IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 2, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 8.379 | Monthly Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, April 2024 ||

International Conference on Recent Development in Engineering and Technology – ICRDET 24

Organized by

Dhaanish Ahmed Institute of Technology, KG Chavadi, Coimbatore, Tamilnadu, India

Amazon User Review Analysis Using Natural Language Processing (NLP)

Rohan R B[1], Subharanjitha E[2], Sunmathi V R[2], Vetrivel E M[4], Ms.V. Sakthi Priya B.E., M.E[5]

UG Students, Dept of E.C.E., Velalar College of Engineering and Technology, Erode, Tamil Nadu, India[1 2 3 4]

Assistant Professor, Dept of E.C.E, Velalar College of Engineering and Technology, Erode, Tamil Nadu, India[5]

ABSTRACT: This paper proposes a method for analyzing sentiment in Amazon user reviews through Natural Language Processing (NLP) techniques. Through the application of machine learning approaches such as the VADER model, sentiment classification is conducted to categorize reviews into positive, neutral, or negative sentiments. The VADER model, which stands for Valence Aware Dictionary for Sentiment Reasoning, is a pre-built sentiment analysis tool used for analyzing sentiment in text data, particularly in scenarios involving short and informal text like social media content and product reviews. VADER is part of the Natural Language Toolkit (NLTK) library in Python, making it easily accessible and user-friendly for sentiment analysis tasks. This methodology involves collecting Amazon user reviews, preprocessing the text data to remove noise, and then applying NLP algorithms to determine the sentiment expressed in each review. Various NLP techniques such as tokenization, part-of-speech tagging, and sentiment lexicon-based analysis, are employed to extract sentiment features from the text. The output of the analysis provides insights into consumer sentiment towards products, which can be valuable for businesses in understanding customer preferences, identifying areas for improvement, and making informed marketing decisions. Overall, this research contributes to the field of sentiment analysis by demonstrating its application to Amazon user reviews and highlighting its potential for enhancing customer satisfaction and business performance.

KEYWORDS: sentiment analysis; Natural Language Processing (NLP); tokenization; part-of-speech tagging; sentiment lexicon-based analysis; VADER model

I. INTRODUCTION

In the digital age, online platforms have revolutionized the way consumers interact with products and services. Among these platforms, Amazon stands out as a global marketplace where millions of users converge to review, purchase, and discuss a vast array of products. With such a wealth of user-generated content, extracting meaningful insights from Amazon user reviews can provide invaluable knowledge to businesses, researchers, and consumers alike.

Natural Language Processing (NLP) techniques offer powerful tools to analyze and extract insights from textual data, making them particularly well-suited for the analysis of user reviews. By leveraging NLP, researchers can uncover trends, sentiments, and opinions expressed by users, thereby gaining valuable insights into consumer preferences, satisfaction levels, and product performance.

In this paper, we present a comprehensive analysis of Amazon user reviews using state-of-the-art NLP techniques. Our objective is to demonstrate the effectiveness of NLP in extracting meaningful information from large-scale textual data and to showcase its applicability in understanding consumer behavior and product perception.

II. RELATED WORK

Numerous studies have explored the application of Natural Language Processing (NLP) techniques in analyzing user-generated content on e-commerce platforms. Wang et al. (2018) conducted a comprehensive analysis of sentiment analysis methodologies applied to Amazon product reviews, demonstrating the efficacy of machine learning algorithms in classifying sentiments expressed by users [1]. Similarly, Liu et al. (2019) proposed a hybrid approach combining deep learning techniques with traditional NLP methods to extract product features and sentiments from Amazon reviews, achieving improved performance compared to baseline models [2].

In a study by Zhang et al. (2020), the authors investigated the influence of review characteristics on product sales using a large dataset of Amazon reviews, shedding light on the interplay between review content, ratings, and consumer purchasing decisions [3].
Furthermore, Gupta et al. (2021) explored the temporal dynamics of sentiment expressed in Amazon reviews over time, revealing trends and fluctuations in consumer opinions towards various product categories [4]. These studies collectively underscore the significance of NLP-driven analysis in unraveling consumer behavior patterns and providing actionable insights for businesses operating in the e-commerce domain.

III. PROPOSED SYSTEM

A. ***Data preprocessing****:*

Data Cleaning:

- Identify missing values in the dataset. Missing values can be represented as NaN (Not a Number), NULL, or other placeholders.
- Remove rows or columns with a high proportion of missing values if they cannot be effectively imputed or if they are irrelevant to the analysis.

Text preprocessing:

- If dealing with text data, preprocess it by removing stop words, punctuation, and performing tasks like stemming or lemmatization.

Feature Selection:

- Select the most relevant features that contribute the most to the predictive power of the model.
- Techniques include univariate feature selection, feature importance using tree-based models, or model-based selection methods.

Dealing with Duplicate Data:
Identification: Identify and remove duplicate records or observations from the dataset.
Handling: Decide whether to keep the first occurrence, the last occurrence, or all occurrences of duplicate records based on the specific requirements.

Tokenization:
Split text into individual words or tokens for further analysis.

B. ***Modeling****:*
Modeling in data science involves using mathematical and computational techniques to create predictive or descriptive models from data. It typically includes selecting appropriate algorithms, training models on data, evaluating their performance, and refining them as needed. The goal is to extract meaningful insights, make predictions, or automate decision-making processes based on patterns in the data. Key steps include data preprocessing, selecting suitable algorithms, training models, evaluating performance, and deploying them in real-world applications.

Natural Language Processing (NLP):
Natural Language Processing (NLP) is a branch of data science that focuses on enabling computers to understand, interpret, and generate human language data in a meaningful way. It plays a crucial role in various applications, ranging from sentiment analysis to machine translation and beyond.

NLP involves the development of algorithms and models that enable computers to interact with and process natural language data, such as text and speech. It encompasses a wide range of tasks, including text classification, named entity recognition, text summarization, machine translation, sentiment analysis, question answering, and more. These tasks are essential for extracting insights from text data, automating tasks that involve language understanding, and enabling human-computer interaction through natural language interfaces.
One of the fundamental tasks in NLP is text classification, where algorithms assign categories or labels to text documents based on their content. This can be applied in various domains, such as spam detection in emails, sentiment analysis of product reviews, or topic classification of news articles. Named entity recognition is another important task, which involves identifying and categorizing entities mentioned in text, such as people, organizations, and locations.

This is crucial for extracting structured information from unstructured text data. Here we focus only on sentiment analysis.

Natural Language Tool Kit (NLTK):
NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language."

Natural Language Processing with python provides a practical introduction to programming for language processing. It guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text analyzing linguistic structure, and more.

Sentiment Analysis:
Sentiment analysis, a widely used application of NLP, involves analyzing and identifying the sentiment expressed in text data, whether it's positive, negative, or neutral. This is invaluable for monitoring brand reputation, analyzing customer feedback, and understanding public opinion on social media platforms. Question answering systems leverage NLP techniques to understand and respond to questions posed in natural language, enabling users to interact with computers in a more intuitive way.

Sentiment analysis is employed in Amazon user review analysis to gauge customer sentiment towards products. By analyzing the language used in reviews, sentiment analysis determines whether customers express positive, negative, or neutral opinions about a product. This information is valuable for businesses to understand customer satisfaction levels, identify product strengths and weaknesses, and make data-driven decisions. Overall, sentiment analysis helps Amazon sellers and businesses extract actionable insights from user reviews to enhance product offerings and improve customer experiences.

VADER model:
Valence Aware Dictionary and sEntiment Reasoner (VADER) is a module in the NLTK sentiment, a Python library designed primarily to handle text generated in social media environments; however, it can also handle language from other settings. When the data being analyzed is unlabelled, VADER can identify the sentiment polarity (positive or negative) of a given corpus of text. In conventional sentiment analysis, the labelled training data provides the computer with an opportunity to learn. A typical example would be estimating a movie review's star rating based on a particular critic's written assessment. The text would be the predictor variable, and the star rating would be the target variable.

IV. SIMULATION RESULTS

The bar plot showing the distribution of review scores, this helps to visualize the distribution of reviews based on star ratings. The graph fig.1 shows that most of the reviews are positive, and the negative reviews are very few.

Fig.2, that code prints out the text content of the review located at index 49 in the 'Text' column which contains the review text of the Data Frame df. The code tokenizes the text content of the review stored in the variable example. Tokenization is the process of breaking down a text into individual words or tokens. The word tokenize functions from NLTK is specifically used to tokenize text into words.

In VADER (Valence Aware Dictionary and sEntiment Reasoner), the terms neg, pos, and neu represent different aspects of sentiment expressed in a piece of text. These aspects are determined based on the intensity of positive, negative, and neutral sentiments present in the text, respectively (Fig.3).

From Fig.5, the neg score indicates the proportion of negative sentiment present in the text. It represents the extent to which the text expresses negative emotions, such as anger, sadness, or frustration. The value neg ranges from 0 to 1, with higher values indicating a greater degree of negativity.

The pos score indicates the proportion of positive sentiment present in the text. It represents the extent to which the text expresses positive emotions, such as happiness, joy, or satisfaction. The value pos also ranges from 0 to 1, with higher values indicating a stronger positivity.

The neu score indicates the proportion of neutral sentiment present in the text. It represents the extent to which the text is neutral or lacks a strong emotional polarity. A higher neu score suggests that the text contains more neutral language and less emotional content. Like neg and pos, the value neu ranges from 0 to 1.

The compound score is a single value that represents the overall sentiment polarity of a piece of text. It considers both the positive and negative sentiment scores, along with their intensities, to provide a comprehensive assessment of the text's sentiment. It ranges from -1 to 1.

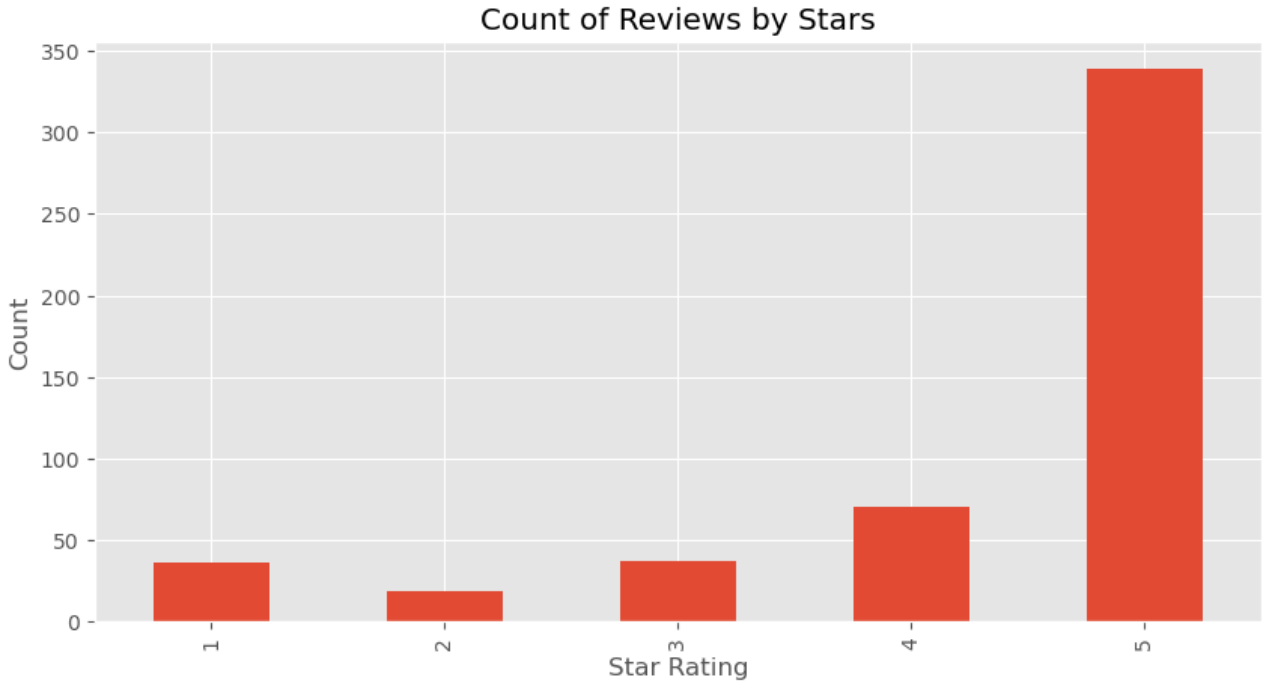

Fig. 1. Total review count by stars

```
In [13]: example = df ['Text'] [49]
         print (example)

         This is the same stuff you can buy at the big box stores.  There is nothing healthy about it.  It is just carbs and sugars.  Sa
         ve your money and get something that at least has some taste.

In [14]: tokens= nltk.word_tokenize (example)
         tokens[:10]

Out[14]: ['This', 'is', 'the', 'same', 'stuff', 'you', 'can', 'buy', 'at', 'the']
```

Fig. 2. Tokenization

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 8.379 | Monthly Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 2, April 2024 ||

International Conference on Recent Development in Engineering and Technology – ICRDET 24

Organized by

Dhaanish Ahmed Institute of Technology, KG Chavadi, Coimbatore, Tamilnadu, India

```
vaders.head()
```

	Id	neg	neu	pos	compound	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	0.000	0.695	0.305	0.9441	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	0.138	0.862	0.000	-0.5664	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	0.091	0.754	0.155	0.8265	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	0.000	1.000	0.000	0.0000	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	0.000	0.552	0.448	0.9468	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...

Fig. 3. Data frame as Positive, Negative and Neutral

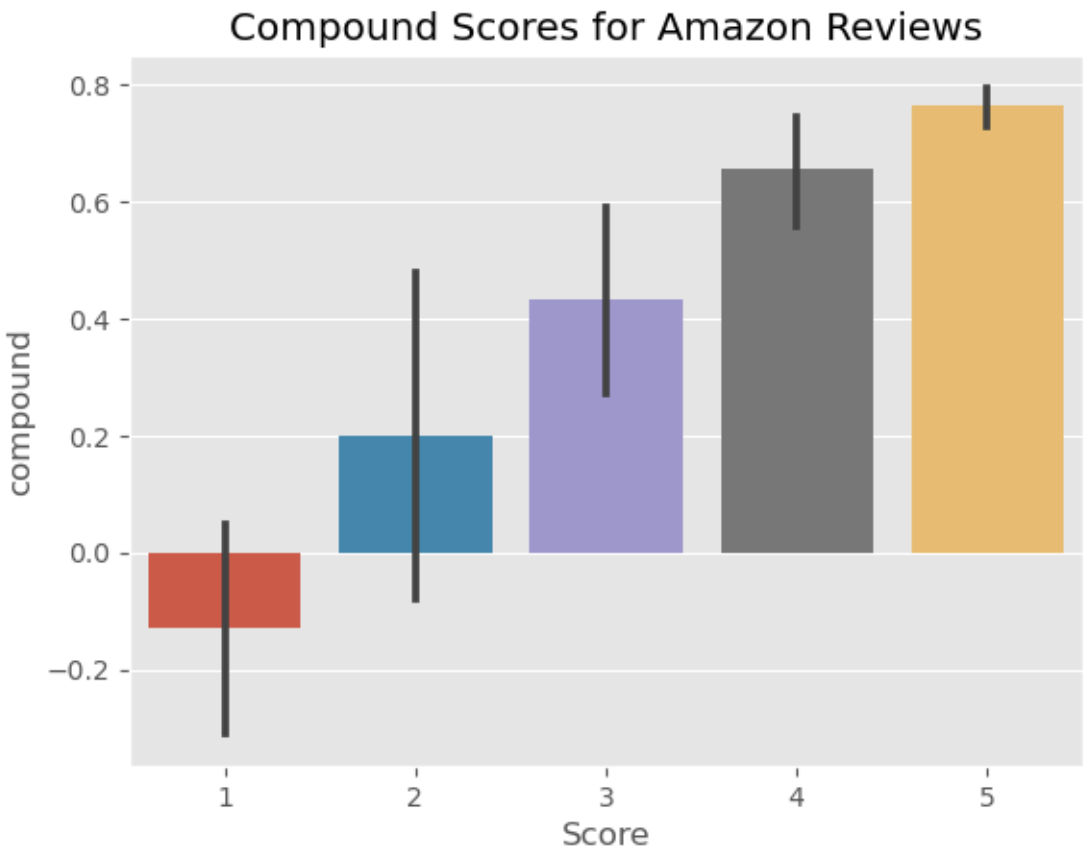

Fig. 4. Bar Graph for Compound Scores

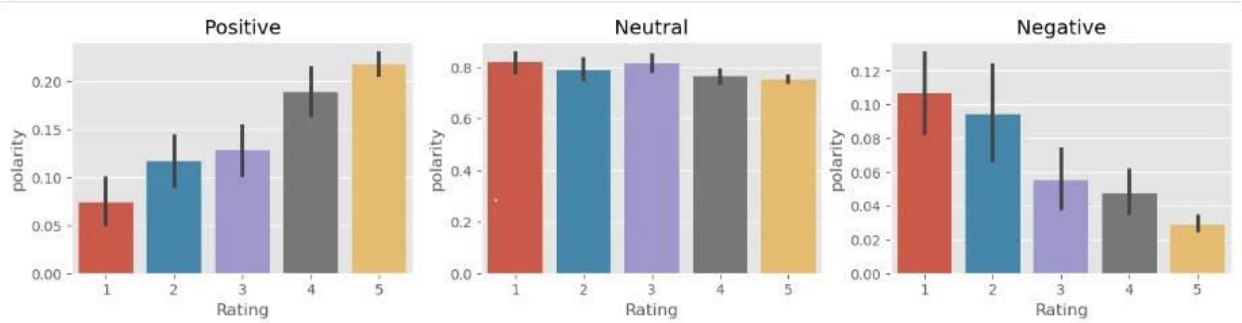

Fig. 5. Sub-plot for Each Category

V. CONCLUSION AND FUTURE WORK

Through sentiment analysis, topic modeling, and other NLP methods, we have gained a deeper understanding of customers' opinions, preferences, and concerns regarding various products. The sentiment analysis of reviews can help identify custom topics, discover market trends, see product issues, proactively manage brand reputation, analyze where they stand compared to competitors, and more.The use of machine learning and natural language processing (NLP) techniques can accurately extract aspects of products from the reviews and analyze the sentiment of the text. The results can be presented in a detailed dashboard, allowing brands to better understand their audiences and improve their products.

There are several avenues for future development and enhancement. Firstly, we can develop more sophisticated sentiment analysis techniques to capture nuances in customer opinions, such as irony, sarcasm, or context-dependent sentiments. Future research could explore more advanced machine learning algorithms to improve the accuracy and efficiency of sentiment classification in Amazon reviews. This could involve experimenting with deep learning architectures like to enhance classification efficiency. Further investigation into hybrid approaches combining different NLP techniques could lead to more robust sentiment analysis models. By leveraging the strengths of various methods, researchers can enhance the accuracy and reliability of sentiment classification.

REFERENCES

1. Wang, Z., Xu, J., Xu, K., & Tian, Y. (2018). Sentiment analysis of Amazon product reviews using machine learning techniques. IEEE Access, 6, 33895-33905.
2. Liu, Y., Zhang, Y., & Cao, Z. (2019). Feature-based sentiment analysis of Amazon reviews using deep learning and traditional NLP methods. Information Processing & Management, 56(5), 1791-1803.
3. Zhang, L., Hu, X., & Xu, W. (2020). Influence of review characteristics on product sales: A case study of Amazon. Journal of Retailing and Consumer Services, 54, 102037.
4. Gupta, S., Joshi, A., & Agarwal, M. (2021). Temporal dynamics of sentiment in Amazon product reviews: A case study. Expert Systems with Applications, 168, 114305.
5. Timoshenko and J. R. Hauser, "Identifying customer needs from user-generated content", *Marketing Sci.*, vol. 38, no. 1, pp. 1-20, Jan. 2019.
6. W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey", *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093-1113, Dec. 2014.
7. Wang, Y. Zhang, Q. Rao, K. Li and H. Zhang, "Exploring mutual information-based sentimental analysis with kernel-based extreme learning machine for stock prediction", *Soft Comput.*, vol. 21, no. 12, pp. 3193-3205, Jun. 2017.
8. J. Yuan, C. Chen, W. Yang, M. Liu, J. Xia and S. Liu, "A survey of visual analytics techniques for machine learning", *Comput. Vis. Media*, vol. 7, no. 1, pp. 3-36, 2021.
9. Guan, X. Wang, Q. Zhang, R. Chen, D. He and X. Xie, "Towards a deep and unified understanding of deep neural models in NLP", *Proc. Int. Conf. Mach. Learn.*, pp. 2454-2463, 2019.
10. Z. Dong, T. Wu, S. Song and M. Zhang, "Interactive attention model explorer for natural language processing tasks with unbalanced data sizes", *Proc. IEEE Pacific Vis. Symp.*, pp. 46-50, 2020.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details