Sentiment Analysis of Restaurant Reviews Using Multiple Algorithm

Priyanka Afini, Trupti Dilhiwala, Hemangini Mehta

Abstract— ***This project focuses on the application of machine learning algorithms—Naive Bayes, SVM, KNN, and Random Forest in the context of sentiment analysis for restaurant reviews. The objective was to discern the most effective algorithm for accurately categorizing sentiments expressed in customer feedback. Through meticulous implementation and thorough evaluation, it was determined that the Naive Bayes algorithm consistently outperformed its counterparts, showcasing superior performance in handling the intricacies of restaurant reviews and sentiment classification. Following this comprehensive analysis, a dedicated model was developed utilizing the Naive Bayes algorithm, leveraging its simplicity and efficiency in handling text-based data. The resulting sentiment analysis model offers a robust solution for extracting valuable insights from restaurant reviews, aiding restaurant owners in understanding customer sentiments and making informed decisions.***

Index Terms— Naive Bayes, SVM (support vector Machine), KNN (k-nearest neighbor), Random Forest.

I. INTRODUCTION

Customer satisfaction lies at the intersection of consumer preferences and reality [1]. Presently, attitude of consumers expresses their opinions through online platforms like Twitter, sharing reviews. The significance of customer reviews on social networks has grown as they can enhance the visibility of a seller's product or service. Social network research is commonly categorized into web-based content mining, structure mining, and usage mining. Web content mining involves analyzing the content generated by users on the social web, contributing valuable insights into consumer sentiments and preferences.

NLP stands for Natural Language Processing, which is a field of artificial intelligence (AI) that focuses on the interaction between computers and humans through natural language [1]. Applications of NLP are widespread and include customer support Chatbots, language translation services, sentiment analysis in social media, voice-activated virtual assistants, and clinical services reviews. [2]

Manuscript received July 11, 2024

Priyanaka Afini, Computer Enginnering, Bhagwan Mahavir Collage of Engineering And Technology, Surat, India

Trupti Dilhiwala, Computer Enginnering, Bhagwan Mahavir Collage of Engineering And Technology, Surat, India

Hemangni Mehta, Computer Enginnering, P.P.Savani University, Kosamba, Surat, India

A crucial focus within certain fields is sentiment analysis, also known as opinion mining. This study area can be further categorized into subjectivity analysis and sentiment analysis. Sentiment analysis finds extensive application in social web research, particularly in pattern recognition and identifying influential review sentiments. This facet has become a significant realm of social research, aiding in identifying key individuals who contribute valuable insights for informed decision-making across diverse domains. The importance of opinion mining within data mining is heightened by these multifaceted issues and applications. As internet data continues to grow, efficient data processing, especially for sentiment analysis and complex natural language tasks, becomes imperative. Machine learning and statistical approaches commonly handle these tasks.

Main objective of this research project is to compare and Classifying restaurant reviews using various NLP algorithms techniques which is useful to provide a valuable insight to the businesses so that they can improve their products and services.

II. LITERATURE REVIEW

Research was using content and sentiment analysis of TripAdvisor datasets scrapped using Webscarpper software. They have studied using two different datasets which are pre and post Covid-19 for quality of service and products. it is a general behavior of customer for most of restaurants and trust in outside food is declined so it cannot predict the positive or negative impact perfectly [3].

Sentiment analysis using characterization into classifications lexicon based and Machine Learning (ML) and Deep Learning (DL) methods. Lexicon based algorithm used a data dictionary to consider tag words into positive and negative. DL methods such as RNN, CNN, and LSTM provides good performance in terms of accuracy [4].

Research approach by using restaurant reviews of the Kaggle dataset observed that SVM model with 70% and 30% training and testing data respectively given a highest accuracy of 77% for the dataset [5].

Yelp reviews are very useful to find nearest restaurants for quality food and review of yelp can be well monitored so accuracy of data is good enough to study sentiment analysis. observed that SVM is generating acceptable outcome by comparing a model to other models like Naive Bayes and KNN [6].

Hybrid method was based on combining various classification methods using arcing classifier and performances of these classifiers are analyzed with keeping in the mind accuracy as a main factor. This method was a combination of using Naïve Bayes (NB), Support Vector Machine (SVM) and Genetic Algorithm (GA) [7].

Research found and used a combined CNN-LSTM architecture for sentiment analysis. The experiment completed with the collected data from websites to create a dataset. Researcher suggested that explore other pre-trained word vectors (such as BERT) [8].

Literature focused on analyzing sentiments expressed in Twitter unstructured data transforming into structure data or valuable data for sentiment analysis. They use 14640 tweets of Six Airlines as data with 15 Attribute. They achieve 75% accuracy and suggested that explore more algorithm like SVM, KNN and logistic Regression [9].

III. PROPOSED WORK

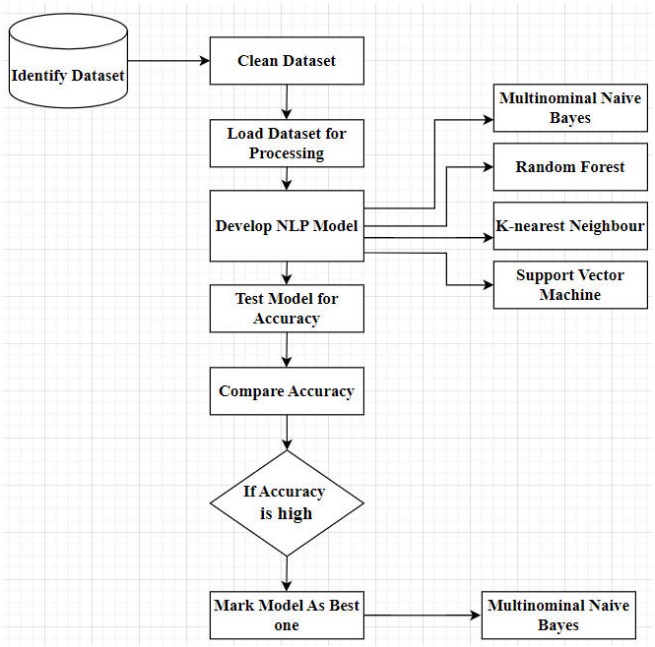

Figure 1 proposed work

IV. METHODE

We will exploit the typical machine learning pipeline. We will first import the dataset. we will perform text preprocessing to convert textual data to numeric data that can be used by a machine learning algorithm like Naïve Bayes, SVM, KNN and Random Forest. Finally, we will use machine learning algorithms to train and test our sentiment analysis models.

A. *Dataset*

After identifying and downloading Kaggle dataset for restaurant reviews in Comma Separated Values (csv) format. dataset will be loaded into the data frame for reading using Panda library in the python.

B. *Preprocessing*

After loading dataset drop the unnecessary column. Text data often contains noise, including irrelevant characters, symbols, or formatting issues. Cleaning helps remove such noise, allowing the sentiment analysis model to focus on meaningful patterns in the text. Sentiment analysis focuses on meaningful words that contribute to the sentiment expressed. We can set X is the input feature (Review) that we give to the model, and Y(Rating) is the output that the model should predict. Remove the missing value by replace Null value with median values of data. Cleaning often involves converting all text to lowercase.

Tokenization, Sentiment analysis relies on breaking down text into smaller units, or tokens. Cleaning ensures that the tokenization process is accurate and that each token represents a meaningful unit of information. Removing irrelevant characters and symbols aids in proper tokenization.

Stemming or lemmatization, which involves reducing words to their base or root form. This process helps in capturing the essence of words and simplifies the analysis by treating different forms of a word as the same. Cleaning ensures consistency in the dataset, making it more reliable for training and testing sentiment analysis models.

C. *Confusion Matrix*

The confusion matrix is typically used in binary classification problems, where there are two classes or categories: one is the "positive" class, and the other is the "negative" class [35].

The confusion matrix can be used to calculate various performance metrics for the classification model, such as accuracy, precision, recall, and F1 score. We need accuracy for comparison of various algorithms so we can use following formula [12].

Accuracy = (TP + TN) / (TP + TN + FP + FN)
Where TP = True Positive, TN= True negative

D. *Multinomial Naive Bayes*

It is a probabilistic algorithm commonly used for classification tasks, particularly in natural language processing (NLP) and text mining. It's a variant of the Naive Bayes algorithm. The input data consists of documents and their corresponding class labels. Each document is represented as a feature vector, where each feature corresponds to a unique term or word in the entire corpus.

In the context of text classification, it calculates the probability of each word occurring in a document given a specific class. The algorithm also calculates the prior probability of each class, representing the likelihood of a document belonging to a particular class based on the overall distribution of classes in the training data [10].

$$P(\text{Class}|X) = (P(X|\text{Class}) \cdot P(\text{Class}))/P(X)$$

P(Class|X) is the posterior probability of the class given the features.
P(X|Class) is the likelihood of the features given the class, which is estimated during training.
P(Class) is the prior probability of the class, also estimated during training.

P(X) is the probability of the features, which acts as a normalization constant and can be ignored for classification purposes

E. Support Vector Machine

Support Vector Machines is a supervised machine learning algorithm commonly used for classification and regression tasks. SVM aims to find a hyperplane that best separates data points belonging to different classes in a high-dimensional space.

In a Support Vector Machine (SVM), the decision boundary is represented by a hyperplane that aims to separate data points belonging to different classes. The standard linear SVM formulation involves finding a hyperplane with the maximum margin. Here's the basic formula:

Given a set of training data (X1, Y1), (X2, Y2) …… (Xn, Yn) in which
Xi is the feature vector for specific point
Yi is the class label which is either +1 or -1.
$f(x)=\text{sign}(\mathbf{w}\cdot\mathbf{x}+b)$
Here, w is the weight vector, x is the input feature vector, and b is the bias term.
$yi\cdot(\mathbf{w}\cdot\mathbf{x}i+b) \geq 1$

This constraint ensures that each data point lies on the correct side of the decision boundary with a margin of at least 1.

F. K-Nearest Neighbor

In the context of sentiment analysis, KNN classifies a text by considering the sentiments of its nearest neighbors. The "neighbors" are other pieces of text in a labeled dataset that are most similar to the input text in terms of features or characteristics. KNN offers a straightforward approach to sentiment analysis by leveraging the concept of similarity among data points. It's a useful algorithm for sentiment analysis tasks, especially in scenarios where interpretability and simplicity are prioritized [11].

G. Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees during training and combines their predictions for improved accuracy and generalization. Each tree in the Random Forest is a decision tree, which is a flowchart-like structure where each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents the final decision. Train the Random Forest model using the preprocessed text data and their corresponding sentiment labels [9].

The model will consist of multiple decision trees, each trained on a different subset of the data. Given a new text, the Random Forest aggregates predictions from each decision tree to arrive at a final sentiment prediction. This can be achieved through voting mechanisms.

The Random Forest algorithm is a powerful tool for sentiment analysis in NLP, leveraging ensemble learning and decision trees to effectively capture complex patterns in textual data and deliver accurate predictions.

V. RESULT AND DISSCUSION

Here we can train and test different machine learning algorithm to find the highest accuracy.

Result of Accuracy for Naïve Bayes algorithm with different Train and test dataset

Train Data %	Test Data %	Accuracy
70	30	90.6%
75	25	90.8%
80	20	90.9%

Table 1 Result of Naïve Bayes algorithm

Result of Accuracy for Random Forest algorithm with different Train and test dataset

Train Data %	Test Data %	Accuracy
70	30	89.6%
75	25	89.5%
80	20	89.4%

Table 2 Result of Random Forest algorithm

Result of Accuracy for KNN algorithm with different Train and test dataset

Train Data %	Test Data %	Accuracy
70	30	79.56%
75	25	80.01%
80	20	82.05%

Table 3 Result of KNN algorithm

Result of Accuracy for SVM algorithm with different Train and test dataset

Train Data %	Test Data %	Accuracy
70	30	89.13%
75	25	89.24%
80	20	89.65%

Table 4 Result of SVM algorithm

If we consider all the result received by various algorithms, we can observe that for the sentiment analysis of restaurant review using Kaggle dataset for training Multinominal Naïve Bayes is giving the best accuracy in all the algorithm. So, model can be prepared for the Multinominal Naïve Bayes for classification and sentiment analysis of the review automatically.

Method	Split Training /Testing %	Highest Accuracy received
Naïve Bayes	80-20	90.9%
Random Forest	80-20	89.4%
KNN	80-20	82.05%
SVM	80-20	89.65%

Table 5 Accuracy of different Algorithm

VI. CONCLUSION

We can conclude that the examination of various machine learning algorithms, namely Naive Bayes, SVM, KNN, and

Random Forest, for sentiment analysis of restaurant reviews has yielded valuable insights. The goal of this investigation was to identify the most effective algorithm for discerning sentiment in the context of restaurant feedback. After meticulous implementation and evaluation, it became evident that the Naive Bayes algorithm emerged as the optimal choice, consistently delivering superior results compared to the other algorithms.

REFERENCES

[1] S. Kannappan, "SENTIMENT ANALYSIS USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING, " Shu Ju Cai Ji Yu Chu Li/Journal of Data Acquisition and Processing 38(2):520-526, April 2023. [Online].

[2] Y. &. T. A. &. S. S. &. Z. R. Wang, "Applications of Natural Language Processing in Clinical Research and Practice," 2019. [Online].

[3] J.-N. &. T. G. &. A. D. Harba, "Exploring Consumer Emotions in Pre-Pandemic and Pandemic Times.A Sentiment Analysis of Perceptions in the Fine-Dining Restaurant Industry in Bucharest, Romania," International Journal of Environmental Research and Public Health, 2021. [Online].

[4] B. P. a. N. S. Anirban Adak, "Sentiment Analysis of Customer Reviews of Food DeliveryServices Using Deep Learning and Explainable ArtificialIntelligence: Systematic Review," 2022. [Online].

[5] S. S. T. G. Tanya Bhatt, "Restaurant Review Analysis using NLP," International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 9 , ISSN: 2321-9653, 2021. [Online].

[6] L. I. S. Sasikala P, "Sentiment Analysis of Online Food Reviews using," International Journal of Pure and Applied Mathematics,Volume 119 No. 15, 2018. [Online].

[7] T. S. Haja Sharieff, "Comparison of Machine Learning Techniques for Sentimental Analysis on Restaurant Reviews," International Journal of Advances in Engineering and Management (IJAEM) Volume 2, Issue 6, pp: 740-743 ISSN: 2395-5252, 2020. [Online].

[8] M. R. B. Z. N. T. Naimul Hossain, "Sentiment Analysis of Restaurant Reviews using Combined CNN-LSTM,"IEEE, 2020. [Online].

[9] Bahrawi, "SENTIMENT ANALYSIS USING RANDOM FOREST ALGORITHMONLINE SOCIAL MEDIA BASED," JOURNAL OF INFORMATION TECHNOLOGY AND ITS UTILIZATION, VOLUME 2, ISSUE 2, DECEMBER-2019. [Online].

[10] R. &. O. S. &. N. L. Olanrewaju, "Multinomial Naïve Bayes Classifier: Bayesian versus Nonparametric Classifier Approach.," 2022.

[11] S. &. P. J. Bhardwaj, "Sentiment Analysis Approach based N-gram and KNN Classifier," 2019. [Online].

[12] J. Brownlee, "Confusion Matrix in Machine Learning," 2020. [Online]. Available:https://machinelearningmastery.com/confusion-matrix-machine-learning/. [Accessed 2 11 2023].