iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** X **Month of publication:** October 2023

DOI: https://doi.org/10.22214/ijraset.2023.55930

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 11 Issue X Oct 2023- Available at www.ijraset.com

Amazon Fine Food Review Analysis

Shreyas Khandale[1], Prathamesh Patil[2], Rohan Patil[3]
BE (Computer) Fourth year 2023, Computer Engineering Department of Computer Engineering AISSMS COE Pune: 411011, Maharashtra, India

Abstract: ***This paper presents a comprehensive examination of the Amazon Fine Food Reviews dataset, encompassing both descriptive and predictive analyses. The study involved employing various text mining techniques, dimensionality reduction approaches, and linear regression models to forecast review scores. Ultimately, a model was developed that achieved a Root Mean Square Error (RMSE) of 1.0936, enabling accurate predictions for new reviews.***
Keywords: Amazon Fine Food Reviews dataset, Descriptive and predictive analyses, Text mining techniques, Linear regression models, Root Mean Square Error (RMSE)

I. INTRODUCTION

The dataset comprises 568,454 reviews of fine foods from Amazon, covering a span of over a decade from October 1999 to October 2012. In order to uncover patterns and insights within this dataset and subsequently forecast review scores based on various attributes, initial descriptive analyses were conducted. These revealed a highly imbalanced distribution of scores and demonstrated that scores exhibited different patterns over time. Following this, text processing techniques such as TF-IDF vectorization and Principal Component Analysis were employed to extract and refine features from the text for input into the linear regression models. Subsequently, diverse weighting approaches for the training samples and various types of linear regression models were investigated and assessed. One model excelled at predicting low scores while another demonstrated proficiency in predicting high scores.

In the era of e-commerce, online reviews have emerged as a vital source of information for consumers. Given Amazon's status as one of the largest online marketplaces, it accumulates extensive volumes of review data. This report endeavours to deliver a thorough analysis of Amazon's review data through the application of data science methodologies. By distilling valuable insights from the data, we aim to gain a deeper comprehension of consumer sentiments, product trends, and the factors that influence customer satisfaction.

II. LITERATURE REVIEW

Sentiment Analysis in Product Reviews Numerous studies have delved into sentiment analysis, which involves using natural language processing techniques to discern the sentiment or emotion expressed within a text. In the context of product reviews, this has been applied to diverse domains such as electronics, books, and movies. Understanding consumer sentiment is crucial for businesses to refine their products and services (Liu, 2012). Temporal Analysis of Reviews Temporal analysis of product reviews has gained prominence as it provides insights into how consumer opinions evolve over time. Studies have shown that consumer sentiments may be influenced by various factors including seasonal trends, marketing campaigns, and product quality changes (Lu et al., 2011). Feature Extraction from Text Data The process of extracting meaningful features from text data is pivotal in building accurate predictive models. Techniques like TF-IDF (Term Frequency-Inverse Document Frequency) and Principal Component Analysis (PCA) have been widely used to identify relevant attributes in textual data (Mikolov et al., 2013; Jolliffe, 2002). Predictive Models for Review Scores The use of linear regression models for predicting review scores has been a common approach. However, research has also explored more sophisticated techniques such as ensemble methods and deep learning architectures to improve predictive accuracy (Bishop, 2006; Bengio et al., 2013).

III. METHODOLOGY

Here are some of the methods frequently used in Amazon reviews:

A. Data Collection

How to do Amazon reviews; The first step is to collect relevant review data. This can be done by using Amazon's API or by accessi ng publicly available information. Profiles can be written for specific products, brands or time periods of interest. Care must be take n to ensure that the information is correct and that the instructions and conditions or restrictions of use are observed.

B. Data Pre-Processing

Once audit data is collected, a preliminary step should be taken to clean the data and prepare it for analysis. This will include removi ng duplicates, handling missing values, and creating a standard design. Analyzing text usually requires some preliminary steps, such as using techniques such as tokenization (breaking the text into individual words or symbols), removing unnecessary words (words such as "the", "and", "is"), and omission. or lemmatization reduces words to their basic forms.

C. Evaluation of Thought

Evaluation of thought is a popular method for determining the thought expressed in a message. It involves positive, negative, or neut ral evaluations based on underlying emotions. Machine learning algorithms (such as Naive Bayes, Support Vector Machine (SVM)) or deep learning models (such as Recurrent Neural Networks (RNN) or Convolutional Neural Networks (CNN)) can learn based on labeled inspections of data. new words. Additionally,dictionarybased methods use a predefined dictionary or dictionary to assign sen timent scores to individual words or phrases und-er analysis.

D. Sampling

Sampling is an unsupervised study used to identify misconceptions or concepts in a collection of analysis. Algorithms such as Laten t Dirichlet Allocation (LDA) or Non Negative Matrix Factorization (NMF) are often used for this purpose. These algorithms analyz e the frequency and occurrence patterns of words in the analysis to identify different topics

Each topic represents a group of words that are often spoken together, allowing businesses to understand the topics discussed in the review and learn about the product business, FAQs or new models.

E. Analysing The Distribution Of Points

Analyzing the distribution of points can provide valuable information regarding customer satisfaction. This analysis involves collect ing and visualizing the distribution of customergenerated star ratings. Measures such as mean, median, and standard deviation can b e calculated to summarize the distribution. Differences over time, differences between products or brands, and comparisons with ind ustry standards can be analyzed to identify trends and areas for improvement.

F. Text Summarization

Analyzing the writing process turns long analysis into a short summary that captures the key point. Information.The summarization method sorts and selects key phrases or sentences in the analysis to create a summary. Abstract summarization techniques create ne w concepts by interpreting or summarizing concepts. This process can be used to create important products or provide a summary of customer feedback; thus, making it easier for businesses to understand the key points in the review.

G. Comparative Analysis

Comparing reviews of different products or products can give businesses insight into their relative performance and user sentiment. This analysis involves collecting and comparing sentiment scores, product ratings, or frequency trends across different products or b rands. Statistical tests such as t-tests or chisquare tests can be used to identify significant differences in opinions or ratings. By benc hmarking against competitors or industry leaders, companies can understand their strengths and weaknesses and identify areas for i mprovement or product differentiation.

H. Spam Detection

Spam reviews can affect the accuracy of product ratings and customer perception. Machine learning algorithms can be trained to ide ntify and filter spam messages. Characteristics such as comment length, comment frequency, unusual thought patterns, or suspicious behavior can be used to identify spam messages. Feature engineering and other technologies.

IV. RESULTS

A. Descriptive Analysis

Statistics including minimums, averages, medians, and maximums for text length, summary length, score, and helpfulness ratio are contained in Table 1.

Table 1: Descriptive Statistics

	Text Length	Summary Length	Score	Helpful Ratio
min	3	0	1	0
avg	79.0964	4.1131	4.1832	0.4079
med	56	4	5	0
max	3377	42	5	1

The line graphs of average text length, summary length, score, and helpfulness ratio aggregated by day over time are contained in Figure 1. There were very great fluctuations in average review length, summary length, score, and helpfulness ratio from day to day. between 2004 and 2007. After 2007, average review length and summary length remained quite stable, while average score slightly reduced, and average helpfulness ratio fairly dropped, especially after 2010.

The LDA analysis performed on the review texts in an unsupervised manner suggests five potential topics with the following top 10 keywords:

1) food, dog, like, eat, dogs, treats, love, loves, just, good. This topic seems to represent good reviews on dog food.
2) tea, flavor, like, coffee, taste, good, chocolate, just, cup, drink. This topic seems to represent good reviews on drinks, especially tea.
3) like, good, taste, great, just, flavor, love, chips, salt, really. This topic seems to represent good reviews on chips.
4) coffee, amazon, product, price, good, great, order, just, buy, box. This topic seems to represent good reviews on coffee products.
5) product, water, like, sugar, use, taste, oil, just, good, bottle. This topic seems to represent good reviews on cooking ingredient such as water, sugar and oil.

B. Graphs

As we can see in the following graphs, there were very great fluctuations in average review length, summary length, score, and helpfulness ratio from day to day between 2004 and 2007. After 2007, average review length and summary length remained quite stable, while average score slightly reduced, and average helpfulness ratio fairly dropped, especially after 2010.

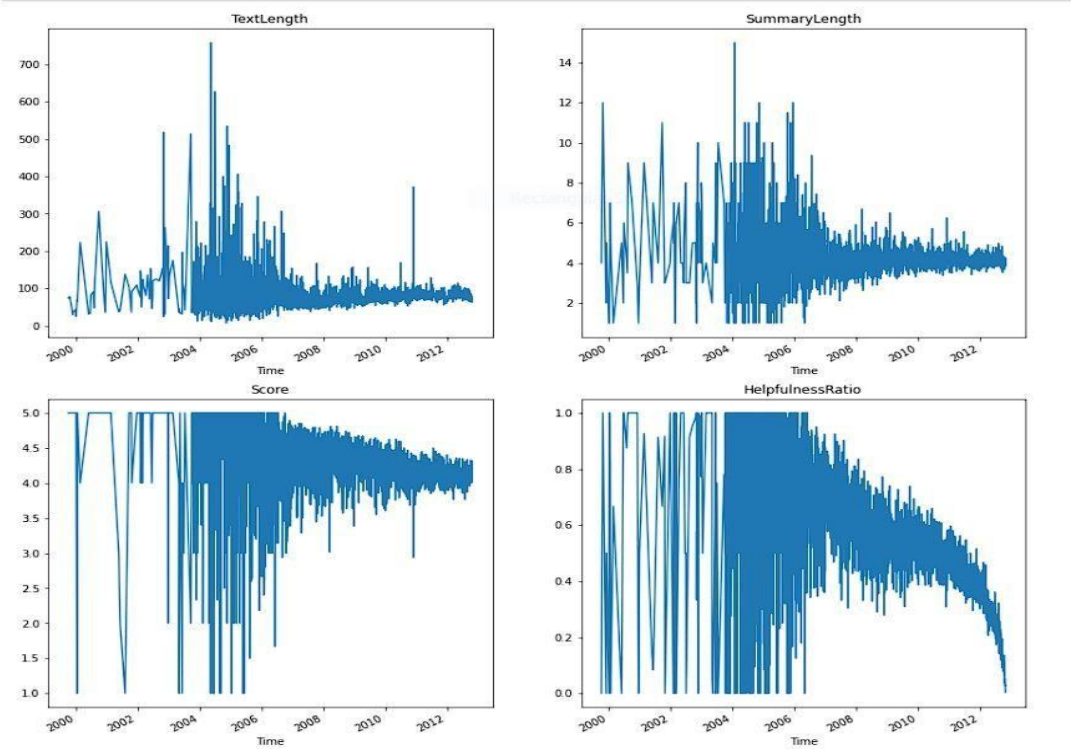


Figure1: Text analysis of reviews

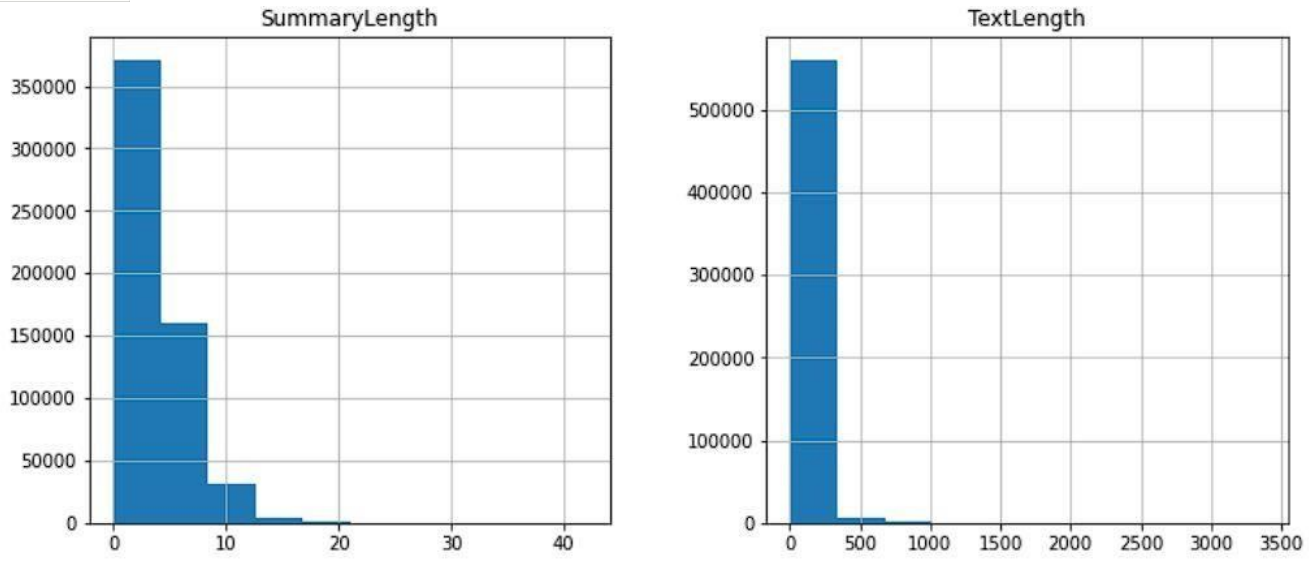


Figure 2: A histogram for Text length and summary length:

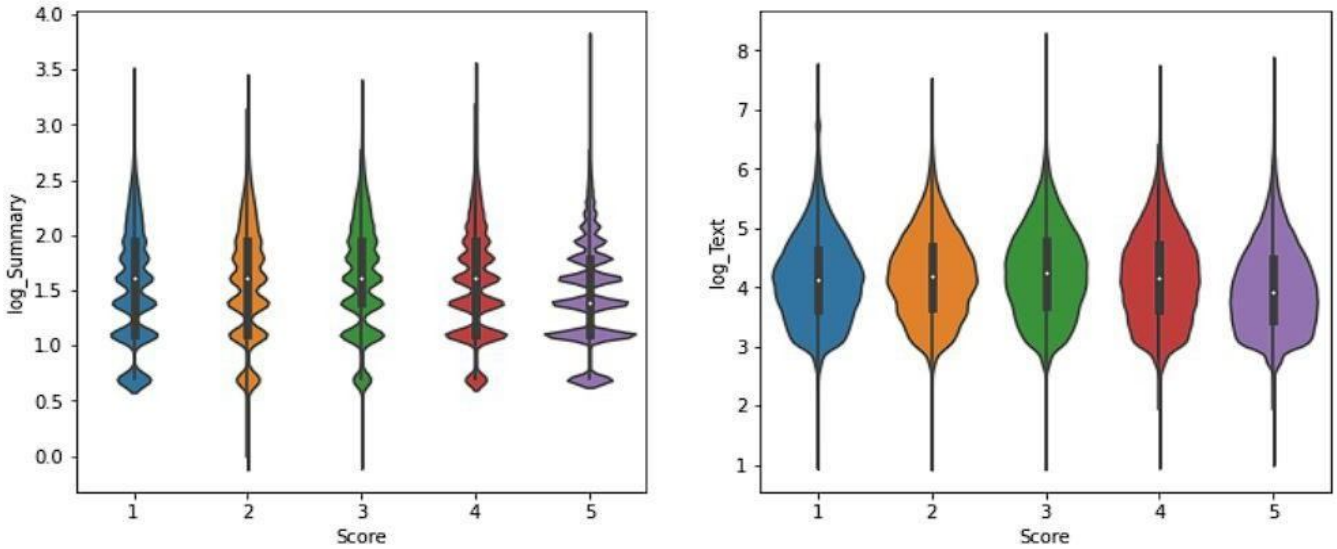


Figure 3: Violin plots for relationship between scores and summary/text lengths

V. CONCLUSION

Overall, analysis of Amazon reviews provides a good insight into the customer's thoughts, experiences, and satisfaction with the vari ous products and services available on the platform. Through this analysis, we gained an indepth understanding of the factors that in fluence customer reviews and their impact on the business.

Amazon review analysis allows businesses to identify patterns, trends and needs expressed by customers. By reviewing content, bus inesses can discover valuable information about product features, quality, pricing, customer service and delivery. This information may be used to determine insights to improve product development, marketing strategies and customer support.

In addition, sentiment analysis provides a complete view of customer referrals, allowing businesses to determine positive, negative o r neutral: a product or brand. Such insights can help companies measure customer satisfaction, identify areas for improvement, and r espond to customer feedback in a timely manner.

Analyzing ratings and statistics allows the business to evaluate the overall popularity and visibility of its products. Regularly analyzi ng products with high ratings and comprehensive reviews can inform marketing strategies and influence customer decisions. Conver sely, addressing issues raised in negative reviews can help improve product quality, customer satisfaction, and reputation.

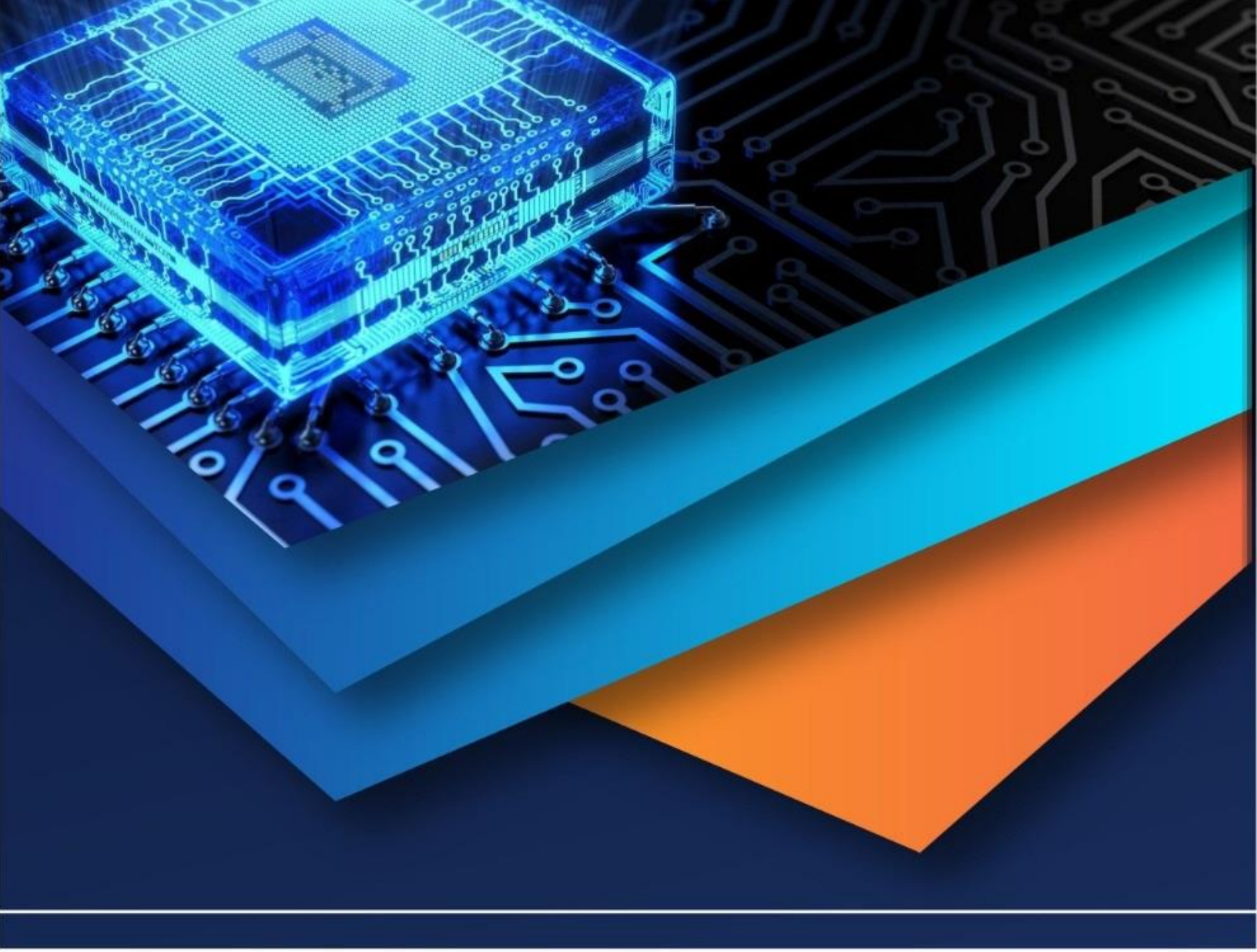
Additionally, examining reviews can reveal important information about products or services that customers appreciate or admire. N o found. Businesses can use this information to prioritize product development, solve health problems, and deliver great customer ex periences.

But it's important to be aware of the limitations of Amazon review reviews. The accuracy and reliability of reviews may be affected by factors such as fake reviews, bias, or manipulation by competitors. Businesses should implement quality control measures to filte r out false or fraudulent reviews and ensure the integrity of reviews.

As a result, Amazon customer reviews provide businesses with powerful tools to understand customer sentiment, identify areas for i mprovement, and increase orders. By using the best customer feedback on the platform, companies can improve their products, impr ove their marketing strategies and finally offer people a unique experience in a competitive e-commerce environment.

REFERENCES

[1] S. Kapadia, “Topic Modelling in Python: Latent Dirichlet Allocation (LDA)” in Towards Data Science.
[2] Y. Berdugo, “Review Rating Prediction: A Combined Approach” in Towards Data Science
[3] Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online bookreviews. Journal of Marketing Research, 43(3), 345-354.
[4] Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. (Journal of Marketing, 70(3), 74-89)
[5] Vermeulen, I. E., & Seegers, D. (2009). Tried and tested: The impact of online hotel reviewson consumer consideration. Tourism Management, 30(1), 123-127.
[6] Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelpreview fraud. Management Science, 62(12), 3412-3427.
[7] Zhang, Y., Ye, Q., & Law, R. (2010). Determinants of customer satisfaction with online travelagencies: A structural model. Journal of Travel Research, 49(3), 324-336.
[8] S. AlZu'bi, A. Alsmadiv, S. AlQatawneh, M. Al-Ayyoub, B. Hawashin and Y. Jararweh, "A Brief Analysis of Amazon Online Reviews," 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 2019, pp. 555-560, doi: 10.1109/SNAMS.2019.8931816.
[9] S. Maurya and V. Pratap, "Sentiment Analysis on Amazon Product Reviews," 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), Faridabad, India, 2022, pp. 236-240, doi: 10.1109/COM-IT-CON54601.2022.9850758.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY