two classes. The algorithm draws the two support vectors considering the last point in both the classes. The support vector assigns and draws the Hyperplane with reference to the support vectors. This hyper plane depends on the arrangement of the two classes to be classified.

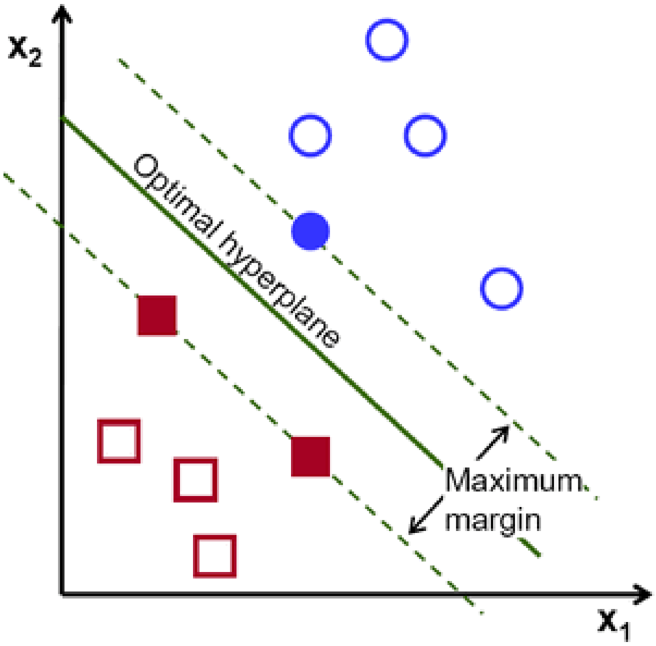

Fig. 3. Support Vector Machine

LIBSVM package is used here to open the .feat files provided in the data set. LIBSVM stands for Library for Support Vector Machines. The LIBSVM package is the SVM package provided by Nanyang Technological University, Singapore. This package is basically written in c++ language and is designed to work in cross platforms. This package is used for classification of two classes in this proposed model.

The support vector machine algorithm is performed for the entire data set and the efficiency is low when compared to the reduced data set. The SVM classification is done using the Linear and Non Linear Kernels. The Non Linear Kernel used in this classification is Radial Basis Function (RBF) kernel. RBF kernel is employed in this proposal as the support vector machines and other models employing the kernel function do not scale well when operating with a data set consisting of large numbers of training samples.

$$K(X, X') = exp[(-||X - X'||^2)/(2\sigma^2)] \quad (4)$$

Here the X and X' are the 2 classes considered and sigma is the kernel parameter. Here we considered $\sigma = 0.72$

G. Random Forest Algorithm

Random Forest Algorithm is an ensemble learning algorithm used for classification of classes. Random Forest algorithm works by constructing multiple decision trees at training time and classifying the classes in the mode of different classes seperable. The process of Bootstrap aggregating is used to improve the efficiency of classification. The process of bootstrapping is also called as Bagging. This process helps in avoiding the overfitting.

Generally a tree is constructed from a random sample which is formed with replacement from a subset of data. The classification is done so as tree tries to split on the best split among the subset pf predictors rather than the complete set of predictors. This helps in case that the values so obtained will atleast find one best match among all. Finally final voting will be done in order to estimate the class in which it belongs to. This helps to maximize the accuracy and prediction because there must be many trees that must be classifying it to correct tree.

The two important variables are number of trees and number of predictors choosen for each trial which is normally equal to the square root of total number of predictors for classification.For testing it does tree splitting with a bootstrap sample and then verification or validation is done by the sample called as out of bag sample which is selected as shown in figure below.

Random Forest Algorithm is performed in the Classification learner toolbox of MATLAB software. This toolbox is used to extract the variables from the workspace and classify the data by considering the labels as response and the dataset as predictor. Since this algorithm uses decision trees and the data set available is huge, few data is heldout and the classification is done.

VI. EVALUATION MEASURES

There are 3 major evaluation measures considered for measuring the performance of the classification method used with the given data set in the proposed system. The various performance measures are Precision, Recall and Accuracy.

A. Precision

In this method of information extraction, precision is defined as the fraction of the extracted data which are

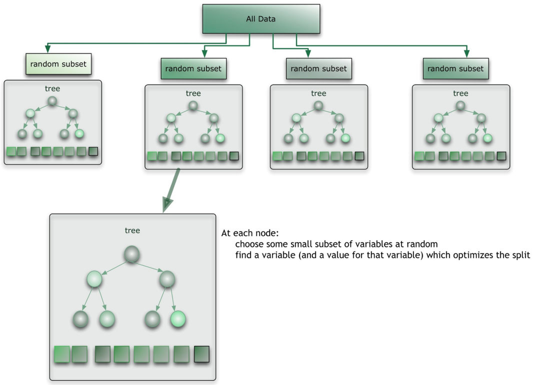
All Data
random subset
random subset
random subset
random subset
tree
tree
tree
tree
tree
At each node:
choose some small subset of variables at random
find a variable (and a value for that variable) which optimizes the split

Fig. 4. Random Forest Tree

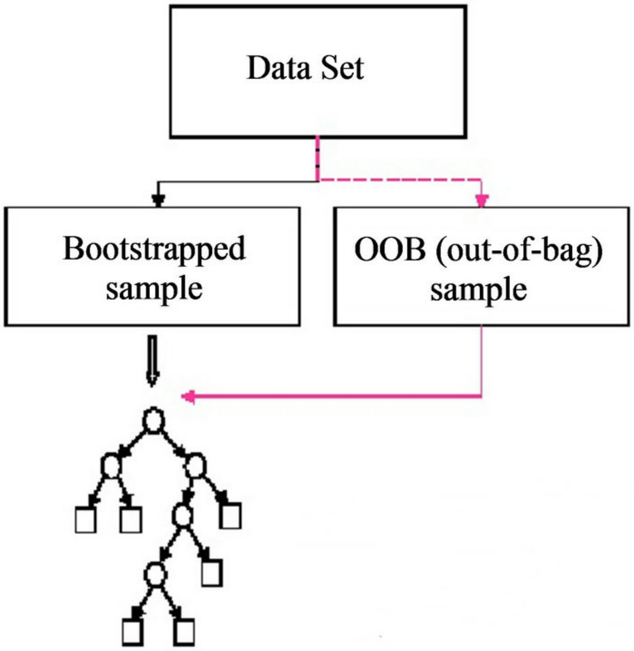

Fig. 5. Bootstrap and out of bag samples for random forest

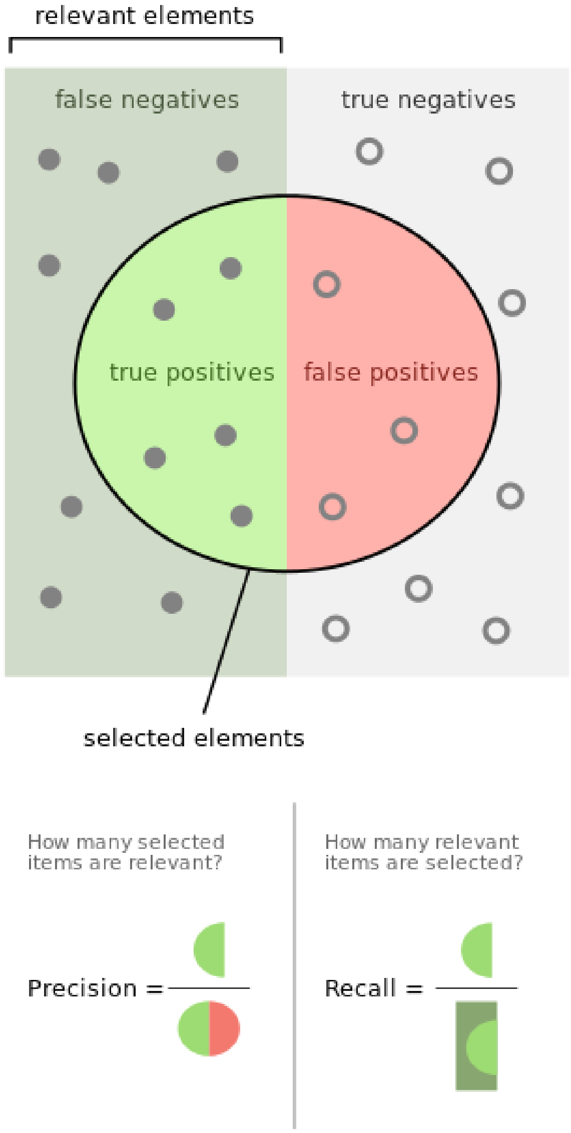

Fig. 6. Figure illustrating precision and recall

relevant to the given class

$$Precision = |(R \cap r)/r| \tag{5}$$

R = Relevant data
r = Retrieved data

B. Recall

Recall is defined as the fraction of data which are relevant to the class are successfully retrieved

$$Recall = |(R \cap r)/R| \tag{6}$$

R = Relevant data
r = Retrieved data

C. Accuracy

Accuracy is defined as the number of true results (both true positives and true negatives) among the total number of cases examined. This includes true positives, true negatives, false positives and false negatives

VII. RESULTS

The Support Vector Machine technique is performed with the entire data set intially without having any Pre processing techniques. The efficiency obtained in this method is comparetively low. The reason being the availability of words which does not contribute towards any sentiment and the agent unfortunately classified the words to a particular class.

The Data is now rolled over the pre processing algorithms and the data is filtered. also the features is reduced and unwanted data set is removed. Now this data set is given to the Support Vector Machine. The efficiency is now

TABLE I
CLASSIFICATION EFFICIENCY

CLASSIFIER	ACCURACY	PRECISION	RECALL
SVM With Complete Data-set	0.8503	0.8412	0.8620
LOGISTIC REGRESSION	0.893	0.8817	0.9020
LINEAR SVM	0.9038	0.9108	0.8955
RANDOM FOREST	0.9902	0.9914	0.98912
NON LINEAR SVM	0.99353	0.99904	0.9960

improved a lot and it is around 99 Percentage.

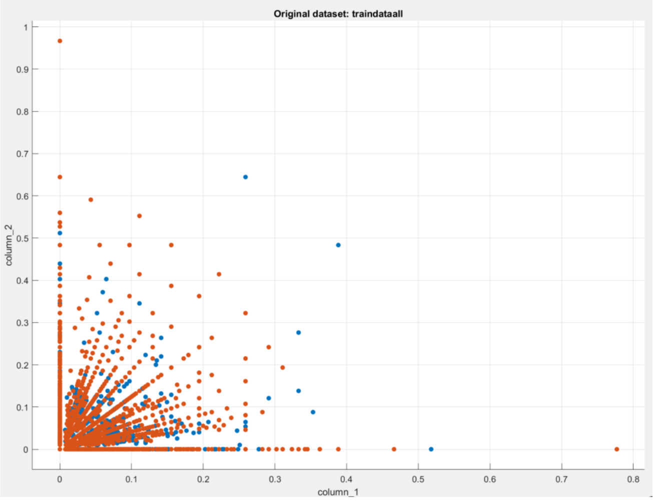

Fig. 7. Figure illustrating test data

Similarly the Random Forest Algorithm is implemented with the data set and the efficiency obtained from this classifier is pretty close to that of the Support Vector Machine.

Now the Accuracy, precision and Recall percentages for the algorithm is calculated and tabulated so as to make the results easy for comparison. The Baseline Algorithm logistic regression results are compared with both the algorithms and it is seen that the Non Linear RBF kernel SVM gave the highest classification accuracy. The results are shown in table 1.

DISCUSSION

RBF kernel used in the non linear SVM gave much higher accuracy than the linear SVM as the given data is not linear seperable. When using the linear SVM to the given model, the output efficiency was low when compared with all other accuracies. The linear regression did not classify the two classes in the given data set. linear Regression classify the classes by drawing a straight line and naming one class of data points as class 1 and the other class of data points as class 0. If there is a data point far to the end of the group points, then the linear regression tries to draw a linear line which seperates the two classes. while drawing this line, the data of the classes gets classified wrong. The classified output values from the linear regression becomes decimal values greater than 0 and decimal values greater than 1. But in our case, we had only two classes 0 and 1 and the classifier classified it with different decimal values.

This is the case we use logistic regression. Logistic regression is used to predict values between 0 and 1. It consideres a imaginary line seperating the two classes with the different values and classifying one group as class 0 and the other group as class 1.

The main problem that we faced in this proposed experiment is that, since the data set is too big, the processing time for the MATLAB took more and the initial pre processing of the data set took around 12 hours of time. Someone the MATLAB crashes as the data set is greater than 1 GB in memory and the processor fails in handling this big data set. The entire computer freezes and some times turns off and does not restart for the next 15 to 20 minutes. We used to put the data set into process the over night by putting the computer to do not sleep mode. This is the major hard time we had with this given data set.

After reducing the data set by removing the unwanted data, the processing time was much better but still it took around 4-6. hours. AFter reducing the data set, the size of the new data set reduced to 1 GB and then the processing time improved a little. But even after reduction, the computer crashed many times while handling huge memory. While using the toolbox for classification, the training of models took a lot of time and at the end ’failed’. This happened for logistic regression. we tried working in College labs and library systems. The computer crashed every time. The CPU utilization became full and entire computer froze. During this case, an addition of good GPU memory is required in order to prevent computer from crashing.

CONCLUSION

The proposed model is implemented by collecting the data set from the Stanford University website, pre-processing is done on the data set thereby reducing the unwanted informations and signals from the raw data set and the filtered data is sent to the classifier to perform the classification algorithm. The different classifiers like Logistic regression, Support Vector Machine and Random forest algorithms are performed with the filtered data set and the efficiency of the classifier is obtained and compared. The Precision and

Recall is calculated along with accuracy of every classifier. On comparing the efficiency of every algorithm, we could see that the Non Linear SVM using RBF kernel gave the maximum accuracy when compared to other algorithms. The random forest algorithm also gave the similar accuracy. Both Random Forest and RBF kernel SVM gave a really good efficiency for Sentiment Analysis

FUTURE WORK

On having a successful classification of the huge IMDb review data set, we are planning to work with the data sets of other social media like Facebook, Twitter and work on various different analysis like spam filter, fake identity etc.

ACKNOWLEDGMENT

We thanks Professor Dr. Sildomer T. Monteiro for his immense guidance with this project. The Teaching Assistant Utsav Gewali was really helpful in framing our ideas into picture.

References

[1] G. O. Young, Synthetic structure of industrial plastics (Book style with paper title and editor), in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 1564.

[2] W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123135.

[3] H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.

[4] B. Smith, An approach to graphs of linear forms (Unpublished work style), unpublished.

[5] E. H. Miller, A note on reflector arrays (Periodical styleAccepted for publication), IEEE Trans. Antennas Propagat., to be publised.

[6] J. Wang, Fundamentals of erbium-doped fiber amplifiers arrays (Periodical styleSubmitted for publication), IEEE J. Quantum Electron., submitted for publication.

[7] C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[8] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style), IEEE Transl. J. Magn.Jpn., vol. 2, Aug. 1987, pp. 740741 [Dig. 9th Annu. Conf. Magnetics Japan, 1982, p. 301].

[9] M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.

[10] J. U. Duncombe, Infrared navigationPart I: An assessment of feasibility (Periodical style), IEEE Trans. Electron Devices, vol. ED-11, pp. 3439, Jan. 1959.

[11] S. Chen, B. Mulgrew, and P. M. Grant, A clustering technique for digital communications channel equalization using radial basis function networks, IEEE Trans. Neural Networks, vol. 4, pp. 570578, July 1993.

[12] R. W. Lucky, Automatic equalization for digital communication, Bell Syst. Tech. J., vol. 44, no. 4, pp. 547588, Apr. 1965.

[13] S. P. Bingulac, On the compatibility of adaptive controllers (Published Conference Proceedings style), in Proc. 4th Annu. Allerton Conf. Circuits and Systems Theory, New York, 1994, pp. 816.

[14] G. R. Faulhaber, Design of service systems with priority reservation, in Conf. Rec. 1995 IEEE Int. Conf. Communications, pp. 38.

[15] W. D. Doyle, Magnetization reversal in films with biaxial anisotropy, in 1987 Proc. INTERMAG Conf., pp. 2.2-12.2-6.

[16] G. W. Juette and L. E. Zeffanella, Radio noise currents n short sections on bundle conductors (Presented Conference Paper style), presented at the IEEE Summer power Meeting, Dallas, TX, June 2227, 1990, Paper 90 SM 690-0 PWRS.

[17] J. G. Kreifeldt, An analysis of surface-detected EMG as an amplitude-modulated noise, presented at the 1989 Int. Conf. Medicine and Biological Engineering, Chicago, IL.

[18] J. Williams, Narrow-band analyzer (Thesis or Dissertation style), Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.

[19] N. Kawasaki, Parametric study of thermal and chemical nonequilibrium nozzle flow, M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.

[20] J. P. Wilkinson, Nonlinear resonant circuit devices (Patent style), U.S. Patent 3 624 12, July 16, 1990.