Vol. XVII & Issue No. 09 September - 2024
INDUSTRIAL ENGINEERING JOURNAL

GAUGE SENTIMENTS ON NEW INCOME TAX REFORMS IN INDIA WITH TWITTER DATA

Anima Srivastava & Tanveer J. Siddiqui
Department of Electronics and Communication, University of Allahabad.
Email: anima2424@rediffmail.com / tjsiddiqui@allduniv.ac.in

Abstract

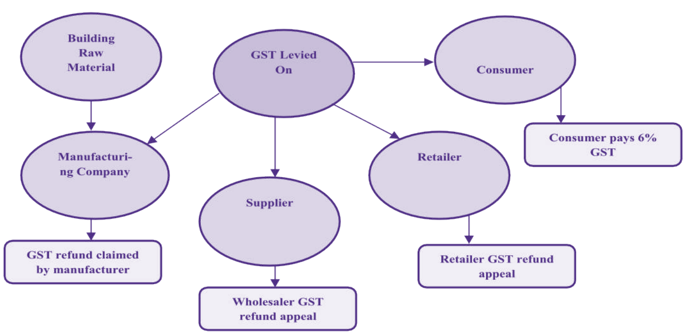
The initiation of Goods and Service Tax was one of India's most notable economic revolutions leading to extensive debates. It ushered a massive platform via Social Media websites for the general public to throw spotlight on their opinions of GST. Ascertaining public sentiment with these opinions will help harness future reforms. We accumulated GST-related tweets from last day of April 2018 until 1st of May 2018, and analyzed their polarization stimulated with this detailing. We also studied six different classifiers: Ridge Classifier, Linear SVC, Logistic Regression, Perceptron, K-Nearest Neighbor and Decision Tree. Using tf-idf feature form test runs were conducted for each classifier. The effectiveness of each scenario was assessed and found that ridge classifier observed the maximum accuracy 96%. The performance assessment is made with other available work on GST. All classifier performed better with tf-idf feature with the existing feature permutations also findings are compared with other existing works on identical dataset.

Keywords: Goods and Service Tax (GST), linear classification, Supervised Machine Learning.

1. INTRODUCTION

Sentiment Analysis is the process of analyzing text automatically to determine people's feelings [1] sentiments, attitudes and emotions towards certain products, services, events, organizations, individuals etc. Nowadays, social media websites have become a hub of opinionated content [2,3]. According to the statistics published on statista.com the number of social network users in India in 2016 was 168 million and the prediction is that it will reach to 258 million in 20191. People share their thoughts, experiences, views, and emotions on these websites on all kinds of topics on a regular basis. The opinions expressed on these websites provide valuable feedback on product, policies, services, movies, individuals, etc. [4]. This information is quite useful for companies, service providers, individuals, policy makers, government, political parties and celebrities. However, analyzing this huge volume of opinionated content manually is a herculean task. This has made automatic sentiment analysis or opinion mining a hot topic of research. Both machine learning and knowledge-based approaches have been used to automatically analyze textual data to know its polarity [5,6]. This paper focuses on analysis of tweet data related to GST (Goods and Service Tax) to identify the polarity of the sentiments expressed in it.

Figure-1 Mechanism of GST



The mechanism of GST levied is shown in Figure-1.GST mitigates the inadequacy in indirect taxes and improves tax compliance which in turn reduces the heavy taxes imposed on end customers by its cascading effect. Consequently, the GST is levied on manufactures, wholesaler, retailer and consumer out of which only consumer has to pay 6% in GST and the rest of lot claim it back, GST is a single tax that replaces all indirect taxes charged by central and state government of India [7]. GST was levied on manufacturer, wholesaler, retailer and consumer (Figure-1). It aims to combat the inadequacies of indirect tax and to improve tax compliance. However, the induction of GST invited a lot of criticism from a section of society who blame GST for the slowdown in the economy. Consequently, certain reforms have been made and the government is open for future reforms. Knowing the sentiments of general public may be of great interest for government in shaping the future reforms. Some earlier works reported on GST sentiment analysis [8,9]. In order to take a step forward in this direction, we investigate and report the performance of six different supervised machine learning classifiers including Ridge Classifier (RC), Linear Support Vector Classifier (LSVC), Logistic Regression (LG), Perceptron(P), K-Nearest Neighbor (KNN) and DT (Decision Tree) on GST tweet data. The rest of the paper is organized as follows: We'll go over some of the existing efforts in the next section involving GST sentiment analysis. Literature survey was discussed in section 2. Section 3 presents the details of the data preparation and the classifier used in this work. In section 4, the experimental investigations have been made and in section 5, the outcomes are compared. Finally, in section 6, conclusions are drawn.

2. LITERATURE SURVEY

Earlier work on GST sentiment analysis includes [10, 11, 12, 13, and 14]. Ganguly and Roy [10] analyzed opinions expressed about GST using Twitter data. They have collected all the tweets and re-tweets about GST from the day of its announcement till one day later (July 1st, 2017, to July 2nd 2017). The sentiment polarity is computed using the method presented in [15]. A cut-off of 0.25 is used to categorize tweets as positive or negative. Tweets with polarity score less than 0.25 are considered negative. The study also investigates the social connection among users who have expressed their opinions by building a directed graph based on data collected. In this graph, nodes correspond to users and a connection between two nodes tells that user have responded or retweeted posts. The clustering coefficient and length of average path in the resulting network was found to be 0.103 and 1.109 respectively, indicating that most nodes are not connected but are closer together. According to the polarity analysis, 38 percent of people support GST and 62 percent oppose it.

Gautam & Yadav [11] used WordNet based semantic analysis [16] to improve the results of supervised classifier. To classify product reviews, they used three distinct classifiers: Support Vector Machine, Maximum Entropy, and Nave Bayes. The maximum accuracy was observed using NB classifier. The output of the NB classifier was then used to label semantically related word as positive and negative. The semantic relatedness was derived using WordNet. The effectiveness of the classifiers was measured based on accuracy, precision and recall.

Tomar et. al. [12] used SVM to classify GST tweets. They experimented with two different models. The first model was trained on IMDB dataset while the second model was trained using a combination of dataset mainly composed of IMDB dataset, manually annotated tweets on GST. Both the models were tested on GST dataset collected from twitter. They reported an accuracy of 73.28% using model two (IMDB+ domain specific dataset). Implementation is done by using a modern open-source data platform Waikato Environment for Knowledge Analysis (WEKA) [17].

By combining manually annotated GST-related tweets with the IMDB dataset's labelled reviews, domain and time specific characteristics were used in the training dataset. They used two different models and evaluated precision, recall, f1-score and accuracy. The model-1 was trained on the IMDB movie review dataset and tested on GST-related tweets. The model-2 was developed and validated using IMDB dataset + Twitter dataset. GST related tweets collected from twitter microblog.

Das & Kolya [13] used NB classifier to tag tweets into one of the five categories: most positive, positive, normal, negative and most negative. Emojis were also considered in sentimental rating generation. They collected approximately 30,000 tweets from Twitter Streaming API and analyzed people's opinion about GST using Naïve Bayes algorithm. The dataset comprises of 10 days tweets on GST during implementation phase of GST in India. The sentiment rating for each of the five categories is reported on a 10-point scale (1 to 10).

Chaudhary and Paulose [14] proposed a new opinion mining method and model using Stanford CoreNLP, on newspaper headlines .Three different variants of support vector classification classifiers were used namely linear SVM, TF-IDF + linear SVM and Stochastic Gradient Descent (SGD). They evaluate the performance by three different models: Model A, Model B and Model C. Model B with bigram feature secure (91.52%) highest accuracy among all model used. We compare the performance of five different supervised classifiers on GST tweet data using uni-gram feature, bigram feature and combination of uni-gram and bigram. The dataset consists of 1897 tweets. The best performing case is compared with existing works on GST dataset.

3. METHODOLOGY

As no GST data was publicly available at the time of this work, data collection was a prerequisite. We collected tweets about GST using twitter API.

3.1 Dataset Preparation: We crawled twitter messages related to Goods and Service Tax (GST) from April 30, 2018 until May, 1, 2018 via the streaming API in keyword tracking mode using python client Tweepy. The keywords used are: #gst, #CGST, #SGST, #gst tax, #gstbenefits, #onenationonetax, #dualgst. We dropped non-English words occurring in these tweets. Only micro blog messages in English were retained. The data thus obtained contains re-tweets as well. This increases the size of the data but no new information. Therefore, we remove all duplicate re-tweets. We obtained 200 KB twitter messages comprising of 1897 tweets.

3.2 Classifiers Used: We use different supervised machine learning algorithms to evaluate sentiment on the GST dataset collected from twitter microblog [18]. These classifiers are discussed below:

3.2.1 Ridge Classifier: A modification of linear regression called ridge regression modifies the loss function to simplify the model. Ridge regression is a technique for evaluating multiple regression data with multicollinearity. Despite the fact that least squares forecasts are unbiased in the context of multicollinearity, their wide variances render them possibly erroneous. In order to reduce the standard errors, ridge regression slightly slants the regression estimates. This method is used when the independent variables are significantly linked. L2 regularization is carried out, and it entails a penalty proportional to the square root of the size of the coefficients

Minimization goal=LS Obj+ *(Sum of square of coefficients) (1)

This change involves the addition of a compensation component equal to the square of the magnitude of the coefficients. Loss function is determined by adding ordinary least square (OLS) and alpha (squared coefficient values). We must choose alpha as the parameter in the loss function shown above. Low alpha values can lead to over-fitting while high alpha values may lead to under-fitting. Scikit Learn's Ridge class is used to create a ridge regression model. To reduce the subsequent cost function:

$$(y - X\beta)^T (y - X\beta) + \lambda\beta^T\beta \qquad (2)$$

λ is a value given by user input (or by a grid search, or whatever). Note that here we use λ, scikit-learn uses α. β is a vector of weights, βi, assigned to each of the features to produce a finished model.

3.2.2 Logistic Regression: One kind of analysis is logistic regression that is used to classify data and to figure out how different independent variables interact. It is a probabilistic classifier and uses a logistic function to model the probability that describes the possible outcome of a single trial. It works when the assumed variable is dual (binary two class- 0 or 1 classification), free from missing values and all predictors are independent of each other. The outcome of logistic regression is determined by taking the event's log odds in (P/1P), where P is the probability of the event. As a result, P is always between 0 and 1. The formula (3) of logistic regression says that to find P, exponential of a+bx is added to one (1) and is branched out with exponential of a+bx. whereas the formula (4) says that e to the power of – (a+bx) is branched out by one (1) only to get P.

$$\mathrm{P} = \frac{\mathrm{e}^{\mathrm{a+bX}}}{1 + \mathrm{e}^{\mathrm{a+bX}}} \tag{3}$$

$$\mathrm{P} = \frac{1}{1 + \mathrm{e}^{-(\mathrm{a+bX})}} \tag{4}$$

3.2.3 Perceptron: The Perceptron is an algorithm for linear classification. This suggests that it learns a decision boundary that splits two classes using a feature space line called a hyperplane. As a result, it works well for situations where the classes can be efficiently divided by a line or linear model, referred to as linearly distinguishable problems. The model's coefficients, or input weights, are trained using the stochastic gradient descent optimization technique. For classification in binary format with two classes, the Perceptron method is a machine learning strategy. It is a member of a group of neural network models, arguably the most fundamental. It is composed of an individual node or neuron that determines the class from a sequence of incoming input. To do this, a bias and the weighted sum of the inputs are calculated (set to 1). The model's activation is the weighted sum of its input as given in equation (5).

$$\text{Activation} = \text{Weights} * \text{Inputs} + \text{Bias} \tag{5}$$

If the activation is larger than 0.0, the model will create 1.0; if it is less than 0.0, it will produce 0.0.
Predict One (1): If Activation > 0.0
Predict Zero (0): If Activation <= 0.0
Prior to using a model, it is best procedure to normalize or standardize the data since model inputs, like those for logistic and linear regression, are multiplied with model coefficient.

3.2.4 Decision Tree: Decision tree is considered amongst the most influential approach for supervised class of machine learning. It is simple to understand and comprehend. It can be used for both categorical and numerical data. The output of the decision tree is expressed as a sequence of rules which are used for classification task. Sometimes, DT learning can produce a complex tree that does not generalize well. DTs can be unbalanced because little dissimilarity in the data might generate completely different tree. The decision tree learning algorithm uses a measure called information gain to build a decision tree. Knowledge improvement is estimated in terms of entropy of the initial set and the split obtained after testing an attribute. The entropy of a sample S is mathematically defined as:

$$E(S) = \sum_{i=1}^{c} - p_i \, log_2 \, p_i \tag{6}$$

3.2.5 Linear Support Vector Classifier (LSVC): A Linear Support Vector Classifier's goal is to categorize or split the data you provide by returning the "best fit" hyperplane. You may then add specific characteristics to the classifier to get the "predicted" class after acquiring the hyperplane. The LSVC uses a linear kernel function to conduct classification and does well with a lot of samples. When compared to the SVC model, the LSVC adds more parameters including the loss function and penalty normalization, which applies "L1" or "L2." Since LSVC is dependent on the kernel linear technique, the kernel method cannot be modified.

3.2.6 K-NN: The supervised learning method serves as the foundation for the K-Nearest Neighbor algorithm, also known as the KNN algorithm. The K-NN algorithm operates under the presumption that similar items exist nearby. Because of this, the K-NN method uses attribute resemblance among additional data points and points in the training set (existing cases) to forecast the value of the target data points. In general, the K-NN approach determines the value of the most recent data point by comparing it to the values in the training dataset. Although the K-NN technique is applicable to both regression and classification issues, it is most frequently used for classification issues.

3.3 Proposed Algorithm

1. Tokenize tweet data
2. Pre-process the dataset to remove stop words, hashtag, and re tweets.
3. Train the classifiers using NLTK toolkit
4. 4.Apply the trained classifiers on test data
5. Final result

Implementation of this work is done by using NLTK (Natural Language Toolkit) [19]. It's a suite of techniques for regression, clustering, classification, association, data pre-processing, and other tasks. We experiment with tf-idf feature scheme with all six classifiers on GST dataset.

4. EXPERIMENTS AND RESULTS

We do a train-test split on the dataframe's X and Y components. GST twitter dataset, as explained in section 3.1, and train test split () method were used to divide our data into train and test sets for each of the six classifiers. A set of data was used to fit the model. The training dataset is what it is called the data set that the model was trained on. The model notices and takes note of this information. Our data must first be divided into features (X) and labels (y) for analysis. The X trains, X test and Y train, Y test components of the data frame are separated. Using the X train and Y train sets, the model is trained and fitted. Using the X test and y test sets, the model is evaluated to check if it accurately predicts the outputs and labels. It is possible to explicitly test the size of the train and test sets. The test sets should be less extensive than the training ones. In our study, 25% of the data were used for testing, and 75% were used for training tests.4.1 Evaluation Process: Accuracy, precision, recall and f-score are the indicators used as the performance metrics for the classifiers. These measures are calculated using confusion matrix (Table-1).

Table 1 : Confusion Matrix for Performance Measure

	P (Predictable)	**N** (Not Predictable)
P (Actual)	**TP**	**FN**
N (Not Actual)	**FP**	**TN**

FP= False Positive, FN= False Negative, TP= True Positive, TN= True Negative
FP is the overall number of incorrectly made positive predictions.
The number FN represents the overall number of incorrect negative predictions.
TP stands for the total number of correctly predicted positive outcomes.
TN denotes the overall number of accurately predicted negative outcomes.
Accuracy (A) is the simplest performance metric. It may be derived from the confusion matrix (Table 1) using the formula (7). It is the ratio of accurately expected observations to all observations.

$$\text{Accuracy (A)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (7)$$

In order to calculate precision, divide the overall number of positive predictions (P) by the proportion of correct positive predictions. Positive predictive value (PPV) is another name for it (PPV). The highest precision is 1.0, while the lowest is 0.0.

$$\text{Precision (P)} = \frac{\text{TP}}{\text{TP+FP}} \quad (8)$$

Recall is calculated for classifier as it indicates the ability of classifier to classify the positive samples, greater the value of recall, classifier is said to predict more positive samples. It is determined as the sum of the true positive samples and the false negative samples divided by the true positive samples. The true positive rate, or recall ®, is another name for it (TPR). The highest sensitivity is 1.0, while the lowest is 0.0.

$$\text{Recall(R)} = \frac{\text{Rate}}{\text{Recall}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (9)$$

The F-score is often referred to as the F-Measure or F1Score. It's a precision and recall harmonic mean.

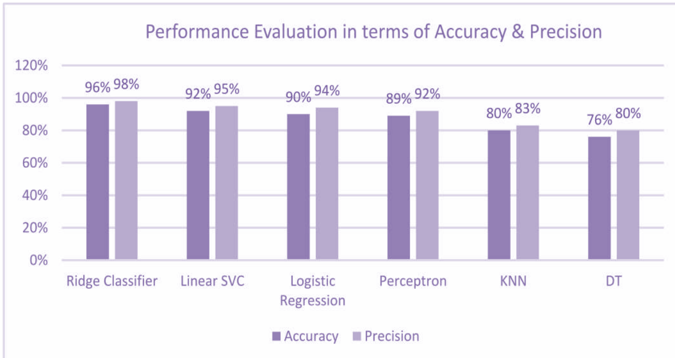
$$(10) \quad \text{F} - \text{Measure(F1)} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.2 Experimental Result: Each classifier was tested using tf-idf. With GST dataset the best result was obtained using tf-idf. The best performing case is reported in Table2. The highest accuracy of 96% was obtained using Ridge classifier with our dataset. Figure-2 shows comparison graph of these classifiers.

Table 2 : Result Evaluation with Classifiers

Classifier	Accuracy	Precision
Ridge Classifier	96%	98%
Linear SVC	92%	95%
Logistic Regression	90%	94%
Perceptron	89%	92%
K-Nearest Neighbor (K-NN)	80%	83%
Decision Tree (DT)	76%	80%

Figure 2. Performance comparison of techniques in terms of accuracy& precision.



5. DISCUSSION

Table3 summarizes the classifiers used in this work for GST sentiment analysis and compare the best performing cases with other best-of-breed analysis approaches existing in literature. According to the evaluation results in Tables 2 and 3, the performance of the classifiers utilized in this study is comparable with other cutting-edge methods presented in [10, 11, 12, 13, 20]. Our best performing case (Ridge Classifier) outperforms the best performing classifiers reported in [10, 11, 12, 13, and 20] in terms of accuracy. However, the GST dataset used in all these works is different. Hence, in [5], SVM with model-1 and model-2 have been evaluated on two datasets. In order to further enhance accuracy, we have used tf-idf with six supervised classifiers for linear classification on GST dataset. We have obtained a maximum accuracy of 96% using Ridge classifier and observed 98% precision. The improvements are quite significant with RC, LSVC, P, K-NN except LG and DT.

Table 3 : Comparison of Classification Approaches

S. No.	Author	Methods	Features	Dataset	Accuracy	Precision	Re-call
1.	Our	**RC*,** LR, P, K-NN, LSVC, DT.	**Term frequency-inverse document frequency***	TwitterAPI(GST) *	96%	98.02%	94.62%
2.	Alec Go et. al. [20]	NB, **ME***, SVM	Unigram,bigram,**unigram+bigram***, unigram+POS	Twitter API (product/brand)	83.0%	---------	------
3.	Das &Kolya [13]	NB	Particular feature from class of features	Twitter API(GST)	Calculate Range, TF, CF, Zipf Score		
4.	Ganguly& Roy [10]	Polarity Method	Structural features	Twitter API(GST)	38% +ve, 62% -ve	...	...
5.	Gautam & Yadav [11]	**NB*,** ME, SVM	Unigram feature technique, **WordNet***	Product Review	89.9% (WordNet)	88.3%	44.3%
6.	Tomer et. al. [12]	SVM	Linguistic based feature	IMDB+(GST) Twitter*, IMDB	73.28%	73.09%	73.67%

6. CONCLUSION

In governments new taxation scheme that is based on the theme of all for one and one for all. Goods and Services Tax is an integrated tax that applies to all goods and services. It combines federal and state taxes into a single tax that is collected. However, its introduction has been quite debatable. To understand the public opinion on the new taxation system is crucial for shaping future reforms. This work attempts to identify this opinion by analyzing polarity of tweets made about GST. We investigated six classifiers -Ridge Classifier, Linear Support Vector Classifier, Logistic Regression, Perceptron, K-Nearest Neighbor and Decision Tree, using tf-idf feature and examined the performance in terms of accuracy. The observed accuracy is 96%, 92%, 90%, 89%, 80% and 76% respectively. RC achieved highest accuracy among all the classifiers. All performs better on identical dataset.

REFERENCES

[1] *Bhadane C, Dalal H, Doshi H (2015), "Sentiment analysis: Measuring opinions", International Conference on Advanced Computing Technologies and Applications (ICACTA-2015), Procedia Computer Science 45 (2015) 808 – 814 ELSEVIER.*

[2] *Singh S. K. and Sachan M. K. (2019), "SentiVerb system: classification of social media text using sentiment analysis", Multimedia Tools and Applications https://doi.org/10.1007/s11042-019-07995- India (2019)78:32109 –32136.*

[3] *Siddiqui, T.J (2011), "Utilizing Sentiments in Online Contextual Advertising", In Online Multimedia Advertising: Techniques and Technologies, IGI Global, pp.32-37.*

[4] *Pak A., Paroubek P., Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S., Rosner M., Tapias D., "Twitter as a corpus for sentiment analysis and opinion mining" (2010), Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta (May 2010), pp. 19–21.*

[5] *5. Pang B., Lee L. and Vaithyanathan S. (2002), "Thumbs up? Sentiment Classification Using Machine Learning Techniques", In EMNLP '02: Empirical methods in natural language processing, arXiv preprint cs/0205070. 2002 May 28, pages 79–86. ACL, 2002.*

[6] *Liu B. (2012), "Sentiment Analysis and Opinion Mining", Human Language Technologies (Synthesis Lectures), 5 (1), 1–167(2012).*

[7] *http://gstcouncil.gov.in/sites/default/files/01012019-GST Concept-and-Status.pdf*

[8] *CnossenSijbren; Preparing the way for a modern GST in India, Int Tax Public Finance DOI 10.1007/s10797-013 9281-0, © Springer Science+Business Media New York 2013.*

[9] *Madan A., Arora R., Roy N. R. (2017), "Sentiment Analysis of Indians on GST", International Conference on Recent Developments in Science, Engineering and Technology REDSET 2017: Data Science and Analytics pp 568-575.*

[10] *Ganguly M. and Roy S. (2018), "A Social Network Analysis of Opinions on GST in India within Twitter", ICDCN '18, January 4–7, 2018, Varanasi, India, c 2018 Association for Computing Machinery, ACM ISBN 978-1-4503-6397 6/18/01https://doi.org/10.1145/3170521.3170539.*

[11] *Gautam G. and Yadav D. (2014), "Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis", In Proceedings of Seventh International Conference on Contemporary Computing (IC3), IEEE, pp. 437-442, 2014.*

[12] *Tomar N., Manav R., Srivastava R., Ahuja B. (2018), "Opinion Mining of GST Implementation using Supervised Machine Learning", International Journal of Computer Applications Volume 180 – No.49, June 2018, (0975 – 8887).*

[13] *Das S. and Kolya A. K. (2017), "Sense GST: Text Mining & Sentiment Analysis of GST Tweets by Naive Bayes Algorithm", Third International Conference on Research in Computational Intelligence and Communication Network (ICRCICN) 2017 Nov 3, IEEE, pp. 239-244.*

[14] *Chaudhary J. R. and Paulose J. (2019), "Opinion mining on newspaper headlines using SVM and NLP", International Journal of Electrical and Computer Engineering (IJECE) June 2019, Vol. 9, No. 3, pp. 2152-2163.*

[15] *Barnaghi P., Ghaffari P., Breslin J.G. (2016), "Opinion Mining and Sentiment Polarity on Twitter and Correlation Between Events and Sentiment", Second International Conference on Big Data Computing Service and Applications, (BigDataService) 2016 Mar 29, IEEE, pp. 52 57. DOI 10.1109/BigDataService.2016.36.*

[16] *Turney P. D.; Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), arXiv preprint cs/0212032.2002 Dec11, pp. 417-424.*

[17] *Hall et al. (2009), "The WEKA Data Mining Software: An Update, SIGKDD Explorations newsletter, Volume 11, Issue 1, pp. 10-18.*

[18] *Devika, M. D., Sunitha C, Amal G. (2016), "Sentiment Analysis: A Comparative Study on Different Approaches", Procedia Computer Science 87 2016, pp. 44 – 49, doi: 10.1016/j.procs.2016.05.124.*

[19] *Bird S., Klein E., and Loper E. (2009)," Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit", O'Reilly Media, Inc.",2009 un-12, (https://www.nltk.org/book/).*

[20] *Go, A., Bhayani, R. and Huang, L. (2009), "Twitter sentiment classification using distant supervision" CS224N Project Report, Stanford, 1(2009), p.12.*