ISSN: 2278-4632 Vol-13 Issue-02 Nov 2023

MINING EMOTIONS FROM AMAZON PRODUCT FEEDBACK: A NATURAL LANGUAGE PROCESSING APPROACH

Dr. Vaka Murali mohan[1], Polemoni Sai bhavya[2], B. Mahesh3, B. Nithin[4], Y. SVS Suhaas [5]

[1]Principal, [2,3,4,5]Students

[1, 2,3,4,5]Department of Artificial Intelligence and Machine Learning

Malla Reddy Institute of Technology and Science, Hyderabad, India.

Email Id: vakamuralimohan@gmail.com, polemonisailubhavya@gmail.com, Maheshmahi7182@gmail.com nithingoud2222@gmail.com, Venkatasaisuhaas@gmail.com

I.ABSTRACT

In the current era of online shopping, digital product reviews hold immense influence over consumer decision-making. Platforms like Amazon provide a space for customers to share their experiences with various products, making it easier for potential buyers to determine which products are worth investing in. Our research delves into the world of sentiment analysis, utilizing Amazon product reviews to gain insights that closely resemble real-world conversations. Consider a scenario where a tech-savvy individual is seeking to purchase a new smartphone from Amazon. They sift through countless reviews, each containing a mix of praise, criticism, and personal anecdotes. Our research examines this scenario on a larger scale, analysing and processing millions of Amazon product reviews across various categories. Using advanced natural language processing techniques, including advanced deep learning models and rule-based systems, we analyse these reviews to understand how customers feel about certain attributes of the product. Just as our hypothetical consumer considers factors such as camera quality, battery life, and usability, our analysis also differentiates sentiment related to specific product attributes. Our analysis offers a diverse range of perspectives that reflect the multifaceted nature of consumer preferences and expectations. We explore the intricacies of consumer psychology, capturing the essence of user experience in the digital marketplace. Additionally, we investigate the impact of various factors, such as review length, user demographic profile, and product category, on expressed sentiment. This information serves as a guide for companies looking to improve their products and helps consumers make informed decisions amidst the abundance of online reviews. Although our research takes place in the digital realm, it has tangible implications for consumer decision-making. We acknowledge the limitations of our approach and discuss potential future research directions in sentiment analysis, highlighting the growing significance of customer sentiment in shaping product choices in the digital age. In essence, our research emphasizes the undeniable importance of sentiment analysis in deciphering consumer preferences. This reflects the daily life experiences of individuals navigating the dynamic e-commerce landscape.

II. INTRODUCTION

Over the past few decades, online markets have gained popularity. As a result, sellers and marketers ask buyers to share their opinions on the products they have purchased. This has led to the creation of millions of reviews across the internet every day, covering various products, services, and locations. As a result, the internet has become the most significant source of ideas and opinions about products and services. In recent years, there has been a significant amount of research in the field of sentiment analysis, with efforts focused on understanding customer feelings. This is due to the shift of the consumer products market towards the internet, changing the direction of the shopping experience and providing a wealth of information on product usage created by users. This differs from traditional marketing activities that rely on word of mouth and advertising. Before cyberspace, we used to ask others about a product or service before buying something. In recent years, e-commerce has taken the lead. Sites like Amazon, eBay, Walmart, etc. provide services for buyers to make online purchases. For companies to improve their business, it is necessary to know buyers' opinions about their products or services. Now there is no need to ask others about the product before buying. Our dataset includes customer reviews and ratings extracted from consumer reviews on Amazon products. We extracted features from the dataset and built some supervised models based on it. These models include not only traditional algorithms such as naive algorithms Bayes, linear support vector machines, K-nearest neighbours, but also deep learning metrics such as Recurrent Neuralization, Convolutional neural networks, and networks. The purpose of this study is to investigate a small part of this big problem: the positive and negative attitudes towards Amazon products and customer behaviour. Sentiment analysis attempts to determine the characteristics of text that demonstrate its nature and context, such as positive, negative, objective, subjective, etc. The goal of this article is to classify the positive and negative customer reviews of various products and build a supervised learning model for a massive polarization number of reviews.

III. DATA SETS

To gain comprehension of the organization and framework of a dataset, let us examine an Amazon Review example [1,2]. In Figure 1, Amazon user reviews encompass four key components:

Summary: Magazine name

Magazine content: The actual content of the review.

Rating: Users rate products on a scale of 1 to 5.

Usefulness: The number of people who found this review helpful.

Analysing these aspects will aid in comprehending and deducing insights from reviews.

Figure 1 Actual Amazon customer review sample

A. Raw data

our raw data was downloaded from the datafinti website, which includes 3,500 records with over 20 columns such as product ID, category, subcategory, review title, revision text, reviewer name, and more. However, the data set contains a lot of 0 values, incorrect prices, and formatting, as well as outliers and noise. In Figure 2, you can observe some columns from the original Amazon consumer product review dataset where certain values are null and in the wrong format.

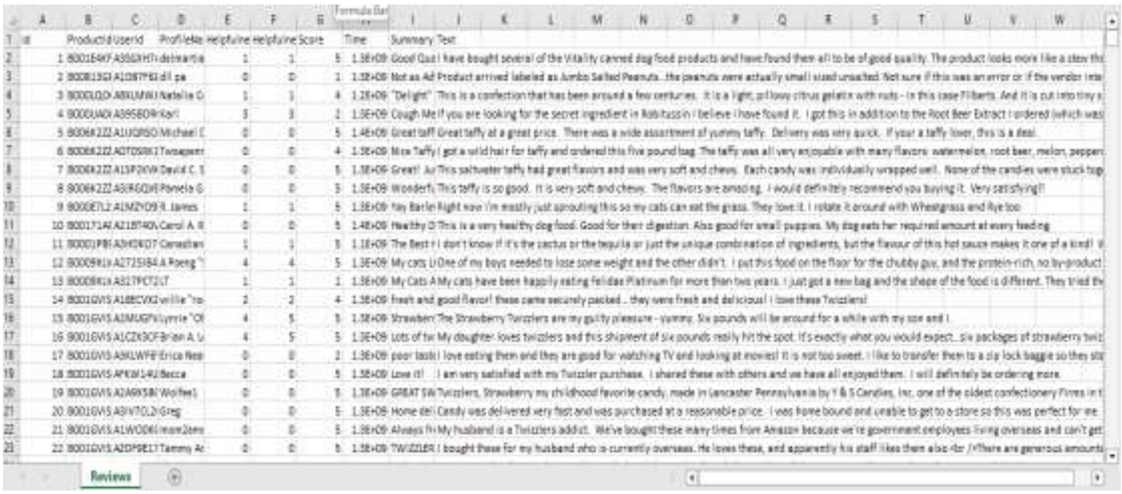

Figure2 sample of raw data

B. Processed data

We trimmed the raw data using Excel, as it had over 20 columns, but reduced it to 10 columns for analysis. Additionally, we formatted some data and price columns using built-in Excel functions for proper formatting. To handle outliers, we used a single built-in Excel function. We removed duplicates and "No" entries using a table filter, including, and excluding with the "Figures" function. The normalized data used for analysis is shown in Figure 3, which had fewer columns than the original dataset at the time.

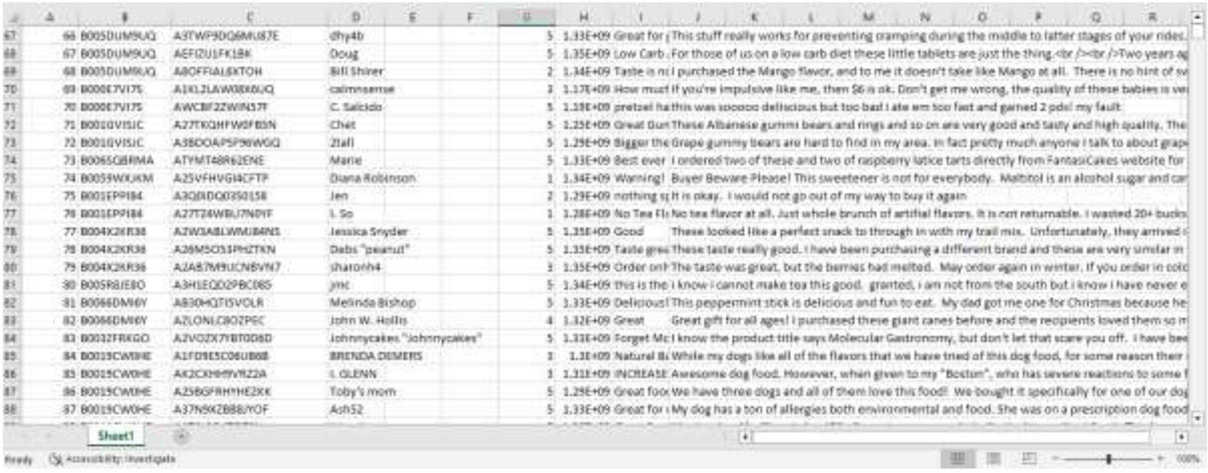

Figure3 Processed data for analysis

C. Data preprocessing

We obtained our data from Amazon consumer reviews, which includes various data points. The data encompasses various examples such as product type, name, text and product reviews. To start our project, we extracted the rating and review columns as they are crucial for our analysis. After examining the data, we found some data points without notes. We removed them. Additionally, we created a score distribution diagram for an overview of the dataset. The diagram revealed that we have five classes, denoted from 1 to 5, with varying distribution. We noticed that grades 1 and 2 have a small amount of data, while grade 5 has more than 20,000 reviews, making the distribution unbalanced. As an example, from our dataset, a review states "So far this product has not disappointed me. My kids love using it, and I love the ability to easily monitor and control what they see", which received a rating of '5'.

D. Data resampling

In order to address the imbalance in our dataset, we conducted tests by resampling the data. Resampling is a common technique used to handle imbalanced data. For this project, we repeatedly sampled data from layers 1, 2, and 3 as these classes had much fewer samples than the other two classes. As a result, we repeated the initial evaluations of labels 1, 2, and 3 15 times in our training set. However, it is important to note that there are many repeating patterns in the training set which may cause the model to over-adapt.

IV. SENTIMENT CLASSIFICATION AND ANALYSIS

Sentiment analysis involves using natural language processing, text analytics, computational linguistics, and biometrics to identify, extract, quantify, and analyse subjective information related to emotional states. [3,4,5] It helps determine the emotional tone or sentiment expressed in a post, such as reviews, tweets, or comments. There are three main approaches to sentiment analysis: rule-based, machine learning, and deep learning, each using different methods.

Product reviews on Amazon are analysed to determine the emotional tone of customer feedback. The analysis helps identify whether feedback is positive, negative, or neutral, providing valuable information to both sellers and consumers.

Sentiment classification has been tested in various domains, including cinema evaluation, tourism destination evaluation, and product evaluation (Liu et al., 2007; Pang et al., 2009; Ye et al., 2009). [6] Two primary approaches used for sentiment classification are vocabulary-based methods and machine learning methods.

I. Sentiment classification using machine learning methods

A multitude of articles have been published pertaining to machine learning, [7,8,9] a field that utilizes various algorithms for sentiment classification. This section seeks to explore several key aspects of machine learning. The overarching aim of machine learning is to create an algorithm capable of optimizing system performance utilizing sample data. The solution for machine learning involves sentiment analysis and comprises two main stages. The initial stage involves learning the model from the training data, while the second stage involves the classification of unseen data utilizing the trained model (Khairnar &Kinikar, 2013).[10] Machine learning algorithms can be classified into three distinct types: supervised learning, semi-supervised learning, and unsupervised learning.

ii. sentiment classification using lexicon basic method

There is another unsupervised method called the lexicon-based method, which uses word and sentence annotations. [11,12] This method calculates the sentiment score for each text by using a dictionary of words and expressions that convey emotions. The vocabulary-based method is the simplest approach to identify the sentiment of an assessment document, where a quantity-based approach is used. By using a vocabulary resource with positive and negative annotations of words and sentences, we can determine the polarity of the critique. If the number of positive words is greater than the number of negative words, the sentiment of the results is positive. Conversely, if there are more negative feeling words than positive, the overall sentiment of the text is negative [13].

A. feature extraction

Our project delved into two distinct types of features. The first involved a traditional approach where we constructed a dictionary based on common words and indexed each word. To achieve this, we established a threshold for the word dictionary at six occurrences and proceeded to collect 4,223 words from our entire database. Each rating was then converted into a vector, where each value represented the number of times a word appeared. Although we attempted to modify the threshold and dictionary length, we found that an increase in dictionary length had little impact on accuracy.

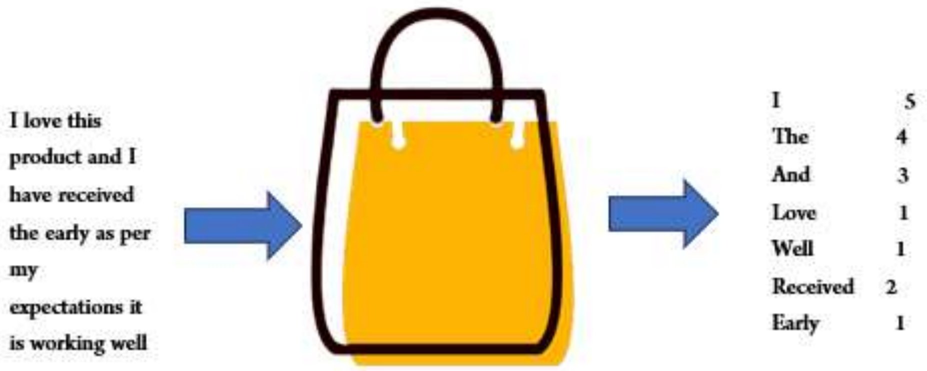

Figure 4 An illustration of the bag of words model

The second type of feature involved a 50-d Glove2 dictionary pre-trained on Wikipedia. Our goal was to capture the essence of every word in this section. We represented each review's average vector of the 50-D glove vector of all individual words comprising the reviews. Since machine learning algorithms only function with numeric vectors of fixed length rather than plain text, we parsed the input data, which in this case involved text data. This text-to-feature method is commonly referred to as the Bag of Words model, which offers a widely used method for feature extraction. This approach operates by producing different bags of words that emerge in the training dataset, where each word is associated with a unique number. This number indicates the occurrence of each word in the document. Figure 4 provides a simple example of the Bag of Words model, so-called because it disregards the position of words in the rejected document.

B. Monkey learn API

Monkey Learn is a robust text analytics platform that offers both individuals and businesses a comprehensive suite of tools and APIs for processing natural language. By leveraging its impressive capabilities in generating valuable insights, automating text-related tasks, and enabling informed decision-making based on text data, the platform has become a leading solution for NLP tasks. Users can take advantage of MonkeyLearn's pre-defined models or create custom models to analyse text data for sentiment, classification, and other NLP tasks. The platform's APIs and tools are widely used for extracting insights from text data and automating text analysis. Our team successfully utilized the MonkeyLearn API for Python programming in order to perform sentiment analysis. Through training our models on over 150 comments that were either positive, negative, or neutral, we were able to assign ratings of 4 and 5 for positive comments, 1 and 2 for negative comments, and 3 for neutral comments.

C. Positive sentiments

When customers write Amazon product reviews expressing satisfaction, enjoyment, or approval, it is considered positive sentiment. Recognizing positive sentiment in these reviews is crucial for businesses as it aids in comprehending what customers appreciate and how it can be used to enhance marketing and product development. Figure 5 demonstrates that our trained model can accurately detect positive sentiment when analysing review text from our dataset

Figure 5 Positive sentiment

D. Negative sentiments

In the realm of customer reviews on Amazon, expressions of dissatisfaction, disappointment, or criticism are classified as negative sentiments. For enterprises, it is imperative to recognize these adverse emotions as they facilitate the identification of areas for improvement and the addressing of customer concerns. Figure 6 is an accurate and confident representation of this negative sentiment when analysing a product. To enhance our confidence in the identification of negative sentiments, we can train our model with a more comprehensive range of negative words

Figure 6 Negative sentiment

E. Neutral sentiments

When analysing product reviews on Amazon, neutral emotions refer to reviews that do not strongly express positive or negative feelings. These reviews tend to describe the product or experience without indicating clear satisfaction or dissatisfaction. Figure 6 demonstrates neutral sentiment in terms of accuracy and confidence. To increase the accuracy and dependability, we can train our model with more sentences that are positive, negative, or neutral.

V. LITERATURE SURVEY AND COMPARATIVE ANALYSIS

1. In 2002, Pang, Lee, and Vaithyanathan conducted a study aimed at categorizing movies into positive and negative categories using supervised learning techniques, specifically Support Vector Machines (SVM), Naïve Bayes, and maximum entropy classification. The researchers experimented with various features and found that the best results were achieved by machine learning algorithms that utilized a

"bag of words" as features in these classifiers. All three techniques resulted in highly accurate outcomes.

2. In the publication titled "Like or Dislike? Semantic Disambiguation Applied to Unsupervised Review Classification," authored by Turney, P. D. in 2002, the author explores the application of semantic disambiguation in the classification of unsupervised reviews. The paper was presented at the 40th Annual Conference of the Association for Computational Linguistics and can be found on pages 417-424.
3. In their 2016 publication titled "Modelling Fashion Trend Evolution Using Single-layer Collaborative Filtering," Lui and McAuley explored the evolution of fashion trends. Their study was presented at the 25th International Conference on World Wide Web, and it involved the use of a single-layer collaborative filtering approach. The authors' findings shed light on the dynamics of fashion trends and the role of user preferences in shaping them.
4. A study conducted by Chavalit and Chu (2005) compared supervised machine learning algorithms to an unsupervised approach known as semantic orientation in the context of film criticism. The results indicated that the latter approach was more effective and reliable. This finding has significant implications for the field of machine learning and underscores the importance of ongoing research in this area
5. Wang, X., Hu, and Bing, L. (2016) presented a paper titled "Remote Monitoring for Emotion Classification with Binary Values" during the 54th Annual Conference of the Association for Computational Linguistics. The authors explored the use of remote monitoring to classify emotions using binary values. Their research aimed to provide a practical solution for emotion analysis by leveraging the benefits of remote monitoring. The paper delves into the methodology used, outlining the approach taken to classify emotions, and the results obtained. The research conducted in this paper has the potential to contribute to the field of emotion analysis and provide insights into how remote monitoring can be used to classify emotions in a practical context.
6. Jindal and Liu (2008) conducted an analysis on spam and comments in their conference paper. Their research provides valuable insights for both businesses and academics, shedding light on the prevalence and impact of these phenomena. Jindal and Liu's study emphasizes the importance of understanding and addressing problematic online behaviours. Such behaviours can have significant ramifications for online communities and communication. Consequently, their work highlights the need for continued research and attention in this area. Effective strategies must be developed for mitigating the negative effects of spam and other forms of online disruption.

VI. METHODOLOGIES

A.SVM

An effective supervised learning method for solving sentiment classification problems is the Support Vector Machine (SVM) technique. [14] This method utilizes a decision plan where labelled training data is placed and an optimal hyperplane is outputted to divide the data into different groups or classes. The best hyperplane is the one that separates the classes with the greatest amplitude, which is achieved by selecting a hyperplane that maximizes its distance from the data on each layer. SVM is a popular machine learning algorithm used in sentiment analysis, particularly in Amazon product review analysis. It is especially effective in text classification tasks such as sentiment analysis [15].

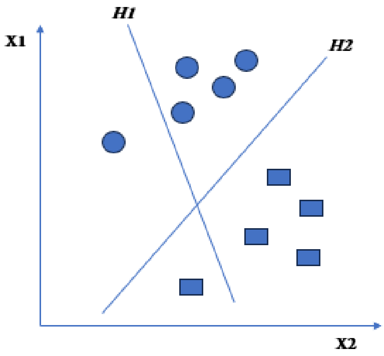

Figure 7 H1 does not separate the classes.

H3 separates them with the max margin

Linear support vector machine

The linear support vector machine (linear SVM) algorithm is a commonly employed method for analysing product reviews on Amazon for sentiment analysis. It utilizes a linear kernel for classification and is highly favoured for its simplicity, efficiency, and effectiveness, especially when confronted with high-dimensional text data. Essentially, the linear SVM creates a classifier by separating labelled data sets. Geometrically speaking, it aims to maximize the minimum distance between two types of points - circle and x - in each space, which ultimately results in maximizing profits. The optimization problem that the SVM aims to solve is presented below.

$$\arg\max_{\gamma,w,b} \frac{1}{2}||w||^2$$
$$s.t. y^i(w^T x + b) \geq 1, i = 1, 2, ...m$$

GLM

The Generalized Linear Model (GLM) is a linear regression model that is specifically designed to fit generalized linear models to data by maximizing the log-likelihood. [16,17,18] The GLM is a highly flexible and reliable method for model classification, making it an easily understandable tool for a wide range of business and academic applications. By optimizing the log-likelihood function, the GLM can provide accurate predictions and insights into a variety of complex datasets, making it an invaluable tool for statistical analysis and modelling.

B.Naive bayes

One of the widely known techniques in machine learning is Naive Bayes [19,20], which is surprisingly powerful despite its simplicity. This classifier is based on Bayes theorem and assumes that features, usually words in text classification, are independent of each other. Although this hypothesis is incorrect in some cases, Naive Bayesian classifiers have proven to be surprisingly effective, as shown by Rish in 2001. Naive Bayes is one of the most popular algorithms for generative tools and classification problems. To improve its performance, we integrate Laplace smoothing into our model. The formula below is used to predict an example:

$$p(x_1, ..., x_k|y) = \prod_{i=1}^{k} p(x_i|y)$$

$$\hat{y}^{(i)} = \arg\max_{j} \prod_{i=1}^{k} p(x_i|y=j)\phi(j)$$

To present revised texts, you can use an array of non-negative integers and a model p(xi|y) with a multinomial distribution. Alternatively, using a glove dictionary to present the review text does not require non-negative integers, so we chose to use the model p(xi|o) with a Gaussian distribution [21,22,23].

C. K-Nearest neighbour

K-NN is a learning algorithm based on a non-parametric version that can classify data points based on how close they are to labelled samples in a feature space. he K-nearest Neighbour (KNN) method is a type of non-parametric classification that has seen widespread use in recent times [24,25]. To make a prediction using this method, it searches for the K = n closest neighbours to the input and assigns the majority class of those neighbours. The distance between each neighbour is calculated using Euclidean distances, which measure the similarity between each data point**.**

$$\hat{f}(x) = \frac{1}{K} \sum_{x \in N_K(x)} y_i$$

E. Long short-term memory

Long-Short Term Memory, commonly known as LSTM, is a type of recurrent memory unit that is part of a neural network called RNN. [26,27] An LSTM unit consists of a cell, an entry gate, an exit gate, and a forgetfulness gate. The cell's primary function is to store values for an indefinite amount of time, while the three gates regulate the flow of incoming and outgoing information. LSTM networks are highly effective in processing, classifying, and predicting time-based data, especially when there are unknown time lags between significant events in the time series.

VII. CORRELATION

When analysing data, correlation refers to the degree of association between two variables. There are three possible outcomes of a correlational study: positive correlation, negative correlation, and no correlation. Utilizing correlation analysis can be a useful tool when examining the sentiment of product reviews on Amazon. This can help determine the relationship between various variables, such as sentiments and aspects of the review

A. Correlation between number of helpful reviews and price

In Figure 7, a trend line is used to show the correlation between the number of helpful reviews and price. The trend line indicates that there is a negative correlation between these two variables. The colours in the graph provide additional details about the correlation. However, this correlation is inconclusive for this sample set. While the overall rating is increasing, the number of useful ratings is decreasing. The graph that shows the correlation between ratings and usefulness numbers has mixed results. At rating 5, there is an increase in helpfulness, but at 2, there is a decrease from 1. Rating 1 shows a larger increase in helpfulness than ratings 2 and 3. Based on this information, it is likely that there is a negative correlation in this scatter plot.

B. Hypothesis testing using P-value

In this section, we will examine the data from the perspective of the null hypothesis and see what insights we can gather. We will focus on the rating and hope count attributes in the Amazon dataset for our sentiment analysis. Figure 8 displays the relationship between these attributes. However, a scatter plot does not offer much information except that the highest scores are increasing. Interestingly, even a rating below 1 can provide a hopeful perspective. To further understand the distribution of ratings, we can refer to Figure 8 which reveals a bias towards higher ratings.

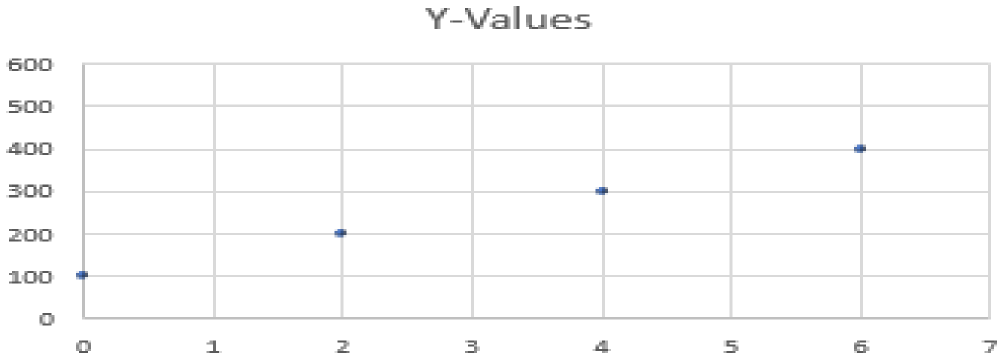

Figure 8 Rating v/s Hopeful number

VIII. RESULT

This section focuses on the research results obtained from the models employed. The accuracy value, which indicates the percentage of correctly classified test datasets, was used to evaluate the performance of the algorithms used. In both cases, the algorithms achieved accuracy rates above 90%, with the Support Vector Machine (SVM) performing better than the Naive Bayes method. However, when the summary was applied, the Naive Bayes method produced better results.

To implement the revised text with input features of 4223-d, several models were used, including multinomial Naive Bayes, SVM with a linearity kernel, SVM with an RBF kernel, KNN-4, 5, 6, and LSTM. KNN-5 outperformed the remaining two models, with SVM with a linear kernel slightly better than SVM with an RBF kernel. It is worth noting that SVMs with linear kernels exhibited an overfitting problem, which is indicated by the significant gap between training accuracy and testing accuracy.

Finally, LSTM was found to be the most efficient among all models in terms of test accuracy. This information is critical for businesses or academic researchers interested in the development of models for classification tasks.

IX. DISCUSSION

During training, KNN has a higher computational complexity compared to Naive Bayes and SVM. This is because the KNN algorithm requires the calculation of distances between all training and evaluation data points, leading to longer processing time. However, increasing the length of the dictionary for KNN does not significantly affect accuracy. If the dictionary threshold is reduced, the dictionary length will increase. LSTM produces slightly better results than other conventional machine methods due to the high number of parameters. The experiment shows that well-trained learning machines with adequate training data can perform classification well. While SVM tends to perform better than Naive Bayes in terms of accuracy, the difference is not significant, and both algorithms can achieve over 90% accurate classification.

X. CONCLUSION

The findings of our research indicate that most Amazon products in our dataset have been rated with four or five stars. As a result, the information presented in our chart may not provide sufficient insight into products rated with one, two, or three stars. Our investigation also revealed that low-priced products tend to receive more high ratings than high-priced products, suggesting that consumers are more likely to review cheaper products.

Furthermore, we observed a low total number of product reviews on Amazon in 2014, with a significant increase from 2016 to 2017. Our statistical analysis included a computation of the correlation between product price and the number of helpful reviews, which demonstrated a negative correlation between the two variables. This means that as prices rise, the number of helpful reviews tends to decrease, and vice versa. Additionally, we conducted chi-square tests on products in various categories, such as office supplies and hardware. For office supplies, we selected a sample of 250 values due to the limited number of products in this category, and our observations revealed a significant difference between the expected and observed values with a small p-value, which invalidated our null hypothesis. Similarly, for hardware, we used a sample of 1,000 products and found a significant discrepancy between the expected and observed values, which also invalidated our null hypothesis. It is worth noting that our data is biased towards high scores of four and five, which presented challenges in formulating a sound hypothesis for our chi-square tests. For future research, we plan to expand our work by incorporating more products, categories, and comparing Amazon products with other electronic products from different companies. Additionally, we intend to leverage advanced visualization techniques to represent our dataset more effectively.

XI.REFERENCES

[1]. T. Pinch and F. Kesler, How Aunt Amy Gets Her Free Lunch: A Study of the Top-Thousand Customer Reviewers at amazon.com, 2011.

[2]. McAuley. (2016). Amazon productdata. [Online]. Available: http://jmcauley.ucsd.edu/data/amazon/

[3]. [3]Bhatt, A. Patel, H. Chheda, and K. Gawande, "Amazon review classification and sentiment analysis," International Journal of Computer Science and Information Technologies, vol. 6, no. 6, pp. 5107–5110, 2015

[4]. X. Fang and J. Zhan, "Sentiment analysis using product review data,"Journal of Big Data, vol. 2, pp. 1–14, 2015

[5]. Subhabrata Mukherjee. Sentiment analysis of reviews

[6]. Priyank Pandey, Manoj Kumar, and Prakhar Srivastava. Classificationtechniques for big data: A survey. In Computing for Sustainable GlobalDevelopment (INDIACom), 2016 3rd International Conference on, pages3625–3629. IEEE, 2016.

[7]. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In Proceedings of the

[8]. ACL-02 conference on Empirical methods in natural language processingVolume 10, pages 79–86. Association for Computational Linguistics, 2002

[9]. K. Dhola and M. Saradva, "A comparative evaluation of traditional machine learning and deep learning classification techniques for sentiment analysis", Proc. 11th Int. Conf. Cloud Computing. Data Sci. Eng., pp. 932-936, Jan. 2021.

[10]. D. Guan, "Sentiment classification of financial online reviews based on machine learning algorithm", Proc. Int. Conf. Educ. Netw. Inf. Technol. (ICENIT), pp. 289-293, Sep. 2022.

[11]. Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. Computationallinguistics, 37(2):267–307, 2011.

[12]. R. D. Tan, K. Piad, A. Lagman, J. Victoriano, I. Tano, N. S. Gabriel, et al., "LMS content evaluation system with sentiment analysis using lexicon-based approach", Proc. 10th Int. Conf. Inf. Educ. Technol. (ICIET), pp. 93-98, Apr. 2022.

[13]. W. Suwanpipob, N. Arch-int and M. Wattana, "A sentiment classification from review corpus using linked open data and sentiment lexicon", Proc. 13th Int. Conf. Inf. Technol. Electrical Eng. (ICITEE), pp. 19-23, Oct. 2021.

[14]. C. Cortes and V. Vapnik, "Support-vector networks", Mach. Learn., vol. 20, pp. 273-297, Apr. 1995.

[15]. C. C. Aggarwal and C. Zhai, Mining Text Data, Cham, Switzerland: Springer, 2012

[16]. I. Trofimov and A. Genkin, "Distributed coordinate descent for generalized linear models with regularization", Pattern Recognition. Image Anal., vol. 27, no. 2, pp. 349-364, Apr. 2017

[17]. P. McCullagh, "Generalized linear models", Eur. J. Oper. Res., vol. 16, no. 3, pp. 370-384, 1984

[18]. M. Goldburd, A. Khare and C. D. Tevet, "Generalized linear models for insurance rating", Proc. Casualty Actuarial Soc., pp. 1-122, 2016.

[19]. M. Abbas, K. Ali, S. Memon, A. Jamali, S. Memonullah and A. Ahmed, "Multinomial Naïve Bayes classification model for sentiment analysis", Int. J. Computer. Sci. NetworkSecure., vol. 19, no. 3, pp. 62, Mar. 2019.

[20]. A. A. Farisi, Y. Sibaroni and S. A. Faraby, "Sentiment analysis on hotel reviews using multinomial Naïve Bayes classifier", J. Phys. Conf. Ser., vol. 1192, Mar. 2019

[21]. M.-R. Amini and N. Usunier, "Semi-supervised learning" in Learning With Partially Labeled and Interdependent Data, Cham, Switzerland:Springer, pp. 3-18, 2015.

[22]. Sebastian Raschka. Naive bayes and text classification i-introduction andtheory. arXiv preprint arXiv:1410.5329, 2014.

[23]. Irina Rish. An empirical study of the naive bayes classifier. In IJCAI 2001workshop on empirical methods in artificial intelligence, volume 3, pages41–46. IBM, 2001.

[24]. S. Hota and S. Pathak. Knnclassifier-based approach formulti-class sentiment analysis of twitter data. In InternationalJournal of Engineering Technology, pages 1372–1375. SPC,2018

[25]. L. Dey, S. Chakraborty, A. Biswas, B. Bose and S. Tiwari, "Sentiment analysis of review datasets using Naïve Bayes- and K-NN classifier", Int. J. Inf. Eng. Electron. Bus., vol. 8, no. 4, pp. 54-62, Jul. 2016.

[26]. F. Li, M. Zhang, G. Fu, T. Qian and D. Ji, "A Bi-LSTM-RNN model for relation classification using low-cost sequence features", arXiv:1608.07720, 2016.

[27]. Q. Huang, R. Chen, X. Zheng and Z. Dong, "Deep sentiment representation based on CNN and LSTM", Proc. Int. Conf. Green Informat. (ICGI), pp. 30-33, Aug. 2017.