iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** I **Month of publication:** January 2024

DOI: https://doi.org/10.22214/ijraset.2024.58091

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 I Jan 2024- Available at www.ijraset.com

Analysing the Sentiments using Natural Language Processing

Aryan Saraf[1], Rashika Bhardwaj[2], Rajneesh Thakur[3], Bhumika[4]
Computer Science EngineeringChandigarh University, Mohali, India

Abstract: ***Sentiment analysis is an approach to identify, extract, quantify, and study affective states and subjective data that makes use of text analysis, computational linguistics, biometrics, and natural language processing. Customer materials, online and social media content, and medical materials all commonly use. Deep language models have made it possible to analyze more difficult data domains, such as news texts. The fundamental task is to categorize a text based on its polarity, that is, whether the expressed opinion is positive, negative, or neutral. In advanced sentiment classification, emotions like surprise, fear, anger, disgust, sadness, and pleasure are all looked at. Psychological studies and the General Inquirer are two sources of sentiment analysis's antecedents. Challenges in sentiment analysis include the possibility of opinion words being positive in one state being negative in another, and the fact that people may express opinions differently. Developing algorithms to identify and classify opinion or sentiment in electronic text is a goal.***

Keywords: computational linguistics, Sentimental analysis. Deep language models, Natural language processing

I. INTRODUCTION

In recent years, sentiment analysis, a multidisciplinary field at the nexus of biometrics, text analysis, computational linguistics, and natural language processing, has experienced remarkable growth and application across multiple domains. Affective states and subjective information embedded in textual data are systematically identified, extracted, quantified, and analyzed in this dynamic field of research. There are various applications for this tool, such as analyzing sentiments in customer-generated content such as reviews and surveys, monitoring sentiments on social media, and its utilization in healthcare for tasks ranging from marketing and customer service to clinical medicine. Deep learning language models have made sentiment analysis more capable of handling even more difficult data domains, such as news texts, where writers frequently use less direct language to convey their thoughts and feelings. Sentiment analysis's primary task is to classify text's polarity at different granularities, such as the document, sentence, feature/aspect, or feature level Finding out if a statement, paragraph, or particular entity feature/aspect expresses a neutral, positive or negative opinion is the goal. Furthermore, sentiment analysis has advanced beyond the simple categorization of polarities to include the identification of complex emotional states like surprise, fear, anger, sadness, and pleasure. The General Inquirer, which offered early insights into quantifying design in textual data, and psychological research that attempted to infer a person's mental state based on an analysis of their verbal nature are the forerunners of sentiment analysis. Sentiment analysis is not without its difficulties, though. Firstly, the interpretation of opinion words can be context-dependent, with a word considered positive in one context possibly bearing a negative connotation in another. Secondly, human expressions of opinions are often complex and variable, posing a significant challenge for automated analysis. Unlike traditional text processing, which assumes that slight differences in text do not alter the overall meaning, human expressions can be inherently contradictory. Considering these challenges, the pursuit of developing algorithms capable of accurately identifying and classifying sentiment in electronic text remains an ongoing endeavor. In recent years, numerous applications and enhancements of sentiment analysis algorithms have been proposed, each striving to address the intricacies and complexities of sentiment analysis. The motive of this research paper is to have a comprehensive analysis of the development of sentiment analysis algorithms, looking into possible design choices and assessing the algorithms' performance in various processing scenarios. Through exploring the developments and difficulties in this area, we hope to add to the current conversation about sentiment analysis and its growing range of uses.

.

II. LITERATURE REVIEW

Opinion mining, which is another term for sentiment analysis, is a field of study that looks at people's beliefs, attitudes, and feelings regarding a variety of topics, such as individuals, products, businesses, and services.[1] A variety of methods and software tools are being developed for sentiment analysis, and the ability of publicly available online services to classify and score text according to sentiments is being examined and contrasted.

Sentiment analysis is challenged by the subtlety and complexity of language use, nonstandard and innovative language, and the absence of paralinguistic information. The construction of valence- and emotion-association lexicons is done manually and automatically, and research into sentiment composition and sentiment detection in figurative and metaphorical language is still underway.[2] Sentiment analysis is especially helpful for analyzing viewpoints that are expressed in user-generated content, such social media posts, tweets, and product reviews. Using the Naïve Bayes model, sentiment analysis was conducted regarding COVID-19 vaccinations in the Philippines. Twitter is used to collect statistics on Filipinos' opinions of the government's initiatives in the country. Over the past ten years, sentiment analysis has gained popularity as a study topic in the domains of data mining and natural language processing. In tasks involving sentiment analysis, deep neural network (DNN) models have demonstrated encouraging outcomes. Long short-term memory (LSTM) models and their derivatives, like the gated recurrent unit (GRU), have garnered interest in sentiment analysis. The large dimensionality of feature space and the equal weighting of various features in LSTM are addressed by the suggested ABCDM model.[3] Predicting a text's sentiment polarities—positive, negative, or neutral—is the goal of sentiment analysis. Arabic sentiment analysis is challenging because of the language's ambiguity, diversity of dialects, rich morphology, lack of contextual information, lack of express sentiment words in implicit text, and complexity. [4] Deep learning is regarded as the most advanced model in Arabic sentiment analysis since it has demonstrated success in sentiment analysis. Tweets in the English and Filipino languages were divided into positive, neutral, and negative polarity using the Naïve Bayes model. During the COVID-19 pandemic, online learning also became more popular, raising questions about its efficacy and acceptability for both teachers and students. [5] Using a Twitter dataset, the study analyses people's opinions regarding e-learning. Deep learning methods are used to achieve this. Many learning models, such as random forest and SVM classifier, are used for sentiment classification. With Bag of Words (Bow) features, these models achieve a high accuracy of 0.95. [6] To determine the issues with e-learning, such as children's difficulties understanding online education, lagging efficient networks for online education, and uncertainty about campus opening dates, topic modelling is done. Reddit users' opinions on online courses are generally divided, but when posts about online classes are the subject of particular attention, opinions become more intense. [7] While educators' sentiments are more logical and upbeat, learners' tendencies are more negative, reflecting feelings of melancholy and self-doubt. The study emphasizes the necessity of a longer monitoring period in order to determine the efficacy of problem-solving techniques in online courses. [8] In order to promote learner engagement and the development of a community of practice, it is recommended that educators post more frequently. Teachers' posts are essential in fostering discussion.

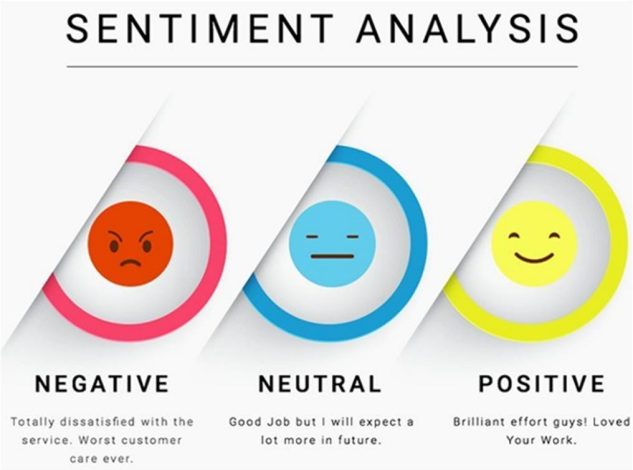


Fig.1 Classification Algorithms
Source- devpost.com

III.PROPOSED SOLUTION

The development of a method to precisely analyses the sentiments on the supplied datasets or feedback is the suggested remedy in the study article on sentimental analysis based on the random forest classification algorithm. The principal objective is to enhance the sentimental analysis tool's efficiency through the application of machine learning techniques and the random forest algorithm. The system will concentrate on gathering crucial data, such as product reviews and customer feedback, in order to ascertain which of the sentiments in the provided data are most likely to be positive or negative.

To achieve this, a large dataset containing a range of datasets and content for the analysis and output will be used to build a robust machine learning model. The accuracy of this model is its most crucial feature; it gathers resources and selects the best answer based on the situation at hand. The model will undergo multiple training and validation processes, each based on a different dataset, in order to confirm its viability and accuracy. The datasets will be analyzed and their patterns comprehended by this system. This algorithm was selected because of its ensemble learning technique, which improves the output and model's accuracy. Following submission of the feedback to the system, the random forest algorithm will analyse the data and provide a list of likely positive and negative responses. The analysis of trends in the commercial and private sectors will be aided by the suggested solution. After analyzing the output provided by this model, it will assist businesses and other manufacturers in raising the caliber of their products. Other use cases include the ability for app developers to fully work on the application's future scopes and upgrades, fix bugs, and enhance the program in accordance with user or customer requirements. A SQL database will also house the collected feedback data and product quality forecasts to ensure simple data management and accessibility. By implementing this recommended fix, the research hopes to significantly improve the Sentimental Analysis Model's accuracy and efficacy. Early and precise product reviews and analysis can lead to earlier reviews and accurate decision outcomes. This system has the potential to completely transform the review and recommendation industry, ultimately benefiting users and consumers, by providing a dependable and flexible tool for sentiment analysis.

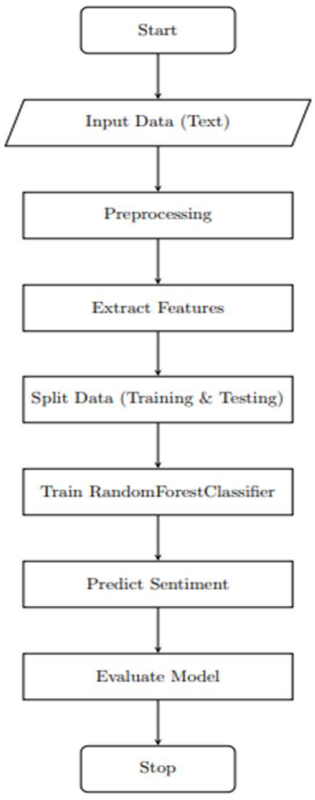


Fig.2 Sentiment Analysis Flowchart using Random Forest Classifier Model [15]

A. Natural Language Processing

What precisely is "natural language processing"? Individuals frequently ask questions like this one. In short, natural language processing is an artificial intelligence field that is expanding quickly. It allows computers to automatically understand and synthesize human language, allowing them to work with text data in a way that is comparable to human comprehension. It distinguishes itself by extracting meaning from written or spoken words for tasks like text analysis and chatbot conversations.

"Natural Language Processing is the eloquent fusion of computer science and linguistics, unlocking the potential for machines to decipher and communicate with the intricate tapestry of human language." – Yoshua Bengio

B. Types Of Algorithm Used In Sentimental Analysis

Sentiment analysis, of course, involves using machine learning algorithms to automatically categorize text data into different sentiment categories, such as positive, negative, or neutral. In this process, a range of machine learning algorithms are commonly employed for sentiment analysis tasks:

1) *Random Forests:* This ensemble learning technique creates multiple decision trees during training, with the mode of the classes being generated as the prediction. Random forests can effectively handle large amounts of text data and complex feature interactions, making them valuable for sentiment analysis tasks.
2) *Support Vector Machines (SVM):* SVM is a supervised learning algorithm that divides text into various sentiment categories by employing a hyperplane to classify data points. Its ability to handle complex classification tasks and non-linear relationships between features makes it an effective choice for sentiment analysis applications.
3) *Logistic Regression:* By estimating the likelihood that a text will fall into a specific sentiment category, logistic regression—a popular statistical model for binary classification tasks—can also be used in sentiment analysis. Its simplicity and interpretability make it popular in various sentiment analysis scenarios.
4) *Naive Bayes Classifier:* [10] This probabilistic classifier assumes that a feature's presence in a class is independent of the presence of other features, based on the Bayes theorem. Because naive Bayes classifiers are straightforward and effective, they can handle massive amounts of text data processing for sentiment analysis applications.
5) *Recurrent Neural Networks (RNNs):* RNNs, designed for sequential data processing, are adept at capturing context and dependencies in text data. They are instrumental in understanding the sentiment of a text within the context of surrounding words, making them suitable for sentiment analysis tasks.
6) *Convolutional Neural Networks (CNNs):* CNNs are typically used for image recognition, but by treating text data as one-dimensional sequences, they can be modified for sentiment analysis as well. Their proficiency in identifying specific patterns and characteristics in textual data is advantageous for sentiment analysis assignments.

C. Why Random Forest Is Superior Then Other Algorithm In Sentimental Analysis

Random Forest emerges as a superior algorithm in the domain of sentiment analysis due to a variety of distinctive advantages that render it exceptionally well-suited for processing textual data and effectively classifying sentiments. Its superiority over other algorithms in sentiment analysis is primarily attributed to the following key factors:

1) *Effective Handling of Large Feature Sets:* Random Forest demonstrates remarkable proficiency in handling an extensive array of features, thereby making it an ideal candidate for sentiment analysis tasks involving intricate textual data with numerous attributes. This capacity allows the algorithm to capture complex relationships and patterns within the data, leading to more precise and accurate sentiment classification outcomes.
2) *Robustness Against Overfitting:* Leveraging an ensemble learning approach through the construction of multiple decision trees, Random Forest effectively addresses and mitigates overfitting challenges that often arise in intricate sentiment analysis tasks. By amalgamating predictions from various trees, it can deliver more generalized and dependable sentiment classification results, thereby minimizing the risk of overfitting to noise prevalent in the data.
3) *Variance Reduction:* Through the amalgamation of multiple decision trees by means of randomly selected features and data samples, Random Forest curtails the variance in sentiment analysis predictions. This characteristic ensures that the algorithm yields more stable and consistent sentiment classification outcomes, thereby enhancing the reliability and dependability of the analysis results.
4) *Proficient Handling of Non-linear Relationships:* Random Forest adeptly manages non-linear relationships inherent within textual data, enabling it to grasp intricate interactions among diverse sentiment indicators. This capability renders it especially suitable for sentiment analysis tasks entailing nuanced and context-dependent language patterns, thereby enabling the extraction of meaningful insights from the data.
5) *Resilience to Noisy Data:* Random Forest exhibits a higher degree of resilience in the face of noisy data and outliers in comparison to certain other algorithms, thereby making it more robust when confronted with noisy textual data. Its capability to aggregate predictions derived from multiple decision trees aids in mitigating the impact of noisy data points, thereby ensuring more accurate and reliable sentiment classification, even in the presence of inconsistencies or inaccuracies within the text.

IV.IMPLIMENTATION

The study of how emotions are conveyed in textual material, such as news articles, reviews, and social media posts, is known as sentiment analysis. It belongs to the natural language processing field. Its primary objective is to classify the sentiment as positive, negative, or neutral. Numerous industries, such as social media, market research monitoring, and customer feedback analysis use this technology. The sentiment analysis workflow consists of several key stages. It commences with data collection, where a dataset of text documents is assembled, each tagged with a sentiment label. Data quality and variety greatly influence model performance.

Data preprocessing follows, involving tasks like tokenization, lowercase conversion, and the removal of stop words, punctuation, and special characters. Stemming and lemmatization are additional options to standardize text.

Feature extraction is crucial for enabling machine learning models to work with text. TF-IDF (Term Frequency-Inverse Document Frequency) vectorization and word embeddings like Word2Vec and GloVe are common techniques.

Model selection is a significant decision, with options ranging from traditional machine learning algorithms (e.g., Logistic Regression, Naive Bayes) to deep learning models (e.g., RNNs, CNNs, Transformers like BERT).

Model training occurs on a dataset split into training and testing sets. The model learns to associate text features with sentiment labels, with fine-tuning and hyperparameter adjustments.

Performance evaluation uses metrics like accuracy, precision, recall, and the F1-score. Post-training refinements enhance accuracy.

Following successful model training, deployment takes place, often through APIs or web services. Periodic retraining with new data ensures the model remains current, and ongoing monitoring identifies and rectifies inaccuracies.

In summary, sentiment analysis plays a crucial role in helping businesses and organizations gain insights from textual data, make informed decisions, and enhance customer experiences. Its applications span multiple sectors, making it an integral part of contemporary data analysis and decision-making processes.

V. METHODOLOGY

In this research, we describe a thorough approach for doing sentiment analysis, a key job in Natural Language Processing (NLP). To accomplish effective sentiment analysis, we combine natural language processing approaches with the powerful Random Forest algorithm. This approach provides a systematic and adaptable framework for assessing sentiments in text data, allowing insights into the emotional tones conveyed in diverse channels such as social media and consumer reviews. To guarantee the robustness of our technique, we use a broad dataset acquired from Kaggle, which includes text-based material from many platforms, and divide it into training, testing, and validation sets to allow for thorough examination. Data collection is the first step in the approach, and we get text-based datasets via Kaggle. These datasets are derived from a variety of sources, including social media sites and consumer reviews, and provide a broad and realistic depiction of text data with a range of emotional expressions. Our sentiment analysis model's reliability and generalizability are ensured by the inclusion of training, testing, and validation datasets. Subsequently, the collected data undergoes meticulous cleansing and preparation, a critical step to ensure uniformity and data integrity. We employ the Pandas library to clean and format the data, thereby creating a consistent and high-quality dataset for analysis. [11] Textual data visualization techniques, facilitated by Matplotlib and Seaborn, play a pivotal role in this phase. These visualization tools aid in gaining a deeper understanding of the distribution of sentiments within the Kaggle dataset, enabling us to identify prominent words and phrases associated with different emotional tones. Following that, the obtained data is meticulously cleansed and prepared, which is a vital step in ensuring consistency and data integrity. The Pandas library is used to clean and format the data, resulting in a consistent and high-quality dataset for analysis. Textual data visualization tools, made possible by Matplotlib and Seaborn, are critical at this phase. These visualization tools help us obtain a better grasp of the sentiment distribution within the Kaggle dataset, allowing us to spot popular words and phrases linked with various emotional tones. Feature extraction is a critical phase in the process. We use the scikit-learn CountVectorizer to convert the text input into numerical feature vectors that the Random Forest algorithm can analyze. The Random Forest method was chosen for our model due to its success in text categorization tasks. Prior to model training, the hyperparameter tuning phase leverages the GridSearchCV tool to systematically explore a variety of hyperparameter values in order to discover the ideal configuration for the Random Forest model. This painstaking method guarantees that our sentiment analysis algorithm is fine-tuned for optimal performance. The assessment step, which is the last in the methodology, is where the model's performance is methodically assessed. We employ a range of performance metrics, such as accuracy, precision, recall, and F1-score, to evaluate the model's effectiveness.[9] Testing on an independent dataset and using cross-validation techniques evaluate the model's performance. The assessment measures used are determined by the specific objectives of the sentiment analysis task, ensuring that the model satisfies the appropriate performance requirements. Our study article extensively covers the model's performance, providing a complete analysis of its capabilities.

The findings serve as a foundation for essential changes and fine-tuning, leading to the ongoing advancement of sentiment analysis across a variety of text-based datasets and applications. In conclusion, this work proposes a careful and adaptable sentiment analysis approach that successfully combines data pretreatment, text visualization, text processing, feature extraction, model selection, and rigorous assessment.[12] This approach establishes the Random Forest algorithm as a reliable analytical tool suitable for a broad range of text-based datasets and applications, while also advancing the field of sentiment analysis through the use of several Python modules and specialized tools.

VI.RESULT

In this project, the individual sentiments of the classification model are assessed. The categorization models were created using the review data, and their accuracy was assessed by contrasting the predicted and actual reviews produced by the models. With an accuracy of about 96%, the decision tree-based classification model has been found to be the best performing model. F1-score and recall, the other two factors, also provided prediction scores of roughly 96% and 97%. The impact of each factor on the estimated amount was investigated. It was found that for every algorithm used, a person's review and rating status had the greatest impact on the prediction.

```
Accuracy_score:  0.961
Precision_score:  0.9616228070175439
Recall_score:  0.9532608695652174
--------------------------------------------------------------------
              precision    recall  f1-score   support

           0       0.96      0.97      0.96      1080
           1       0.96      0.95      0.96       920

    accuracy                           0.96      2000
   macro avg       0.96      0.96      0.96      2000
weighted avg       0.96      0.96      0.96      2000
```

Fig 4: Result

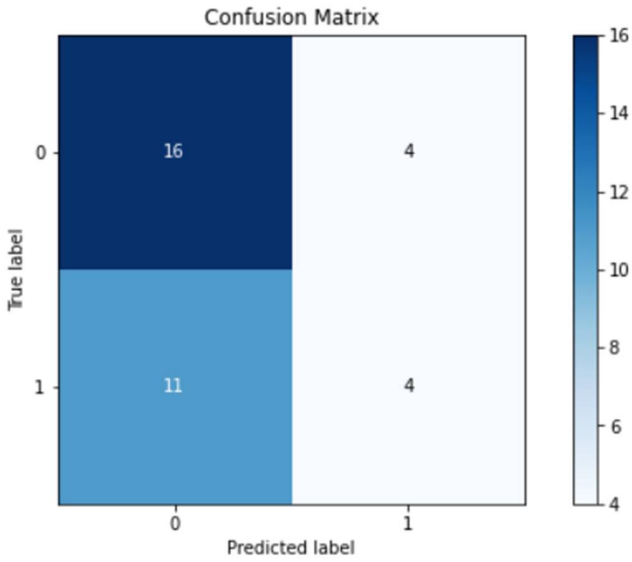


Fig5. Confusion matrix

VII. CONCLUSION AND FUTURE WORK

Finally, we used a product review dataset from Kaggle.com and a range of machine learning classification models to predict feelings based on certain variables. We acquired a fantastic accuracy rate of 96% in determining the emotions represented in comments through the evaluation of our classification models. Several major criteria emerged from our analysis, with a person's review and star rating standing out as the most influential variables in all of the algorithms used. This finding emphasizes the importance of individual input and star ratings in sentiment prediction.

The study emphasizes the importance of sentiment analysis in product assessment, focusing on comprehensive analysis of product attributes rather than just comments and conditions. Future data collected can be used to predict sentiments effectively, benefiting individuals and companies in making informed purchasing decisions. Future research should consider additional factors influencing consumer sentiments, such as product price, brand reputation, and delivery speed, and extend the analysis to different product types and industries. Collaborative efforts between product owners and data scientists can enhance sentiment prediction models and facilitate data-driven product improvements.

REFERENCES

[1] Chakravarthi, B. R., Priyadharshini, R., Muralidaran, V., Suryawanshi, S., Jose, N., Sherly, E., & Mccrae, J. P. (2020). Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text. https://en.wikipedia.org/wiki/Malayalam

[2] Huang, F., Li, X., Yuan, C., Zhang, S., Zhang, J., & Qiao, S. (2021). Attention-Emotion-Enhanced Convolutional LSTM for Sentiment Analysis. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2021.3056664

[3] Kumar, S., Gahalawat, M., Roy, P. P., Dogra, D. P., & Kim, B. G. (2020). Exploring impact of age and gender on sentiment analysis using machine learning. Electronics (Switzerland), 9(2). https://doi.org/10.3390/electronics9020374

[4] Marcec, R., & Likic, R. (2021). Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. Postgraduate Medical Journal. https://doi.org/10.1136/postgradmedj-2021-140685

[5] Naseem, U., Razzak, I., Khushi, M., Eklund, P. W., & Kim, J. (2021). COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis. IEEE Transactions on Computational Social Systems, 8(4), 976–988. https://doi.org/10.1109/TCSS.2021.3051189

[6] Wang, L., Niu, J., & Yu, S. (2020).IEEE Transactions on Knowledge and Data Engineering, 32(10), 2026– 2039. https://doi.org/10.1109/TKDE.2019.2913641

[7] Wang, T., Lu, K., Chow, K. P., & Zhu, Q. (2020). COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model. IEEE Access, 8, 138162–138169. https://doi.org/10.1109/ACCESS.2020.3012595

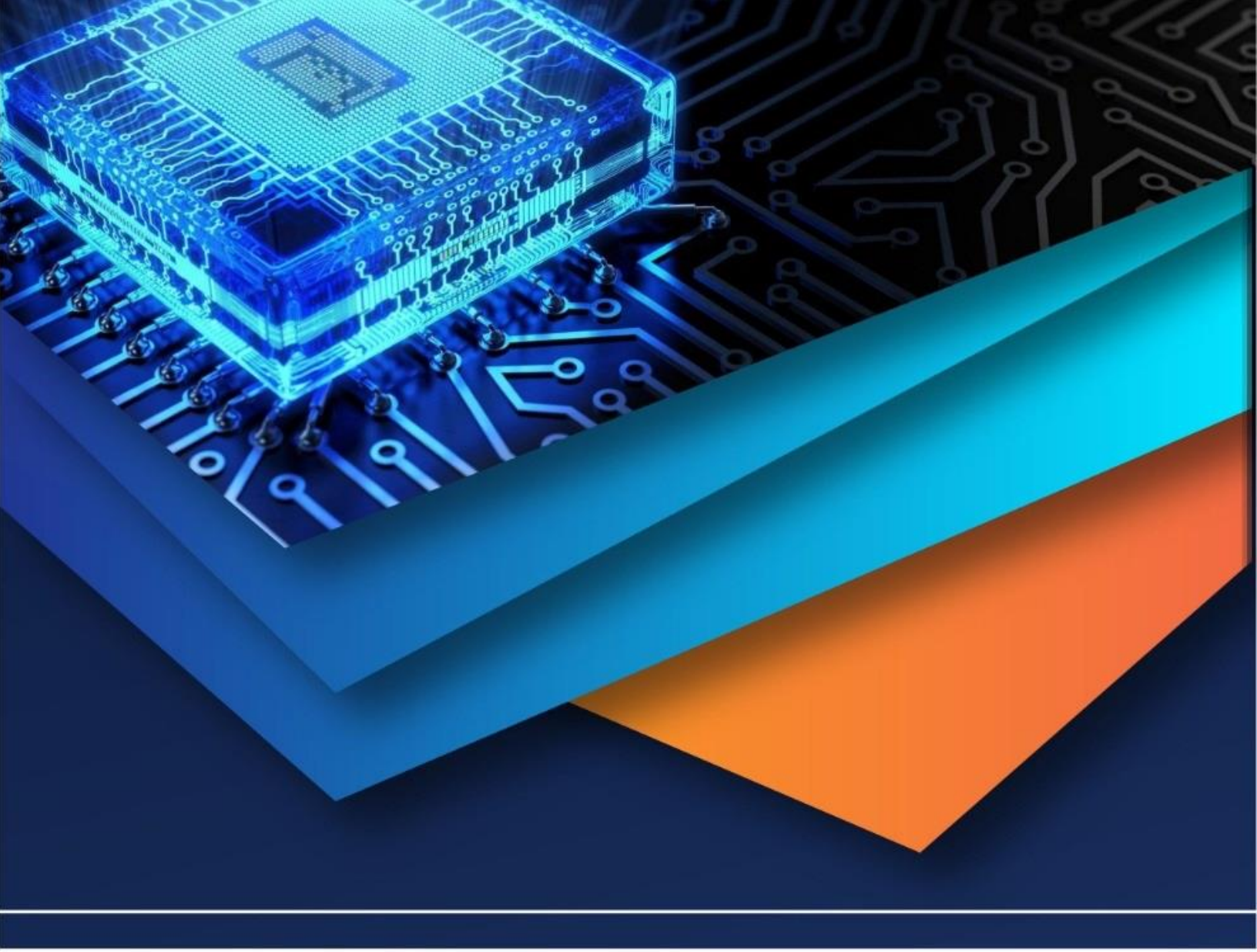
[8] Yang, L., Li, Y., Wang, J., & Sherratt, R. S. (2020). Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning. IEEE Access, 8, 23522–23530. https://doi.org/10.1109/ACCESS.2020.2969854

[9] Zulfadzli Drus, Haliyana Khalid, Sentiment Analysis in Social Media and Its Application: Systematic Literature Review, Procedia Computer Science,Volume 161,2019, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.11.174.

[10] S. Smetanin, "The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives," in IEEE Access, vol. 8, pp. 110693-110719, 2020, https://doi.org/10.1109/ACCESS.2020.3002215.

[11] Pooja, Bhalla, R. A Review Paper on the Role of Sentiment Analysis in Quality Education. SN COMPUT. SCI. 3, 469 (2022). https://doi.org/10.1007/s42979-022-01366-9

[12] Siddhaling Urologin, "Sentiment Analysis, Visualization and Classification of Summarized (IJACSA), 9(8), 2018. http://dx.doi.org/10.14569/IJACSA.2018.090878



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY