2021 International Conference on Emerging Smart Computing and Informatics (ESCI)
AISSMS Institute of Information Technology, Pune, India. Mar 5-7, 2021

Comparative Study of Machine Learning Algorithms for Twitter Sentiment Analysis

Yash Indulkar
Information Technology
Thakur College of Science & Commerce
Mumbai, India
yashindulkar31@gmail.com

Abhijit Patil
Information Technology
Thakur College of Science & Commerce
Mumbai, India
abhijitpatil976@gmail.com

2021 International Conference on Emerging Smart Computing and Informatics (ESCI) | 978-1-7281-8519-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ESCI50559.2021.9396925

***Abstract*—Sentiment Analysis is important to understand various aspects of human emotions through different modes, the modes can be, either by understanding the text or analyzing it for obtaining the desired outputs. The three algorithms considered for sentiment analysis are Logistic Regression, Multinomial Naïve Bayes & Random Forest on the Uber & Ola datasets. The number of tweets extracted from Twitter is 3000. These tweets are cleaned & tokenized using python. The main factor of this paper is Google word2Vec, as the tokenized tweets are transformed with vocabulary from Google Word2Vec. Using this immense dataset of words, helped tokenized words to create a better vocabulary and understanding. Finally, the accuracy and the Mean Cross-Validation Accuracy (MCVA) was generated for all the three algorithms which are used to check if it was giving proper results to the trained data. Visualization was created for understanding the accuracy of three algorithms, which in turn helped to select the most accurate algorithm among others. The programming language used in this for pre-processing & analysis is Python.**

Keywords—Twitter, Uber, Ola, Machine learning, Python, Logistic Regression, Multinomial Naïve Bayes, Random Forest, Google Word2Vec.

I. INTRODUCTION

Machine learning is a branch of artificial intelligence that is concerned with the construction of models from data & presenting it in an understandable format. It is an application of artificial intelligence that deals with training the machine based on algorithms making it self-learn a pattern of solving problems. The machine then gets trained with a new input data given to it and trains the new data to the already trained data. Recently, the field of machine learning has seen a rise in the popularity of probabilistic and statistical models. This paper focuses on showing the effects of machine learning algorithms, Logistic Regression, Multinomial Naïve Bayes, and Random Forest. We gather 3000 tweets from Uber & Ola and we first clean it then the tweets are applied to these machine learning algorithms and data from the tweets is trained. The output is given in the form of positive or negative sentiments and accuracies. The algorithms look for the words resembling or close to the sentiments & then compare it as well. It then categorizes it & does analysis in the form of graphs, diagrams. The data on which we have applied the algorithm is training data. For particular datasets, in our case for the first 500 tweets, we train the data. Then we increase the number of tweets. We then next train 1000 tweets & increment it to 3000 tweets. The first 500 trained tweets then are used as trained data, the new datasets or tweets are the testing data, and this testing or training data is compared with the tested or trained data. The approach we have used here is to use these supervised machine algorithms to understand how accurately can we relate the data and fetch accuracies. The experimental results have Word2Vec representation of words closely similar to each other. These words are given a vector value and based on the distance between the words the data is trained. Logistic Regression uses a probabilistic approach to count the word from an observation. Multinomial Naïve Bayes is based on the Bayes theorem. Multinomial Naïve Bayes finds a frequency of a word in a particular text by using multinomial distribution. The third algorithm we have used is the Random Forest. Random forest works on the principle of the decision tree. Using a decision tree, we can find the problem area.

II. LITERATURE REVIEW

Machine Learning is more like an application of AI (Artificial Intelligence) which can learn automatically and understand. It is used widely today to enable machines to perform various tasks voluntarily. Machine learning has applications in image processing, medical diagnosis, prediction, classification, etc. There are various machine learning algorithms in machine learning that give the desired outputs as required. Based on the results we can train our machine to deal with likewise data & generate outputs. This data is understood in two forms. Training data and testing data. Training data is the data that is used by the algorithm to learn. Testing data is the data on which we want to perform the analysis. The algorithms can be supervised or unsupervised. The reason we choose these three algorithms was that we saw the potential these machine learning algorithms could have. Today data can be efficiently used, understood and conclusions can be made. The work on previous topics motivated us to use these algorithms for sentiment analysis of Uber & Ola. What motivated us to select Uber & Ola as datasets? Uber has around 7 million users (as of Aug 18) in India and the Ola cab service has 23.9 million users (as of Nov 2019). This makes us understand the vastness of data these companies have. With so much data around the cab services, these companies are on the data peaks. This is what motivated us to understand what sentiments do people have about these two cab services. Understanding how much of an impact machine learning has created on topics of different domains we studied it and laid out the foundation of our work. Starting with the first paper by Jiang Su, Jelber Sayyad-Shirabad, Stan Matwin. This paper focused on how Multinomial Naïve Bayes was effectively used to find large-scale text classification [3]. The paper named "Microblog Sentiment Analysis Bases on Cross-media Bag-of-words model" by "Min Wag", "Donglin Cao", "Shaozi Li, Rongrong Ji" in the year July 2014. This paper focused on the use of a bag of word models giving us sentiment analysis on cross-media [4]. This motivated us for

displaying outputs in the Word2Vec approach to find similar models. For understanding how Multinomial Naïve Bayes creates an impact for our data sets, we referred to the paper "Multinomial Naïve Bayes for Text Categorization Revisited" by "Ashraf M. Kibriya", "Eibe Frank", "Bernhard Pfahringer", "Geoffrey Holmes" throws light upon how Multinomial Naïve Bayes effects text categorization [3]. The paper in 2018 by the famous author regarding the Google Word2Vec, based on a text by Rudkowsky, Elena shared the use of Word2vec [8]. An article on Logistic Regression by Stanford in the "Speech & Language Processing" brings to notice the use of this machine learning. Other research paper was, "Twitter Sentimental Analysis & Algorithm Comparison for Uber & Ola Using 'R", this paper utilized Machine Learning Algorithm such as SVM & Naïve Bayes, based on the same datasets which were published by "J Anthal" et al, it showed how the Uber & Ola datasets were performing with 2 different classification algorithm and what accuracy was getting generated, this was the stepping stone for applying machine learning algorithms on such datasets, the methodology was simplest and future work was having greater scope [14]. Again, the research paper which was named "Sentiment Analysis of Uber & Ola using Deep Learning,", this particular paper that was published by "Y Indulkar & A Patil", showed how the deep learning algorithms such as Non-Deep Neural Networks & Convolutional Neural Networks worked on such datasets, this research-based on deep learning for further another future scope that was achieved [5]. These were the things that motivated us to do sentiment analysis using these three ML algorithms. Hence, we selected Logistic Regression, Multinomial Naïve Bayes, and Random Forest algorithms to see what results do we get on our data sets.

III. METHODOLOGY

This section of the paper deals with different methods used to obtain the desired outputs which are explained in the experimental results. The methodology is divided into 3 sub-sections that are:

A. Logistic Regression

Logistic Regression is a model that can be used for both the regression & classification problems, in this research it has been used for the classification of tweets based on Positive (1) & Negative (0) with the help of a Logistic curve that is shown in Fig 1. The regression model can be divided into Linear & Logistic. The Logistic Regression counts the probability of a word occurring for randomly selected observations versus the probability that the word does not occur. This curve for Logistic Regression is like a Sigmoid.

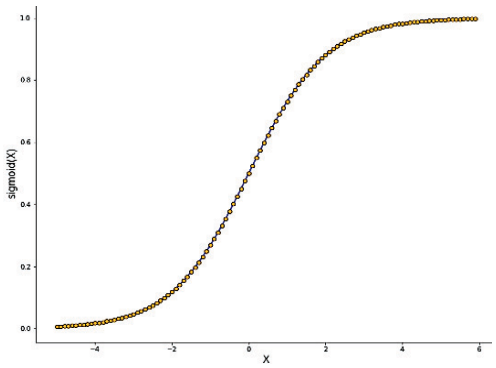

Fig. 1. Logistic Curve for Regression Analysis

B. Multinomial Naïve Bayes

Multinomial Naïve Bayes is the part of the probability algorithm based on the Naïve Bayes theorem. Bayes theorem is one of the probabilistic algorithms developed by the Reverend Bayes. It's a classification algorithm based on Bayes Theorem. Multinomial Naïve Bayes uses the multinomial distribution to find the frequency of the particular word occurring in the given text. It can be calculated with the formula (1) below, which shows the probability standards followed by Multinomial Naïve Bayes.

$$\Pr(j) = \log \pi_j + \sum_{i=1}^{|V|} \log(1 + f_i) \log(\Pr(i|j)) \quad (1)$$

C. Random Forest or Random Decision Forest

Random Forest or Random Decision Forest is an ensemble learning method that can be used for the Classification or Regression based on the requirements. Decision Tree consists of a root, splitting, decision nodes & leaf which together form a decision tree structure that can be used for identifying the problem area. Combining multiple decision trees, the random forest algorithm works, which gives better results due to multiple trees which helps in calculating the mean results from the overall trees. It can be well understood with Fig 2.

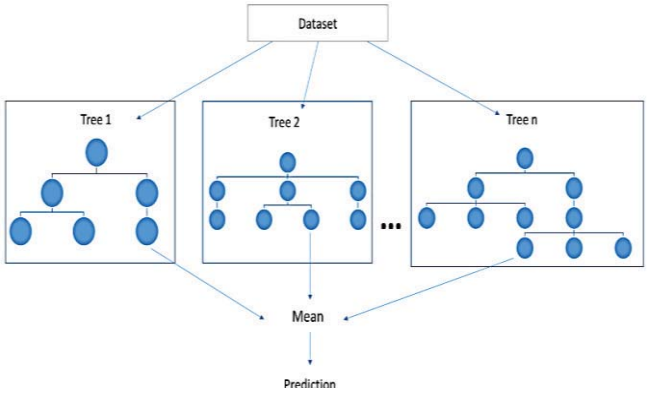

Fig. 2. Decision Tree Representation from Root

IV. EXPERIMENTAL RESULTS

The experimental section of this research paper consists of various tests and results obtained by performing iterations. This paper consists of three different classification algorithms for the Twitter datasets, which are Uber & Ola with 3000 Tweets, which were extracted using Twitter API. The Sentimental Analysis of these tweets generated Positive & Negative sentiments which in turn were converted to a binary variable. The Twitter Sentimental Analysis for Binary Variable are categorized as below:

1.) Positive = 1
2.) Negative = 0

The vocabulary was generated with the help of Google Word2Vec. The Google Word2Vec is a two-layer Neural Network that processes the tweet by vectorizing the words. The input given to the Word2Vec is a text corpus and the output generated is the vectors. The purpose of this is to group the vectors with similar words. In this paper, the words distinguished by the distance can be understood with the help of Fig 3.

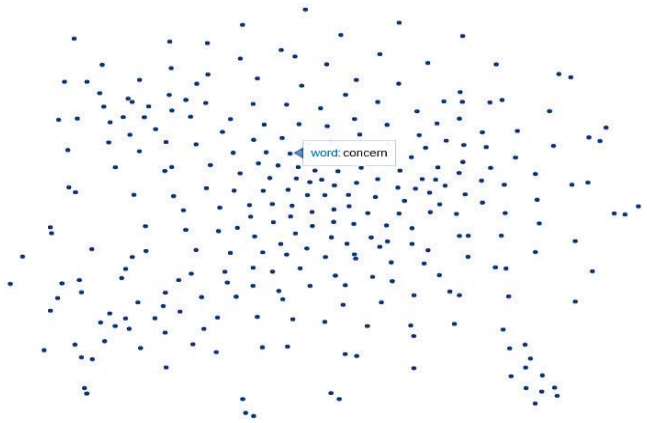

Fig. 3. Google Word2Vec Map Plots

It can be seen that the word: Concern has similar words in the nearest vector. The distance is directly proportional to the similarity of words. Table I shows the similarity obtained for the datasets with the distance from the words like 'message' & 'email'.

TABLE I. GOOGLE WORD2VEC DISTANCE

Sr. No of Words	Words	Distance Similarity
1	Around	0.37331557273864745
2	Via	0.3613654673099514
3	Ola	0.34762924909591671
4	Help	0.3421033024787905
5	Service	0.32505300641059878

The data was then divided into Training & Testing fragments, the train data consist of 80 % overall data & the test data consists of the remaining 20 % overall data. It was then Vectorized with TfidfVectorizer, using the feature extraction to fit & transform on the training dataset & to only transform on the testing datasets. This created the features that were important for understanding the text properly. The number of features varied from the number of data used. That can be visualized for each data ranging from 500 to 3000 that has been used in this research. The number of features of each data can be understood in Fig 4. It can be observed that the line is linear with both the Uber & Ola datasets, as the datasets are increasing the number of features also increases linearly. The plot for features is based on the data and the number of features.

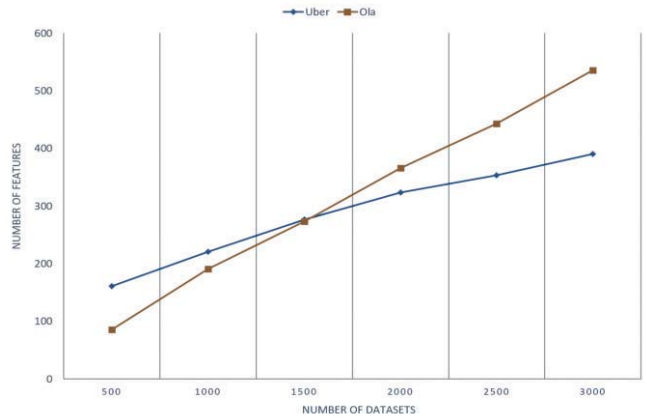

Fig. 4. Vocabulary Size for Datasets based on Features

It can be observed that the number of features for both Uber & Ola gets on increasing with the growth in the data, the highest number of features for Uber data is 391 for 3000 Tweets & similarly for the Ola data is 536. The Three Classification algorithms used are Logistic Regression, Multinomial Naïve Bayes & Random Forest for classifying the sentiments that are in binary. To understand the test score accuracy on testing data the Mean Cross-Validation Accuracy (MCVA) was calculated with the Cross-Validation (CV) of 5. The Mean Cross Validation Accuracy has the formula that is shown in Fig 5 below.

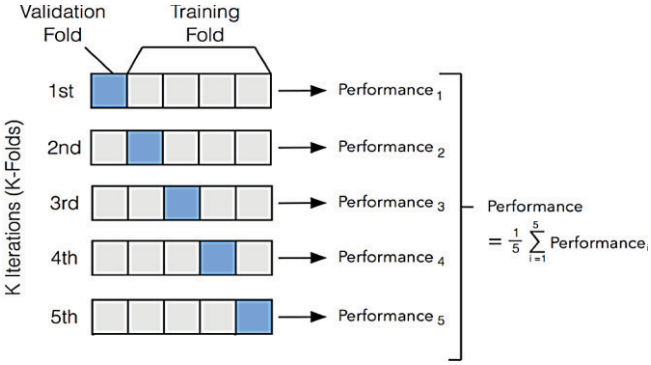

Fig. 5. Mean Cross-Validation for 5 Folds

That generated different accuracy on all three algorithms for different datasets in Uber can be seen in Table II below.

TABLE II. MEAN CROSS VALIDATION ACCURACY FOR UBER DATASETS

Datasets	Logistic Regression	Multinomial Naïve Bayes	Random Forest
500	88.0 %	88.0 %	91.0 %
1000	90.0 %	85.0 %	93.0 %
1500	90.0 %	85.0 %	94.0 %
2000	95.0 %	90. 0 %	95.0 %
2500	92.0 %	86.0 %	95.0 %
3000	93.0 %	87.80 %	95.0 %

The visual representation of the Mean Cross-Validation for Uber data can be observed from the below Fig 6, it can be seen that the highest MCVA for the uber data was achieved by Random Forest (In Green), the percentage was 95 % which was constant throughout the last three datasets (2000, 2500, 3000) as the iteration increased.

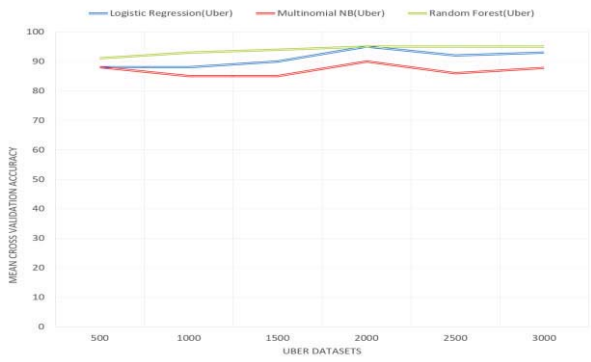

Fig. 6. Mean Cross Validation Accuracy Visualization for Uber

Similarly, the generated different accuracy on all three algorithms for different datasets in Ola can be seen in Table III below.

TABLE III. MEAN CROSS VALIDATION ACCURACY FOR OLA DATASETS

Datasets	Logistic Regression	Multinomial Naïve Bayes	Random Forest
500	77.0 %	65.0 %	70.0 %
1000	74.0 %	70.0 %	70.0 %
1500	74.0 %	68.0 %	75.0 %
2000	75.0 %	75.0 %	85.0 %
2500	77.0 %	75.0 %	80.0 %
3000	79.0 %	75.80 %	85.0 %

Similarly, the visual representation of the Mean Cross-Validation for Ola data can be observed from the below Fig 7, it can be observed that the highest MCVA for the ola datasets was achieved by Random Forest (In Green) followed by the Logistic Regression, the percentage was 85 % which was not constant throughout the last three datasets (2000, 2500, 3000) as the iteration increased.

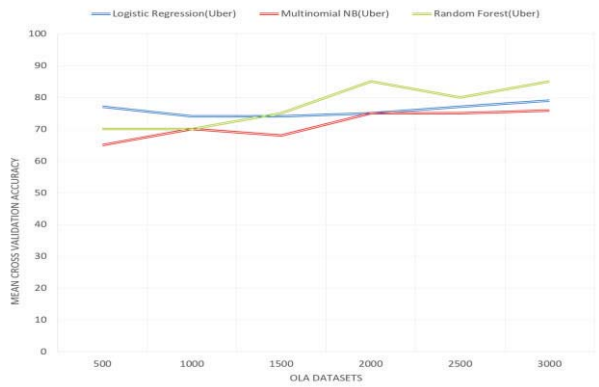

Fig. 7. Mean Cross Validation Accuracy Visualization for Ola

The Mean Cross-Validation Accuracy gives the mean value for the training data that should match the testing data in terms to get the proper results. The model generated different MCVA for different algorithms on different datasets that were ranging from 500-3000. To understand what the testing accuracy obtained can be seen from Table IV & Table V.

TABLE IV. TESTING ACCURACY OBTAINED FOR UBER DATASETS

Total No of Datasets	Logistic Regression	Multinomial Naïve Bayes	Random Forest
500	88.0 %	88.0 %	91.0 %
1000	89.5 %	84.5 %	93.0 %
1500	91.0 %	84.0 %	94.3 %
2000	94.3 %	89.5 %	93.5 %
2500	92.0 %	86.0 %	95.6 %
3000	92.7 %	87.8 %	96.3 %

TABLE V. TESTING ACCURACY OBTAINED FOR OLA DATASETS

Total No of Datasets	Logistic Regression	Multinomial Naïve Bayes	Random Forest
500	63.0 %	64.0 %	68.0 %
1000	70.5 %	70.0 %	70.5 %
1500	69.0 %	67.3 %	72.3 %
2000	74.0 %	72.5 %	74.0 %
2500	79.2 %	74.6 %	79.0 %
3000	81.9 %	75.5 %	83.4 %

The Mean Cross-Validation Accuracy can be compared with the above table to understand where the algorithm lacks the testing data if the percentage varies with a large difference. It can be seen that in the case of the Ola dataset the accuracy varied differently as compared to the Uber datasets for all three algorithms. Accuracy generated for the three algorithms on 3000 datasets for Uber & Ola can be visualized with the help of Fig 8 & Fig 7 respectively.

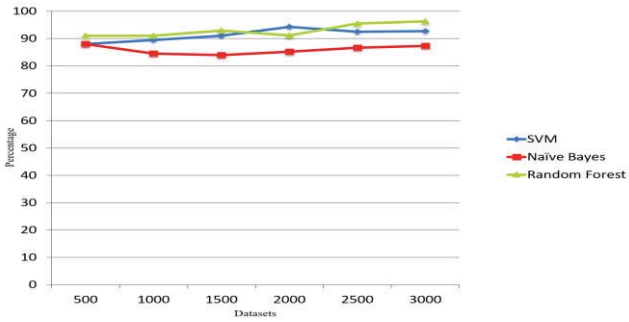

Fig. 8. Accuracy Comparison for Uber Datasets

The above figure shows the accuracy comparison for different algorithms used in this research on the Uber datasets, it can be seen that the most accuracy was achieved for the Random Forest algorithm, in the end, followed by SVM and Naïve Bayes. The data were gradually increased with the range of 500 tweets per training data. The overall accuracy was ranging between 80 % to 97 % respectively. This shows the interpretability of how the algorithm performed for the particular datasets over other algorithms that were used.

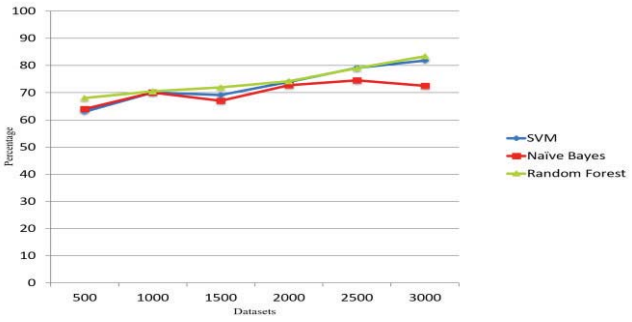

Fig. 9. Accuracy Comparison for Ola Datasets

It can be seen that the highest accuracy from the three algorithms used for Uber & Ola datasets is given by Random Forest in both the cases with

1. Random Forest (Uber) - 96.3 %
2. Random Forest (Ola) - 83.4 %

V. CONCLUSION

The study aims to understand the various classification algorithm with different datasets such as Uber & Ola. The regression algorithm used for binary classification is Logistic Regression & the classification algorithms used are Multinomial Naïve Bayes & Random Forest. It can be observed that from the three algorithms used, the best accuracy was generated from Random Forest for both the respective datasets. The Random Forest gave better accuracy because it created multiple decision trees and then calculated a mean value from all the decision trees. For the Uber dataset, the best accuracy was 96.3 % and for the Ola datasets, it was 83.4 %. It can be seen that in the case of Ola the accuracy was quite low within all the intervals as compared to Uber because the dataset was not that precise and it contained many words that were not in the standard language which made it difficult to be understood by the machine, all these algorithms generated better accuracy because the vocabulary was trained with the help of Google Word2Vec which made the inter-relational words more significant.

VI. FUTURE WORK

We understood from this research about the datasets that, data correctness is really important factor in text classification. The more accurate or precise the data sets the better the outputs for classification. For example, in this paper, the data set Uber cab service gave more accurate results because the data related to uber (tweets) was clean and understandable as compared to that of Ola. So whichever data set we want to do sentiment analysis on should have accurate or correct data. We can improve data sets and help algorithms understand for which language the data (tweet) has to be trained and accordingly understand the sentiments more accurately & correctly. We can further have an area or location-based sentiment analysis for a better understanding of customer reviews according to the area they use the services in which can be extracted by co-ordinates. Random Forest proved to be the best algorithm in our case with the highest accuracy of 96.3% for Uber data sets and 83.4% for Ola data sets. Data dependency for text classification depends on what data is presented for which analysis has to be done.

REFERENCE

[1] Deho, B. Oscar, et al. "Sentiment analysis with word embedding." 2018 IEEE 7th International Conference on Adaptive Science & Technology (ICAST). IEEE, 2018.

[2] Su, Jiang, Jelber S. Shirab, and Stan Matwin. "Large scale text classification using semi-supervised multinomial naive bayes." *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011.

[3] Kibriya, Ashraf M., et al. "Multinomial naive bayes for text categorization revisited." *Australasian Joint Conference on Artificial Intelligence*. Springer, Berlin, Heidelberg, 2004.

[4] Wang, Min, et al. "Microblog sentiment analysis based on cross-media bag-of-words model." *Proceedings of international conference on internet multimedia computing and service*. 2014.

[5] Y. Indulkar and A. Patil, "Sentiment Analysis of Uber & Ola using Deep Learning," 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2020, pp. 21-27, doi: 10.1109/ICOSEC49089.2020.9215429.

[6] Mäntylä, Mika V., Daniel Graziotin, and Miikka Kuutila. "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers." *Computer Science Review* 27 (2018): 16-32.

[7] Da Silva, Nadia FF, Eduardo R. Hruschka, and Estevam R. Hruschka Jr. "Tweet sentiment analysis with classifier ensembles." *Decision Support Systems* 66 (2014): 170-179.

[8] Rudkowsky, Elena, et al. "More than bags of words: Sentiment analysis with word embeddings." *Communication Methods and Measures* 12.2-3 (2018): 140-157.

[9] Tang, Duyu, et al. "Sentiment embeddings with applications to sentiment analysis." *IEEE transactions on knowledge and data Engineering* 28.2 (2015): 496-509.

[10] Rezaeinia, Seyed Mahdi, et al. "Sentiment analysis based on improved pre-trained word embeddings." *Expert Systems with Applications* 117 (2019): 139-147.

[11] G. Murray, E. Hoque, G. Carenini, Chapter 11 - Opinion Summarization and Visualization, Editor(s): Federico Alberto Pozzi, Elisabetta Fersini, Enza Messina, Bing Liu, Sentiment Analysis in Social Networks, Morgan Kaufmann, 2017, Pages 171-187, ISBN 9780128044124.

[12] Giatsoglou, Maria, et al. "Sentiment analysis leveraging emotions and word embeddings." *Expert Systems with Applications* 69 (2017): 214-224.

[13] Sarlan, Aliza, Chayanit Nadam, and Shuib Basri. "Twitter sentiment analysis." *Proceedings of the 6th International conference on Information Technology and Multimedia*. IEEE, 2014.

[14] J. Anthal, Y. Indulkar, A. Upadhyay, A. Patil, Twitter Sentimental Analysis & Algorithm Comparison for Uber & Ola Using 'R', Vol. 13 No. 1s (2020): Vol. 13 No.1s (2020) Special Issue.

[15] Pandarachalil, Rafeeque, Selvaraju Sendhilkumar, and G. S. Mahalakshmi. "Twitter sentiment analysis for large-scale data: an unsupervised approach." *Cognitive computation* 7.2 (2015): 254-262.