Sentiment Analysis of Top Colleges Using Twitter Data

Rajendra M[1], Gagana L[2], Aishwarya Anand[3]
[1] Asst. Professor, Dept of Computer Science Engineering
[2, 3] Dept of Computer Science Engineering
[1, 2, 3] Atria Institute of Technology

Abstract- *In today's world, opinions and reviews accessible to us are one of the most critical factors in formulating our views and influencing the success of a brand, product or service. With the advent and growth of social media in the world, stake holders often take to expressing their opinions on popular social media, namely Twitter. While Twitter data is extremely informative, it presents a challenge for analysis because of its humongous and disorganized nature. This paper is a thorough effort to dive into the novel domain of performing sentiment analysis of people's opinions regarding top colleges in India. Besides taking additional preprocessing measures like the expansion of net lingo and removal of duplicate tweets, a probabilistic model based on Bayes' theorem was used for spelling correction, which is overlooked in other research studies. This paper also highlights a comparison between the results obtained by exploiting the following machine learning algorithms: Naïve Bayes and Support Vector Machine and an Artificial Neural Network model: Multilayer Perceptron. Furthermore, a contrast has been presented between four different kernels of SVM: RBF, linear, polynomial and sigmoid.*

Keywords- Sentiment Analysis, Machine Learning, Neural Network, Opinion Mining, Natural Language Processing, Twitter.

I. INTRODUCTION

Social Media has captured the attention of the entire world as it is thundering fast in sending thoughts across the globe, user friendly and free of cost requiring only a working internet connection. People are extensively using this platform to share their thoughts loud and clear. Twitter is one such well known micro-blogging site getting around 500 million tweets per day [1]. Each user has a daily limit of 2,400 tweets and 140 characters per tweet [2]. Twitter users post (or 'tweet') every day about various subjects like products, services, day to day activities, places, personalities etc. Hence, Twitter data is of great germane as it can be used in various scenarios where companies or brands can utilize a direct connection to almost each of their client or user and thereby, improve upon their product. Consider a dissatisfied costumer of a telecommunication company voicing out his/her grievances about a particular plan he/she is subscribed to. Twitter also serves as a huge platform for users to know more and get direct comments about a product or a service in which they are interested [3]. Opinions and reviews in the form of tweets from customers, potential users and critics can easily influence the image and consequently, demand of a product/service being provided by a company. Hence, whether the stakeholder's opinion is positive/negative about their offering becomes a crucial and pressing question for the organization to ask and monitor.

According to fig. 1, roughly 34,582,000 out of an estimated 176,805,000 of the 18-23 year old age group in India receive higher education which equates to about 19.56% of the age group [4]. Many reputed government and private collegesinIndiaaimtowardsprovidingaclasseducationtotheirstudents and follow different ideologies, pedagogies and examination procedures. It becomes highly important for the interestedstudenttoevaluatethechoicesavailabletohim/herinselectinga college that not only furnishes the student with the desired academic or professional prowess but also equipshim/her with the right kind of learning tools according to his/her capabilities. Three of the premier colleges in India, namely the All India Institute of Medical Sciences (A.I.I.M.S.), the Indian Institute of Technology (I.I.T.) and the National Institute of Technology (N.I.T.) [5] have been analyzed to find the user's sentiment pertaining to the perception of these colleges and the magnitude of these opinions.

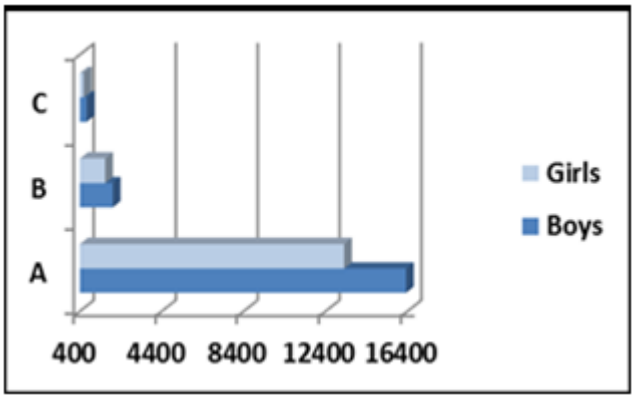

Fig. 1. Plot showing the enrollment of boys and girls of the 18-23 year old age group in Higher Education programs (based on the report compiled by the Ministry of Human Resource and Development in India). Here, A: All Categories, B: Scheduled Castes and C: Scheduled Tribes.

II. RELATEDWORK

Sentiment Analysis has been of avid interest to researchers lately. A lot of work has been put into it and there is a vast domain of its applications. A number of studies focus upon the popularity and reviews of products and services offered by different organizations. Arora, Li and Neville used Lexicon based Sentiment analysis on various smart phone brands to judge their popularity and reviews in the range of sentiment scores from -6 to 6 [6]. Similarly, Choi, Lee, Park, Na and Cho used sentiment analysis for laundry washers and televisions[7].

Researchers have also been working upon prediction of accuracy of tested dataset using Machine Learning algorithms. Kanakaraj and Guddeti used Natural Language Processing Techniques for sentiment analysis and compared Machine Learning Methods and Ensemble Methods to improve on the accuracy of the classification [8]. Bahrainian and Dengel compared different supervised, unsupervised methods along with their hybrid method (combining supervised and unsupervised methods) which outperformed other methods[9]. Pak and Paroubek performed Sentiment Analysis using formulas of Entropy and Salience and also implemented Naïve Bayes and SVM [10]. Shahheidari, Dong and Bin Daud used a Naïve Bayes classifier for classification and tested it for news, finance, job, movies and sports taking into consideration data mining on the basis of two emoticons (:) and :( ) [11]. Neethu

M. S. and Rajasree R used twitter posts on electronic products, compared the accuracy between different machine learning algorithms and further improved the accuracy by replacing repeated characters with two occurrences, including a slang dictionary and taking emoticons into consideration [12].

In addition, the area of neural networks has been investigated for performing sentiment analysis on benchmark datasets consisting of online product reviews. Bespalov, Bai, Qi and Shokoufandeh carried out binary classification on Amazon and TripAdvisor datasets using a Perceptron classifier and obtained one of the lowest error rates among their experiments of 7.59 and 7.37 on the two data sets respectively[13]. Jotheeswaranand Koteeswaran performed binary classification on the IMDB dataset by employing a Multi-layer Perceptron Neural Network and using Decision Tree-based Feature Ranking for feature extraction and a hybrid algorithm (based on Differential Evolution and Genetic Algorithm) for weight training, thereby obtaining a maximum classification accuracy of 83.25% [14]. Socher et al introduced a Semantic Treebank and a Recursive Neural Tensor Network which improves state of the art accuracy on binary classification from 80% to 85.4% on the movie data set introduced by Pang and Lee [15]. Santos and Gatti developed a deep convolutional neural network and obtained an accuracy of 85.7% and 86.4% on the aforementioned Stanford Sentiment Treebank and Stanford Twitter Sentiment Corpus (which is bounded by its classification based on emoticons) respectively[16].

III. PROCEDURE

A. Preprocessing of tweets

In this study, twitter data concerning three of the top colleges in India was obtained in JSON format for the duration of a month from 19 June, 2015 to 19 July, 2015. Unique tweets referring to A.I.I.M.S., I.I.T. and N.I.T. were extracted in order to reduce the bias of user opinions, eliminate redundant data and minimize the frequency of tweets which may be spam or fake reviews.

The tweets also provide information about the user, location, time-zone et cetera. In order to segregate the user opinion from user information, preprocessing was performed on the tweets. Removal of URLs, repeated letters in sequence which occurred more than twice with two of the same letter, ASCII escape sequences for Unicode characters, uninformative symbols and some but not all punctuations from the tweets was performed in order to sustain emoticons in the tweet. A dictionary of over 113,800 words was created in order to distinguish between words of English language and ambiguous words.

Expansion of SMS lingo, emoticons and abbreviations in net speak has been performed in order to include user opinions fitted rigidly under the constraint of 140 characters by referencing a slang dictionary which contains roughly 5,200 slang words and incorporates about 270emoticons.

B. Spelling Correction

Many a times, people unknowingly misspell words which may take away them eaning of the sentence. The ubiquitous approach of replacing more than two occurrences of a letter with two occurrences of the same letter is not a complete solution as misspellings may occur from the user's finger slipping to a nearby letter or the user's spelling the word phonetically. Therefore, these words have been corrected to the best possible effort by applying a probabilistic model based on Bayes' theorem which shows roughly 86.67% accuracy [17]. The best possible alternative to a spelling mistake is chosen in keeping with(1):

$$maxP(c|\ i) = maxP(i|c)P(c) \quad (1)$$

where c is the correct suggested alternative for the spelling error and i is the spelling error or typo. This equation calculates the maximum probability of the occurrence of the suggested c by the algorithm given i was typed which is equivalent to the product of maximum probability of the occurrence of i given c was typed and the probability of the occurrence of c. Here, the occurrences of c and i are considered as independent events.

Further, stop-words which are considered to be common words in the English language and do not contribute towards the sentiment of a sentence were removed with referral from a corpus of stop-words and also from the dictionary meant to test for ambiguous words in order to prevent redundant checks producing a resultant dictionary consisting of 113,253 English words. Thereafter, ambiguous words were removed which are either non-lexical conversation sounds, not part of the English language or badly misspelled typos beyond correction. Some tweets consisted entirely of URLs, stop-words, ambiguous words or a combination of these and hence, after the preprocessing, resulted in entirely blank lines which had to be removed.

C. Lexicon-based Sentiment Analysis using AFINN-111 word list

The final number of tweets obtained for the sentiment analysis after preprocessing was 1,295 for AIIMS, 1,794 for IIT and 1,833 for NIT; accounting for a total of 4,922 unique tweets. The tweets were then classified as positive, negative or neutral on the basis of all the unigram features present in the word list generated by Nielsen for each university separately. This approach has been shown to have accuracy of roughly 58.1% for three class problems [18] and to be one of the best present-day techniques to implement sentiment analysis[19].

In the sentiment analysis that was performed using the Nielsen lexicon, a magnitude of 0 was considered as neutral valence, greater than 0 was considered as positive valence while less than zero was considered as negative valence. Based on this, the resultant data from the above step was segregated into positive and negative opinions, while the neutral opinions were removed as the aim was to obtain strong and notable opinions on these colleges and neutral data would not contribute to this. 75 percent of the data thus procured was used as training data for the machine learning algorithms – Naïve Bayes and Support Vector Machine and the Artificial Neural Network model – Multilayer Perceptron while the rest was used as testing data for these algorithms and model.

D. Implementing Naïve Bayes algorithm for sentiment analysis

Rather than using Nielsen's word list as the prominent features for classification,chi-squared test was used to select the best k features for the polarity classification where k ranged from 10 to 15,000. The Naïve Bayes classifier utilizes all the features in these best k features and makes the 'naïve' assumption of independence of these features from each other. The Bayes rule employed in this classifier is represented by(2):

$$P(l|F)=P(l))\prod^{k}_{i=1} P(fi|lj) \quad (2)$$

where *l* is the class label and *F* is the feature vector defined as $F = \{f1, f2, \ldots, fk\}$.

Accuracies were determined by plotting Receiver Operating Characteristic (ROC) curves which plot the true positiverateasafunctionofthefalsepositiverateatvariousthresholdsettings.

It is essential to compare these rates and not the frequency of the classified predictions for a more precise take on the predicted sentiments. It was observed that the greatest accuracy was obtained for the best 1,000 features for each university and was lesser for other number of features as depicted by fig. 2. This could be because the classifier requires a number of training samples that is logarithmic in the number of features to fit a Naïve Bayes model[20].

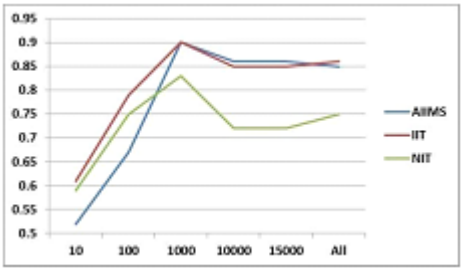

Fig. 2. The Y axis depicts the accuracy obtained by plotting ROC curves on all 3 data sets while the X axis depicts the number of best features chosen for the prediction.

E. Implementing Support Vector Machine algorithm for sentiment analysis

In a two-class polarity classification, Support Vector Machine isolates the two classes using a hyperplane with a decision boundary derived from (3):

$$w^T\Phi(x) + b = 0 \quad (3)$$

where w signifies the weight vector, Φ is the implicit embedding of data into a high dimensional feature space, x is the input sample feature and b is bias.

This aids in the choice of an optimal hyperplane which reduces the loss of data samples while simultaneously amplifying the margin. For exploiting the Support Vector Machine (SVM) algorithm, it was integral to choose an effective kernel. For text classification which is concerned with sparse data vectors, linear kernel is the optimum choice [21]. Still, as choice of kernel is still an open question in literature and results differ from one dataset to the other [22-23], a comparison was made using the following four kernels with optimal parameters in Python: Gaussian Radial Basis Function (R.B.F.) kernel, linear kernel, polynomial kernel and sigmoid kernel. Machine learning algorithms are sensitive to parameter optimization that is, different parameter values can significantly change the performance. Hence, the gamma parameter for RBF and sigmoid kernel was setto 0.9. The gamma parameter essentially represents the extent of influence of a single training example. High values of gamma represent a small radius whereas low values represent a large radius around that sample. It is crucial to choose an optimal value for gamma as large values of gamma may lead too verfitting where as small values of gamma may not capture the curve of the data. The degree of the polynomial kernel (a parameter irrelevant to other kernels) was set to 2 while gamma was fixed at 1.5. The degree represents the kernel function intended for fitting the data. It should be as low as possible to ensure low number of inflection points and thereby, avoid overfitting. The other parameters (penalty parameter, probability, shrinking etc.) were not modified or reflected no improvement following modification, and there fore, maintained default values.

The RBF and polynomial kernels performed relatively poorly, while the sigmoid and linear kernels performed well consistently on all three data sets as illustrated by fig. 3. The poorer prediction of the RBF kernel could be owing to the reason that an overly simplified model results in poor results whereas a complex model leads to overfitting. The solution to this is using a regularization coefficient which though yields good generalization, affects the exactness of the interpolation. The comparatively poor performance of the polynomial kernel could be due to there being a small difference between the number of positive and negative classifications [24]. Also, since the number of features is very large, it is not required to map the data to a higher dimensional space as non-linear mapping would not improve the performance and would also need more resources for training [25]. Further, the rationale of the positive performance of the sigmoid kernel is that an SVM model with a sigmoid kernel can be compared to a two-layer perceptron neural network [26] which can yield approximate solutions to extremely complex problems[27].

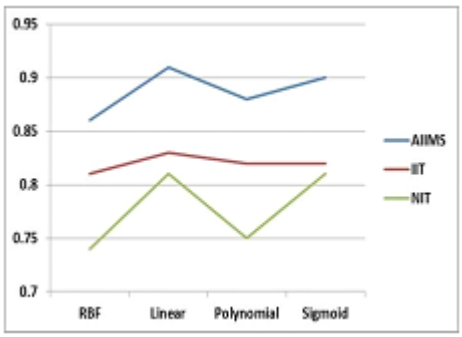

Fig. 3. The Y axis depicts the accuracy obtained by plotting ROC curves on all 3 data sets while the X axis depicts the different SVM kernels chosen for the prediction.

F. Implementing Neural Network for sentiment analysis

Artificial Neural Networks (ANNs) are a family of models in machine learning and cognitive science which are based upon existing neurological networks of the biological world. An ANN model utilized for supervised learning is the Multi-Layer Perceptron (MLP). It is a feed forward model that maps data onto a set of pertinent outputs. The training data is transmitted to the input layer which is first transferred onto the hidden intermediate layers and then to the outer layers as depicted by fig. 4. As the number of hidden layers increase, the generalization ability of the neural network is enhanced albeit a higher number of hidden layers leads to performance

issues. MLP NN (Neural Network) is implemented in two basic steps:-

a> Feed forward propagation
b> b> Backpropagation

In MLP NN, the primary step involves learning features from a feed forward propagation algorithm followed by minimising the cost function through backpropagation.

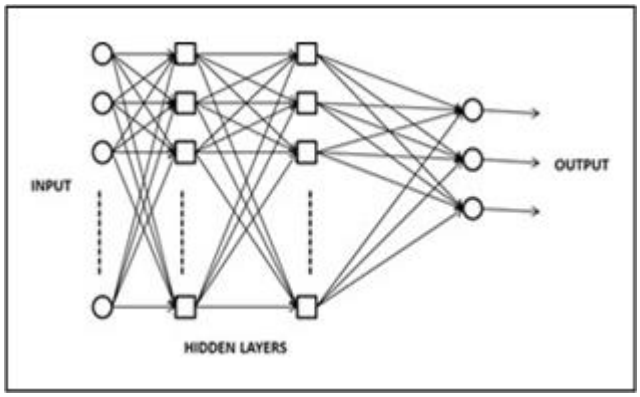

Fig. 4. Multilayer Perceptron with two hidden layers.

Both activation functions used in the current application are sigmoid, and are described by

$$\emptyset(y)=\tan(hi) \quad \textbf{(4)}$$

$$\emptyset(yi)=\frac{1}{(1+e—vi)} \quad \textbf{(5)}$$

Here yiis the output of the i^{th} node(neuron) and viis the weighted sum of the input synapses. (4) is the hyperbolic tangent which ranges from -1 to 1 while (5) is sigmoid and ranges from 0 to1.

The cost function for MLP NN is given by

$$j(8)=\sum^{N}\sum^{K} yilog(h8x^{i})—(1—y^{i})(1—)log(h8x^{i}) \quad (6)$$

i=1 k=1 k k k

where Θ signifies weights, h8x is the activation function, *y* refers to the output vector, *m* is the no. of trainee examples, *k* is the no. of output units and *x* signifies the input vector/trainee examples set.

To minimize the cost function, backpropagation or backward propagation is utilized. In backward propagation, the derivative is found in a backward fashion i.e., the error for the final layer is computed first and then the errors for previous layers are computed. During this process, the weights are updated simultaneously.

IV. RESULTS

The general sentiment derived from the dataset regarding the three colleges AIIMS, IIT and NIT were, as follows: a total of 399 tweets were regarded as positive, 231 as negative and 665 as neutral for AIIMS. 564 tweets were classified as positive, 364 as negative and 866 as neutral for IIT. 501 tweets were considered as positive, 350 as negative and 982 as neutral for NIT. As previously mentioned, a magnitude of 0 was considered as neutral valence, greater than 0 was considered as positive valence while less than zero was considered as negative valence. Table I displays a few of the statistics obtained.

TABLE I. STATISTICS ON THE SENTIMENTS EXTRACTED FROMTWEETS

College	Ratio of positive to negative tweets	Average positive sentiment
AIIMS	1.73	4.56
IIT	1.55	2.93
NIT	1.43	2.94

AIIMS had the highest positive average sentiment and the ratio for positive to negative tweets. This translates to the observation that the positive tweets about AIIMS are more positive in the magnitude of their sentiment and also indicates that AIIMS is talked about positively more than it is talked about negatively the most among the three institutions.

The predictions made by the machine learning algorithms showed high accuracy. For measuring accuracy, ROC curves were constructed which plot the true positive rate as a function of the false positive rate at various threshold settings. Simply put, true positive rate depicts the number of samples predicted to be positive which were also positive in actuality. It is computed as the ratio of true positives to total positives. Whereas, false positive rate signifies the number of samples which were actually negative, but were predicted to be positive and is defined as the ratio of false positives to total negatives. The diagonals (red dotted line) in fig. 5a, 5b, 5d, 5e, 5g and 5h depict the ROC curve of a classifier which makes entirely random guesses for a large sample. An ROC curve above this diagonal represents good predictive results whereas one below the diagonal represents poor classification results (in which case the predictions may be reversed for better classification). By further extending this logic, larger the area under the ROC curve, more accurate is the predictive power of the algorithm. For the AIIMS dataset, Naïve Bayes achieved an accuracy of 90%, SVM achieved an accuracy of 91% and

Neural Network obtained an accuracy of 92.6% as depicted by fig. 5a, 5b, 5c.

Naïve Bayes attained an accuracy of 90% for the IIT dataset, SVM attained 83% accuracy and Neural Network obtained an accuracy of 89.6% as shown by the fig. 5d, 5e and5f.

Naïve Bayes accomplished 83% accuracy for the NIT data set, SVM accomplished 81% accuracy and Neural Network obtained an accuracy of 87.6% as illustrated by fig. 5g, 5h,

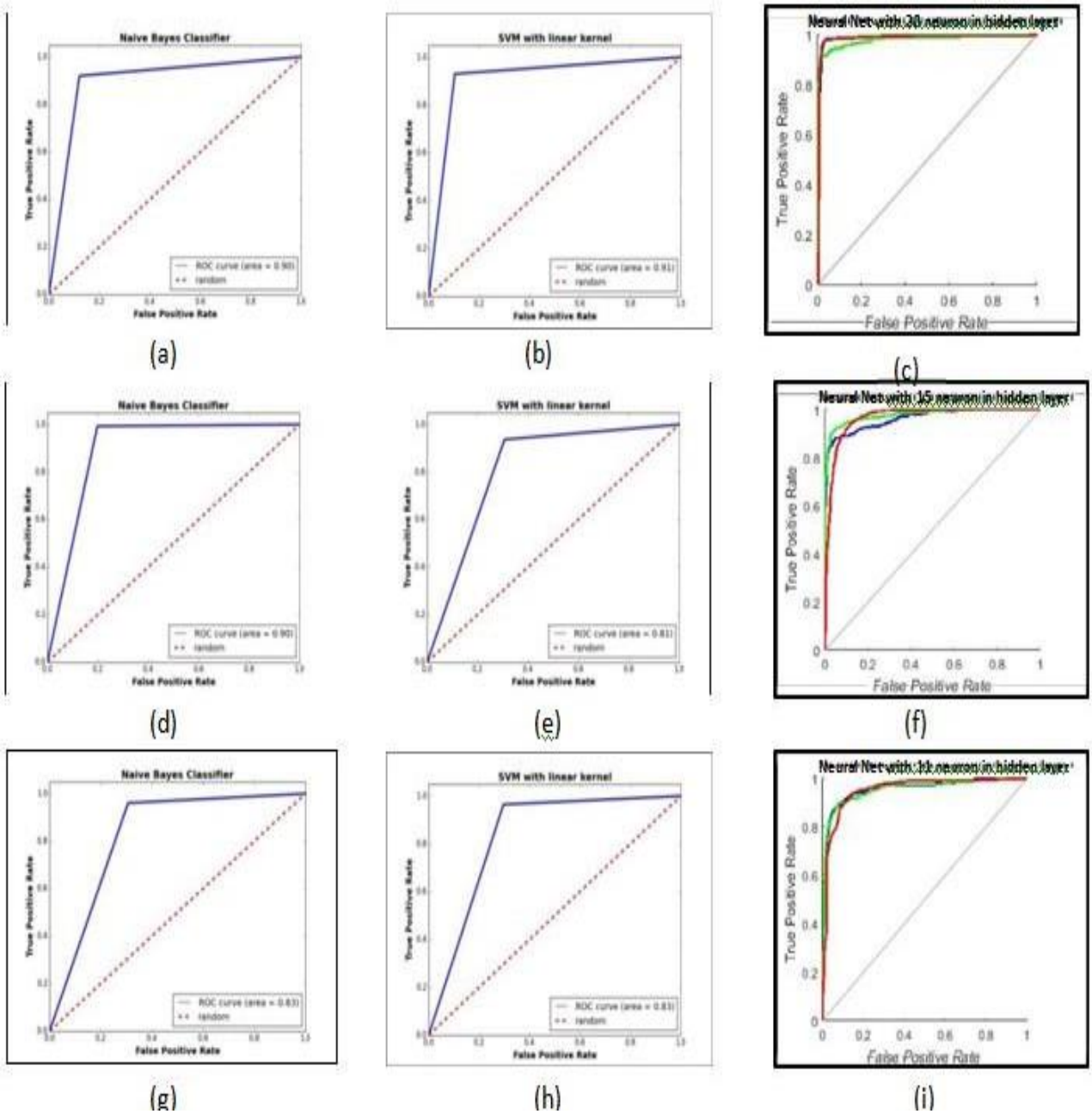

Fig.5. Receiver Operating Characteristic (ROC) Curves representing accuracy obtainedbyusingdifferentMachineLearningalgorithmsandNeuralNetwork model.

a) ROC for AIIMS using Naive Bayes, b) ROC for AIIMS using SVM, c) ROC for AIIMS using MLP NN, d) ROC for IIT using Naïve Bayes, e) ROC for IIT using SVM, f) ROC for IIT using MLP NN, g) ROC for NIT using Naive Bayes, h) ROC for NIT using SVM, i) ROC for NIT using MLPNN.

V. CONCLUSION

In conclusion, AIIMS is the most positively talked about college among the premier institutes of Indiaon Twitter. Comparison of the machine learning algorithms and ANN model suggests that MLP NN outperforms or matches the performance of Naïve Bayes which in turn, performs better than or almost equal to SVM on the three college datasets. Also, the most efficacious choice of kernel for SVM to perform text classification linear. Sentiment analysis is an effective way of classifying the opinions formulated by people regarding any topic, service or product. Automation of this task makes it easier to deal with the massive amount of data being produced by social websites like Twitter on a real-time basis. Polarity classification, in turn, aids in understanding the reception of a product or service, for instance, colleges in this case. Machine learning algorithms like Naïve Bayes and Support Vector Machine and an ANN model like Multilayer Perceptron yield promisingly accurate predictions on unseen data.

Multilayer Perceptron Neural Network surpasses the results yielded by the machine learning algorithms owing to its highly accurate approximation of the cost function, ideal number of hidden layers and learning the relationship among input and output variables at each step. Naïve Bayes outperforms Support Vector Machine for the purpose of textual polarity classification which is interesting because the model used by Naïve Bayes is simple (use of independent probabilities) and the probability estimates produced by such a model are of low quality. Yet, the classification decisions made by the Naïve Bayes model portray a good accuracy because each time a decision with the higher probability is being made[28].

REFERENCES

[1] Twitter Usage / Company Facts, https://about.twitter.com/company

[2] Posting a tweet, https://support.twitter.com/articles/15367-posting-a- tweet

[3] KingR.A.,Racherla P.and BushV.D., What We Know and Don't Know about Online Word-of-Mouth: A Review and Synthesis of the Literature, Journal of Interactive Marketing, vol. 28, issue 3, pp. 167- 183, August 2014

[4] Ministry of Human Resource Development, http://mhrd.gov.in/statist

[5] India's Best Colleges, 2015, http://indiatoday.intoday.in/bestcolleges/2015/

[6] Arora D., Li K.F. and Neville S.W., Consumers' sentiment analysis of popular phone brands and operating system preference using Twitter data: A feasibility study, 29th IEEE International Conference on Advanced Information Networking and Applications, pp.680-686, Gwangju, South Korea, March 2015

[7] ChoiC.,LeeJ.,ParkG.,NaJ.andChoW.,Voice of customer analysis for internet shopping malls, International Journal of Smart Home: IJSH, vol. 7, no. 5, pp. 291-304, September2013

[8] Kanakaraj M., Guddeti R M.R., Performance Analysis of Ensemble Methods on Twitter Sentiment Analysis using NLP Techniques, 9th IEEE International Conference on Semantic Computing, pp.169-170, Anaheim, California,2015

[9] Bahrainian S.-A., Dengel A., Sentiment Analysis and Summarization of Twitter Data", 16th IEEE International Conference on Computational Scienceand Engineering,pp.227-234,Sydney,Australia,December2013

[10] Pak A. and Paroubek P., Twitter as a Corpus for Sentiment Analysis and Opinion Mining, 7th International Conference on Language Resources and Evaluation, pp. 1320-1326, Valletta, Malta,May2010

[11] Shahheidari S., Dong H., Bin Daud M.N.R., Twitter sentiment mining: A multidomain analysis, 7th IEEE International Conference on Complex, Intelligent and Software Intensive Systems,pp.144-149, Taichung, Taiwan, July 2013

[12] Neethu M. S. and Rajasree R., Sentiment Analysis in Twitter using Machine Learning Techniques, 4th IEEE International Conference on Computing, Communications and Networking Technologies, pp.1-5, Tiruchengode, India, 2013

[13] BespalovD.,BaiB.,QiY.,and Shokouf andeh A.,Sentiment classification based on supervised latent n-gram analysis, 20th ACM international conference on Information and knowledge management, pp.375-382, New York, USA, 2011

[14] Jotheeswaran J. and Koteeswaran S., Decision Tree Based Feature Selection and Multilayer Perceptron for Sentiment Analysis, Journal of Engineering and Applied Sciences, vol. 10, issue 14, pp.5883-5894, January 2015

[15] Socher R., et al, Recursive Deep Models for Semantic Compositionality Overa Sentiment Treebank, Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, October2013. Tiruchengode, India, 2013

[16] BespalovD.,BaiB.,QiY.,andShokoufandehA.,Sentimentcla ssification based on supervised latent n-gram analysis, 20th ACM international conference on Information and knowledge management, pp.375-382, New York, USA, 2011

[17] Jotheeswaran J. and Koteeswaran S., Decision Tree Based Feature Selection and Multilayer Perceptron for

Sentiment Analysis, Journal of Engineering and Applied Sciences, vol. 10, issue 14, pp.5883-5894, January 2015