SENTIMENT-BASED INSIGHTS INTO AMAZON MUSICAL INSTRUMENT PURCHASES

Alla Ammulu[1], Maridu Bhargavi [2], Ande Mokshagna[3],
Bollimuntha Manasa[4], and Parasa Ganesh[5]

[1]Department of CSE, Vignan's Foundation for Science, Technology and Research, Vadlamudi, Guntur, Andhra Pradesh, India
[1]allaammulu0@gmail.com
[2]bhargaviformal@gmail.com
[3]mokshagnaande55@gmail.com
[4]manasabollimuntha1656@gmail.com
[5]ganeshparasa2005@gmail.com

Abstract

The project aims to analyze Amazon music instrument reviews to understand customer satisfaction in terms of key sentiments related to the products. Online shopping has made customer reviews increasingly relevant for both buyers and sellers. The project analyzes unprocessed text data using Natural Language Processing (NLP), which includes steps like text cleaning, tokenization, stop-word removal, stemming, and lemmatization. Word embeddings or term frequency-inverse document frequency (tf-idf) are used to encode the large dataset. A method of sentiment analysis on user reviews is described, with the dataset going through preliminary NLP work. Sentiments are classified using various models, evaluated based on accuracy, cross-validation scores, and classification reports. Visualization graphics depict sentiment distribution, emphasizing the potential of NLP for better sentiment analysis performance.

Keywords : Sentiment Analysis,Machine Learning Mod- els,NLP,Text Classification,Model Evaluation and Visualization

1 Introduction

In recent years, sentiment analysis has gained great importance and has typically been used in understanding customer behavior, especially in e-commerce, where consumer reviews significantly impact selling strategies and business practices[1]. This paper aims to describe a way in which reviews from the users can be classified based on rating using various methods of technology to evaluate the emotions of the consumers. Homogeneous, multilabel, multi class data concerning musical instruments reviews is in this regard processed using NLP techniques such as stemming, stopword removal, and TF-IDF in order to change unstructured text data to a structured format that machine learning makes sense of.

Some of the machine learning models that are executed include, Gradient Regression, Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Gradient Boosting Machine (GBM), Naive Bayes, and LightGBM are a few examples of other. These models are produced using pre-established rating cut-off points, at which they are trained and evaluated to anticipate an outcome of the review being positive, neutral or negative. To assess the accuracy of the various models and thereby develop an effective model for the purposes of sentiment prediction, parameters such as accuracy scores, cross validation and classification reports are put into place. Also, the features distribution of ratings and other text features are presented in graphs to help capture the overall trend of user feedback better.

Table 1: The dataset used in the implementation contains a list of columns and their descriptions

Attributel	**Description**
Reviewer Id	Uniqu identifier for reviewer
Asin	a unique product identification number
Reviewer Name	The reviewer's full name
Helpful	The reviewer's rating of helpfulness
Review Text	Text that the user submitted for approval
Overall	The user's rating of the product
Summary	The user's summary or title of the review
Unix Review Time	Unix timestamp for the information and the moment the review was sent in
Review Time	Information and the time the review was sent in

This study's main goal is to evaluate the various machine learning models' efficacy in doing sentiment analysis and giving companies a trustworthy tool for examining client comments. It is evident that integrating different machine learning classifiers with nlp techniques makes the analysis of user emotion very effective. This method makes the selection of the most optimal model for solving practical problems easier, as it is possible to perform comparisons of numerous models to evaluate how effective each scan classification algorithm is in completing sentiment analysis tasks[2]. This technique can help businesses categorize reviews more effectively and employ the results obtained to enhance customer service, create improved products, and finally increase the number of loyal customers. This will help them make better decisions in trying to enhance customer satisfaction and improve their product offerings.This is roughly the customer review for each of the 10,000 or so musical instruments present in the Amazon has to be predicted about their residual feelings either positive or negative. The raw data has been extracted from the Kaggle dataset for Amazon Musical Instruments [3]. The dataset consists of 9 columns and 10262 rows. The list of the columns is given in Table 1 and a short description is also attached for each column.

2 Literature Survey

Zhang and Yin focus on issues related to domain-specific sentiment analysis in e-commerce[1].Information Processing & Management.The paper explores domain-specific problems in sentiment analysis related to musical instruments, offering direction for feature extraction in applied contexts involving complex items such as musical instruments.

Rashika Mishra and colleagues have performance of reaching 92.8% accuracy using CNN based architecture with U-Net network on ISBI 2017 dataset. Redha Ali and others presented a way to use VGG19 and U-Net structure and achieved 93.6% on ISIC 2018 dataset. Use features extracted from grab cut segmentation of melanomas and SVM based study was conducted.

Wanliang et al., Xinyu W et al., and Xinyu X et al. [3] investigated Ama using conventional machine learning methods in conjunction with the deep learning algorithm RNN.

Yadav et al. and Jain et al. [4] employed a non-standard feature-based method in 2020 to perform sentiment analysis on Flipkart and Amazon product reviews. In fact, it fared better than TF, TF-IDF, and Naive Bayes Meth .

In 2017, Hu et al. [5] combined the use of the LSTM model with the application of keyword vocabulary to do sentiment analysis on brief texts.

3 Methodology

The methodology will include sentiment analysis and eval- uation of a machine learning model on text data regarding reviews of musical instruments including data preprocessing, feature extraction, training, and comparison.

3.1 Data Pre-processing

We carried out a number of data pre-processing procedures, which are described below, in order to get the dataset ready for sentiment analysis: Handling Missing data: The reviewerName and review have missing values. An empty string has been used in the text field for consistency and to avoid any errors during text processing. Preprocessing: Text Cleaning Normalizing the Textual Material Removing extra and redundant characters, conversion of text to lowercase, removal of stop words, tokenization, and stemming are the default text pre-processing function that we utilized. These steps in text pre-processing normalize the textual material and even reduce the number of vocabulary that might be subjected to proper analysis.

The processed text was then fed into a new column called stemmed content.Sentiment Labelling: Four classes of sentiment were identified from the rating of the review (overall column): Positive (ratings 4 and 5), Neutral (rating 3 and), Negative (ratings 1 and 2). This transformation was conducted using a custom method with the sentiment class of every review stored in a new column called Sentiment.

3.2 Data Augmentation

This means that we put the stratified sampling methodology to good use in solving the class imbalance anomaly identified with our sentiment analysis data set. Class imbalance is a problem in multi-class classification wherein certain classes dominate the dataset, so if left unchecked, this would have significantly affected model performance. For this particular data set, we had the three classes of sentiment: neutral, negative, and positive.
The given dataset is balanced after the data augmentation process. Since all the nine classes of the data are equal in size, the overall dataset is balanced as well. Therefore, it would improve the performance of the model if this balanced data is to be tested on it. This will perform better with data augmentation. Data augmentation is useful in overfitting cases, or those where the train data is fewer.

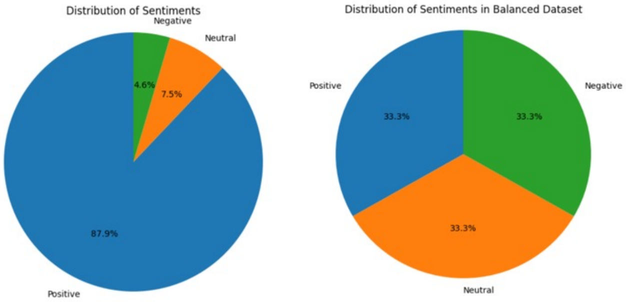


Figure 1: Rating distribution histogram

3.3 Feature Extraction

One of the most important phases in improving context understanding is feature extraction. The text needs to be transformed into usable features after proper cleaning and preparation for modeling. This method primarily aims at finding characteristics that positively affect the results of the categorization. In our study, we have applied two feature extraction machine learning methods, namely Bag-of-Words model and TF-IDF.

(TF-IDF): We applied the technique TF-IDF to emphasize more relevant words and de-emphasize stop words. This technique gave a weight to each phrase based on how many times it appeared in any given document relative to

how many times it occurred throughout the entire corpus. While doing so, the technique of TF-IDF highlights especially revealing words for the mood.
Term Frequency:
tf(t,d) = amount of words in d, count of t in d
Document Frequency:
df(t) = incidence of t in written records
Inverse Document Frequency:
df(t) = N(t)
where
N(t) = Total number of documents that contain the term t
df(t) = Document frequency of a term t

Therefore,
tfidf(t, d) = tf(t, d) * idf(t)

The frequency of a phrase in a given document is determined by TF. It is only the total words in a text divided by the frequency with which a phrase occurs in that text. Often occurring terms in a manuscript are underlined prominently. IDF determines a term's rarity across a set of texts. Then, there would be penalties for terms that appear in every document. IDF and TF combined to form TF-IDF. To determine a term's TF-IDF score in a document, calculate each term's TF and IDF scores and multiply them together.

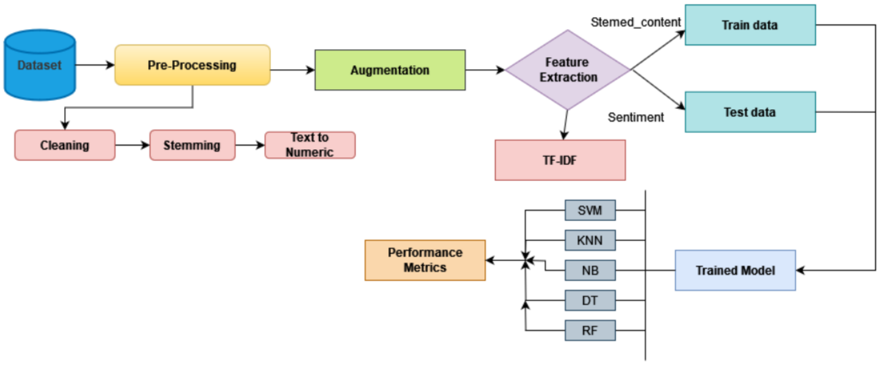


Figure 2: Proposed model flowchart

Algorithm 1 Sentiment Analysis of Amazon Music Instrument Reviews Processx

```
Input: Dataset
$D = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i$ are features and $y_i \in \{0, 1\}$ 0 = negative, 1 = positive (0 = legitimate, 1 = fraud)
Output: Best-performing model with metrics
Step 1: Preprocessing
- Remove any unnecessary columns and clean the text data.Convert the text data into numeric form using TF-IDF. $D$:
    $$D = Dtfidf = TFIDF(D) \quad (1)$$
- Split $D$ into train (80%) and test (20%) sets:
    $$D_{train}, D_{test} = split(D, 0.8, 0.2) \quad (2)$$
Step 2: Balancing the dataset using SMOTE
- Apply SMOTE to handle class imbalance on the training data:
    $$D_{train}^{balanced} = SMOTE(D_{train}) \quad (3)$$
Step 3: Feature Selection using TF-IDF
-We applied the technique TF-IDF to emphasize more relevant words and de-emphasize stop words:
    $$tfidf(t, d) = tf(t, d) * idf(t) \quad (4)$$
    Step 4: Apply machine learning models
    - Train the following models on $F_k$:
    M = { Logistic Regression, Random Forest, Gradient Boosting, KNN, SVM, Decision tree, Light GBM, Naive bayes }
    Step 5: Model Evaluation
    - For each model $m \in M$, evaluate using the following metrics:Accuracy Precision Recall F1-Score
- Light GBM and decision tree provides the best performance for these metrics
Return Best model with metrics
```

3.4 Model Selection

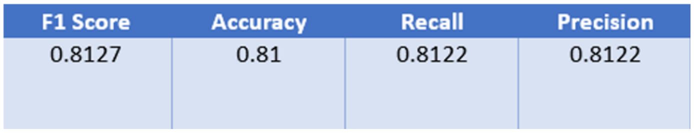
We assessed many machine learning algorithms to categorize sentiment into positive, negative, and neutral groups during the Model Selection stage. The following algorithms were taken into consideration: Logistic Regression: Very intuitive and familiar, applicable to multi class problems. SVMs: Function optimally even when the dimensions are very large.Random Forest: It is an ensemble technique that reduces overfitting and increases accuracy.Naive Bayes:

Good, feature-independent method for text categorization.

To test the performance of the model, we applied k-fold cross-validation testing over accuracy, precision, recall, and F1-score metrics. The implementation of the models was relatively easy using the scikit-learn toolkit, and the best model was further tuned for deployment purposes.

3.4.1 Decision Tree Classifier

learning, they are mainly applied for classification problems, but this is not to say that regression problems cannot be solved with it. This is a tree structured classifier with internal nodes substituting for attributes of the dataset, whereas the nodes represent decision rules and leaf nodes represent each outcome.



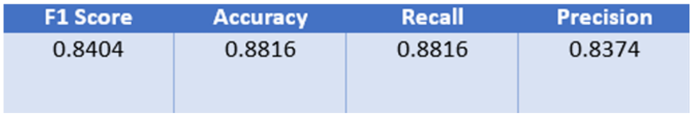
F1 Score	Accuracy	Recall	Precision
0.8127	0.81	0.8122	0.8122

Figure 3: Prediction Metrics for Decision Tree

The accuracy produced is 81.9 when used with TF-IDF. Using hyperparameter tuning and cross-validation, a high accuracy of 96.00 was achieved, the highest accuracy in this investigation. Table 4 displays the prediction metrics for cross-validation.

3.4.2 Light GBM

It is developed by Microsoft, an open-source, distributed, high-performance gradient boosting framework in focus on accuracy, scalability, and efficiency. Predicated on decision trees, it decreases memory utilization and enhances the efficacy of the model. In addition, LightGBM generates trees through methods involving histograms, which happens to be more efficient than all the other competing gradient boosting frameworks. These optimization methods give LightGBM an advantage, making it more efficient and more superior to other competing gradient boosting frameworks. Some other competitive optimizations include leaf-wise tree development and efficient data storage formats.



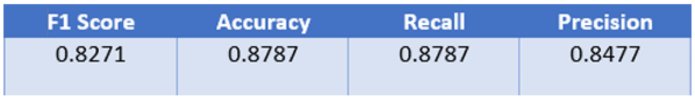
F1 Score	Accuracy	Recall	Precision
0.8404	0.8816	0.8816	0.8374

Figure 4: Prediction Metrics for LightGBM

The accuracy produced is 88.16 when used with TF-IDF. Using hyperparameter tuning and cross-validation, a high accuracy of 95.33 was achieved, the highest accuracy in this investigation. Table 4 displays the prediction metrics for cross-validation.

3.4.3 Logistic Regression

It applies algorithms, such as logistic regression, which is one of the most applied machine learning techniques. The approach uses an already existing set of independent factors when making a prediction of the category dependent variable.



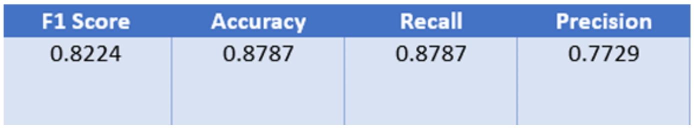
F1 Score	Accuracy	Recall	Precision
0.8271	0.8787	0.8787	0.8477

Figure 5: Prediction Metrics for Logistic Regression

The accuracy produced is 87.8% when used with TF-IDF. Using hyperparameter tuning and cross-validation, a high accuracy of 97.33% was achieved, the highest accuracy in this investigation. fig 5 displays the prediction metrics for cross-validation.

3.4.4 SVM (Support Vector Machine)

The support vector machine is a machine learning algorithm that uses either linear or nonlinear classification, strong regression, and outlier identification. Text classification, picture classification, handwriting identification, face detection, spam detection, gene expression analysis, and anomaly detection are a few applications of SVM. Because SVMs can learn in a high-dimensional space and take nonlinear interactions into account, they are widely used.



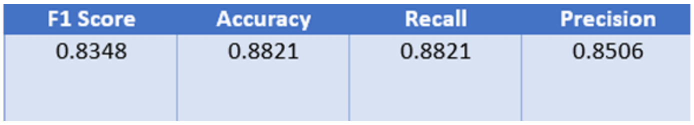
F1 Score	Accuracy	Recall	Precision
0.8224	0.8787	0.8787	0.7729

Figure 6: Prediction Metrics for SVM

The accuracy produced is 87.8% when used with TF-IDF. Using hyperparameter tuning and cross-validation, a high accuracy of 97.33% was achieved, the highest accuracy in this investigation. Table 6 displays the prediction metrics for cross-validation.

3.4.5 Gradient boosting model

In gradient boosting, a new model is learned at each step to minimize the loss function, which could be the previous model's cross-entropy or mean squared error. Gradient boosting, then, is a very effective boosting technique that turns a lot of weak learners into strong learners. Using the predictions from the current ensemble, the procedure calculates the gradient of the loss function at each iteration and trains a new weak model to maximize this gradient. Once the new model's predictions are included in the ensemble, the process is repeated until the stopping condition is met.



F1 Score	Accuracy	Recall	Precision
0.8348	0.8821	0.8821	0.8506

Figure 7: Prediction Metrics for Gradient boosting

The accuracy produced is 88.2% when used with TF-IDF. Using hyperparameter tuning and cross-validation, a high accuracy of 96.00% was achieved, the highest accuracy in this investigation. Table 7 displays the prediction metrics for cross-validation.

4 Metrics

The performance of our ensemble model is measured according to several essential metrics, which are spelled out in the section that follows:

A. Confusion Matrix

This matrix is also responsible for documenting how accurate and incorrect the various predictions made by the model were when the evaluation was done on the test data. The Confusion Matrix is widely used to measure classification models, where it is used to predict a categorical label for each input instance. Some of the named quantities in it include True Positive(TP), True Negative(TN), False Positive(FP), and False Negative(FN).

B. Precision

Precision is a measure of how well a model predicts favorable outcomes. The ratio of all the model's positive predictions to the actual positive forecasts is its definition.

$$Precision = \frac{A}{A+B} \tag{5}$$

C. Recall

Recall measures a classification model's ability to identify each relevant instance in a dataset. The percentage of true positive (TP) cases relative to the total number of false negative (FN) and true positive (TP) cases is what matters.

$$Recall = \frac{A}{A+C} \tag{6}$$

D. F1-Score

The F1-score is used to evaluate a classification model's overall performance. The harmonic mean of precision and recall is what it is.

$$F1Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{7}$$

E. Accuracy

The performance of the model is evaluated based on its accuracy. It is computed as the ratio of all instances to all accurate occurrences.

$$Accuracy = \frac{A + D}{A + D + B + C} \tag{8}$$

Variable Definitions

A = Correctly Predicted Positives (TP)
B = Incorrectly Predicted Positives (FP)
C = Missed Positives (FN)
D = Correctly Predicted Negatives (TN)

5 Results and Discussions

This research study demonstrated that SVM, Na¨ıve Bayes, K-NN, Light-GBM, Logistic Regression will get the best result on this dataset with accuracy ranges above 90+%. The essential aim of this research has been such that most reviews concerning the products should be correctly categorized and analyzed to let us determine which products customers love or detest. Therefore, we will now compare the classifiers that we used in our implementation with the ones already in use in this sector. Models which generalize well will be good and include logistic regression and SVM[. The model found to be overgeneralizing and the model reached the highest training accuracy.

Table 2: Performance Metrics of the Proposed Models

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Decision tree	0.96	0.81	0.81	0.81
Light GBM	**0.95**	**0.856**	**0.88**	**0.83**
Gradient Boosting	0.96	0.85	0.88	0.83
Logistic regression	0.97	0.84	0.87	0.82
SVC	0.97	0.77	0.87	0.82
KNN	0.96	0.77	0.87	0.82

A comparison of the results from the second set in comparison with the first set from all three models shows an improvement.The largest change is found in Logistic Regression (from 0.88% to 0.97%).In the second set, KNN and Decision Tree are equally good at achieving an accuracy of 96%, with a slight edge to Decision Tree for the first set. The probable adjustments to model settings, data quality, feature engineering, and training protocols were done in light of the significant increase in accuracy achieved with the second set.

6 Conclusion

The investigation concludes that feature engineering and preprocessing have a major influence over the performance of the machine learning model. Among all the used models, the best generalization skills were presented by the Gradient Boosting and LightGBM models, which are quite good for this kind of issue. In the future, more advanced NLP techniques can be further used to improve accuracy in sentiment predictions, ensemble learning to get better models, and hyperparameter tuning to optimize better.

7 References

[1] Amazon Musical Instruments Reviews Dataset for musical instruments review and ratings in Kaggle (https://www.kaggle.com/eswarchandt/amazon-music- reviews)

[2] Chattopadhyay, A., Basu, M. (2022). Unsupervised Learning Based Brand Sentiment Mining using Lexicon Approaches–A Study on Amazon Alexa. Indian Journal of Data Mining (IJDM), 2(1), 15-20.

[3] Sánchez-Franco, M. J., Arenas-Márquez, F. J., Alonso-Dos-Santos, M. (2021). Using structural topic modelling to predict users' sentiment towards intelligent personal agents. An application for Amazon's echo and Google Home. Journal of Retailing and Consumer Services, 63, 102658

[4] Khurana, P., Gupta, B., Sharma, R., Bedi, P. (2024). A sentiment-guided session-aware recommender system. The Journal of Supercomputing, 1-40.

[5] Vishwakarma, S., Garg, D., Choudhury, T., Singh, T. P. (2023, December). Amazon Sales Sentiment Prediction and Price Forecasting Using Facebook Prophet. In International Conference on Cyber Intelligence and Information Retrieval (pp. 93-105). Singapore: Springer Nature Singapore.

[6] Marwat, M. I., Khan, J. A., Alshehri, D. M. D., Ali, M. A., Ali, H., Assam, M. (2022). Sentiment analysis of product reviews to identify deceptive rating information in social media: A sentideceptive approach. KSII Transactions on Internet and Information Systems (TIIS), 16(3), 830-860.

[7] Amorim, I. B. M. (2023). Using sentiment analysis to predict Amazon ratings: a comparative study using dictionaries approaches (Doctoral dissertation).