Twitter Sentimental Analysis & Algorithm Comparison for Uber & Ola Using 'R'

Jyotsna Anthal[1], Anand Upadhyay[2], Yash Indulkar[3], Abhijit Patil[4]

[1,2,3,4] Thakur College of Commerce & Science, University of Mumbai, India

Abstract

Twitter is a social networking site where a large number of users are actively present. Data with hashtags are popular widely on twitter, hence twitter has large amounts of datasets where user tweet their reviews. These sentiments are used to understand what opinion do people have about a product or service through their tweets. The datasets used in this system for extracting tweets are 'UBER & OLA'. We understand reviews of people towards different products or services which in turn gives business insights as to what changes can be done or incorporated. It also helps us in the analysis of market trends and monetization. In this paper, we propose two models for sentiment analysis based on Naïve Bayes and Support Vector Machine (SVM). Its purpose is to analyze sentiments more effectively. This system uses R- statistical programming language to generate outputs. Further, this paper represents the outputs in Word Cloud. The two classifier algorithms are machine learning algorithms in which we compare their overall accuracy, precision and recall values.

Keywords: *Twitter, Sentimental Analysis,Uber, Ola, Algorithm comparison, SVM, Naive Bayes, Crowd Sourcing, Social Media, R-Programming Language.*

1. Introduction

A language is a powerful tool which help in expressing emotions. Sentiment analysis is text mining which helps a business to understand what social sentiment people have about their brand or product. It uses natural language processing and data mining techniques to the solve real world problems[2]. Besides getting insights about a brand through user reviews businesses could be improved. With those insights a brand can decide how successful is the new product launched, how customers react to the product or service. Are they satisfied or they aren't?[3]. The tweets are basically the reviews of people and are bifurcated into two sentiments positive and negative in this paper. Since the user's tweet in languages which they are comfortable most of the tweets have texts which are difficult to clean. The datasets which this paper is using are 'UBER' & 'OLA'. R-programming language is used in this project. R is a statistical programming language used for computing and data analysis. The reason for selecting this programming language is that it gives better results for analysing and understanding the data precisely as it contains different types of packages for example e1071[6]. This paper uses Machine Learning algorithm techniques which are "SVM" (Support Vector Machine) & "Naïve Bayes". These two are classification algorithms that classify the data into different categories and are a part of supervised machine learning. The purpose of selecting these algorithms are, they give better results for text classification [9].

2. Literature Review

The advent of the internet has helped in the wide share of information. Today information is available on various social media platforms. People express their reviews, suggestions on such social media platforms. These reviews can be studied for analysis of market trends. Twitter is one such platform that was formed in the year 2006 by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams. Since then Twitter has largely grown to a community of 300 million active users & 145 million daily active users[1]. This gives us an idea of why a large number of businesses have started paying close attention to data collection from twitter. Having a vast variety of users from different social interests & domains adds to the vastness of the community. With this vastness & available technology, researchers started twitter sentiment analysis from gathered data. In a research paper "Tapping into the Power of Text Mining" in 2005 by Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang explained the importance of how we can use text to extract useful information to establish a relationship between the words[5]. This helps us to understand the power of text & how can we use data from texts to understand a relationship. With this researcher started extracting data from twitter to understand how the largely available data can be put to use & generate an opinion from their tweets. The previous work from Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining" in the year 2010 helped to further throw the light of how can twitter sentiments help in generating an opinion[6]. Further advancement in this field led to the use of Machine Learning algorithms & how can they be applied on twitter datasets. In our paper, the machine learning algorithms used in the proposed method are “SVM” (Support Vector Machine) & “Naïve Bayes” both fall in the category of supervised machine learning algorithms. Now both these algorithms help in text classification which has helped us in classifying the texts from our datasets “Uber” & “Ola”. We have classified it in categories in our method as positive & negative. This paper has an algorithm applied on datasets “Uber” & “Ola” on the extracted twitter sentiments. After algorithm application, we have outputs which generate Word cloud, algorithm comparison. What motivated to select the datasets “Uber & Ola” & the algorithms will be explained further in the next block.

3. Motivation

Since the use of twitter sentiment analysis has widely been showcased in other domains of datasets like movie review system, disease prediction, etc[1,2]. We felt the domain of cab services can be largely benefited if twitter sentimental analysis is done. Uber has about 110 million users and fulfils about 17 million rides per day(as of May’19) in India we have about 5 million users(as of August’17) & Ola has about 150 million users with about 2 million rides each day with about 23.9 (as of November’19) million users in India hence it adds to the uniqueness of the datasets. With this much availability of data, these two companies are sitting on the peaks of data about users. With these large numbers we felt the curiosity to understand people’s reviews of the cab services from twitter & how can sentiment analysis helps to understand it better for further improvements. The algorithms we choose were SVM & Naïve Bayes because SVM’s can be used in multi-dimensional data sets where data points are referred to as vectors. It is more often used in regression and classification problems. Naive Bayes follows a probabilistic approach to solve problems considering different factors[8]. More about the algorithms & process is explained below in Methodology and Experimental Results.

4. Methodology

A Diagrammatic representation of the sentimental analysis is explained below.

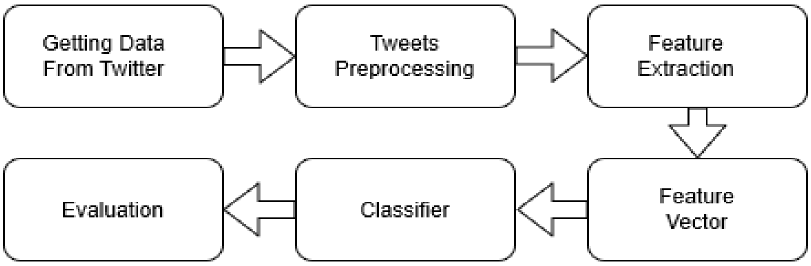


Figure 1. Sentimental Analysis Flowchart

The above Figure 1 consists of various phases in the form of flowchart starting from the Extraction of tweets to the Visualization of tweets. The first phase is “Getting Data From Twitter” that is connection of Twitter API to extract tweets. The second phase is “Tweets Preprocessing” that is after extraction the tweets are in unstructured format that needs to be taken care of with the help of cleanning, stemming, transformation. The third phase is “Feature Extraction” which is important as it decides which feature to be selected for the process. The “Feature Vector” is done for multiple features clubed together for classification. “Classifier” is the phase where different classification algorithms are applied for the final visualization. The “Evaluation” is the last phase where the output extracted can be evaluated for knowledge purpose.Below now the two algorithms are explained for this paper.

The Algorithm considered for classification purpose is SVM & Naïve Bayes

4.1 SVM (Support Vector Machine) Classification Algorithm

The Support Vector Machine can be described as binary classifier. It attempts to find a hyperplane that can separate two class of data by the largest margin.There are 2 types of separation in SVM the one is Linear & the other is Non-Linear [4]. The Non-Linear classification can be done with the help of kernel trick. Kernel plays an important role in separation of data as in nonlinear margin cannot be drawn in 2-D it has to be lifted in higher dimension where the data can be separated that is 3-D plane. Take Figure 2 as it can be observed that two different classes indicating circle & square are used, SVM creates a hyperplane that divides the two classes with the maximum margins.

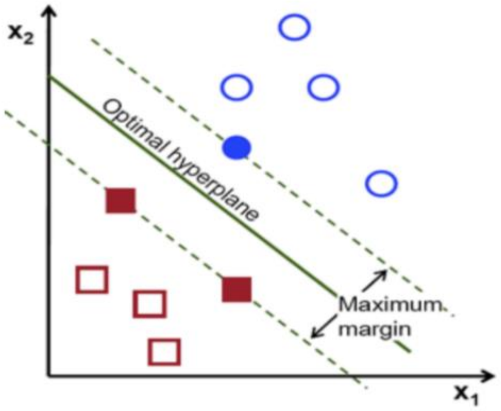
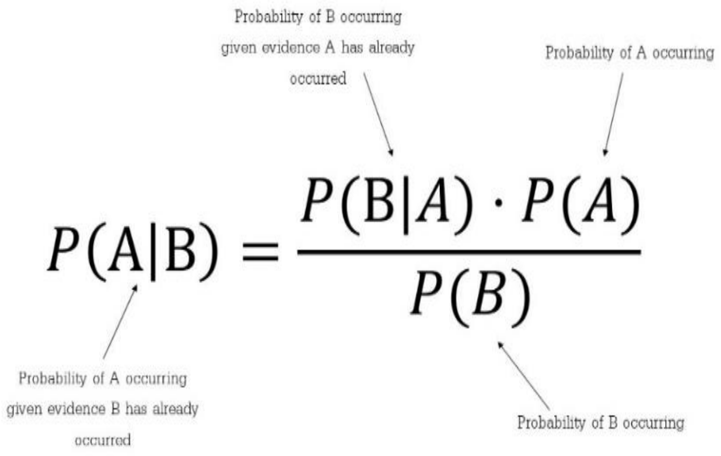


Figure 2. Operation of SVM Algorithm

4.2 Naïve Bayes Classification Algorithm

Naïve Bayes is also a classification algorithm that is based on the principle of Bayes Theorem. Naïve Bayes is not a single algorithm but a collection of algorithms that gives the probability of event occurring. The principle that is followed by this algorithm is that every pair of features that has been classified is independent of each other. The probability of the features is considered with the probability of individual feature occurring divided by the probability of the remaining feature.This states the Bayes' Theorem on which Naïve Bayes' is made. As the features are considered independent, the algorithm will give individual results of each variables to perform differently from other algorithms. Take Figure 3 as it can be observed that conditional probability of B that A has already occurred is multiplied with probability of A divided by probability of B.



$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{B}|A) \cdot P(A)}{P(B)}$$

Probability of B occurring given evidence A has already occurred
Probability of A occurring
Probability of A occurring given evidence B has already occurred
Probability of B occurring

Figure 3. Bayes Theorem Formula

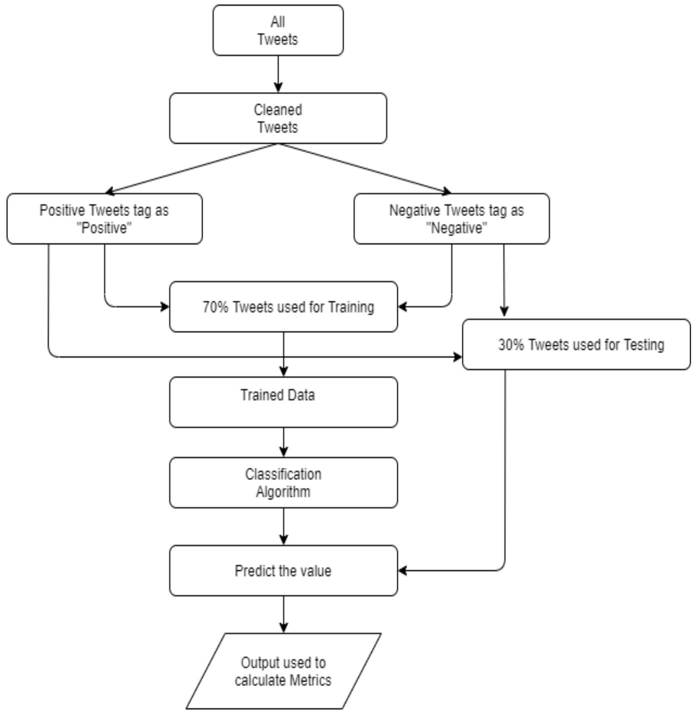

All Tweets
Cleaned Tweets
Positive Tweets tag as "Positive"
Negative Tweets tag as "Negative"
70% Tweets used for Training
30% Tweets used for Testing
Trained Data
Classification Algorithm
Predict the value
Output used to calculate Metrics

Figure 4. The Classification Process Flowchart

From above Figure 4 it can be observed that the classification flowchart consists of various steps and that leads to the final calculation of metrics. Metrics is nothing but the accuracy in terms of algorithm. The cleaned tweets are divided in training & testing with the percentage of 70/30.

Steps in flowcharts are explained below with numerical format

1.) Data fram of Positive & Negative tweets is created.
2.) Total number of tweets in the data frame are 3000.
3.) It is been divided into 2 separate data frame.
4.) The one with maximum tweets is Training data frame that is 70% of 3000.
5.) The other is Testing data frame that is 30% of the 3000.
6.) Once the data is splitted it has been trained with different classification algorithms.
7.) The trained data is then tested with testing data to check what accuracy is generated.
8.) The accuracy is then explained with the help of Confusion Matrix.

5. Experimental Results

.In this paper we reviewed the sentiments of people using tweets extracted from twitter. These sentiments helped in understand the perception of people towards the company. Finally, the Word Cloud was generated of Uber & Ola displaying the words that were frequently used. The word cloud can be seen in Figure 5 for Uber & Figure 6 for Ola datasets respectively.

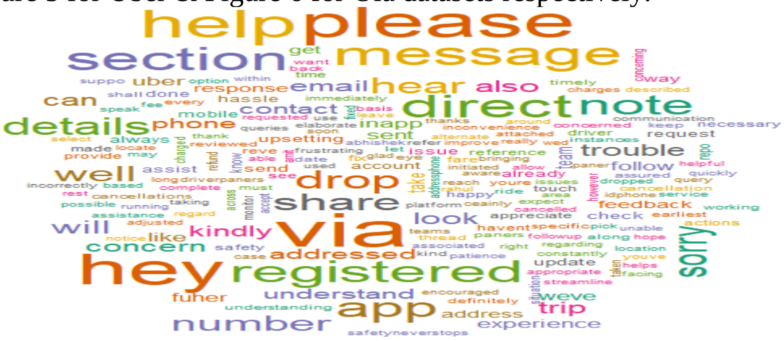


Figure 5. Word Cloud of Uber

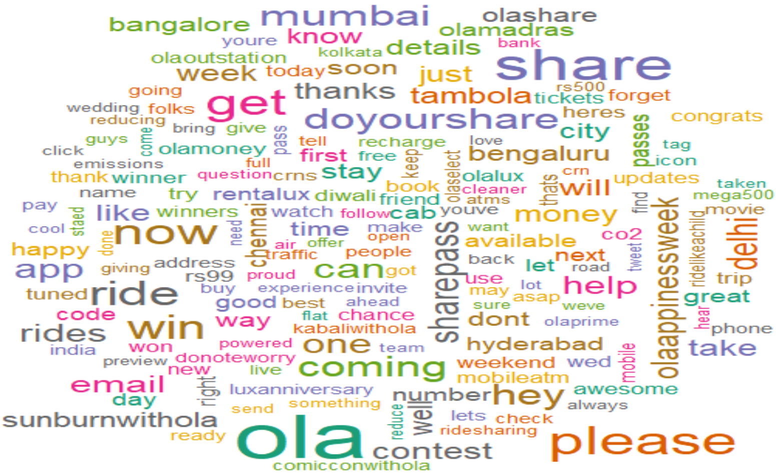


Figure 6. Word Cloud of Ola

The Secondary purpose of this paper was to categorize the data (Tweets).
Categorization of data was done in 2-category:

i. Positive
ii. Negative

Then this data was trained with 70% of partitioning. The rest 30% was used for testing.
After the training was done it was tested to check how good data was trained.
To understand the results, Accuracy was generated with the help of Confusion Table.
The formula for accuracy is

$$\text{Accuracy} = \frac{a+b}{a+b+c+d} \qquad (1)$$

Table 1. The Confusion Matrix Table

	True PosTweets	**True Neg Tweets**
Predicted PosTweets	a	b
Predicted Neg Tweets	c	d

The above table shows the confusion matrix, which is the table where accuracy can be calculated based on values obtained in respective cells. The Overall accuracies of two algorithms is indicated in Table 2 and Figure 7.

Table 2. Overall Accuracy Comparison on 3000 Data

Number of Test	**Number of Tweets Used**	**Accuracy**			
		SVM (Uber)	**Naïve Bayes (Uber)**	**SVM (Ola)**	**Naïve Bayes (Ola)**
1.)	3000	84.87%	86.65%	73.19%	73.64%

In above table 2 the overall accuracy of two algorithms on two different datasets is calculated with the help of confusion matrix. It can be seen that the SVM in both the datasets performed good but in compare to Naïve Bayes on both the datasets was not that good, as Naïve Bayes in both datasets outperformed SVM. The graphical representation can be observed in the below Figure 7.

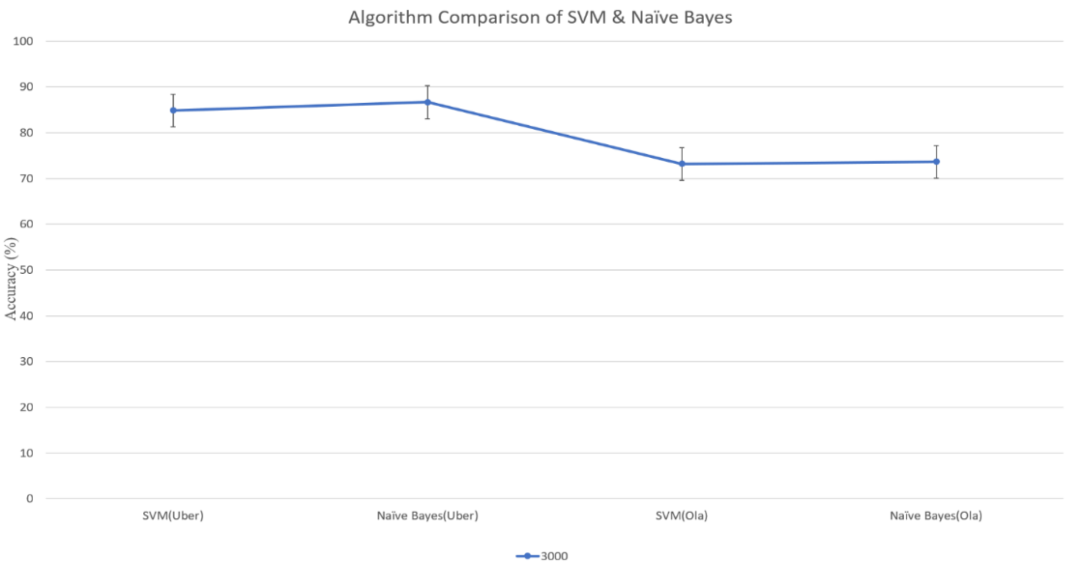


Figure 7. The Diagrammatic representation of accuracies in the experiment

Table 3. Results of Correct Tweets & Incorrect Tweets

Number ofData	Algorithm Used	Dataset	Correct Tweets	Incorrect Tweets
3000	SVM	Uber	763	136
		Ola	654	245
	Naïve Bayes	Uber	790	109
		Ola	661	238

The above table shows the correct tweets & incorrect tweets of the two datasets which are Uber & Ola on two different algorithms that are SVM & Naïve Bayes.

6. Conclusion

The focus of the paper is to evaluate the accuracy between the two classification algorithms and understand what accuracy is been generated also to understand the sentiments of the people with the help of sentimental analysis. In this paper, the two algorithms are compared for sentimental classification of tweets. The experimental data showed that the classifier yield better results for the Uber datasets which was trained with Naïve Bayes, similarly the Ola datasets yield good results in caseof Naïve Bayes. As we can see that Naïve Bayes was dominant in both the cases with accuracy of 86.65% in the case of Uber & 73.64% in the case of Ola. Thus, it can be said that the Naïve Bayes is better algorithm that can be used to classify the Uber & Ola datasets. Finally, the sentiments of the people through tweets are shown with the help of word cloud. Word cloud is the visual representation of the word that are used most in the tweets making us understand what people want to convey in the message. It can be helping the particular organization to understand their people and to make the business even better through sentimental understanding.

7. Future Work

This project focuses on performances in terms of accuracies of "Naïve Bayes" and "SVM" machine learning algorithms. The datasets we used in the project are "Uber" and "Ola" cab services. These machine learning algorithms; the outputs were displayed and statistical data. Likewise, we can use these machine learning algorithms for different data sets as well. We can help and improve businesses by using the data from popular food delivery apps like "Zomato" and "Swiggy". We can use the tweets of the customers who give reviews about the services and generate statistical data considering the positive and negative sentiments of the customers. In the case of our datasets that are "Uber" and

“Ola” we can further find out ‘Area-based customer sentiments’ and give improvised results to “Uber” and “Ola”. They can further focus on areas that need to improve services to the user. For example: If we come to know an area has given outputs which have more negative sentiments we can focus on that area and find out the reason. We can investigate what can we deliver more to such customers.Also, this paper was concerned on shallow learning this can be also done on deep learning with different datasets or the same.

References

[1] Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, Sweta Tiwari. "Sentiment Analysis of Review Datasets Using Naïve Bayes‘ and K-NN Classifier", International Journal of Information Engineering and Electronic Business, 2016.

[2] P.Kalaivani, “Sentiment Classification of Movie Reviews by supervised machine learning approaches” Indian Journal of Computer Science and Engineering (IJCSE) ISSN: 0976-5166 Vol. 4 No.4 Aug-Sep 2013.

[3] "Progress in Computing, Analytics and Networking”, Springer Science and Business Media LLC, 2018.

[4] Manish N. Tibdewal, Swapnil A. Tale. "Multichannel detection of epilepsy using SVM classifier on EEG signal”, 2016 International Conference on Computing Communication Control and automation (ICCUBEA), 2016.

[5] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, “Tapping into the Power of Text Mining”, Journal of ACM, Blacksburg, 2005.

[6] Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining." *LREc*. Vol. 10. No. 2010. 2010.

[7] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." *CS224N project report, Stanford* 1.12 (2009): 2009.

[8] Jadav, Bhumika M. and Vimalkumar B. Vaghela. “Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis.” (2016).

[9] Pennacchiotti, Marco, and Ana-Maria Popescu. "A machine learning approach to twitter user classification." *Fifth international AAAI conference on weblogs and social media*. 2011.

[10] Lina L. Dhande and Dr. Prof. Girish K. Patnaik, “Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier”, IJETTCS, Volume 3, Issue 4 July-August 2014, ISSN 2278-6856.