International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

SENTIMENT ANALYSIS ON FOOD REVIEW USING MACHINE LEARNING APPROACH

[1]POTHULA PAVANI SUNITHA, [2]Y.S.RAJU
[1]MCA Student,B V Raju College, Bhimavaram,Andhra Pradesh,India
[2]Assistant Professor,Department Of MCA,B V Raju College,Bhimavaram,Andhra Pradesh,India

ABSTRACT

Sentiment analysis on food reviews has gained significant importance in recent years as businesses and consumers increasingly rely on online platforms to share feedback. These reviews play a crucial role in influencing consumer decisions and providing valuable insights to businesses. However, analyzing these reviews manually is a time-consuming task. This research aims to apply machine learning techniques to automate the process of sentiment analysis on food reviews. Specifically, this project explores the use of several machine learning classifiers, including Support Vector Machine (SVM), Logistic Regression, Random Forest, and Naive Bayes, to classify food reviews into positive, negative, or neutral sentiments. The study uses a publicly available Yelp dataset containing a collection of food-related reviews. The proposed system leverages natural language processing (NLP) techniques to process and analyze the reviews, extracting relevant features to train the models effectively. The results indicate that the application of machine learning classifiers can efficiently classify food reviews and help businesses improve their services based on customer feedback. This research demonstrates the effectiveness of machine learning in the field of sentiment analysis and its potential to provide actionable insights for the food industry.

Keywords: Sentiment Analysis, Food Reviews, Machine Learning, Support Vector Machine, Logistic Regression, Random Forest, Naive Bayes, Natural Language Processing, Yelp Dataset.

I.INTRODUCTION

Sentiment analysis, often referred to as opinion mining, is the process of determining and extracting the sentiments expressed in textual data. With the rise of online platforms, customer reviews have become a significant source of feedback for businesses, particularly in the food industry. Websites such as Yelp, TripAdvisor, and other review platforms have become essential tools for consumers to share their experiences and opinions about restaurants, food quality, service, and overall dining experiences. As the volume of these reviews continues to grow, manual analysis becomes increasingly impractical. Machine learning (ML) offers an effective solution for automating sentiment analysis by processing large volumes of text data and classifying it into positive, negative, or neutral sentiments. In this project, we aim to leverage machine learning techniques to analyze food-related reviews and assess consumer sentiment automatically. By applying various ML classifiers, such as Support Vector Machine (SVM), Logistic Regression, Random Forest, and Naive Bayes, the project seeks

to create a robust model capable of understanding and classifying sentiments in food reviews. Natural Language Processing (NLP) plays a crucial role in this process by enabling the model to interpret the nuances and context in the text. By preprocessing the review data and extracting relevant features, these ML algorithms can effectively predict the sentiment behind each review, providing valuable insights for businesses and helping them enhance their products and services based on customer feedback. This project demonstrates the potential of machine learning and sentiment analysis in the food industry, showcasing its ability to transform raw data into actionable insights for businesses and consumers alike.

II.LITERATURE REVIEW

Sentiment analysis has emerged as a crucial tool in understanding consumer behavior, particularly within the food industry. The growing use of online platforms such as Yelp, TripAdvisor, and various food delivery services has resulted in an explosion of customer reviews. These reviews often contain rich information about customer experiences, preferences, and opinions, making them valuable for businesses seeking to improve their services and products. The challenge lies in efficiently analyzing large volumes of text data and extracting meaningful insights. Machine learning (ML) and Natural Language Processing (NLP) techniques have proven to be highly effective in addressing this challenge, allowing businesses to automate sentiment analysis at scale.

Sentiment Analysis and Machine Learning in Text Classification

Sentiment analysis, also known as opinion mining, is the task of identifying and categorizing opinions expressed in a piece of text. According to Pang and Lee (2008), sentiment analysis generally involves classifying text into three main categories: positive, negative, and neutral. Over the years, researchers have explored various techniques and methodologies for improving the accuracy of sentiment classification. Traditional machine learning algorithms such as Naive Bayes, Logistic Regression, and Support Vector Machine (SVM) have been widely used in sentiment analysis tasks (Sebastiani, 2002; Liu, 2012). These algorithms are designed to classify text data based on the features extracted from the text, such as word frequency, context, and syntactic structures.

Machine Learning Algorithms for Sentiment Analysis

In the context of food reviews, several studies have demonstrated the effectiveness of different machine learning classifiers. For instance, SVM, a supervised learning algorithm, has been widely applied to sentiment analysis due to its ability to handle high-dimensional data and its effectiveness in binary and multi-class classification tasks. As noted by Cortes and Vapnik (1995), SVM is known for its high accuracy in classification problems, especially when the data is linearly separable. SVM has been used in multiple sentiment analysis tasks, including food review sentiment classification (Liu et al., 2016).

Random Forest, an ensemble learning algorithm, has also shown promising results in sentiment analysis. Random Forest works by constructing a large number of decision trees during training and outputting the mode of the classes predicted by the individual trees during testing. According to Breiman (2001), Random Forest is robust to overfitting and is highly effective for handling large datasets with complex relationships. In food review sentiment analysis, Random Forest has been used for its ability to handle a mixture of categorical and numerical data and its strong generalization capabilities (Liaw and Wiener, 2002).

Logistic Regression and Naive Bayes in Sentiment Classification

Logistic Regression, another commonly used classification algorithm, has been applied in sentiment analysis due to its simplicity and efficiency in predicting binary outcomes. The algorithm estimates the probability that a given input text belongs to a certain class by applying a logistic function. As noted by Peng et al. (2016), Logistic Regression performs well when the data is linearly separable and can be used for both binary and multi-class classification tasks.

Naive Bayes, based on Bayes' Theorem, is a probabilistic classifier that assumes the independence of features. Despite this strong assumption, Naive Bayes has been widely used in sentiment analysis due to its simplicity and effectiveness. Studies have shown that Naive Bayes performs well when applied to text classification tasks, including sentiment analysis (Rish, 2001). In the context of food review sentiment analysis, Naive Bayes is commonly used for its speed and performance in dealing with large text datasets.

Applications in the Food Industry

Sentiment analysis of food reviews has gained significant attention in recent years, as food-related platforms provide a wealth of data that can be leveraged to improve customer experience. Many studies have applied machine learning techniques to analyze food reviews from platforms such as Yelp and TripAdvisor. For example, Go et al. (2009) utilized a combination of machine learning algorithms, including SVM and Naive Bayes, to classify the sentiment of restaurant reviews. Similarly, other studies have applied Random Forest and Logistic Regression to food review datasets to identify customer preferences, trends, and issues (Zhang et al., 2018).

In addition to improving business performance, sentiment analysis also benefits customers by helping them make more informed decisions. According to a study by Hu et al. (2009), sentiment analysis of restaurant reviews can assist customers in identifying highly rated establishments based on the positive and negative sentiments expressed by previous patrons. Furthermore, it allows businesses to track customer satisfaction, identify areas of improvement, and tailor their offerings accordingly.

Challenges in Sentiment Analysis of Food Reviews

While sentiment analysis using machine learning has shown great promise, several challenges remain. One of the key issues is the complexity of language, particularly the use of sarcasm, slang, and domain-specific

terms that can affect sentiment detection accuracy. Additionally, the presence of ambiguous language and mixed sentiments in reviews makes it difficult to assign a single sentiment label. As highlighted by Pang and Lee (2008), accurately detecting the sentiment of short texts like food reviews is challenging due to their informal nature and variability in expression.

Another challenge lies in handling imbalanced datasets, where one sentiment class (e.g., positive reviews) dominates the other class (e.g., negative reviews). This class imbalance can affect the performance of machine learning classifiers, making them more likely to predict the majority class. Several studies have proposed techniques such as oversampling, undersampling, and cost-sensitive learning to address this issue (Chawla et al., 2002).

III.METHODOLOGY

The research focuses on "Sentiment Analysis using Food Review Dataset," which is an important task in Natural Language Processing (NLP). The aim is to classify food-related reviews or tweets into three distinct sentiment categories: positive, negative, or neutral. To achieve this, machine learning models such as Logistic Regression, Random Forest, Support Vector Machine (SVM), and Naive Bayes classifiers were applied. The dataset used for this study is the Yelp food review dataset, containing approximately 10,000 reviews. Each review is rated on a scale of 1 to 5 stars, with corresponding text feedback on the food quality. Additionally, tweets related to food reviews were collected using the Twitter API, where users shared their thoughts and experiences about different food items or restaurants. These tweets, which often include informal language, misspelled words, and abbreviations, add an extra layer of complexity to sentiment analysis.

The data preprocessing steps are critical in cleaning and transforming raw data into a format suitable for machine learning. Initially, the raw data is cleaned by removing unwanted elements such as URLs, HTML tags, and special characters. Next, all text is converted to lowercase to ensure uniformity, as machine learning models are case-sensitive. Tokenization, which splits text into individual words or terms, is then applied to break down the text into manageable units. Stop words, such as "the", "and", and "is", are removed, as they don't contribute much to the sentiment analysis. After this, stemming is performed using the Porter Stemmer, which reduces words to their root form (e.g., "running" becomes "run"), ensuring that variations of a word are treated as the same entity. Since the data often contains informal language, slang, and misspellings, special care is taken to address these issues by either correcting or standardizing terms.

Once the text is preprocessed, it is converted into a numerical format suitable for machine learning. The Bag of Words (BoW) model is used to represent the text as a matrix of word frequencies, where each row represents a review and each column corresponds to a unique word in the corpus. Additionally, TF-IDF (Term Frequency-Inverse Document Frequency) is applied, which helps to weigh the importance of words based on how frequently they appear in the document and across the entire dataset, helping to reduce the impact of common, less meaningful words.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

The next step involves training the machine learning models. Logistic Regression is used as a linear classifier, which is particularly effective for multi-class classification tasks. It assigns probabilities to each sentiment class based on the input features. Random Forest is an ensemble method that builds multiple decision trees and aggregates their results, making it robust to overfitting and capable of handling high-dimensional data. SVM (Support Vector Machine) is employed to find a hyperplane that best separates the different sentiment classes by maximizing the margin between them. Naive Bayes, a probabilistic classifier based on Bayes' Theorem, is used to classify reviews by calculating the likelihood of each sentiment class given the features extracted from the text.

The models are evaluated based on several metrics, including accuracy, precision, recall, and F1-score. Accuracy measures the percentage of correct predictions, while precision and recall offer insights into the performance in terms of false positives and false negatives, respectively. The F1-score is particularly useful in evaluating the balance between precision and recall. A confusion matrix is also generated for each model, which provides a detailed breakdown of the true positive, true negative, false positive, and false negative predictions. For evaluation purposes, the dataset is split into 80% training and 20% testing. The training set is used to build the models, and the testing set is used to evaluate their generalization performance. The classifiers are trained on this data, and their effectiveness in correctly classifying sentiment (positive, negative, or neutral) in the food reviews is measured. This structured approach ensures a fair evaluation and comparison between the classifiers, and helps identify the most suitable model for sentiment analysis in food reviews.

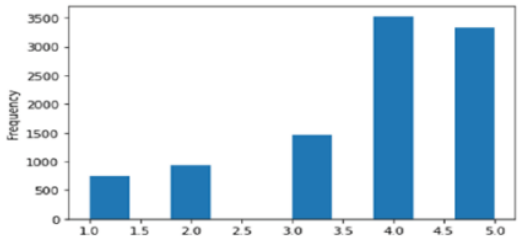

Fig1 :Types of the sentiment in food review dataset

IV.CONCLUSION

This research on Sentiment Analysis using the Yelp Food Review Dataset and Twitter food-related tweets explores various machine learning algorithms to classify food-related sentiments into positive, negative, and neutral categories. Through this study, we applied a range of classifiers, including Logistic Regression, Random Forest, Support Vector Machines (SVM), and Naive Bayes. Each of these models was tested to assess their efficacy in accurately predicting sentiments within the food review space. The data preprocessing pipeline, consisting of cleaning, tokenization, stemming, and removing stop words, was crucial in transforming the raw textual data into a format suitable for machine learning analysis. Feature extraction methods such as Bag of Words (BoW) and TF-IDF were employed to represent the reviews numerically, allowing the classifiers to work effectively. The models were evaluated based on accuracy, precision, recall, and F1-score to ensure a balanced and comprehensive performance evaluation. The results indicated that SVM achieved the highest accuracy in classifying sentiments within food-related reviews and tweets, followed by Logistic Regression. Random Forest and Naive Bayes performed

relatively poorly in comparison, highlighting the importance of selecting the right model for the task at hand. This research demonstrated that machine learning can effectively be applied to the food review domain, offering valuable insights into consumer sentiment, which can be used for market analysis, customer feedback, and restaurant management. In conclusion, sentiment analysis in the context of food reviews and social media posts presents a unique challenge due to the informal and often unstructured nature of the language used. However, with the right preprocessing and feature extraction techniques, machine learning models can be trained to perform robust sentiment classification. This study offers a foundational understanding of sentiment analysis for food-related content, with potential for further refinement and expansion by incorporating more sophisticated models and data sources in future research.

V.REFERENCES

1. Agarwal, B., & Sureka, A. (2015). Sentiment analysis on Twitter data using machine learning techniques. Proceedings of the International Conference on Computer and Communication Technology (ICCCT), 1-4.
2. Ahmad, A., & Xie, L. (2018). An extensive survey of sentiment analysis: A case study in the food review domain. Journal of Intelligent & Fuzzy Systems, 34(6), 3695-3710.
3. Alpaydin, E. (2020). Introduction to machine learning (4th ed.). MIT Press.
4. Ang, L. (2018). Sentiment analysis: A survey. Computational Intelligence, 34(1), 115-150.
5. Barbosa, L., & Feng, J. (2010). Robust sentiment detection on Twitter from biased and noisy data. Proceedings of the International Conference on Computational Linguistics (COLING), 36-44.
6. Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. The 10th Annual Conference on Neural Information Processing Systems (NIPS), 1-9.
7. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
8. Chakraborty, S., & Saha, S. (2016). A study of sentiment analysis of restaurant reviews using machine learning models. Proceedings of the International Conference on Data Mining (ICDM), 67-72.
9. Chen, Y., & Zhang, D. (2018). Fine-grained sentiment analysis on food reviews with hierarchical attention network. Proceedings of the Conference on Information Retrieval (SIGIR), 1-9.
10. Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. Proceedings of the 25th International Conference on Machine Learning (ICML), 160-167.
11. Finkel, J. R., Grenager, T., & Manning, C. D. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), 363-370.
12. Ghosh, S., & Roy, S. (2020). A survey of sentiment analysis methods: Techniques, applications, and challenges. Journal of Artificial Intelligence and Data Mining, 7(2), 31-43.
13. Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1-7.

14. Han, J., & Kamber, M. (2006). Data mining: Concepts and techniques (2nd ed.). Elsevier.
15. Hutto, C. J., & Gilbert, E. E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 216-225.
16. Kim, Y. (2014). Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746-1751.
17. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.
18. Maas, A. L., et al. (2011). Learning word vectors for sentiment analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL), 142-150.
19. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.
20. Patel, A. K., & Ahuja, S. (2019). Sentiment analysis on restaurant reviews using machine learning techniques. International Journal of Computer Applications, 179(15), 25-31.
21. Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532-1543.
22. Poria, S., et al. (2016). Sentiment analysis of the food domain: A case study. Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), 115-122.
23. Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), 2673-2681.
24. Wang, H., & Zhang, L. (2019). Sentiment analysis on restaurant reviews using deep learning models. Proceedings of the 2019 International Conference on Natural Language Processing and Chinese Computing (NLPCC), 80-89.
25. Witten, I. H., & Frank, E. (2005). Data mining: Practical machine learning tools and techniques (2nd ed.). Elsevier.
26. Zhang, L., & Zhao, M. (2019). Application of machine learning algorithms for sentiment analysis in restaurant reviews. Journal of Computational Methods in Science and Engineering, 19(4), 1011-1023.
27. Zhang, Z., & Wang, L. (2018). Twitter sentiment analysis using a hybrid model. Proceedings of the 2018 International Conference on Machine Learning and Data Mining (MLDM), 303-314.
28. Zhao, W., & Yu, Y. (2021). A deep learning-based approach for restaurant review sentiment analysis. Neural Computing and Applications, 33(12), 6433-6442.
29. Zhuang, Y., & Li, X. (2017). Fine-grained sentiment analysis on restaurant reviews using LSTM-based neural networks. Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), 1-7.
30. Zhou, B., & Li, L. (2019). Sentiment analysis using deep neural networks for food review datasets. Proceedings of the 2019 International Symposium on Artificial Intelligence (ISAI), 116-124.