

Machine Learning Application in Sentiment Analysis

Natural Language Processing Bachelor Project

By

**Mohamed Yunes & Omar Fedawy**

Submitted to the School of Arts & Science of the Lebanese International University

In part of fulfillment of the requirements for the degree of

**BACHELOR OF SCIENCE IN** **COMPUTER SCIENCE AND INFORMATION TECHNOLOGY**

**Supervised by: Dr Ali Sabra**

Spring 2024-2025

Table of Contents

[Chapter 1: Introduction 3](#_Toc198759738)

[1.1 Problem 3](#_Toc198759739)

[1.2 Sentiment Analysis & Machine Learning 3](#_Toc198759740)

[1.3 Process 4](#_Toc198759741)

[Chapter 2: Implementation 5](#_Toc198759742)

[2.1 Preprocessing 5](#_Toc198759743)

[2.2 Training 7](#_Toc198759744)

[2.3 Testing 9](#_Toc198759745)

[2.3.1 Random Forest 9](#_Toc198759746)

[2.3.2 Testing With Other Models 10](#_Toc198759747)

[2.3.3 Evaluation 12](#_Toc198759748)

[2.3.4 Comparing Results 13](#_Toc198759749)

[Chapter 3: Conclusion 14](#_Toc198759750)

[3.1 Analyzing results 14](#_Toc198759751)

[3.2 Recommendation 14](#_Toc198759752)

[References 15](#_Toc198759753)

**Table of figures:**

[Figure 1: Label Sentiment Conversion 6](#_Toc198579046)

[Figure 2: Sentiment Distribution 6](#_Toc198579047)

[Figure 3: Training the model 7](#_Toc198579048)

[Figure 4: Metrics for Random Forest 9](#_Toc198579049)

[Figure 5: Confusion matrix for random forest 9](#_Toc198579050)

[Figure 6: Logistic Regression Matrix 10](#_Toc198579051)

# Chapter 1: Introduction

## 1.1 Problem

We are in a data-driven age where a company’s number one priority is figuring out the needs of its customer base. The need for data analysis in the corporate and marketing world is higher than ever, and fields of data science like natural language processing have come a long way. Companies can use data science for research and analysis and to gauge the public sentiment regarding events and products and experiences. One of these applications is opinion mining, which can be carried out using a number of different methods including machine learning.

## 1.2 Sentiment Analysis & Machine Learning

Sentiment analysis (or opinion mining) is a process in natural language processing used to determine the emotional tone or attitude expressed in a piece of text. It classifies the sentiment into several degrees (in this case positive, neutral and negative.)

It is widely applied in fields such as marketing, customer service, and social media monitoring to understand public opinion, customer satisfaction, and brand perception. By analyzing reviews, comments, or posts, sentiment analysis helps organizations make data-driven decisions and improve user experience by gaining insight into how people feel about products, services, or topics.

Machine learning is one of the methods used to perform sentiment analysis. Machine learning is the process of training algorithms on data to identify patterns and make predictions or decisions without being explicitly programmed. It involves data collection, preprocessing, model training, evaluation, and deployment (Mitchell, 1997).

## 1.3 Process

This project aims to study the viability of machine learning for sentiment analysis, which involves several key steps: collecting and preprocessing textual data (e.g., cleaning, tokenizing, and vectorizing using methods like TF-IDF), labeling the data with sentiment classes (e.g., positive, negative, neutral), and then training a supervised learning model such as Random Forest, Logistic Regression, or Support Vector Machine. Then the models’ performance will be evaluated based on several key metrics to determine the viability of machine learning in sentiment analysis.

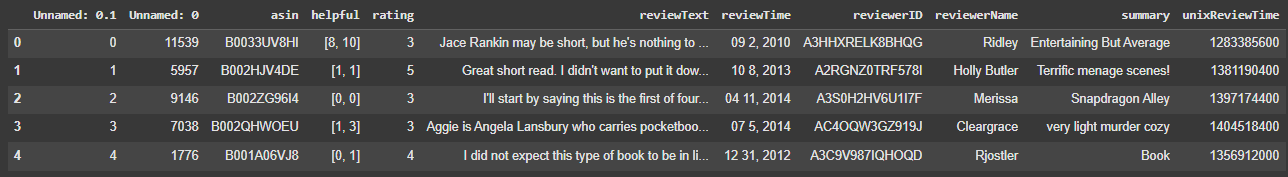
# Chapter 2: Implementation

## 2.1 Preprocessing

Preprocessing is the process of cleaning and preparing raw text data so it can be effectively used in natural language processing. It involves three steps.

1. **Importing the dataset:**

The used dataset that was used in this project was found on [Kaggle](https://www.kaggle.com/datasets/datafiniti/consumer-reviews-of-amazon-products) and it has the following attributes:

 Then the dataset is loaded into a data frame using the pandas library

1. **Cleaning the dataset:**

The dataset has many features but for this project we only need two of them: rating and review-text. We must also make sure there are no missing values (text or rating) since the model would not know what to do with those. We drop any rows with null values from the data frame.

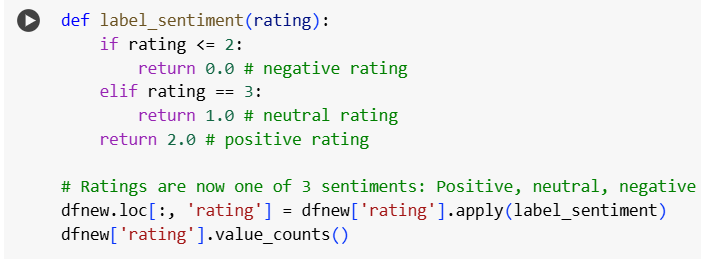
1. **NLP formatting:**

Now we need to clean up the review texts so they are more understandable to the machine learning model. First we remove stop words (unimportant filler words) and then we lemmatize the words so that they return to their lexical root forms, as shown in figure 1. This will allow the model to relate words between reviews much more accurately and comprehensively. Finally we normalize the ratings into 3 numbers (0, 1, 2) so it is easier for the model to categorize them. (figure 2)

After normalizing the ratings into 3 sentiments (positive, neutral, negative) it becomes obvious that there is a bias in the data where there are more positive sentiments than the rest. (figure 3) This will affect the accuracy of the model, however there are ways to work around this discrepancy.

Figure 1: Label Sentiment Conversion



Figure 2: Cleaned Reviews

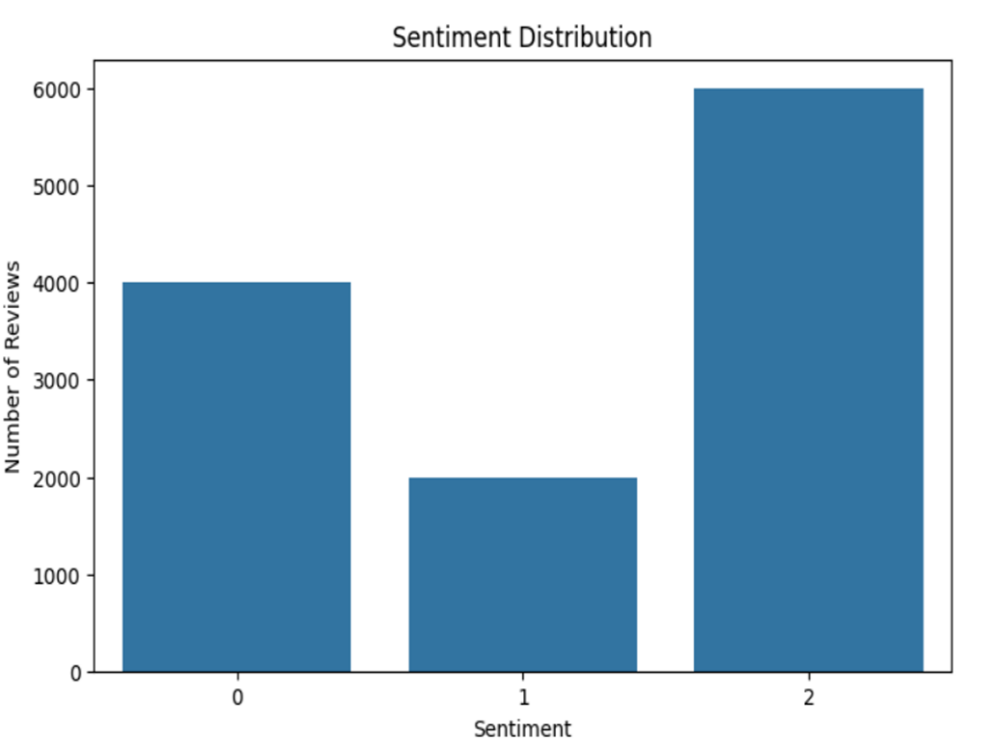


Figure 3: Sentiment Distribution

## 2.2 Training

After the dataset is processed for machine learning purposes, the model must be trained on the data. The process is as follows:

1. Before training the model we must vectorize the reviews to numerically represent the text so our model can understand it. TF-IDF will do this for us.

2. After that we split the model using the TRAIN-TEST-SPLIT Provided by the SCIKIT LEARN library. This splits the dataset into a portion for training and a portion for testing later on.

3. Finally, we train the model using the training set. At first we will be using RANDOM FOREST CLASSIFIER .

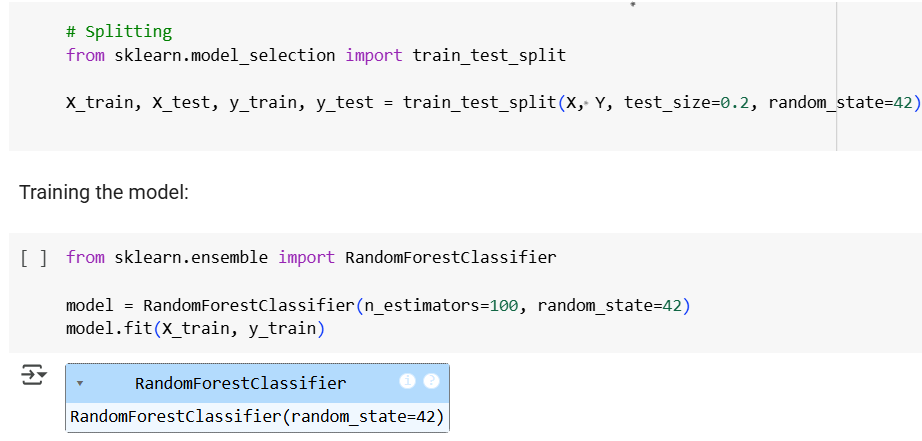


Figure 4: Training the model

This is how Random forest classifier algorithm works:

1.Multiple Decision Trees:

The algorithm creates many decision trees, not just one.

Each tree is trained on a random subset of the training data (this is called bootstrapping).

2.Random Feature Selection:

When building each tree, only a random subset of features is considered at each split.

This introduces variety among the trees and reduces overfitting.

3.Voting for Classification:

When making a prediction, each tree gives a vote.

The final output is based on majority voting where the class with the most votes becomes the prediction.

## 2.3 Testing

### 2.3.1 Random Forest

After training the model on our dataset, we can evaluate several factors about the results. We use the testing portion of the dataset to evaluate the model.

We start with the accuracy which turns out to be approximately 70% which is a somewhat satisfying result. However, since our dataset is a little skewed towards positive reviews, accuracy alone is not a sufficient metric for evaluating this model. We need to be sure that the model is consistently able to determine sentiment.

After using the precision score, recall score, and f1 score we get the following results:

•Precision: 66.67%

•Recall: 70.79%

•F1: 64.66%

The precision means more true flags, the recall means less false flags, and the f1 score is a harmonic mean of the latter two. The f1 score is the most important metric when evaluating a classification model.

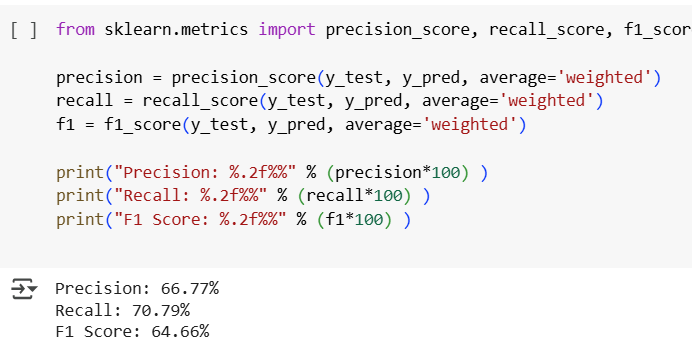


Figure 5: Metrics for Random Forest

A confusion matrix is a table used to evaluate the performance of a classification model. It compares the actual labels with the predicted labels and shows how well the model is performing.

The confusion matrix for the random forest model shows that while it is great at identifying positive sentiment, it is very weak at correctly identifying positive and neutral sentiment.

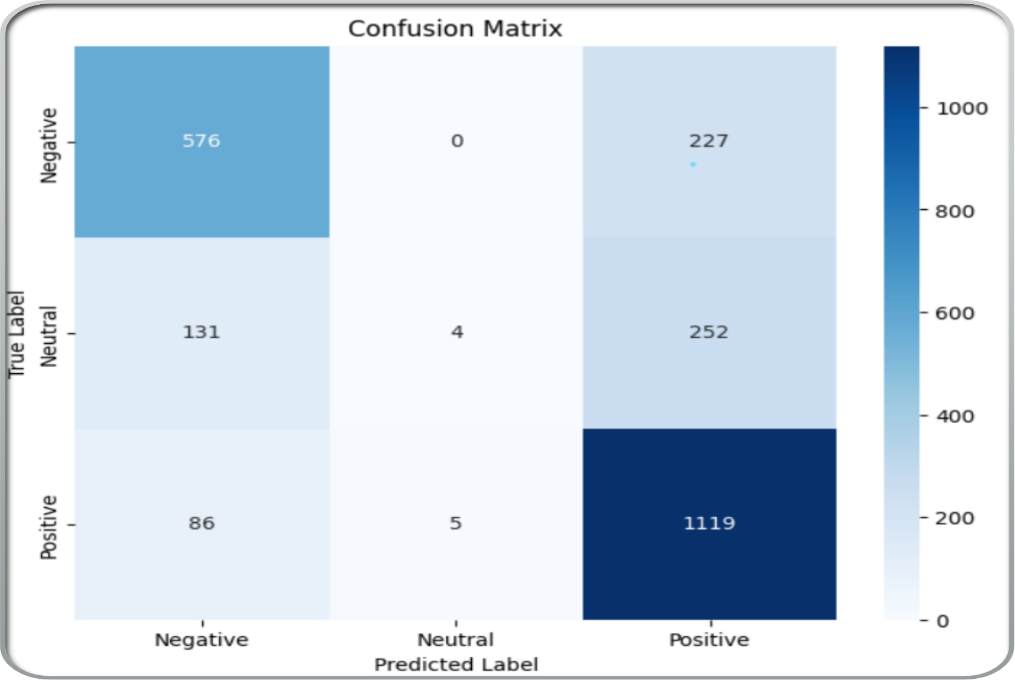


Figure 6: Confusion matrix for random forest

### 2.3.2 Testing With Other Models

Now we use another model provided by SCIKIT LEARN which is LOGISTIC REGRESSION leading to the following results:

•Accuracy: 75.46%

•Precision: 72.80%

•Recall: 75.46%

•F1:72.46%

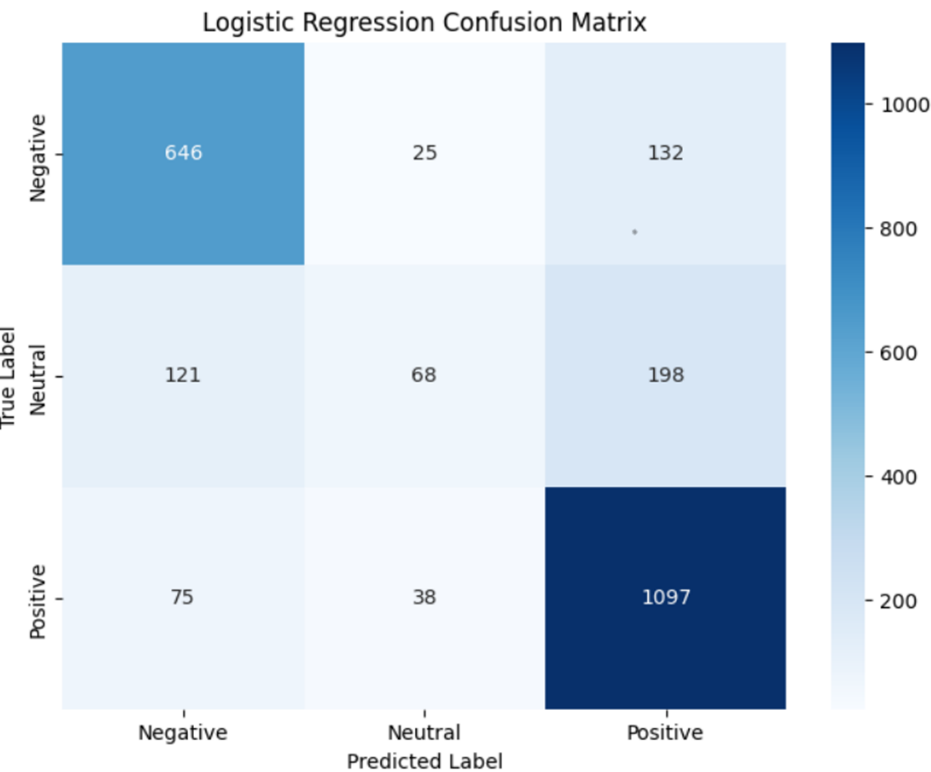


Figure 7: Logistic Regression Matrix

SVM is yet another classification model, after training the model we get the following results:

1.Accuracy: 75.08%

2.Precision: 72.09%

3.Recall: 75.08%

4.F1: 72.35%

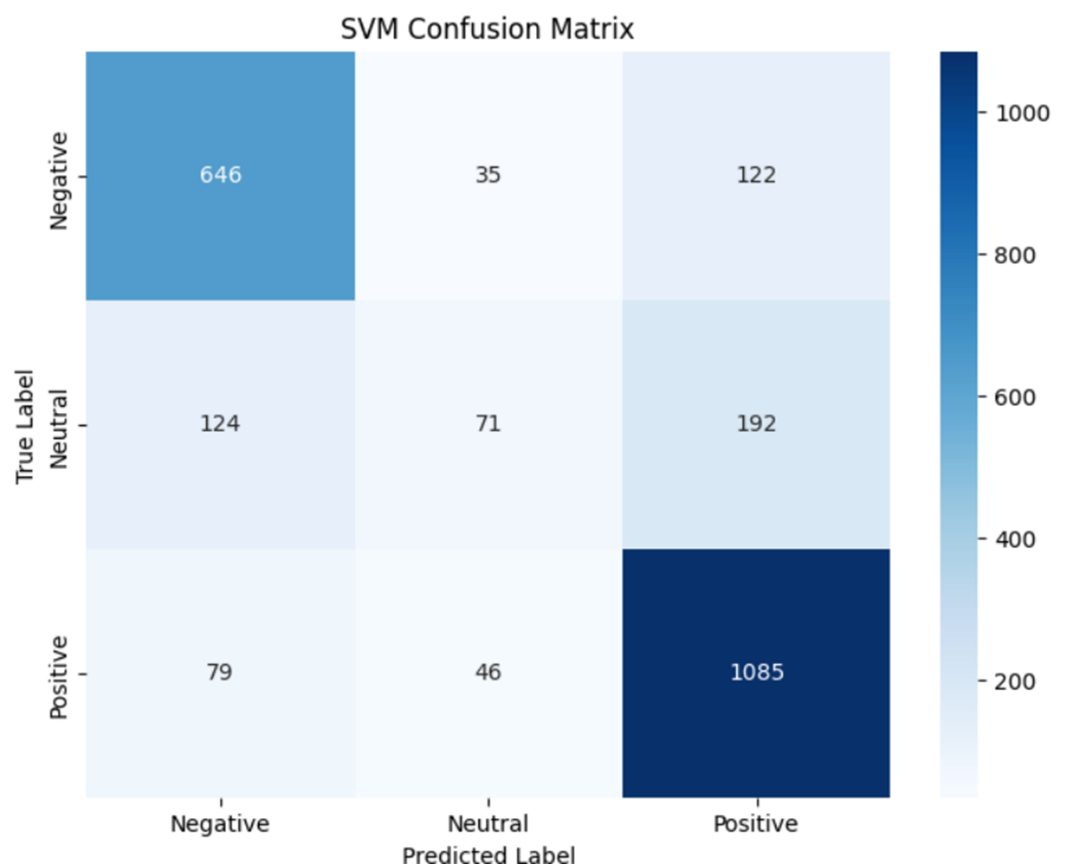


Figure 8: SVM Confusion Matrix

### 2.3.3 Evaluation

To test the viability of each model, we decided to run the following individual strings on each model:

1. "This book was great and reading it was a very joyful experience "
2. "This book was okay, I enjoyed some parts but most of it was uninteresting"
3. "This book was terrible and i did not enjoy reading it"
4. "Lorem Ipsum"
5. "لقد استمتعت بقراءة هذا الكتاب"

The expected outcome for 1 and 5 is positive, for 2 it is neutral, for 3 is negative, while 4 does not have an expected sentiment. We predicted that the models might misidentify the 5th sentence as the dataset only had English words for the models to work with.

The results for each model are as follows:

**Random Forest:**

1. Positive
2. Positive
3. Negative
4. Positive
5. Positive

**SVM:**

1. Positive
2. Neutral
3. Negative
4. Positive
5. Positive

**Logistic Regression:**

1. Positive
2. Neutral
3. Negative
4. Positive
5. Positive

All of the models correctly identified the sentiments except for random forest which misidentified the second sentence as negative.

### 2.3.4 Comparing Results

After testing several models it is evident that SVM and Logistic Regression are more suited than Random forest for sentiment analysis purposes, with logistic regression coming out on top. We determined this from

# Chapter 3: Conclusion

## 3.1 Analyzing results

After training and testing Support Vector Machine (SVM), Random Forest, and Logistic Regression models for sentiment analysis, the results show that machine learning provides a reasonable approach for classifying sentiment in text data.

Among the models tested, Logistic Regression performed the best in terms of accuracy and consistency. However, the overall effectiveness was influenced by the imbalance in the dataset, which likely affected the precision and recall across sentiment classes.

## 3.2 Recommendation

For future study, we would first make sure to pick a larger and more varied dataset to work off of. Ensuring a balanced and representative dataset will be key to building a more reliable and robust sentiment analysis system. Secondly we recommend exploring other methods of sentiment analysis application, like deep learning.

## References

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

"What is Machine Learning?". IBM. 22 September 2021.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.