Sentiment Analysis on Amazon Reviews

September 6, 2023

Abstract

Reviews can significantly impact a company's reputation in the market, potentially influencing its overall business outcomes, either positively or negatively. This is especially crucial for companies that operate primarily through e-commerce platforms. Hence, it is vital for companies to pay close attention to customer reviews.Sentiment Analysis, often referred to as "opinion mining," is a significant procedure in Natural Language Processing (NLP) which serves the purpose of ascertaining the emotional tone of a provided text and categorizing it into positive, negative, or neutral perspectives.In this paper, sentiment analysis methodology is presented for classifying amazon reviews which utilizes a large dataset of reviews and employs Multinomial Naive Bayesian (MNB), Support Vector Machine (SVM), Maximum Entropy (ME), and Logistic Regression as the primary classifiers by the authors. With the aid of machine learning, we employed a supervised learning approach to an extensive Amazon dataset in order to categorize it based on sentiment polarity, achieving a high level of accuracy for the results.Here, we utilized the Kaggle dataset that includes a substantial volume of reviews and associated metadata which comprises customer review and ratings on amazon products.

Index terms: Sentiment Analysis, Machine learning, Natural Language Processing, Naive Bayesian (MNB), Support Vector Machine (SVM), Maximum Entropy (ME), Logistic Regression,feature extraction, text classification.

1 Introduction

The long-term viability of businesses like amazon is heavily contingent on their capacity to effectively fulfill customer requirements.Countless individuals share their opinions about various services or prod- ucts through different platforms like social networking sites, blogs, or popular review sites that will comes up just with a google search. So,in recent times it has become a common practice to search for reviews before making a purchase decision.Hence, disseminating customer reviews and feedback about online products or services can significantly impact the perceptions of new customers regarding the organization. Additionally, Amazon can examine these reviews to determine their authenticity and assess whether they might be part of a competitor's scheme to manipulate perceptions.Also, con- ducting a thorough analysis of customer sentiments empowers Amazon's business enthusiasts to gain deeper insights into the market, enabling them to make informed decisions that preemptively address customer needs and concerns, like betterment or elimination of any existing product, increasing or decreasing monetary value for marketing reach etc. Sentiment analysis leverages natural language pro- cessing techniques and text analysis to delve into what customers are conveying, how they articulate their thoughts, and the underlying meaning behind their expressions.The primary objectives of this paper is to extract the sentiments conveyed in customer reviews and analyze these senti-

various classifiers were employed iteratively we utilized the labeled datasets to conduct further processing while the extracted features were classified through different classifiers.We utilized a combination of two feature extraction methods: the bag of words approach and the tf-idf and Chi-square approach to enhance the accuracy of our results

2 Background

Businesses today, particularly those in the e-commerce industry like Amazon, confront a significant challenge: understanding and using the power of consumer reviews. A company's reputation and financial performance may be made or broken by customer feedback, which is frequently published on a variety of internet platforms. Reviews are now frequently consulted and relied upon by customers before making judgments about purchases, making them an essential component of the modern marketplace. However, because this enormous amount of customer feedback is made up primarily of unstructured data, it can be difficult for firms to gather insightful information. Here comes the role of sentiment analysis, a division of Natural Language Processing (NLP). Sentiment analysis, sometimes known as "opinion mining," is an effective approach for identifying the emotional undertone of text and classifying it as neutral, positive, or negative. In our study "Sentiment Analysis on Amazon Reviews," we set out to explore the potential of sentiment analysis within the context of Amazon's enormous review ecosystem. We make use of a sizable dataset from Kaggle, a community for data science fans that includes a wide range of reviews and related metadata. Our objective is straightforward: to use computational linguistics and machine learning methods, such as Naive Bayes, Multinomial Naive Bayes, Support Vector Machine (SVM), and Decision Tree, as major classifiers to categorize these reviews based on sentiment. We seek to categorize a sizable number of reviews based on sentiment polarity by utilizing supervised learning approaches and sophisticated feature extraction techniques like the bag of words approach and TF- IDF. The goal of this research is to give organizations the resources they require to get a thorough understanding of client attitudes, empowering them to make defensible choices that satisfy customer demands, enhance current goods or services, and improve marketing tactics. In conclusion, sentiment analysis is a powerful tool that, by revealing the hidden sentiments inside customer reviews, can help firms flourish in today's cutthroat business environment. Our study goes into the complexities of sentiment analysis and offers businesses a road map for efficiently utilizing client feedback. We hope to enable businesses like Amazon to make data-driven decisions that improve customer satisfaction and propel success in the digital era through this investigation.

3 Related work

Many research papers regarding opinion mining or sentiment analysis have already been explored by many researchers. For identifying interests of the customers about products, sentiment analysis played an important role at the documentation level. The authors in the work [3], developed a business model by analyzing the result of the sentiments they got from the reviews on amazon. Also, they focused on detecting the emotions of the customers from the reviews and fake reviews. Besides, they worked on finding the gender of the customers based on the name given. Multinomial Naive Bayesian(MNB) and Support Vector Machine(SVM) were used as the dominant classifiers in their work. From the paper [4], the author enlisted the help of a supervised machine learning algorithm for foreseeing the textual based review ratings on a given numerical scale. For the work, 70 percent of data were used as training data and the rest 30 percent as testing data for cross validation. Applying different classifiers as their algorithm, the author was determined to get accurate results of the values. The author in paper [5], choose reviews from the section of books and 2 kindle on amazon to perform sentiment analysis for research. Mainly, naive bayesian(NB) and decision list classifiers were applied for determining whether to label the reviews given as positive or negative. For paper [6], the author applied Naive Bayesian, Support Vector Machine and Maximum Entropy as the main classifiers that created a system which portrayed the sentiment of the reviews in the chart. By using data scraping from amazon url, they preprocessed the data gotten from it. In the paper, they did not show the result of accuracy as they only gave a summary of the product review. As a result, they depicted the result in the form of a statistical chart. In paper [7], the author used simple algorithms as they got high accuracy on support vector machine. But those algorithms were incapable of performing well on a large number of datasets. So, they applied a simple algorithm of support vector machine, logistic regression and decision tree method. The author in paper [8], used tf-idf to speculate the given ratings from a bag of words. Besides, they did not use many classifiers. So, with the help of root mean square error and linear regression model they got the result. Authors in paper [9], designed a model for the prediction of the ratings of the product dependent on a bag of words from rating text for testing used unigrams and bigrams. Upon testing, unigrams gave more precise results than bigrams. The percentage of unigrams was 15.89 percent better than the result of bigrams. In the paper [10], the authors used different types of selection techniques for doing research on sentiment analysis. Because of stop words and removing special characters, they performed preprocessing after gathering data from Amazon. Besides , they applied naive bayes as the main classifier for phrase level, single word and multiword feature selection. As a result, they summed up that the result of phrase level was better than of single word and multiword with the help of Naive Bayes. Moreover, they only applied the algorithm of Naive Bayes as their classifiers. So, by using only Naive Bayes , we would not be able to get a proper result. As a result, we decided to use ideas from the above mentioned works which would give us higher accuracy in our

research. We were able to make better decisions by making the best use of ideas from other related works. So, we were able to get enhanced results of our tasks for applying an active learning approach. As a result , the accuracy of our result was more accurate than other mentioned works.

4 Literature Review

With the help of machine learning algorithms, classification of textual data for sentiment analysis is done easily for the detection of positive and negative reviews about the products on amazon. By using machine learning algorithms, we concentrated more on the preciseness of the classification of the reviews. So, we extracted data from the reviews on amazon site and created a business model by surveying the result of the reviews. Based on the data, we detected positive and negative reviews of the customer by using Multinomial Naive Bayesian(MNB), Support Vector Machine(SVM), Maximum Entropy(ME) and Logistic Regression as the main classifier. By adapting a supervised machine learning algorithm, we predicted the ratings of textual reviews on a given numerical scale. In the paper we used natural language processing for classifying the textual reviews. To label the reviews given by the customers whether it is positive or negative, Naive Bayesian and Decision list is used as classifiers. The main objective is to create a system that portrays the sentiment of the reviews in the chart. Maximum Entropy came under the category of probabilistic classifier and it did not take into account the elements independent of each other. Also added that Maximum Entropy solved a huge amount of text classification problems. So, it is very popular in the field of sentiment analysis. Naive Bayesian constructed classifiers that assigned a class label to the problem examples. Depending on the value of the features being defined, Naive Bayesian expressed a vector form where labels came from the finite sets.Despite not being a standard algorithm, Naive Bayesian is dependent on the principles being used. For classification problems, another algorithm called Support Vector Machine is used widely. Besides, they are popular for being especially effective at separating data points that belong to various classes using the ideal hyperplane, making them particularly well-suited for classifi- cation issues. They distinguished the classes very effectively. On the other hand Logistic Regression known as the statistical approach and machine learning algorithm on the idea of probability is used to solve classification problems. It is applicable if the dependent variable is categorical. To determine whether the output of the reviews are positive, neutral or negative, we used logistic regression. As a result to get more accurate results for our research paper, we used logistic regression, Support Vector Machine, Naive Bayesian and Maximum Entropy for the prediction of amazon reviews.

5 Methodology

5.1 Collected Data

In our effort to conduct sentiment analysis on Amazon reviews, we made use of a carefully curated dataset obtained from Kaggle, a well-known community for data science enthusiasts. We used the Kaggle dataset for our investigation, which contains a sizable number of reviews and related metadata, to build a strong sentiment analysis framework. 21 columns and 34,660 rows make up this dataset, which depicts various facets of the review data in each row and column. We concentrated on using 10 essential features from this large dataset for our particular investigation.Our sentiment analysis targets extracting valuable insights from the Kaggle dataset, illuminating the sentiments expressed within Amazon reviews by focusing on these 10 carefully selected features. This group of features provides a targeted and effective way to analyse consumer sentiment on Amazon and discover the variables affecting how they feel about specific products.

5.2 Proposed Methodology

One of Amazon's most valuable and established features are customer reviews. For the purpose of determining the polarity of a review, we have used the reviews.text and reviews.rating features from our dataset.We have preprocessed our data by removing null values from our data and replaced them with spaces, as well as any extraneous features, since If they are not handled properly, missing values could introduce bias and errors into the analysis. Missing values have the potential to skew statistical metrics like means, medians, and correlations. There are a finite number of categorical variables, most of which take the form of "strings" or "categories."Data preparation is made more challenging by categorical values' lack of numeric values.Algorithms can find relationships and patterns in the data by converting category values into numerical representations, which are widely used to describe qualitative aspects.. Additionally, categorical variables like reviews.text are included in our project. The code in our project converts the "reviews.text" column from a string to an integer type. CountVectorizer and TfidfTransformer are used to convert the text data (reviews) into numerical characteristics that machine learning models can use. The dataset was separated into training and training dataset 70 and 30 percent, respectively, and scrambled at random. The accuracy was then assessed using machine learning approaches such as SVM, logistic regression, Decision tree, naive bayes, and random forest algorithm.

5.3 Data Analysis

Data analysis is a vital phase in any machine learning project, especially one focusing on sentiment analysis from Amazon reviews, where data quality and preprocessing play a critical role.In this por- tion of the paper, we present a complete examination of the dataset utilised in this work, including its char-

acteristics, preprocessing steps,model training and testing and come up with strategies to attain our goal. Two significant concerns such as handling missing values and handling categorical variables have stand out in our data analysis process and have ramifications on the quality and dependability of our conclusions.Missing values within our dataset, particularly in the "reviews.rating" column, provide a considerable issue. These null values can bring biases and mistakes into our analysis, influencing statistical measurements like means and correlations. Moreover, during model training and testing, these missing variables might lead to errors, hurting the performance of the machine learning models. To overcome this issue, we apply approaches like deleting null entries or imputing them with accept- able values, ensuring the integrity of our dataset.Categorical variables, such as the "reviews.text," are non-numeric in nature, complicating data preparation. However, translating these category values into numerical representations is necessary for machine learning algorithms to recognise patterns and relationships effectively. We address this difficulty by applying text vectorization algorithms, including CountVectorizer and TfidfTransformer, which turn textual data into numerical representations suited for analysis. Moving beyond data preprocessing, we tested the performance of various machine learn- ing models. Naive Bayes, Logistic Regression, Support Vector Machine (SVM), and Decision Tree classifiers were applied for model training and testing. Our results reveal that the SVM and logistic regression models regularly outperform others, obtaining a remarkable accuracy rate of roughly 93 percent. These models have proved their effectiveness in capturing sentiment trends among Amazon 5 reviews. In summary, our data analysis highlights the significance of robust data pretreatment to handle missing values and categorical categories effectively. Furthermore, our model evaluation reveals both SVM and logistic regression classifiers are well-suited for our sentiment analysis project, contin- uously generating excellent accuracy rates. These findings give a solid platform for making informed decisions based on Amazon reviews and deriving important insights from customer sentiment.

6 Prototype Implementation

We incorporated a variety of machine learning methods for sentiment analysis into our prototype implementation. Naive Bayes: One of the essential machine learning methods that helps with classification problems is the Naive Bayes algorithm. When training high-dimensional datasets for text classification, it is developed from Bayes' probability theory. We started with the data splitting, partitioning our dataset into training and testing sets at a 90percent to 10 percent ratio. Then, in order to turn textual input into numerical features, we used Count Vectorization. We then used the vectorized training data to train a Multinomial Naive Bayes classifier, and we used this model to predict the test data. In addition to accuracy calculations, we produced a thorough classification report and used a confusion matrix to display performance. Support Vector Machine (SVM): Support Vector Machine SVN is utilized for regression

and clas- sification. It finds a hyperlane that maximizes the margin between the data while simultaneously effectively classifying the data into different groups. For this method's initial step, we used the scikit- learn module to build a pipeline for text classification. This pipeline included SVM classification, TF-IDF transformation, and Bag-of-Words transformation. We used this pipeline to make predictions on the test data after training it on the training set of data. Finally, we calculated and displayed the SVM classifier's prediction accuracy. Decision Tree: Decision Tree is a supervised learning method that can be applied to problems related to regression and classification. It is a tree-structured classifier, where internal nodes stand in for the dataset's features, branches for the rules of classification, and each leaf node for the result. We used tools like scikit-learn in our implementation and built a decision tree classifier. This classifier was created to make predictions using features extracted from the training set of data. Recursively dividing the data into subsets based on particular feature criteria allowed the decision tree to learn to make decisions, eventually forming a tree-like structure. For our analysis, the tree's depth and structure were set up as needed. The decision tree classifier was used to predict sentiment labels for the test data after it had been trained. Multinomial Naive Bayes: For text classification issues in Natural Language Processing (NLP), Multinomial Naive Bayes (MNB) is a popular machine learning approach. For issues involving text data with discrete properties, such as word frequency counts, it is especially helpful. We started this method by dividing the dataset into training and testing subgroups. The text was then converted into numerical features using Count Vectorization. We trained a Multinomial Naive Bayes classifier using this vectorized training data, which is effective at handling discrete data like word counts. The test data's sentiment labels were then predicted using the trained Multinomial Naive Bayes model. In order to examine the performance of the model, we also prepared a classification report with precision and recall metrics and evaluated the accuracy of these predictions. This prototype implementation highlights how we apply a variety of machine-learning approaches to conduct sentiment analysis, and it demonstrates how the code may be modified to match particular data and project requirements.

7 Result Analysis

We used Tf-idf and countVectorization for counting polarity and relative frequency of a word.Text preprocessing, svm classifier, Multinomial NB, Decision Tree, Logistic Regression, Natural Language Matplotlib, and Seaborn are the ideas and models that are used in this article. All of these techniques have been combined in a novel way in our research to examine consumer re-perspectives where we evaluate how accurate the models we used are. We identify the review's attitude as either positive, neutral, or negative based on the comparison of the applied models' accuracy levels. Our results show that the research's outcomes are more pleasing.With the help of three matrices—precision, recall, and f1 score—we develop a classification report to evaluate the model's quality.1.

The percentage of correctly made positive predictions compared to all positive predictions is known as precision. The percentage of correctly predicted positive outcomes compared to all actual positive outcomes is known as recall. The precision and recall components of the F1 Score are weighted harmonic means. Better models are those that are closest to 1. Score for F1: 2* (Precision * Recall) / (Precision + Recall) By evaluating a classification model's performance using these three criteria, we may determine how effectively it can forecast results for a certain response variable.

Confusions matrix and classification result for different types of model are given:

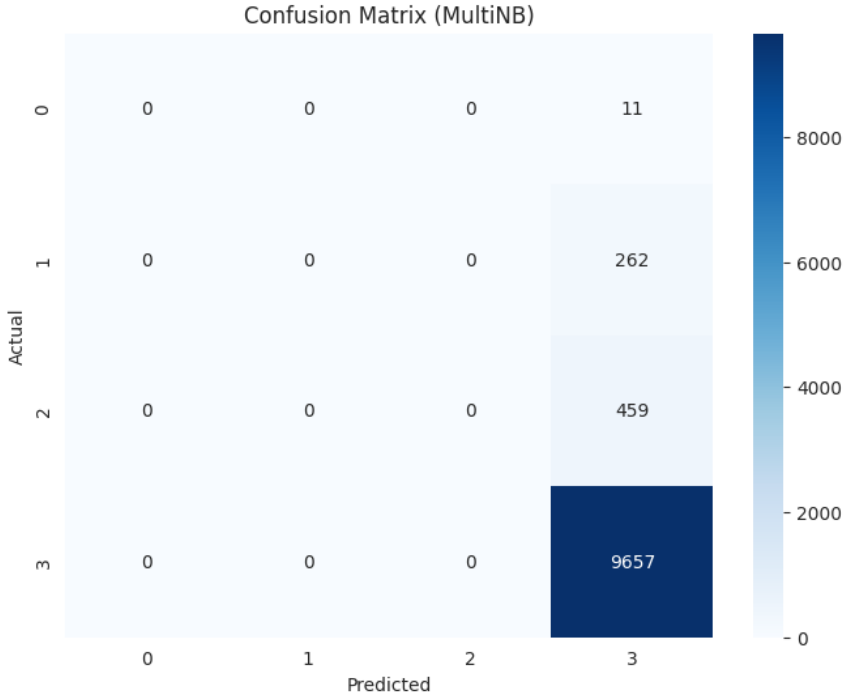

Figure 1: multinomailNB confusion matrix.

The naïve Bayes theorem states that the accuracy rate is 93 percent. In comparison, the accuracy rate for SVM is almost identical to that of logistic regression at 93 percent. The decision tree classifier also has a 90 percent accuracy rate. Therefore, we can say that MultinomialNB, SVM, and logistic regression models are the best for our project, whereas decision tree is the worst.

Classification Report (MultiNB):

	precision	recall	f1-score	support
	0.00	0.00	0.00	11
Negative	0.00	0.00	0.00	262
Neutral	0.00	0.00	0.00	459
Positive	0.93	1.00	0.96	9657
accuracy			0.93	10389
macro avg	0.23	0.25	0.24	10389
weighted avg	0.86	0.93	0.90	10389

Figure 2: MultinomialNB classification report.

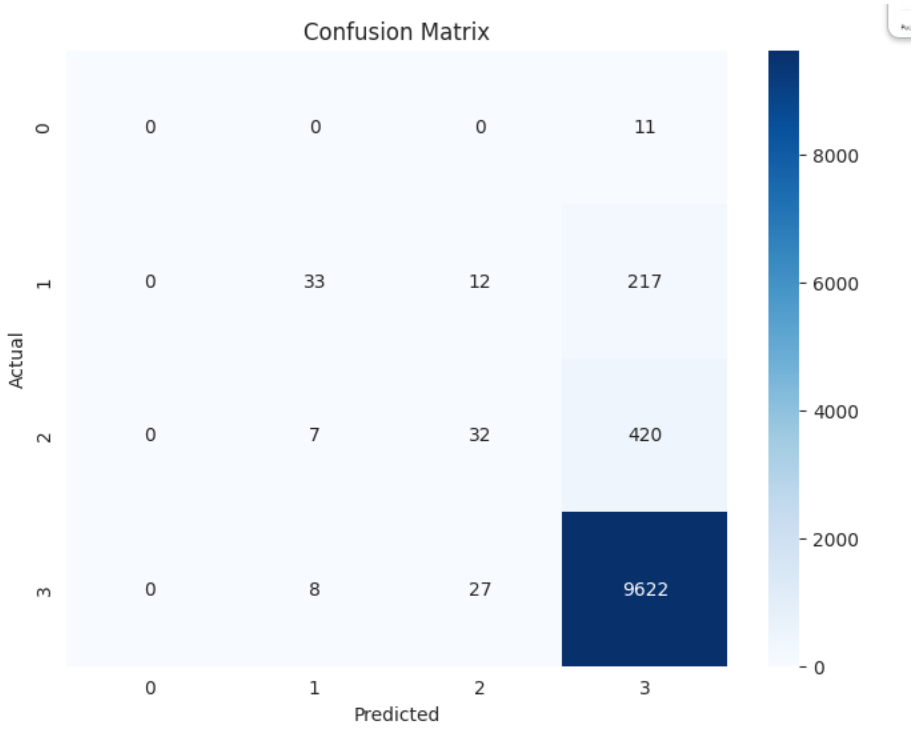

Figure 3: Logistic confusion matrix.

Classification Report (LinearSVC):

	precision	recall	f1-score	support
	0.00	0.00	0.00	11
Negative	0.61	0.21	0.31	262
Neutral	0.45	0.11	0.18	459
Positive	0.94	0.99	0.97	9657
accuracy			0.93	10389
macro avg	0.50	0.33	0.36	10389
weighted avg	0.91	0.93	0.91	10389

Figure 4: Logistic classification report.

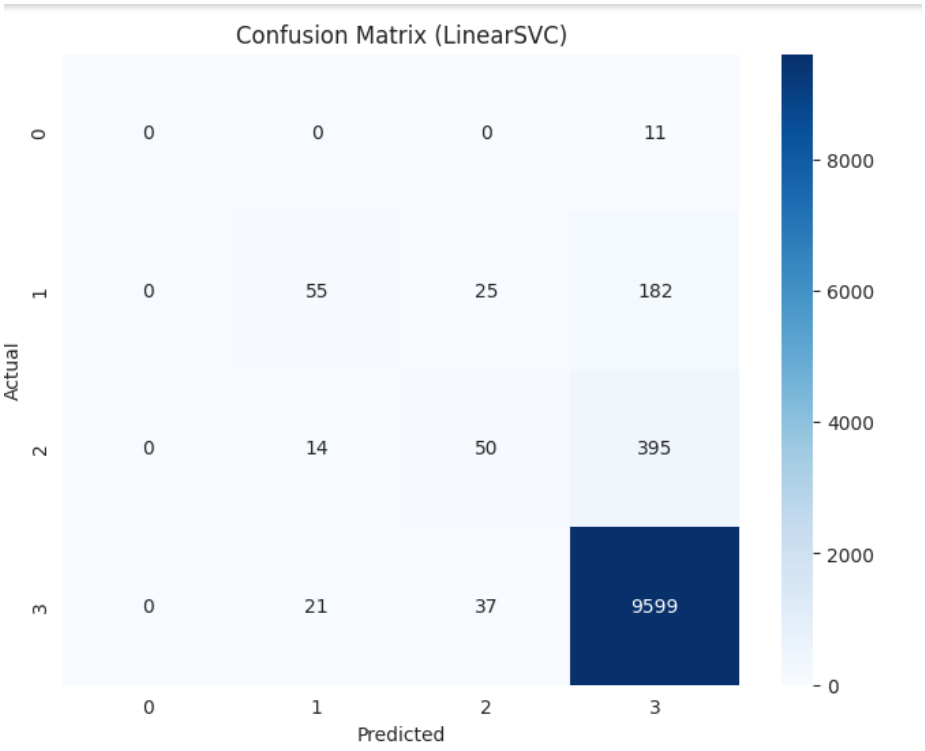

Figure 5: SVM Confusion matrix.

Classification Report (LinearSVC):

	precision	recall	f1-score	support
	0.00	0.00	0.00	11
Negative	0.61	0.21	0.31	262
Neutral	0.45	0.11	0.18	459
Positive	0.94	0.99	0.97	9657
accuracy			0.93	10389
macro avg	0.50	0.33	0.36	10389
weighted avg	0.91	0.93	0.91	10389

Figure 6: SVM classification report.

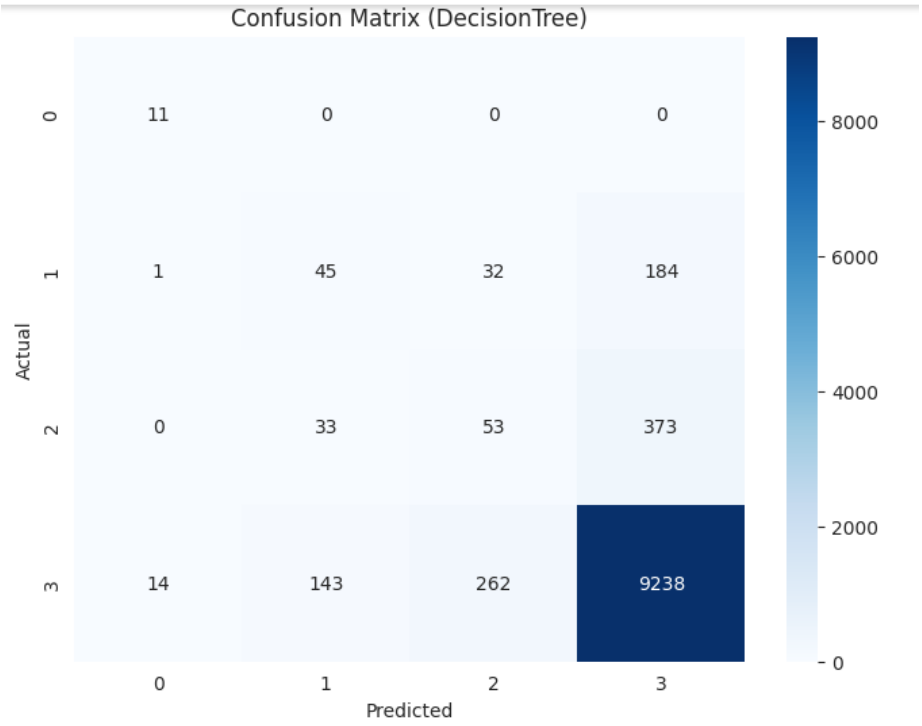

Figure 7: Decision Tree Confusion Matrix.

Classification Report (Decission tree):

	precision	recall	f1-score	support
	0.38	1.00	0.55	11
Negative	0.25	0.21	0.23	262
Neutral	0.16	0.13	0.14	459
Positive	0.94	0.96	0.95	9657
accuracy			0.90	10389
macro avg	0.43	0.57	0.47	10389
weighted avg	0.89	0.90	0.90	10389

Figure 8: Decision Tree classification report.