2023 6th International Conference on Recent Trends in Advance Computing (ICRTAC) | 979-8-3503-9470-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICRTAC59277.2023.10480757

Sentiment analysis on Zomato customer reviews using Random Forest Classifier

1st Nousheen Begum
Department of CSE,
B V Raju Institute of Technology
Narsapur, India
21211a05k3@bvrit.ac.in

2nd M. Ruthika
Department of CSE,
B V Raju Institute of Technology
Narsapur, India
21211a05f9@bvrit.ac.in

3rd N. Lakshmi Deepika
Department of CSE,
B V Raju Institute Of Technology
Narsapur, India
21211a05j2@bvrit.ac.in

4th M. Susanna Sucharitha
Department of CSE,
B V Raju Institute Of Technology
Narsapur, India
21211a05h1@bvrit.ac.in

5th V. Punna Rao
Department of CSE
B V Raju Institute Of Technology
Narsapur, India
punnarao.v@bvrit.ac.in

***Abstract*—Sentiment analysis which is also referred as Opinion mining is the process of determining emotion behind a text or message. Zomato, a prominent player in food and restaurant industry and has become a best choice for gauging customer's sentiments and preferences. It hosts a plethora of customer reviews that provide valuable insights into the food experiences of its users. This paper presents a comprehensive analysis of Zomato customer reviews which utilizing a Random Forest classifier to discern sentiments associated with restaurant visits. And its primary objective is to develop an effective sentiment analysis model capable of categorizing customer reviews into positive, negative, or neutral sentiments accurately. In order to gain these objectives, initially, we leverage a dataset from Kaggle Hyderabad restaurants, comprising a diverse range of customer reviews. Then we employ natural language processing (NLP) techniques like TF-IDF to preprocess and extract valuable features from textual data. Followed by, splitting the dataset into training set(70%) and testing set(30%). Subsequently, a Random Forest classifier is trained using the training set, harnessing the power of ensemble learning to enhance sentiment classification accuracy. Finally, the model is evaluated by the metrics(Accuracy, Precision, Recall). The results obtained are average accuracy of 93% and average precision of 93%, recall of 87%, showcases the effectiveness of the Random Forest classifier in accurately categorizing customer sentiments within the context of Zomato reviews. This helps us provide insights into the factors that influence sentiment, helping restaurant owners and managers understand the key drivers behind customer's contentment. Additionally, this paper explores the implications of sentiment analysis in the context of restaurant recommendations, enabling personalized dining suggestions for users based on their preferences.**

***Index Terms*—Sentiment analysis, Random forest Classifier, opinion mining, polarity, Natural Language Processing(NLP).**

I. Introduction

Sentiment interpretation, is the process of extracting and comprehending the sentiment expressed in a text document. The user's viewpoint is expressed in the form of textual data. Investigating and assessing an opinion's sentiment is a critical undertaking[1]. Analysis is one of the most active study areas, and it is also frequently explored in the field of data mining. Sentiment analysis is widely employed in practically all economic and social arenas since opinions are at the heart of most human actions and behaviours. Sentiment analysis is in high demand due to its effectiveness.

978-1-6654-6483-3/23/$31.00 ©2023 IEEE

Zomato is a website for finding and searching restaurants. It currently conducts business in 23 nations, including the US, Australia, and India. It shows user reviews and ratings in addition to restaurant details like menus and photos taken by past patrons[2]. The customers mostly check reviews and ratings to order any product online, this makes emotion analysis a key point in the field of study.

Food ordering and delivery (FOD) is a novel business concept that has resulted in the launch of various online enterprises in the e-commerce age. Online meal ordering and delivery is successful because it bridges the gap between the establishment and the customer. A consumer searches for a restaurant, filters the available items and cuisines, and orders food through a mobile phone application[3]. The online meal ordering/delivery system is dependent on the mobile application and operates based on the location of the customer. To comprehend how computers and natural (human) languages interact, researchers employ a branch of artificial intelligence (AI) and machine learning called natural language processing (NLP). Sentiment Analysis uses NLP which evaluates peoples's writing and to decide whether they like or dislike something.[3]

Sentiment Analysis is divided into the following stages:

- **Pre-Processing Phase:** Cleaning of data to reduce noise.
- **Feature Extraction:** The keywords are assigned a token, which is then analysed.
- **Classification Phase:** These keywords are classified

based on various algorithms.

The primary goal of this article is to investigate how opinion mining may be used to the history of express users in order to obtain their review aspect and matching sentimental propensity. We present an opinion mining system for online reviews that employs review sentences as input in this study. Businesses can obtain insight into their consumers' opinions and experiences with their products by analysing the sentiment of product reviews, which can be utilised to improve and adapt their product design and marketing strategy. Machine model is trained huge amount of data and labeled with the right sentiment. The machine learning models learn patterns from the data and use them to automatically classify new reviews.

Sentimental analysis benefits in product reviews and is advantageous in business realm. First, We can visualize the results using things like word clouds or bar charts or scatter plots[5]. By using these visualizations, companies can better understand how sentiment is distributed across reviews and determine the most popular positives and negatives of the product. Second, sentiment analysis helps business administrations to identify their problems and improve their product. By analyzing how customers feel about different aspects of your product, you can determine which features and elements of your product customers find most attractive and which areas need improvement. This information can be used to make appropriate changes to desired product.

In our study, we aim to examine the sentiment of customers towards restaurants featured on this food platform, with a particular focus on reviews of restaurants in Hyderabad. To achieve this, we employ the Random Forest Classifier to gain insights from our analysis and present a comprehensive case study of our discoveries. The outcomes of this research could be valuable for restaurant owners in adapting to customer preferences, meeting demand, and making informed decisions about pricing specific food items.[6]. In this paper, Literature survey, problem statement, methodology, result, discussion and conclusion are followed in the next sections respectively.

II. Literature Survey

Sentiment analysis is useful for determining whether or not users enjoy something. Zomato is a restaurant review application. The rating includes a restaurant review, which can be used for sentiment analysis. The review is pre-processed by changing all words to lowercase letters, tokenizing than eliminating numerals and punctuations from the review, stopping words are then removed. Random forest classifiers are defined by criteria such as precision, recall, and accuracy. Linear SVC, Sentiment analysis is the computational identification and categorization of both favourable and adverse views from a piece of text in the context of information retrieval. With such a massive number of data, it is necessary for data warehouses and relational databases to evaluate and use adequate data for their functionality, which has become a big concern. Spark is the most efficient huge data framework, exceeding other optimisations such as Hadoop[2].

These processes involve gathering datasets via various visualisations, data preparation, feature extraction via Spark MLLIB, and finally, model evaluation via train test split using multiple binary classification metrics. Various demos were done out for analysing sentiments with large datasets using Spark MLLIB and various classification approaches such as NB, Linear SVC, and LR. Sentiment analysis of data is particularly valuable for expressing the viewpoint of the masses, groups, or individuals. In the form of messages, blog posts, status updates, postings, and so on, social media and other online communities contain huge amounts of data[3] .In this paper, we used many techniques to examine movie reviews, including Nave Bayes, K-Nearest Neighbour, and Random Forest. They proposed using a combination of grid search and grading to calculate the average sentiment ratings from the Rhetoric Structure Theory (RST) tree. The Navie Bayes, K-Nearest Neighbour, and Random Forest algorithms were applied to the data set. The observations and results are presented in the tables below[3].

The article proposes an opinion mining structure to analyse reviews on the internet that takes comment phrases as input and completes two separate sorts of analysis tasks: sentiment classification and clustering analysis. Identifying emotive phrases and giving them greater weight can improve sentiment categorization accuracy. We take ED-TextRank (TextRank Based on Sentiment Dictionary) attributes and extract them from review sentences[4]. This study's main objective is to detect review elements and accompanying emotive tendencies by applying opinion mining research to previous Express user reviews. In this essay, a framework concept is presented for handling express businesses' internet reviews. There are major two key works in this area. First, it proposes a word2vec-based VC-Word2vec algorithm to partition reviews. Second, this study enhances the traditional TextRank method for categorising the divided reviews into various emotive domains[4]. Numerous algorithms exist in the sentiment analysis sector to address NLP issues. Additionally, this study demonstrates that Support vector machines (SVM) perform with higher accuracy than Naive Bayes and Maximum Entropy approaches. In this study, it is found that sentiment analysis, also known as opinion mining, is crucial when deciding whether or not to purchase a particular good or service[5].

This study aims to use sentiment analysis on an Amazon mobile phone dataset to predict review polarity. It explores the effectiveness of modern NLP models like BERT for handling complex categorization issues. Multiple machine-learning models, including Logistic Regression, Naive Bayes, Random Forest, and Bi-LSTM, are applied with various feature extraction techniques. The study reframes the problem as binary classification by removing the neutral class and re-training the top-performing model. The investigation involves multiclass and binary classification, combining GloVe and Bi-LSTM, and using BERT for representation[6]. Thorough explanations are provided for each stage of sentiment polarity categorization and POS analysis, focusing on pre-processing, pre-filtering, biasing, and data correctness. The study uses product reviews from Amazon.com, considering star ratings to determine sentiment. It collects all subjective content for analysis instead of just the objective content, defining a sentimental statement as having at least one positive or negative word. Part-of-speech (POS) tagging is carried out using the max-entropy POS tagger from the Penn Treebank Project [7].

In this article, the usage of Word2Vec model as features in SVM-based sentiment analysis of Indonesian product reviews is explored. The process involves three steps: preprocessing, Word2Vec model construction, and SVM classification. Preprocessing includes case folding, tokenization, and cleaning. Word2Vec is used to create word vector representations, which are then employed for categorization using SVM. The method shows lower accuracy compared to other techniques due to limited training data for Word2Vec, indicating the need for more examples to improve word representation and accuracy [8].

The proposed method consists of two primary steps: establishing the model and utilizing the constructed model to deliver real-time analytics in an e-commerce application, performing sentiment analysis on product reviews. Data preprocessing involves tokenization, stop word elimination, adding POS tags, and stemming. Negative phrase identification is used to evaluate product qualities in online e-commerce. The paper distinguishes between two categories of phrases: negation-of-adjective (NOA) and negation-of-verb (NOV), using Support Vector Machines (SVM) as the chosen categorization model. The SVM machine learning technique is employed to classify customer reviews as positive or negative, achieving high accuracy as assessed by recall, accuracy, F1, and ROC AUC [9].

This paper proposes a sentiment polarity categorization technique for a substantial collection of internet reviews of Instant Videos, classifying the polarity into five classes (Strongly Positive, Positive, Neutral, Negative, and Strongly Negative). The pre-processing stage involves sentence verification, tokenization, removal of white spaces, stop words, emotions, html tags, new line tags, and special symbols. It uses Natural Language Tool Kit (NLTK) for review tagging. Lexicons like WordNet provide numerical emotion ratings for negativity, positivity, and objectivity, helping calculate sentence and review scores. The final review ratings are categorized into a 5-star rating category. Various classifiers like Naive Bayes, Gradient Boosting, Decision Tree, Support Vector Machine, Random Forest, and Sequence to Sequence Recurrent Neural Network have been employed in the scientific community for automatic classification, opening up room for further development. However, automated sentiment analysis still faces limitations, including handling other styles like sarcasm [10].

III. Problem Statement

- The authors addresses the task of accurately classifying sentiments (positive, negative, neutral) expressed in Zomato customer reviews, which are important for guiding restaurant decisions and enhancing user experiences
- This paper leverages the Random Forest classifier and natural language processing techniques to develop a robust sentiment analysis model capable of providing valuable insights from Zomato reviews, thereby supporting data-driven decisions in the restaurant industry.
- Because unstructured text is inherently complicated, it can be difficult to analyze restaurant reviews. To achieve high classification accuracy, it is essential to take into account linguistic nuances, context-dependent attitudes, and the wide range of terminology used by reviewers.

IV. Proposed Work and methodology

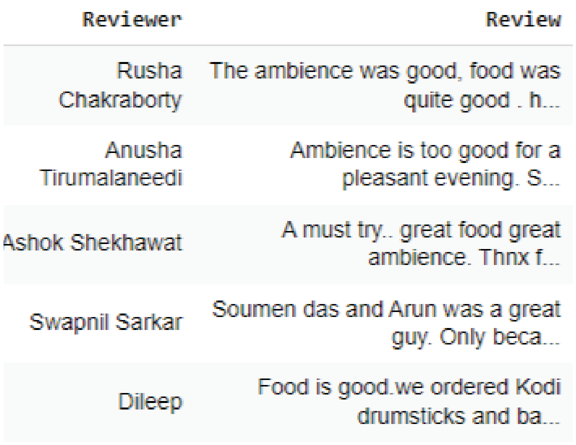
This paper takes customer reviews data from Kaggle, Hyderabad Zomato customer reviews. This data is processed and Random forest classifier in Scikit-learn is used to perform sentimental analaysis. It is analysed using accuracy and precision-recall.

A. Data Collection

The authors utilized a dataset from Kaggle containing customer reviews focused on restaurants in Hyderabad. Kaggle, a trusted platform for datasets, provided valuable real-world customer feedback from a diverse range of dining establishments in the city. Providing an understanding of the attitudes and preferences of diners, this information served as the basis for our sentiment analysis. Data quality was ensured through thorough review, establishing a reliable base for our analysis.

B. Data Preprocessing

In order to get the raw text data ready for sentiment analysis, we went through a number of preprocessing steps. First, we got rid of any irrelevant data, like URLs, special characters, that don’t really help sentiment analysis. The reviews were then broken up into individual words or phrases using tokenization, allowing for more in-depth analysis. We used stop-word removal methods to improve the textual data’s quality and consistency. The stop words like ”the,” ”and,” or ”is” are removed because they do not contain significant sentiment-related information. The Natural Language Toolkit (NLTK) library successfully removes these stop words. Additionally, we reduced words to their base or root form using stemming or lemmatization methods. This procedure ensures that words with similar meanings are consistently represented and reduces word variation. Again, NLTK was used to implement stemming or lemmatization, which helped the reviews have a more consistent representation of words.



Reviewer	Review
Rusha Chakraborty	The ambience was good, food was quite good . h...
Anusha Tirumalaneedi	Ambience is too good for a pleasant evening. S...
Ashok Shekhawat	A must try.. great food great ambience. Thnx f...
Swapnil Sarkar	Soumen das and Arun was a great guy. Only beca...
Dileep	Food is good.we ordered Kodi drumsticks and ba...

Fig. 1. Data set collected

C. Random Forest Classifier

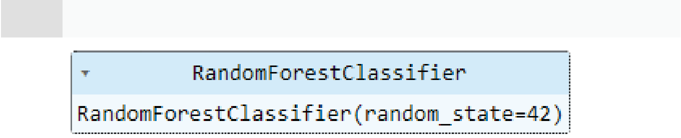
Following preprocessing, the reviews' textual content is transformed into numerical features using techniques like TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF assigns weightage to words based on their frequency within a specific review relative to their frequency across the entire dataset. This feature representation allows the classifier to capture the importance of words or phrases in determining sentiment.

Once the features are prepared, a Random Forest classifier is initialized and trained on the transformed data. The Random Forest model, an ensemble learning method, leverages a collection of decision trees to collectively make predictions. Each tree in the forest examines different aspects of the reviews' feature space, and through a voting mechanism, the forest combines their outputs to arrive at a final sentiment prediction.

In practice, this means that when a new review is introduced to the trained Random Forest classifier, the model assesses the importance of each word or phrase in the review within the context of the dataset. By aggregating these insights from multiple decision trees, the classifier generates a sentiment prediction, categorizing the review as positive, negative, or neutral based on the accumulated evidence from the reviews in the training dataset. This methodology enables automated sentiment analysis, aiding in understanding customer sentiments and guiding decision-making processes for businesses in various domains, including the restaurant industry.

V. Result

We built a machine model using 80% of the training data and leaving the rest 20% of the data to serve as the testing dataset to assess the model's performance. The Random Forest Classifier was employed for prediction. Following the predictions, we evaluated the model's performance using various metrics, including accuracy and the generation of a confusion matrix. The accuracy metric quantifies the ratio of correctly predicted labels to the total number of instances in the test set. Meanwhile, the confusion matrix is a tabular representation that presents the counts of true positive, true negative, false positive, and false negative predictions.



```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

Fig. 2. Example of a figure caption.

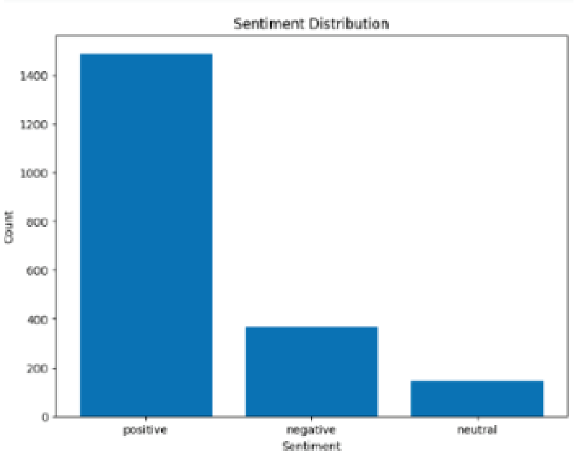


Fig. 3. Sentiment distribution in the dataset.

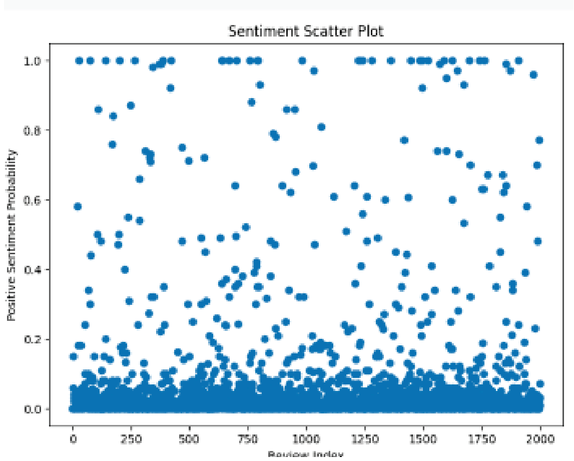


Fig. 4. Sentiment Scatter Plot.

TABLE I
Confusion Matrix

value	Precision	Recall
Positive	93	99
Negative	92	89
Neutral	96	73
Average	93	87

A. Accuracy

This machine model yields outcomes with an accuracy of 93.6%. Twenty percent of the statistics explain it.

B. Confusion Matrix

As seen in the table, the confusion matrix is a statistic that displays the true and prediction errors of data derived from algorithmic outcomes.1. Table 1:confusion matrix

TABLE II
Confusion Matrix

value	Positive	Negative	Neutral
Positive	166	6	196
Negative	6	101	37
Neutral	15	8	1465

VI. Discussion and future Work

In this paper authors applied a Random Forest classifier to perform sentiment analysis on Zomato customer reviews, aiming to uncover valuable insights into customer sentiments towards restaurants and dining experiences. The results demonstrated the effectiveness of this approach, with a high accuracy rate in classifying sentiments as positive, negative, or neutral. Our research findings indicate that sentiment analysis can provide essential information for both restaurant owners and Zomato platform administrators to understand customer satisfaction and make data-driven decisions for service improvement. Furthermore, the use of Random Forest showed promising results, indicating its suitability for sentiment analysis tasks in the context of restaurant reviews.

Sentiment analysis employs different approaches, such as Random Forest and Deep Learning models, each with unique strengths and considerations. Random Forest, an ensemble learning technique, constructs decision trees and utilizes feature engineering, like word frequency and TF-IDF scores, providing decent interpretability by showcasing feature importance. On the other hand, Deep Learning models, like RNNs or CNNs, excel in learning intricate patterns from raw text, offering superior performance, especially with ample data. However, they require substantial computational resources, lack interpretability compared to Random Forest, and demand expertise in tuning and architecture design. The choice between these approaches hinges on factors like dataset size, interpretability needs, and the trade-off between accuracy and complexity desired for effective sentiment analysis. However, challenges such as sarcasm detection and handling unstructured data remain, suggesting avenues for further investigation.

Future research can explore the integration of advanced natural language processing techniques, such as deep learning models like recurrent neural networks (RNNs) or transformer-based models like BERT, to improve sentiment analysis accuracy, especially in dealing with complex linguistic nuances. Additionally, incorporating temporal analysis to track sentiment changes over time and geographical analysis to understand regional variations in customer feedback would enhance the scope of this study. Moreover, expanding the dataset to include reviews from other restaurant review platforms and social media sources could yield more comprehensive insights.

Furthermore, the application of sentiment analysis could extend beyond mere classification to identify specific aspects of customer experience (e.g., food quality, service, ambiance) and their impact on overall sentiment, providing even more actionable information for restaurant owners.

VII. Conclusion

In conclusion, this research contributes to the field of sentiment analysis by successfully applying a Random Forest classifier to Zomato customer reviews. Our study highlights the importance of sentiment analysis in the restaurant industry, offering insights into customer satisfaction and areas for improvement. While Random Forest yielded promising results, future work should explore more advanced NLP techniques and consider temporal and geographical factors to provide a more nuanced understanding of customer sentiments. This research underscores the potential for data-driven decision-making in the restaurant business, ultimately leading to enhanced dining experiences and improved customer satisfaction.

References

[1] Baid, Palak, Apoorva Gupta, and Neelam Chaplot. ”Sentiment analysis of movie reviews using machine learning techniques.” International Journal of Computer Applications 179, no. 7 (2017): 45-49.

[2] Shivaprasad, T. K., and Jyothi Shetty. ”Sentiment analysis of product reviews: a review.” In 2017 International conference on inventive communication and computational technologies (ICICCT), pp. 298-301. IEEE, 2017.

[3] Zhang, Zhibin, Hong Li, and Wendong Yu. ”Fine-grained opinion mining: An application of online review analysis in the express industry.” In 2017 3rd IEEE International Conference on Computer and Communications (ICCC), pp. 1498-1503. IEEE, 2017.

[4] Fauzi, M. Ali. ”Word2Vec model for sentiment analysis of product reviews in Indonesian language.” International Journal of Electrical and Computer Engineering 9, no. 1 (2019): 525.

[5] Jabbar, Jahanzeb, Iqra Urooj, Wu JunSheng, and Naqash Azeem. ”Real-time sentiment analysis on E-commerce application.” In 2019 IEEE 16th international conference on networking, sensing and control (ICNSC), pp. 391-396. IEEE, 2019.

[6] Jonathan, Bern, Jay Idoan Sihotang, and Stanley Martin. ”Sentiment Analysis of Customer Reviews in Zomato Bangalore Restaurants Using Random Forest Classifier.” In Abstract Proceedings International Scholars Conference, vol. 7, no. 1, pp. 1719-1728. 2019.

[7] Kausar, Samina, X. U. Huahu, Waqas Ahmad, and Muhammad Yasir Shabir. ”A sentiment polarity categorization technique for online product reviews.” IEEE Access 8 (2019): 3594-3605.

[8] Pandey, Prashant, and Nitasha Soni. ”Sentiment analysis on customer feedback data: Amazon product reviews.” In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), pp. 320-322. IEEE, 2019.

[9] Ahmed, Hafiz Muhammad, Mazhar Javed Awan, Nabeel Sabir Khan, Awais Yasin, and Hafiz Muhammad Faisal Shehzad. ”Sentiment analysis of online food reviews using big data analytics.” Hafiz Muhammad Ahmed, Mazhar Javed Awan, Nabeel Sabir Khan, Awais Yasin, Hafiz Muhammad Faisal Shehzad (2021) Sentiment Analysis of Online Food Reviews using Big Data Analytics. Elementary Education Online 20, no. 2 (2021): 827-836.v

[10] AlQahtani, Arwa SM. ”Product sentiment analysis for amazon reviews.” International Journal of Computer Science & Information Technology (IJCSIT) Vol 13 (2021).

[11] Gupta, Rahul, Syed Sameer, Harsha Muppavarapu, Murali Krishna Enduri, and Satish Anamalamudi. ”Sentiment analysis on Zomato reviews.” In 2021 13th International Conference on Computational Intelligence and Communication Networks (CICN), pp. 34-38. IEEE, 2021.

[12] Abhang, Rutuja Deepak, Bhakti Deepak Bailurkar, Sakshi Shailesh Save, Pradnya Dilip Ingale, and Manali Yashwant Patekar. ”Zomato Review Analysis Using Machine Learning.” In 2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT), pp. 1-5. IEEE, 2023.

[13] Yanfi, Yanfi, Yaya Heryadi, Lukas Lukas, Wayan Suparta, and Yulyani Arifin. ”Sentiment Analysis of User Review on Indonesian Food and Beverage Group using Machine Learning Techniques.” In 2022 IEEE Creative Communication and Innovative Technology (ICCIT), pp. 1-5. IEEE, 2022.

[14] Sarkar, Ansh, Aronya Baksy, and Vinay Kirpalani. ”Analysis of zomato services using recommender system models.” In 2021 International Conference on Intelligent Technologies (CONIT), pp. 1-5. IEEE, 2021.

[15] Laksono, Rachmawan Adi, Kelly Rossa Sungkono, Riyanarto Sarno, and Cahyaningtyas Sekar Wahyuni. ”Sentiment analysis of restaurant customer reviews on tripadvisor using naïve bayes.” In 2019 12th international conference on information & communication technology and system (ICTS), pp. 49-54. IEEE, 2019.