Methods

This paper will demonstrate the analysis and implementation of training a sentiment analysis dataset. The performance is evaluated on one dataset. The sentiment from these reviews were predicted using supervised machine-learning algorithms, more specifically, logistic regression and random forest. The methods used in this research were trained and tested with google collab with the use of some basic libraries. The algorithms used were trained using regular runtime CPU'S, therefore high-power GPU's are not essential to run them.

Dataset

The dataset, consisting of several Amazon customer reviews, was sourced from Kaggle in a fastText format.

The dataset, having been pre-split into a training and a test set, consists of two sections: the input text (which are the customer reviews) and the output labels (which are the star ratings). The output labels are split into two categories: '__label__1' and '__label__2' [4]. The '__label__2' category comprises of positive 4- and 5-star reviews, while the '__label__1' category compromises of negative 1- and 2-star reviews [4].

Due to the large size of the dataset, the number of processing samples were reduced significantly. The training set was reduced to 8,000 samples, while the test set was reduced to 2,000 samples. Therefore, a total of 10, 000 samples were used in training each model.

Test Data Set: consists of 400,000 reviews, split equally between the negative and positive	
Attribute	Details
Output Labels	This consists of two labels: __label_1 and __label__2
Input Text	This section contains text reviews that expresses customer's sentiment.

Train Data Set: consists of 3,600,000 reviews, split equally between the negative and positive	
Attribute	Details
Output Labels	This consists of two labels: __label_1 and __label__2
Input Text	This section contains text reviews that expresses customer's sentiment.

Pre-processing

A YouTube tutorial by [8] was used as a basic reference for building a preprocessing function. Pre-processing a dataset is very important when training an artificial intelligence model. It is important because it tailors the dataset to its user's specifications, and removes unnecessary data, that complicates a model's performance and accuracy. Therefore, as a part of preprocessing, the dataset was formatted from a 'txt.bz2' back to a '.csv' file. This was done because of the ease, accessibility, and readability of using '.csv' files to train machine learning model as compared to the compressed '.bz2' format.

Furthermore, stop words, special characters and punctuation marks are removed from the input data. All words are turned to lowercase. Sentences are divided into individual words (tokenized), affixed words are broken down to their root form (stemmed) and converted to numerical data (vectorized) before finally going through the training and testing phase. At this point, the data must be split into training and test sets if it has not already.

Also, the output labels in the dataset were changed from '__label__1' and '__label__2' to '0' and '1' respectively. This has been done for several reasons. Firstly, it is essential because the chosen models need numerical labels as input for training and testing. Also, it is essential because numerical labels are needed for evaluating and plotting a model's performance metrics.

Some python libraries used in the code include:

1. The nltk library: This imported in the stop words and PorterStemmer functions.
2. The sklearn library: This imported in the TfidfVectorizer, LogisticRegression and RandomForestClassifier models, train_test_split, mean_squared_error, confusion_matrix and accuracy_score functions.
3. The csv and bz2 Libraries: the dataset was converted from .bz2 format using both the bz2 and csv libraries.
4. The pickle library: This library saved and loaded the trained models for future predictions.
5. The Matplotlib library: This enabled the plotting of both model's confusion matrix.

Algorithms

The algorithms of choice are logistic regression and random forest classifier. They are both popular machine learning models renowned for their performance in classification and regression tasks, therefore why they are the perfect algorithms for training this dataset. The logistic regression algorithm is used for classification and probability prediction [6], while the random forest classifier consists of multiple decision trees that filter out input data to reach a final output [7].

Logistic Regression Model

A YouTube tutorial by [8] was used as a basic reference for the model's parameters. The LR model was imported from the sklearn library therefore only its 'n_jobs' and 'max_iter' hyperparameters were finetuned. The 'n_jobs' value was set to -1, meaning that the model is allowed to use all CPU cores while training the dataset to speed up the training process. The max_iter value was set to 1000.

Random Forest Model

The RF model was imported from the sklearn library therefore only its 'n_estimators' and 'random_state' hyperparameters were finetuned. The 'n_estimators' parameter sets the number of decision trees used in the model. The 'n_estimators' value was set to 100, meaning that the model will have 100 decision trees. The random_state value was set to 1000.

Each algorithm will be trained and tested using the same pre-processed dataset. Subsequently, the algorithms will be evaluated on their accuracy, precision, recall, f1, root mean squared error (RMSE), false positive rate (FPR), and false negative rate (FNR) scores.

Precision is the ratio of true positives to the sum of classified positives in the dataset; recall is the ratio of true positives to the sum of actual positives in the dataset; and the F1 score is the average harmonic mean of the recall and precision [11]. Therefore, having a high recall, precision and F1 score means that a model is performing well.

The FPR is the ratio of false positive data to the total amount of negative data while, the FNR is the ratio of false negative data to the total amount of positive

data. Therefore, having a low FPR and FNR indicates that a model is performing well.

The formula for precision is TP/ (TP + FP).

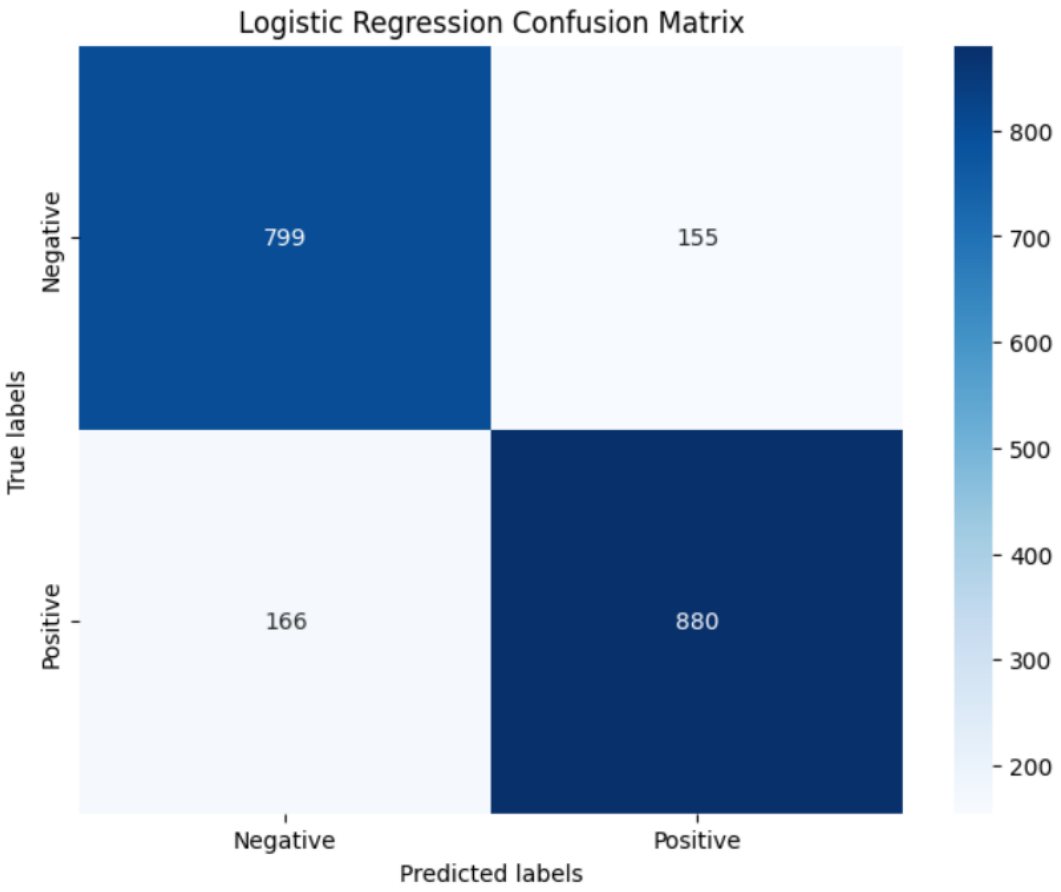
The formula for recall is TP/ (TP + FN).

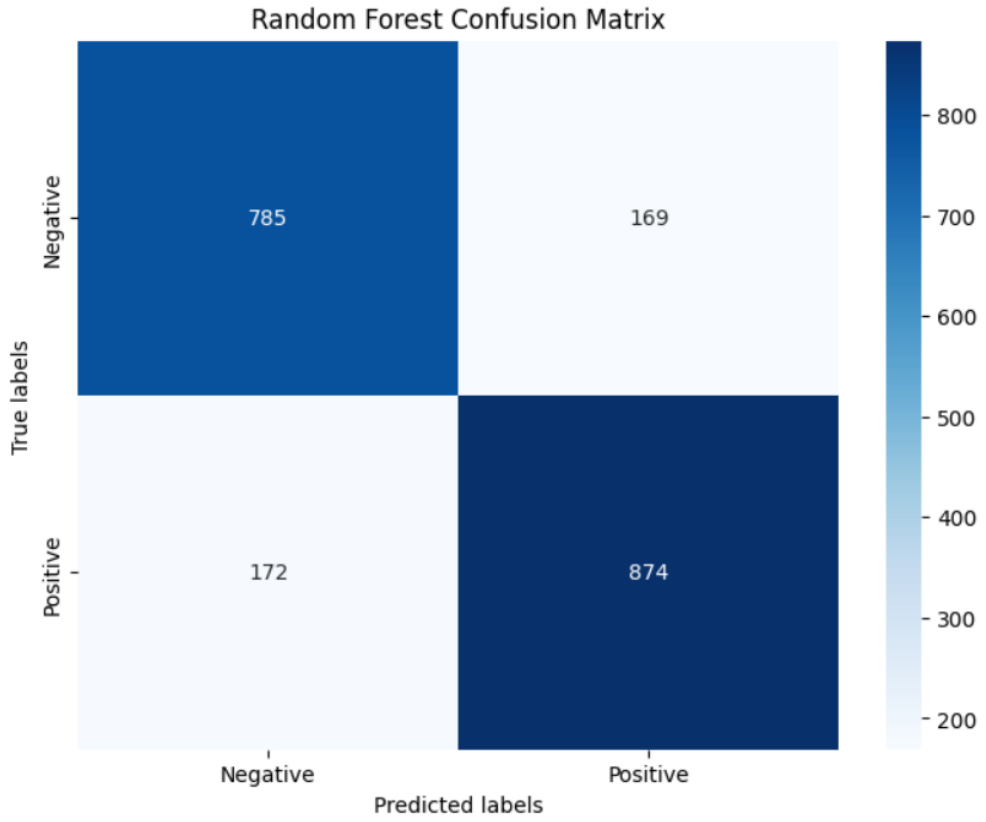
The formula for F1 score is 2(recall x precision)/ recall + precision.

Results and Discussion

The following are the training and testing results of both models:

Model	Training Accuracy (%)	Test Accuracy (%)	Precision	Recall	FPR	FNR	F1 Score	RMSE
Logistic Regression	92.44	83.95	0.85	0.84	0.16	0.16	0.85	0.4
Random Forest Classifier	100	82.95	0.84	0.84	0.18	0.16	0.84	0.41


Logistic Regression Confusion Matrix
Negative
Positive
True labels
799
155
166
880
Negative
Positive
Predicted labels
800
700
600
500
400
300
200



These results suggest that the RF model performs reasonably well, with a test accuracy of 82.95%, but is clearly overfitting. On the other hand, the LR model performs slightly better than the RF, with a test accuracy of 83.95%, and does not overfit.

Looking at the confusion matrixes, the RF model predicted 874 positive labels and 785 negative labels correctly, while the LR model predicted 880 positive labels and 799 negative labels correctly. This resulted in the LR model having a precision of 0.85, a FNR of 0.16, RMSE of 0.4 and a FPR of 0.16, while the RF model had a precision of 0.84, a FNR of 0.16, RMSE of 0.41 and a FPR of 0.18.

From the statistics above, it can be concluded that sentiment analysis can successfully measure consumer satisfaction from UGC. However, this raises significant ethical and societal concerns regarding privacy. According to Karoo and Chitte [12], analysing individuals' sentiment without their consent, though the data is public, can be deemed an invasion of privacy. Therefore, it is crucial

that corporations obtain proper informed consent from individuals which ultimately builds trust and transparency.

Model	Training Accuracy (%)	Test Accuracy (%)	Precision	Recall	FPR	FNR	F1 Score	RMSE
Logistic Regression	92.44	83.95	0.85	0.84	0.16	0.16	0.85	0.4
Random Forest Classifier	100	82.95	0.84	0.84	0.18	0.16	0.84	0.41

References

[4] A. Bittlingmayer, “Amazon Reviews for Sentiment Analysis”, 2020 [Online]. Distributed by Kaggle. https://www.kaggle.com/datasets/bittlingmayer/amazonreviews

[5] E. Y. Boateng and D. A. Abaye, “A review of the logistic regression model with emphasis on medical research,” *Journal of Data Analysis and Information Processing*, vol. 07, no. 04, pp. 190–207, 2019. doi:10.4236/jdaip.2019.74012.

[6] IBM. “What is logistic regression?” ibm.com. https://www.ibm.com/topics/logistic-regression (accessed April 29, 2024).

[7] IBM. “What is random forest?” ibm.com. https://www.ibm.com/topics/random-forest(accessed April 29, 2024).

[8] GeeksforGeeks. *Twitter Sentiment Analysis (NLP) | Machine Learning Projects | GeeksforGeeks*. (13 Nov 2023). Accessed: April 23, 2024. [Online Video]. Available: https://www.youtube.com/watch?v=4YGkfAd2iXM