Sentiment Analysis of Amazon Customer Reviews: An Ensemble Learning Approach with Data Augmentation

Abhinav Palanivel *Student, IB DP Canadian Internation School, Bangalore,*

Dr. Nandini N. *Associate Professor, Department of Computer Science and Engineering,*

Dr. Ambedkar Institute of Technology, Bengaluru, India.

---***---

Abstract - *This study examines the effectiveness of ensemble learning methods, specifically bagging and boosting, for data augmentation and sentiment analysis of Amazon product evaluations. Businesses have to evaluate the sentiment of online reviews in order to understand their customers' needs and develop their products. Sentiment analysis is difficult, though, due to the use of casual language and subtleties like slang and sarcasm. We investigate the comparative performance of ensemble learning models over individual models and investigate how resampling strategies can enhance the algorithms' capacity to handle unseen data points through data augmentation. Our results show that in the sentiment categorization of Amazon shoe reviews, Gradient Boosting performs better than other ensemble learning models such as Random Forest and Ada Boost. By demonstrating the value of ensemble techniques and data augmentation for processing informal language, this study advances sentiment analysis.*

Key Words: *Gradient Boosting, Random Forest, Ada Boost, Bagging, Boosting and Ensemble Model*

1. INTRODUCTION

Online reviews are now a priceless resource for both companies and customers. By the means of sharing consumer experiences and the provision of comprehensive product and service information, these evaluations assist prospective customers in making well-informed selections. Businesses can gain important insights into consumer happiness, product strengths and shortcomings, and overall brand impression by conducting sentiment analysis on internet reviews. Businesses can improve the way develop their products, target their marketing campaigns more effectively, and ultimately increase customer happiness by carefully examining these reviews.

Sentiment analysis of product evaluations on Amazon, however, poses a special difficulty. Due to the overwhelming amount of casual language and evaluations that users frequently use, typical machine learning models operate in a difficult context. Slang, irony, and conflicting feelings frequently mask the genuine sentiment expressed in a review, which can result in misunderstandings and imprecise sentiment analysis. Even though Gaussian Naive Bayes, Random Forest Classifiers, Support Vector Machines, and Logistic Regression are strong machine learning algorithms, they may find it difficult to accurately capture these subtleties in order to tackle these issues, this work examines how well ensemble learning methods — more especially, stacking — work when combined with data augmentation to analyze sentiment in Amazon product evaluations. Since Amazon product reviews are so common and have so much data available, we decided to concentrate on them. Our goal is to create a strong model that can reliably classify consumer sentiments and negotiate the complexity of informal language by analyzing sentiment in this big dataset. Our hypothesis is that ensemble learning can outperform individual models by leveraging the strengths of numerous models, so overcoming their limits. In addition, we investigate data augmentation using resampling methods to possibly strengthen the model's capacity to handle unknown data points and increase generalizability.

This paper makes several significant contributions to the field of sentiment analysis. We start by discussing the unique difficulties involved in sentiment analysis of Amazon product reviews. Secondly, we demonstrate the potency of ensemble learning combined with stacking as a strong solution to these problems. Third, we examine how data augmentation by resampling affects the model performance, offering important information about its possible advantages for this particular task. The rest of this paper will describe our all-inclusive methodology, show the outcomes of our trials, and evaluate the various models' performances. After that, we discuss our findings, examining how they might affect sentiment analysis of Amazon product evaluations and suggesting possible lines of inquiry for future study.

1.1 Literature review

Sentiment analysis approaches are being studied by an increasing number of studies to better comprehend customer attitudes in online product reviews. Xing Fang et al. in 2015 presented a general sentiment analysis procedure for the purpose of categorizing the sentiment of Amazon product reviews. Promising findings are obtained in this work that investigates sentiment polarity categorization at the sentence and review level [1]. Xeenia

Singla et al. in 2017 examined sentiment analysis as a method for categorizing favorable or negative online product reviews. They demonstrate the efficacy of machine learning for sentiment categorization by comparing the results of Naive Bayes, Support Vector Machines, and Decision Trees on a dataset of more than 4,000 reviews [2]. Karthiyayini. T et al. in 2017 introduced a new method calle**d** Senti, which uses the current natural language processing APIs to parse and project the comparative accuracy levels in order to analyze the sentiments of Amazon product evaluations, particularly the Meta dataset [3].

Chauhan et al. in 2017 investigates methods of summarizing product reviews using sentiment analysis. Their research demonstrates how feature-wise analysis may be used to produce unbiased summaries of customer sentiment from vast amounts of web reviews [4]. Rajkumar S. Jagdale et al. in 2018 used machine learning to analyze sentiment analysis of product evaluations on an Amazon dataset that included a variety of categories. In terms of camera reviews, their research shows that Naive Bayes has the best accuracy (98.17%), demonstrating the potency of machine learning in sentiment classification [5]. Ang Liu et al. in 2018 introduced a design framework for deriving customer 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 [6]. Rajesh Bose et al. in 2018, examined sentiment in more than 500,000 fine cuisine reviews on Amazon in 2018, in order to further understand customer behavior. Their research focuses on classifying emotions and pinpointing areas where product satisfaction might be raised by using sentiment lexicons and word clouds [7].

Wassan et al. in 2021 presented a sentiment analysis method that concentrated on the attributes of the products mentioned in online reviews. Through their efforts, marketers can better understand customer preferences by extracting sentiment at the aspect level from Amazon reviews [8]. Bickey Kumar Shah et al. in 2021 used machine learning methods (Logistic Regression, Naive Bayes, Random Forest), in order to categorize reviews as good, neutral, or negative. Based on their findings, Random Forest performs better in terms of sentiment classification accuracy than other techniques [9]. Arwa S. M. AlQahtani et al. in 2021 investigated sentiment analysis of Amazon product evaluations using a variety of text vectorization methods (Bag-of-Words, TF-IDF, GloVe) and machine learning algorithms (Logistic Regression, Random Forest, Naive Bayes, Bi-directional LSTM, BERT). Their research demonstrates how well deep learning techniques, such as BERT, function for online review sentiment classification [10].

1.2 Our Novelty

By exploring ensemble learning, in particular bagging and boosting, to improve robustness in capturing the subtleties of informal language seen in online reviews, this work adds uniqueness to sentiment analysis. Furthermore, we investigate how class imbalance problems frequently present in review datasets might be addressed by data augmentation using resampling approaches, which may result in sentiment analysis models that are more broadly applicable. This integrated strategy presents a novel way to improve sentiment analysis for unstructured text data.

2. METHODOLOGY

This section details the methods employed to analyze the customer reviews.

2.1 Dataset

Customer reviews for shoes sold on Amazon UK are available in the Data.world database. It contains information that was scraped from the amazon site, such as the product name, reviewer name, review content, rating, and timestamps with 6823 samples.

2.2 Exploratory Data Analysis

1) Data Preprocessing: Our first investigation of the data involved figuring out its properties and distribution. In all, 6823 samples were obtained. Fig. 1. illustrates how the distribution of review ratings, which range from one to five stars, revealed the user's opinion. In order to guarantee a clean dataset, we removed duplicate reviews and resolved missing values by adding up all of the null values in each review across the board. Stop words were removed to cut down on noise, and punctuation marks were removed to concentrate on the main idea.

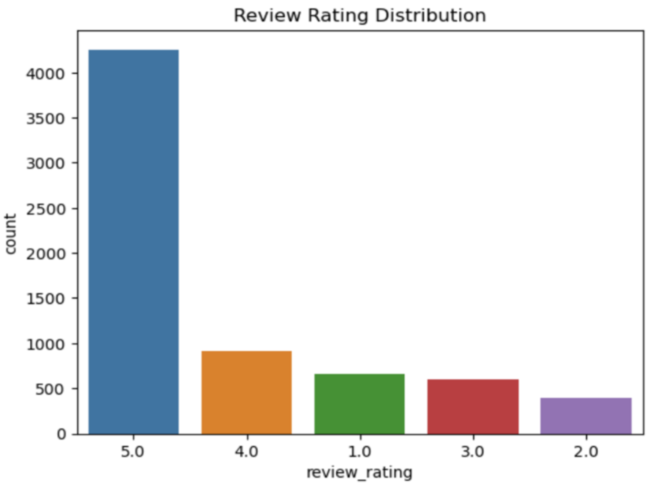

Fig -1: Distribution of the Review Rating

Furthermore, since lemmatization takes into account the part of speech and offers a more accurate and nuanced representation of the terms used in the reviews, we chose it over stemming. For the reviews, it is used to normalize words to their base form while maintaining their intended meaning. Lastly, we created a word cloud, as shown in Fig. 2. to get a general idea of the terminology utilized.

2) Sentiment Analysis: Sentiment analysis explores the textual data's emotional undertone. There are two main components to it: polarity and subjectivity. If a statement conveys sentiments, views, or ideas that are not supported by facts, it is considered subjective. Conversely, polarity expresses the text's general attitude, which might be neutral, positive, or negative.

comfortable
shoe
great
size
love
bought
perfect
pair
super
wear
feet
sneaker
boot
fit
look
color
good
really
foot
time
und
de
running
walking
comfortable shoe
running shoe

Fig -2: Word Cloud of the Reviews

These are two closely related aspects. Subjective statements are generally more prone to positivity or negativity, whereas objective assertions, or facts, are usually impartial. Sentiment analysis offers a more thorough understanding of the emotional undercurrents in a text by examining both subjectivity and polarity.

We focus on the compound score, which is a single floating-point value ranging from -1 (most negative) to +1 (most positive). This score reflects the overall sentiment of the text. The polarity and subjectivity categories are used in this paper as per table I and II respectfully. We have visualized the density of the polarity and subjectivity for the reviews as in Fig. 3. and the variation of polarity over subjectivity for the dataset as in Fig. 4.

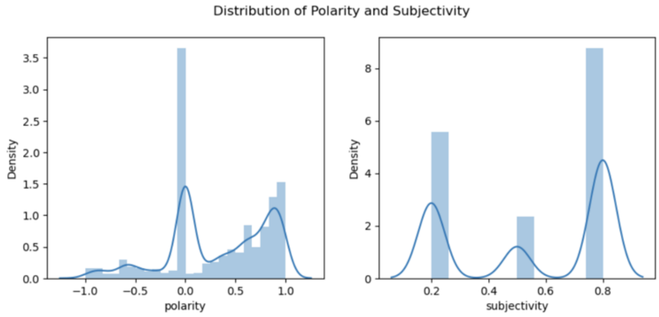

Fig -3: Distribution of the Review Rating

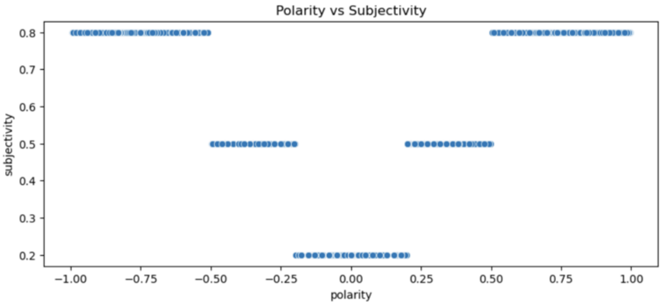

Fig -4: Distribution of the Review Rating

The lexical landscape of the text data is revealed by visualizing word frequencies and n-grams. Finding common keywords and maybe stopping words can be achieved by looking at the most used words.

Table -1: POLARITY CATEGORIZATION

Compound Score Range	Polarity
Score = 0	Neutral
0 < Score ≤ 0.3	Weakly Positive
0.3 < Score ≤ 0.6	Positive
Score > 0.6	Strongly Positive
0 > Score ≥ -0.3	Weakly Negative
-0.3 > Score ≥ -0.6	Negative
Score < -0.6	Strongly Negative

Table -2: SUBJECTIVITY ESTIMATION

Absolute Compound Score Range	Subjectivity Score
0 ≤ abs(Score) ≤ 0.2	Low (0.2)
0.2 < abs(Score) ≤ 0.5	Moderate (0.5)
abs(Score) > 0.5	High (0.8)

Words with the lowest frequency may contain mistakes or jargons exclusive to a certain domain. Unigram analysis establishes the framework, while bigram and trigram analyses provide word combinations that reveal phrases, collocations, or even colloquial idioms. In this paper the most frequent words, least frequent words, unigrams, bigrams and trigrams are visualized as in Fig. 5, 6, 7, 8 and 9 respectively.

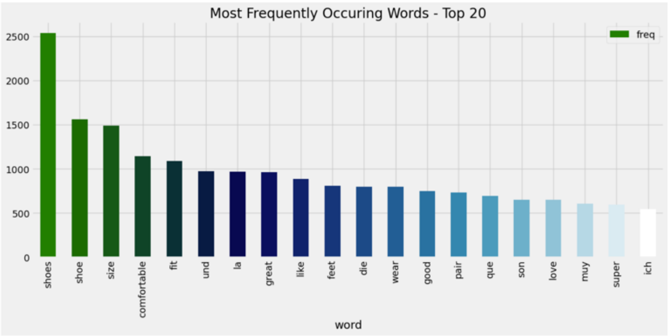

Fig -5: Most Frequently Occurring Words

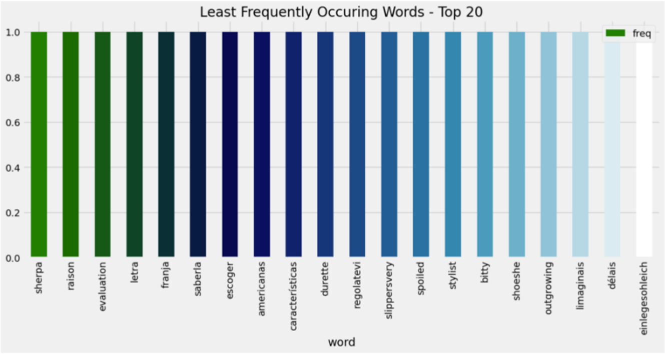

Fig -6: Least Frequently Occurring Words

Also, based on a user-supplied search phrase, sentiment analysis is performed on a set of reviews, and the findings are shown in both a general sentiment category and a comprehensive percentage breakdown for each sentiment type; it is visualized in Fig. 10.

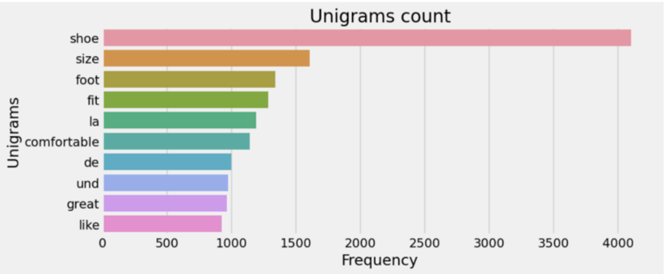

Fig -7: Top 20 Unigrams

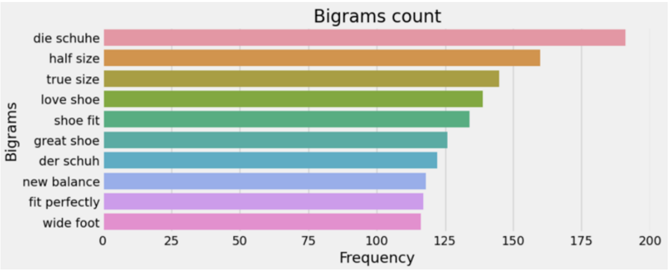

Fig -8: Top 20 Bigrams

3) Vector Embeddings: Each review text is vectorized, or converted into a vector of integers, with each location potentially denoting a word or textual feature that has been taken out. Methods such as TF-IDF (Term Frequency-Inverse Document Frequency) take into account a word's overall value across all reviews in addition to its frequency inside a review. Machine learning algorithms can determine the similarities or distances between reviews once they are stored as vectors and estimates reviews with similar sentiments.

4) Data Augmentation: This technique in case of imbalance in the classes. Here, we experimented with two methods namely, SMOTE (Synthetic Minority Oversampling Technique) and NearMiss (Under-sampling Technique) to address this problem and chose the best performing method to proceed further. The former effectively increases its representation in the training set and reduces bias, thus improves the model's learning of the minority class while the latter brings the size of the majority class closer to that of the minority class by removing data points from it. By doing this, the model is not overloaded with the majority class and is free to concentrate on absorbing information from the less frequent class as well. The observations are mentioned in the table III.

5) Models: With its state-of-the-art performance across a range of tasks, ensemble methods have emerged as a key component of machine learning. Three well-known ensemble methods are given in this section: Ada Boost, Gradient Boosting, and Random Forests.

Table -3

LABEL CLASS DISTRIBUTION			
Class	Original Dataset	Over-Sampled	Under-Sampled
0	4255	4255	2568
1	2568	504	504

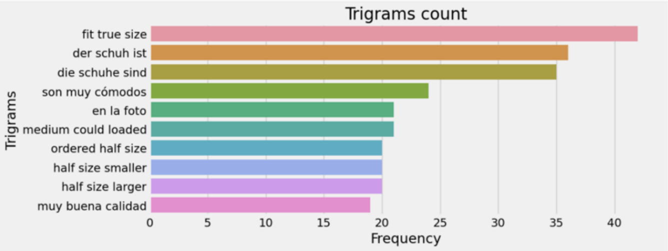

Fig -9: Top 20 Trigrams

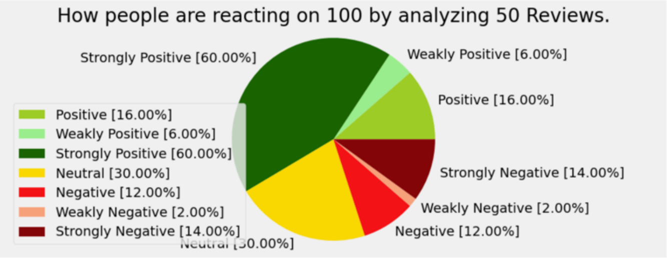

Fig -10: Polarity for Specific Number of Reviews

Decision Trees are supervised learning models that are nonparametric and generate predictions by repeatedly dividing the data according to attributes. The model chooses the feature and threshold that best divides the data into the target classes at each split. With each node representing a decision rule and the leaves representing the final classifications, this procedure creates a structure like a tree. Decision Trees are effective, but they can overfit, especially when working with high-dimensional data. Random Forests use bagging, or bootstrap aggregating, to solve the overfitting problem with Decision Trees. Several Decision Trees are trained using this ensemble technique on arbitrary subsets of the data (with replacement). Every tree learns a classification model on its own, and the final prediction is produced by averaging the individual forecasts (majority vote for classification, regression by averaging, etc.). Compared to single Decision Trees, Random Forests achieve better generalization and lower variance by utilizing the diversity of these independently trained trees.

Another effective ensemble technique that makes use of a boosting strategy is Gradient Boosting. In contrast to bagging, which trains trees separately, boosting trains models in a sequential manner. The goal of each new model in the ensemble is to fix the mistakes produced by the ones before it. To do this, data points that the previous model misclassified are given larger weights, which forces the new model to focus more on those occurrences. They have the ability to handle complex datasets and provide better results. During the ensemble building process, a particular kind of boosting technique called Ada Boost dynamically modifies the weights of the data points. It aims to improve the categorization of previously misclassified instances, much to Gradient Boosting.

On the other hand, Ada Boost does this by direct manipulation of data point weights. Higher weights are applied to data points in the ensemble that have been repeatedly misclassified by earlier models, ensuring that learning from these difficult cases is given priority in later models. Ada Boost is especially good at managing complicated and noisy datasets.

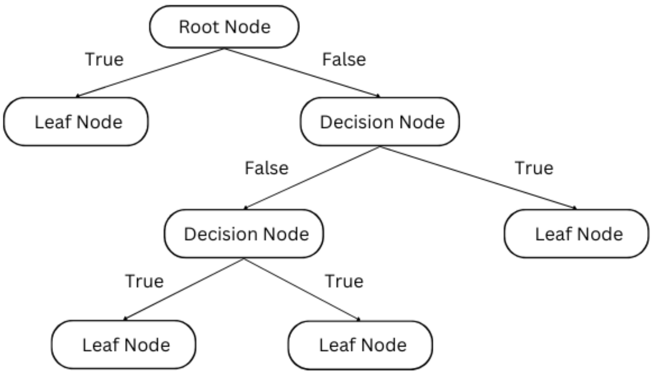

Fig -11: Workflow Diagram

2.3 Proposed Approach

In the initial stage, the dataset with 6823 samples is pre-processed by removing stop-words and punctuation, dropping out the duplicate values, and handling the missing values. Vader Sentiment Lexicon uses sentiment intensity analyzer to analyze the sentiment of the reviews with polarity and subjectivity. Feature Engineering is performed by creating new features to improve the model performance like character count, word count, sentence length, polarity, subjectivity and word density for each of the 6823 customer reviews as shown in Fig. 11. Also to avoid overfitting Synthetic Minority Oversampling Technique (SMOTE) is utilized to resample the data and equalize the sentiment distribution. In this paper, Decision Tree Classifier, Random Forest Classifier, Ada Boost Classifier and Gradient Boost Classifier machine learning models are used to train the resampled data and get the desired outcomes.

All four models selected hyperparameters place equal emphasis on avoiding overfitting and striking a balance between model complexity. To ensure reproducibility throughout training, they are all based on a random state, which is set to either 0 or 42. The default configurations are used by Decision Tree. To prevent overfitting and limit individual tree complexity, Random Forest uses 200 trees (growing in complexity) at a maximum depth of 10. Both Ada Boost and Gradient Boosting use 200 trees and a 0.1 learning rate. Their ability to understand intricate decision limits is facilitated by this setup, and the smaller learning rate helps avoid huge updates during training, which may lessen overfitting.

2.4 Results and Discussions

We used stratified train-test splitting to divide the preprocessed data (X resampled, Y resampled) into training and testing sets in order to assess the performance of our models. By preserving the class distribution in both sets, this method guarantees a more reliable evaluation. To ensure reproducibility, we employed a random state of 42 and a test size of 20% (1702 samples). As a result, the following shapes were included in the training and testing sets:

- Training set: X train (6808 samples, 2500 features), Y train (6808 labels)

- Testing set: X test (1702 samples, 2500 features), Y test (1702 labels)

This clarifies the data splitting process and provides details about the resulting set sizes and characteristics.

Table -4

CLASSIFICATION REPORT FOR DECISION TREE CLASSIFIER MODEL				
Class	Precision	Recall	F1-Score	Support
0	0.63	0.91	0.74	891
1	0.81	0.41	0.55	811
Accuracy	0.67			1702
Macro Avg	0.72	0.66	0.65	1702
Weighted Avg	0.71	0.67	0.65	1702

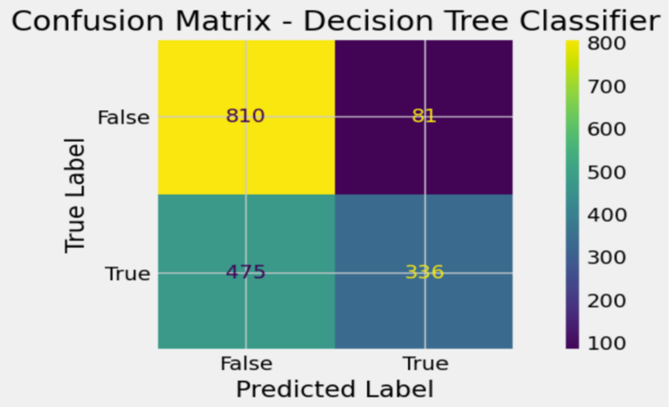

Fig -12: Confusion Matrix of Decision Tree Classifier

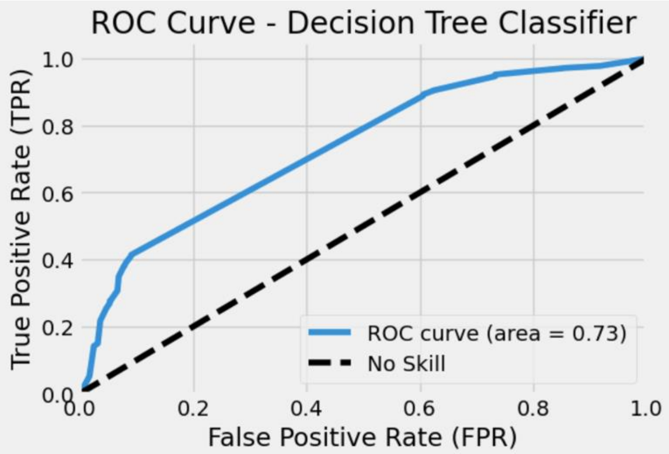

Fig -13: ROC-AUC Curve of Decision Tree Classifier

Precision is the percentage of anticipated positive cases that were actually positive, which is represented by this metric. Recall shows how well the model recognizes real positive cases. The F1-score provides a fair assessment of recall and precision. The precision, recall and F1-score are shown in Table IV, V, VI and VII. This shows that the Gradient Boost Classifier has a good nature of recognizing the positive cases than the other two models.

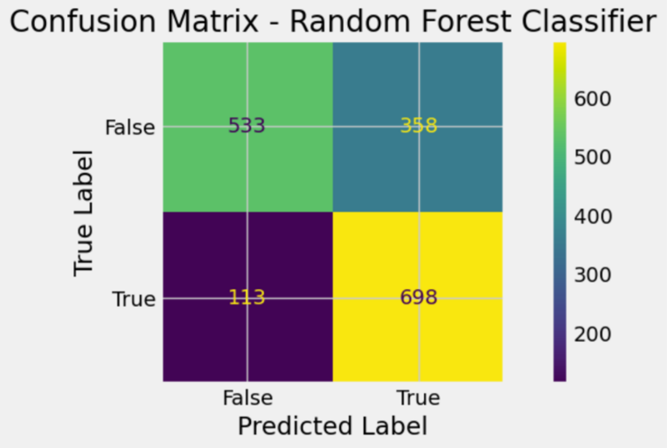

Fig -14: Confusion Matrix of Random Forest Classifier

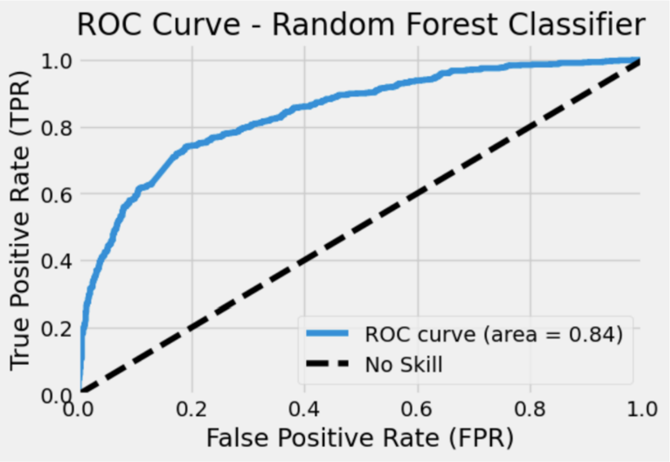

Fig -15: ROC-AUC Curve of Random Forest Classifier

Table -5

CLASSIFICATION REPORT FOR RANDOM FOREST CLASSIFIER MODEL				
Class	Precision	Recall	F1-Score	Support
0	0.83	0.60	0.69	891
1	0.66	0.86	0.75	811
Accuracy	0.73			1702
Macro Avg	0.74	0.73	0.82	1702
Weighted Avg	0.75	0.72	0.72	1702

Table -6

CLASSIFICATION REPORT FOR GRADIENT BOOST CLASSIFIER MODEL				
Class	Precision	Recall	F1-Score	Support
0	0.83	0.82	0.83	891
1	0.81	0.81	0.81	811
Accuracy	0.82			1702
Macro Avg	0.82	0.82	0.82	1702
Weighted Avg	0.82	0.82	0.82	1702

The testing accuracy is 67%, 72%, 82% and 78% for Decision Tree Classifier, Random Forest Classifier, Gradient Boost Classifier and Ada Boost Classifier respectively as shown in Table IV, VI and V. The confusion matrix is described in Fig. 12, 14, 16 and 18 for the Decision Tree Classifier, Random Forest Classifier, Gradient Boost Classifier and Ada Boost Classifier respectively. This tell us that the Decision Tree Classifier has (810 + 336 = 1146) correctly predicted values, the Random Forest Classifier has (533 + 698 = 1231) correctly predicted values, the Gradient Boost Classifier has (735 + 659 = 1394) correctly predicted values and the Ada Boost Classifier has (720 + 607 = 1327) correctly predicted values. This shows Gradient Boost outperforms the other models.

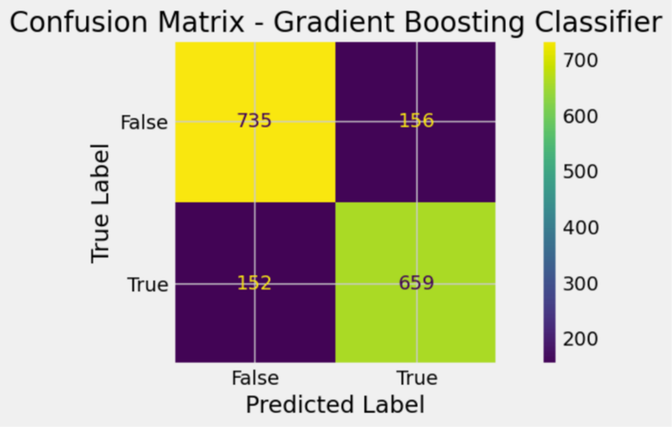

Fig. 16. Confusion Matrix of Gradient Boost Classifier

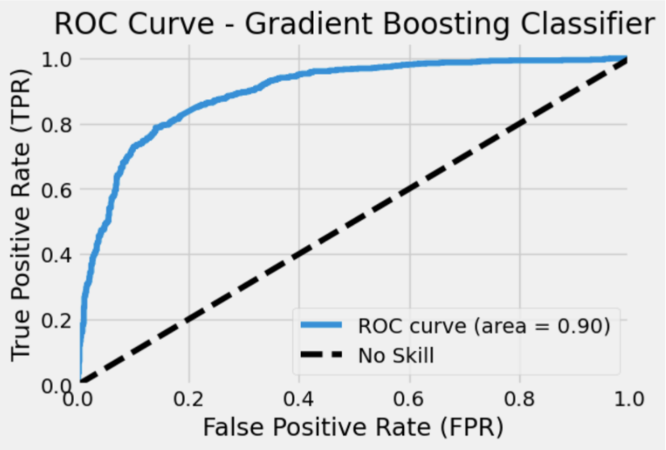

Fig -17: ROC-AUC Curve of Gradient Boost Classifier

When assessing binary classification models, one popular performance indicator is the Receiver Operating Characteristic (ROC) AUC curve. The y-axis displays the True Positive Rate (TPR), while the x-axis displays the False Positive Rate (FPR). The classification model's overall performance is shown by the AUC (Area Under the Curve). An AUC value of 0.90 for Gradient Boost Classifier shows that this model outperforms the Decision Tree Classifier, Random Forest Classifier and Ada Boost Classifier models with 0.73, 0.84 and 0.85 AUC values respectively as shown in Fig 13, 15, 17 and 19.

Thus overall, the ensembling of the Decision Tree Classifier where the residual error of the model is fed to the second model in the sequence with a depth of 10 and with 200.

Table -7

CLASSIFICATION REPORT FOR ADA BOOST CLASSIFIER MODEL				
Class	Precision	Recall	F1-Score	Support
0	0.78	0.81	0.79	891
1	0.78	0.75	0.76	811
Accuracy	0.78			1702
Macro Avg	0.78	0.78	0.78	1702
Weighted Avg	0.78	0.78	0.78	1702

Estimators and a learning rate of 0.1, also can be described as Gradient Boost Classifier outperforms the other ensembling techniques of the Decision Tree Classifier, namely the Random Forest Classifier where a majority vote from each of the individual Decision Tree model is taken or the Ada Boost model where the weighted majority vote from each of the individual Decision Tree model is considered.

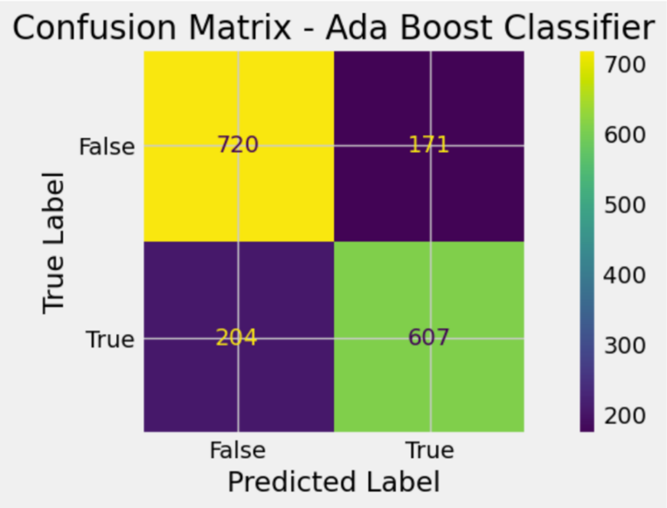

Fig -18: Confusion Matrix of Ada Boost Classifier

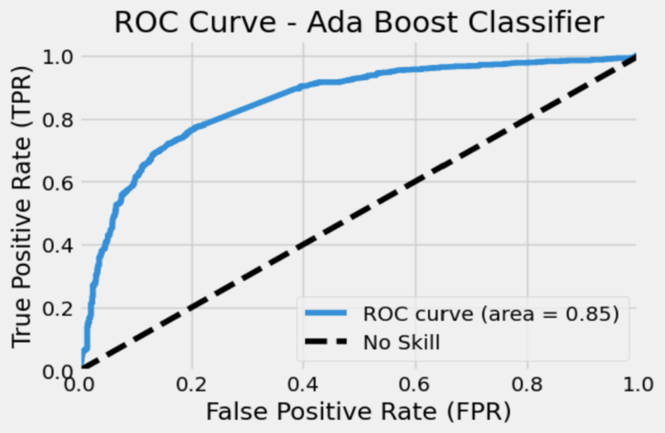

Fig -19: ROC-AUC Curve of Ada Boost Classifier

3. CONCLUSION

We performed exploratory data analysis and identified the sentiment of the reviews with polarity and subjectivity. For our binary classification job, Gradient Boosting proved to be the best ensemble classification model. It produced the highest F1-score and testing accuracy due to its balanced performance, which included good recall and precision for affirmative cases. This was validated by the confusion matrix and the ROC AUC curve, which showed how well the model distinguished between the classes and could accurately detect both positive and negative cases. Even though Gradient Boosting showed encouraging results, more research might focus on hyper-parameter tuning to enhance the system even more and investigate different ensemble techniques or deep learning approaches to find areas where performance could be improved.

REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[6] 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.

[7] 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.

[8] 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.

[9] B. K. Shah, A. K. Jaiswal, A. Shroff, A. K. Dixit, O. N. Kushwaha, and N. K. Shah, “Sentiments detection for amazon product review,” in 2021 International conference on computer communication and informatics (ICCCI). IEEE, 2021, pp. 1–6.

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