EESTech 2024 - Hackathroners

We chose to use the “Sentiment140” dataset (https://www.kaggle.com/datasets/kazanova/sentiment140/data), which contains 1.6 million tweets along with annotations about each of their sentiment (positive or negative).

Unsupervised task - Clustering

We used k-means clustering as part of EDA, clustering users based on the mean sentiment of all their tweets. We validated our clustering algorithm, using the Davies-Bouldin cluster separation measure.

Supervised task - Sentiment classification

We mapped positive sentiments to 0 and negative to 1 and dropped the unnecessary columns. We then removed mentions, URLs, 4 strings starting with ampersand, multiple dashes as well as leading and trailing whitespaces. We also removed duplicates and split our dataset using a total of 200,000 random tweets from the dataset (100,000 of each sentiment class) into 80,000 train and 20,000 test data for each class. We vectorized our dataset and fed it to our model.

We built a feed-forward MLP with 4 hidden layers, using batch normalization, dropout and the ReLU activation function, except for the output layer, where we used the sigmoid activation function.

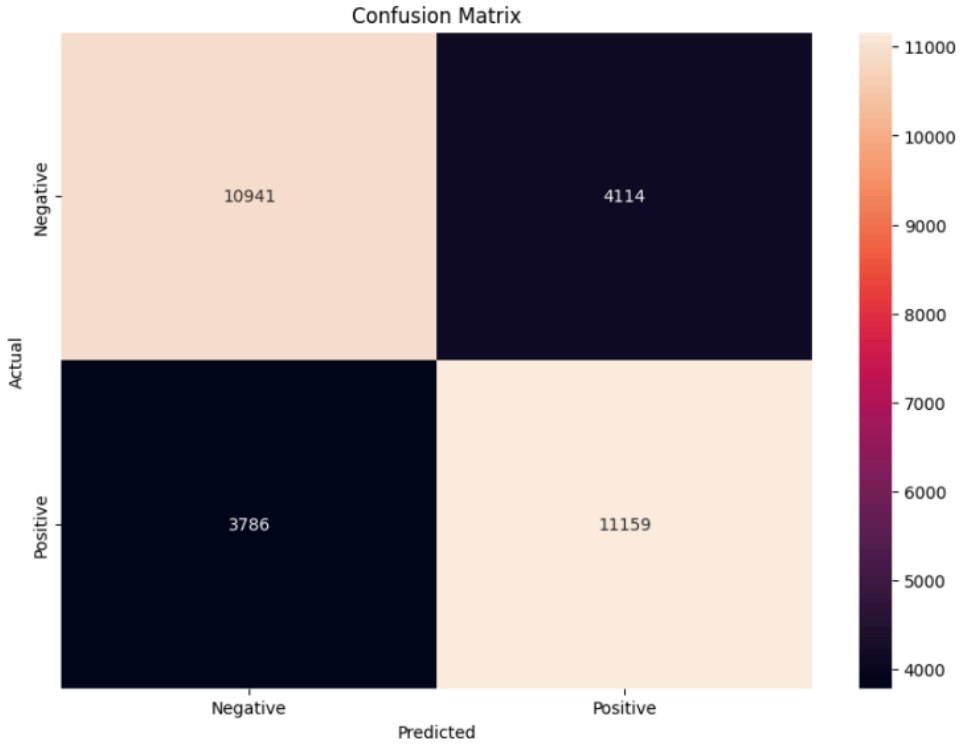
We achieved the following results:

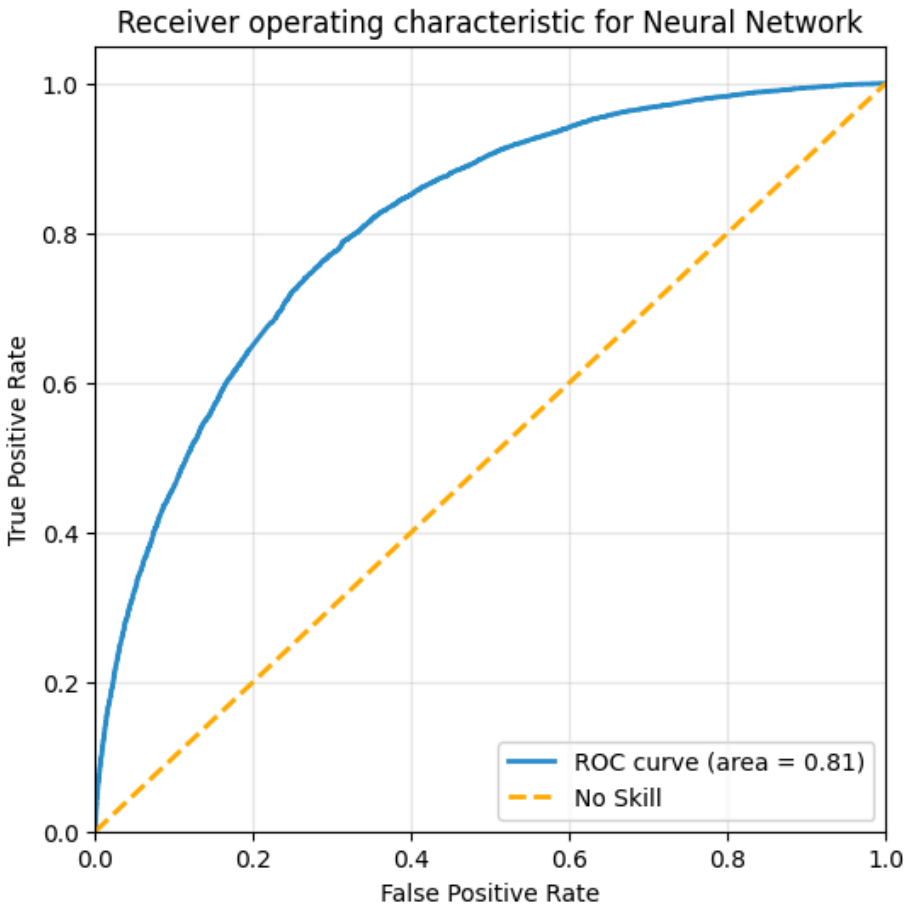
Train Classification Report

	precision	recall	f1-score	support
Negative	0.84	0.83	0.84	34945
Positive	0.84	0.84	0.84	35055
accuracy			0.84	70000
macro avg	0.84	0.84	0.84	70000
weighted avg	0.84	0.84	0.84	70000

Test Classification Report

	precision	recall	f1-score	support
Negative	0.74	0.73	0.73	15055
Positive	0.73	0.75	0.74	14945
accuracy			0.74	30000
macro avg	0.74	0.74	0.74	30000
weighted avg	0.74	0.74	0.74	30000




Receiver operating characteristic for Neural Network
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.81)
No Skill