International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Using Data Mining Techniques to Analyze the Customers Reaction towards Social Media Advertisements

Ashutosh Bansal, Saleena B., Prakash B

Abstract: ***As social media is in boom, it is becoming very easier for customers to share their views and comments and express their feelings regarding any products which are present in online social media. . If these data can be analyzed efficiently different suggestions can be provided to the company regarding to improvise their products sale. It becomes easier for the company to understand the customer's reaction after seeing the advertisements of the products posted on social media. This research focuses on analyzing the sentiments of customers based on the comments and reviews of products available in Facebook. Sentimental Analysis is performed to analyze the customer comments as positive, negative and neutral and later they are labeled as 0 or 1. After the labeling process, a comparative analysis is performed using different classification algorithms. The classification algorithms used are K Nearest Neighbors (KNN), Support Vector Machine (SVM) and Naïve Bayes Classifier. The classification algorithm with the highest accuracy is identified to predict the sales of online products.***

Index Terms: Social media, classification, reviews, opinion mining, sentiment analysis, feedback

I. INTRODUCTION

Facebook social media application is used by about 2.23 billion users during October 2018. In Twitter, the total number of tweets done in every 4 days is nearly billion or 6000 tweets/seconds. Compared to traditional media sources, social media information is distributed around the world. Social media plays a very important role in predicting the future. Before the existence of social media people got the information through traditional media but were not able to give feedback regarding that information. Now as social media is trending, people post there reviews, comments and feedbacks of the product they have used .There are lots of feedback given by the customers in different contents of Facebook as comments and reviews. This feedback is very much necessary because this feedback is analyzed properly that results in some prediction that what will happen next..

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Ashutosh Bansal*, II Year MCA, School of Computing Science and Engineering, VIT - Chennai Campus
Saleena B, Associate Professor, School of Computing Science and Engineering, VIT - Chennai Campus
Prakash B, Assistant Professor, School of Computing Science and Engineering, VIT - Chennai Campus

Online advertisement in social media is one of the methods that companies or organization use for increasing the sales of the product. Before launching a new phone they post the advertisement of that phone in social media, a large number of users are interacting with social media from different part of the world see the advertisements in social media make comments and give there reviews about that phone

This comments, likes and reviews are analyzed properly by the Company. With the help of these data, the company predicts whether the customers like the product or not. Data mining techniques can be used for extracting the data from huge data sets and also helps us in predicting facts related to the dataset. For mining the information, different techniques are used like association, clustering and classification. It's a real challenge to choose an appropriate algorithm for analysing data according to the requirements. In Facebook lots of advertisements are posted per day, these advertisements contain lots of user data like comments of the customer which may be either positive or negative. The main aim of this research is to analyse the data which are extracted from Facebook. Sentiment analysis is performed on the extracted comments to analyse if the product is liked by the customers or not, and after sentiment analysis the prediction is done to predict the sales of the product.Based on this analysis, changes are incorporated in to the products according to customer's requirements.

II. LITERATURE REVIEW

The objective of this research is toanalyse the customer's reaction for a particular product and to predict the sales of the product for which the advertisement is done. The algorithms for analysis were chosen after an exhaustive literature review. Ping Feng Pai and Chia HsinLiv have proposed a methodology for the prediction of sales of vehicles per month in USA [5]. The labelling of the data of Twitter has been done for analysing and predicting the sales and stock market data. The algorithm which is used by the author was Least Square Support Vector Regression (LSSVR). The importance of opinion mining and sentiment analysis was described byShaidShayaa[8]. The other techniques which were introduced by the author Elvira Popescu for the prediction of student's academic performance [3] was K Nearest Algorithm. This is done by analysing the student's interaction with social media. The author AnithaAnandan[1] has introduced lots of techniques for improving the recommendation system.

New techniques were identified for social media recommendation system. The algorithms which were used are K mean for clustering and K Nearest Neighbours for classification. The author describe that data mining consists of three levels of discovery design, translating them with the end goal to check then convince lastly utilizing the example to take care of business issues [12]. M Negnevitsky has described the techniques applied for marketing strategies for getting better benefits. The strategies discussed were direct marketing, trending business etc [11]. Direct marketing is all about the customer interaction towards the product that how much interest the customer is showing for purchasing the product and trend marketing means marketing the products according to the trends in the business world. R.I Morgan has described the email business communication between the companies. The customers are getting direct emails from the company and can access the mail directly [10]. The technique for identifying the popularity of any company and how much the company is popular, which brand is the best brand is discussed by Berry and Linoff. It also finds a correlation between two companies [9].

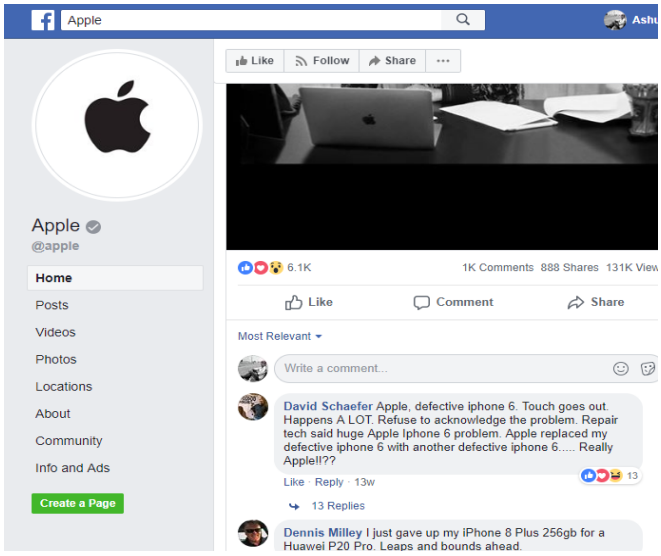
III. COMPARATIVE ANALYSIS

Three classification algorithms namely Naïve Bayes, Support Vector Machine and K Nearest Neighbor are chosen to find the accuracy of predicting the customer's reaction. With the help of python, the classification is done to find accuracy. After finding the accuracy of all these three algorithms the comparative analysis is done between these algorithms. The algorithm which has the highest accuracy is used for further prediction of the sales.

IV. PROPOSED METHODOLOGY

1. *Dataset*

The comments and feedback given by customers for the advertisements posted in the Facebook in the Apple page are present in the form of text. These comments are extracted with the help of Face Pager tool. This tool extracts all the comments and stores it into the database.



2. *Sentiment Analysis:*

Anorganization is said to succeed only when it satisfies the customer's needs. The most important thing for an organization is to identify what the customer's think about their company and their product. Based on the feedback of the customer, a company can make changes. For improving all this purpose the feedback and reviews given by the customers are analyzed, so that process of analyzing the reviews or feedback will result in some opinion which is called as mining or sentiments analysis. It's a process of identifying the attitude of the customers towards the particular product. The reviews can be positive, negative or neutral. Sentiment analysis is performed using the following steps: Tokenization, Cleaning the data, removing the stop words and Labelling the data.

Tokenization:

The process of separating the paragraph into a group of individual statements and the process of separating statements into a group of individual words, this division of each word of the sentence into tokens is called Tokenization.

Cleaning the data:

The next step in the sentiment analysis after tokenization is cleaning the data. Cleaning the data means removing the irrelevant, inaccurate data from the dataset so that data analysis can be done properly.

Removing the stop words:

The next step in sentiment analysis is removing the words like the, that, was, is, he, sheetc ,called stop words which are not important for analysis.

Labelling:

After removing all the characters and symbols which are not useful for analytics purpose the labelling is done with the remaining words of a particular sentence, labelling of the sentence either can be positive or negative or neutral, so positive sentence is labelled as +1 and negative sentence are labelled as -1 and neutral sentence are labelled as 0.

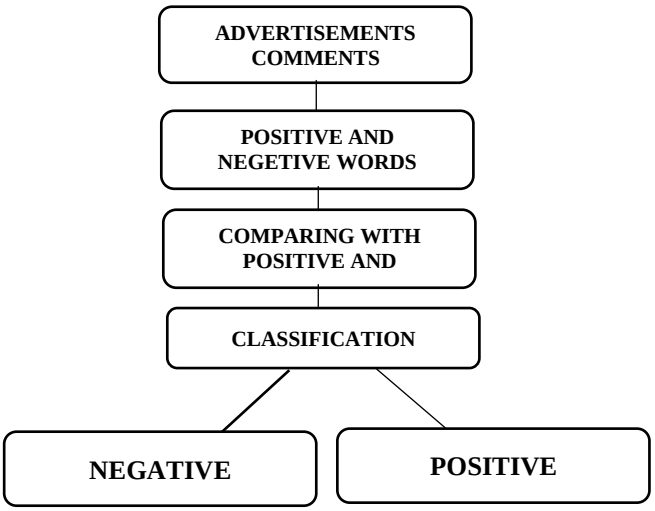


Figure 1 Sentiment Analysis

3. *Feature Extraction*

Documents are information which is present in the form of text.

Retrieval Number: B1700078219/19©BEIESP
DOI: 10.35940/ijrte.B1700.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The text can be of following types such as text which comes in the form of user comments or feedback the text came in the form of mail. This text can be called a complete observation.Corpus is nothing but it can be called as the complete documents or data set. Tokens are the division of the entire corpus into individual words the process of converting corpus into tokens is called tokenization. Feature extraction is the process in which the feature is extracted from the documents. Extraction of features means conversion of text into a numerical value as for doing the classification different types of mathematical operations are performed and these mathematical operations are not performed in a sentence or string so if it is converted into numerical value the mathematical calculation and classification techniques can be performed.

Input:

(["This phone is having good functionality", "The functionality is good but the cost is very high", "So due to functionality the sale is high"])

Features Extracted:

['but', 'cost', 'due', 'functionality', 'good', 'having', 'high', 'is', 'phone', 'sale', 'so', 'the', 'this', 'to', 'very']

Matrix:

```
[[0 0 0 1 1 1 0 1 1 0 0 0 1 0 0]
 [1 1 0 1 1 0 1 2 0 0 0 2 0 0 1]
 [0 0 1 1 0 0 1 1 0 1 1 1 0 1 0]]
```

Figure 2: Bag of Word Model

As shown in the above Figure 2 features are extracted from the input comment. Words which are present in this sentence is considered as 1 in the matrix and all the other values are considered as 0.

4. Classification

The technical fields which are in trending nowadays are Machine Learning and Artificial Intelligence. The most important aspects of Machine learning are classification and prediction.

Naïve Bayes Classifier:

Naïve Bayes is a simple but surprisingly powerful algorithm for predictive analytics. It is a classification technique based on Bayes theorem with an assumption of independence among predictors. It comprises of two parts which are Naïve and Bayes. It assumes that the presence of the particular feature in a class is unrelated to the presence of any other feature. Even if this feature depend on each other or upon the other features all of these properties independently contribute to the probability whether the comment is positive or negative.

Bayes theorem:

Given a hypothesis H and evidence E, Bayes theorem states that the relationship between the probability of the hypothesis before getting the evidence P(H) and the probability of the hypothesis after getting the evidence P(H|E) is

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)}$$

Origin of these Bayes theorem:

As it is given: -

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} \quad \text{--- (1)}$$

$$P(Y|X) = \frac{P(Y \cap X)}{P(X)} \quad \text{--- (2)}$$

On combining equation (1) and (2)

$$P(X \cap Y) = P(X|Y).P(X) = P(Y|X).P(Y)$$

Final Equation:

$$P(X|Y) = \frac{P(Y|X).P(Y)}{P(X)}$$

Support Vector Machine:

Support Vector Machine (SVM) is an accurate algorithm among all the classifier algorithms because the mathematical calculations performed in SVM is very good and accurate. SVM is the easiest way to represent over fitting. SVM is the only algorithm which works with a large number of features. Even though a large number of features are used it has very less computation.SVM works on the extremes of the dataset and after analysing it draws a line which is called as hyper plane and the extreme points which are on both sides of the hyper plane are called as support vectors.SVM is the algorithm which divides the two-class accurately (hyper plane/line).In Support Vector Machine only support vectors play important role whereas all other training examples are ignored.

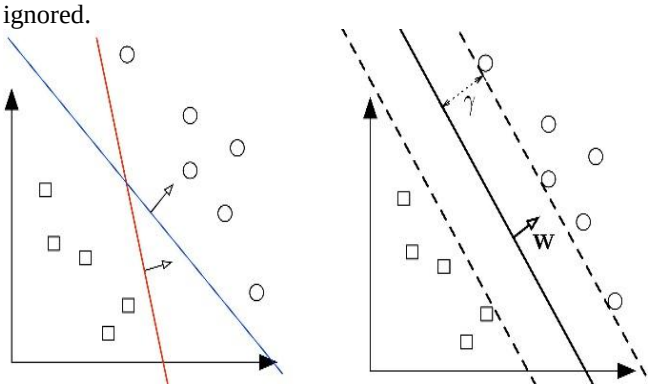


Figure 3 Support Vector Machine

K Nearest Neighbors (KNN):

K Nearest Neighbor is one of the easiest and simplest algorithms among the entire classification algorithms. KNN uses complete dataset while it is training the dataset or when the model is prepared. Whenever a prediction is in the requirement or it is done KNN searches the K most similar instances in the entire training dataset and data which comes out after this is returned as a prediction. KNN algorithm stores all the cases which are available and classify the new data. The main point which KNN describes is that if something is similar to the neighbour they are predicted to be one of them among theneighbour. K in KNN denotes the number of nearest neighbours which are voting class of the new data or the testing data.

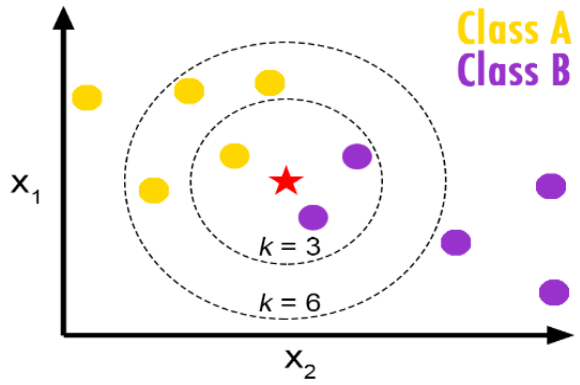


Figure 4 K Nearest Neighbors

As it is seen in the above figure that the majority is class B so we can say that for K=3 star belongs to class B.

V. EXPERIMENTAL DISCUSSION:

This section discusses about how the classification algorithms were implemented and evaluated. 2000 comments were extracted from the Apple page of the Facebook and are stored in the MS-Excel in tsv format. For classification, the dataset is divided into two parts they are training dataset and testing dataset, 80% of the data is given as the training data for training the model and 20% of the dataset are given as testing data for testing the model. After dividing the data three different classification algorithms are applied on the same dataset for obtaining the confusion matrix. Confusion matrix is just the collection of true positive, true negative, false positive and false negative. After that with the help of this confusion matrix comparative analysis is done between all three classification algorithms. After doing the classification the confusion matrix obtained are depicted in Table 1 and results obtained are depicted in Figure 5,6 and 7.

Table 1: Confusion Matrix of Classification Algorithms

CLASSIFICATION ALGORITHM	CONFUSION MATRIX	ACCURACY
SUPPORT VECTOR MACHINE	$\begin{bmatrix} 122 & 21 \\ 39 & 218 \end{bmatrix}$	63.5%
NAÏVE BAYES CLASSIFIER	$\begin{bmatrix} 117 & 26 \\ 114 & 113 \end{bmatrix}$	57.5%
K NEAREST NEIGHBORS	$\begin{bmatrix} 107 & 36 \\ 50 & 207 \end{bmatrix}$	78.5%

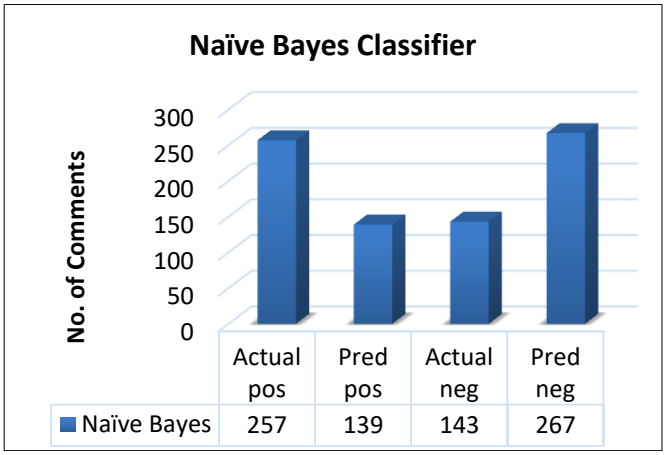


FIGURE 5 Comparison Between the Actual Data and Predicted Data

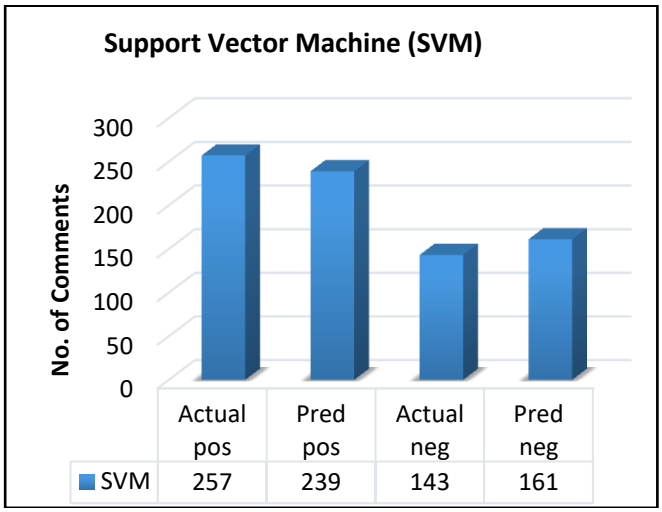


Figure 6 Comparison between the actual data and predicted data

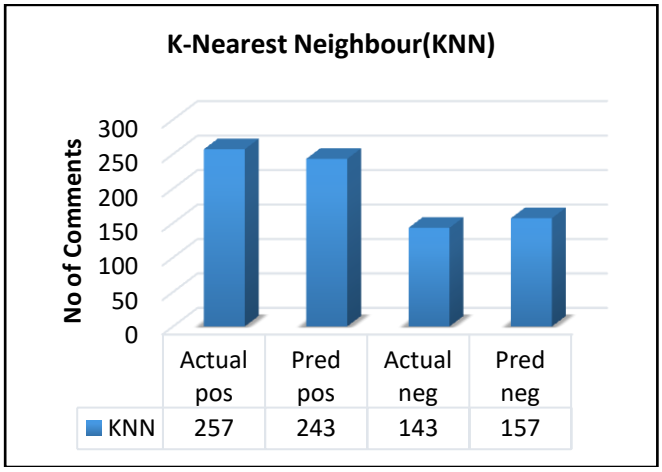


Figure 7 Comparison between the actual data and predicted data

After doing the comparative analysis with all the three algorithms the results of accuracy obtained are depicted in figure 8.After comparing the accuracy of all the three algorithms the results obtained are Naïve Bayes has 59.25% accuracy, Support Vector Machine has 65.25% accuracy and K Nearest Neighbour has 80.05% accuracy.

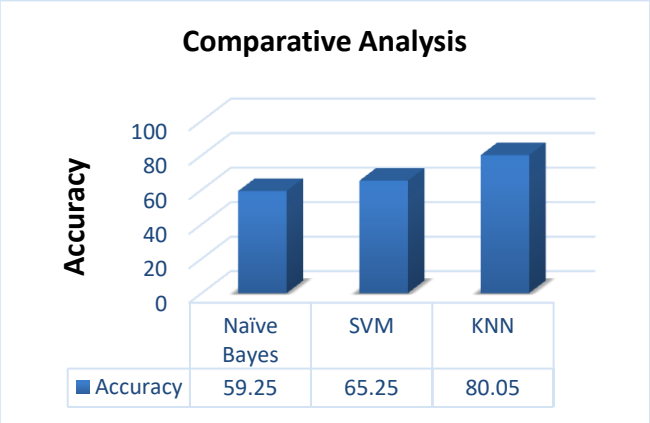


Figure 8 Comparison between the three algorithms

The comparative analysis proves that KNN has the highest accuracy in classifying the sentiments of customers.

VI. CONCLUSION

In this research work, different types of classification algorithms are used for testing the model prepared with the dataset extracted from the social media application that is Facebook and comparative analysis is done according to the accuracy obtained with this algorithm. For analysing the sentiments of the comments or reviews sentiment analysis technique is used. The results obtained by different algorithms are represented in the form of a bar graph. The highest accuracy of classification algorithm obtained is used further for the prediction process. In Future the research can be extended to analyse and extract the data which are present in social media in different languages. For example there is lots of data present in Urdu language if this data is extracted and analyzed there are lots of information hidden in that data which can be obtained only if the languages other than English can be analyzed for information mining.

REFERENCES

1. ANITHA ANANDHAN 1, LIYANA SHUIB1, MAIZATUL AKMAR ISMAIL1, AND GHULAM MUJTABA "Social Media Recommender Systems: Review and Open Research Issues", IEEE Transactions on Knowledge and Data Engineering ,date of publication February 27, 2018
2. MUBASHIR ALI , SHEHZAD KHALID, AND MUHAMMAD HASEEB ASLAM, "Pattern Based Comprehensive Urdu Stemmer and Short Text Classification", IEEE Translation and Content Mining, date of publication December 28, 2017
3. ELVIRA POPESCU 1, (Member, IEEE), AND FLORIN LEON2 "Predicting Academic Performance Based on Learner Traces in a Social Learning Environment" , IEEE Transactions on Knowledge and Data Engineering, date of publication November 20, 2018
4. Deokgun Park, Seungyeon Kim, Jurim Lee, JaegulChoo, Nicholas Diakopoulos, and NiklasElmqvist, Senior Member, IEEE:" ConceptVector: Text Visual Analytics via Interactive Lexicon Building using Word Embedding", IEEE Transactions on Visualization and computer graphics, date of publication JANUARY 2018
5. PING-FENG PAI , (Senior Member, IEEE), AND CHIA-HSIN LIU "Predicting Vehicle Sales by Sentiment Analysis of Twitter Data and Stock Market Values" , IEEE Transactions on Knowledge and Data Engineering ,date of publicationOctober4,2018
6. KUN GAO AND YIWEI ZHU: "Deep Data Stream Analysis Model and Algorithm With Memory Mechanism", IEEE Transactions on Knowledge and Data Engineering, date of publication September 27,2018
7. WEI LU, HONGBO SUN, JINGHUI CHU, XIANGDONG HUANG , (Member, IEEE), AND JIEXIAO YU:A Novel Approach for Video Text Detection and Recognition Based on a Corner Response Feature Map and Transferred Deep Convolutional Neural Network, date of publication July 2, 2018
8. SHAHID SHAYAA1, NOOR ISMAWATI JAAFAR 2, SHAMSHUL BAHRI2, AININ SULAIMAN 2, PHOONG SEUK WAI 2, YEONG WAI CHUNG2, ARSALAN ZAHID PIPRANI 2, AND MOHAMMED ALI AL-GARADI2:Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges, IEEE Transactions on Knowledge and Data Engineering, date of publication June 28, 2018
9. M. J. A. Berry and G. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management 3rd ed., John Wiley and Sons Ltd., Publication, UK, 2016.
10. D. J. Hand, H. Mannila and P. Smyth, Principles of Data Mining, MIT Press, Cambridge, MA, 2015.
11. J. Han and M. Kamber, Data Mining Concepts and Techniques Second Edition, Morgan Kaufmann Publishers, United States of America, 2016.
12. R. I. Magos and C. A. Acatrinei, Designing Email Marketing Campaigns - A Data Mining Approach Based onConsumer Preferences, A nalesUniversitatisApulensis Series Oeconomica, 17(1), 2015, 15-30.