Sentiment Analysis of Amazon Reviews Data Using Logistic Regression

P. Arumugam*, S.R Sakthi Malaviga**

[A] *Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India*

Abstract

Amazon reviews give Indian consumers an opportunity to express their thoughts and personal experiences with products, assisting new clients in making educated purchases and providing sellers and manufacturers with insightful feedback. By identifying the emotional tone and viewpoints conveyed in text data, sentiment analysis is a tool used in data mining to extract insights from huge amounts of data. This process yields useful information for business analysis and decision-making. In this work, sentiment analysis is applied to review datasets through the use of logistic regression for classification. This technique aids in the logical understanding of consumer opinions by classifying the sentiment of reviews. Consequently, the study's results were successfully classified using logistic regression, and the model's accuracy and reliability in sentiment prediction were confirmed by a thorough assessment of its performance. These methods enable the model to be applied in a wide range of socioeconomic domains, including public opinion tracking, consumer feedback analysis, and market research. It helps companies and decision-makers make data-driven decisions, which eventually improves goods, services, and public opinion.

Keywords: Amazon reviews, sentiment analysis, logistic regression, accuracy and prediction

1. Introduction

The international technology giant Amazon.com, Inc., with its headquarters in Seattle, Washington, has an Indian subsidiary called Amazon India. Bengaluru, Karnataka, is home to Amazon India's headquarters, which it opened in June 2013. With a market share of more than 30 % the company is among the biggest e-commerce businesses in India. Products from electronics, literature, apparel, home and kitchen, grocery, and more are available on Amazon India. In addition, the business provides a variety of services, including Amazon Web Services, Amazon Pay, and Amazon Prime. Sentiment analysis is a branch of natural language processing and data mining that focuses on identifying the sentiment or emotional tone of text by removing subjective information from it. Analysing and comprehending the beliefs, attitudes, and feelings expressed in textual data is the aim of this field. Sentiment analysis uses methods like machine learning and linguistic analysis to classify text as positive, negative, or neutral. This allows it to provide insightful information on a variety of topics, including social media monitoring, customer feedback analysis, and business. This field is essential to comprehending public opinion and sentiment regarding goods, services, or events since it uses online textual data. Kaura, Sukhadan, et al. (2016)(18) investigated the use of sentiment analysis in decision support systems, concentrating on the binary classification of movie reviews using logistic regression. filled in the gaps in the field's algorithm selection. Suresh Dara (2017)(5) researched sentiment analysis in stock market prediction, displaying excellent accuracy in categorizing sentiments from social media and attaining top results with SVM using Twitter data and VADER. Kataresada Ketaren et al. (2017)(10) Investigated bid/no-bid decisions in construction firms, utilizing logistic regression models to achieve an 87 % prediction accuracy rate. determined important variables, such as experience and market competition. Heuristic Abhilasha Tyagi et al. (2018)(1) Used k-fold cross-validation and logistic regression to analyze sentiment in social media. For quicker sentiment classification, the Effective Word Score heuristic was introduced. The related literature provided insights into a variety of sentiment analysis applications across a wide range of domains, such as natural language processing, stock market prediction, bid/no-bid decisions in construction companies, social media sentiment evaluation, and sentiment classification in product reviews. These research demonstrated the effectiveness of machine learning techniques, including Logistic Regression, Support Vector Machines (SVM), and Naive Bayes, in extracting sentiment from textual data. The concepts and techniques from the similar studies served as a framework for my research study, making it easier to use a unique strategy designed especially for sentiment analysis of Amazon reviews using logistic regression.

2. Objective of the study

- Develop a sentiment analysis model using logistic regression to classify Amazon reviews into positive or negative sentiments based on the textual content of the reviews.

- The objective is to accurately predict the sentiment of a review to help users quickly understand its tone.

*Corresponding author

**Principal corresponding author

Email addresses: sixfacemsu@gmail.com (P. Arumugam), sakthimalaviga33521@gmail.com (S.R Sakthi Malaviga)

- Evaluate the performance of the logistic regression model using relevant metrics such as accuracy, precision, recall, and F1-score.

3. METHODOLOGY

3.1. DATA COLLECTION

In this study, the Amazon review dataset, comprising 4914 text reviews collected from GitHub, is examined. These reviews encompass a wide array of products found on the Amazon platform and include text feedback generated by users alongside ratings. The substantial size of the dataset indicates its potential as a valuable resource for conducting sentiment analysis. This dataset is deemed suitable for sentiment analysis, allowing for the exploration of customer sentiments and opinions regarding the diverse range of products available on Amazon's platform, thus facilitating informed decision-making for both consumers and sellers. To gain insights into the distribution of overall ratings ranging from 1 to 5, a bar chart was employed. This visual representation allowed for a clear and concise depiction of the frequency of each rating within the dataset.

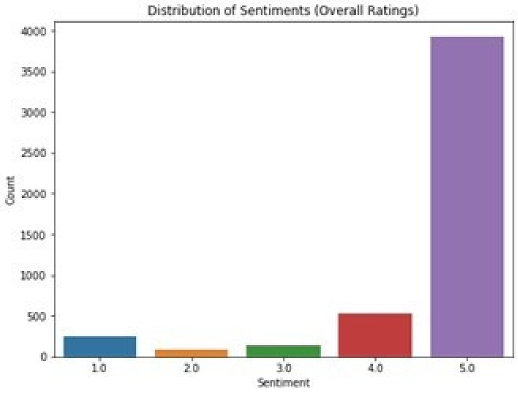

Figure 1: Distribution of Overall Ratings

accordance with their appearance in the text and place any table notes below the table body. Please avoid using vertical rules and shading in table cells.

The fig.1, The distribution of overall ratings was visualized in a bar chart, revealing interesting patterns. Notably, the rating of 5 had the highest frequency, while the rating of 2 was considerably less common. Further Correlation Heatmap was utilized to examine the presence or absence of multicollinearity, a critical assumption check. Fig.2 is the visualization of the correlation plot. There was a strong positive correlation between the total number of votes and the helpfulness of reviews, indicating that more votes tend to be associated with reviews marked as helpful. On the other hand, there was a weak negative correlation between the time difference and overall rating, suggesting that as time passed, the overall rating tended to decrease slightly.

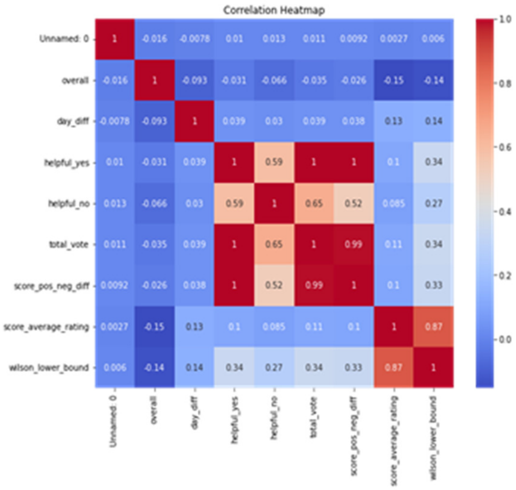

Figure 2: Correlation plot of the dataset

3.2. TEXT PROCESSING

In this study of sentiment analysis, text preprocessing was executed as an essential preliminary step to enhance the quality and reliability of subsequent analytical processes. The purpose of text preprocessing was to prepare the Amazon review dataset by systematically performing various operations on the textual content. Punctuation removal, including the elimination of commas, periods, and other punctuation marks, was carried out to ensure that the text was devoid of extraneous symbols. Conversion of all text to lowercase was performed to maintain uniformity in the dataset, as it was crucial for accurate word analysis.

3.3. SENTIMENT FEATURE EXTRACTION

In this study, the subsequent step involves the extraction of sentiment features from the pre-processed data in a systematic manner. This process begins by analysing the pre-processed text data word by word, assigning sentiment scores to each word using a pre-existing sentiment lexicon. The sentiment scores are then aggregated for each review, producing an overall sentiment score. This score indicates whether a review is perceived as positive, negative, or neutral. The output of this step is a comprehensive sentiment analysis of the Amazon review dataset, enabling the classification of reviews into these sentiment categories. This categorization serves as a valuable resource for understanding customer sentiments and opinions about various products.

3.4. SENTIMENT CLASSIFICATION VIA LOGISTIC REGRESSION

Following the extraction of sentiment features, sentiments are classified using logistic regression. The assumptions of logistic regression were thoroughly examined, and it was determined that these assumptions are met, allowing the logistic regression model to be successfully applied to the data. The essential assumptions, such as linearity, independence of observations, and absence of multicollinearity, were verified and found to be valid. A Correlation Heatmap was utilized to examine the presence or absence of multicollinearity, a critical assumption check. This heatmap visualizes the relationships between variables by displaying correlation coefficients in this step, the sentiment features are employed as input variables, and a logistic regression model is applied to categorize sentiments. The model assigns each review into one of the sentiment categories, be it positive, negative, or neutral, based on the features identified during the extraction process.

3.5. DATA SPLITTING

In the logistic regression phase, the dataset is divided into two subsets: a training dataset and a testing dataset. The training dataset is used to train the logistic regression model, allowing it to learn from the extracted features and their corresponding sentiments. The testing dataset, on the other hand, remains untouched during the training process and is solely employed to evaluate the model's performance.

3.6. MODEL VALIDATION AND EVALUATION

Following the model fitting, the next step involves validating the model using various metrics, including accuracy, recall, precision, and the F1-score. These metrics collectively assess the model's performance in accurately classifying sentiments. Additionally, a Receiver Operating Characteristic (ROC) curve is generated to visually represent the model's ability to discriminate between different sentiment classes. The ROC curve provides a graphical depiction of the model's true positive rate and false positive rate at various threshold settings. The model's accuracy in classifying sentiment based on textual reviews was a primary focus of our analysis. The Receiver Operating Characteristic (ROC) curve displayed an impressive Area Under the Curve (AUC) value of 0.97. This indicates that our binary classification model excels at distinguishing between positive and negative sentiments, demonstrating high accuracy, sensitivity, and specificity. An AUC close to 1 signifies strong model performance.

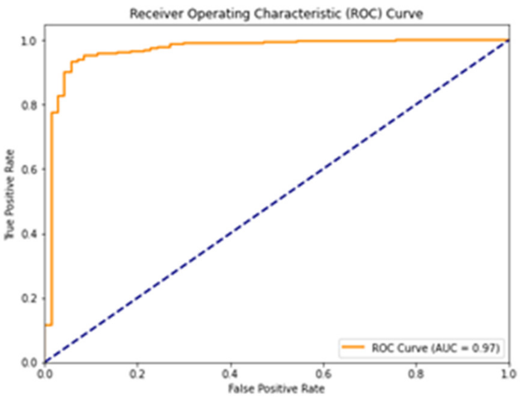

Figure 3: The Receiver Operating Characteristic (ROC) curve

Our model achieved an accuracy of 0.96, implying that it correctly predicted the sentiment of 96% of instances in the test dataset. This metric showcases the overall effectiveness of our model in classifying sentiments.

With a precision of 0.97, our model performed exceptionally well in identifying true positive instances among those predicted as positive. This indicates that 97% of the reviews classified as positive were indeed positive.

The model achieved a high recall of 0.99, implying that it correctly identified 99% of the actual positive instances in the dataset. This metric underscores the model's ability to capture positive sentiments effectively.

The balanced F1-score of 0.98 combines precision and recall, providing an overall assessment of our model's effectiveness. An F1-score close to 1 suggests that our model maintains a balance between precision and recall, making it reliable for sentiment classification.

4. RESULTS AND DISCUSSION

This study employed a bar chart (Fig.1) to analyze the distribution of overall ratings (ranging from 1 to 5). The visualization revealed that a rating of 5 was the most frequent, while a rating of 2 was less common. A correlation heatmap (Fig.2) was then utilized to assess multicollinearity, indicating a strong positive correlation between total votes and review helpfulness, and a weak negative correlation between time difference and overall rating. The focus then shifted to the model's sentiment classification accuracy. The Receiver Operating Characteristic (ROC) curve demonstrated an impressive Area Under the Curve (AUC) of 0.97, highlighting the model's proficiency in distinguishing between positive and negative sentiments. The model achieved an overall accuracy of 96%, with a precision of 97%, indicating high correctness in identifying positive instances. A recall of 99% underscored the model's effectiveness in capturing positive sentiments, while a balanced F1-score of 0.98 affirmed the model's reliability in sentiment classification, maintaining a harmonious balance between precision and recall.

5. CONCLUSIONS

In conclusion, the study has provided a comprehensive analysis of sentiment in Amazon reviews, employing a range of preprocessing, feature extraction, and logistic regression modelling techniques. The sentiment analysis model successfully classified sentiments, offering valuable insights into customer feedback. The dataset's distribution of ratings was visualized, offering a broad perspective on customer satisfaction. Additionally, the exploration using a Correlation Heatmap confirmed the absence of multicollinearity, affirming the integrity of the analysis. The fitted model has been evaluated using the validation matric and further confirmed the model's effectiveness. This study's findings indicate a powerful sentiment analysis tool, which holds promise for real-world applications and decision-making processes

Acknowledgements

Thanks to Manonmaniam Sundaranar University Tirunelveli, the second author has received grants from Manonmaniam Sundaranar University Tirunelveli, Tamilnadu under the scheme of Bharat Rathna Dr. M. G. Ramachandran Centenary Research Fellowship

References

[1] Abhilasha Tyagi & Naresh Sharma, ”Sentiment Analysis using Logistic Regression and Effective Word Score Heuristic”,International Journal of Engineering and Technology,20-23, 2018.

[2] Akanksha Halde, Aditi Uttekar, and Amit Vishwakarma, ”Sentiment Analysis on Amazon Product Reviews” Published in the International Journal of Engineering and Technology. DOI: 10.14419/ijet.v8i5.23706,2021.

[3] Ammar Rashed Hamdallah ,”Amazon Reviews Using Sentiment Analysis”, Published in the International Journal of Information Technology in. DOI: 10. 1007/s41870-022-00629-z,2022.

[4] Arwa S. M. AlQahtani, ”Product Sentiment Analysis for Amazon Reviews” by Published in the International Journal of Computer Applications, DOI: 10. 5120/ijca2020933782, 2020.

[5] Dara, S.A sentiment analysis of food review using logistic regression, International Conference on Machine Learning and Computational Intelligence, 1-6, 2017.

[6] Fayyad, Usama, Piatetsky-Shapiro, Gregory, & Smyth, Padhraic, ”From Data Mining to Knowledge Discovery in Databases.” AI Magazine, 17(3), 37-54,1996.

[7] Haque, T.U., Saber, N.N., & Shah, F.M. Sentiment analysis on large scale Amazon product reviews, In 2018 IEEE International Conference on Innovative Research and Development (ICIRD), 1-6. IEEE,2018.

[8] Han, J., & Kamber, M. Data mining: Concepts and techniques: Morgan Kaufmann,2012.

[9] Hu, Minqing, & Liu, Bing, ”Mining and Summarizing Customer Reviews.” Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 168-177,2004

[10] Kataresada Ketaren & Novdin M. Sianturi ,”Decision Making Modelling with Logistic Regression Approach”,International Journal of Applied Engineering Research,9067-9073,2017.

[11] Kothari, C.R. ”Research Methodology : Methods and Techniques.” New Age International,1985.

[12] Liu, B., Hu, M., & Chen, S. A survey of text-based sentiment analysis methods. ACM Computing Surveys (CSUR), 51(4), 85,2019.

[13] Liu, Bing.”Sentiment Analysis and Opinion Mining.” Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers,2012.

[14] Mishra, R.K., Srivastava, S., & Rathore, S.S. Logistic regression for sentiment analysis: A gentle introduction. International Journal of Computer Applications, 174(11), 1-6, 2020.

[15] Nousi, C., & Tjortjis, C. A methodology for stock movement prediction using sentiment analysis on Twitter and Stocktwits data. International Hellenic University, 2021

[16] Paknejad, S. Sentiment classification on Amazon reviews using machine learning approaches, International Journal of Computer Applications, 181(11), 1-8,2018.

[17] Shmueli, G., Patel, N.R., & Bruce, P.C. Data mining for business analytics: concepts, techniques, and applications in R. Wiley, 2016.

[18] Sukhnandan Kaura & Rajni Mohanaa, ”Prediction of Sentiment from Textual Data Using Logistic Regression Based on Stop Word Filteration and Volume of Data”,International Journal of Control Theory and Applications,2016.

[19] Tanjim Ul Haque, Nudrat Nawal Saber, and Faisal Muhammad Shah, Sentiment Analysis on Large Scale Amazon Product Reviews” by. Published in the International Journal of Computer Science and Information Security,DOI: 10.18178/ijcsis.2023.11.1.001, 2023.

[20] Wankhade, M., Rao, A.C.S., Dara, S., & Kaushik, B. A sentiment analysis of food review using logistic regression, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10(4), 228-235, 2020.

[21] Wen, C., Wu, J., & Chen, D. Analysis of text emotion based on logistic regression model. In 2022 IEEE 5th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), 183-188,IEEE,2022.

[22] Witten, Ian H., Frank, Eibe, Hall, Mark A., & Pal, Christopher J. ”Data Mining: Practical Machine Learning Tools and Techniques.” Morgan Kaufmann,2016.

[23] Zhang, X., Liu, Z., & Wang, Q. A comparative study of logistic regression and deep learning models for sentiment analysis. arXiv preprint, arXiv:2103.02435, 2021.

[24] Zhang, Y., Yang, Z., & Zhang, J. Logistic regression for sentiment analysis of customer reviews. Expert Systems with Applications, 188, 115915,2022.