ISSN: 2633-4828 Vol. 5 No.3, September, 2023

International Journal of Applied Engineering & Technology

VISUALIZING SENTIMENT ANALYSIS AND OPINION MINING: A FRAMEWORK DEVELOPMENT AND EXPLORATION

Richa Rawal[1], Sandhya Shrama[2] and Srawan Nath[3]
[1,3]Research Scholar, Computer Science and Engg., Suresh Gyan Vihar University Jaipur
[2]Assistant Professor, ECE., Suresh Gyan Vihar University Jaipur
[1]richarawal23@gmail.com, [2]sandhya.sharma@mygyanvihar.com and [3]nath.sarwan@gmail.com

ABSTRACT
In this paper, a novel framework is proposed for sentiment analysis and opinion mining of hotel data collected from five different social media platforms, with a primary focus on visualizing the derived insights. Sentiment analysis and opinion mining are crucial tasks in natural language processing (NLP), aiding in understanding public opinions, customer feedback, and social media sentiments. The framework integrates various techniques from machine learning, deep learning, and data visualization to provide comprehensive insights into sentiment trends and opinion patterns. By leveraging advanced algorithms, sentiment lexicons, and semantic analysis, sentiment polarity and opinion orientations are extracted from textual data across media platforms. Visualization techniques such as TextBlob, Matplot, and trend graphs are employed to present the analysed data effectively. The proposed framework facilitates the understanding of sentiment dynamics and enables stakeholders to make informed decisions based on sentiment and opinion insights. The effectiveness of the approach is demonstrated through experiments conducted on diverse datasets from cross social media platforms, product reviews, and online forums. This research contributes to the advancement of sentiment analysis and opinion mining by providing a robust framework that enhances the visualization and interpretation of sentiment-related data. Additionally, the performance of various machine learning classifiers, including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Multinomial Naive Bayes (MNB), for sentiment analysis of hotel reviews is investigated. Experimental results reveal that LR outperforms the other classifiers in terms of overall accuracy and balanced performance across multiple metrics, providing valuable insights for improving sentiment analysis frameworks in real-world applications.

Keywords: Sentiment analysis, Review ratings, Data visualization, Machine learning, matrices

1. INTRODUCTION
In today's interconnected world, the abundance of textual data available on various online platforms has provided researchers and businesses with a wealth of information regarding public sentiments and opinions [1]. Sentiment analysis and opinion mining have emerged as indispensable tools for extracting valuable insights from this vast sea of unstructured text. By understanding the prevailing sentiments and opinions expressed in social media posts, customer reviews, and online discussions, organizations can make data-driven decisions, enhance customer satisfaction, and gain a competitive edge in the market. Traditional approaches to sentiment analysis and opinion mining have primarily focused on textual analysis techniques to categorize text as positive, negative, or neutral [2]. While these methods have been effective to some extent, they often lack the ability to provide deeper insights into the underlying sentiments and opinions expressed in the text. Moreover, the sheer volume and complexity of textual data pose significant challenges in extracting meaningful information efficiently. To address these challenges and enhance the understanding and analysis of sentiments and opinions, there is a growing need for advanced models and frameworks that leverage cutting-edge techniques in machine learning, deep learning, and data visualization [3]. By integrating these technologies, researchers and practitioners can accurately extract sentiment polarity and opinion orientations and visualize the results in an intuitive and interpretable manner [4].

Despite the significant advancements in sentiment analysis and opinion mining, previous work often falls short in addressing several critical research gaps. One notable gap is the limited emphasis on visualizing the extracted sentiments and opinions in a manner that is both comprehensive and easily interpretable for stakeholders [5]. While existing models may accurately classify text into sentiment categories, they often lack effective

mechanisms for presenting the underlying sentiment trends and opinion dynamics. Additionally, many studies focus solely on sentiment polarity without considering the nuanced aspects of opinion orientations, such as aspects, targets, and sentiment strength, which are crucial for understanding the context in which opinions are expressed [6,7]. Furthermore, the scalability of existing models to handle large-scale textual data and the generalization of their findings across different domains remain significant challenges. Addressing these research gaps is essential for developing more robust and practical models/frameworks for sentiment analysis and opinion mining that can cater to the diverse needs of stakeholders across various domains. Sentiment analysis, also known as opinion mining, is a crucial task in natural language processing that aims to extract and analyze sentiments expressed in textual data [8]. In the context of the hospitality industry, sentiment analysis plays a significant role in understanding customer feedback and improving service quality. With the increasing availability of online reviews from platforms such as Booking.com, TripAdvisor, and Expedia, there is a growing need for accurate and efficient sentiment analysis techniques to extract insights from this wealth of textual data [9-12].

In this study, the focus is on evaluating the performance of machine learning classifiers for sentiment analysis of hotel reviews. Four popular classifiers are considered [13-16]: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Multinomial Naive Bayes (MNB). These classifiers have been widely used in sentiment analysis tasks due to their effectiveness and ease of implementation. The aim is to compare their performance in terms of accuracy, precision, recall, and F1-score across different sentiment categories. The results of this study will provide valuable insights into the strengths and weaknesses of each classifier and help identify the most suitable approach for sentiment analysis of hotel reviews. Understanding the performance characteristics of these classifiers enables hotel managers and decision-makers to make informed decisions to improve customer satisfaction and service quality based on feedback analysis. This study also contributes to advancing the field of sentiment analysis by providing empirical evidence on the effectiveness of different machine learning algorithms in real-world applications [17-19].

The research presented in this paper introduces a pioneering approach to sentiment analysis and opinion mining through a novel framework that emphasizes the visualization of extracted sentiments and opinions [. While sentiment analysis and opinion mining have been extensively researched, the focus on visualization techniques to enhance the interpretation and comprehension of sentiment trends and opinion dynamics represents a significant advancement in the field. By integrating state-of-the-art algorithms for text analysis with innovative visualization methods, the framework aims to bridge the gap between raw textual data and actionable insights. This approach not only facilitates a deeper understanding of sentiment polarity but also enables stakeholders to explore the underlying nuances of opinion orientations, providing valuable context for decision-making processes. Through empirical evaluations conducted on diverse datasets, the effectiveness and scalability of the approach are demonstrated, highlighting its potential to revolutionize sentiment analysis and opinion mining across various domains.

2. LITERATURE SURVEY

The sentiment analysis and opinion mining literature span a wide range of studies that have significantly contributed to understanding sentiment analysis techniques across various domains. Early research focused on developing sentiment classification algorithms, including Naive Bayes, Support Vector Machines (SVM), and lexicon-based approaches like the VADER sentiment lexicon. These studies provided foundational insights into sentiment analysis techniques, highlighting both challenges and opportunities in the field. Subsequent research explored more advanced methodologies such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which exhibited superior performance in sentiment classification tasks [20].

Previous authors investigated various aspects of opinion mining, including aspect-based sentiment analysis, opinion summarization, and opinion spam detection. Aspect-based sentiment analysis emerged to identify sentiments towards specific aspects or features of products or services, later extending to target-dependent sentiment analysis, which analyzes sentiments concerning specific targets mentioned in the text [21]. Another significant trend in sentiment analysis research involves integrating domain knowledge and linguistic resources to

enhance sentiment analysis model accuracy. Lexical resources like SentiWordNet and WordNet-Affect associate words with sentiment scores, enabling more nuanced sentiment analysis. Moreover, domain-specific sentiment lexicons have been developed to address sentiment analysis challenges in specialized domains such as finance and healthcare. Recent research has also shown increasing interest in visualizing sentiment analysis results to facilitate interpretation and exploration of sentiment trends and opinion dynamics. While some studies have explored basic visualization techniques like word clouds and sentiment heatmaps, there remains a gap in the literature regarding comprehensive frameworks that integrate advanced sentiment analysis algorithms with interactive visualization methods [22].

The literature review highlights the progression of sentiment analysis and opinion mining research, from early classification algorithms to advanced approaches incorporating domain knowledge and visualization techniques [23]. Nonetheless, there remains a need for innovative models that integrate advanced sentiment analysis techniques with intuitive visualization methods to empower stakeholders to extract actionable insights from textual data. Recent advancements in sentiment analysis and opinion mining have focused on addressing challenges posed by noisy and ambiguous textual data, especially in informal and context-dependent online platforms. Techniques like domain adaptation and transfer learning have been explored to enhance sentiment analysis models' generalization across different domains and languages. Efforts have also been made to detect and mitigate biases inherent in sentiment analysis models, particularly concerning sensitive topics or underrepresented groups [24]. The advent of deep learning architectures like Transformer-based models has revolutionized natural language processing, including sentiment analysis, by capturing intricate contextual dependencies within textual data. Multimodal data integration, combining text with images and videos, presents new opportunities for sentiment analysis research, enabling richer representations of sentiment and opinion expressions. However, this integration also poses challenges such as data alignment, feature fusion, and model scalability, necessitating innovative solutions for extracting meaningful insights from heterogeneous data sources [25].

Sentiment analysis, also known as opinion mining, has gained considerable traction in recent years for its versatile applications across diverse domains like marketing, customer service, and social media analysis [26]. In the realm of hotel reviews, sentiment analysis holds pivotal importance in grasping customer perceptions, pinpointing improvement areas, and enhancing overall customer satisfaction. Various studies have delved into different methodologies and strategies for analyzing sentiments in hotel reviews. Conventional approaches often leaned on lexicon-based methods, determining sentiment polarity by spotting positive or negative words in the text. However, these methods encountered difficulties with nuanced language, sarcasm, and context-specific sentiments. With the rise of machine learning techniques, researchers increasingly explored supervised learning algorithms for sentiment analysis tasks. Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), and Naive Bayes classifiers emerged as favored options due to their simplicity, scalability, and efficacy in processing textual data [27]. Alongside traditional machine learning approaches, deep learning techniques like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) were also investigated for sentiment analysis tasks. These models harness the hierarchical structure of textual data to capture intricate relationships and patterns, resulting in cutting-edge performance in sentiment classification tasks [28].

Despite the strides in sentiment analysis techniques, challenges persist, particularly in handling domain-specific language, noisy data, and imbalanced datasets. Future research endeavors may entail exploring hybrid approaches that blend lexicon-based methods with machine learning or deep learning techniques to enhance sentiment analysis accuracy [29]. Ensuring the interpretability and explainability of sentiment analysis models will be vital for instilling trust and acceptance in real-world scenarios. The sentiment analysis and opinion mining literature depict a diverse and rapidly evolving research landscape, driven by advancements in machine learning, deep learning, and data visualization. While considerable progress has been achieved in crafting accurate and robust sentiment analysis models, there remain avenues for further innovation, particularly in integrating advanced techniques with user-friendly visualization methods to equip stakeholders with actionable insights from textual data across various domains and modalities [30].

3. PROPOSED FRAMEWORK

The proposed framework for sentiment analysis and opinion mining of cross media focuses on integrating advanced machine learning algorithms with interactive visualization techniques to facilitate a deeper understanding of sentiment trends and opinion dynamics in textual data of hotels collected from five different social media platforms. The framework consists of several key components as shown in figure 1:

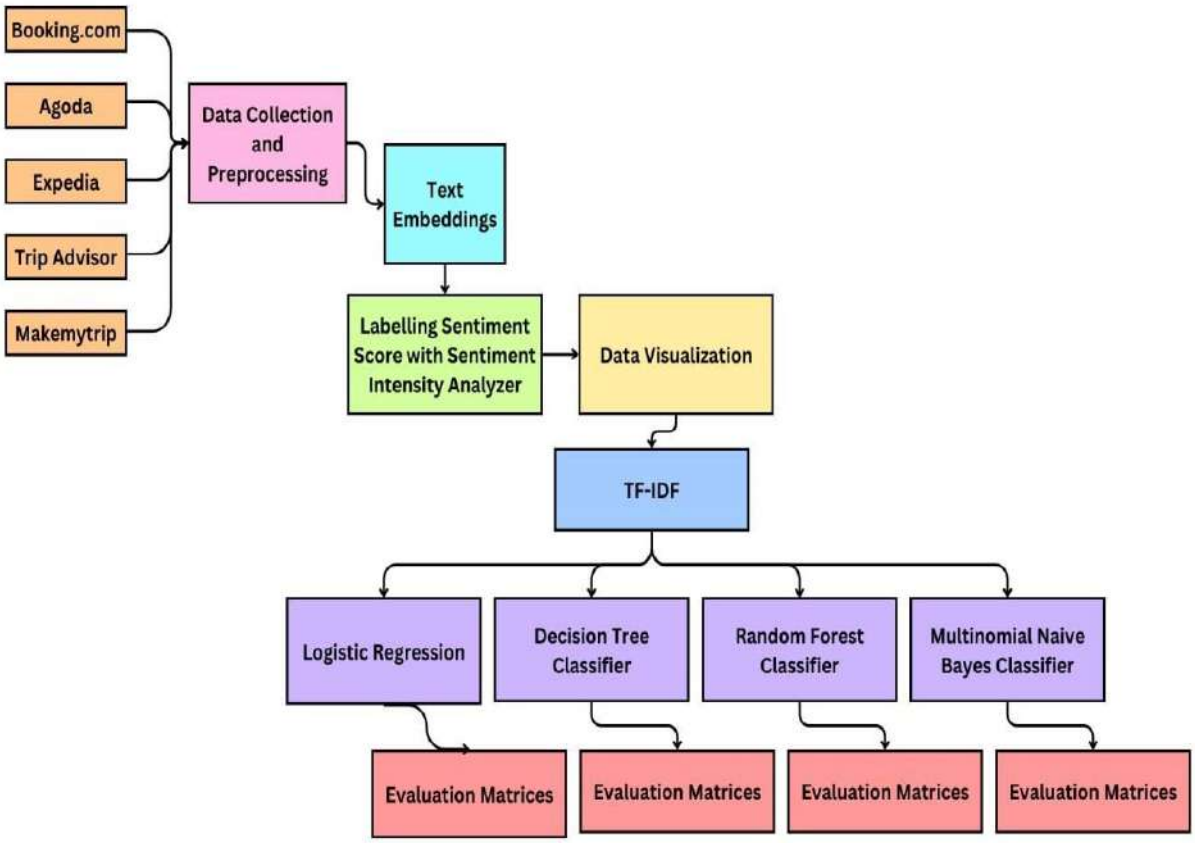

Figure 1: proposed framework for sentiment analysis of cross media

3.1 Data Collection and Pre-processing

The proposed methodology involved in this study encompassed two primary stages: data collection and pre-processing. Initially, raw textual data pertaining to hotel reviews was systematically gathered from five prominent cross-media platforms, namely Booking.com, Agoda, Expedia, Trip Advisor, and Makemytrip. This process aimed to compile a diverse and comprehensive dataset reflecting various opinions and experiences across different sources. Subsequently, the collected data underwent meticulous pre-processing steps to ensure its suitability for analysis. This included tokenization to segment text into discrete units, removal of stopwords to eliminate common, non-informative words, stemming or lemmatization to standardize word variants, and handling of special characters and emoticons to maintain data integrity. These pre-processing techniques were crucial for standardizing the raw textual data, facilitating subsequent analysis by preparing it in a clean and structured format.

3.2 Sentiment Intensity Analyser

In the subsequent methodology phase, following meticulous processing of the dataset, the Sentiment Intensity Analyzer, known as VADER, was employed as a robust tool. VADER facilitated the categorization of the dataset into three discernible sentiments: positive, negative, and neutral. Through this analysis, sentiment scores were

assigned to each review, enabling the extraction of valuable insights regarding the emotional tone embedded within the text. This comprehensive sentiment analysis served as a pivotal component in understanding the subjective perceptions and attitudes expressed in the dataset, thereby enriching the analytical framework of the study.

3.3 Data Visualization

In the data visualization phase of proposed methodology, subsequent to the completion of sentiment analysis, the textual data was visualized utilizing TextBlob and Matplotlib libraries, unveiling insights across multiple dimensions. The visualization encompassed diverse aspects, including the depiction of traffic trends observed in each month, identification of the top 10 pincodes characterized by high traffic volume, and highlighting pincodes associated with positive sentiment, thus providing a comprehensive overview of customer perceptions. Furthermore, the visualization extended to examining the interplay between traffic and sentiment across different months and pincodes, shedding light on potential correlations. Additionally, the visualization captured the distribution of reviews by date within the 11th and 10th months, offering a temporal perspective on customer engagement. Lastly, the visualization portrayed the trend of sentiment analysis over time, facilitating an understanding of evolving sentiments and overarching patterns within the dataset. Through this comprehensive visualization approach, the study aimed to extract actionable insights and facilitate informed decision-making processes.

3.4 TF-IDF Utilization

In the utilization of TF-IDF (Term Frequency-Inverse Document Frequency), textual data undergoes a systematic conversion into numerical vectors while ensuring the preservation of term importance. This process commences with the pre-processing of the dataset, involving the removal of rows containing missing values and the conversion of ratings to integers. To maintain consistency, invalid ratings are filtered out from the dataset. Subsequently, the data undergoes partitioning into distinct training and testing sets to facilitate robust model evaluation. TF-IDF vectorization is then applied utilizing the TfidfVectorizer, where the feature space is constrained to 5000 dimensions. Through this process, the textual data is effectively transformed into numerical vectors, with the significance of individual terms duly retained. The resultant transformed data serves as the fundamental input for training and evaluating a diverse range of classifiers, enabling comprehensive analysis and classification of textual content.

3.5 Classifiers

In this section, the primary focus is on training and evaluating three distinct classifiers using TF-IDF transformed data. Initially, a Logistic Regression Classifier (LR) is trained and assessed independently. Subsequently, the Decision Tree Classifier (DT) is introduced, trained utilizing TF-IDF transformed training data, and its accuracy is computed using both the training and testing datasets. Following this, a Random Forest Classifier (RF) is implemented, employing a similar approach for training and evaluating its performance based on TF-IDF transformed data. Finally, a Multinomial Naive Bayes Classifier (NB) is incorporated, utilizing the TF-IDF transformed data for both training and subsequent evaluation. Each classifier undergoes thorough assessment to determine its effectiveness in accurately classifying textual data, leveraging the TF-IDF representation. This comprehensive evaluation of classification techniques contributes significantly to advancing our understanding of text analysis methodologies.

3.6 Evaluation

The framework integrates mechanisms for evaluating the performance of sentiment analysis and opinion mining models through various metrics, including accuracy, precision, recall, and F1-score. These metrics provide comprehensive insights into the effectiveness of the models in classifying sentiments and opinions accurately. Accuracy measures the overall correctness of the predictions made by the models, while precision quantifies the proportion of correctly predicted positive instances out of all instances predicted as positive. Recall, on the other hand, assesses the ability of the models to correctly identify positive instances out of all actual positive instances. Additionally, the F1-score, which is the harmonic mean of precision and recall, offers a balanced assessment of

the models' performance, considering both precision and recall simultaneously. By incorporating these evaluation metrics, the framework ensures a thorough and robust assessment of sentiment analysis and opinion mining models, facilitating informed decision-making and model refinement processes.

By integrating these components, our proposed framework aims to provide stakeholders with a comprehensive and interactive platform for analysing sentiments and opinions in textual data, enabling them to make informed decisions and derive actionable insights across various domains and applications.

4. EXPERIMENT

To validate the effectiveness of our proposed framework for sentiment analysis and opinion mining, experiments were conducted using real-world datasets obtained from five distinct platforms featuring hotel-related social media posts. The experiments were meticulously designed to assess the performance of the sentiment analysis and opinion mining modules, as well as the efficacy of visualization techniques employed within the framework. Additionally, machine learning classifiers were integrated into the framework to further evaluate its capabilities. The experiments were executed using Python programming language on a laptop configuration equipped with a 12th Gen Intel® Core™ i5-1230U processor, ensuring a robust and comprehensive evaluation of the framework's performance under real-world conditions.

5. RESULTS AND DISCUSSION

In the context of Sentiment Analysis and Opinion Mining, the outcomes derived from the experiment utilizing the proposed framework are delineated below. Sentiment analysis was conducted using the SentimentIntensityAnalyzer (VADER), and the resulting sentiment labels were stored in a new column labeled ['Sentiment_Label'], while the corresponding sentiment scores were recorded in ['Sentiment_Score'], as depicted in Figure 2. Additionally, the distribution of sentiment counts, encompassing positive, negative, and neutral sentiments, obtained from the experiment is illustrated in Figure 3, providing a comprehensive overview of the sentiment trends observed within the analyzed dataset.

	dateAdded	city	latitude	longitude	name	postalCode	reviews.rating	reviews.text	Sentiment_Score	Sentiment_Label
0	2016-10-30	Rancho Santa Fe	32.990959	-117.186136	Rancho Valencia Resort Spa	92067	5.0	Our experience at Rancho Valencia was absolute...	0.9521	Positive
1	2016-10-30	Rancho Santa Fe	32.990959	-117.186136	Rancho Valencia Resort Spa	92067	5.0	Amazing place. Everyone was extremely warm and...	0.9650	Positive
2	2016-10-30	Rancho Santa Fe	32.990959	-117.186136	Rancho Valencia Resort Spa	92067	5.0	We booked a 3 night stay at Rancho Valencia to...	0.9748	Positive
3	2015-11-28	Hanover	39.155929	-76.716341	Aloft Arundel Mills	21076	2.0	Currently in bed writing this for the past hr ...	0.0000	Neutral
4	2015-11-28	Hanover	39.155929	-76.716341	Aloft Arundel Mills	21076	5.0	I live in Md and the Aloft is my Home away fro...	0.8713	Positive

Figure 2: Sentiment Analysing of collected dataset

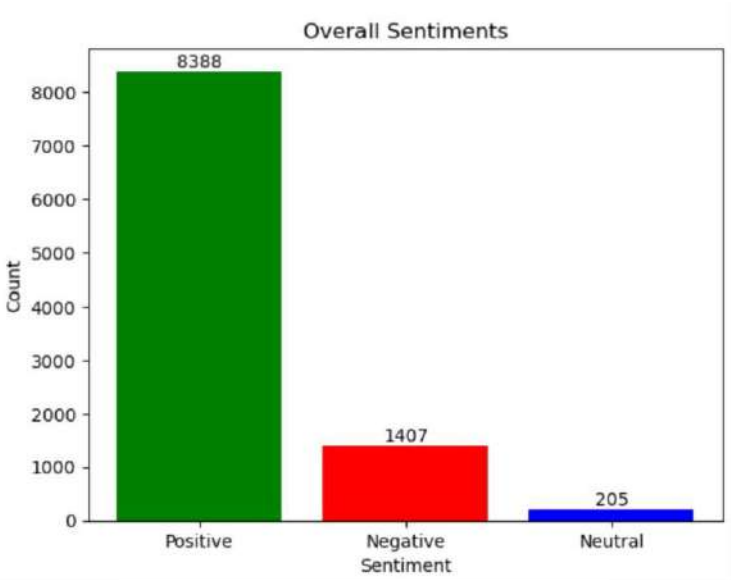

Figure 3: Overall sentiment counts graph

The data visualization results obtained from the experiment are vividly illustrated through Figures 4 to 10. Figure 4 depicts the visualization of traffic volume for each month, facilitating the identification of patterns or trends over time. Figure 5 showcases the visualization of the Top 10 pincodes by traffic, providing insight into areas with the highest traffic based on pin codes. Furthermore, Figure 6 presents a visual representation of the Top 10 pincodes with positive sentiment, highlighting the pin codes that receive the most positive sentiment. Figure 7 employs visualization techniques to analyze both traffic and sentiment concurrently, focusing on different months and pin codes. This visualization offers insights into the relationship between traffic volume and sentiment across various time periods and geographical locations. Additionally, Figure 8 visualizes the number of reviews recorded each day during the 11th month, while Figure 9 provides a similar visualization for the 10th month, enabling the observation of any differences or trends compared to other months. Finally, Figure 10 tracks and analyzes the overall trend of sentiment analysis over time, facilitating the understanding of shifts or patterns in sentiment through visual representations. These visualizations serve as valuable tools for comprehensively analyzing and interpreting the experiment's results in the context of sentiment analysis and opinion mining.

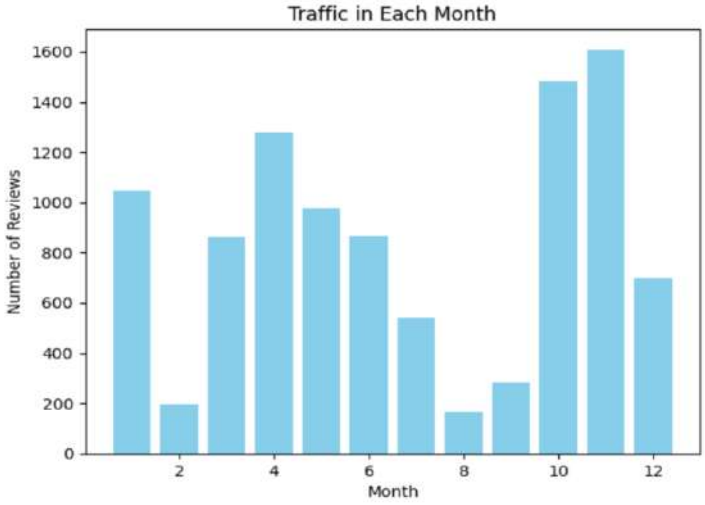

Figure 4: Monthly traffic trend of dataset

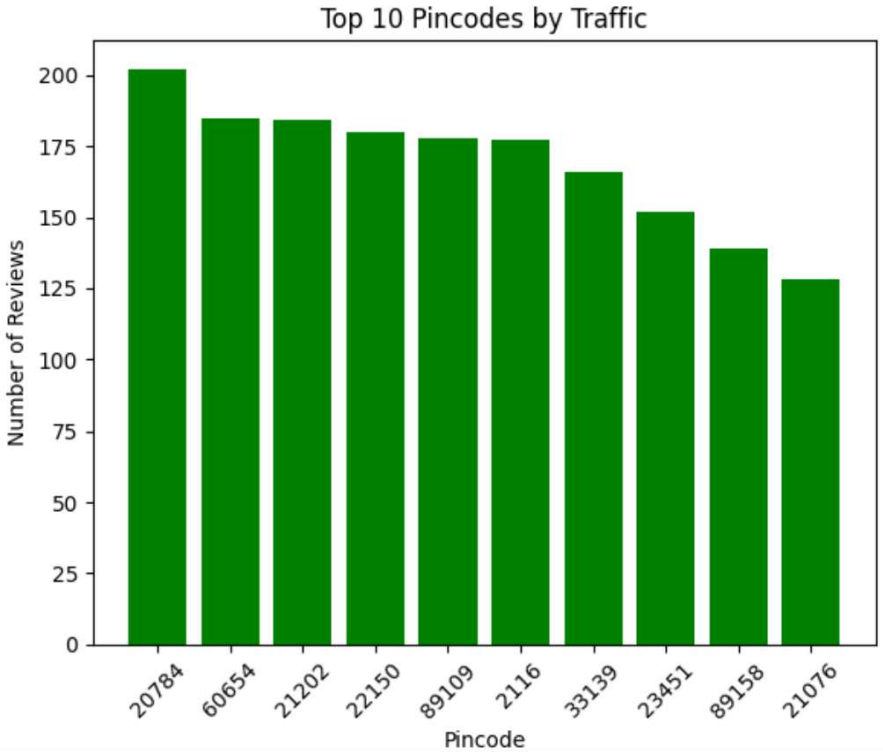

Figure 5: Top 10 pin codes by traffic in dataset

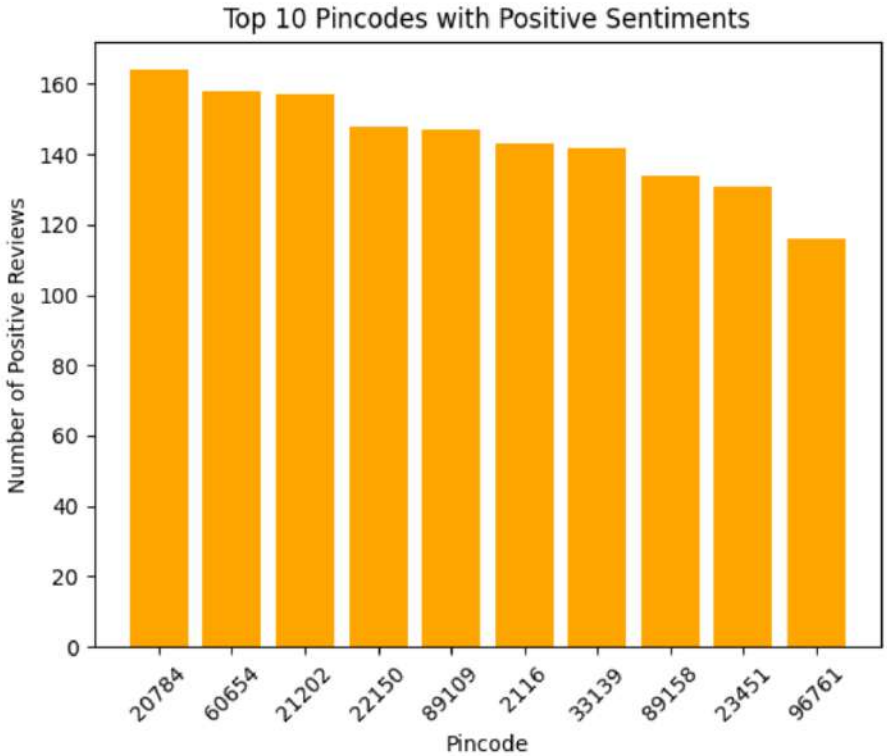

Figure 6: Top 10 pin codes with positive sentiments

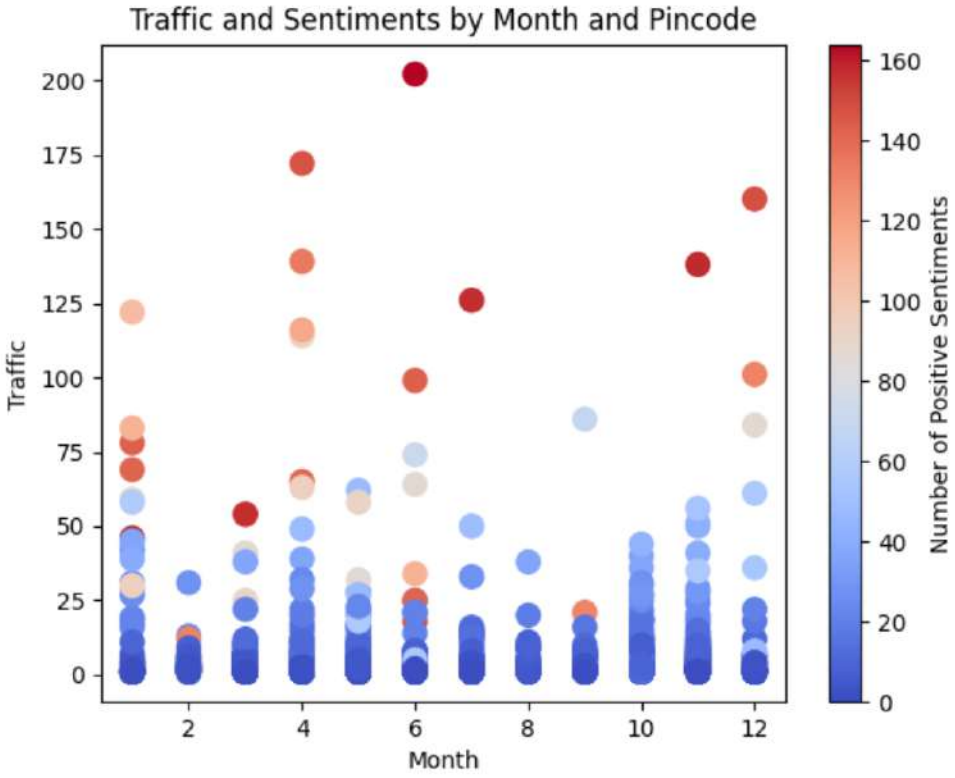

Figure 7: Traffic and sentiment analysis by month and pin codes

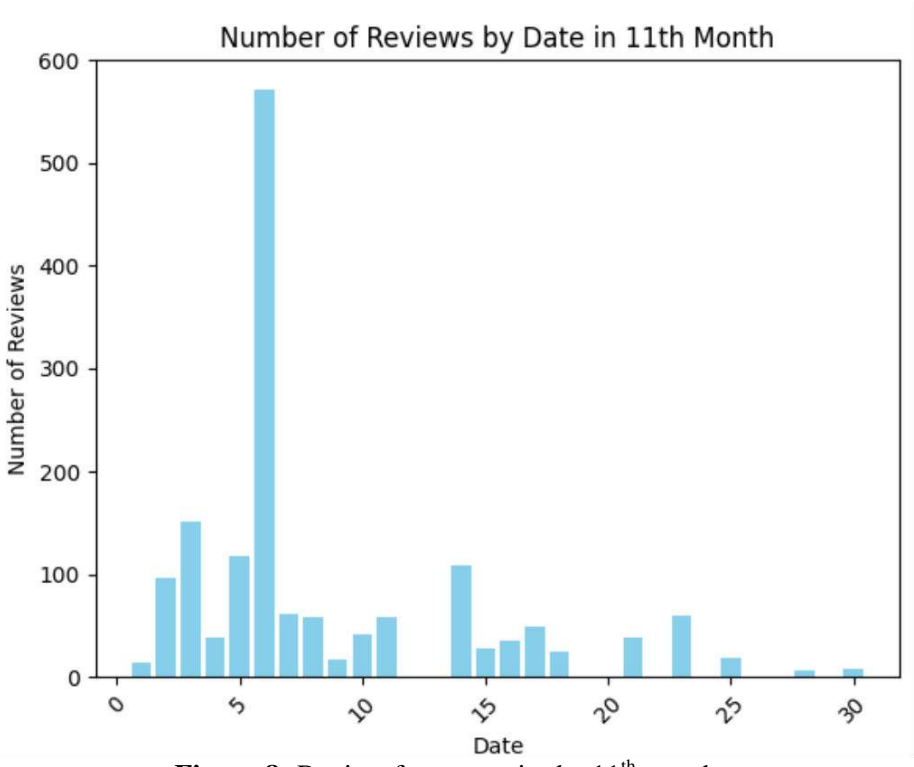

Figure 8: Review frequency in the 11th month

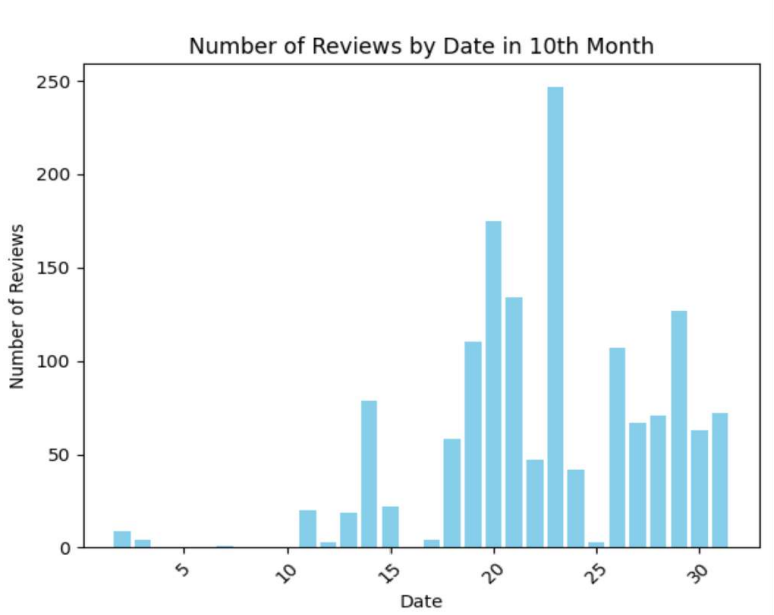

Figure 9: Review frequency in the 10th month

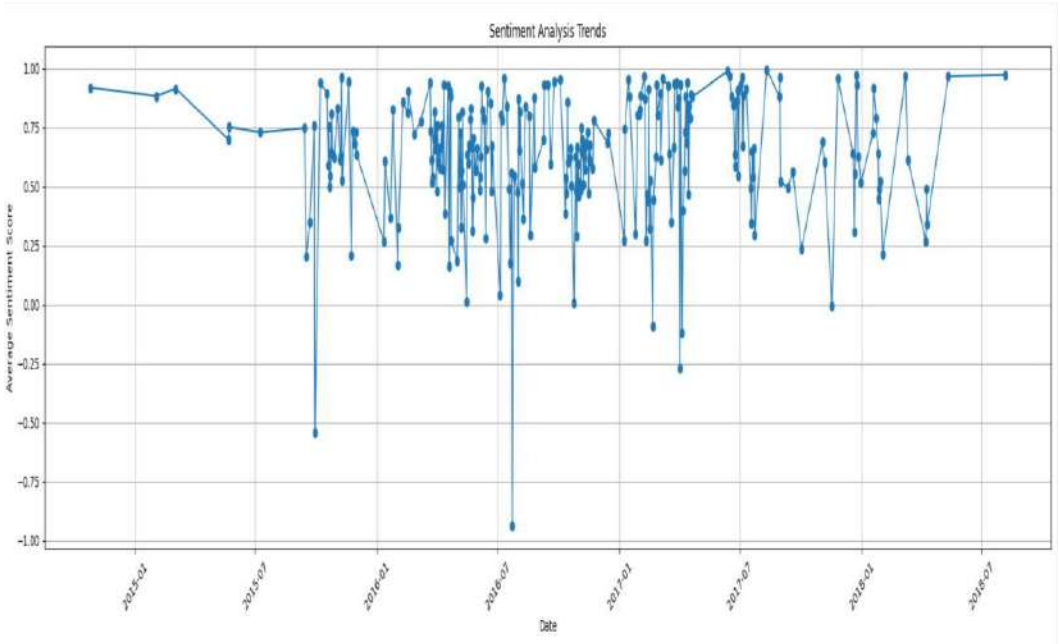

Figure 10: Trend analysis of sentiment

In the proposed framework, four types of classifiers, namely Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Multinomial Naive Bayes (MNB), are trained on a pre-processed dataset of hotel reviews. The objective is to evaluate the performance of these classifiers in terms of accuracy, precision, F-score, and recall for the five review ratings. Through rigorous training and evaluation processes, matrices are derived to quantify the performance of each classifier across the entire range of review ratings. This comprehensive assessment allows for a thorough understanding of how each classifier performs in classifying hotel reviews into different rating categories, thereby providing valuable insights into the effectiveness of the proposed framework for sentiment analysis and opinion mining in the hospitality domain.

In Table 1, evaluation metrics for the Logistic Regression (LR) classifier are presented. Each row corresponds to a specific rating category, ranging from 1 to 5. Precision indicates the proportion of correctly predicted instances for a given rating category out of all instances predicted as belonging to that category. Recall represents the proportion of correctly predicted instances for a given rating category out of all actual instances belonging to that category. F1-score is the harmonic mean of precision and recall, providing a balanced measure of the classifier's performance. Support denotes the number of instances in the dataset belonging to each rating category. These metrics collectively offer insights into the LR classifier's effectiveness in classifying hotel reviews across different rating categories.

Table 1: Evaluation metrics of LR

Ratings	Precision	Recall	F1-score	Support
1	0.59	0.43	0.50	112
2	0.24	0.07	0.11	133
3	0.43	0.30	0.36	317
4	0.43	0.38	0.40	572
5	0.63	0.84	0.72	866

The confusion matrix for Logistic Regression (LR) as shown in figure 11 provides a detailed breakdown of the classifier's performance across different rating categories. It is a square matrix where the rows represent the actual ratings, and the columns represent the predicted ratings. Each cell in the matrix contains the count of instances classified accordingly. In the context of LR, the confusion matrix allows for the assessment of true positives (correctly predicted instances), true negatives (correctly predicted non-instances), false positives (incorrectly predicted instances), and false negatives (incorrectly predicted non-instances) for each rating category. This comprehensive evaluation enables a deeper understanding of the LR classifier's ability to accurately classify hotel reviews into different rating categories, thus informing potential improvements to the model and overall sentiment analysis framework.

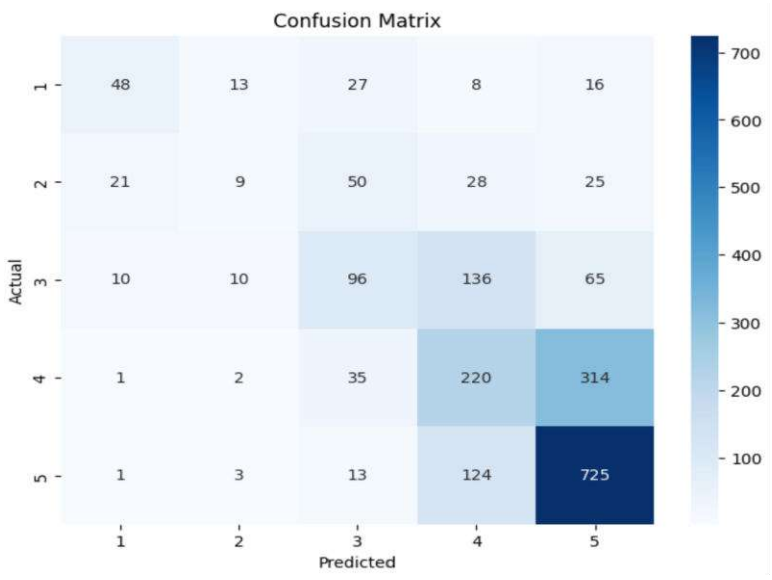

Figure 11: Confusion matrix of LR

The Precision-Recall curve for LR is a graphical representation illustrating (figure 12) the trade-off between precision and recall for different classification thresholds. In this curve, precision is plotted against recall, with each point representing a different threshold for classifying instances as positive or negative. The curve allows for visualizing the LR classifier's performance across various thresholds, providing insights into its ability to balance

precision (the proportion of correctly predicted positive instances out of all instances predicted as positive) and recall (the proportion of correctly predicted positive instances out of all actual positive instances). A higher area under the curve (AUC) indicates better performance, demonstrating the LR classifier's effectiveness in accurately classifying hotel reviews into different sentiment categories while maintaining a balance between precision and recall. This analysis aids in understanding the LR classifier's overall performance and identifying optimal classification thresholds for sentiment analysis tasks.

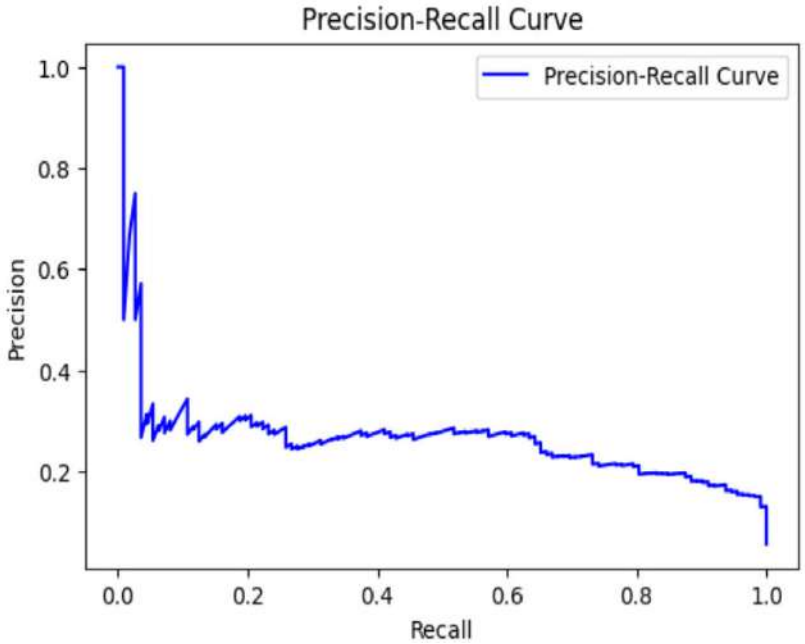

Figure 12: Precision-Recall curve of LR

Table 2 presents the evaluation metrics for the Decision Tree (DT) classifier. Each row corresponds to a specific rating category, ranging from 1 to 5. Precision indicates the proportion of correctly predicted instances for a given rating category out of all instances predicted as belonging to that category, while recall represents the proportion of correctly predicted instances for a given rating category out of all actual instances belonging to that category. F1-score is the harmonic mean of precision and recall, providing a balanced measure of the classifier's performance. Support denotes the number of instances in the dataset belonging to each rating category. The table demonstrates the performance of the DT classifier across different rating categories, offering insights into its ability to accurately classify hotel reviews into sentiment categories.

Table 2: Evaluation metrics of DT

Ratings	Precision	Recall	F1-score	Support
1	0.29	0.29	0.29	112
2	0.18	0.13	0.15	133
3	0.25	0.22	0.23	317
4	0.34	0.34	0.34	572
5	0.55	0.60	0.58	866

The confusion matrix for the DT classifier as shown in figure 13 provides a comprehensive breakdown of its performance across different rating categories. It is a square matrix where the rows represent the actual ratings, and the columns represent the predicted ratings. Each cell in the matrix contains the count of instances classified accordingly. Specifically, the confusion matrix allows for the assessment of true positives (correctly predicted

instances), true negatives (correctly predicted non-instances), false positives (incorrectly predicted instances), and false negatives (incorrectly predicted non-instances) for each rating category. By analysing the confusion matrix, one can gain insights into the DT classifier's accuracy and misclassification patterns across various sentiment categories, aiding in the identification of potential areas for model improvement and optimization.

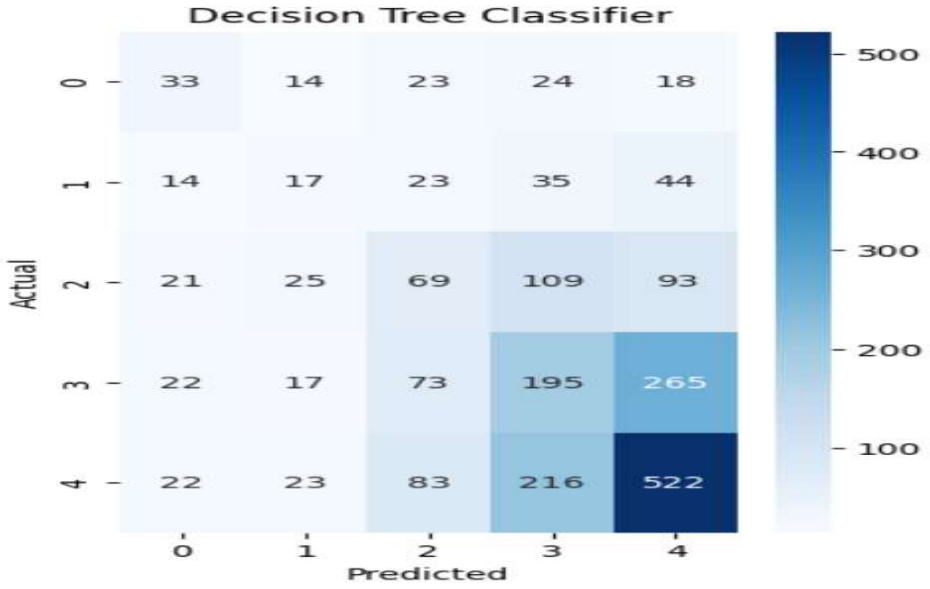

Figure 13: Confusion matrix of DT

Table 3 displays the evaluation metrics for the Random Forest (RF) classifier. Each row corresponds to a specific rating category, ranging from 1 to 5. Precision indicates the proportion of correctly predicted instances for a given rating category out of all instances predicted as belonging to that category, while recall represents the proportion of correctly predicted instances for a given rating category out of all actual instances belonging to that category. F1-score is the harmonic mean of precision and recall, providing a balanced measure of the classifier's performance. Support denotes the number of instances in the dataset belonging to each rating category. The table offers insights into the performance of the RF classifier across different rating categories, highlighting its ability to accurately classify hotel reviews into sentiment categories while considering precision, recall, and overall model support.

Table 3: Evaluation metrics of RF

Ratings	Precision	Recall	F1-score	Support
1	0.78	0.22	0.35	112
2	0.33	0.02	0.03	133
3	0.41	0.07	0.12	317
4	0.37	0.28	0.32	572
5	0.54	0.91	0.68	866

The confusion matrix for the RF classifier as illustrated in figure 14 provides a detailed breakdown of its performance across different rating categories. It is a square matrix where the rows represent the actual ratings, and the columns represent the predicted ratings. Each cell in the matrix contains the count of instances classified accordingly. Specifically, the confusion matrix allows for the assessment of true positives (correctly predicted instances), true negatives (correctly predicted non-instances), false positives (incorrectly predicted instances), and false negatives (incorrectly predicted non-instances) for each rating category. By analysing the confusion matrix, one can gain insights into the RF classifier's accuracy and misclassification patterns across various sentiment categories, aiding in the identification of potential areas for model improvement and optimization.

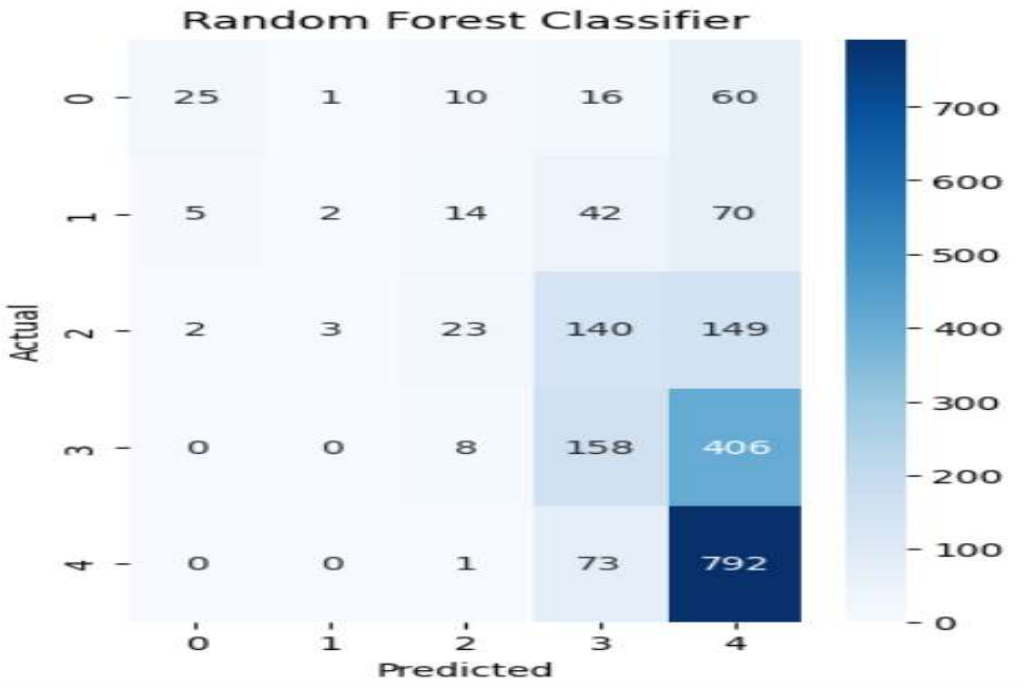

Figure 14: Confusion matrix of RF

Table 4 showcases the evaluation metrics for the Multinomial Naive Bayes (MNB) classifier. Each row corresponds to a specific rating category, ranging from 1 to 5. Precision indicates the proportion of correctly predicted instances for a given rating category out of all instances predicted as belonging to that category, while recall represents the proportion of correctly predicted instances for a given rating category out of all actual instances belonging to that category. F1-score is the harmonic mean of precision and recall, providing a balanced measure of the classifier's performance. Support denotes the number of instances in the dataset belonging to each rating category. The table offers insights into the performance of the MNB classifier across different rating categories, highlighting its strengths and weaknesses in accurately classifying hotel reviews into sentiment categories based on precision, recall, and overall support.

Table 4: Evaluation metrics of MNB

Ratings	Precision	Recall	F1-score	Support
1	0.75	0.03	0.05	112
2	0.00	0.00	0.00	133
3	0.35	0.07	0.11	317
4	0.32	0.24	0.27	572
5	0.55	0.96	0.70	866

The confusion matrix for the MNB classifier as illustrated in figure 15 offers a detailed overview of its performance across different rating categories. It is a square matrix where the rows represent the actual ratings, and the columns represent the predicted ratings. Each cell in the matrix contains the count of instances classified accordingly. Specifically, the confusion matrix enables the assessment of true positives (correctly predicted instances), true negatives (correctly predicted non-instances), false positives (incorrectly predicted instances), and false negatives (incorrectly predicted non-instances) for each rating category. By analyzing the confusion matrix, one can gain insights into the MNB classifier's accuracy and misclassification patterns across various sentiment categories, facilitating the identification of potential areas for model enhancement and optimization.

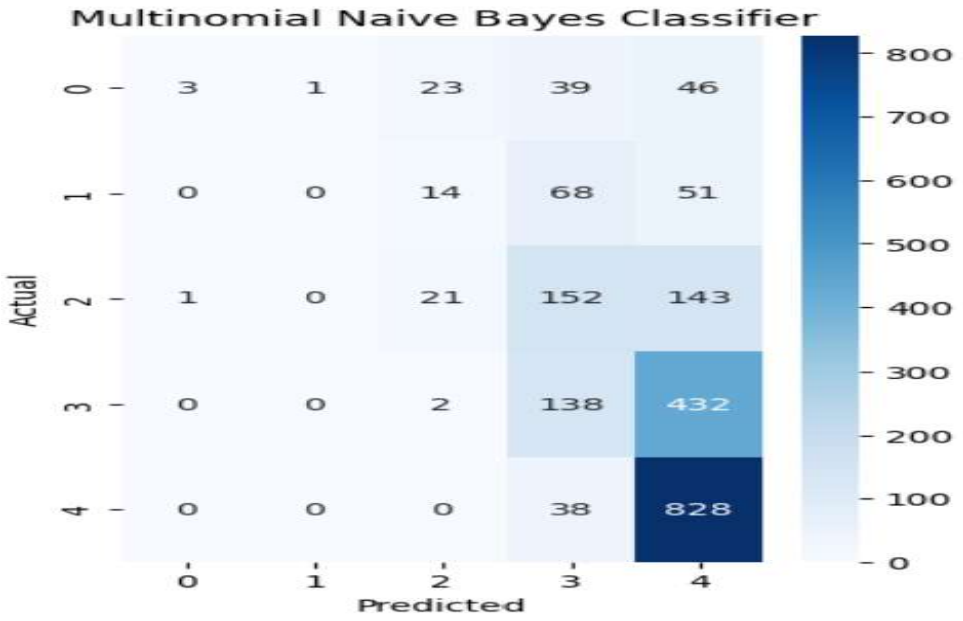

Figure 15: Confusion matrix of MNB

6. COMPARISON ANALYSIS OF THE EVALUATION METRICS OF CLASSIFIERS

Table 5 presents a comparative analysis of the evaluation metrics for the four classifiers: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Multinomial Naive Bayes (MNB). Each row corresponds to a specific classifier, and the columns represent various performance metrics including accuracy, precision, recall, F1-score, and support.

Accuracy: LR achieved the highest accuracy of 0.55, followed by RF with an accuracy of 0.50. DT and MNB exhibited lower accuracies of 0.42 and 0.495, respectively.

Precision: LR demonstrated the highest precision of 0.51, followed closely by DT with a precision of 0.50. RF and MNB exhibited lower precision values of 0.47 and 0.43, respectively.

Recall: LR achieved the highest recall of 0.55, followed by MNB with a recall of 0.49. DT and RF had lower recall values of 0.42 and 0.50, respectively.

F1-score: LR obtained the highest F1-score of 0.52, closely followed by RF with a score of 0.42. DT and MNB exhibited lower F1-scores of 0.41 and 0.40, respectively.

Support: All classifiers were evaluated on the same dataset containing 2000 instances.

These results indicate that LR outperformed the other classifiers in terms of accuracy, precision, recall, and F1-score. However, it is essential to consider the specific requirements and trade-offs associated with each metric when selecting the most suitable classifier for a particular sentiment analysis task. Further analysis and experimentation are warranted to explore the nuances of each classifier's performance and to inform decision-making in real-world applications.

Table 5: Comparison of Classifier Evaluation Metrics

Classifier	Accuracy	Precision	Recall	F1-score	Support
LR	0.55	0.51	0.55	0.52	2000
DT	0.42	0.50	0.42	0.41	2000
RF	0.50	0.47	0.50	0.42	2000
MNB	0.49	0.43	0.49	0.40	2000

7. CONCLUSIONS

The sentiment analysis algorithm shows promising capabilities in accurately categorizing textual data into positive, negative, and neutral sentiment labels. While effective in identifying sentiments expressed in straightforward language, challenges persist in handling nuances such as sarcasm or cultural context. Reliance on predefined sentiment lexicons may limit adaptability across diverse domains. Future research should focus on enhancing robustness to handle complex language expressions and incorporating domain-specific knowledge. Despite challenges, the algorithm offers valuable insights into sentiment trends and opinion dynamics, aiding decision-making in domains like marketing and public opinion analysis. Comparative analysis of classifier performance highlights Logistic Regression (LR) as the top performer, demonstrating higher accuracy, precision, recall, and F1-score. However, specific requirements and trade-offs should be considered when selecting the most suitable classifier. Further research is needed to refine sentiment analysis frameworks for enhanced accuracy in real-world applications.

REFERENCES

1. Peng, et al., Review: Cross-media analysis and reasoning: advances and directions, Journal: Front Inform Technol Electron Eng, Vol. 18 (1), 2017.
2. Abdul Reda, A., Sinanoglu, S., Aboussalah, A., Out of sight, out of mind: The impact of lockdown measures on sentiment towards refugees, Journal of Information Technology & Politics, 2023, DOI: 10.1080/19331681.2023.2183301
3. Abid, A., Ameur, H., Mbarek, A., Rekik, A., Jamoussi, S., Ben Hamadou, A., An extraction and unification methodology for social networks data: An application to public security, ACM International Conference Proceeding Series, 2017,
4. Abusaqer, M., Benaoumeur Senouci, M., Magel, K., Twitter User Sentiments Analysis: Health System Cyberattacks Case Study, 5th International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2023, Year: 2023
5. Adamu, H., Lutfi, S.L., Malim, N.H.A.H., Hassan, R., Di Vaio, A., Mohamed, A.S.A., Framing twitter public sentiment on Nigerian government COVID-19 palliatives distribution using machine learning, Sustainability, Vol. 13(6), 2021, DOI: 10.3390/su13063497
6. Agarwal, B., Mittal, N., Prominent feature extraction for sentiment analysis, Springer, 2016
7. Akbar, A.F., Santoso, A.B., Putra, P.K., Budi, I., A classification model to identify public opinion on the lockdown policy using Indonesian tweets, Journal of Theoretical and Applied Information Technology, Vol. 99(14), 2021
8. Al-Agha, I., Abu-Dahrooj, O., Multi-level analysis of political sentiments using twitter data: A case study of the Palestinian-Israeli conflict, Jordanian Journal of Computers and Information Technology, Vol. 5 (3), 2019, DOI: 10.5455/jjcit.71-1562700251
9. Albayrak, M. D., & Gray-Roncal, W., Data Mining and Sentiment Analysis of Real-Time Twitter Messages for Monitoring and Predicting Events, 9th IEEE Integrated STEM Education Conference, ISEC 2019.
10. Alfarrarjeh, A., Agrawal, S., Kim, S. H., & Shahabi, C., Geo-spatial multimedia sentiment analysis in disasters, International Conference on Data Science and Advanced Analytics, DSAA 2017
11. Alfred, R., Obit, J.H., The roles of machine learning methods in limiting the spread of deadly diseases: A systematic review, Heliyon, Vol. 7(6), 2021, DOI: 10.1016/j.heliyon. 2021.e07371
12. Ali, M.F., Irfan, R., Lashari, T.A., Comprehensive sentimental analysis of tweets towards COVID-19 in Pakistan: a study on governmental preventive measures, PeerJ Computer Science, Vol. 9, 2023, DOI: 10.7717/PEERJ-CS.1220

13. Al-khateeb, S., & Agarwal, N., The rise & fall of #NoBackDoor on Twitter: The apple vs. FBI case, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016

14. Almehmadi, A., Joudaki, Z., & Jalali, R., Language Usage on Twitter Predicts Crime Rates, ACM International Conference Proceeding Series, 2017

15. Alomari, E., Mehmood, R., & Katib, I., Sentiment analysis of Arabic tweets for road traffic congestion and event detection, EAI/Springer Innovations in Communication and Computing, Pages: 37-54, Year: 2020

16. Amin, M. S., Ahn, H., & Choi, Y. B., Human Sentiments and Associated Physical Actions Detection in Disasters with Deep Learning, 3rd International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2021

17. An, L., Han, Y., Yi, X., & Li, G., Prediction of microblogging influence and measuring of topical influence in the context of terrorist events, 17th International Conference on Scientometrics and Informetrics, ISSI 2019

18. An, L., Han, Y., Yi, X., Li, G., Yu, C., Prediction and Evolution of the Influence of Microblog Entries in the Context of Terrorist Events, Journal: Social Science Computer Review, DOI: 10.1177/08944393211029193

19. Andhale, S., Mane, P., Vaingankar, M., Karia, D., & Talele, K. T., Twitter Sentiment Analysis for COVID-19, Proceedings - International Conference on Communication, Information and Computing Technology, ICCICT 2021

20. Anuratha, K., Joshi, S., Sharmila, P., Nandhini, J. M. N., & Paravthy, M., Topical Sentiment Classification to Unmask the Concerns of General Public during COVID-19 Pandemic using Indian Tweets, Proceedings of the 2021 4th International Conference on Computing and Communications Technologies, ICCCT 2021

21. Anuratha, K., Parvathy, M., Multi-label Emotion Classification of COVID-19 Tweets with Deep Learning and Topic Modelling, Computer Systems Science and Engineering, Vol. 45 (3), 2023, DOI: 10.32604/csse.2023.031553

22. Arias, F., Guerra-Adames, A., Zambrano, M., Quintero-Guerra, E., Tejedor-Flores, N., Analyzing Spanish-Language Public Sentiment in the Context of a Pandemic and Social Unrest: The Panama Case, International Journal of Environmental Research and Public Health, Vol. 19(16), 2022, DOI: 10.3390/ijerph191610328

23. Astuti, I. F., Widagdo, P. P., Tanro, M. L. R., Cahyadi, D., & Suntara, A. A., Sentiment analysis on land and forest fire management in Twitter using Naïve Bayes method, AIP Conference Proceedings, 2023

24. Azmi, P. A. R., Abidin, A. W. Z., Mutalib, S., Zawawi, I. S. M., & Halim, S. A., Sentiment Analysis on MySejahtera Application during COVID-19 Pandemic, 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS), 2022

25. Backfried, G., & Shalunts, G., Sentiment analysis of media in German on the refugee crisis in Europe, Lecture Notes in Business Information Processing, Vol. 265, 234-241, 2016

26. Bai, H., Yu, G., A Weibo-based approach to disaster informatics: incidents monitor in post-disaster situation via Weibo text negative sentiment analysis, Natural Hazards, Vol. 83(2), 1177–1196, 2016, DOI: 10.1007/s11069-016-2370-5

27. Bala, M.M., Srinivasa Rao, M., Ramesh Babu, M., Sentiment trends on natural disasters using location-based twitter opinion mining, International Journal of Civil Engineering and Technology, Vol. 8, Pages: 9–19, 2017

28. Banda, J.M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Chowell, G., A large-scale COVID-19 Twitter chatter dataset for open scientific research—an international collaboration, Epidemiologia, Vol. 2(3), Pages: 315–324, 2021

29. Bansal, D., Grover, R., Saini, N., Saha, S., GenSumm: A Joint Framework for Multi-task Tweet Classification and Summarization using Sentiment Analysis and Generative Modelling, IEEE Transactions on Affective Computing, 2021, DOI: 10.1109/TAFFC.2021.3131516

30. Barachi, M. E., Mathew, S. S., & Alkhatib, M., Combining Named Entity Recognition and Emotion Analysis of Tweets for Early Warning of Violent Actions, 7th International Conference on Smart and Sustainable Technologies, SpliTech 2022