Sentiment Analysis of Twitter Data

Firoz Khan, Apoorva M, Meghana M, Pavan Kumar P Shimpi, Rakshanda B K
Department of information science,
GMIT, Davangere

***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, stakeholders 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 pre-processing 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*— Naïve Bayes, O*pinion, mining

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.

II. LITERATURE SURVEY

Sentiment Analysis has been of keen 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 Trip Advisor 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 datasets respectively [13]. Jotheeswaran and 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 dataset 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. METHODOLOGY

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 separate the user opinion from user information, pre-processing 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. The processing of tweets is explained in the fig 3.1.

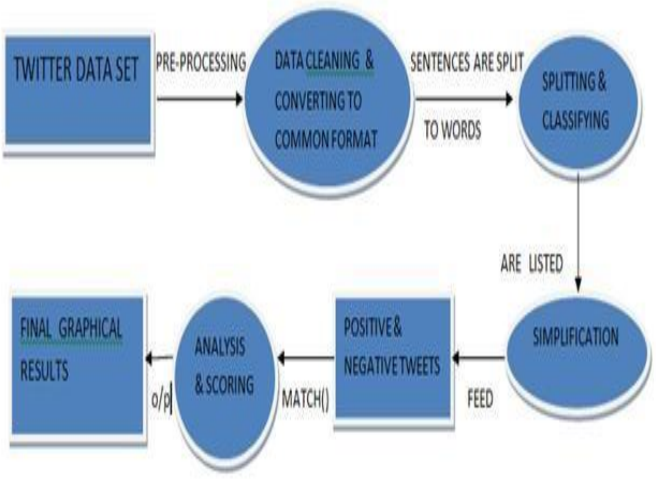


Fig 3.1: Processing of tweets

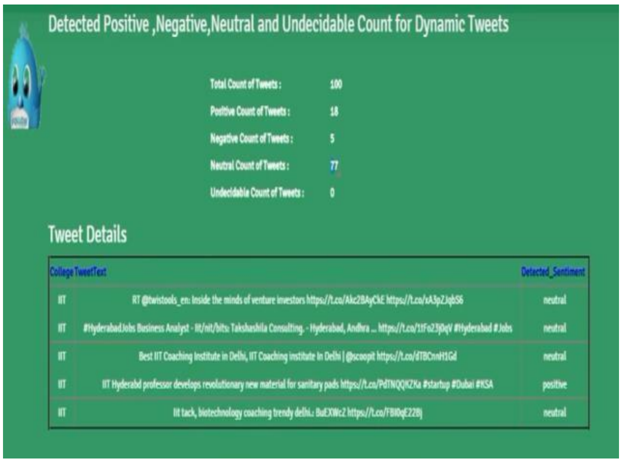
observation that the positive tweets about AIIMS are more positive in the magnitude of their sentiment and also indicates that AIIMS is talked about positively more than it is talked about negatively the most among the three institutions. The predictions made by the machine learning algorithms showed 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.

IV. .RESULTS

The general sentiment derived from the dataset regarding the three colleges AIIMS, IIT and NIT were, as follows: a

V. CONCLUSIONS

In conclusion, AIIMS is the most positively talked about college among the premier institutes of India on 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 is 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 out performs Support Vector Machine


Detected Positive ,Negative,Neutral and Undecidable Count for Dynamic Tweets

Total Count of Tweets :	100
Positive Count of Tweets :	18
Negative Count of Tweets :	5
Neutral Count of Tweets :	77
Undecidabla Count of Tweets :	0

Tweet Details

College	TweetText	Detected_Sentiment
IIT	RT @twistools_en: Inside the minds of venture investors https://t.co/Akc2BAyCkE https://t.co/xA3pZJqbS6	neutral
IIT	#HyderabadJobs Business Analyst - IIt/nit/bits: Takshashila Consulting. - Hyderabad, Andhra ... https://t.co/1tFo23j0qV #Hyderabad #Jobs	neutral
IIT	Best IIT Coaching Institute in Delhi, IIT Coaching institute In Delhi | @scoopit https://t.co/dTBCnnH1Gd	neutral
IIT	IIT Hyderabd professor develops revolutionary new material for sanitary pads https://t.co/PdTNQQKZKa #startup #Dubai #KSA	positive
IIT	IIt tack, biotechnology coaching trendy delhi.: BuEXWc2 https://t.co/FBI0qE22Bj	neutral

Fig 4.1. Final output of processed tweets

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.

VI. References

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

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

[3] King R. A., Racherla P. and Bush V. 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] 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

[5] Choi C., Lee J., Park G., Na J. and Cho W., Voice of customer analysis for internet shopping malls, International Journal of Smart Home: IJSH, vol. 7, no. 5, pp. 291-304, September 2013

[6] 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

[7] Bahrainian S.-A., Dengel A., Sentiment Analysis and Summarization of Twitter Data", 16th IEEE International Conference on Computational Science and Engineering, pp. 227-234, Sydney, Australia, December 2013

[8] 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, May 2010

[9] 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

[10] 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

[11] Bespalov D., Bai B., Qi Y., and Shokoufandeh 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

[12] 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

[13] Socher R., et al, Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, October 2013.

[14] dos Santos C. N. and Gatti M., Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts, 25th International Conference on Computational Linguistics, pp. 69-78, Dublin, Ireland, August 2014.

[15] Nielsen F.A., Making sense of microposts, Finn Årup Nielsen blog, https://finnaarupnielsen.wordpress.com/tag/sentimentanalysis/

[16] Koto F. and Adriani M., A Comparative Study on Twitter Sentiment Analysis: Which Features are Good?, Natural Language Processing and Information Systems, Lecture Notes in Computer Science vol. 9103, pp. 453-457, June 2015

[17] Ng A.Y., Jordan M. I., On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes, Advances in Neural Information Processing Systems vol. 14, pp. 841-848, 2002

[18] Karatzoglou A., Meyer D., Hornik K., Support vector machines in R., Journal of Statistical Software, vol. 15, issue: 9, April 2006

[19] Rajendran S., Kalpana B., A Comparative Study and Choice of an Appropriate Kernel for Support Vector Machines., International Journal of Soft Computing and Engineering (IJSCE), vol. 1, issue: 5, November 2011

[20] Abakar K. A. A., Yu C., Performance of SVM based on PUK kernel in comparison to SVM based on RBF kernel in prediction of yarn tenacity, Indian Journal of Fibre and Textile Research, vol. 39, pp. 55-59, March 2014

[21] Salazar D. A., Velez J. I., Salazar J. C., Comparison between SVM and Logistic Regression: Which One is Better to Discriminate?, Colombian Journal of Statistics, Special Issue Biostatistics, vol. 35, no. 2, pp. 223- 237, June 2012

[22] Hsu C.-W., Chang C.-C., Lin C.-J., A practical guide to support vector classification, National Taiwan University, Taipei, April 2010

[23] Girma H., A Tutorial on Support Vector Machine, Center of Experimental Mechanics, University of Ljubljana, 2009

[24] Multilayerperceptron,https://en.wikipedia.org/wiki/Multilayer _percepti on

[25] Manning C. and Raghavan P., Introduction to information retrieval, New York: Cambridge University Press, 2008.