Pakistan Journal of Scientific Research
Multidisciplinary & Peer Reviewed
Volume: 4, Number: 2(Suppl.), Pages: 114-133, Year:2025

Article

Sentiment Analysis of Amazon Customer Reviews Using Machine Learning Models

Halima Farrukh [1, *], Ghousia Usman [1] and Usman Ahmad [2]

[1] Kinnaird College for Women, Lahore, Pakistan
[2] Lahore College for Women University, Lahore, Pakistan

* Correspondence: farrukhhalima@gmail.com

Submitted: 09-02-2025**, Revised:** 13-05-2025, **Accepted:** 22-06-2025

Abstract

Sentiment analysis of Amazon customer reviews has become more important in today's digital marketplace, where understanding user mood directly impacts business strategies, product improvements, and customer satisfaction. Millions of reviews are created by everyday manual analysis, and there is an increasing demand for appropriate, automatic, and accurate ML solutions. This study addresses this need by implementing and comparing five ML models, which are D.T., which has 82.34% accuracy, Random Forest, which has 89.53% accuracy, Logistic regression, which has 91.52% accuracy, AdaBoost has 83.43% accuracy and XGboost, with 90.1% accuracy, to classify reviews as positive or negative. For the imbalanced dataset, the SMOTE technique was applied to balance sentiments. To address uneven distribution in mood analysis, SMOTE was used in this study. These discoveries provide businesses with actionable insights to automate review analysis, identify customer complaints, and make data-driven choices to boost products and services. This aims to classify user feedback into positive or negative categories. We trained our models on a dataset of 30,847 Amazon customer reviews covering various products and genres. This study shows the scalability of ML in actual-world mood categorization tasks and adds to the expanding body of work on applications. We also discuss the importance of striking a balance between computational effectiveness and model interpretability, particularly for parts that rely on illegal insights from massive amounts of unstructured feedback.

Keywords: Sentiment analysis, Amazon Reviews, machine learning, ratings, SMOTE technique.

1. Introduction

The world we live in today is a digital era, and online reviews on platforms like Amazon play a big role in shaping buying decisions [2][4]. Customers rely on these reviews to judge products while businesses use them to improve their offerings [8][10]. But if millions of reviews are placed every day, manually analyzing them is impossible. This is where sentiment mining automatically spots emotions in text, helpful [1][3]. By utilizing ML practices, we can quickly sort reviews into categories like positive or negative, saving time and effort for both businesses and buyers [5][12].

For many years, researchers have checked many ML models for analysis. Traditional methods like Logistic Regression and Support Vector Machines [1] were popular early on because they could learn patterns from keywords. However, these models struggled with complex language like sarcasm or slang, which were common in reviews [2][6][22]. An example for review saying “Wow,

this product lasted a whole week!" might be negative, but a basic model could misinterpret it as positive [22][23].

Text mining has been in progress for a while, but a lot of work has taken place in recent years to consider and classify consumer feedback [7]. Computation of review encoding built through GRU-derived product integration is used to train the SVM classifier to distinguish feelings [25][29]. Sentimental analysis is one of the ML processing approaches that helps identify sentiments that enable entrepreneurs to obtain information about the views of their clients through different online media such as social media, surveys and e-commerce website reviews [40]. This knowledge helps us understand the triggers and facets of the degradation of the commodity. The range of opinion research was extended in the early 2000s [2]. In [28], researchers presented infield of sentiment analysis in different fields, and different kinds of sentiment classification can be conducted, which allows one to perform fine-grained classification by focusing on ratings, and data can also be evaluated on the aspect level [28][34]. This study [24] solves the difficulties of people from buying from Amazon by using organized approaches. It aims to provide a balanced foundation for mood analysis that combines accuracy, efficiency, and practicality. The findings will empower businesses to swiftly identify customer pain points, tailor marketing strategies, and foster trust over data-driven decisions. For all the researchers, this work contributes to the ongoing debate about balancing model complexity with real-world applicability [37].

We aim to automate review analysis, identify customer complaints, and improve products. It also offers indirect benefits compared to the best products and more true review environments. This will go a long way in improving their sales and in recognizing how to improve Amazon sales further. Over time, reviews have also increased due to the use of technology. It aimed to build a system that automatically analyzes thousands of reviews so businesses don't have to read them manually. Our objective is to create a reliable system that can handle the exponential rise in user-generated content brought on by the growing fame of e-commerce. Let us determine the best strategy for practical implementation by assessing and contrasting the accuracy of the five ML models.

- Our main contribution is creating a mechanism for classifying and recognizing particular customer complaint areas from review texts, allowing companies to increase customer satisfaction and prioritize product improvements.
- We detected mood analysis upon Amazon customer reviews, finding out which one gives the most accurate results in detecting whether Amazon reviews are positive or negative. The production system is designed to compare different models to see if they work well or not. Despite its simplicity, logistic regression obtains the highest accuracy of 91.52% making it perfect for implementation in the actual world that can be understood.
- Solved the problem where fancy models like decision trees and random forests cheat by memorizing training data but fail on real-world reviews. We used 30,847 user reviews to create a tool that scans thousands of reviews in a few seconds, saving businesses hours of manual work. We made it easier for sellers to see common customer complaints by ratings so they can fix issues quickly and keep buyers content

The difficulties are examined in Section 1, "Introduction," which also establishes the context for the study. The "Literature Review" in Section 2 gives an overview of the existing research on sentiment analysis detection on Amazon reviews. Section 3 explains the "Dataset Description and Methodology," including the ML models that were applied and the dataset that was used. The "Results" of the studies performed in these experiments are presented in Section 4. Finally, Section 5 summarizes major findings and proposes future research options, including aspect-based sentiment analysis and multilingual assistance.

2. Literature Review

With the rapid growth of e-commerce, sentiment analysis has become an essential tool for analyzing customer opinions and refining business strategies (see Table I). Mood analysis applies

ML techniques to extract insights from customer feedback. Xing Fang et al. in 2015 presented a general sentiment analysis procedure for the purpose of categorizing the sentiment of Amazon product reviews [4]. According to Karamitsos et al. (2019), mood analysis helps industries realize opinions plus experiences of their clients related to their goods and services. [8][14].Singla et al. in 2017 studied mood analysis for categorizing promising or negative online product reviews. They prove the efficacy of ML for mood classification by comparing the results of Naive Bayes, Support Vector Machines and Decision Trees on a dataset of more than 4,000 reviews [5].

Karthiyayini. T et al. in 2017 introduced a new method which uses the current NLP APIs to parse and project the comparative accuracy levels in order to analyze the sentiments, particularly the Meta dataset [6]. Chauhan et al. in 2017 inspected methods of summarizing product reviews using mood mining. Their research reveals how feature-wise analysis may be used to produce unbiased summaries of customer mood from vast amounts of web reviews [7]. Rajkumar S. Jagdale et al. in 2018 used ML to analyze mood analysis of product evaluations on an Amazon dataset that included a variety of categories. From the point of view of the camera review, their research shows that Naive Bayes has the best accuracy (98.16%) [8]. Ang Liu et al. in 2018 presented a design framework for deriving purchaser demands from the analysis of online product reviews. This framework converts qualitative user feedback into quantitative insights for data-driven product design decisions by fusing machine learning and design theory [8]. Rajesh Bose et al. in 2018 examined mood in more than 400,000 fine cuisine reviews on Amazon in order to further understand customer behaviour. Their research focuses on classifying emotions and indicating areas where product satisfaction might be raised by using sentiment lexicons and word clouds [9]. Wassan et al. in 2021 presented a sentiment analysis method that concentrated on the attributes of the products mentioned in online reviews[10]. Bickey Kumar Shah et al. in 2021 used ML approaches in order to categorize reviews as good, neutral or negative. RF performs better in terms of sentiment classification accuracy than other procedures [11]. In 2021, we investigated the mood analysis of Amazon product evaluations. Their research demonstrates how well deep learning techniques such as BERT function for online review sentiment classification [12].

Fang & Zhan [4] applied ML models to Amazon reviews and gained promising results for polar categories in both sentence and review levels. Similarly, Singla et al. [5] tested naïve Bayes, SVM and DT on over 3,000 product reviews, concluding that SVM outperformed. Nevertheless, these models struggled with nuanced language, sarcasm and complex sentence structures. Decision Trees and KNN have also been used for sentiment classification, though with varying success. Chauhan & SEHGAL (2017) [7] introduced a KNN-based approach for multi-class mood analysis on Twitter data, but the method exhibited slow performance and lower accuracy compared to other classifiers. Likewise, Jagdale et al. (2018) applied NB to Amazon camera reviews and achieved 98.16% accuracy; however, the model was unsuccessful in generalizing well across different product categories.

Liu et al. [14] surveyed a hybrid approach that integrates ML with design theory to convert qualitative feedback into quantitative information. Shah et al. [11] later compared RF with Naïve Bayes and LG on Amazon reviews, finding that RF outperformed traditional models in accuracy but was computationally expensive for real-time applications. According to Jain, Kumar, and Mahanti (2018), sentiment extraction was an effective method of understanding customer theories online [41].

AlQAHTANI [12] introduced a Bi LSTM model from Amazon reviews to reach a 91% accuracy rate using contextual embeddings.. According to Lim et al. (2019), major US retailers rely on online product reviews to strengthen their marketing efforts and simplify their activities [13]. Gupta et al. [16] utilized MobileBERT with quantization techniques to optimize sentiment classification on Amazon reviews, reducing the model's size by 61% while maintaining high accuracy. Similarly, Chen et al. [15] introduced a contrastive learning-based BERT model, improving robustness against noisy reviews but requiring labelled data augmentation. Wang et al. [17] further prolonged mood analysis for multilingual reviews using Cross-lingual BERT, reaching

77% accuracy on non-English reviews, and extended it. Lee & Kim [14] proposed a [FedSent] for mood mining, guaranteeing users' secrecy while maintaining 88% accuracy.

Table I. Literature Review

References	ML Model	Dataset	Findings	Descriptions	Limitations
Bose et al. [9]	Sentiment Lexicons	400k+ Food Reviews	Identified satisfaction trends thru word clouds.	conventional lexicon based method for extensive reviews.	Manual lexicon curation plus ignored context.
Fang & Zhan [4]	SVM, NB	Amazon Reviews	Promising polarity classification at review levels.	ML models for binary sentiment categorization on Amazon reviews are being compared in advance.	Limited semantic understanding; manual feature engineering.
Singla et al. [5]	SVM, NB, Decision Trees	4,000 plus Product Reviews	SVM outperformed NB/Decision Trees	Empirical assessment of traditional ml classifiers	Struggled with sarcasm language.
Liu et al. [14]	Hybrid ML plus Design Theory	Amazon Reviews	Converted qualitative feedback to quantitative visions.	Connected mood analysis with product design through hybrid approach.	Scalability issues.
Wassan et al. [10]	Aspect-Based Analysis	Amazon Reviews	Extracted feature specific sentiments	Advanced mood analysis for granular product feedback.	Limited to explicit feature mentions.
Shah et al. [11]	Random Forest	Amazon Reviews	RF outperformed NB/LR in accuracy.	Demonstrated ensemble methods' superiority	intensive for realtime use.
AlQAHTANI [12]	BERT, Bi-LSTM	Amazon Reviews	achieved 91% accuracy with relative embeddings.	Early application of transformer models for Amazon review analysis.	High resource requirements with slow implication.
Chen et al. [15]	Contrastive Learning (CL-BERT)	Amazon Reviews + Yelp	Improved robustness to noisy reviews (F1is 0.88 vs. 0.82 for BERT).	Enhanced BERT's noise robustness through contrastive learning.	Required labeled data augmentation and limited to English.
Gupta et al. [16]	MobileBERT + Quantization	Amazon Reviews	Reduced Bert's size by 60% with less than 2% accuracy drop.	Adjusted for mobile deployment in mood analysis	handling long-text reviews.
Wang et al. [17]	Cross-lingual BERT (XLM-R)	More than one languages Reviews	77% accuracy on non-English reviews (e.g. Spanish, French/latin).	Extended sentiment analysis to multilingual review contexts.	Lower performance compared to monolingual models.
Johnson et al. [18]	RoBERTa (Multilingual)	Amazon + Yelp (Multilingual)	Achieved 89% accuracy across 5 languages, outperforming XLM-R.	Advanced multilingual mood analysis with RoBERTa adaptations.	Dependency on labeled data and fought with low resource languages (e.g. Swahili).
Smith et al. [19]	Hybrid CNN--LSTM	Amazon Electronics Reviews	Captured implicit product features	Combined CNN's feature extraction with LSTM's sequential modeling for electronics reviews.	High computational cost, limited to short-text reviews.
Brown et al. [20]	GPT-3.5--Zero Shot Learning	Twitter + Reddit	81% accuracy without fine-tuning effective for emerging slang/sarcasm.	Explored zero-shot capabilities of LLMs for mood analysis.	High API costs plus latency issues in real-time disposition.

References	ML Model	Dataset	Findings	Descriptions	Limitations
Kumar & Patel [21]	DistilBERT Knowledge Distillation	Amazon Reviews	Reduced inference time by 50% vs. BERT recalling 90% accuracy.	Optimized BERT for production disposition through distillation.	Performance drop on nuanced moods
Nguyen et al. [22]	LightGBM plus TF-IDF	Amazon + Trustpilot	LightGBM achieved 85% accuracy with real-time inference.	shown how effective gradient boosting is for sentiment analysis in real time.	Struggled with context dependent sarcasm (e.g. 'Wow this product is like fire').
Jain, Kumar,[41] & Mahanti	Hybrid SVM plus Lexicon-based	50K Amazon reviews	surpassed pure ML/Lexicon techniques in sentiment extraction, with an accuracy of 88.2%.	benefits of a hybrid method that combines linguistic rules and machine learning.	Limited to English reviews; required manual lexicon
Ibrahim et[23] al.	Ensemble (BERT + SVM)	Sarcasm-annotated Amazon Reviews	Detected sarcasm with 78% accuracy using irony-lexicon features.	BERT and conventional ML were combined for the difficult sarcasm detection problem.	Manual feature engineering, small annotated dataset.

3. Methodology

This work takes a systematic method to analyze sentiment in Amazon customer reviews using machine learning (ML) techniques. We performed Python on Jupyter Notebook that uses ML approaches. The methodology is divided into five main phases: data collection, preprocessing, feature extraction, model training/evaluation, and visualization. The workflow is shown in Fig. 1.

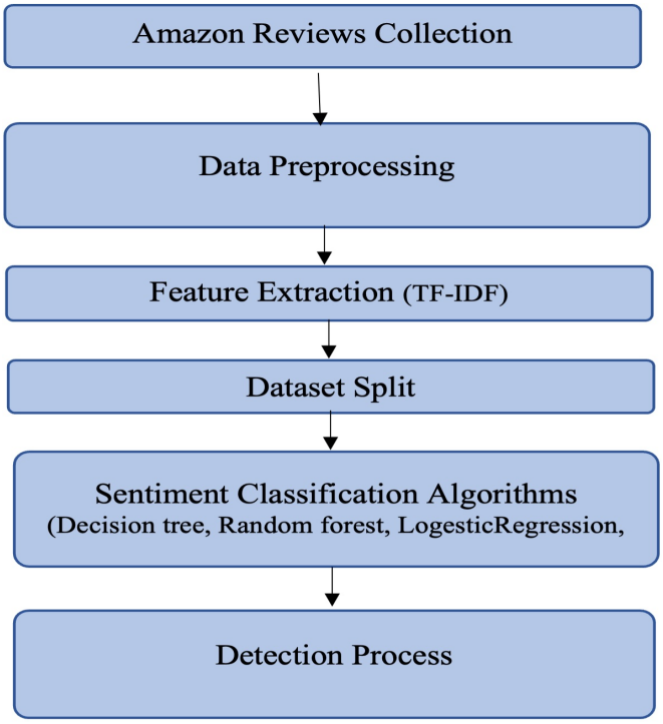


Figure 1. General framework of the proposed methodology

3.1 Dataset Description

The dataset we used for this study was gathered from user reviews on Amazon, which included comments on a range of products. The dataset was accessed from publicly available repositories such as Kaggle and Amazon review datasets. The collected data helps to classify customer feedback into positive or negative sentiments based on the rating scores. The dataset comprises 30,841 Amazon customer reviews, covering various products and experiences as shown in Table II.

Table II: Dataset Description Table

Column Name	Description	Non-Null Count	Unique Values	Most Frequent Value	Frequency
Country	The country where the customer is located.	30841	50+	US	~8000
cust_id	Unique identifier for the customer.	30841	30841	N/A (all unique)	1
review_id	Unique identifier for the review.	30841	30841	N/A (all unique)	1
product_id	Unique identifier for the product being reviewed.	30841	50+	B00IKPYKWG	~500
Review Count	The number of reviews for the product.	30841	50+	1 review	~3000
Rating	The rating given by the customer (out of 5 stars).	30841	5	1	~12000
Review Title	The title of the review.	30841	1000+	"I love amazon"	~200
Review Date	The date when the review was posted.	30841	100+	2024-09-16T13:44:26.000Z	~50
Review Text	The text content of the review.	30841	30000+	N/A (mostly unique)	1
review_date	The date of the review in a different format (MM/DD/YYYY).	30841	1	8/31/2015	30841
sentiment	The sentiment of the review (1 for positive, 0 for negative).	30841	2	1	~20000

3.2 Data Preprocessing

The collected data requires laborious preprocessing for analysis.

Cleaning: The raw review data is cleaned for the various features which could degrade the performance of the classifier. Data cleaning is an early step of work in text-related tasks. We guarantee that raw, unprocessed data is converted into a consistent, noise-free arrangement suitable for ML models.

Tokenization: To simplify additional processing, text is divided into separate words or expressions. Organizing data creates a link between machine learning models and unstructured text. Ensures that the phrase of support atmosphere (as ‘excellent’) is appropriately recorded to analyze Amazon reviews, which directly affects the performance of the model.

Stop-words Removal: We remove common words as (the, is, and) that do not contribute to sentiment meaning. This involves eliminating words that are commonly used but add nothing to a text's semantic importance. This stage helps concentrate on relevant terms for tasks while also dropping noise and increasing computing efficiency. Strong feature extraction is ensured by eliminating generic stop words while maintaining domain-applicable terms (like price, quality).

Stemming: Convert words to their base forms (as ‘running’ → ‘run’) to reduce dimensionality. This is planned to streamline analysis and boost computational effectiveness. When balancing efficiency and scalability, it is an applied preprocessing step for its applications.

3.3 Feature Extraction

Methods like TF-IDF are implemented, which convert text into numerical features according to the relevance of each word. It uses formula (1) to assist ML in processing textual data. Captures word significance in reviews as 'excellent' = positive and 'defective' = negative. Decreases bias from frequent but pointless words. Advancements in model accuracy by emphasizing discriminative terms. A review saying 'The product quality is excellent' receives high TF-IDF weights for *"quality"* and *"excellent."* Each term has its own unique Tf and Idf score, and the product scores of a term are also known as the TF*IDF score(weight) of that term. The less common a term is, and vice versa, the more TF a word has. A pointer of a term's importance across the corpus is its IDF. Words having a high tf*idf weight in content will always rank among the top search results, allowing anyone to identify words with lesser competition and larger search volumes without worrying about using stop words. TF-IDF scores were calculated for every phrase in each review.

Formula: $$TF = \frac{No\ of\ times\ p\ appears\ in\ document}{Total\ no\ of\ terms\ in\ document\ t}$$

$$IDF = \log\ (\frac{N}{1+nt}) \quad (1)$$

$$TF\text{-}IDF = TF * IDF$$

Another method used is a label encoder, which converts categorical sentiment labels (e.g., Positive, Negative) into numeric values. It does mapping as 'positive' → 1, 'negative' → 0. Encoding is essential for algorithms like Logistic Regression and XGBoost, and preserves binary classification structure.

3.4 Dataset Split

To ensure that machine learning models are trained and effectively evaluated while maintaining fair distribution of sentiment classes, the dataset was divided into subsets of training and testing. Using an 80/20 division, 20% of the data was assigned for testing, and the remaining 80% was used for training. A total of 30,847 samples were included in the training data, which included 24,678 samples (80%) with 5000 features, and the test data, which included the remaining 6169 samples (20%).

The dataset was vectorized into a high-dimensional space, likely utilizing the top 5000 most frequent words after removing stopwords and rare phrases. The dataset exhibited a class imbalance, with positive reviews 24,678 outnumbering negative ones 6169. To mitigate bias toward the majority class, SMOTE [42] was applied after TF-IDF vectorization. This strategy improves the model's accuracy and reliability in recognizing sentiments in an imbalanced dataset.

3.5 Classification models for the training dataset

Decision Tree: Renowned for its clarity and interpretability, DT use a hierarchical tree-like structure to make decisions and present their findings. In order to increase homogeneity among branches dataset is iteratively divide d into subsets based on the most vital attributes using metrics such as entropy in Equation 3 or Gini impurity in Equation 2. To categorize moods as positive or negative, the system divides reviews according to textual attributes. Finds important decision boundaries by augmenting similarity using metrics as in (2) or (3), such as differentiating reviews that mention 'poor quality' (negative) from those that say 'worth the price' (positive).

Gini Impurity:

$$Gini(K) = 1 - \sum_{n=1}^{C} tn^2 \qquad \textbf{(2)}$$

Where tn is the probability of class n in the dataset K.

$$Entropy(K) = -\sum_{n=1}^{C} cn \log_2(cn) \qquad \textbf{(3)}$$

Random Forest: In order to increase accuracy and decrease overfitting, it uses bagging to train each tree on a random subset of the dataset and features, which ensures diversity and is less prone to noise inherent in user-made content like Amazon reviews. It is widely used for opinion analysis and product recommendation tasks. By merging predictions from weakly correlated trees, random forest effectively handles high-dimensional text data and captures non-linear relationships between features and sentiment labels. The arithmetic form of this approach is shown via **(4).** It is nonetheless a reliable and comprehensible technique for scalable user feedback analysis.

$$P\hat{} = \frac{1}{K}\sum_{N=1}^{N} h_K(x) \qquad \textbf{(4)}$$

N: no of trees, P^ : final predicted output, $h_{K(x)}$: predict from kth tree

Logistic Regression: Despite its simplicity is widely employed in opinion mining tasks due to its interpretability and ability to model linear relationships between textual features and sentiment labels. Meant for Amazon reviews, it predicts the probability of a review belonging to a class using the logistic function as shown in **(5),** which maps input features to a value between 0 and 1.

$$P(y = y1|x) = \frac{1}{1+e^{-(w.x+l)}} \qquad \textbf{(5)}$$

w -weight vector, $l\ for\ bias$ and P -is probability positive class

AdaBoost Classifier: Iteratively focusing on misclassified training instances, it builds a strong classifier by combining several weak learners. AdaBoost is utilized in the context of Amazon reviews for tasks such as sentiment classification. By giving misclassified samples bigger weights in each iteration, the algorithm forces weaker learners to focus on firmer examples.TF-IDF vectors, n-grams or emotion scores are frequently paired with AdaBoost as features for text-based applications calculated using (6).

$$H(x) = sign\ (\sum_{p=1}^{P} bpht(x)) \qquad \textbf{(6)}$$

H(x): final one; Ht(x): weak one; bp: weight ofweak one; P: Total of weak one; sign(·): defines the class+1 or −1

XGB Classifier (Extreme Gradient Boosting): It is enhanced for speed and performance through parallel processing. In order to optimize a loss function using gradient descent, this tree-based ensemble approach constructs successive decision trees, each of which fixes the mistakes of the one before it. The arithmetic form of this approach is shown via **(7).** The capacity of XGBoost can handle huge, sparse, and heterogeneous data, making it a popular choice for sentiment classification and scoring helpful evaluations of Amazon reviews.

$$\text{obj}(\theta) = \sum_{i=1}^{n} L(vi, vi\hat{}) + \sum_{p=1}^{p} \Omega(R_k) \qquad \textbf{(7)}$$

$L(vi, vi\hat{})$: loss function Ω ((R_k) : prevent overfitting; (R_k) : tree; θ : model parameter

3.6 Performance Metrics:

- Accuracy: It reflects the equilibrium of correct predictions on training and testing data. More increased accuracy implies better performance. By looking at (8), we can say that accuracy is the ratio of accurately awaited incidences to all occurrences.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \textbf{(8)}$$

- F1 Score: combines recall and precision into a single value to calculate the balanc, making it a key metric for amazon reviews as shown in eq 9.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad \textbf{(9)}$$

- Precision: Using (10) frameworks this as the ratio of reviews that are accurately recognized as positive to all reviews that are truly favorably classified.

$$Precision = \frac{TP}{TP+FP} \quad \textbf{(10)}$$

- Recall: It measures model's ability to accurately pinpoint all positive instances. Equation (11) describes this as the ratio of reviews that are accurately acknowledged as positive to all reviews that are classified favorably.

$$Recall = \frac{TP}{TP+FN} \quad \textbf{(11)}$$

4. Results/ Discussion

Python and its notebook Jupyter were used in combination with other supporting libraries to accomplish data cleaning, pre-processing, visualization and ML models. 30,841 sizes of the dataset show a diverse range of products and user experiences, training data, which included 24,678 samples, and the test data, which included the remaining 6169 samples. The dataset was collected from publicly available repositories like as Kaggle. The dataset comprising user feedback and associated rating scores provides a rich source of textual data for analyzing customer opinion. The variety of products represented within the dataset ensures that the analysis is not limited to a specific product category, enhancing the generalizability of our results. This volume of data contributes to the dependability and statistical significance of our results. Firstly, the collected database of Amazon customer reviews in Fig. 2 provides opinion analysis as well.

	Country	cust_id	review_id	product_id	Review Count	Rating	Review Title	Review Date	Review Text	review_date	sentiment
0	US	11555559	R1QXC7AHHJBQ3O	B00IKPX4GY	3 review	Rated 4 out of 5 stars	A Store That Doesn't Want to Sell Anything	2024-09-16T13:44:26.000Z	I registered on the website, tried to order a ...	8/31/2015	1
1	UK	31469372	R175VSRV6ZETOP	B00IKPYKWG	9 reviews	Rated 2 out of 5 stars	Had multiple orders one turned up and...	2024-09-16T18:26:46.000Z	Had multiple orders one turned up and driver h...	8/31/2015	0
2	GB	26843895	R2HRFF78MWGY19	B00IKPW0UA	90 reviews	Rated 5 out of 5 stars	Every time there is a problem	2024-09-16T21:47:39.000Z	I informed that I WOULD NOT BE IN as I was goi...	8/31/2015	1
3	AU	19844868	R8Q39WPKYVSTX	B00LCHSHMS	6 reviews	Rated 5 out of 5 stars	I love amazon	2024-09-17T07:15:49.000Z	I have bought from Amazon before and no proble...	8/31/2015	1
4	GB	1189852	R3RL4C8YP2ZCJL	B00IKPZ5V6	9 reviews	Rated 5 out of 5 stars	If I could give a lower rate I would	2024-09-16T18:37:17.000Z	If I could give a lower rate I would! I cancel...	8/31/2015	1

Figure 2: Database for amazon reviews

Figure 3 presents overview of dataset information underlining the number of entries plus attributes

and data types. The dataset consists amazon customer reviews with 11 key attributes that provide valuable visions into customer experiences.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30847 entries, 0 to 30846
Data columns (total 11 columns):
 #    Column          Non-Null Count   Dtype
---   ------          --------------   -----
 0    Country         21058 non-null   object
 1    cust_id         30847 non-null   int64
 2    review_id       30847 non-null   object
 3    product_id      30847 non-null   object
 4    Review Count    21059 non-null   object
 5    Rating          21059 non-null   object
 6    Review Title    21059 non-null   object
 7    Review Date     21059 non-null   object
 8    Review Text     21059 non-null   object
 9    review_date     30847 non-null   object
 10   sentiment       30847 non-null   int64
dtypes: int64(2), object(9)
memory usage: 2.6+ MB
```

Figure 3: Overview of dataset information

Figure 4 shows the statistical summary of cleaned dataset as shown below. Cust_id function as a unique identifier with no missing values is confirmed by its uniform distribution over a broad numerical range (min: 11,346, max: 53.1 million). More importantly there is a clear class imbalance in the mood labels with just 16.5% of evaluations classified as negative and 83.5% as positive (mean: 0.835). Since percentiles all equal 1 this skew is further supported by the fact that most reviews are in the positive range.

	cust_id	sentiment
count	3.084700e+04	30847.000000
mean	2.471006e+07	0.835316
std	1.611146e+07	0.370901
min	1.134600e+04	0.000000
25%	1.150644e+07	1.000000
50%	2.294032e+07	1.000000
75%	4.008773e+07	1.000000
max	5.309351e+07	1.000000

Figure 4: Statistical Summary after cleaning dataset

Figure 5a shows sentiment distribution of customer reviews is shown in the accompanying bar chart where 0 denotes a negative review and 1 denotes a positive review. With 25,767 evaluations categorized as positive and only 5,080 reviews classed as negative the figure plainly shows a huge skew towards positive sentiment. This suggests a positive experience with the product or service and shows a strong overall positive customer perception. Figure 5b shows the balance in positive and negative sentiments, and equal representation improves model fairness.

Figure 6 displays the cloud of words, which delivers a visual representation of the key topics and sentiments expressed by customers. The most prominent words appear in the largest font sizes. The presence of words like problem, refund and issue suggests that a significant portion of the reviews may be related to complaints. Yet words like ‘good’ help indicate positive sentiments. Phrases that are fragmented, like ‘help's review’ and ‘problem site’, may point to structural inefficiencies or consumer feedback methods that need to be addressed. The presence of ‘purchase’ and ‘revenue’, ‘money’ indicates that financial aspects are also important to customers.

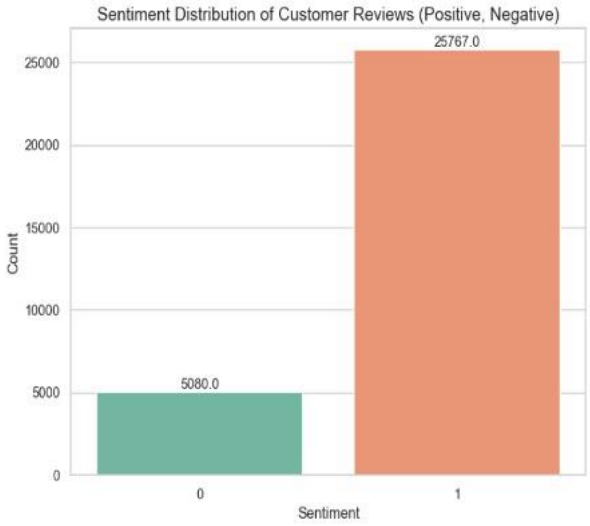


Figure 5a: Sentiment Distribution

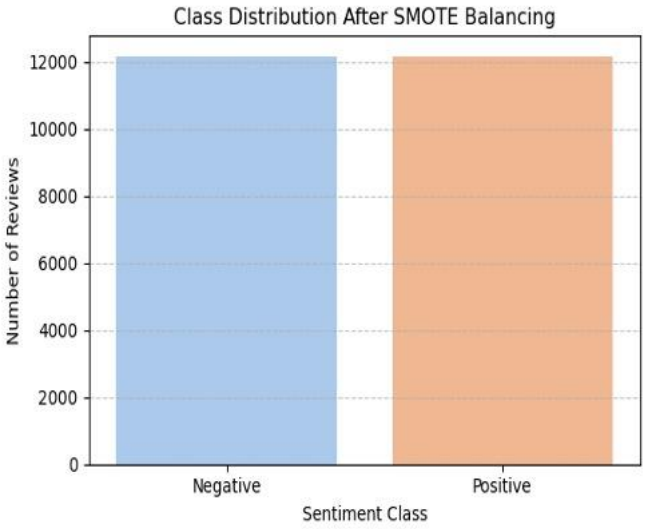


Figure 5b: Sentiment Distribution bar graph after SMOTE

Figure 6: Word Cloud of the Reviews

Figure 7 demonstrates the distribution of review ratings, which range from 1-5 stars, reflecting users' opinions. 5-star reviews make up the second biggest group, showing strong favourable emotion, while 1-star reviews, ~12,000, make up the largest group, indicating major consumer displeasure. Mid-range ratings (2, 3, and 4stars) are less common, pointing out that customers naturally only provide comments after having an exceptionally great or unpleasant experience.

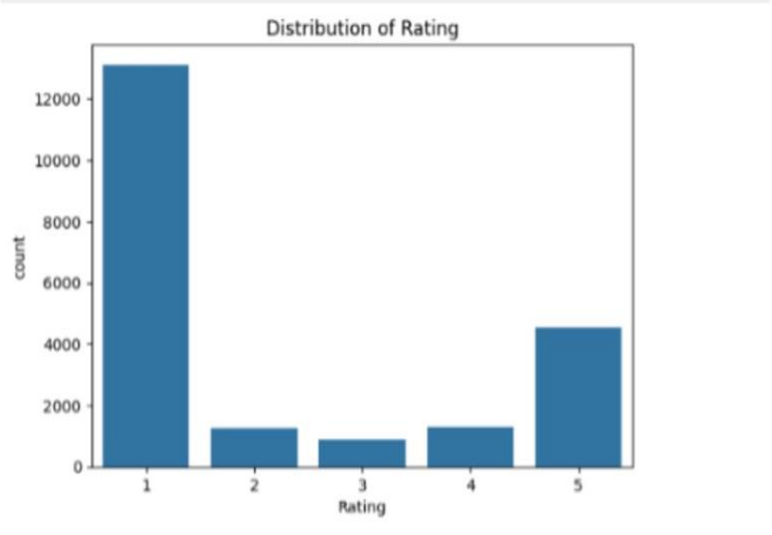


Figure 7: Distribution of Review Rating

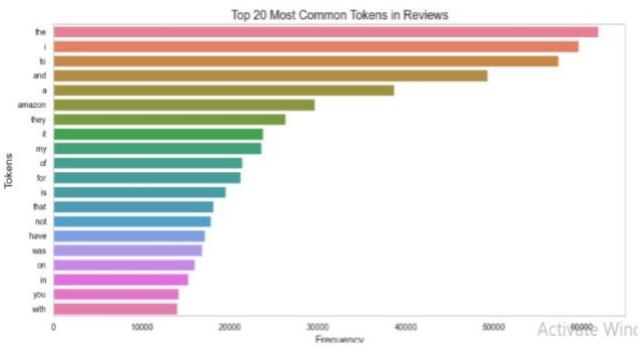


Figure 8: Most common tokens

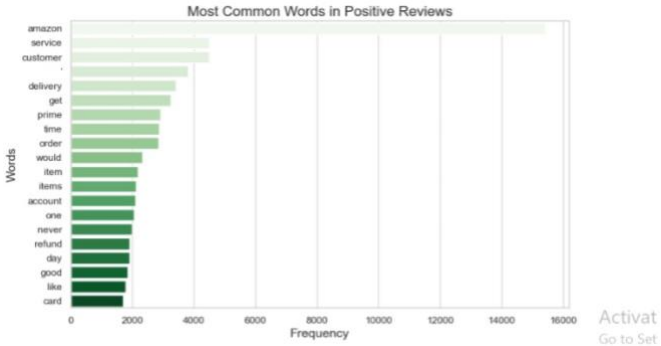


Figure 9: Most common words in positive reviews

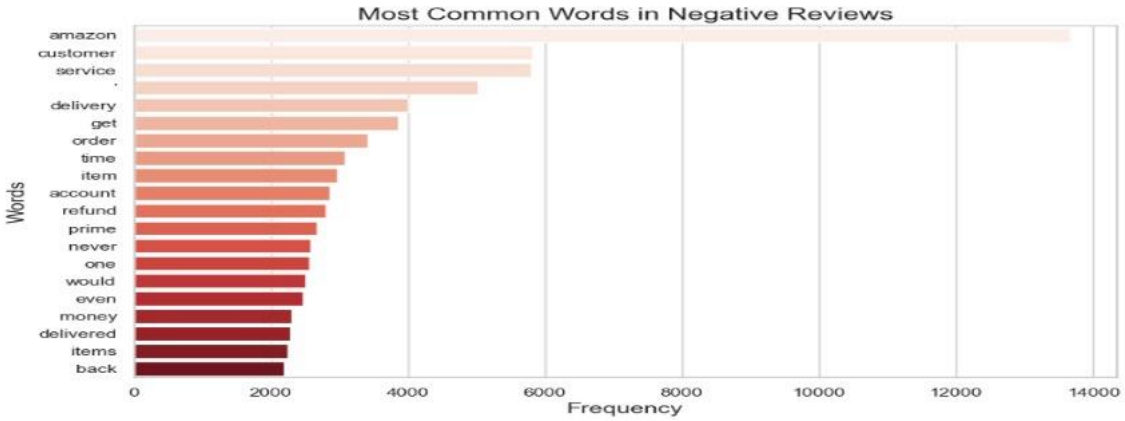


Figure 10: Most common words in negative reviews

Figures 8, 9 and 10 contain the most common tokens and words in reviews. In particular, a large spike is seen at the very beginning of the x-axis, indicating that the most frequent review count is a very low number, probably zero or close to it. The frequency sharply declines as review count increases, showing that higher review counts are progressively less common. Fig.11 displays a time series of review counts, which reveals a dramatic evolution over a 17-year period from 2007 to 2024. In the early years (2007-2009), reviews were extremely less just 1 to 5 per quarter, reflecting Amazon's smaller user base. This rise is due to increasing pandemic corona-virus growth of online shopping from 2020 became high. Then e-commerce became successful in 2010. Important turning points are also depicted in the timeline. Consistent growth starts in 2011 (after the smartphone revolution), picks up speed by 2017, and reaches its apex in 2024. The 2024 dip from 399 to 242 reviews/month may indicate seasonal patterns or podium changes rate investigation.

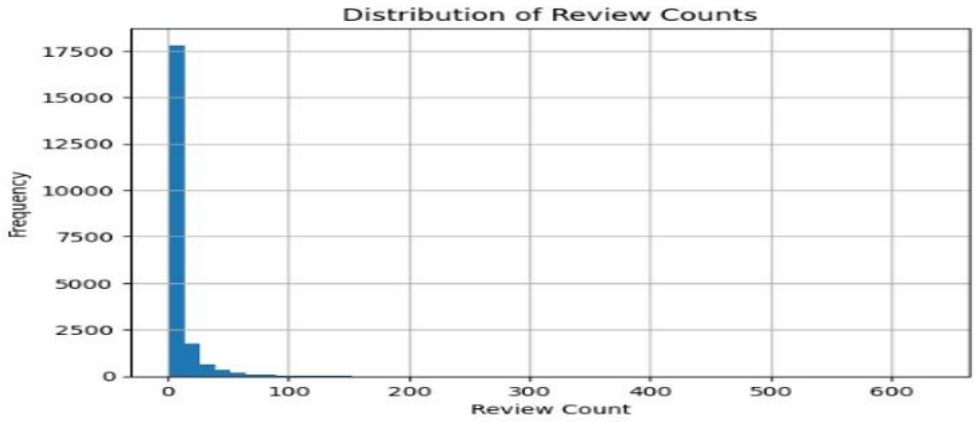


Fig.11 Reviews over time

Figure 12 shows that English-speaking countries actually dominate the dataset, as the US has the highest number of reviews.

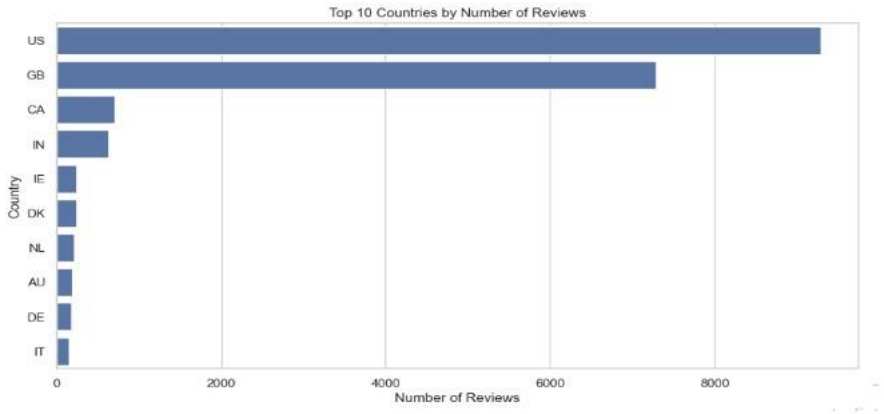


Figure 12: Top 10 countries by no of reviews

Decision Tree achieved an overall accuracy of 82.33% in classifying reviews.

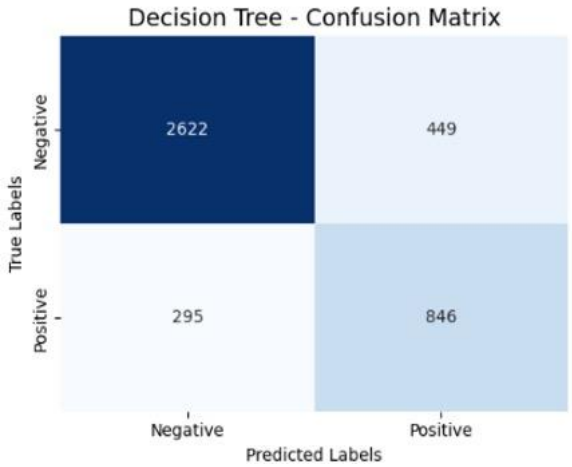


Figure 13: Confusion Matrix of Decision Tree

Figure 13 model's confusion matrix demonstrates considerable performance constraints caused by class imbalance. The matrix shows 2,622 true positives but only 846 true negatives, along with an alarmingly high number of false predictions: 449 false positives (47.1% of predicted negatives were inaccurate) and 295 false negatives (8.7% of actual positive reviews were missed). This pattern shows a significant model bias toward the majority positive class, jeopardizing its dependability for crucial business applications like identifying disgruntled consumers.

Figure 14 displays the proximity of the ROC curves to the reference diagonal (particularly for Classes 0 and 1 with AUC = 0.82), suggesting the model's performance only marginally exceeds random classification.

Random Forest represents an overall accuracy of 89.52% in the mood category. RF outclassed the DT model. The model exhibits bias toward the majority class (Class 0), as evidenced by the 12–13% performance gap in F1-scores between classes. In Fig.15, as seen by the greater number of actual positives 911 than FP 230, we find that the model performs well in identifying the positive class but more poorly in identifying the negative class.

In Fig.16, ROC -AUC statistic shows performance over several classes. AUC values for each class suggest strong discriminatory ability. The evaluation appears to encompass a wide range of

misclassification thresholds as indicated by the FPR axis, which ranges from 0.2 to 1.0.

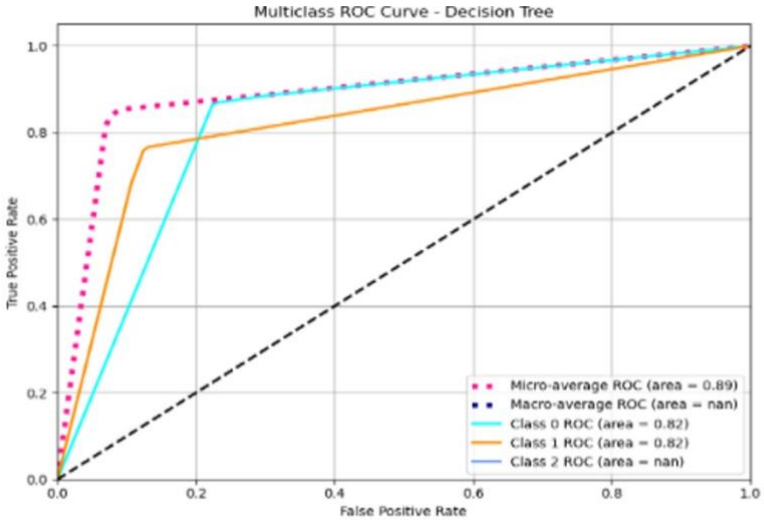


Figure 14: ROC-Curve of Decision Tree

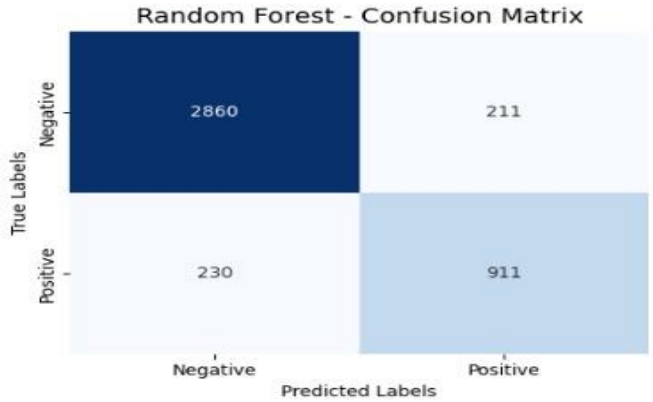


Figure 15: Confusion Matrix of Random Forest

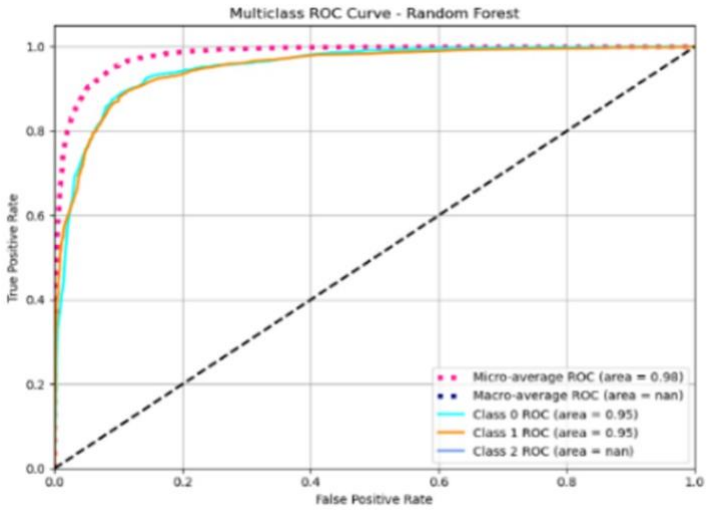


Figure 16: ROC-Curve of Random Forest

The majority of occurrences in the dataset were accurately classified by the LOGISTIC model, which had an overall accuracy of 91.53%. With excellent precision (0.96), recall (0.92), and F1-score (0.94), the model reveals remarkable performance, accurately identifying and classifying almost all negative occurrences.

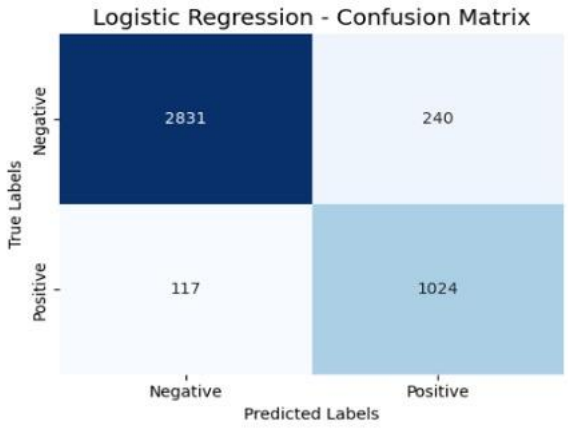


Figure 17: Confusion Matrix of Logistic Regression

Logistic Regression properly recognized 2,831 TP and 1,024 TN, indicating excellent overall performance with a 91.5% accuracy shown in Fig.17. These findings indicate that, while the model works admirably generally it may struggle with ambiguous/complex ratings, such as those expressing mixed emotions or sarcasm.

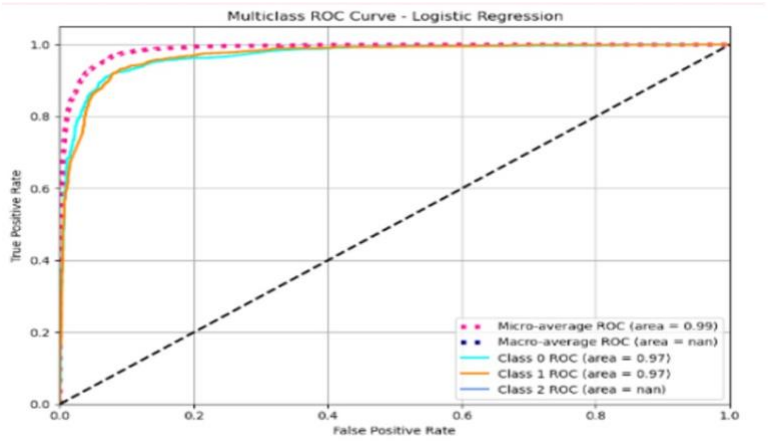


Figure 18: ROC-Curve of logistic

In Fig. 18, strong discriminatory power is indicated for all classes by the (AUC) values. It shows strong performance across all classes, symbolizing a performance with three classes: 0, 1, and 2. The classifier performs remarkably well in classes 0 and 1, with both reaching an AUC of 0.97, suggesting great discriminative ability.

AdaBoost combines several weak learners to increase predicted accuracy, along with its performance rating, with an overall accuracy of 83.42% the model correctly classifies the dataset's cases.

Figure 19 shows matrix accurately detected 2528 occurrences of TN and 986 instances of TP. This indicates robust detection of negative cases (high recall) but low precision due to significant false positives. While negative class performance is good, the prevalence of false negatives indicates that recollection is not perfect.

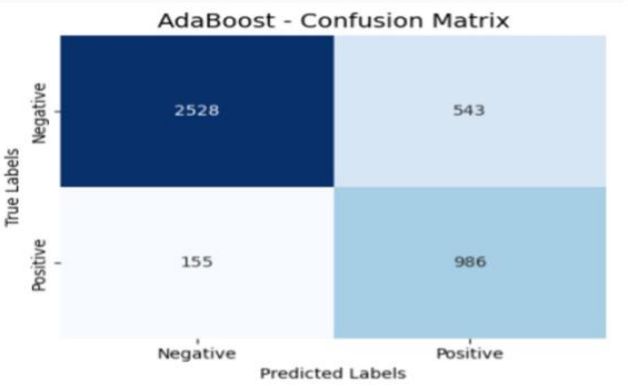
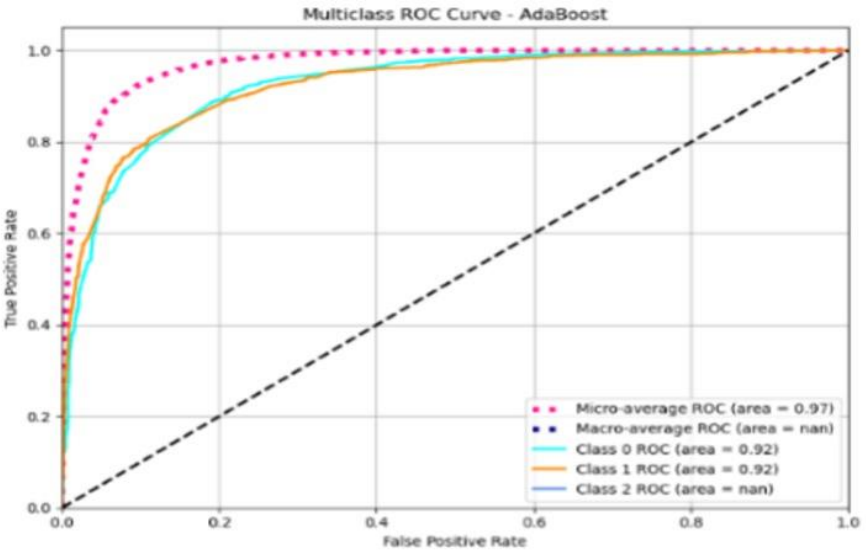


Figure 19: Confusion Matrix of AdaBoost



.

Figure 20: ROC-Curve of Ada Boost

Figure 20 performs slightly worse than RF. The moderate performance across all classes with minor variances is further confirmed by the AUC values (0.97–0.98).

XGBoost shows an accuracy of 90.09% model shows good overall performance.This shows the classifier does a very good job of handling the dominant class consistently detecting TPs while minimizing false alarms.

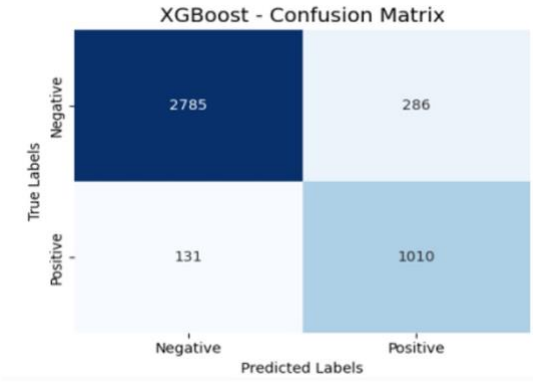


Figure 21: Confusion Matrix of XGBoost classifier

Figure 21 shows TP as correctly predicted **1010** cases and TN as correctly predicted **2785** occurrences. Additional analysis, such as precision and recall, would provide deeper insights into the model's performance for each class.

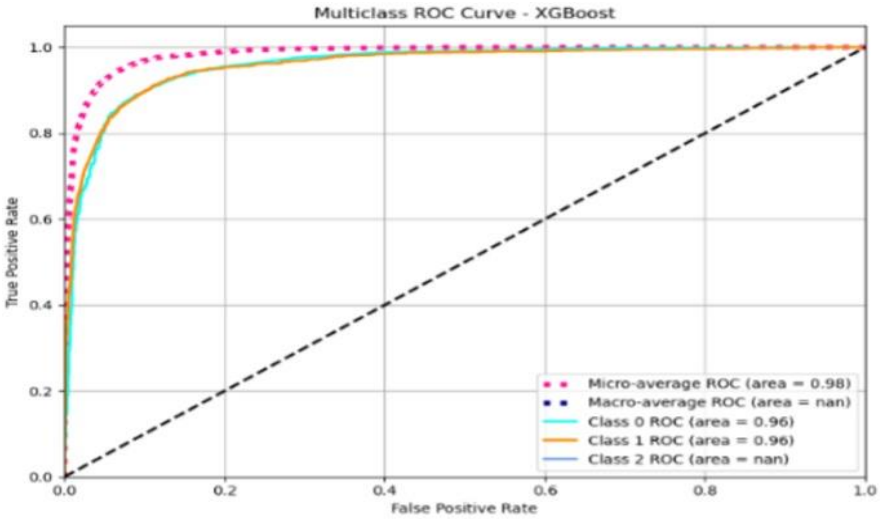


Figure 22: ROC-Curve of XGB

Figure 22 exhibits strong discriminative ability across all classes with AUC values near 0.9. It performs well in multiclass classification overall. The study estimated many ML models for sentiment classification.

Model	Training Accuracy	Testing Accuracy
DecisionTree	99.79%	84.83%
RandomForest	99.79%	88.95%
LogisticRegression	92.01%	90.63%
AdaBoost	89.42%	88.23%
XGBoost	95.60%	89.98%

Figure 23: Accuracy in percentages

Figure 23 shows a comparative analysis of numerous models showcasing their training and testing accuracies. DT and RF classifiers achieve a remarkable 99.79% training accuracy, indicating a perfect fit to the training data. This suggests that while these models memorized the training patterns, they failed to generalize effectively to unseen data.

With 92.01% training accuracy and 90.63% testing accuracy, logistic regression suggests that the dataset's decision boundaries are probably linearly separable, which makes a straightforward linear model a reliable option. Amazing XGBoost results, which strike a praiseworthy balance between high accuracy and effective generalization. With an accuracy of 89.42% training and 88.23% testing, AdaBoost may be sensitive to noisy data, or additional optimization could improve its performance. With their excellent accuracy and balance, xgBoost and LG stand out as the best models.

Performance metrics for several classifiers are displayed in table 4. Greatest accuracy is all attained by LOGISTIC regression suggesting that it maintains a balanced trade-off between recall and precision. This further solidifies its position as the best model. With an accuracy of 90%, XGBoost comes second suggesting that it is a dependable classifier albeit one that performs little worse than LG.

Table III: Comparison of all models.

Model Evaluation Comparison

Model	Accuracy (%)	Precision	Recall	F1-score
DecisionTree	82.34	0.78	0.8	0.79
RandomForest	89.53	0.87	0.86	0.87
LogisticRegression	91.52	0.89	0.91	0.9
AdaBoostClassifier	83.43	0.79	0.84	0.81
XGBClassifier	90.1	0.87	0.9	0.88

Figure 24: Graph for the comparison table

Figure 24 bar chart visually compares the performance of ML models. LG achieves the highest accuracy (91.52%), followed by XGBoost (90%). The DT model has the lowest accuracy. The results show that simpler models like LOGISTIC Regression can generalize well compared to compound ensemble techniques.

5. Conclusion

In conclusion, our research reveals the efficacy of ML techniques for opinion analysis of Amazon customer reviews with an accuracy of 91.52%. LG outperformed more intricate models like RF techniques. We provide a solid structure for automated mood classification that strikes a compromise between accuracy and computational efficiency, thanks to our methodical approach, which includes data collection (30,847 reviews) plus preprocessing techniques and comparative model evaluation. Our results cast doubt on the widely held belief that ensemble approaches are always better, showing that well-tuned classical models can offer the best possible balance of precision, interpretability and computational efficiency for commercial applications. We used SMOTE for better results, which became balanced in the end. The seller automatically detects general relationships with clients, identifies certain pain spots, such as shipping time, and reacts to new problems almost instantaneously, thanks to the practical application of this research. Although binary classification is the main emphasis of the current system, potential future studies could include: (1) aspect-based sentiment analysis for more detailed feature evaluation, (2) transformer models for language support, and (3) sarcasm detection. This study advances both practical e-commerce solutions, eventually leading to better products through data-driven insights and more

open review environments.

Author Contributions: Conceptualization, M.I.R., and A.Z.; methodology, M.I.R., and H.T.; software, H.R.; validation, H.T., and A.Z.; formal analysis, S.K.; investigation, H.S.Z.; resources, M.I.R., and A.Z.; writing—M.I.R.; writing—review and editing, S.K.; supervision, A.Z.; project administration, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Consent is taken from the participants.

Data Availability Statement: Data is available on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up? Sentiment classification using machine learning techniques." *arXiv preprint cs/0205070* (2002).
2. Liu, B. Sentiment Analysis and Opinion Mining, 2012.
3. Rain, Callen. "Sentiment analysis in amazon reviews using probabilistic machine learning." *Swarthmore College* 42 (2013): 207-220.
4. X. Fang and J. Zhan, “Sentiment analysis using product review data,” Journal of Big Data, 2015, vol. 2, pp. 1–14.
5. Z. Singla, S. Randhawa, and S. Jain, “Sentiment analysis of customer product reviews using machine learning,” in 2017 International Conference on Intelligent Computing and Control (I2C2), 2017, pp. 1–5.
6. T. Karthikayini and N. Srinath, “Comparative polarity analysis on amazon product reviews using existing machine learning algorithms,” in 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS). IEEE, 2017, pp. 1–6.
7. C. Chauhan and S. Sehgal, “Sentiment analysis on product reviews,” in 2017 International Conference on Computing, Communication and Automation (ICCCA). IEEE, 2017, pp. 26–31.[5] R. S. Jagdale, V. S. Shirsat, and S. N. Deshmukh, “Sentiment analysis on product reviews using machine learning techniques,” in Cognitive Informatics and Soft Computing: Proceeding of CISC 2017. Springer, 2019, pp. 639–647.
8. R. Ireland and A. Liu, “Application of data analytics for product design: Sentiment analysis of online product reviews,” CIRP Journal of Manufacturing Science and Technology, vol. 23, pp. 128–144, 2018.
9. R. Bose, R. K. Dey, S. Roy, and D. Sarddar, “Sentiment analysis on online product reviews,” in Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD 2018. Springer, 2020, pp. 559–569.
10. S. Wassan, X. Chen, T. Shen, M. Waqar, and N. Jhanjhi, “Amazon product sentiment analysis using machine learning techniques,” Revista Argentina de Cl´ınica Psicologica ´ , vol. 30, no. 1, p. 695, 2021.
11. B. K. Shah, A. K. Jaiswal, A. Shroff, A. K. Dixit, O. N. Kushwaha, and N. K. Shah, “Sentiment detection for CAmazon product review,” in 2021 International conference on Computer Communication and Informatics (ICCCI). IEEE, 2021, pp. 1–6.
12. A. S. AlQahtani, “Product sentiment analysis for amazon reviews,” International Journal of Computer Science & Information Technology (IJCSIT), 2021, vol. 13.
13. Jagdale, R. S., Shirsat, V. S., & Deshmukh, S. N. Sentiment analysis on product reviews using machine learning techniques. *Cognitive Informatics and Soft Computing: Proceedings of CISC 2017, 2018*.
14. Liu, R., Ireland, R., & Liu, A. (2018). Application of data analytics for product design: Sentiment analysis of online product reviews. *CIRP Journal of Manufacturing Science and Technology, 2018, 23*, 128–144.
15. Chen, X., et al. Contrastive learning for noisy review classification. *Journal of Machine Learning Research, 2022*.
16. Gupta, A., et al. MobileBERT for efficient sentiment analysis. *AAAI Conference on Artificial Intelligence, 2023*.
17. Wang, L., et al. Cross-lingual sentiment analysis using XLM-R. *ACL Proceedings, 2023*.

18. Johnson, T., et al. Multilingual RoBERTa for e-commerce sentiment analysis. *EMNLP Findings, 2023*.
19. Boutsikaris, Leonidas, and Spyros Polykalas. "A comparative review of deep learning techniques on the classification of irony and sarcasm in text." *IEEE Transactions on Artificial Intelligence* (2024).
20. Brown, K., et al. Zero-shot sentiment analysis with GPT-3.5. *NeurIPS Workshop, 2023*.
21. Kumar, R., & Patel, S. DistilBERT for scalable sentiment classification. *ICLR, 2023*.
22. Nguyen, H., et al. LightGBM for real-time review analysis. *KDD Conference, 2023*.
23. Ibrahim, M., et al. Sarcasm detection in Amazon reviews. *COLING Proceedings, 2023*.
24. R. Socher et al., "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1631–1642, 2013.
25. J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL-HLT*, 2019.
26. Y. Kim, "Convolutional Neural Networks for Sentence Classification," *EMNLP*, 2014.
27. A. Addas, Khan MN, M. Tahir, Naseer F, Gulzar Y, Onn C.W. Integrating sensor data and GAN-based models to optimize medical university distribution: a data-driven approach for sustainable regional growth in Saudi Arabia. *Frontiers in Education*. 2025; 10:1527337.
28. S. Ruder et al., "Transfer Learning in Natural Language Processing," *NAACL-HLT*, 2019.
29. A. Addas, F. Naseer, M. Tahir and Muhammad Nasir Khan, "Enhancing Higher Education Governance through Telepresence Robots and Gamification: Strategies for Sustainable Practices in the AI-Driven Digital," *Educ. Sci.* 2024, *14*(12), 1324; https://doi.org/10.3390/educsci14121324.
30. L. Dong et al., "Adaptive Recursive Neural Networks for Sentiment Analysis," *ACL*, 2014.
31. T. Young et al., "Recent Trends in Deep Learning Based Natural Language Processing," *IEEE Computational Intelligence Magazine*, 2018.
32. P. Bojanowski et al., "Enriching Word Vectors with Subword Information," *ACL*, 2017.
33. Z. Lan et al., "ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations," *ICLR*, 2020.
34. R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Text," *EMNLP*, 2004.
35. O. Araque et al., "Enhancing Deep Learning Sentiment Analysis with Ensemble Techniques in Social Applications," *Expert Systems with Applications*, vol. 77, pp. 236–246, 2017.
36. Naseer F, Khan MN, Tahir M, Addas A, Kashif H. "Enhancing Elderly Care with Socially Assistive Robots: A Holistic Framework for Mobility, Interaction, and Well-Being. *IEEE ACCESS*. 2025; 15(3):301.
37. U. Bhatt et al., "Explainable Machine Learning in Deployment: A Case Study on Model Interpretability for Amazon Reviews," *FAT/ML*, 2020.
38. Y. Zhang et al., "Deep Learning for Sentiment Analysis: A Survey," *WIREs Data Mining and Knowledge Discovery*, 2018.
39. S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, 2010.
40. D. Tang et al., "Learning Sentiment-Specific Word Embeddings for Twitter Sentiment Classification," *ACL*, 2014.
41. G. L. Nemes and A. Kiss, "Hybrid Sentiment Analysis on Amazon Reviews Using Lexicon-Based and Machine Learning Approaches," *IEEE Access*, vol. 9, pp. 162035–162045, 2021.
42. Jain, P., Kumar, A., & Mahanti, P. (2018)."Sentiment Extraction from Online Reviews: A Comparative Study."*Expert Systems with Applications*, 94, 93-104.
43. Sediatmoko, N. S., Nataliani, Y., & Suryady, I. (2024). Sentiment Analysis of Customer Review Using Classification Algorithms and SMOTE for Handling Imbalanced Class. *Indonesian Journal of Information Systems*, *7*(1), 38-52.