Aspect-Based Sentiment Analysis of Amazon Product Reviews Using Machine Learning Models and Hybrid Feature Engineering

1st Md Rezwane Sadik
Department of Decision Sciences
University of South Dakota
Vermillion, South Dakota, USA
rezwane@iut-dhaka.edu

2nd Umma Hafsah Himu
Department of Statistics
University of Rajshahi
Rajshahi, Bangladesh

3rd Ifrat Ikhtear Uddin
Department of Computer Science
University of South Dakota
Vermillion, SD, USA
ifratikhtear.uddin@coyotes.usd.edu

4th Md Abubakkar
Department of Computer Science
Midwestern State University
Dallas, TX, USA
mabubakkar@ieee.org

5th Fazle Karim
Department of Computer science
Daffodil International University
Dhaka, Bangladesh
fazle4492@diu.edu.bd

6th Yousuf Abdullah Borna
Department of Computer science
California State University,Fullerton
Fullerton, California, USA
yousuf26@csu.fullerton.edu

***Abstract*—While sentiment analysis is a popular and significant research trend, aspect-based sentiment analysis (ABSA) requires more focus from researchers. The customer reviews of headphones and Bluetooth devices on Amazon are the main subject of this study. Several machine learning (ML) algorithms are used in the study, including Support Vector Machine (SVM), k-nearest Neighbors (KNN), Random Forest (RF), Naive Bayes (NB), Decision Tree (DT), Logistic Regression (LR). Additionally, a hybrid feature engineering technique combining TF-IDF (Term Frequency-Inverse Document Frequency) and word n-gram is applied, specifically utilizing word n-gram (1,4) in conjunction with TF-IDF. The results of evaluating these methods showed that, with an accuracy of 91%, SVM with hybrid word n-gram (1,3) produced the best outcomes. The research dataset exhibits imbalance, which is addressed by using the Matthews Correlation Coefficient (MCC) as an additional performance metric. This results in a score of 0.77. The results show that aspect-based sentiment analysis is effective in gaining insightful information from customer reviews of headphones and Bluetooth devices on Amazon. The SVM algorithm and the designated hybrid feature engineering technique perform better than the other.**

***Index Terms*—Sentiment analysis, Aspect base, Emotion, Hybrid feature engineering, Machine learning, Product reviews**

I. Introduction

Consumer behavior within the dynamic and intricate realm of e-commerce is shaped by various elements, with customer attitudes playing a pivotal role. An indispensable technique in understanding and categorizing these attitudes is sentiment analysis (SA), a method employed to assess consumer feedback automatically [1]. Particularly, consumer decisions are heavily influenced by online reviews, notably those on platforms like Amazon. These evaluations often serve as proxies for product quality and significantly impact consumer behavior. Baek et al. point out that factors such as review rating, reviewer reputation, and review substance contribute to the usefulness of these reviews [2]. Moreover, the reliability and utility of the platform itself significantly influence consumer trust in and reliance on these reviews, as discussed by Holleschovsky and Constantinides [3].

Online reviews, representing the collective opinions of consumers, play a crucial role in driving product sales [4]. Gaurav and Kumar propose a Consumer Contentment Rating System utilizing SA to gauge consumer satisfaction with specific product aspects [5]. Similarly, Sharaff et al. offer a method to identify product attributes and categorize reviews based on these attributes, aiding consumers in making informed purchase decisions [6]. These studies collectively underscore the dynamic nature of customer sentiments, extending beyond product attributes or overall ratings. Recognizing consumer sentiments, both general and specific to product qualities, is imperative for devising effective commercial strategies and enhancing customer satisfaction. Another work focused on the significance of understanding and leveraging consumer emotions in product design to enhance customer happiness and influence purchase decisions [7]. Within the context of e-commerce, comprehending customer sentiment necessitates delving into aspect-based sentiment analysis (ABSA). ABSA surpasses traditional SA by identifying and evaluating sentiments linked to specific product attributes or features. While traditional SA may only provide an overall sentiment score for a product review, ABSA allows for a more nuanced understanding.

In structuring this paper, we have segmented it into several sections to provide a comprehensive exploration of our research. In Section II, we review diverse approaches to SA. Subsequently, Section III outlines our methodology, including the crucial preprocessing steps and hybrid feature engineering techniques employed in our research. Section IV delves into the experimental outcomes of the machine learning (ML) models we utilized, shedding light on the performance metrics considered. Finally, we conclude by presenting our findings and highlighting the novelty of our approach in addressing another aspect within the realm of ABSA.

II. Related Work

Lee (2018) found that high star ratings and detailed descriptions significantly impact potential customers' perceptions, while Mudambi (2010) highlighted the importance of review extremity, review depth, and product type. Willemsen (2011) further emphasized the role of argumentation, review valence, and expertise claims in determining the perceived usefulness of reviews.

These findings collectively suggest that a combination of factors, including the nature of the product, the content of the review, and the reviewer's expertise, contribute to the perceived usefulness of online reviews on Amazon.

In recent years, there has been a surge in research articles focused on opinion mining, sentiment analysis, and the evaluation of product reviews. This burgeoning field places a significant emphasis on leveraging advanced technologies such as Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning to extract nuanced emotions from user feedback. The intricate use of computational methods plays a pivotal role in recognizing, comprehending, and estimating sentiments and viewpoints expressed in evaluations. The integration of ML, NLP, and deep learning techniques stands out as a crucial aspect, enhancing the precision and effectiveness of sentiment analysis. This, in turn, yields valuable insights into consumer sentiments.

The evolution of Amazon review analysis has been a subject of multifaceted exploration. Lee et al. [8] investigated the impact of review format and content on the perceived usefulness of a product, highlighting the significance of thorough descriptions and high star ratings. Rashid and Huang's work [9] emphasized the influential role of user reviews in shaping online consumer decisions. Their exploration involved applying visualization tools to the Amazon User Review Dataset, shedding light on the dynamics of the Amazon Review system. Alqahtani [10] delved into sentiment analysis, employing ML techniques to classify reviews and determine optimal models. Collectively, these studies underscore the dynamic nature of Amazon reviews, necessitating continuous evaluation and adaptation.

Various methodological approaches have been employed for sentiment analysis of Amazon reviews. Rathor et al. [11] compared the effectiveness of ML methods and found that Support Vector Machines (SVM) yielded superior results with weighted unigrams. Sinnasamy and Sjaif [12] achieved high accuracy using N-gram and a team-based approach, particularly with TF-IDF and SVM. Ejaz et al. [13] argued in favor of lexicon-based methods over ML approaches. Kumar et al. [14] adopted an integrated strategy, combining sentiment polarity with ratings, and identified the efficacy of NB, Linear SVM, and Logistic Regression classifiers. D'souza [15] pinpointed the limitations of the Bag-of-Words method, while Elmurngi [16] concentrated on detecting biased reviews. Collectively, these studies underscore the diversity and potential of methodological techniques for refining sentiment analysis of Amazon reviews.

While advancements have been made, there remains a consensus among researchers such as D'souza and Sonawane [15], and Elmurngi et al. [16] regarding the need for more precise sentiment analysis. D'souza specifically critiques the limitations of the Bag-of-Words method, while Elmurngi focuses on the identification of biased reviews.

III. Methodology

The methodology is illustrated in the accompanying Figure 1, which presents a flow chart depicting the sequential steps from data preprocessing to model evaluation in a structured and systematic manner.

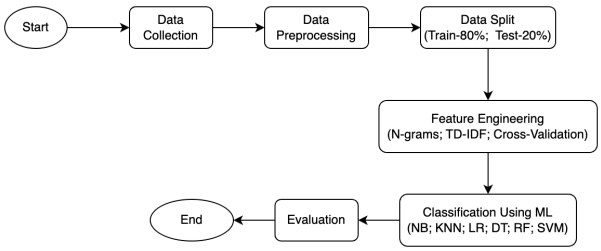

Fig. 1. Methodology of our research

A. Dataset Description and Preprocessing

The dataset used in this study comprises 14,337 Amazon reviews, each accompanied by star ratings, sourced from the Kaggle repository[1]. The dataset is specifically designed for Aspect-Based Sentiment Analysis (ABSA) and focuses on the ten most recent Bluetooth earphone devices as of mid-2019. It includes key information such as ReviewTitle, ReviewBody, ReviewStar (customer-assigned star ratings), and Product (product name). Before proceeding to subsequent steps, the dataset underwent preprocessing.

The dataset was structured based on our research goal, where we categorized it into positive and negative classes. Raw data from the CSV file was loaded into a pandas DataFrame to create a refined dataset [17]. To ensure data integrity, rows with missing values in critical columns ('ReviewStar,' 'ReviewBody,' and 'Product') were eliminated. Neutral reviews (star value of 3) were excluded to improve the model's ability to distinguish between positive and negative sentiments. Subsequent steps involved text cleaning, including tasks such as expanding contractions, removing HTML tags, URLs, non-alphanumeric characters, emoticons, and extra white spaces. Language filtering was applied to retain only English reviews, utilizing the 'langdetect' library. Additional data cleaning involved discarding rows with empty and NaN values in relevant columns. A new column labeled 'sentiment' was generated based on star ratings, categorizing reviews as 'positive' for ratings 4 and 5 and 'negative' for all other ratings. Brand names and model numbers were removed to eliminate potential biases. Misspelled words were addressed, and word replacement functions were incorporated to enhance sentiment analysis accuracy. Words starting with 'awe,' 'fan,' 'excel,' 'perfect,' and 'wonder' were substituted with their corresponding positive counterparts. Stop words were removed, and Lemmatization techniques [18] were applied. It is noteworthy that our custom stopwords set excludes specific terms such as 'not' and 'ok' to align with the specific requirements of our preprocessing approach. Simultaneously,

[1]source: https://www.kaggle.com/datasets/shitalkat/amazonearphonesreviews/data

lemmatization was applied to the processed text to enhance the coherence of the feature set for subsequent analytical procedures.

Data preprocessing and sentiment labeling 0 -1

```
Load dataset into a pandas DataFrame
Handle null values
```

row in DataFrame `Clean text in row['text_column']` `Remove HTML tags` `Remove special characters, punctuation, and numbers` `Remove extra white spaces` `Remove stopwords`

row in DataFrame `Label sentiment based on row['rating_column']` `row['rating_column']` is in [4, 5] `Set sentiment label to 'positive'` `Set sentiment label to 'negative'`

B. Feature Engineering

We utilized TF-IDF vectorizer [19] and word N-gram [20] in our research. TF-IDF is defined as the term frequency multiplied by the inverse document frequency. A prime example of an appropriate input representation is the TF-IDF vectorizer, which assesses a word's presence in a document instead of relying solely on raw numbers. The frequency of a particular word appearing in a given document is referred to as its frequency. Conversely, Inverse Document Frequency takes into account all papers containing that term. Within the field of NLP, word n-grams are collections of n words extracted consecutively from a given text or speech corpus. By encapsulating local linguistic nuances, these n-grams serve as micro-contextual windows, unraveling intricate inter-word relationships within a document. This detailed study produces a wealth of knowledge helpful in solving various NLP problems. The use of n-grams in linguistic exploration greatly aids in decoding complex syntactic and semantic complexities among different languages. It allows for a more detailed investigation of the complex fabric underlying word relationships. This enhances our understanding of the intricacies involved in linguistic communication by revealing patterns and dependencies unique to a given language. In our research, we applied different n-gram techniques, dividing the sentence based on the value of N, where we used N=1,2,3,4.

C. Classification

In our classification system, we employed Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes (NB), k-Nearest Neighbors (KNN), Decision Tree (DT), and Random Forest (RF) [21]. All classifiers are explained below, and the parameters used in the classifiers are shown in Table I. For sentiment analysis on unbalanced datasets, selecting the right method is essential for producing relevant and reliable findings. LR is the classifier created for binary classification. The One-vs-Rest (OvA) technique allows logistic regression to be modified for multiclass problems [22]. Using this method, for a dataset having C-level courses, one class is designated as the positive class, and the remaining classes are negatives in each of the C-distinct logistic regression models trained. The probability of each class having an instance is computed, and the class with the largest probability is predicted [23]. It functions well when there is a linear relationship between the characteristics and the target variable, which is why we utilized it. On the other hand, NB is predicated on the independence of characteristics and the Bayes theorem. It is easy to use, quick, and effective with high-dimensional data. The method used to calculate the likelihood that a document will belong to a specific class. When using the OvA strategy, C binary SVM classifiers are trained for a C-class problem, treating one class as the positive class and the remaining classes as the negative class. As we saw in class, there are distinct margins so the SVM classifier was employed. It functions best when there is a clear boundary between the classes. When it comes to binary and multiclass classification tasks, DT is a potent and easily interpreted ML algorithm. Recursively dividing the feature space according to the values of the input features is how the algorithm decides what to do. The choice to partition the data is made at every node in the tree, and this process is carried out until a stopping criterion—such as a predetermined tree depth or a minimum number of samples per leaf node—is satisfied [24]. The decision rule entails traversing the tree from the root to a leaf node depending on the feature values of a particular instance. Each leaf node of the tree corresponds to a distinct class label. The majority class linked to the leaf node then determines the ultimate prediction. An effective ensemble learning algorithm that performs well in multiclass classification tasks is called RF. It works by building a collection of decision trees, each of which is trained using a bootstrapped dataset sample [25]. At each split, it adds unpredictability by taking into account a smaller subset of features. The diversity of the trees enhances the model's ability to generalize to new data and reduces overfitting. A majority vote among the trees determines the final classification. During prediction, each tree independently offers a class prediction. When it comes to capturing intricate relationships in the data, robustly handling outliers and missing values, and revealing the significance of individual features, Random Forest is incredibly effective.

TABLE I
CLASSIFIERS AND PARAMETERS

Classifier	Parameters
LR	Regularization (C), Penalty (L1, L2), Solver ('liblinear', 'lbfgs')
KNN	Number of neighbours (k), Distance metric (Euclidean, Manhattan), Weighting (Uniform, Distance-based)
DT	Criterion (Gini, Entropy), Maximum depth, Minimum samples per leaf
RF	Number of trees, Maximum depth, Minimum samples per leaf
SVM	Kernel (Linear, RBF), Regularization (C), Gamma (for RBF kernel)

IV. RESULT AND DISCUSSION

We strategically used a variety of machine learning (ML) models in our sentiment analysis process, along with careful pre-processing methods and feature engineering, to maximize the textual data for fine-grained sentiment extraction. We will explain our model outcomes in detail in this section. Firstly, we have calculated F-score, accuracy, precision, sensitivity, and specificity. The model's accuracy in making positive predictions and its capacity to capture positive instances fully is measured by precision. To provide a fair assessment,

specificity quantifies how well the model detects instances of negative sentiment. A comprehensive evaluation is guaranteed by the F-score, which is a harmonic mean of precision and sensitivity. Furthermore, accuracy is a basic indicator of overall correctness. Taking into account class imbalances, we use the Matthews Correlation Coefficient (MCC) as a critical metric to capture the model's performance and account for the imbalance in the dataset. The detailed performance of all models is described in Table II, where we also included different N-gram outcomes in our result.

TABLE II
PERFORMANCE OF ALL MODELS USING DIFFERENT N-GRAM TECHNIQUES

Model	N-gram	Accuracy	Precision	Sensitivity	Specificity	F1 Score	MCC
NB	1-1 grams	0.808042	0.834338	0.992792	0.307282	0.769948	0.469790
	1-2 grams	0.794639	0.832953	0.997379	0.245115	0.745768	0.427470
	1-3 grams	0.783629	0.826765	0.998034	0.202487	0.726465	0.386941
	1-4 grams	0.776927	0.821849	0.998034	0.177620	0.714498	0.359982
KNN	1-1 grams	0.855912	0.851672	0.927916	0.660746	0.852190	0.619963
	1-2 grams	0.871709	0.868603	0.933814	0.703375	0.869073	0.663972
	1-3 grams	0.869315	0.865835	0.938401	0.682060	0.865751	0.655367
	1-4 grams	0.872188	0.868969	0.945609	0.673179	0.867938	0.661545
LR	1-1 grams	0.879368	0.879984	0.945609	0.673179	0.872382	0.679137
	1-2 grams	0.877932	0.883458	0.945609	0.673179	0.868441	0.676796
	1-3 grams	0.869315	0.876984	0.945609	0.673179	0.857484	0.653023
	1-4 grams	0.865965	0.874084	0.945609	0.673179	0.853307	0.643469
SVM	1-1 grams	0.892772	0.890872	0.945609	0.673179	0.891284	0.721655
	1-2 grams	0.909047	0.907708	0.945609	0.673179	0.907961	0.764656
	1-3 grams	0.909526	0.908247	0.945609	0.673179	0.908531	0.766166
	1-4 grams	0.907611	0.906372	0.945609	0.673179	0.906711	0.761575
DT	1-1 grams	0.823839	0.822693	0.945609	0.673179	0.823235	0.549633
	1-2 grams	0.817616	0.822719	0.945609	0.673179	0.819756	0.548975
	1-3 grams	0.813787	0.826438	0.945609	0.673179	0.818190	0.555556
	1-4 grams	0.782671	0.799654	0.945609	0.673179	0.788608	0.486712
RF	1-1 grams	0.886070	0.887293	0.945609	0.673179	0.879679	0.698174
	1-2 grams	0.880804	0.881173	0.945609	0.673179	0.874171	0.683184
	1-3 grams	0.882719	0.884356	0.945609	0.673179	0.875658	0.688828
	1-4 grams	0.882240	0.882725	0.945609	0.673179	0.875740	0.687257

Different patterns were seen when classifiers from different N-gram approaches were evaluated for aspect-based sentiment identification [26]. With unigram features, Naive Bayes (NB) obtained 0.808 accuracy, 0.834 precision, 0.993 sensitivity, 0.307 specificity, 0.770 F1 Score, and 0.470 MCC. The accuracy fell to 0.795, precision to 0.833, sensitivity to 0.997, specificity to 0.245, F1 Score to 0.746, and MCC to 0.427 for NB with bigram characteristics. Regarding trigram characteristics, NB demonstrated 0.784 accuracy, 0.827 precision, 0.998 sensitivity, 0.202 specificity, 0.726 F1 Score, and 0.387 MCC. NB showed a sensitivity of 0.998, specificity of 0.178, F1 Score of 0.714, accuracy of 0.777, precision of 0.822, and MCC of 0.360 with quadgram features. With unigram features, K-Nearest Neighbors (KNN) obtained 0.856 accuracy, 0.852 precision, 0.928 sensitivity, 0.661 specificity, 0.852 F1 Score, and 0.620 MCC. With bigram characteristics, the results increased to 0.872 for accuracy, 0.869 for precision, 0.934 for sensitivity, 0.703 for specificity, 0.869 for F1 Score, and 0.664 for MCC. KNN preserved 0.869 accuracy, 0.866 precision, 0.938 sensitivity, 0.682 specificity, 0.866 F1 Score, and 0.655 MCC for trigram features. KNN maintained 0.872 accuracy, 0.869 precision, 0.946 sensitivity, 0.673 specificity, 0.868 F1 Score, and 0.662 MCC with quadgram features. With unigram features, Logistic Regression (LR) produced the following results: accuracy of 0.879, precision of 0.880, sensitivity of 0.946, specificity of 0.673, F1 Score of 0.872, and MCC of 0.679. With bigram characteristics, the LR accuracy was 0.878, the precision was 0.883, the sensitivity was 0.946, the specificity was 0.673, the F1 Score was 0.868, and the MCC was 0.677. LR demonstrated 0.869 accuracy, 0.877 precision, 0.946 sensitivity, 0.673 specificity, 0.857 F1 Score, and 0.653 MCC with trigram features. LR displayed the following quadgram features: MCC of 0.643, F1 Score of 0.853, specificity of 0.673, sensitivity of 0.946, accuracy of 0.866, and precision of 0.874. With unigram features, Support Vector Machine (SVM) obtained 0.893 accuracy, 0.891 precision, 0.946 sensitivity, 0.673 specificity, 0.891 F1 Score, and 0.722 MCC. With bigram characteristics, SVM achieved 0.909 accuracy, 0.908 precision, 0.946 sensitivity, 0.673 specificity, 0.908 F1 Score, and 0.765 MCC. SVM was able to retain 0.910 accuracy, 0.908 precision, 0.946 sensitivity, 0.673 specificity, 0.909 F1 Score, and 0.766 MCC with trigram features. SVM demonstrated 0.908 accuracy, 0.906 precision, 0.946 sensitivity, 0.673 specificity, 0.907 F1 Score, and 0.762 MCC using quadgram features. With unigram features, Decision Tree (DT) obtained 0.824 accuracy, 0.823 precision, 0.946 sensitivity, 0.673 specificity, 0.823 F1 Score, and 0.550 MCC. The DT using bigram characteristics showed a decline in accuracy to 0.818, precision to 0.823, sensitivity to 0.946, specificity to 0.673, F1 Score to 0.820, and MCC to 0.549. DT demonstrated 0.814 accuracy, 0.826 precision, 0.946 sensitivity, 0.673 specificity, 0.818 F1 Score, and 0.556 MCC with trigram features. DT showed 0.783 accuracy, 0.800 precision, 0.946 sensitivity, 0.673 specificity, 0.789 F1 Score, and 0.487 MCC with quadgram features. Utilizing unigram characteristics, Random Forest (RF) obtained 0.886 accuracy, 0.887 precision, 0.946 sensitivity, 0.673 specificity, 0.880 F1 Score, and 0.698 MCC. With bigram characteristics, the RF achieved 0.881 accuracy, 0.881 precision, 0.946 sensitivity, 0.673 specificity, 0.874 F1 Score, and 0.683 MCC. RF was able to retain 0.883 accuracy, 0.884 precision, 0.946 sensitivity, 0.673 specificity, 0.876 F1 Score, and 0.689 MCC with trigram features.RF showed 0.882 accuracy, 0.883 precision, 0.946 sensitivity, 0.673 specificity, 0.876 F1 Score, and 0.687 MCC with quadgram features.

TABLE III
THE BEST PERFORMER N-GRAM TECHNIQUE FOR A MODEL

Name	N-grams	Acc.	Sensitivity	Specificity	Prc.	F1	MCC
NB	1-1	0.81	0.99	0.31	0.83	0.77	0.47
KNN	1-2	0.87	0.93	0.70	0.87	0.87	0.66
LR	1-1	0.88	0.95	0.67	0.88	0.87	0.68
SVM	1-3	0.91	0.95	0.67	0.91	0.91	0.77
DT	1-1	0.82	0.95	0.67	0.82	0.82	0.55
RF	1-1	0.89	0.95	0.67	0.89	0.88	0.70

The result shows that different models performed well for different N-gram techniques. Which N-gram technique performed well for which model, that is summarized in Table III. The table showed that unigram performed best for NB, LR, DT, and RF, among other N-gram techniques. For KNN, bigram performed the best, and trigram performed best for the SVM classifier. In our research, among all models, SVM with trigram hybrid feature engineering techniques performed best. The NB classifier achieved 81% highest accuracy, which is 10% lower than SVM. The best accuracy performance of the KNN classifier is 87%, which is 4% lower than the SVM classifier [27]. LR and RF achieved 88% and 89% accuracy, respectively. Among all of the models, NB performed the worst. Furthermore, extensively evaluating the model for the unbalanced situation, we have calculated the MCC value. When the model produces no errors, its MCC value is 1,

which denotes flawless prediction. Perfect inverse prediction is indicated by an MCC of -1, which indicates the total errors made by the model. Random prediction is shown by an MCC of 0. We got the highest MCC value of 0.77 by SVM, which shows that there is a substantial positive connection between the actual and anticipated categories. This implies that the model performs well overall in binary classification tasks, with its predictions agreeing well with the genuine labels. An MCC of 0.77 is regarded as high and suggests that the model is functioning well, giving confidence to the precision of its predictions. The result demonstrated that SVM performed well in our research proved by several evaluation criteria where NB performed the worst. The confusion matrix of the best model is shown in Figure 2.

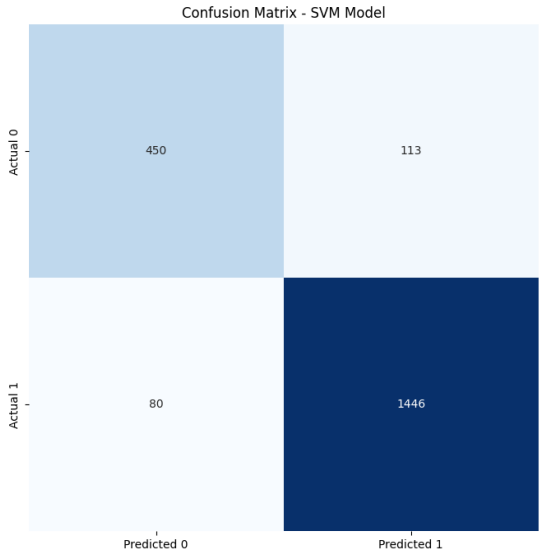

Fig. 2. Confusion matrix of SVM(1-3 Feature).

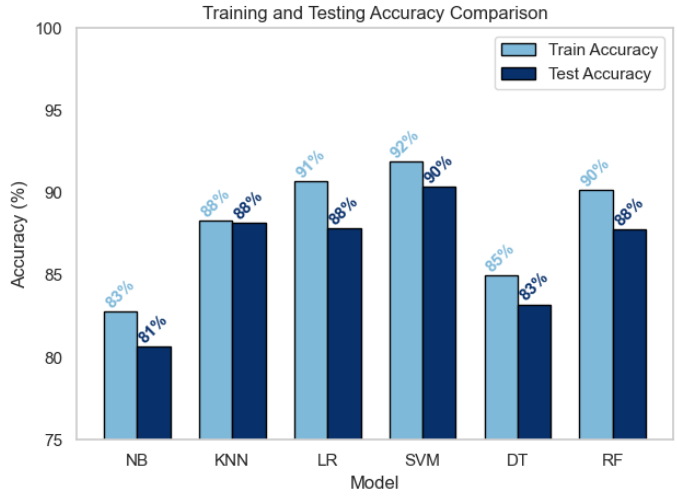

Fig. 3. Training and testing accuracy of all models.

We identified True Positives, False Positives, True Negatives, and False Negatives as the four crucial metrics within this matrix [28] [29]. True Positives, which total 450, are the number of times our model correctly predicted positive feelings when those sentiments were real. On the other hand, False Positives, which register at 113, represent instances in which the model incorrectly identified negative emotions as positive. True Negatives represent the 1446 cases in which our model identified negative sentiments correctly. Finally, False Negatives identifies the 80 instances in which the model mislabeled positive emotions as negative. Finally, the testing and training accuracy are presented in Figure 3.

V. Conclusion

The research explores sentiment analysis in Amazon product reviews, employing various machine learning (ML) models and N-gram techniques. The study underscores the importance of aspect-based sentiment analysis (ABSA) in understanding user feedback, particularly within Amazon reviews. Models such as Naive Bayes (NB), Logistic Regression (LR), Decision Trees (DT), and Random Forest (RF) demonstrated reliable performance with uni-gram features, while k-Nearest Neighbors (KNN) exhibited improved accuracy using 1-2 grams. Despite acknowledging limitations such as an unbalanced dataset and a limited number of data points, our evaluation, considering metrics like Matthews Correlation Coefficient (MCC), proved crucial for handling imbalances. Support Vector Machine (SVM), utilizing 1-3 grams, showed exceptional performance, achieving the highest MCC of 0.77.

In light of our findings, it is essential to note that different feature engineering approaches performed better in different scenarios, as outlined in our research [30]. Furthermore, our study highlighted the pervasive issue of an unbalanced dataset in product review analysis, underscoring the need for its careful consideration and mitigation.

As a step towards future work, we recommend applying deep learning techniques to further enhance ABSA in Amazon consumer reviews. Incorporating deep learning methodologies can potentially provide more nuanced insights into consumer sentiments, thereby refining the evaluation process [31]. This addition would contribute to advancing the field and establishing a more precise framework for ABSA in online product reviews.

References

[1] T. Islam, M. A. Sheakh, M. R. Sadik, M. S. Tahosin, M. M. R. Foysal, J. Ferdush, and M. Begum, "Lexicon and deep learning-based approaches in sentiment analysis on short texts," *Journal of Computer and Communications*, vol. 12, no. 1, pp. 11–34, 2024.

[2] H. Baek, J. Ahn, and Y. Choi, "Helpfulness of online consumer reviews: Readers' objectives and review cues," *International Journal of Electronic Commerce*, vol. 17, no. 2, pp. 99–126, 2012.

[3] E. Constantinides and N. I. Holleschovsky, "Impact of online product reviews on purchasing decisions," in *International Conference on Web Information Systems and Technologies*, vol. 2. SCITEPRESS, 2016, pp. 271–278.

[4] T. Islam, A. Kundu, R. J. Lima, M. H. Hena, O. Sharif, A. Rahman, and M. Z. Hasan, "Review analysis of ride-sharing applications using machine learning approaches," *Computational Statistical Methodologies and Modeling for Artificial Intelligence*, pp. 99–122, 2023.

[5] K. Gaurav and P. Kumar, "Consumer satisfaction rating system using sentiment analysis," in *Digital Nations–Smart Cities, Innovation, and Sustainability: 16th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2017, Delhi, India, November 21–23, 2017, Proceedings 16*. Springer, 2017, pp. 400–411.

[6] A. Sharaff and A. Soni, "Analyzing sentiments of product reviews based on features," in *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 2018, pp. 710–713.

[7] G. Vaidya and P. C. Kalita, "Understanding emotions and their role in the design of products: an integrative review," *Archives of Design Research*, vol. 34, no. 3, pp. 5–21, 2021.

[8] S.-G. Lee, S. Trimi, and C.-G. Yang, "Perceived usefulness factors of online reviews: a study of amazon. com," *Journal of computer information systems*, vol. 58, no. 4, pp. 344–352, 2018.

[9] A. Rashid and C.-y. Huang, "Sentiment analysis on consumer reviews of amazon products," *International Journal of Computer Theory and Engineering*, vol. 13, no. 2, p. 7, 2021.

[10] A. S. AlQahtani, "Product sentiment analysis for amazon reviews," *International Journal of Computer Science & Information Technology (IJCSIT) Vol*, vol. 13, 2021.

[11] A. S. Rathor, A. Agarwal, and P. Dimri, "Comparative study of machine learning approaches for amazon reviews," *Procedia computer science*, vol. 132, pp. 1552–1561, 2018.

[12] N. N. A. Sjaif *et al.*, "Sentiment analysis using term based method for customers' reviews in amazon product," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, 2022.

[13] A. Ejaz, Z. Turabee, M. Rahim, and S. Khoja, "Opinion mining approaches on amazon product reviews: A comparative study," in *2017 International Conference on Information and Communication Technologies (ICICT)*. IEEE, 2017, pp. 173–179.

[14] D. Kumar, R. Aggarwal, S. Raghuvanshi, and S. Chand, "An integrated approach for amazon product reviews classification using sentiment analysis," *International Journal of Computer Applications & Information Technology*, vol. 12, no. 2, pp. 317–324, 2020.

[15] S. R. D'souza and K. Sonawane, "Sentiment analysis based on multiple reviews by using machine learning approaches," in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2019, pp. 188–193.

[16] E. I. Elmurngi and A. Gherbi, "Unfair reviews detection on amazon reviews using sentiment analysis with supervised learning techniques." *J. Comput. Sci.*, vol. 14, no. 5, pp. 714–726, 2018.

[17] M. A. Hossain Raju, T. Imam, J. Islam, A. A. Rakin, M. N. Nayyem, and M. S. Uddin, "An ontological framework for lung carcinoma prognostication via sophisticated stacking and synthetic minority oversampling techniques," in *2024 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, 2024, pp. 125–130.

[18] T. Islam, A. Kundu, N. Islam Khan, C. Chandra Bonik, F. Akter, and M. Jihadul Islam, "Machine learning approaches to predict breast cancer: Bangladesh perspective," in *International Conference on Ubiquitous Computing and Intelligent Information Systems*. Singapore: Springer Nature Singapore, April 2021, pp. 291–305.

[19] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: Tf-idf approach," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*. IEEE, 2016, pp. 61–66.

[20] G. Gledec, R. Šoić, and Š. Dembitz, "Dynamic n-gram system based on an online croatian spellchecking service," *IEEE Access*, vol. 7, pp. 149 988–149 995, 2019.

[21] M. Hasan, J. Islam, M. Ahmed, and M. M. Hasan, "Prediction of colon cancer using densenet121, cnn, and resnet50 machine learning models and using image processing techniques," in *2023 International Conference on Artificial Intelligence Robotics, Signal and Image Processing (AIRoSIP)*, 2023, pp. 296–301.

[22] T. Islam, A. Vuyia, M. Hasan, and M. M. Rana, "Cardiovascular disease prediction using machine learning approaches," in *2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*. IEEE, April 2023, pp. 813–819.

[23] B. P. Ghosh, T. Imam, N. Anjum, M. T. Mia, C. U. Siddiqua, and M. A. I. Mamun, "Advancing chronic kidney disease prediction: Comparative analysis of machine learning algorithms and a hybrid model," 2024.

[24] M. S. Tahosin, M. A. Sheakh, T. Islam, R. J. Lima, and M. Begum, "Optimizing brain tumor classification through feature selection and hyperparameter tuning in machine learning models," *Informatics in Medicine Unlocked*, vol. 43, p. 101414, 2023.

[25] T. Islam, A. Kundu, T. Ahmed, and N. I. Khan, "Analysis of arrhythmia classification on ecg dataset," in *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*. IEEE, April 2022, pp. 1–6.

[26] K. S. Sharif, M. M. Uddin, and M. Abubakkar, "Neurosignal precision: A hierarchical approach for enhanced insights in parkinson's disease classification," in *2024 International Conference on Intelligent Cybernetics Technology Applications (ICICyTA)*, 2024, pp. 1244–1249.

[27] I. U. Ahamed, A.-A. Hossain, T. Imam, and J. Islam, "A multimodal analytical approach to alzheimer's disease diagnosis using machine learning and convolutional neural networks on mri datasets," in *2024 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, 2024, pp. 32–37.

[28] M. Hasan, J. Islam, M. A. Mamun, A. A. Mim, S. Sultana, and M. S. H. Sabuj, "Optimizing cervical cancer prediction, harnessing the power of machine learning for early diagnosis," in *2024 IEEE World AI IoT Congress (AIIoT)*, 2024, pp. 552–556.

[29] A. A. Rakin, M. N. Nayyem, K. S. Sharif, A.-A. Hossain, and R. Arafin, "A comprehensive framework for advanced machine learning and deep learning models in cervical cancer prediction," in *2024 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, 2024, pp. 1–6.

[30] M. A. Hossain, S. Bin Shawkat, K. S. Sharif, M. I. Hossain, H. Asmani, and M. M. Rahman, "Precisioncardio: A comprehensive machine learning approach for accurate prediction of heart failure trajectory," in *2024 IEEE 30th International Conference on Telecommunications (ICT)*, 2024, pp. 1–4.

[31] M. Sarkar, M. Reja, M. Arif, A. Uddin, K. Sharif, M. Tusher, S. Devi, M. Ahmed, M. Bhuiyan, M. Rahman *et al.*, "Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications," *International Journal on Computational Engineering*, vol. 1, no. 3, pp. 62–67, 2024.