
# CSE422: Artificial Intelligence
# Project Report
# Project Title: Book Review Analysis

| Group No:06, Lab Section:10, Summer 2023 | |
|---|---|
| **ID** | **Name** |
| 20101488 | Mubtasim Saadid Ahmed |
| 20101458 | Md Abu Sayem |
| 20341013 | SK Fahema Mahajabin |

# Table of Contents

## *Introduction:*

Sentiment analysis is a computational technique that can be used to determine the emotional tone from any given text or sentence. By analyzing the words, phrases, and context used in a text, sentiment analysis algorithms can identify whether the commenter's opinion is positive or negative. This technique is widely used in natural language processing and has a variety of applications, including studying product reviews. By analyzing product reviews using sentiment analysis, companies can quickly and easily identify which reviews are positive or negative, and use this information to improve their products or services. It is an effective way to gain insights into people's opinions towards different topics, making it a valuable tool for businesses to use.

## *Dataset Description:*

The source of our dataset is the "Amazon Kindle Book Review" available only on Kaggle, https://www.kaggle.com/datasets/meetnagadia/amazon-kindle-book-review-for-sentiment-analysis. It is a fairly simple dataset with 9 features that are asin, helpful, overall, reviewTime, reviewerID, reviewerName, summary, unixReviewTime. This dataset mainly consists of user reviews for various books in Amazon Kindle Store that are used for sentiment analysis in our project. The dataset allows for sentiment analysis modeling to determine the overall sentiment of reviews, enabling the classification of reviews as positive, negative, or neutral based on the text and associated ratings. By analyzing the most commonly mentioned authors, book titles, and other attributes, insights can be gained about popular genres, authors, and trends in Kindle book reviews. With the helpfulVotes column, it's possible to build models to predict the likelihood of a review being considered helpful by other users.The dataset's date column enables temporal analysis to understand how sentiment has evolved over time for different books and genres. There are 30000

data points in the dataset meaning there are about 30000 rows consisting of numerical and categorical variables. Looking at the histogram, we can see that the dataset is a little bit unbalanced.

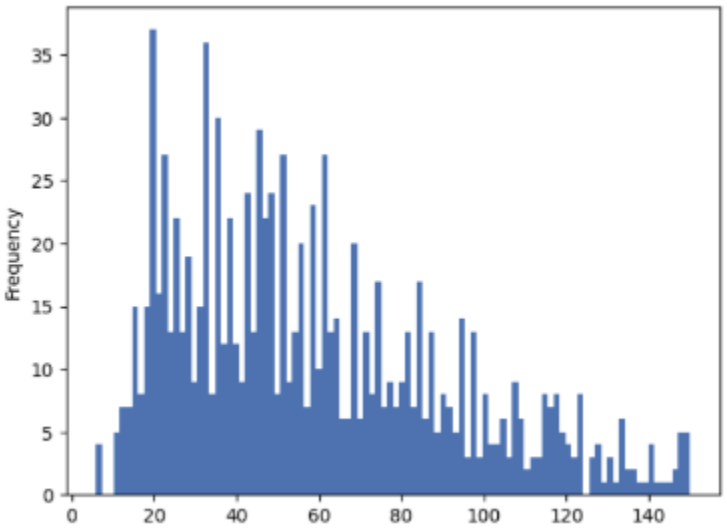



Fig. 1. Imbalanced Dataset Visualization

In this dataset, there are no NULL values present. We already discussed the unbalanced length of the reviews or statements left by the customers. It is hard to carry out proper sentiment analysis through a huge text. As a result, we balance our data by removing all reviews that exceed 64 characters. Moreover, the reviews must not be too short either and thus, we also removed all reviews that are less than 18 characters. Although we can still see some spikes in the histogram, relatively we end up with a more balanced dataset that provides us more accurate results.

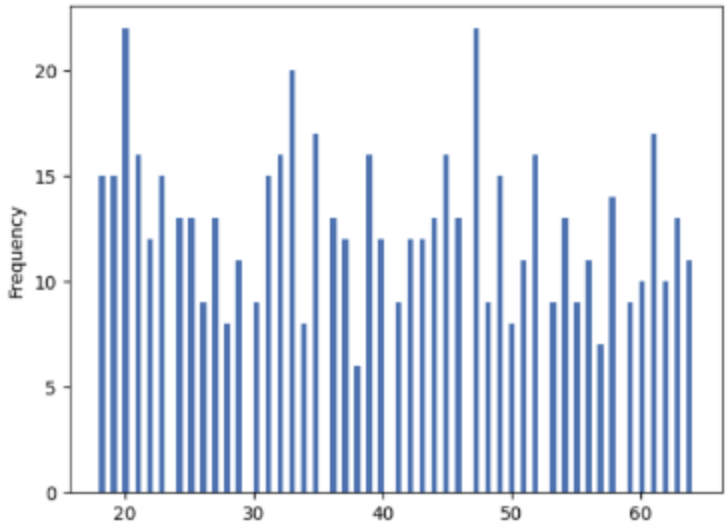



Fig. 2. More Balanced Dataset

Lastly, since the reviews in our dataset are in normal human language, it is difficult for machines to understand and work with them. Therefore, we have used natural language process-ing (NLP) techniques to pre-process the long texts and make them more understandable for the machines. Tokenization technique, which involves breaking down the reviews into individual words or tokens has been used here. We used Regular Expression and Natural Language Toolkit libraries along with the stopword removal and stemming modules for this function. We then represent the reviews as numerical feature vectors using the Bag-of-Words model, which counts the frequency of each word in the review. Feature scaling is not required in our study since the sentiments are being expressed by 0 (for negative) and 1 (for positive) only. There aren't any absurd values or change-worthy deviations in the dataset. Since there are no values that are outside of this range or any extreme outliers, there is no need for any data normalization or standardization techniques.

In order to train the machine learning models, we used the train test split function from the scikit-learn library. We set the test size parameter to 0.3, which means that 30 % of the rows in our

dataset were randomly selected to be used as a test set, while the remaining 70 % were used as a training set. We also set the random state parameter to 0, which ensures that each time the function is run, it produces the same set of random numbers. This randomizes the selection of rows for the test and training sets, which helps to ensure that the data is split in a consistent and reproducible way.

## *Model Training Testing:*

1) Naive Bayes:

   Naive Bayes is another popular classification algorithm used in machine learning for solving binary and multiclass classification problems. It is based on Bayes' theorem and the assumption of independence between features, which means that each feature is considered independent of every other feature. The Naive Bayes algorithm works by calculating the probability of a given data point belonging to a particular class based on the probabilities of the individual features. It calculates the conditional probability of each feature given the class, and then uses Bayes' theorem to calculate the probability of the class given the features. The class with the highest probability is then assigned as the prediction for the given data point. The Naive Bayes algorithm is popular due to its simplicity and efficiency, as it requires a relatively small amount of training data and can handle high-dimensional data well. It is commonly used in spam filtering, sentiment analysis, and text classification tasks. One of the main advantages of the Naive Bayes algorithm is its ability to handle missing data, as it simply ignores the missing values during training and classification. However, its assumption of independence between features may not hold true in all cases, which can lead to suboptimal performance. Nonetheless, it

remains a widely used and effective algorithm for classification problems. In our study we used Gaussian Naive Bayes,

$$P(X_i|C) = \frac{1}{\sqrt{2\pi\,\mathrm{Var}(C_i)}} \exp\left(-\frac{(X_i - \mathrm{Mean}(C_i))^2}{2\,\mathrm{Var}(C_i)}\right)$$

We got,

Best_score: 0.90

Best_alpha: 10.00

Best_min_df: 0.00100

2) Logistic regression:

One of the advantages of Logistic Regression is that it is a relatively simple and interpretable model. It can be easily implemented and trained on large datasets. Additionally, it provides useful information such as the odds ratio and the coefficients of the input variables, which can help in understanding the relationship between the input variables and the outcome. The logistic regression model used the bag-of-words approach to create a feature vector for each review, which was used to classify the sentiment of the review as either positive or negative. The logistic function (also known as the sigmoid function) was used to map the output of the linear regression model to a value between 0 and 1, which represented the probability of the review being positive. If the probability was greater than or equal to 0.5, the review was classified as positive, otherwise, it was classified as negative.

We got,

Training Accuracy w/ TfidfVectorizer: 0.95

Testing Accuracy w/ TfidfVectorizer: 0.93

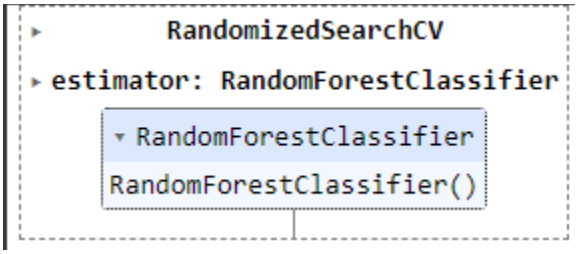
[[ 390  560]

 [ 116 8435]]

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.77 | 0.41 | 0.54 | 950 |
| 1 | 0.94 | 0.99 | 0.96 | 8551 |
| accuracy | | | 0.93 | 9501 |
| macro avg | 0.85 | 0.70 | 0.75 | 9501 |
| weighted avg | 0.92 | 0.93 | 0.92 | 9501 |

3) Random Forest Classifier:

Random Forest is capable of performing both Classification and Regression tasks. It is capable of handling large datasets with high dimensionality. It enhances the accuracy of the model and prevents the overfitting issue. Here we will visualize the training set result. To visualize the training set result we will plot a graph for the Random forest classifier. The classifier will predict yes or No for the users who have either reviewed or not as we did in Logistic Regression. The above image is the visualization result for the test set. We can check that there is a minimum number of incorrect predictions without the Overfitting issue. We will get different results by changing the number of trees in the classifier.

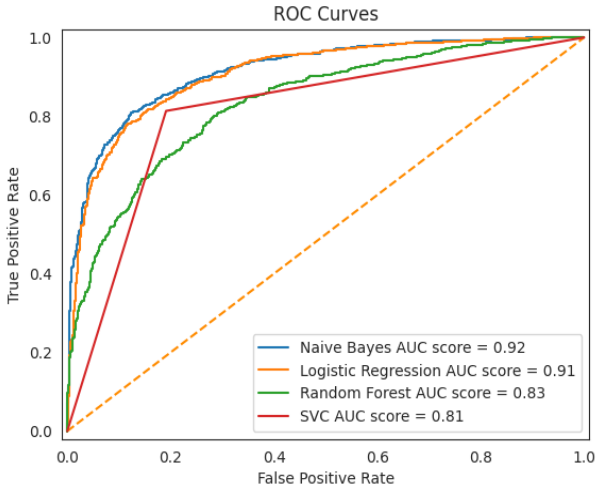


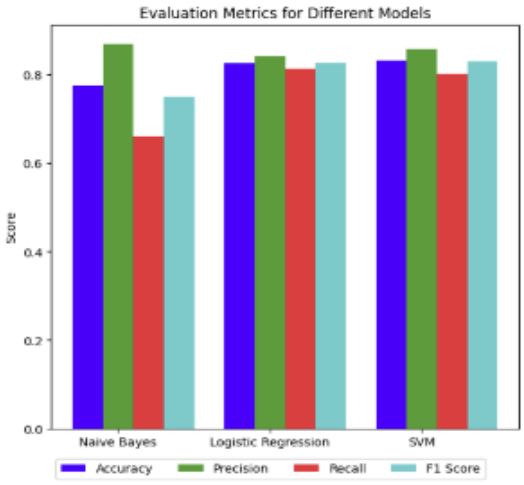
4) Linear SVC:

Support Vector Machine (SVM) is yet another powerful machine learning algorithm used for both classification and regression problems. In classification, SVM finds the best hyperplane that separates the data points of different classes with maximum margin, while in regression, it tries to find the best fitting line with maximum margin. The main idea behind SVM is to map the input data into a high-dimensional feature space where the data is more separable, and then find the hyperplane that best separates the data points. This is achieved by maximizing the margin between the hyperplane and the data points. The hyperplane that maximizes the margin is called the optimal hyperplane. In SVM, the decision boundary is defined by a set of support vectors, which are the data points that are closest to the hyperplane. These support vectors are used to calculate the distance between the hyperplane and the data points. The SVM algorithm uses a kernel function to map the input data into a high-dimensional feature space. The most commonly used kernel functions are linear, polynomial, radial basis function (RBF), and sigmoid. The formula for the optimal hyperplane in SVM is given by:

$$w^T * x + b = 0$$

Model Comparison Analysis:




| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Naive Bayes | 0.7765363128491162 | 0.8695652173913043 | 0.6593406593406593 | 0.75 |
| Logistic Regression | 0.8268156424581006 | 0.8999999999999999 | 0.8171868131368132 | 0.8268156424581005 |
| SVM | 0.8324022346368715 | 0.8588235294117647 | 0.8021978021978022 | 0.8295454545454546 |

| Confusion Matrix | TP | FP | FN | TN |
|---|---|---|---|---|
| **Naive Bayes** | 79 | 9 | 31 | 60 |
| **Logistic Regression** | 74 | 14 | 17 | 74 |
| **SVM** | 76 | 12 | 18 | 73 |

From our study, it was apparent that in the case of sentiment analysis on customer reviews, SVM and Logistic Regression outperforms Naive Bayes model easily. However, we observed a close performance between SVM and Logistic Regression. We used four of the most popular performance metrics for classification problems and those are accuracy, precision, recall and the F1score.

## *CONCLUSION:*

We have explored four different machine learning algorithms commonly used in classification problems: Logistic Regression, Naive Bayes, Support Vector Machines and Random Forest Classifier. Naive Bayes is another popular algorithm that works well with text classification problems. It is fast, efficient, and produces highly accurate results. It assumes that all features are independent, which may not always be the case, and can result in lower accuracy for more complex datasets. Logistic Regression is a simple yet powerful algorithm that works well with binary classification problems. It is easy to implement, has low computational requirements, and produces easily interpretable results. However, it may not work well with complex datasets and may suffer from overfitting. Support Vector Machines are highly versatile algorithms that can be used for both classification and regression problems. They work well with high-dimensional datasets, have high accuracy, and can handle non-linear decision boundaries. However, they may be computationally intensive and may require extensive tuning of hyperparameters. It is important to experiment with different algorithms and evaluate their performance before making a final decision.

The results demonstrated that all three models were able to accurately classify the sentiment of the reviews, with the support vector machine model performing slightly better than the others. The accuracy achieved by the models ranged from 83% to 95%, indicating that they were able to effectively capture the sentiment of the reviews. In conclusion, this study highlights the potential of machine learning algorithms in the field of sentiment analysis. By accurately analyzing customer reviews, businesses can gain valuable insights

into customer satisfaction and areas for improvement. The use of machine learning algorithms can also save time and resources, as they can process large amounts of data much faster than humans.