VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

# Statistical Natural Language Processing

## (CSE6060)

**Mini Project - Review 1**

**Social network using streaming Sentiment Analysis**

**Date of submission**

13th June, 2020

**Supervised by,**

Prof. Arivoli A.

| Registration Number | Name |
|---|---|
| 19MCS0028 | Sumit Pravin Rathi |
| 19MCS0021 | Patel Rutu Manish |

# Social network using streaming Sentiment Analysis

## Problem Statement

Twitter is the popular blogging site where thousands of people exchange their thoughts daily in the form of tweets. The characteristics of tweet is to be short and simple way of expressions. We can perform the sentiment analysis on twitter data using many machine learning algorithms. This research paper will focus on techniques of sentiment analysis where we will perform how to extract tweets from twitter, preprocess them and finally classify them based on emotions. Eventually we will compare different sentiment analysis techniques and also the approaches containing twitter dataset. For the final product Naïve Bayes analysis is used for classification.

## Literature Survey

Semantic analysis has been in the limelight for many years now with many public domains using the method to analyse various kinds of data. The paper title "Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis" tells the procedure of januradha@vit.ac.in analysing and filtering the data required for sentiment analysis to be used in the sentiment analysis of twitter data. The methods used here such as Naive Bayes, Maximum Entropy and Support Vector Machine gives us a clear understanding of the accuracy differences between the classifications methods. Wordnet is used as the secondary dataset for lexical and semantic analysis, it is a comprehensive list of words with a pre-recorded positive and negative scores. The results achieved in this paper shows that wordnet achieves a maximum accuracy of 89.9% followed by Naive Bayes analyser scoring 88.2%,SVM with 85.5% and Maximum entropy with the least accuracy of 83.8.[1] Sentiment Analysis in Twitter using Machine Learning Techniques talks in detail about the pre-processing measures the authors took to identify emoticons and slang words from tweets and filtering the data raw to get an accurate result for the machine learning classification techniques used. The paper compares the results from different classification techniques such as Naive Bayes, SVM, Maximum Entropy classifier and Ensemble model with all of the above classifiers combined. The accuracy scores obtained by the training model shows that SVM, Maximum entropy and Ensemble model all score higher than Naive Bayes with all of them scoring 90% and Naive Bayes scoring 89% but the major difference is observed in the recall values with Naive Bayes scoring 91% while others score 93% concluding that a better pre-processing method will improve the model towards higher accuracy.[2]

As an application of SVM the paper Application of SVM in the sentiment analysis of twitter dataset goes into detail about the probabilistic latent semantic analysis using fisher kernel with the classification algorithm as SVM. A fisher kernel uses a fisher score method which is estimation of a parameter by the Gaussian mixture model using a Expectation Maximization algorithm. A convergence model is then developed using the kernel with SVM to provide better accuracy results. The comparison is between HIST- SVM, FK-SVM and PLSA-SVM which uses different kernel under SVM classifier. The experiments reveal that SVM classifier with fisher kernel has an 88% precision score comparing to 83% of PLSA-SVM model thus concluding that the fisher kernel greatly improves upon the primary kernel of SVM classifier.[3] A TfidfVectorizer which uses a term frequency inverse domain frequency values is discussed in “A TfidfVectorizer and SVM based sentiment analysis framework for text data corpus”. This method analyses the frequency of each word in a text corpus. This vectorized dataset is then fed into a SVM classifier to train the model. Different parameters such as max_df which is the maximum value of a document frequency, sublinear_tf which is a scaling parameter for tf to avoid the dominance of certain words. This proposed model is then compared with other available classification techniques such as decision trees, RNN, KNN, Naive Bayes and a normal SVM. The results show that the accuracy of the proposed model is 91% whereas the other models peak at 87% with KNN being the least at 72% concluding that better tf parameter values always lead to better accuracy results. [4]

Comparison of Feature Weighting in SVM Performance for Sentiment Analysis of Jakarta BRT gives an inside view of Jakarta's Bus Rapid Transit System and how the service uses public forums to analyse the sentiment of users on the bus service and issues faced by the users. This method uses different tokenization methods such as TF-IDF which stands for Term Frequency Inverse Different Frequency, TF-OR which stands for Term Frequency Odd Ration weighing method, TF-RF for Term Frequency Relevance Frequency and TF-CHI which stands for Term Frequency Chi-Square method. The study employs SVM for classification and uses a local dataset manually labelled by the authorities to divide positive and negative comments. Different tokenization schemes are then compared with TF-IDF giving the highest average mean accuracy of 79% while others come close to this value.[5] Sentiment Analysis on Twitter Data Using Support Vector Machine uses a broad range of dataset classification with phrase level, document level and aspect level classification. This is particularly useful for contextual analysis since most of the sentiment analysis uses word frequencies to determine the sentiment. A FB-LDA filter is used to remove the noise in the tweets by filtering out the stop words, unnecessary words like retweets and count. The final accuracy score of the sentiment analysis using this algorithm gives an accuracy score of 74% while this algorithm performs exceptionally well for Amazon product review dataset with an accuracy score of 97.5%. [6]

Sentiment Analysis of Twitter Data uses a unigram model to obtain the sentiment of a tweet. A unigram model is a tree kernel model which uses trees to represent the sentiment and is compared with other unigram models which requires more than 10,000 features to perform. It provides a less accurate score than the proposed model which uses only 100 features. It contains all kinds of data ranging from emoticons to tags. All the languages are considered and are converted to English using google translate for training and the emoticons are identified by using official wiki sources to say if the emoticon is positive or negative thus providing a bigger dataset than many sentiment analysis models. The tree kernel is designed using the Partial Tree method with prior polarity scoring method. The tree is divided into subtrees with polar and non-polar data on either side of the tree. The results show that the accuracy is 73% for the kernel model which concludes that the proposed model with 100 features is more accurate than the unigram model.[7] Sentiment Analysis and Summarization of Twitter Data proposes a hybrid classifier containing an aspect detector, an improved novel hybrid polarity detection system and a ranking algorithm. The baseline algorithm being SVM with a unigram model which is being compared with an unsupervised Polarity Detection algorithm. The model uses a pre-processing system which eliminates all the unnecessary items from the data corpus while also detecting the useful ones. A target word is checked for the sentiment which is used to gather a large information pool such as positive word followed by the target word, negative word followed by the target and many more. Evaluation results shows that the hybrid model achieves an accuracy score of 89.78% while the unsupervised achieving 81% and baseline SVM model achieving 86.70%.[8]

Multi-Class Text Sentiment Analysis discusses in detail about the different ML classifiers for sentiment analysis. This includes SVM, Decision tree, Random forest, Multi- layer perceptron, KNN, Quadratic discriminant analysis and Gaussian Naive Bayes. It also includes various CNN methods with different dimensions and neural networks. Making an unbalanced dataset balanced word2ve is used to convert text into numerical form for neural networks to understand. In CNN the architecture includes a convolutional layer with varying number of channels, followed by ReLU and Max-Pool activation layers; in the final layer, the model uses softmax function to output the probability predictions for each class. The experiment results show that textCNN with word2vec having the highest accuracy with 91.4% as the score while SVM with RBF kernel which stands for radial basis function kernel has the next best accuracy score of 83.7%. [9] The paper titled “A BigData approach for sentiment analysis of twitter data using Naive Bayes and SVM Algorithm” uses the sentiment analysis method for analysing the student feedback for the teachers. This is collected from many different platforms leading to an accumulation of big data. The proposed method uses both SVM and naive Bayes to detect the sentiment of the user and analysing the classifier which has the best accuracy score and then assigning that model higher priority. [10]

The paper “Effective Sentiment Analysis of Social Media Datasets using Naive Bayesian Classification” introduces a combination of a unigram and a bigram model to extract words from data and is trained using the Multinomial Naive Bayes classifier. The model uses a huge collection of data from different sources like movie reviews, twitter dataset, emoticon datasets, and sentiment lexicons. However the model is trained two times one with neutral sentiment which gives an accuracy of 65% for unigram and without neutral sentiment the accuracy obtained is 81.25% leading to better results. [11] Sentiment Analysis Using Naïve Bayes Classifier uses Naive Bayes analyser to check the probability of word count in the text corpus. The algorithm which uses Naive bayes calculates the frequencies of the word in the document and calculates the conditional probability of the keyword and then a uniform distribution is performed to avoid zero frequency problem after which the higher probability value is assigned to that particular document. The paper goes into detail about the accuracy of the positive and negative of the tweets as the desired output is achieved with an accurate prediction of whether the tweet is positive or negative. [12]

## Comparison of existing algorithms

| Author | Data set | Algorithm | Accuracy (%) |
|---|---|---|---|
| Geetika Gautam | Customer Review Twitter Dataset | Naive Bayes<br>Maximum Entropy<br>Maximum Entropy<br>Semantic Analysis (Word Net) | 88.2<br>83.8<br>85.5<br>89.9 |
| Neethu M.S. | Twitter posts about electronic products | Naive Bayes<br>SVM<br>Maximum Entropy<br>Ensemble Learning | 89.5<br>90.0<br>90.0<br>90.0 |
| Sayed-Ali Bahrainian | Twitter data on Smartphones | Unigram Feature, SVM, NB, MaxEnt Hybrid Approach | 89.78 |
| Dhiraj Gurkhe | Twitter Data | Unigram<br>Bigram<br>Uni+Bigram | 81.2<br>15.0<br>67.5 |
| Apoorv Agrawal | 11875 manually annotated tweets | Unigram<br>Senti-features<br>Kernel<br>Unigram+Senti-features<br>Kernel+Senti-features | 71.35<br>71.27<br>73.93<br>75.39<br>74.61 |

## References

[1] Gautam, Geetika, and Divakar Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." 2014 Seventh International Conference on Contemporary Computing (IC3). IEEE

[2] Neethu, M. S., and R. Rajasree. "Sentiment analysis in twitter using machine learning techniques." 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). IEEE,.

[3] Han, Kai-Xu, et al. "Application of Support Vector Machine (SVM) in the Sentiment Analysis of Twitter DataSet." Applied Sciences 10.3 (2020): 1125.

[4] Kumar, Vipin, and Basant Subba. "A TfidfVectorizer and SVM based sentiment analysis framework for text data corpus." 2020 National Conference on Communications (NCC). IEEE, 2020.

[5] Damayanti, Nourma Reizky, Teguh Bharata Adji, and Guntur Dharma Putra. "Comparison of Feature Weighting in SVM Performance for Sentiment Analysis of Jakarta BRT." Journal of Physics: Conference Series. Vol. 1196. No. 1. IOP Publishing, 2019.

[6] BholaneSavita, D., and Deipali Gore. "Sentiment analysis on twitter data using support vector machine." International Journal of Computer Science Trends and Technology (IJCST)–Volume 4: 365.

[7] Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." Proceedings of the Workshop on Language in Social Media (LSM).

[8] Bahrainian, Seyed-Ali, and Andreas Dengel. "Sentiment analysis and summarization of twitter data." 2013 IEEE 16th International Conference on Computational Science and Engineering. IEEE.

[9] Hong, Jihun, Alex Nam, and Austin Cai. "Multi-Class Text Sentiment Analysis." (2019).

[10] Jadon, Priyanshu, Deepshikha Bhatia, and Durgesh Kumar Mishra. "A BigData approach for sentiment analysis of twitter data using Naive Bayes and SVM Algorithm." 2019 Sixteenth International Conference on Wireless and Optical Communication Networks (WOCN). IEEE, 2019.

[11] Gurkhe, Dhiraj, Niraj Pal, and Rishit Bhatia. "Effective Sentiment Analysis of Social Media Datasets using Naive Bayesian Classification." International Journal of Computer Applications 975.8887: 99.

[12] Kavya Suppala, Narasinga Rao. “Sentiment Analysis Using Naïve Bayes Classifier” International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8 June, 2019