Sentiment Analysis on Business Data using Machine Learning

Mrs Usha G R
Department of ISE
SDM Institute of Technology
Ujire, INDIA
ushagr2@gmail.com

Dr. Dharmanna L.
Department of ISE
SDM Institute of Technology
Ujire, INDIA
dharmannasdmit@gmail.com

2021 Second International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE) | 978-1-6654-8297-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSTCEE54422.2021.9708593

***Abstract*—Sentiment Analysis, also known as Opinion Mining, is the systematic identification, extraction, quantification, and study of affective states and subjective information using natural language processing and text analysis. For applications ranging from marketing to customer service, sentiment analysis is commonly used on reviews and survey answers, online and social media, and healthcare materials.**

The aim of this study is to determine the accuracy of the various algorithms for sentiment analysis of product reviews. For this study, 4,444 online product reviews were received from an e-commerce site.

Keywords— Sentiment analysis, Classification

I. INTRODUCTION

Sentiment analysis is the process of extracting meaningful information from unstructured and unorganized text content on social media platforms, blogs, and comments. Since the popularity of online markets in recent decades, online sellers and merchants have asked their customers to share their opinions on the things they sell. Every day, millions of comments are made on the Internet concerning various products, services, and locations. As a result, the Internet has become the most essential source of product and service ideas and opinions.

However, as the number of product assessments available increases, it becomes more difficult for potential customers to make the right decision about whether to purchase the product or not. On the one hand, different perspectives on the same product can be confusing; on the other hand, ambiguous comments can make it even more difficult for customers to make the proper decision. The requirement to assess these contents appears to be critical for all e-commerce businesses in this case. Every day, people shop online through e-commerce sites, where they may read thousands of evaluations from other consumers about the things they want. These reviews can assist purchasers comprehend practically every facet of the product by providing helpful thoughts about its attributes, quality, and suggestions. This is advantageous not only to customers, but also to businesses, but also helps marketers who make products to better understand consumers and their needs. Analyzing the sentiment tendency of customer evaluation can not only provide a reference for other customers but also help business on the e-commerce platform to improve service quality and customer satisfaction.

Sentiment analysis plays a vital role in analyzing and understanding the communication that occurs in transcription. Sentiment evaluation and category is a computational examine that tries to resolve issues through extracting subjective information (consisting of reviews and feelings) from a textual content given in natural language. From natural language processing, textual content evaluation, computational linguistics, and biometrics, extraordinary techniques had been used to resolve this problem. In current years, machine learning techniques have end up famous in comment and semantic evaluation because of their simplicity and accuracy.

To analyze the sentiment analysis, we use Machine Learning as a tool. By training the machine learning tool with an example of emotions in text, machine learning automatically learns how to detect sentiment. Based on the reviews shared by customers on social media platforms, sentiment analysis plays a key role in understanding the shopping experience of a particular site and also determines the importance of the shopping experience obtained when buying. products in a particular store.

Machine learning aims to develop an algorithm that optimizes system performance by using sample data. The sentiment analysis solution use of machine learning includes major steps. The first step is to analyze the model from the training facts and the second step is to categorize the invisible facts with the help of the training model. Machine learning algorithms are divided into specific categories: supervised learning, semi-supervised learning, and unsupervised learning. The project is considering the use of supervised machine learning methods to conduct emotional ratings of online reviews, and govern the complete semantics of buyer analyses by rating them positive, negative, and neutral and to know the accuracy of each learning methods.

II. LITREATURE REVIEW

Ying Fang, Hai Tan *et al*. [1] proposed a multi-strategy sentiment analysis based on Naïve Bayes Classifier and SVM. The opinion of customer is expressed in Chinese phases but due to the fuzziness of Chinese characters Machine learning techniques can not represent the opinion of articles.

Pushpitha kenchana Sari, Andry Alamsyah *et al.* [2] used Tokopedia, one of the largest e-commerce services in Indonesia. The author applies Naive Bayes classification method due to its high precision and support for big data processing. The results show that the dimensions of

978-1-6654-8297-4/21/$31.00 ©2021 IEEE

personalization and reliability require more attention because they have high negative emotions. At the same time, the dimensions of trust and network design have a high confident sentiment, which means that you have good service.

Supryadi,Antoni Wibowo *et al.* [3] committed to data mining and machine learning algorithms to detect the intensity of emotions and analyze them. These algorithms are reasonable and effective for sentiment analysis. Even simple algorithms can handle large data sets well, as can the naive Bayesian method.

Salinca and Andreea [4] proposed various automatic sentiment classification methods, using two feature extraction methods (preprocessing and feature extraction) and four machine learning models (Naive Bayes, Linear SVC, Logistic Regression and Stochastic Gradient Descent). The author gives examples of comparative studies on the effectiveness of ensemble methods for examining sentiment classification.

Zeenia Singla, Suchchandran Randhawa et al. [5] evaluated that opinions are categorized into fine and bad sentiments by using Sentiment Analysis. Out of the numerous classification models, Naïve Bayes, Support Vector Machine (SVM) and Decision Tree had been hired for the class of opinions. The assessment of models is accomplished with the use of 10-Fold Cross-Validation.

Manvee Chauhan and Divakar Yadav [6] tested the effectiveness of machine learning techniques for sensory classification. Machine learning (ML) is divided into supervised methods and unsupervised methods. Compared with unsupervised learning that does not require prior training, supervised learning is often more accurate because each classifier is trained on a set of representative data called a corpus.

Alessia, D'Andrea et al. [7] offers an overview of the distinct sentiment classification strategies and tools for sentiment analysis. From this general description, the author categorizes the characteristics / techniques and advantages / limitations and methods of the different techniques used for sentiment analysis.

Muhammad Marong and Nowshanth K [8] describe the normal distribution involved in sentiment analysis methods, such as dictionary-based methods and monitored machine learning. However, sentiment analysis is mainly used in e-commerce retail business, which allows operators to improve their business operations.

Najma Sultana and Pintu Kumar [9] claimed that sentiment analysis is based on text analysis, but there are certain challenges in finding the precise polarity of sentences. Therefore, in order to find the polarity of a sentence, an automated data analysis technique is needed. By demonstrating the Data Filtration Algorithm, Algorithm for Machine Learning implementation, and the proposed algorithm to perform Sentiment Analysis, the polarity of the data in the Web and their classifiers can be identified.

The feature selection technique was used by Tommy Willianto and Suprayadi [10] to improve classification performance. Terms The sentiment of the documents was analysed the use of Frequency-Inverse Document Frequency (TF-IDF) characteristic extraction, Backward Elimination function selection, and 5 distinct classifiers (Nave Bayes, Support Vector Machine, K-Nearest Neighbor, Decision Tree, and Random Forest).

Data mining and machine learning methods were utilised by BhanuSree Reddy and Uma Precilda Jaidev et al. [11] to determine sentiment strength and sentiment analysis. A study examines the importance of sentiment analysis in social networks, using data collecting, text preparation, classification algorithms, and performance evaluation outcomes to obtain high levels of accuracy in the polarity of opinions.

Hemalatha, G P Sharadhi Varma, and colleagues [12] provided a method that collects tweets from social media websites and offers an enterprise intelligence view. The sentiment analysis tool has layers: a fact processing layer and a sentiment analysis layer. The facts processing layer is liable for facts accumulating and facts mining, while the sentiment analysis layer gives the facts mining consequences thru an application.

With this we observed that, different authors used particular or some combination of algorithms to carry out sentiment analysis. So, we worked on common data set in five different approaches of supervised learning to know which method gives better accuracy for the reviews data.

III. SYSTEM DESIGN

A. User Interaction System Design

The design of the system is done as per the requirements of the project. The system classifies the review by using tools from Machine Learning. Furthermore, the use of NLTK is used to do the task in the natural language process and also supports multiple languages like English, Chinese, etc. to do classification text or data into something meaningful. The dataset used for the application is a CSV dataset containing 35,000 reviews. It is an excellent database for machine learning methods while needing minimal effort in preprocessing and classification.

B. Prediction System Design

The sentiment analysis was carried out using machine learning techniques. Naive Bayes Classifier (NBC), Support Vector Machine (SVM), Logistic Regression, Decision Tree, and Random Forest Classifier are the algorithms employed.

C. Naïve Bayes Classifier (NBC)

For sentiment analysis applications, the Naive Bayes Classification method is often used as a starting point. The core principle behind Naive Bayes is to use the joint prospects of words to find the probability of classes given to texts.

Given the structured characteristic vector ($x_1 \ldots$ xn) and the magnificence Ck. Bayes theorem is said mathematically as the subsequent relationship

$$P(C_k|x_1, \ldots, x_n) = \frac{P(C_k)\, P(x_1, \ldots, x_n|C_k)}{P(x_1, \ldots, x_n)} \quad (1)$$

According to the "Naive", conditional independence assumptions, for the given elegance C_k every characteristic of vector x_i is conditionally unbiased of each different characteristic x_j for $i \neq j$

$$P(x_i|C_k,x_{1,\dots,}x_n) = P(x_i|C_k) \quad (2)$$

Thus, the relation can be simplified to

$$P(C_k|x_1,\dots,x_n) = \frac{P(C_k)\prod_{i=1}^{n} P(x_1|C_k)}{P(x_{1,\dots,}x_n)} \quad (3)$$

Since $P(x_1,\dots, xn)$ is constant, if the values of the function variables are known, the subsequent class rule may be used:

$$P(C_k|x_1,\dots,x_n) \propto P(C_k)\prod_{i=1}^{n} P(x_i|C_k) \quad (4)$$

$$\hat{y} = argmax_k P(C_k)\prod_{i=1}^{n} P(x_i|C_k) \quad (5)$$

The assumptions that distinct naive Bayes classifiers make approximately the distribution of P(xi |Ck), while P(Ck) is generally described because the relative frequency of sophistication C_k in the training dataset, are the primary variations among them.

The multinomial distribution is parameterized by vector $\theta_{ki}=(\theta_{k1},\dots,\theta_{kn})$ for each class C_k, where n is the number of features (i.e. the size of the vocabulary) and θ_{ki} is the probability $P(x_i|C_k)$ of feature i appearing in a sample that belongs to the class C_k.

A smoothed variant of maximum likelihood, i.e. relative frequency counting, is used to estimate the parameters θ_k:

$$\widehat{\theta_{ki}} = \frac{N_{ki}+\propto}{N_k+\propto n} \quad (6)$$

Where *Nki* is the quantity of times feature *i* seems in a sample of class *k* in the training set *T*, and *Ny* is the over-all count of all features for class C_k. Setting $\alpha=1$ is called Laplace smoothing, while $\alpha<1$ is called Lidstone smoothing.

Thus, the final decision rule is defined as follows:

$$\hat{y} = argmax_k(ln\, P(C_k)+\sum_{i=1}^{n} ln\frac{N_{ki}+\alpha}{N_k+\alpha n}) \quad (7)$$

D. Support Vector Machine (SVM)

SVM is a supervised machine learning approach that may be used to resolve issues in classification and regression. Regression predicts a continuous value, at the same time as classification predicts a label/group. Support Vector Machine (SVM) classifier works by creating a line that divides the dataset into support vectors by leaving the larger margin as possible between points. As shown in figure 1, Because line A has a bigger margin than line B, points divided by line A must travel much further to cross the division than points divided by B, hence we would choose line A in this scenario.

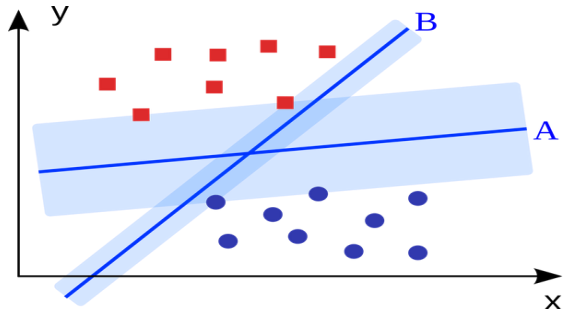

Figure 1. An illustration of SVM method

If we do not know anything about the facts, SVM is a totally beneficial method. It can be used to keep facts which include images, text, and audio, amongst different things. It can be implemented to data that isn't evenly distributed and whose distribution is unknown. The SVM contains a very important technique called as kernel, and we can solve any complex problem by using the associated kernel function. Kernel allows you to select a function that is not always linear and can take on different shapes depending on the data it works with, making it a non-parametric function.

E. Logistic Regression

Logistic regression is a supervised machine learning technique for classification problems. The goal of the model is to learn and approximate a mapping function f (xi) = y from input variables {x1, x2, xn} to output variable (y). Because the model predictions are iteratively assessed and corrected in contradiction of the output values until a suitable performance is obtained, it is called supervised.

F. Decision Tree Classification

A decision Tree is a supervised learning approach that can be used to resolve each classification and regression problem, but it is most generally employed to resolve classification issues. Internal nodes constitute dataset attributes, branches constitute decision rules, and every leaf node gives the conclusion in this tree-structured classifier. Leaf nodes are the output of these selections and do now no longer include any greater branches, while Decision nodes are used to make any decision and have numerous branches. The Figure 2 explains the general structure of a decision tree:

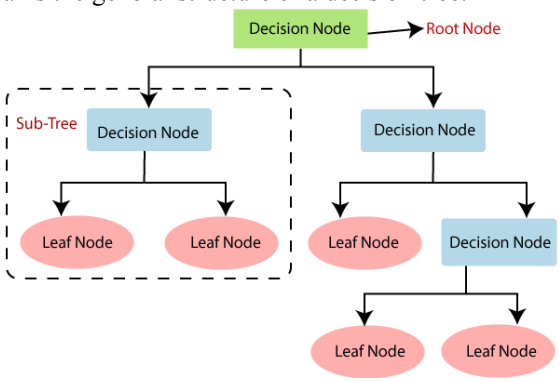

Figure 2: Structure of decision tree

G. Random Forest Classification

In machine learning, Random Forest can be used to resolve each classification and regression problem. It is based on ensemble learning, which is a technique of integrating numerous classifiers to resolve a complex hassle and increase the model's performance. Random Forest is a classifier that mixes some of the decision trees on different subsets of a dataset and averages their consequences to grow the dataset's predictive accuracy. Instead of relying on a single decision tree, the random forest collects the forecasts from every tree and predicts the final output based on the bulk votes of predictions.

IV. IMPLEMENTATION

Implementation is the process of carrying out or executing a project according to a set of instructions in order to complete it and achieve the expected objectives. This method includes all processes and actions involved in carrying out the project plan and achieving the project's goals and objectives.

A. Split into Train and Test

- We divided the dataset into training and test datasets before diving into it.
- Our long-term goal is to develop a sentiment analysis classifier.
- We used a stratified split at the opinions rating to keep away from training the classifier on facts that became unbalanced.

```
split = StratifiedShuffleSplit(n_splits=5, test_size=0.2)
for train_index, test_index in
    split.split(dataAfter,dataAfter["reviews.rating"]):
```

B. Data Exploration (Training Set)

We utilized regular expressions to clean the dataset of any bad characters, and then ran a preview following the cleaning procedure.

```
reviews = strat_train.copy()
reviews.head(2)
```

C. Sentiment Analysis

With the features available, we created a classifier that can detect the sentiment of a review. The act of evaluating a piece of text for views and feelings is known as sentiment analysis. There are numerous real-world applications for sentiment analysis, such as determining how customers feel about a product or service.

Set Target Variable (Sentiments)

```
def sentiments(rating):
if (rating == 5) or (rating == 4): return "Positive"
elif (rating == 2) or (rating == 1): return "Negative"
```

Extract Features

The following steps are used to extract the features from the dataset.

Step 1: Using the Bag of Words technique, turn a dataset into numerical feature vectors.

Step 2: Assign a fixed integer id to each word occurrence in the dataset as X [i, j], where i is the integer indices, j is the word occurrence, and X is our training dataset.

To contrivance the Bag of Words approach, SciKit-Learn's CountVectorizer is used and it performs as follows:

- Text preprocessing: Text preprocessing is done in two levels

  - Tokenization: The process of breaking sentences into words is called Tokenization. We used the following method to tokenize the sentences.

```
import nltk
nltk_tokens = nltk.sent_tokenize(strat_train["reviews.text"])
```

  - Removing Stop words: In the dataset there may be unwanted words like "the", "are", "is", "a", etc. To achieve the better analysis we need to remove these words from the dataset. This can be done as follows

```
stopword = [word for word in tokens if not word in
stopwords.words (' strat_train["reviews.text"])')]
```

- Occurrence counting: Now we need to count the frequency of words. So, construct a feature dictionary using integer indices and word occurrences using the following code
- Feature Vector: Lastly, feature vectors are obtained from the dictionary of text document

```
from sklearn.feature_extraction.text import CountVectorizer
count_vect = CountVectorizer()
```

D. Building a Pipeline from the Extracted Features

The multinomial naive Bayesian algorithm is best suited for word counts where the data is generally represented as word vector counts where the result number X [i, j] is observed during the n attempts, while the non-occurrence of a characteristic i. it is ignored. P (x | y), where x is the characteristic and y is the classifier, is a simplified form of Bayes' theorem in which all characteristics are considered conditionally independent of each other (the classifiers).

```
clf_multiNB_pipe=Pipeline([("vect",CountVectorizer()),("tf
idf",TfidfTransformer()),
("cf_nominalNB", MultinomialNB())])
clf_multiNB_pipe.fit(X_train, X_train_targetSentiment)
```

E. Testing Models

Naive Bayes Classifier

We trained our data with a Naive Bayes classifier. We used a multinomial Naive Bayes classifier that works well with data-based feature vectors. We did this with the MultiNB class:

```
predictedMultiNB = clf_multiNB_pipe.predict(X_test)
np.mean(predictedMultiNB == X_test_targetSentiment)
```

Logistic Regression Classifier

We used logistic regression to create a sentiment classification model, and we trained the model using a review sample dataset as part of the process.

```
predictedLogReg = clf_logReg_pipe.predict(X_test)
np.mean(predictedLogReg == X_test_targetSentiment)
```

Support Vector Machine Classifier

SVM are supervised learning models that are commonly used to classify data. To categorize the reviews in dataset, we used Linear SVM.

```
from predictedLinearSVC =
clf_linearSVC_pipe.predict(X_test)
np.mean(predictedLinearSVC == X_test_targetSentiment)
```

Decision Tree Classifier

Decision trees are a type of classifier in which each node represents a test on a data set attribute, and the children represent the outcomes. The leaf nodes represent the data points' final classes.

predictedDecisionTree =
clf_decisionTree_pipe.predict(X_test)
np.mean(predictedDecisionTree == X_test_targetSentiment)

Random forest Classifier
Random forests construct decision trees from data samples picked at random, obtain predictions from each tree, and then choose the best option. It's also a good indicator of how important the feature is.
predictedRandomForest =
clf_randomForest_pipe.predict(X_test)
np.mean(predictedRandomForest
==X_test_targetSentiment)

- The accuracy is calculated and displayed
- Multinomial Naïve bayes classifier, Logistic regression classifier, Support Vector Machine Classifier, Decision Tree Classifier and Random forest classifier are applied on dataset for evaluating the sentiments

V. Result

The goal of this application/system is to train the reviews dataset so that the system can determine if a given statement deserves a favourable, negative, or neutral assessment. Multinomial Naive Bayes, Logistic Regression, Decision Tree, Random Forest Classifier, and Support Vector Machine models were used to analyse the sentiment of the reviews.

Table 1. Accuracy of classification algorithms

Algorithm	*Accuracy*
Multinomial Naive Bayes	*93.4%*
Logistic Regression	*93.92%*
Support Vector Machine	*93.93%*
Decision Tree	*90.16%*
Random Forest Classifier	*93.50%*

Table 1 represents the accuracy of Multinomial Naive Bayes, Logistic Regression**,** Decision Tree, Random Forest Classifier and Support Vector Machine algorithms. Multinomial Naive Bayes performed with an accuracy of 93.4%, **Logistic Regression** performed with an accuracy of 93.92%, Support Vector Machine performed with an accuracy of 93.93%, Decision Tree performed with an accuracy of 90.16% and Random Forest Classifier with an accuracy of 93.50%.

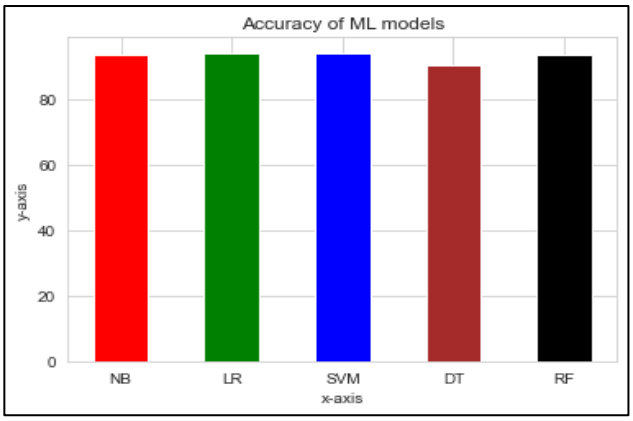

Figure 4. Accuracy of models

The above Bar graph shows the accuracy of classification algorithms.

Figures 5 depict the manual GUI window. The user will provide input in the form of text, and the system will recognize the feeling expressed in the text.

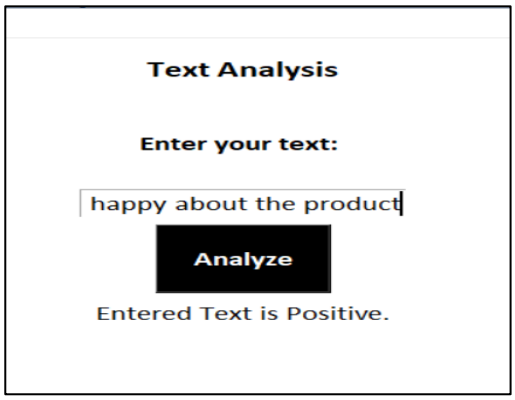

Figure 5. Manual GUI window for text analysis

Figure 6 shows a file analysis with review classification, as well as two graph options: simple and scatter graph.

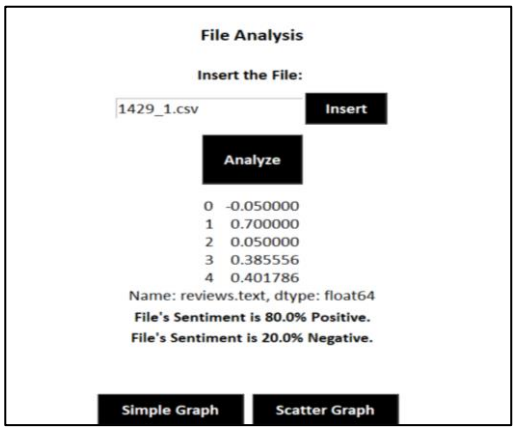

Figure 6. Manual GUI window for file analysis

After the extract feature, the dataset is analyzed by calculating the polarity of the review text. Then, as indicated in the image, plot the graph for polarity of the review text from the dataset analysis procedure. Figure 7 illustrates a basic graph of dataset sentiment analysis, whereas Figure 8 illustrates a scatter representation of dataset sentiment analysis.

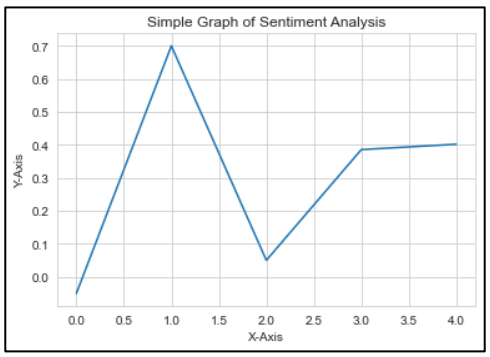

Figure 7. Simple graph of sentiment analysis

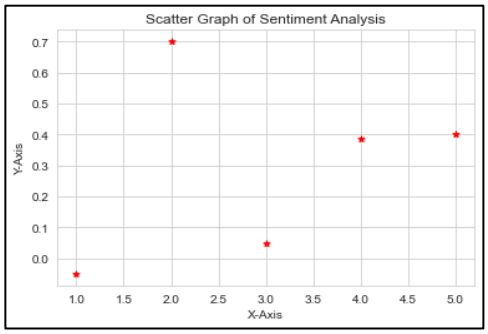

Figure 8. Scatter graph of sentiment analysis

VI. CONCLUSION

Sentiment Analysis on Business Data Using Machine Learning is a project that looks at how to classify reviews based on the sentiments they include. This project focuses on a common sentiment analysis model with four key processes: data splitting into train/test, extract features, design a pipeline from the retrieved features, and fine tune the support vector machine classifier, and covers representative techniques used in those steps.

This project analyzed the sentiment of e-commerce reviews and developed a model to predict the sentiment of a comment in the review text. The distribution and features of the data, as well as the machine learning techniques used, have an impact on the accuracy of machine learning models, especially since the data is text. The predicted accuracy of Support Vector Machine is shown to be the best among the five classifiers i.e., Naive Bayes Classifier, Support Vector Machine, Random Forest, Logistic Regression, and Decision Tree. The accuracy results were cross validated, and among the five models, SVM had the highest accuracy score of 94.08%. Sentiment analysis systems can recognize and evaluate texts automatically and fast, which is important for monitoring public opinion and clients' likes and dislikes in the information age.

REFERENCES

[1] Fang, Ying, Hai Tan, and Jun Zhang.(2018) "Multi-strategy sentiment analysis of consumer reviews based on semantic fuzziness." Ieee Access 6, 20625-20631.

[2] Sari, Puspita Kencana, Andry Alamsyah, and Sulistyo Wibowo(2018) "Measuring e-Commerce service quality from online customer review using sentiment analysis." Journal of Physics: Conference Series. Vol. 971. No. 1. 2018.

[3] Supryadi, Antoni Wibowo, et al.(2018) "Sentimental Analysis on E-commerce Product using Machine Learning and Combination of TF-IDF and Backward Elimination", International Journal of Recent Technology and Engineering, ISSN:2277-3878, Vol. 8, Issue-6,2862-2867.

[4] Salinca, Andreea. "Business reviews classification using sentiment analysis." 2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC). IEEE, 2015.

[5] Zeenia Singh, Suchchandran Randhawa et al, "Sentimental Analysis of Customer Product Reviews Using Machine Learning Approaches", Journal of Network Communications and Emerging Technologies (JNCET), ISSN: 2395-5317, Volume-5, December 2015.

[6] Manvee Chauhan, Divakar Yadav, "Sentiment Analysis of Customer Product Reviews Using Machine Learning approaches", International Conference on Intelligent Computing and Control (I2C2), June 2017, 2017,8321910.

[7] Alessia, D'Andrea et al. "Approaches, tools and applications for sentiment analysis implementation." International Journal of Computer Applications 125.3 (2015).

[8] Muhammad Marong, Nowshath K Batcha, "Sentiment Analysis in E-Commerce: A Review on the Techniques and Algorithm", Journal of Applied Technology and Innovation, ISSN_Number: 2600-7304, 2020.

[9] Najma Sultana, Pintu Kumar, "Sentiment Analysis for Product Review",ICTACT Journal On Soft Computing, ISSN: 2229-6956, IJSC.2019.0

[10] Tommy Willianto, Suprayadi, "Sentiment Analysis on E-commerce Product using Machine Learning and Combination of TF-IDF and Backward Elimination", International Journal of Electrical and Computer Engineering (IJECE),ISSN: 2277-3878, Volume-8 Issue-6, March 2020, 2862-2867.

[11] Bhanu Sree Reddy, and Uma Pricilda Jaidev. "A Review on the Concept of Sentiment Analysis and its Role in Marketing Strategies for E-Commerce." Iioab Journal 7.5 (2016): 216-224.

[12] Hemalatha, GP Saradhi Varma, and A. Govardhan. "Sentiment analysis tool using machine learning algorithms." International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) 2.2 (2013): 105-109.