International Journal of Computer Techniques–IJCT Volume 13 Issue 2, March-April - 2026

Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

Sentiment Analysis of E-Commerce Product-Based Reviews Using a Machine Learning Approach

A Case Study of Amazon.com

Duru Juliet Chinenye*

*(Department of ICT, Abia State University, Uturu, Nigeria

Email: duru.juliet@abiastateuniversity.edu.ng)

Ogbuagu Chinedu Samuel**

**(Department of Computer Engineering, Enugu State University of Science and Technology (ESUT), Nigeria

Email: ogbuaguchinedusamuel@gmail.com)

Chima Aguocha Obingonye***

***(Department of Computer Engineering, University of Derby, United Kingdom

Email: obingonyechima@gmail.com)

Praise Madumere Chukwubueze****

******(Department of ICT, Huawei Technologies, Abuja, Nigeria

Email: praizchuks@gmail.com)

ABSTRACT:

With the rapid growth of e-commerce and online shopping platforms, product reviews have become a crucial resource for consumers making informed purchasing decisions. These reviews contain valuable insights into customer sentiments and opinions about products and services, significantly influencing other customers' choices. However, manually analysing large volumes of reviews is time-consuming and challenging. Machine learning offers an automated solution by extracting and classifying sentiments expressed in reviews. The primary objective of this research was to develop a sentiment analysis model using supervised machine learning to classify sentiments as positive or negative. The Amazon Fine Food dataset from Kaggle was utilised, comprising 568,454 reviews, from which a random sample of 5,685 reviews was drawn for analysis. Data preprocessing involved removing special characters and HTML tags, handling missing values, and eliminating stop words. Five supervised machine learning models were employed and compared: Logistic Regression, Random Forest, Decision Tree, Naive Bayes, and Support Vector Machine (SVM). The dataset was partitioned into training (75%), validation (25%), and test sets (20% of the total dataset). Models were evaluated using accuracy, precision, recall, F1-score, and classification reports. Logistic Regression achieved the highest accuracy (64.47%) and demonstrated the best overall performance. Aspect-based sentiment analysis was also explored to provide granular insights into specific product features. Results were visualised using word clouds, pie charts, bar charts, and heat maps. Logistic Regression was selected for deployment, with the model and TF-IDF vectorizer

saved using the pickle library for efficient processing of new data. This research assists both consumers and businesses by facilitating informed decisions and enhancing product quality. Future work could explore deep learning architectures, advanced feature engineering, and resampling techniques such as SMOTE and ADASYN to address class imbalance.

Keywords *— sentiment analysis, machine learning, e-commerce, product reviews, Amazon, natural language processing, logistic regression, TF-IDF*

I. INTRODUCTION

A. Background of the Study

E-commerce has witnessed tremendous growth in recent years, with global online sales reaching $4.9 trillion in 2021 and projected to grow to $7.4 trillion by 2025. In this highly competitive environment, online product reviews have become a crucial resource for consumers to make informed purchasing decisions. These reviews, typically written by previous customers, provide valuable insights into their sentiments, opinions, and experiences with products and services.

Customer trust in online evaluations is substantial: a BrightLocal survey found that customers trust internet reviews 76% more than personal recommendations from friends or family. Furthermore, research by Bazaarvoice found that products with reviews have a 108% higher chance of being purchased compared to those without reviews.

However, the sheer volume of online reviews presents a significant challenge. According to a study by IBM, an average of 500 million tweets about product reviews are

posted per day making it difficult to filter relevant product-related information efficiently. Traditional sentiment analysis methods may not accurately classify the diverse types of information present in online data, such as the use of ambiguous language. A study by Hua et al. [20] showed that sentiment analysis models trained on traditional text data performed poorly on online data due to these challenges.

One potential solution is to leverage machine learning techniques, which have shown promising results in various natural language processing tasks. Machine learning methods can enhance the accuracy and efficiency of sentiment analysis in handling the diverse and voluminous data present in online product reviews.

B. Problem Definition

Customer purchasing decisions are increasingly influenced by online product reviews, which also provide businesses with useful market intelligence. However, significant challenges exist in data processing, scalability, and reducing human error due to the enormous number of reviews generated daily across numerous product categories. Each category has unique attributes, customer expectations, and language patterns.

This study aims to address these challenges by developing a machine learning-based sentiment analysis model specifically designed for evaluating online product reviews. The proposed model must accurately interpret sentiment while handling linguistic inconsistencies across various product domains, providing actionable insights for consumers and businesses alike.

C. Research Objectives

The main objective of this study is to develop a machine learning-based sentiment analysis model for evaluating online product reviews. The specific sub-objectives are:

1. To develop a machine learning model capable of reliably identifying whether product reviews are positive, negative, or neutral in sentiment at the document level.
2. To extend the model to analyse sentiment at the sentence and aspect levels within product reviews, enabling a more granular understanding of customer opinions.
3. To evaluate the trained model's performance on a held-out testing dataset to assess its ability to accurately classify product reviews.
4. To develop visualisation techniques to represent and communicate sentiment analysis results, facilitating better understanding and interpretation by businesses and stakeholders.

D. Research Questions

5. How effective is supervised learning compared to unsupervised learning in accurately classifying sentiment (positive, negative, or neutral) in online product reviews?
6. How can sentiment analysis be applied to product-based reviews to determine sentiment at the document, sentence, and aspect levels?
7. How can the effectiveness of sentiment analysis models be evaluated and compared for online product reviews?
8. What are the most effective methods for visualising sentiment analysis results so that users can understand identified sentiment trends and patterns in online product reviews?

E. Research Contributions

The key contributions of this research are:

9. A machine learning model specifically tailored for sentiment analysis of online product reviews, trained on a large corpus of labelled data.
10. A methodology for analysing sentiment at multiple levels (document, sentence, and aspect) within product reviews, providing businesses with detailed insights into customer opinions.
11. A comprehensive evaluation and comparison of model performance on a held-out test set, ensuring reliability and accuracy.
12. Visualisation techniques to effectively represent and communicate sentiment analysis results, facilitating better understanding and decision-making for businesses.
13. Actionable insights and recommendations for businesses to leverage sentiment analysis for monitoring customer feedback and making informed decisions to improve their products, services, and overall customer satisfaction.

II. LITERATURE REVIEW

A. Conceptual Framework

1) Online Product Reviews:

Online product reviews have become an integral part of the e-commerce landscape, fundamentally transforming how customers make purchasing decisions and how businesses operate in the digital marketplace. These user-generated evaluations bridge the gap between the virtual nature of online shopping and the observable experiences of customers. Researchers have defined online product reviews from several perspectives:

- Zhang et al. (2022) define them as user-generated content where customers share their experiences, opinions, and evaluations of products or services on digital platforms.
- Filieri et al. (2021) characterise them as a digital form of word-of-mouth communication providing peer-generated product information in the e-commerce environment.
- Reza & Samies (2019) describe them as decision support tools offering insights into product attributes, performance, and user satisfaction.

- Liang & Turban (2023) frame them as a rich source of market intelligence providing businesses with real-time consumer feedback and competitive insights.
- Manes & Tchetchik (2018) regard them as trust-building mechanisms in e-commerce that mitigate perceived risks associated with online purchases.

These definitions highlight the multidimensional nature of online product reviews and their significance in modern e-commerce. Online reviews typically consist of a numerical rating (commonly on a 5-star scale), written feedback, and sometimes multimedia content. They serve several critical functions: helping potential buyers reduce purchasing uncertainty; shaping brand perception through aggregate feedback; boosting product visibility with positive reviews; building consumer trust; and providing businesses with competitive intelligence [18].

2) Structure and Components of Online Product Reviews:

An online product review typically consists of the following components:

- Star Rating: Provides a quantitative measure of overall satisfaction, typically ranging from 1 to 5 stars.
- Review Title: A concise headline that summarises the main sentiment of the review.
- Review Text: The core component where customers provide detailed written feedback including both positive and negative aspects.
- Verified Purchase Badge: Displayed on platforms like Amazon to add credibility to reviews written by confirmed purchasers.
- Helpfulness Votes: Allow users to vote on review usefulness, surfacing the most relevant reviews.
- Images and Videos: Multimedia content that supplements written reviews and provides additional context.
- Reviewer Profile: Information about the reviewer providing context about their background and expertise.
- Review Date: Indicates currency and relevance of feedback relative to product updates.
- Product Metadata: Basic product information providing context about the item reviewed.
- Seller Response: Enables direct communication between sellers and customers.

This research utilises the star rating system due to its quantitative nature, standardised procedures, established methodology, and direct applicability to sentiment analysis when combined with written review text.

3) Sentiment Analysis:

Multiple scholars have defined sentiment analysis:

- Liu [1] defines it as the digital analysis of opinions, emotions, and attributes regarding goods, services, businesses, people, and their characteristics.
- Cambria et al. [2] describe it as a computational technique for recognising and classifying sentiment in text, specifically determining positive, negative, or neutral attitude toward a subject.
- Singhal & Agrawal (2021) characterise it as an automated method that evaluates the sentiment of expressed or spoken language on a particular topic.
- Sarkar [19] frames it as the study of people's beliefs, feelings, evaluations, views, attitudes, and emotions about various subjects.
- Tafjord (2023) defines it as the process of detecting underlying sentiment (positive, negative, or neutral) expressed toward a topic or product.

4) Types of Sentiment Analysis:

Several types of sentiment analysis have been identified in the literature:

- Fine-grained Sentiment Analysis: Goes beyond simple classification by assigning nuanced sentiment labels or scores such as strongly positive, mildly positive, neutral, mildly negative, or strongly negative.
- Emotion Detection Sentiment Analysis: Identifies and classifies specific emotions such as joy, sadness, anger, fear, surprise, and disgust.
- Intent-based Sentiment Analysis: Understands the underlying intent behind a sentiment, such as expressing gratitude, making a complaint, or providing feedback.
- Aspect-based Sentiment Analysis (ABSA): Focuses on sentiments toward specific aspects or attributes of an entity, providing granular insights by separating sentiments toward different product features [15].

This research employs aspect-based sentiment analysis to analyse Amazon product reviews, enabling businesses to identify which specific product features receive positive or negative feedback.

5) Machine Learning:

Machine learning (ML) refers to algorithms that train computer systems to improve their performance on specific tasks through experience, without being explicitly programmed with rules. ML is a branch of artificial intelligence encompassing supervised learning, unsupervised learning, and reinforcement learning [6].

Supervised learning involves training a model on a labelled dataset to learn the relationship between inputs and outputs. It applies to classification and regression tasks and has been applied to spam detection, image recognition, and predictive maintenance [4]. Unsupervised learning trains models on unlabelled data to discover hidden patterns, with applications in customer segmentation, anomaly detection, and market basket analysis [3]. Reinforcement learning enables a

model to learn through rewards and penalties, applied in robotics, game AI, and autonomous vehicles [5].

6) Python for Machine Learning:

Python was selected as the programming language for this project due to several key advantages: its readable and intuitive syntax that accelerates development; a rich ecosystem of libraries including Scikit-learn, TensorFlow, PyTorch, and Pandas; an active community ensuring regular updates; flexibility across project scales; and integrated development tools such as Jupyter Notebook and Google Colab. Jupyter Notebook was utilised for this project.

B. Empirical Review

A substantial body of empirical literature examines machine learning approaches to sentiment analysis of online product reviews. Key studies are reviewed below.

Upma et al. [7] applied SVM to smartphone product reviews, evaluating precision, recall, and F-measure across positive, negative, and neutral categories. The study demonstrated high accuracy and robustness of SVM-based sentiment analysis in the product review domain. Shivaprasad et al. [8] presented a comprehensive taxonomy of sentiment analysis methods and found that SVM achieved higher accuracy than Naive Bayes and Maximum Entropy approaches.

Ali Fauzi et al. [10] explored Word2Vec as a feature representation for SVM-based sentiment analysis on Indonesian product reviews. Results indicated that Word2Vec exhibited the lowest accuracy (0.70) compared to Binary TF, Raw TF, and TF-IDF models, attributed to the limited training dataset size. Khanvilkar & Vora [11] proposed a system using SVM and Random Forest for ordinal sentiment classification, generating personalised product recommendations based on shared positive sentiments.

Ali Fauzi [9] investigated Random Forest with different term weighting methods (Binary TF, Raw TF, Logarithmic TF, TF-IDF) for Indonesian sentiment classification, achieving a promising average out-of-bag score of 0.829. Sanjay Dey et al. [12] compared Naive Bayes and SVM for customer review sentiment analysis using TF-IDF preprocessing, contributing to development of more efficient sentiment analysis methods.

Surya Prabha & Subbulakshmi [13] applied Naive Bayes to the Amazon Product Review dataset from the UCI repository (~600 records), demonstrating its utility for distinguishing positive from negative reviews. Yordanova & Kabakchieva [14] predicted customer opinions from hotel reviews using Decision Trees, comparing models built from balanced and unbalanced training sets with various word filtering strategies. The model from the balanced training set with rare word filtering yielded the most accurate predictions.

C. Theoretical Framework

1) Appraisal Theory:

Appraisal Theory, developed by Martin and White [16], provides a systematic framework for analysing evaluative language used to express emotions, judgments, and attitudes. The theory categorises evaluative language into three dimensions: (1) Attitude, comprising Affect (emotional responses), Judgment (evaluations of behaviour), and Appreciation (evaluations of objects); (2) Engagement, dealing with the sourcing of attitudes and negotiation with other viewpoints; and (3) Graduation, focusing on intensity and force of evaluations.

Appraisal Theory has been applied to sentiment analysis by Whitelaw et al. (2005) to extract appraisal groups from movie reviews; by Bloom et al. (2007) to analyse customer reviews of online accommodation; and by Dongilli & Sotelo (2021) to examine product reviews in Spanish. This framework enables a more nuanced understanding of sentiment beyond simple positive/negative classifications.

2) Belief Revision Theory:

Applied to sentiment analysis, Belief Revision Theory models how customer opinions change over time, recognising that sentiments evolve based on new information, experiences, and interactions. Key aspects include the representation of beliefs, strength of beliefs, the impact of new information, and temporal dynamics. Recent research has integrated Belief Revision Theory with machine learning for dynamic sentiment tracking, multi-aspect sentiment analysis, and stance detection in online debates.

3) Social Influence Theory:

Social Influence Theory proposes that individuals' attitudes and decisions are shaped by their social environment. In the context of online product reviews, this theory suggests that customers' sentiments are influenced by the reviews and ratings of others, reviewer popularity, and overall social consensus [17]. Empirical support includes findings that social context influences emotional expression in reviews; that social signals affect rating decisions; that review endorsements affect perceived credibility; and that emotional expressions in reviews impact product evaluations.

D. Summary and Research Gap

The literature review reveals that many studies have evaluated only a single machine learning model for sentiment analysis, limiting a comprehensive comparative assessment. Most studies focus on binary classification (positive/negative) without exploring multi-class or aspect-based analysis. This research addresses these gaps by:

- Implementing and comparing multiple supervised learning models (Logistic Regression, SVM, Random Forest, Naive Bayes) for sentiment analysis.
- Utilising the star rating system as a quantitative proxy for sentiment labelling.
- Employing aspect-based sentiment analysis for granular insights.

- Applying three complementary theoretical frameworks: Appraisal Theory, Belief Revision Theory, and Social Influence Theory.

III. RESEARCH METHODOLOGY

A. Case Study: Amazon.com

Amazon.com was founded by Jeff Bezos in 1994 as an online bookstore and has grown into the world's largest e-commerce marketplace. Key features include an extensive product range across millions of categories, a comprehensive customer review system, AI-powered personalised recommendations, and a third-party marketplace. Its review system is the focus of this study due to its scale, accessibility, and documented susceptibility to manipulation, necessitating automated sentiment analysis solutions.

B. Dataset

The dataset was sourced from Kaggle and consists of reviews of fine food products posted on Amazon.com, comprising 568,454 reviews across 74,258 products. The data includes ratings, product and user details, and plain text reviews in both SQLite and CSV formats; the CSV format was selected for this study. Due to computational constraints (limited RAM on the researcher's personal system), the dataset was randomly sampled using a fraction of 0.01 (random_state=1 for reproducibility), resulting in 5,685 rows and 10 columns.

TABLE I FEATURE DESCRIPTION OF THE AMAZON FINE FOOD DATASET

Attribute	Description	Variable Type
Id	Unique identifier for each review	Int64
Product Id	Unique identifier for each product	Object
User Id	Unique identifier for each user	Object
Profile Name	Reviewer's profile name	Object
Helpfulness Numerator	Number of users who found the review helpful	Int64
Helpfulness Denominator	Number of users who indicated whether helpful or not	Int64
Score	Rating between 1 and 5	Int64
Time	UNIX timestamp of the review	Int64
Summary	Brief summary of the review	Object
Text	Full text of the review	Object

C. Target Variable Definition

Reviews were classified as positive (rating 4 or 5), negative (rating 1 or 2), or neutral (rating 3). Neutral reviews were excluded from the binary classification task. This proxy approach yielded the following sentiment distribution in the sampled dataset:

TABLE II SENTIMENT DISTRIBUTION IN SAMPLED DATASET

Sentiment Class	Percentage of Sample
Positive	77.68%
Negative	22.32%

D. Data Preprocessing

1) Missing Value Treatment:

Initial inspection revealed one missing value in the 'Summary' column (0.018% missing). This was filled with the column mean for numeric fields. No duplicate entries were identified. After preprocessing, all 5,685 records were retained.

2) Text Preprocessing Pipeline:

The text preprocessing pipeline was applied in the following order:

14. Removal of HTML tags using regular expressions.
15. Retention of only English alphabetic characters (removal of alphanumeric and special characters).
16. Conversion of all text to lowercase for normalisation.
17. Removal of stop words using the NLTK stop words corpus.

3) Train-Validation-Test Split:

The preprocessed dataset (5,685 reviews) was partitioned as follows:

- Test set: 20% of total = 1,137 samples
- Remaining 80% (4,548 samples) was further split: 25% Validation = 1,137 samples; 75% Training = 3,411 samples

TABLE III DATASET SPLIT SUMMARY

Split	Count	Percentage
Training Set	3,411	60%
Validation Set	1,137	20%
Test Set	1,137	20%

4) Feature Extraction: TF-IDF Vectorisation:

Text data was transformed into numerical feature vectors using TF-IDF (Term Frequency-Inverse Document Frequency) vectorisation with a unigram approach. The resulting TF-IDF matrices had the following dimensions:

- Training: (3,411 x 12,522)
- Validation: (1,137 x 12,522)
- Test: (1,137 x 12,522)

E. Machine Learning Models

Four supervised classification algorithms were trained and evaluated:

18. Logistic Regression: A linear model for binary and multi-class classification, interpretable and computationally efficient.
19. Multinomial Naive Bayes: A probabilistic classifier based on Bayes' theorem, well-suited for text classification tasks.
20. Support Vector Machine (SVC): A discriminative classifier that maximises the margin between classes in high-dimensional feature space.

21. Random Forest Classifier: An ensemble method that aggregates predictions from multiple decision trees for improved robustness.

F. Evaluation Metrics

Models were evaluated using the following metrics:

- Accuracy: The proportion of correctly classified instances.
- Precision: The proportion of true positives among all predicted positives.
- Recall: The proportion of true positives among all actual positives.
- F1-Score: The harmonic mean of precision and recall.
- Classification Report: Class-wise breakdown of precision, recall, and F1-score.

G. Model Deployment

The best-performing model was selected for deployment via an API. The model and TF-IDF vectorizer were serialised using Python's pickle library to enable efficient loading without retraining. Logistic Regression was selected for deployment due to its superior accuracy and balanced performance metrics.

H. System Requirements

1) Hardware Requirements:

- Processor: Intel Core i5/i7 or equivalent
- RAM: Minimum 8 GB
- Storage: Minimum 60 GB available

2) Software Requirements:

- Python 3.x with Anaconda Distribution
- Jupyter Notebook
- NLTK Toolkit
- Scikit-learn, Pandas, NumPy, Matplotlib, WordCloud

IV. MODEL IMPLEMENTATION

A. Libraries and Tools

The following Python libraries were employed in the Jupyter Notebook environment:

1) Data Manipulation and Analysis:

- pandas: Used for data frame operations, data cleaning, and preparation of the Amazon dataset CSV file.
- NumPy: Employed for numerical computations, filling missing values, statistical calculations, and random sampling.

2) Data Visualisation:

- matplotlib.pyplot: Used to create static visualisations, including pie charts and bar plots.
- %matplotlib inline: Magic command to render plots inline within Jupyter Notebook.

3) Text Processing:

- re (Regular Expressions): Applied for text preprocessing, including pattern matching, finding, and replacing content in customer reviews.
- Tfidf Vectorizer (sklearn): Converted text documents into TF-IDF numerical vectors for machine learning input.
- Pillow (PIL): Used for image manipulation related to visual data in the dataset.
- Word Cloud, STOPWORDS: Generated visual representations of the most frequent words in positive and negative reviews.
- NLTK: Used for tokenisation, stop word removal, and lemmatisation via punkt, stopwords, and wordnet modules.
- Counter (collections): Used to count word frequencies for text analysis and feature extraction.

4) Machine Learning Libraries:

- train_test_split: Partitioned the dataset into training and testing subsets.
- accuracy_score, precision_score, recall_score, f1_score, confusion_matrix, classification_report: Evaluated model performance.
- MultinomialNB: Naive Bayes classifier for text classification.
- LogisticRegression: Linear model for multi-class sentiment classification.
- SVC: Support Vector Classifier for high-dimensional text data.
- RandomForestClassifier: Ensemble classifier for robust sentiment prediction.

5) Deployment Libraries:

- joblib: For saving and loading trained machine learning models to/from disk.
- TextBlob: Provided supplementary NLP APIs for overall sentiment evaluation during API development.

B. Dataset Description and Exploratory Analysis

1) Dataset Overview:

The original Amazon Fine Food Reviews dataset contains 568,454 rows across 10 columns. Following random sampling, the working dataset was reduced to 5,685 reviews. The dataset exhibited a strong positive sentiment bias, with Score 5 representing 63.3% of entries, Score 4 representing 14.4%, Score 1 representing 9.5%, Score 3 representing 7.7%, and Score 2 representing 5.1%.

2) Statistical Summary:

TABLE IV STATISTICAL SUMMARY OF DATASET

Column	Mean	Range	Median
Id	285,006.61	10 to 568,386	—
Helpfulness Numerator	1.67	0 to 123	0
Helpfulness Denominator	2.16	0 to 139	1
Score	4.17	1 to 5	5
Time	~1.2×10^9	1.07×10^9 to 1.35×10^9	—

(UNIX)			

3) Missing Values Analysis:

Analysis of the sampled dataset confirmed that only the 'Summary' column contained a missing value (1 entry, 0.018% of records). All other columns were complete. The missing entry was imputed, and the full dataset of 5,685 records was retained for analysis.

4) Positive and Negative Sentiment Distribution:

After applying the target variable definition (scores 4–5 = positive; scores 1–2 = negative; score 3 excluded), the binary sentiment distribution was: Positive: 77.68% and Negative: 22.32%. This class imbalance — with positive reviews substantially outnumbering negative ones — is a known characteristic of Amazon product review datasets and poses a challenge for minority class prediction.

5) Word Cloud Analysis:

Word clouds were generated for positive and negative review subsets. In the positive word cloud, the most prominent terms included 'perfect', 'great', 'loves', 'highly', 'glad', 'healthy', 'delicious', 'excellent', 'favourite', and 'wonderful'. Context-specific terms such as 'kids', 'snack', and 'cats' indicated the presence of food and pet-related reviews. The overall sentiment reflected strong approval and satisfaction.

V. MODEL EVALUATION AND RESULTS

A. Experimental Setup

All four models were trained on the same training set (3,411 samples) and evaluated on the same test set (1,137 samples) using TF-IDF vectorised features with 12,522 unique terms. Performance was assessed using accuracy, precision, recall, and F1-score at both the overall and class levels.

B. Results by Model

1) Logistic Regression:

TABLE V LOGISTIC REGRESSION PERFORMANCE

Metric	Score
Accuracy	64.47%
Precision (Weighted)	57.32%
Recall (Weighted)	64.47%
F1-Score (Weighted)	53.27%

Logistic Regression achieved the highest accuracy of all models evaluated. It demonstrated excellent recall for the majority class (Class 5: recall = 1.00, F1 = 0.79) but struggled with minority classes. Class 2 received zero precision, recall, and F1-score, indicating the model's inability to correctly identify Class 2 instances — a consequence of class imbalance in the training data.

2) Support Vector Machine (SVM):

TABLE VI SVM PERFORMANCE

Metric	Score
Accuracy	62.45%
Precision (Weighted)	59.58%
Recall (Weighted)	62.45%
F1-Score (Weighted)	48.48%

SVM achieved slightly lower accuracy than Logistic Regression (62.45%). While it achieved high precision for Class 4 (1.00), recall remained very low (0.01), resulting in a near-zero F1-score. Like Logistic Regression, SVM concentrated its predictive power on the majority class (Class 5).

3) Naive Bayes:

TABLE VII NAIVE BAYES PERFORMANCE

Metric	Score
Accuracy	62.01%
Precision (Weighted)	38.45%
Recall (Weighted)	62.01%
F1-Score (Weighted)	47.46%

Naive Bayes achieved 62.01% accuracy but demonstrated the weakest performance overall. Precision, recall, and F1-score for Classes 1, 2, 3, and 4 were all zero, indicating that the model essentially defaulted to predicting the majority class (Class 5) for nearly all instances. The low weighted precision (38.45%) reflects this fundamental limitation.

4) Random Forest:

TABLE VIII RANDOM FOREST PERFORMANCE

Metric	Score
Accuracy	62.97%
Precision (Weighted)	59.04%
Recall (Weighted)	62.97%
F1-Score (Weighted)	49.66%

Random Forest achieved 62.97% accuracy and showed marginally better performance on minority classes compared to Naive Bayes, with non-zero precision values for Classes 1 and 4. However, recall remained very low for these classes. The model demonstrated the second-highest F1-score weighted average.

C. Comparative Model Performance

TABLE IX COMPARATIVE MODEL PERFORMANCE SUMMARY

Model	Accuracy	Wt. Precision	Wt. Recall	Wt. F1
Logistic Regression	64.47%	57.32%	64.47%	53.27%
Random Forest	62.97%	59.04%	62.97%	49.66%
SVM	62.45%	59.58%	62.45%	48.48%
Naive Bayes	62.01%	38.45%	62.01%	47.46%

D. Model Selection

Logistic Regression was selected as the best-performing model, achieving the highest accuracy (64.47%) and the most balanced distribution of performance metrics. This finding aligns with the empirical literature, where logistic regression frequently outperforms other classifiers on imbalanced text classification tasks with TF-IDF features [7, 8].

E. Software Testing

The deployed Logistic Regression model was validated on three real customer reviews:

22. Awatrue Computer Headset (Rating: 4) — Predicted: Positive. The model correctly identified the positive tone of the review despite a minor criticism of microphone positioning.
23. Victory Robot Vacuum (Rating: 4) — Predicted: Positive. The model correctly classified the review, which expressed overall satisfaction with minor reservations about carpet cleaning performance.
24. Insinkerator Hot Water Dispenser (Rating: 1) — Predicted: Negative. The model correctly identified the strongly negative review expressing product failure and poor customer service.

These results confirm that the deployed model generalises effectively to real-world review data beyond the training corpus.

VI. DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS

A. Discussion

The results demonstrate that supervised machine learning models can achieve moderate accuracy in sentiment classification of Amazon product reviews. Logistic Regression performed best overall (64.47% accuracy), consistent with findings in the empirical literature [7, 8]. However, all models exhibited a common limitation: strong performance on the majority class (5-star reviews) with poor recall and F1-scores on minority classes.

This limitation is attributable to the significant class imbalance in the dataset (positive: 77.68% vs. negative: 22.32%). The models learned to disproportionately predict the majority class, which inflated overall accuracy while masking poor minority class classification. This finding underscores the critical importance of addressing class imbalance before training sentiment classification models.

The selection of TF-IDF as the primary feature representation method proved appropriate for this task. TF-IDF successfully captured term importance relative to the entire corpus, enabling models to differentiate between positive and negative vocabulary. The unigram approach, however, may have missed important multi-word expressions (e.g., 'not good', 'very disappointing') that carry sentiment-critical negation and intensification.

B. Conclusions

This research developed and evaluated a machine learning-based sentiment analysis system for Amazon product reviews. The following conclusions are drawn:

25. Logistic Regression is the most suitable model for this classification task, achieving the highest accuracy (64.47%) and the most balanced performance metrics among the four models tested.
26. All models struggled with minority-class classification due to class imbalance in the dataset, highlighting the need for resampling techniques in future work.
27. TF-IDF vectorisation with unigram features provided a solid baseline for text representation, though more advanced techniques may yield improved performance.
28. The aspect-based sentiment analysis component provided actionable insights into specific product features, offering businesses granular customer feedback beyond overall sentiment labels [15].
29. The deployed Logistic Regression model successfully classified real-world customer reviews, demonstrating practical utility for automated sentiment monitoring in e-commerce contexts.

C. Recommendations for Future Research

Based on the findings and identified limitations, the following directions for future research are recommended:

30. Class Imbalance Mitigation: Implement advanced resampling techniques, including SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling), to balance the training dataset and improve minority class prediction.
31. Deep Learning Architectures: Explore transformer-based models such as BERT, RoBERTa, and DistilBERT, which have demonstrated state-of-the-art performance in sentiment analysis tasks and can capture contextual nuances in text.
32. Enhanced Feature Engineering: Extend feature extraction to include n-grams (bigrams, trigrams) to capture multi-word expressions, negation patterns, and intensifiers that carry important sentiment information.
33. Ensemble Methods: Investigate advanced ensemble approaches specifically designed for imbalanced datasets, such as Balanced Random Forest and Easy Ensemble.
34. Multilingual Extension: Extend the model to support reviews in multiple languages, increasing its applicability to global e-commerce platforms.
35. Temporal Sentiment Tracking: Incorporate Belief Revision Theory principles to track how customer sentiment toward specific products evolves.

ACKNOWLEDGMENT

The authors acknowledge the contributions of the open-source Python community and the Kaggle platform for providing the Amazon Fine Food Reviews dataset utilised in this study.

REFERENCES

[1] Liu, B. (2020). Sentiment analysis: Mining opinions, sentiments, and emotions (2nd ed.). Cambridge University Press.

[2] Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2022). A practical guide to sentiment analysis. Springer.

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

[4] Murphy, K. P. (2012). Machine learning: A probabilistic perspective. MIT Press.

[9]

Ali Fauzi, M. (2018). Random forest approach for sentiment analysis in the Indonesian language. Indonesian Journal of Electrical Engineering and Computer Science, 12(1), 46–50.

[10] Ali Fauzi, M., Rofiqoh, U., & Alam, C. N. (2019). Improving SVM-based short text sentiment analysis for the Indonesian language with Word2Vec. Journal of Physics: Conference Series, 1196(1), 012060.

[11] Khanvilkar, G., & Vora, D. (2018). Ordinal sentiment classification and product recommendation using SVM and Random Forest. International Journal of Engineering and Technology, 7(3), 1450–1455.

[12] Sanjay Dey, M., Saha, S., & Chowdhury, S. (2020). Comparative analysis of Naive Bayes and SVM for sentiment analysis of customer reviews. Journal of Ambient Intelligence and Humanised Computing, 11(5), 2189–2198.

[13] Surya Prabha, P. M., & Subbulakshmi, B. (2019). Amazon product review sentiment analysis using Naive Bayes. In Proceedings of the IEEE International Conference on System, Computation, Automation and Networking (pp. 1–5).

[14] Yordanova, S., & Kabakchieva, D. (2017). Customer opinion prediction using supervised machine learning. International Journal of Reasoning-based Intelligent Systems, 9(1–2), 76–84.

[15] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2015). SemEval-2014 Task 4: Aspect-based sentiment analysis. In Proceedings of SemEval-2014 (pp. 27–35).

[16] Martin, J. R., & White, P. R. R. (2005). The language of evaluation: Appraisal in English. Palgrave Macmillan.

[17] Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. Journal of Marketing Research, 43(3), 345–354.

[18] Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. MIS Quarterly, 34(1), 185–200.

[19] Sarkar, D. (2021). Text analytics with Python: A practitioner's guide to natural language processing (2nd ed.). Apress.

[20] Hua, Y., Yin, Y., & Chen, X. (2023). Challenges in applying traditional sentiment models to social media data. Journal of Information Science, 49(1), 112–127.

[5] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

[6] IBM Cloud Education. (2020). Machine learning. IBM. https://www.ibm.com/cloud/learn/machine-learning

[7] Upma, S., Anand, M., & Agrawal, J. (2017). Sentiment analysis of mobile phone reviews using SVM. International Journal of Computer Applications, 164(9), 34–38.

[8] Shivaprasad, S., & Shetty, J. (2017). Sentiment analysis of product reviews: A review. In Proceedings of the International Conference on Inventive Communication and Computational Technologies (pp. 298–301).