iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59108

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Sentiment Analysis Based on Travelers' Reviews Using the SVM Model with Enhanced Conjunction Rule-Based Approach

Major Dr. V.A. Narayana[1], Belde Pooja[2], Bachupally Akhil Goud[3]
[1]*Professor,* [2,3]*UG Student, Department of CSE, CMR College of Engineering and Technology, Hyderabad, India*

Abstract: ***It is a tough task for tourism management to identify their user reviews and come up with solutions to the advancement of their tourism organizations. There are many social media reviews, and Tourism organizations face the challenge of evaluating numerous social media reviews to find solutions for advancing their organizations. It can be tough to physically assess all those reviews, social media has become a huge trend these days. People are constantly sharing their experiences and opinions on tourist places. By analyzing the sentiment of reviews, it may obtain useful information about the popularity of different tourist destinations. It's a great way to understand what people love about certain places. Sentiment classification is indeed a valuable tool in classifying reviews into different categories, aiding in decision-making. However, it's important to note that reviews often contain noisy content like typos and emoticons, which can affect the accuracy of the algorithms. Taking these aspects into consideration is crucial for achieving more accurate results. Decision-making and multiple class classification of reviews are possible with the application of sentiment classification. Sentiment analysis plays a crucial role in helping tourists make informed decisions about their travel destinations. In this particular paper, Using an Enhanced Conjunction Rule-Based Approach and Support Vector Machine (SVM) machine learning technique, the authors conducted sentiment analysis. They collected the dataset from different tourism review websites to train and evaluate their model. The findings of this paper are significant as they not only provide valuable insights in the field of tourism but also help in identifying the most appropriate algorithm for tourism-related analysis. The findings of this paper are significant as they not only provide valuable insights in the field of tourism but also help in identifying the most appropriate algorithm for tourism-related analysis. This information can greatly contribute to the development and improvement of tourism strategies and decision-making processes.***

Keywords: Sentiment Analysis, Tourism, Support Vector Machine (SVM), Machine Learning, Conjunction Rule Base Approach

I. INTRODUCTION

Nowadays, social media is experiencing rapid growth, with millions of users posting reviews and rating visiting places available on different tourism websites. With the help of Sentiment Analysis, these reviews can be immensely helpful. By properly analyzing the reviews, we can identify trends in the popularity of tourist places, as mentioned in [1,2,3]. The summarized results from sentiment analysis can assist tourists in making decisions about their tour destination and planning. This project's research goal is to create a sentiment analysis system specifically designed for travelers' reviews, using a combination of Support Vector Machine (SVM) modeling and an enhanced conjunction rule-based approach. This tailored system aims to provide valuable insights for travelers based on their reviews. In this study, a feature extraction algorithm was used i.e.., TFIDF Vectorization Algorithm. Three classifiers Support Vector Machine (SVM), SVM with N-grams, and SVM with Enhanced Conjunction Rule-Based classification were also utilized to classify sentiment. Based on variables like precision, F1-score, recall, execution time, and accuracy comparisons of feature extraction and classification algorithm combinations were carried out. In this paper Section II of the paper discusses the sentiment analysis literature surveys that the authors have undertaken. In Section III, the methodology for sentiment analysis has also been outlined, along with features like performance evaluation, visualization, and classification of reviews of tourist attractions. In Section IV, they present the results of their experiment. It's great that they have provided a comprehensive overview of their approach and findings in different sections of their paper.

II. LITERATURE REVIEW

Sentiment analysis of reviews of Tourist Place reviews has been studied and obtained results by many authors. Each author used a different approach when analyzing the reviews. In [1] the author Nur Aliah Khairina Mohd Haris, et.al., have evaluated the sentiment polarity of tourist destination reviews focusing on Taman Negara using an SVM and RF classifier. Contrasting these classifiers' performances shows that SVM is the best option for analyzing sentiment in tourism reviews on social media platforms.

The results achieve an impressive accuracy of 67.97% for the 5-fold cross-validation model, which outperforms RF's accuracy of 63.55%. The project also provides a sentiment analysis dashboard that visually displays visitor reviews of Taman Negara. In [2] the authors Aranyak Maity, et.al., Their study compares N-grams and Conjunction Rule-based Approaches, concluding that the conjunction rule-based approach yields superior results. By enhancing feature-specific sentiment analysis of hotel reviews, this study addresses the challenge of information overload in online feedback, providing a detailed assessment of aspects like food, service, and location. In another study [3], the authors provide an original approach that applies a dual prediction technique. In this method, the unigram and inverse unigram models calculate both positive and negative polarity weights to calculate average spending sentiment. An experimental study [4] shows that in automatic aspect-based sentiment analysis, many important aspects are missed and instead, various irrelevant aspects are extracted and taken into consideration, thereby deteriorating the performance, which can be addressed by providing a pre-built feature list and dictionaries based on those features.

Fang-Zhan [5] even found that implicit neutral emotions were marked as positive. Finding the polarity of a particular target is difficult. For example, even with the same word “rich”, “rich service” has a positive meaning, but “rich food” has a negative meaning. In summary, assigning polarity to a specific target may also require some common sense of aspect-oriented polarity calculations. Timor Kadir et al. [6] describe the growing challenge of processing large amounts of unstructured text. This requires effective text-mining techniques and algorithms to uncover meaningful patterns. Text mining is essential for extracting valuable insights from text data, and is especially important in the biomedical and healthcare fields. Rohit Joshi et al. [7] investigate the use of his Twitter data for sentiment prediction by supervised machine learning algorithms. This research focuses on sentiment analysis of film reviews and uses classifiers such as SVM and Naive Bayes for classification. SVM outperforms other classifiers with an impressive 84% accuracy in predicting sentiment in movie reviews. M.D. Devika et al., [8] Describe sentiment analysis as a process of interpreting user emotions, which falls under the domain of Natural Language Processing (NLP). The rise of internet-based applications has led to an increase in personalized reviews to assist travelers and customers in their decision-making process. Sentiment analysis can prove to be an invaluable tool for extracting and summarizing useful insights from overwhelming online reviews. In a study [9], Twitter data about two major international clothing Using Naive Bayes and the Lexicon Dictionary, brands were contrasted and examined. The purpose of this was to find out how the people felt about the two brands. A different document [10] focused on classifying travelers' ratings into four categories: flight comfort, staff service, food and entertainment, and price.

Bayesian and SVM techniques were used to determine passenger fulfillment in these categories. Additionally, opinions on the Indonesian vaccine were collected on Twitter [11], and support vector machines and random forests were used to predict people's sentiments on vaccines. It's interesting to note that Naive Bayes, SVM, and RF are commonly used as machine learning classifiers in sentiment analysis research. There are various opinion mining strategies, as shown in [12]. Trend management, aspect management, set management. The authors proposed not only tourist destinations but also aspect-based opinion mining. Information on related topics was retrieved from visitor reviews. The ratings were then divided into two categories of positive and negative emotions on various topics. For aspect extraction and opinion trends, they use Tagger and WordNet. They extracted tweets based on that. Positive, negative, and neutral classifications are used. Machine learning helps improve system performance. A learning approach was developed. Text mining is also used in biomedicine. As mentioned in [13], the authors have conducted a study on sentiment analysis and the movie examines the dataset. He noted that prior studies have concentrated on maximum entropy, Naive Bayes, and SVM classification techniques. In this paper, the author has categorized reviews using sentiment classification. The most accurate classifier was the RF classifier with a 90% accuracy rate. The authors of this study [14] used multiple features such as Unigram for the movie review dataset Top 2633, Bigram, Unigrams + Bigrams, POS, adjectives for numerous classification methods, and Unigrams and Unigrams + Position maximum entropy. The authors used sentiment analysis. Naive Bayes and he used SVM at all. He made an accurate comparison. Research has shown that Naive Bayes has the lowest level of accuracy, while SVM provides the highest level of accuracy. Similar to [12], there are various opinion mining strategies. B. Trend management, aspect management, set management. Various Researchers from [15] to [20] have worked on tourist place review analysis using various technologies like machine learning, and opinion mining.

III.METHODOLOGY

The act of computationally examining a text to determine people's thoughts, judgments, points of view, feelings, and sentiment polarity (positive, negative, or neutral) toward topics, situations, events, and themes is known as sentiment analysis [6][7]. Sentiment analysis involves analyzing user opinions from text data, referred to as opinion mining. There are two primary approaches utilized for sentiment analysis i.e., Supervised methodology for machine learning and unsupervised lexicon-based methodology.

Machine learning techniques including SVM, SVM with N-Gram, and SVM with conjunction rule-based are being used to predict feelings from the tourist reviews dataset to close gaps and enhance the performance and accuracy of the predictions. After that, the reviews' performance is assessed.

Sentiment analysis was done in this article using the procedures listed below. Figure 1 displays the architecture of the system. The following procedures are used in this paper to do sentiment analysis.

A. Data Set

In this study, the researchers collected review data from different tourism websites. The data was stored in a CSV format and included review text along with corresponding ratings. They were able to ascertain if the reviewers' sentiments were favorable or negative by examining the text. A rating score greater than 3 is considered a positive review, and a score less than or equal to 3 is considered a negative review.

B. Data Preprocessing

When dealing with social media data, it's crucial to perform data preprocessing to clean up the raw data. This involves several such as lemmatization, stemming, tokenization, and the removal of stop words.

These steps help to refine the data and make it more suitable for sentiment analysis.

1) *Tokenization:* It converts the sequence of reviews into smaller parts which are considered tokens.
2) *Remove Stop Words:* Words that have little to no meaning are known as "stop words." These are the words like 'the', 'a, 'an', 'in', 'of', and 'and'. This study used a custom stop-word list containing words that are unrelated and occur very frequently in the corpus. This reduced the size of the feature vector and improved the performance of the system.
3) *Lemmatization:* It reduces the word to its root form to identify similarities.
   Example: A pile of tokens is converted to a pile.
4) *Stemming:* Discovering the root word of a token is called stemming.
   Example: Token treks will be converted to treks. was performed along with the above steps to improve the performance of the machine learning algorithm. Remove short words, remove punctuation, remove numbers and special characters, and convert to lowercase.

C. Feature Extraction

For feature extraction from evaluation data, count vectorization and TFIDF vectorization algorithms were used. Vectorization of counts is the same as the Bag of Word (BoW) approach. This shows how often the text appears in a particular document and how often it occurs.

TFIDF vectorization is an extension of count vectorization, which also considers inverse document frequencies in parallel to term frequencies.

D. Training Model

The dataset was used for the training model. The study uses support vector machines, SVM with N-grams, and SVM with a joint rule-based classification algorithm on the training validation dataset.

E. Test Model

From the entire evaluation data set, data was used for testing. To predict sentiment polarity, a test was performed on new, unseen reviews. The trained model classifies the sentiment of reviews into two classes: positive and negative.

F. Performance Evaluation

Performance evaluation is a crucial step. In this study, the researchers evaluated the performance utilizing metrics such as recall, execution speed, accuracy, and precision.

It's important to assess how well the model performs and how long it takes to execute. These evaluations help us understand the effectiveness and efficiency of the machine-learning approach.

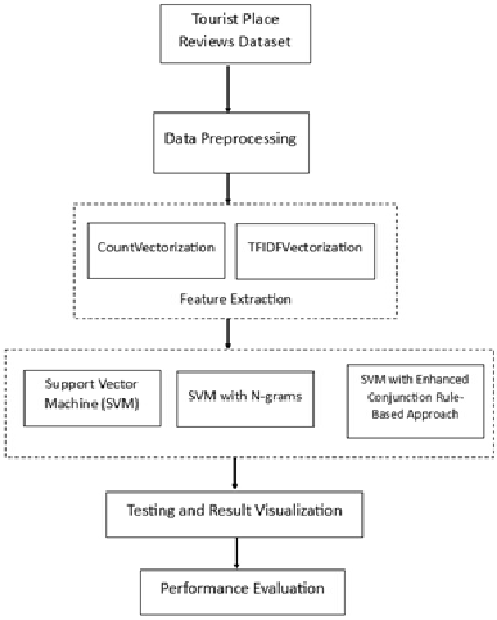


Fig. 1 Data Flow Diagram

In this research, they used to review data from different tourism websites, that was gathered in CSV file format. The information contained review text and related ratings. They classified the reviews as favourable or negative based on these ratings, which indicated the sentiment of the reviews. If the rating was greater than 3, it was considered positive and if it was less than or equal to 3, it was considered negative. It's a clever way to categorize the sentiments expressed in the reviews. The users can log in to the website and can find the tourism data, they have access to search about various places which they want to do analysis. They will get the prediction result based on the data they have searched.

In this research study, the team collected a whopping 800,000 reviews from different tourism websites. Each review was accompanied by corresponding ratings. By analyzing these reviews, they were able to calculate the sentiment expressed in them. It's incredible how they gathered such a large dataset to gain insights into the sentiments of the reviewers.

1) Reviews with ratings higher than 3 indicate a favorable opinion and are assigned a 1

2) If the rating of the review is less than or equal to 3 then the sentiment is negative and marked as 0.

In this way, a labeled dataset for review sentiment analysis was obtained.

Hardware and Software Requirements for Application Development are

a) *Processor Tools:* Intel i5 processor with 32 GB RAM a minimum of 1 TB space on the Hard Disk is needed.

b) *Software Requirements:* The application is developed using Python, with Windows 10 64-bit OS support.

c) *Technologies and Languages:* Python is the primary language used for development.

d) *Dataset Collection:* The data is gathered from Kaggle.

IV. IMPLEMENTATION

Pre-processing the data was necessary to eliminate superfluous and extraneous terms from the reviews because the information gathered from different travel websites was unprocessed. A data preprocessing stop removed words, punctuation marks, and short words. Tokenization, lemmatization, and stemming were also performed. The significance of data pretreatment for feature reduction and enhanced machine learning algorithm performance was acknowledged in this work. They started with data purification and then implemented a feature extraction algorithm from scratch. Count Vectorization and TFIDF Vectorization were used to extract features. The classification algorithm was implemented using the Python sklearn package, which offers various routines for different classifiers. They trained the classification algorithm on the labeled training dataset and evaluated its performance utilizing metrics such as F1 score, recall, and precision.

Review	Class label	Sentiment
Best time is go just after rainy season to enjoy fog & green carpet over the mountain. Good trekking area.	1	Positive
Security at this place is not good at all	0	Negative

Fig. 2 Tourist Place Review Sentiment Analysis Example

V. RESULTS AND DISCUSSION

This section will delve into the results obtained from analyzing positive and negative reviews using an SVM, SVM with n-gram, and SVM with a conjunction rule-based approach for sentiment classification. The dataset from which obtained results has inconsistency. Therefore, researchers balanced the dataset using the SMOTE approach. Following the dataset's use of the Smote. This is the graph that balances out next.

Class distribution before SMOTE: Counter ({1: 760000, 0: 178645})

Class distribution after SMOTE: Counter ({1: 760000, 0: 760000})

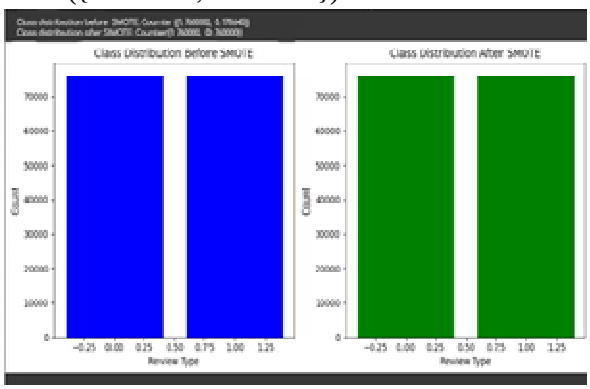


Fig. 3 Data Imbalance Graph

A. Accuracy Score

To be sure, accuracy is a simple, common-sense way to assess performance. It shows the proportion of accurately anticipated observations to all observations. The graphs that compare accuracy are shown in Figures 5 and 6. Visualizing the accuracy can help us understand the performance of the model more easily. It is always helpful to have visual representations to make data analysis more accessible. The formula for accuracy is given below.

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + true\ negatives + false\ negatives + false\ positives}$$

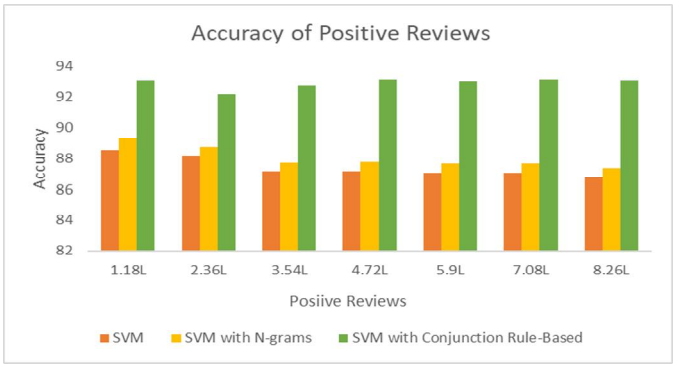


Fig.4 Accuracy of Positive Reviews

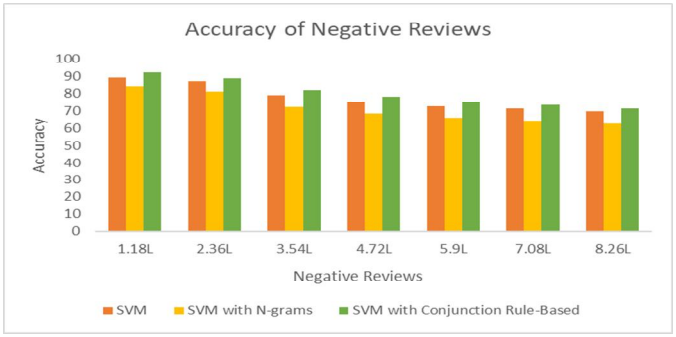


Fig. 5 Accuracy of Negative Reviews

B. Precision

The precision comparison graph, which is shown in Fig. 7, is a valuable visual representation. Predictive value, another name for precision, is an important performance measure. It focuses on the proportion of accurately forecasted positive observations to all scheduled positive observations. The accuracy performance of the model can be understood by looking at a precision comparison graph.

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

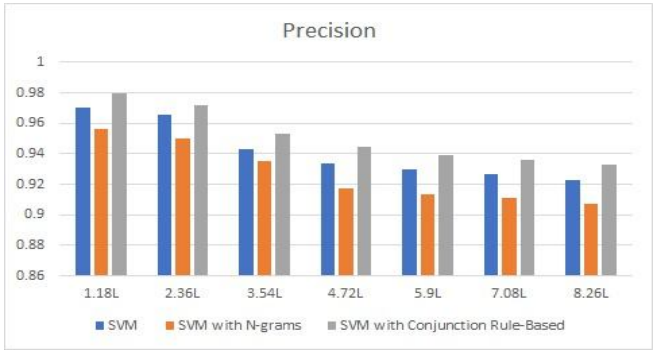


Fig. 6 Precision comparison of algorithms

C. Recall

Recall, sometimes called sensitivity, is an important performance measure. Its main focus is on the proportion of all actual positive observations to all correctly predicted positive observations. Fig. 8 is a recall comparison graph that ought to offer insightful information about the model's memory capabilities.

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

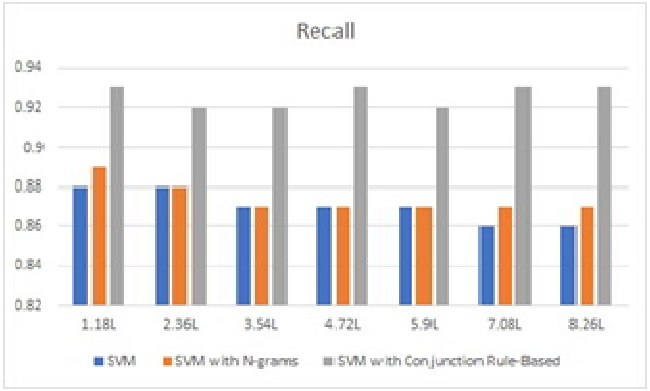


Fig. 7 Recall comparison of algorithms

D. F1-Score

The F1-score is a metric that considers both precision and recall, providing a balanced measure of performance. It takes into account the trade-off between precision and recall, making it a valuable evaluation metric. The F1-score comparison graph is provided in Fig. 9. Visualizing the F1 score can help us understand the overall performance of the model in a balanced way.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

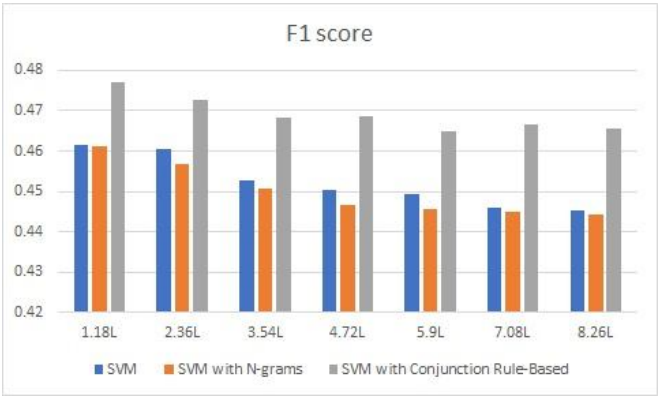


Fig. 8 F1-Score comparison of algorithms

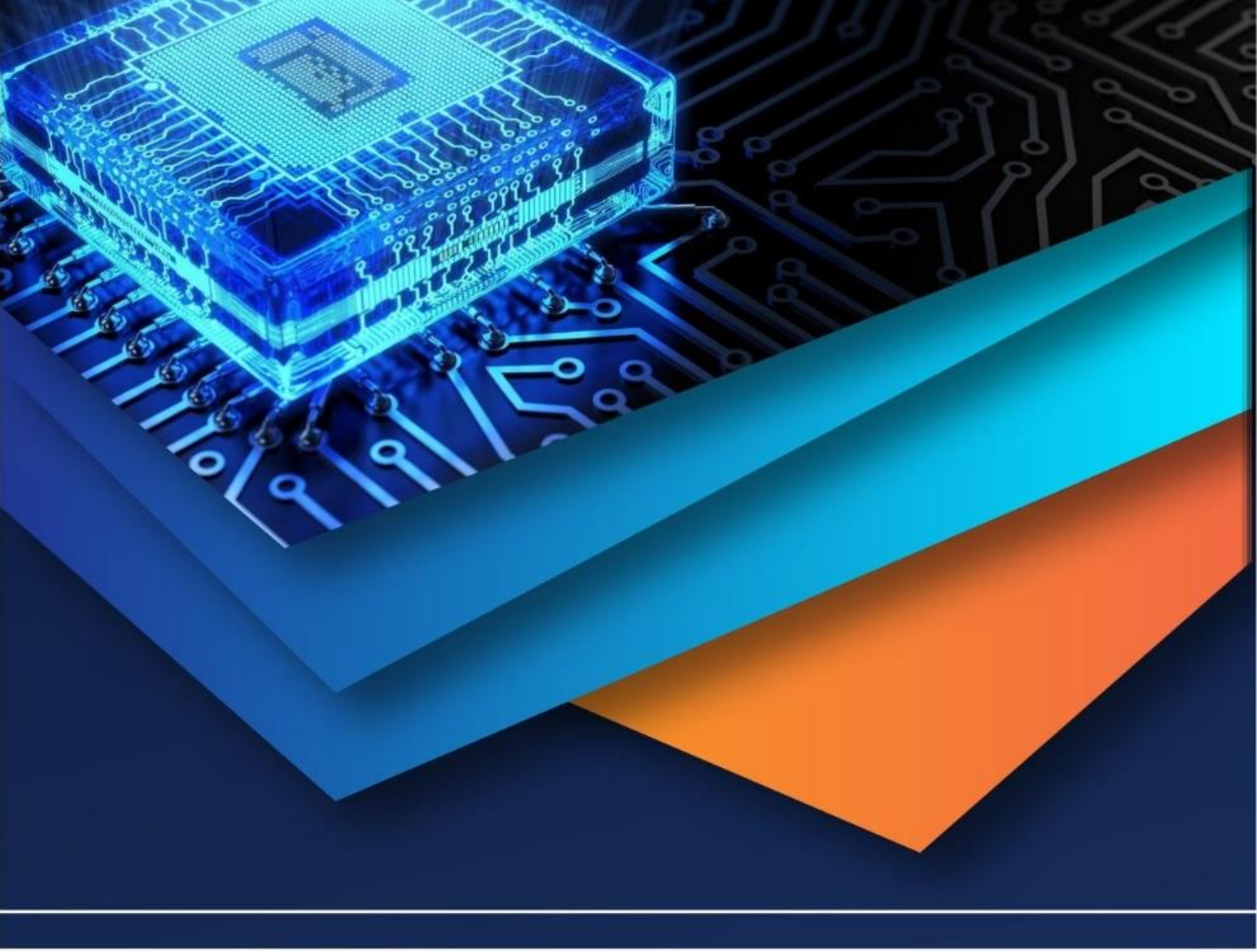
VI. CONCLUSION

In this paper, user reviews are obtained from popular websites for different tourism destinations and the sentiments of the reviews are analyzed. In our proposed model we use a machine learning algorithm named Support Vector Machine (SVM) model to predict the sentiments from tourist place reviews and evaluate the performance using an Enhanced Conjunction rule-based approach. The aggregated outcome will help the users to know which place is suitable for traveling and it helps to identify the users more easily. In future work, we shall try to improve our experimental results accuracy by using a larger data set and gathering data from more websites.

REFERENCES

[1] Khairina Mohd Haris, N. A., Mutalib, S., Ab Malik, A. M., Abdul-Rahman, S., & Kamaliah Kamarudin, S. N. (2023). Sentiment classification from reviews for tourism analytics. International Journal of Advances in Intelligent Informatics, 9(1).
[2] Maity, A., Ghosh, S., Karfa, S., Mukhopadhyay, M., Pal, S., & Pramanik, P. K. D. (2020). Sentiment analysis from travellers' reviews using enhanced conjunction rule based approach for feature-specific evaluation of hotels. Journal of Statistics and Management Systems, 23(6), 983-997.
[3] Anto, M. P., Antony, M., Muhsina, K. M., Johny, N., James, V., & Wilson, A. (2016, March). Product rating using sentiment analysis. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (pp. 3458-3462). IEEE.
[4] Afzaal, M., & Usman, M. (2015, October). A novel framework for aspect-based opinion classification for tourist places. In 2015 Tenth International Conference on Digital Information Management (ICDIM) (pp. 1-9). IEEE.
[5] Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. Journal of Big Data, 2(1), 1-14.
[6] Shaik, K. B., Ganesan, P., Kalist, V., Sathish, B. S., & Jenitha, J. M. M. (2015). Comparative study of skin color detection and segmentation in HSV and YCbCr color space. Procedia Computer Science, 57, 41-48.
[7] Rasool, A., Tao, R., Marjan, K., & Naveed, T. (2019, March). Twitter sentiment analysis: a case study for apparel brands. In Journal of Physics: Conference Series (Vol. 1176, p. 022015). IOP Publishing.
[8] Satria, A. T., & Nugraheni, D. M. K. (2020). Implementation of Integrated Bayes Formula and Support Vector Machine for Analysing Airline's Passengers Review. In E3S Web of Conferences (Vol. 202, p. 15004). EDP Sciences
[9] Afzaal, M., & Usman, M. (2015, October). A novel framework for aspect-based opinion classification for tourist places. In 2015 Tenth International Conference on Digital Information Management (ICDIM) (pp. 1-9). IEEE.
[10] Wankhede, R., & Thakare, A. N. (2017, April). Design approach for accuracy in movies reviews using sentiment analysis. In 2017 international conference of electronics, communication and aerospace technology (ICECA) (Vol. 1, pp. 6-11). IEEE.
[11] Wawre, S. V., & Deshmukh, S. N. (2016). Sentiment classification using machine learning techniques. International Journal of Science and RAllahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919. esearch (IJSR), 5(4), 819-821.
[12] Kaur, H., Mangat, V., & Krail, N. (2017). Dictionary-based sentiment analysis of Hinglish text and comparison with machine learning algorithms. International Journal of Metadata, Semantics and Ontologies, 12(2-3), 90-102.
[13] Nally, R. M., & Fleishman, E. (2004). A successful predictive model of species richness based on indicator species. Conservation biology, 18(3), 646-654.

[14] Garau□Vadell, J. B., Díaz□Armas, R., & Gutierrez□Taño, D. (2014). Residents' perceptions of tourism impacts on island destinations: A comparative analysis. International Journal of Tourism Research, 16(6), 578-585.
[15] Nilashi, M., Yadegaridehkordi, E., Ibrahim, O., Samad, S., Ahani, A., & Sanzogni, L. (2019). Analysis of travellers' online reviews in social networking sites using fuzzy logic approach. International Journal of Fuzzy Systems, 21, 1367-1378.
[16] Iqbal, S., Hassan, S. U., Aljohani, N. R., Alelyani, S., Nawaz, R., & Bornmann, L. (2021). A decade of in-text citation analysis based on natural language processing and machine learning techniques: An overview of empirical studies. Scientometrics, 126(8), 6551-6599.
[17] Sarkar, K., & Bhowmick, M. (2017, December). Sentiment polarity detection in bengali tweets using multinomial Naïve Bayes and support vector machines. In 2017 IEEE Calcutta Conference (CALCON) (pp. 31-36). IEEE.
[18] Laksono, R. A., Sungkono, K. R., Sarno, R., & Wahyuni, C. S. (2019, July). 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.
[19] Mirzaalian, F., & Halpenny, E. (2021). Exploring destination loyalty: Application of social media analytics in a nature-based tourism setting. Journal of Destination Marketing & Management, 20, 100598.
[20] Warsito, B., & Prahutama, A. (2020). Sentiment analysis on Tokopedia product online reviews using random forest method. In E3S Web of Conferences (Vol. 202, p. 16006). EDP Sciences.
[21] https://www.tripadvisor.in/
[22] https://www.kaggle.com



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY