

Sentiment Analysis of Dish Review

Amit Kumar Singh, Prof. Amogh Sanzgiri
Goa Engineering College,
Farmagudi, Goa

Abstract - Sentiment analysis is the mining of opinions and feelings from content through Natural Language Processing (NLP). Sentiment analysis is nothing but categorizing opinions in the given content or documents into "positive" or "negative" or "neutral". In this work, the goal is to predict the sentiment of food reviews in two categories, positive and negative. Our analysis could be a useful tool to help restaurants better understand reviewers' sentiments about food, and can be used for other tasks such as recommender systems and better customer engagement. Studying the opinions of customers helps to determine the people's feelings about a product and how it is received in the market.

Keywords-Customer reviews, Machine Learning, Opinion mining, Natural Language Processing[NLP], Sentiment Analysis.

I.INTRODUCTION

Customer satisfaction is the key in assessing how a product or service of a company meets customer expectations [1] and is an important tool that can give organizations major insights into every part of their business, thus helping them to increase earnings or minimize marketing expenses [2]. Nothing can make customers feel that they are important than asking for their views and valuing their comments. When a customer is asked for any opinion on a product or experience, they feel valued and connected to the organization [3]. In the food industry, customers often look into restaurant reviews before placing their orders[3]. The use of AI in natural language processing (NLP) has immense potential to determine positive, negative, and neutral reviews [4]. Machine learning (ML) and deep learning (DL) techniques are often used interchangeably in AI but have different meanings. At a high level, ML automates analytical model building, and [5] DL is the subset of ML (see Figure 1) concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.

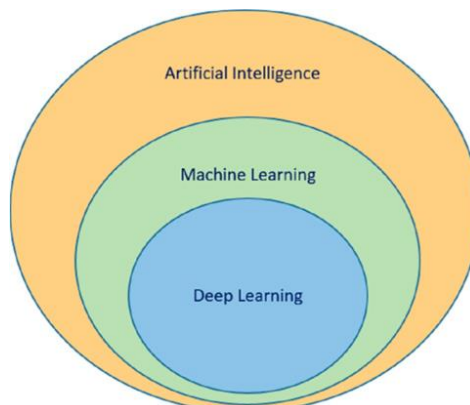


Fig-1: High-level Artificial Intelligence diagram.

If the website does not have any online reviews, then customers may change their decision to order. Having 'no reviews' can be just as detrimental as having negative reviews. Having genuine and positive reviews helps increase the credibility factor. Negative reviews are difficult to handle for any business. They can drive potential customers away from the website and prompt existing customers to question whether they want to re-order. Thus, operators have to remember that they cannot control every customer's experience, mistake, or circumstance. On the bright side, a negative review can provide insights into the customer service's weaknesses and provide opportunities for improvement [6].

The key benefits of sentiment analysis [7] for business are as follows:

- Keeps businesses connected round the clock with the customers.
- Provides business insights to help in decision-making.
- Indicates real-time trends with emotion data.
- Helps improve the business plan of action to gain an advantage over competitors
- Can be conducted on services or products to understand which item is eliciting negative sentiments.
- Provides a great tool for businesses to improve customer service in any domain.

II. PROPOSED SYSTEM

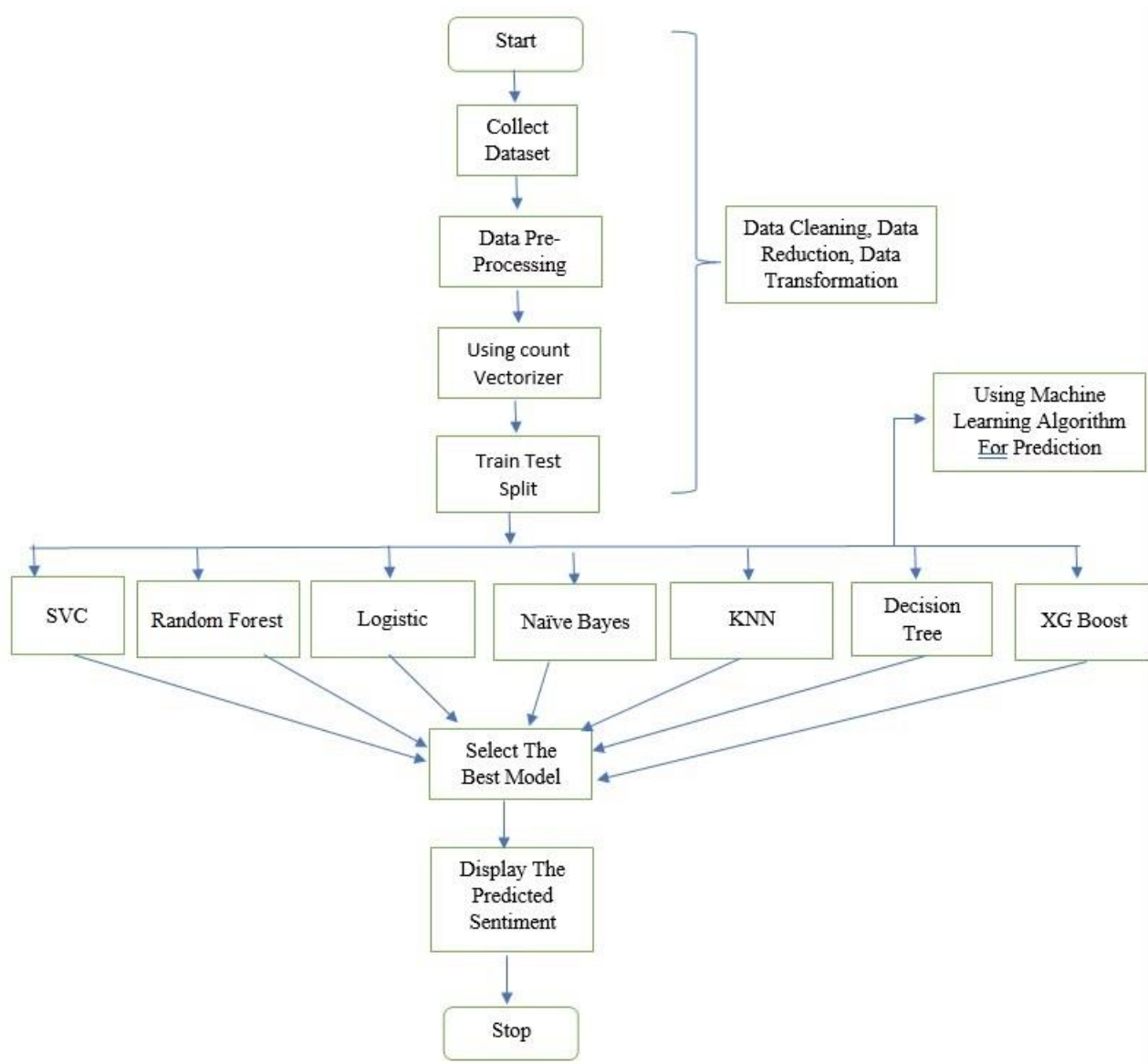


Fig-2: Proposed System Flowchart.

As shown in the above figure, the process starts by collecting the dataset. The next step is to do Data Preprocessing which includes Data cleaning, Data reduction, and Data Transformation. Then, using various machine learning algorithms we will predict the analysis of sentiment. The algorithms involved are SVM, Random Force, Logistics, Naive Bayes, KNN, Decision Tree, and XG Boost. The best model which predicts the most accurate sentiment is selected.

III. CLASSIFICATION ALGORITHM IN SENTIMENTAL ANALYSIS

Sentiment classification is the automated process of identifying opinions in a text and labeling them as positive, negative, or neutral, based on the emotions customers express within them. Sentiment analysis is done using algorithms that use text analysis and natural language processing to classify words as either positive, negative, or neutral. This allows companies to gain an overview of how their customers feel about the brand[1].

Machine learning is a technique to train the system and make the system take decisions of its own. There are various classifiers that are used for the training of the system. They can belong to Supervised, Unsupervised, or reinforcement learning. Supervised classifiers are the classifiers that are given prior training using data input and necessary output, according to which the model learns to predict, and then further classification takes place. Unsupervised learning is the type of learning where the system is given some prior knowledge about the input data, according to which model is created which has to classify the data on its own. Reinforcement learning is the type of machine learning technique that deals with which suitable action to be taken to optimize the

reward in any specific situation. Unlike in a supervised, there is no correct answer given, instead, the decision is made at each task. Figure:3 shows some of the classifiers and the category it belongs to

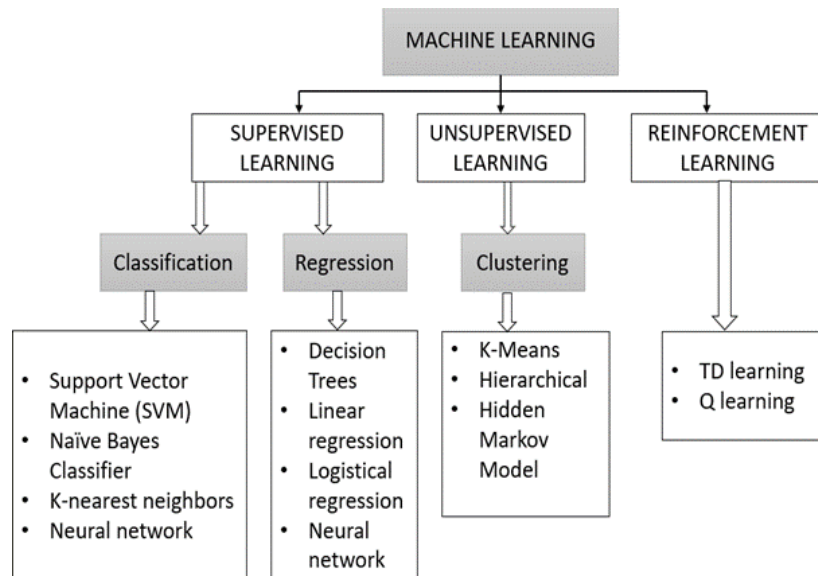


Fig-3: Machine Learning Algorithms

A. Logistic Regression Classifier:

Logistic Regression is a Machine Learning algorithm that is used for classification problems, it is a predictive analysis algorithm based on the concept of probability. It is a classification algorithm used to assign observations to a discrete set of classes. Some of the examples of classification problems are Email spam or not spam, Online transactions Fraud or not Fraud, Tumor Malignant, or Benign. Logistic regression transforms its output using the logistic sigmoid function to return a probability value.

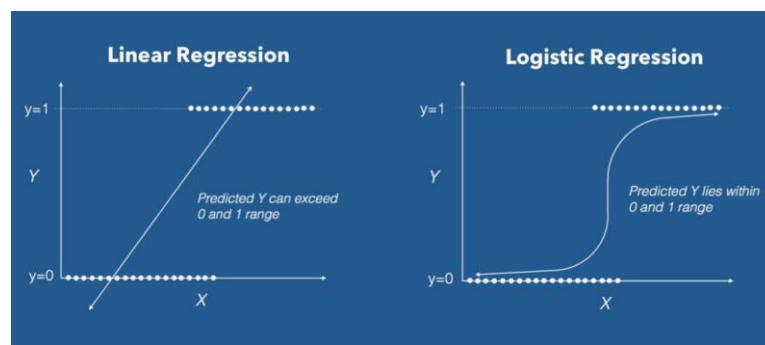


Fig-4: Linear Regression VS Logistic Regression Graph

B. Naive Bayes:

Naive Bayes is the simplest and fastest classification algorithm for a large chunk of data. In various applications such as spam filtering, text classification, sentiment analysis, and recommendation systems, the Naive Bayes classifier is used successfully. It uses the Bayes probability theorem for unknown class prediction. The Naive Bayes classification technique is a simple and powerful classification task in machine learning. The use of Bayes' theorem with a strong independence assumption between the features is the basis for naive Bayes classification. When used for textual data analysis, such as Natural Language Processing, the Naive Bayes classification yields good results.

C. SVM(Support Vector Machine):

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis[2]. SVM is a supervised(feed-me) machine learning algorithm that can be used for both classification and regression challenges. Classification is predicting a label/group and Regression is predicting a continuous value. SVM performs classification by finding the hyper-plane that differentiates the classes we plotted in n-dimensional space.

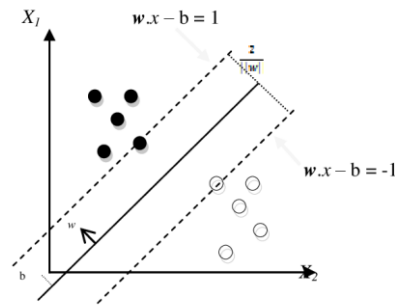


Fig-5: Optimal separating hyperplane between two classes

SVM draws that hyperplane by transforming our data with the help of mathematical functions called “Kernels”.

D. XG-Boost:

XGBoost classifier is a Machine learning algorithm that is applied for structured and tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and boost algorithm. And that means it's a big Machine learning algorithm with lots of parts. XGBoost works with large complicated datasets. XGBoost is an ensemble modeling technique.

E. K- Nearest Neighbor Classifier:

KNN is a straightforward algorithm for predicting a class of a sample. This classifier is supervised learning based on the distance of the sample. The training phase simply saves all training samples with their labels. In order to predict the class for a new test sample, first, it calculates the distance between each training sample and then, keeps the k closest training samples, where $k \geq 1$. Finally, it searches for the label that appears most frequently among these samples. This label is then assigned to this test sample as the predicted result.

F. Decision Tree Classifier:

A Decision tree is a machine learning classifier based on the tree structure. Every node within the tree is related to a specific feature, and the edges from the node separate the data according to the feature's value. Every leaf node binds to a class in the classifier model. The training data is the key point for the information gain (IG) of the feature selection policy. A decision tree simply asks a question and based on the answer (Yes/No), it further split the tree into subtrees.

G. Random Forest Classifier:

A Random Forest classifier is an ensemble algorithm. An ensemble algorithm is the combination of the same or different kinds of algorithms. A set of trees make a forest. Here, Random Forest is a set of decision trees. The voting of each decision tree is taken and the new output case is added to that class that has the highest vote.

III. ADVANTAGES AND DISADVANTAGES OF CLASSIFICATION ALGORITHM

Sr.No.	Algorithm	Advantages	Disadvantages
1	Logistic Regression Classifier	1) Performs properly whilst the dataset is linearly separable.	1) It can be only used to predict discrete functions.
2	Naive Bayes	1) When the assumption of independent predictors holds true it performs better as compared to other models. 2) Easy to implement. 3) Requires a small quantity of data.	1) It assumes that all the attributes are mutually independent hence losing accuracy.
3	Support Vector Machine (SVM)	1) It performs admirably well when there is a clear distinction between classes. 2) It is relatively memory efficient. 3) A small change to the the SVM model is stable.	1)The SVM will underperform when the number of features for each data point exceeds the number of training data samples. 2)Sensitive to noise.

4	K-Nearest Neighbor Classifier	1) Easy to implement. 2) It needs no training before making predictions, new data can be added seamlessly which will not affect the accuracy of the algorithm.	1) It is a lazy learner. 2) It does not learn anything in the training period. 3) Sensitive to noise and missing data.
5	Decision Tree Classifier	1) Its nature is transparent. 2) It allows partitioning of data on a much deeper level.	1) Training is relatively expensive at the complexity and time taken are more. 2) Performs poorly with small data and gives low prediction accuracy.
6	Random Forest Classifier	1) It can produce high dimensional (many features) data without dimensionality reduction and feature selection. 2) Not easy to overfit 3) Fast training speed 4) If a large part of the features is missing, accuracy can still be maintained.	1) Attribute weights of data impact random forests.

Table-1: Advantages And Disadvantages Of Classification Algorithms

IV. DATASET DETAILS

The dataset used for this experiment consists of two columns, the first column consists of the reviews and the second column consists of the reviews given by the users.

'1' indicates a positive review, and '0' indicates a Negative review.

	Review	Liked
0	Wow... Loved this place.	1
1	Crust is not good.	0
2	Not tasty and the texture was just nasty.	0
3	Stopped by during the late May bank holiday of...	1
4	The selection on the menu was great and so wer...	1
5	Now I am getting angry and I want my damn pho.	0
6	Honeslty it didn't taste THAT fresh.)	0
7	The potatoes were like rubber and you could te...	0
8	The fries were great too.	1
9	A great touch.	1

Fig-6: Dataset

V. RESULTS AND DISCUSSION

A. Evaluation Criteria

For classification problems it is common to use a confusion matrix to determine the performance. The confusion matrix for binary classification is built from four terms, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The terms are derived from the predicted class versus the actual class. These are explained below:

True Positive (TP): The expected value matches the actual value. The actual value was positive and the model

predicted a positive value.

False Positive (FP): The expected value was falsely expected. The actual value was negative but the model predicted a positive value,

True Negative (TN): The expected value matches the actual value. The actual value was negative and the model predicted a negative value.

False Negative (FN): The expected value was falsely expected. The actual value was positive but the model predicted a negative value.

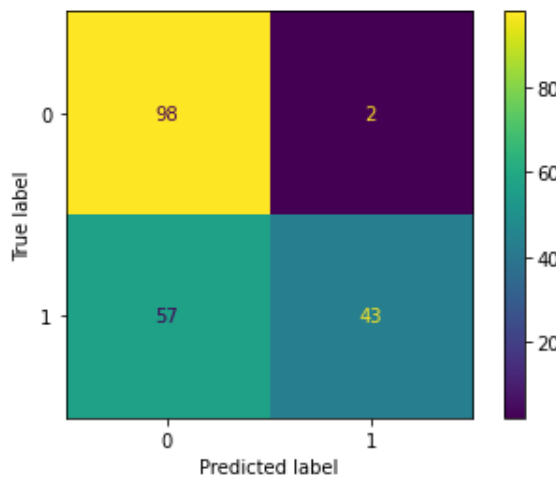


Fig-7: Confusion matrix of Decision Tree model

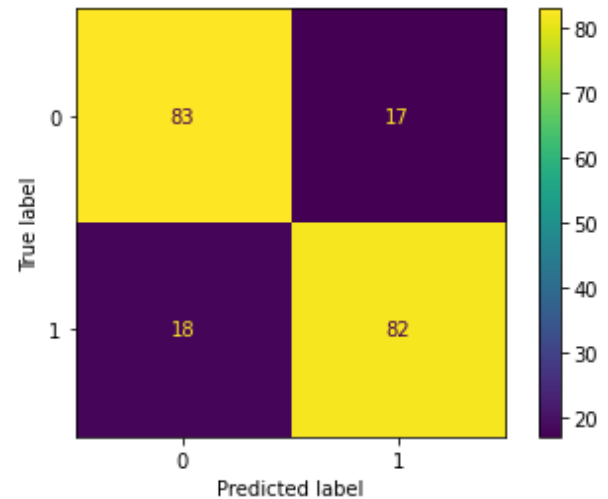


Fig-8: Confusion matrix of Logistic Regression model

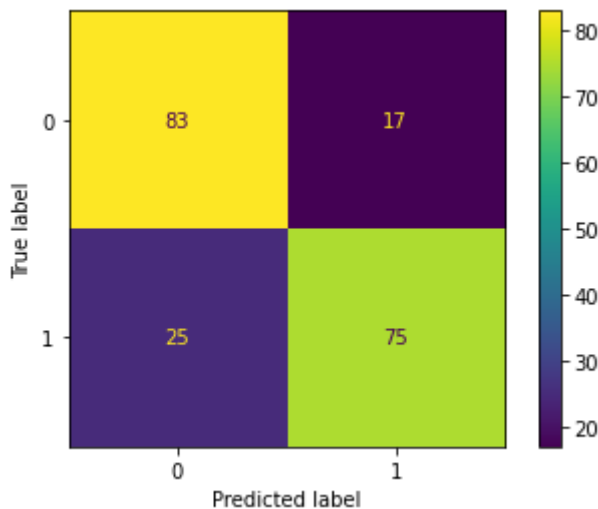


Fig-9: Confusion matrix of XGBoost Classifier Model

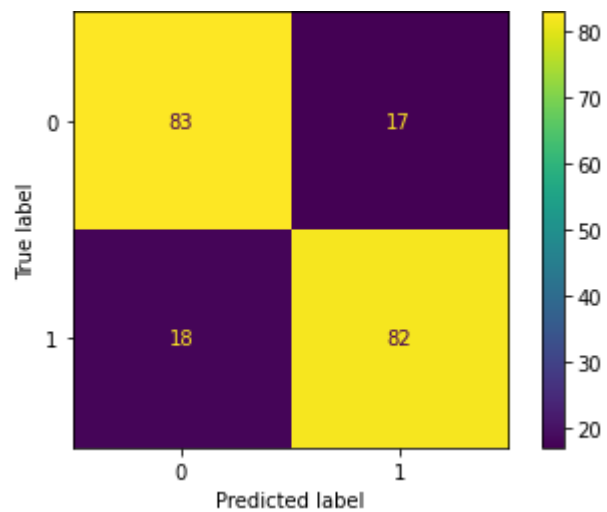


Fig-10: Confusion matrix of SVM classifier model

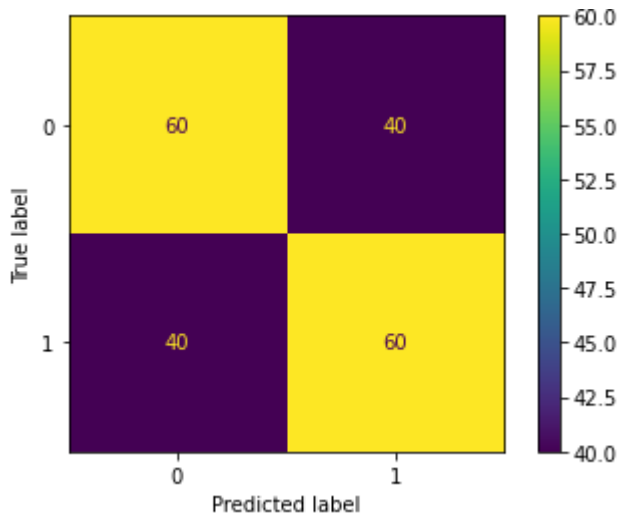


Fig-11: Confusion matrix of KNN classifier model

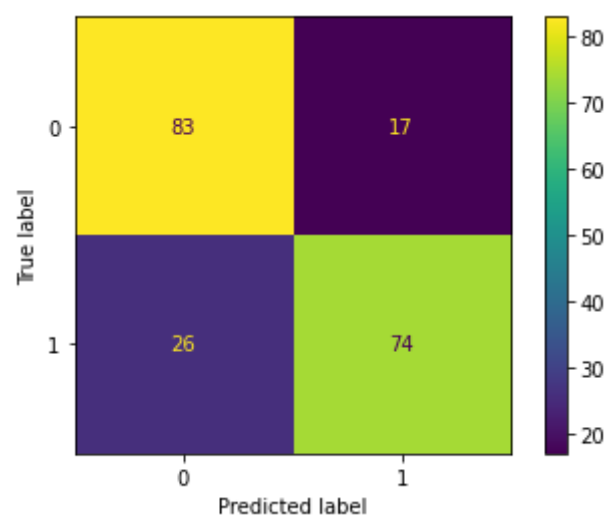


Fig-12: Confusion matrix of Naïve Bayes classifier model

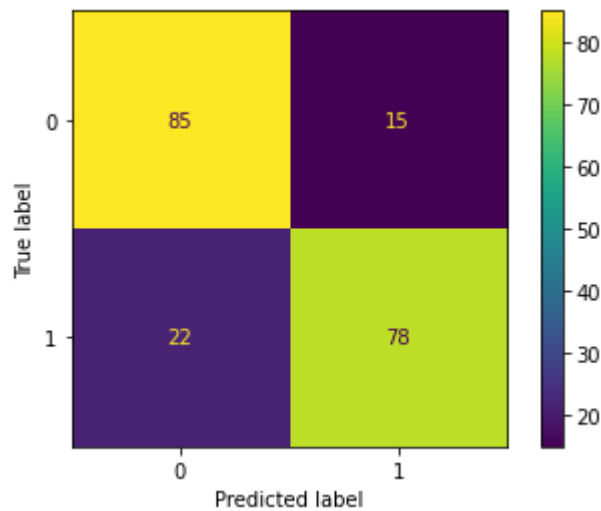


Fig-13: Confusion matrix of Random forest classifier model

B. Result

The accuracy rate of the classifiers implemented is shown in the table below:

	Classifiers	Accuracy
0	Logistic Regression	82.5 %
1	Random Forest Classifier	81.5 %
2	SVM (Linear)	81.0 %
3	XGB Classifier	79.0 %
4	Naive Bayes	78.5 %
5	Decision Tree Classifier	70.5 %
6	K Nearest Neighbors	60.0 %

Fig-14: Accuracy Result

index	Comment	Naive Prediction	Logistic Prediction	RandomForest Prediction	XGB Prediction	DecisionTree Prediction	KNeighbor Prediction	SVC Prediction
0	this dish is very tasty	1	1	1	1	0	0	1
1	i would suggest everyone to visit here atleast once	1	1	1	0	0	0	1
2	Horrible! Don't ever visit again	0	0	0	0	0	0	0
3	I hate this	0	0	0	0	0	1	0
4	I loved the shorma	1	1	1	1	0	1	1
5	this is not good	1	1	0	0	0	1	1

Fig-15: Output Of Classifiers Against User Comments

index	Comment	Predicted Sentiment
0	this dish is very tasty	Positive
1	i would suggest everyone to visit here atleast once	Positive
2	Horrible! Don't ever visit again	Negative
3	I hate this	Negative
4	I loved the shorma	Positive
5	this is not good	Negative

Fig-16: Combining The Results Of Best Three Classifiers To Analyze The Sentiment Of The Comments

CONCLUSION

After doing hyper-parameter tuning on the classifiers model, and training it on the dataset, the best results obtained are from Logistic Regression, followed by Random Forest Classifier. None of the classifiers was able to give accurate results for all the user input comments. In order to overcome this problem, the top three(Logistic, Random, SVM) best-performing algorithms were combined together, and the resultant sentiment was decided to depend upon the combination of individual results obtained from these three algorithms. The results obtained via combined classifiers performed better compared to individual classifiers.

REFERENCES

- [1] Chepukaka, Z.K., Kirugi, F.K, "Service Quality and Customer Satisfaction", at Kenya National Archives and Documentation Service, Nairobi County: Servqual Model Revisited. Int. J. Cust. Relat. 2019, 7, 1.
- [2] Barsky, J.D.; Labagh, R. A," Strategy for Customer Satisfaction", Cornell Hotel Restaur. Adm. Q. 1992, 33, 32–40.
- [3] Suharto, D.; Helmi Ali, M.; Tan, K.H.; Sjahroeddin, F.; Kusdibyo, L. Loyalty toward," Online Food Delivery Service: The Role of E-Service Quality and Food Quality.", J. Foodservice. Bus. Res. 2019, 22, 81–97.
- [4] Geller, Z.; Savi's, M.; Brati'c, B.; Kurbalija, V.; Ivanovic, M.; Dai, W, "Sentiment Prediction Based on Analysis of Customers Assessments in Food Serving Businesses", Connect. Sci. 2021, 33, 674–692.
- [5] LeCun, Y.; Bengio, Y.; Hinton, G," Deep Learning", 2015.
- [6] Hong, L.; Li, Y.; Wang, S," Improvement of Online Food Delivery Service Based on Consumers", Negative Comments. Can. Soc. Sci. 2016, 12, 84–88.
- [7] Nagpal, M.; Kansal, K.; Chopra, A.; Gautam, N.; Jain, V.K," Effective Approach for Sentiment Analysis of Food Delivery Apps", In Advances in Intelligent Systems and Computing; Springer: Singapore, 2020; pp. 527–536.
- [8] Sasikala P, L.Mary Immaculate Sheela, "Sentiment Analysis of Online Food Reviews using Customer Ratings", International Journal of Pure and Applied Mathematics Volume 119 No. 15 2018, 3509-3514.
- [9] Bhanu Chugh, Mayank Pandita Sejal Arya, Tanmay Jain4, G.V.Bhole, "Sentimental Analysis and Visualization of Food Reviews from Zomato using Tableau", International Journal of Advanced Trends in Computer Science and Engineering.
- [10] Hua Feng, Ruixi Lin, "Sentiment Classification of Food Reviews", Department of Electrical Engineering Stanford University.
- [11] Anvar Shathik J.& Krishna Prasad K," A Literature Review on Application of Sentiment Analysis Using Machine Learning Techniques", International Journal of Applied Engineering and Management Letters (IAEM), ISSN: 2581-7000, Vol. 4, No. 2, August 2020.
- [12] Zulfadzli Drugs, Haliyana Khalid," Sentiment Analysis in Social Media and Its Application: Systematic Literature Review", The Fifth Information Systems International Conference 2019.