SENTIMENT ANALYSIS WITH NAIVE BAYES CLASSIFIER

By Mohammad Anas Hussain & Mohammed Abrar Ahmed

STOCK MARKET

WHAT IS IT?

Stock news from Multiple twitter Handles regarding Economic news.

HOW MANY LABELS

- Negative (-1)
- Positive (1)

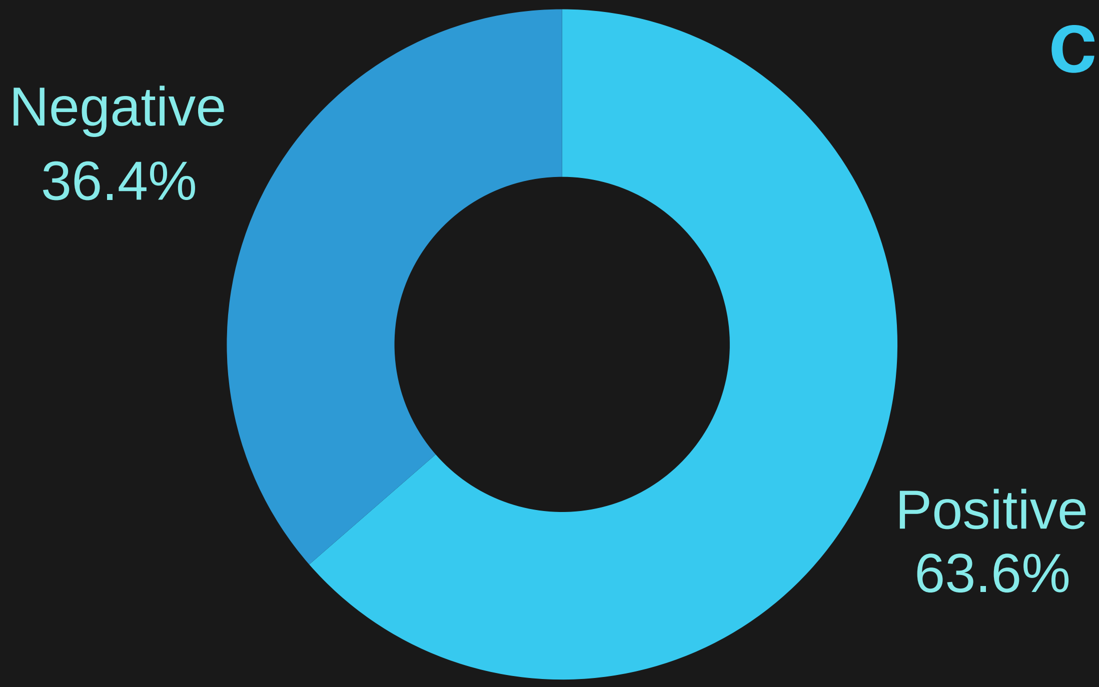
SAMPLES

- Negative count: 2,106
- Positive count: 3,685
- Total: 5,791

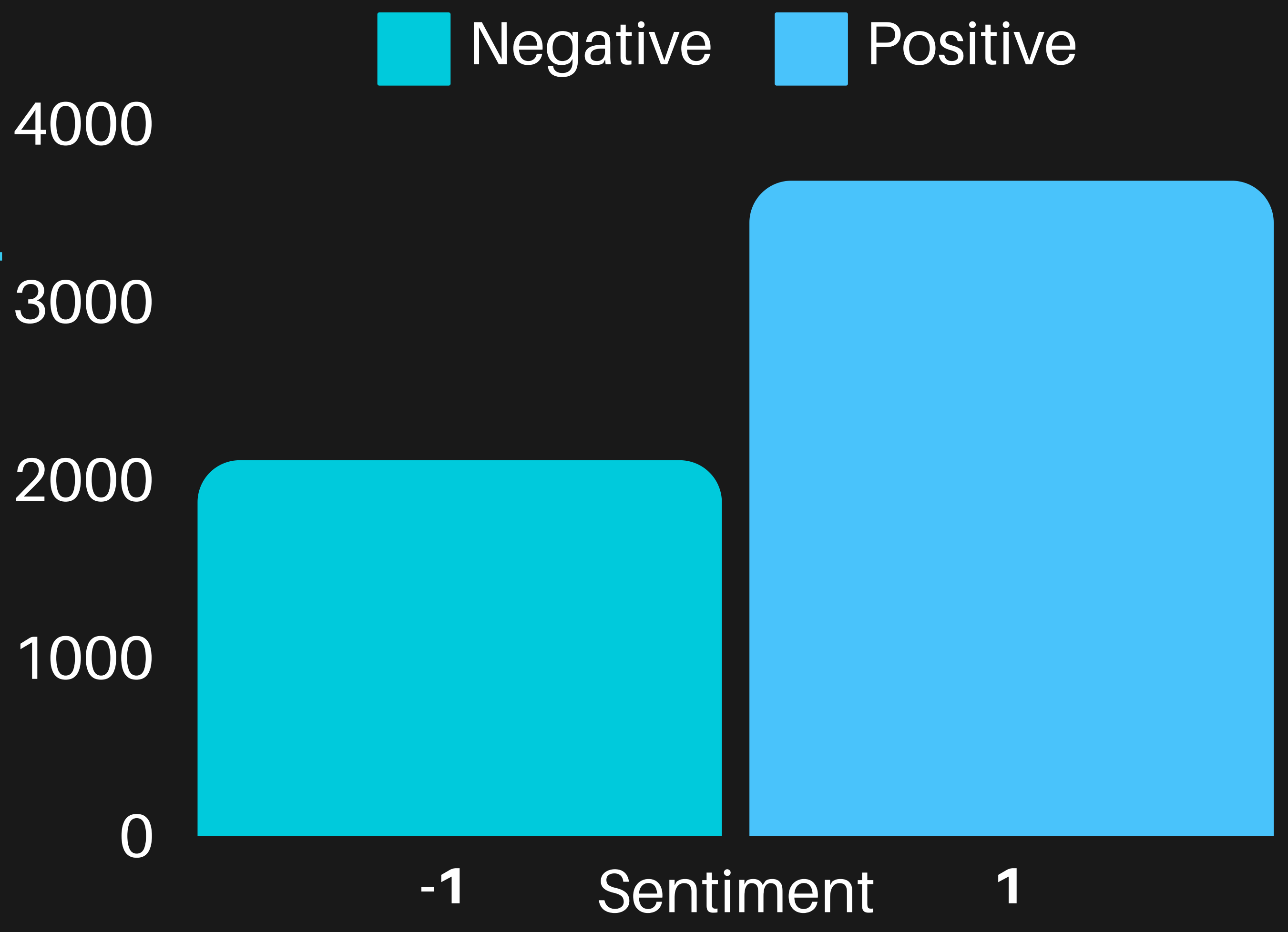
PREPROCESSING

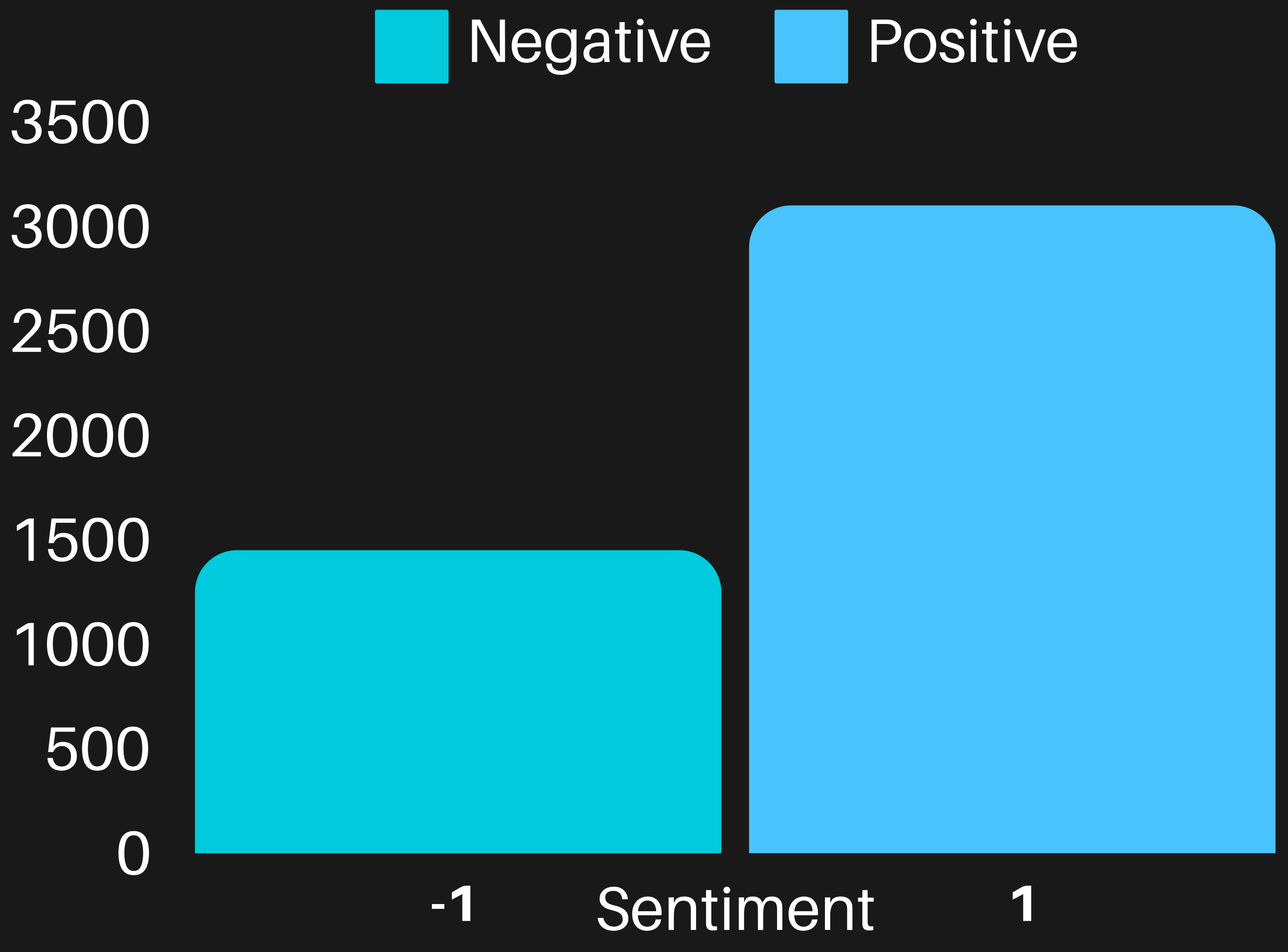
- Remove special characters
- Convert to lowercase

DATASET DETAILS

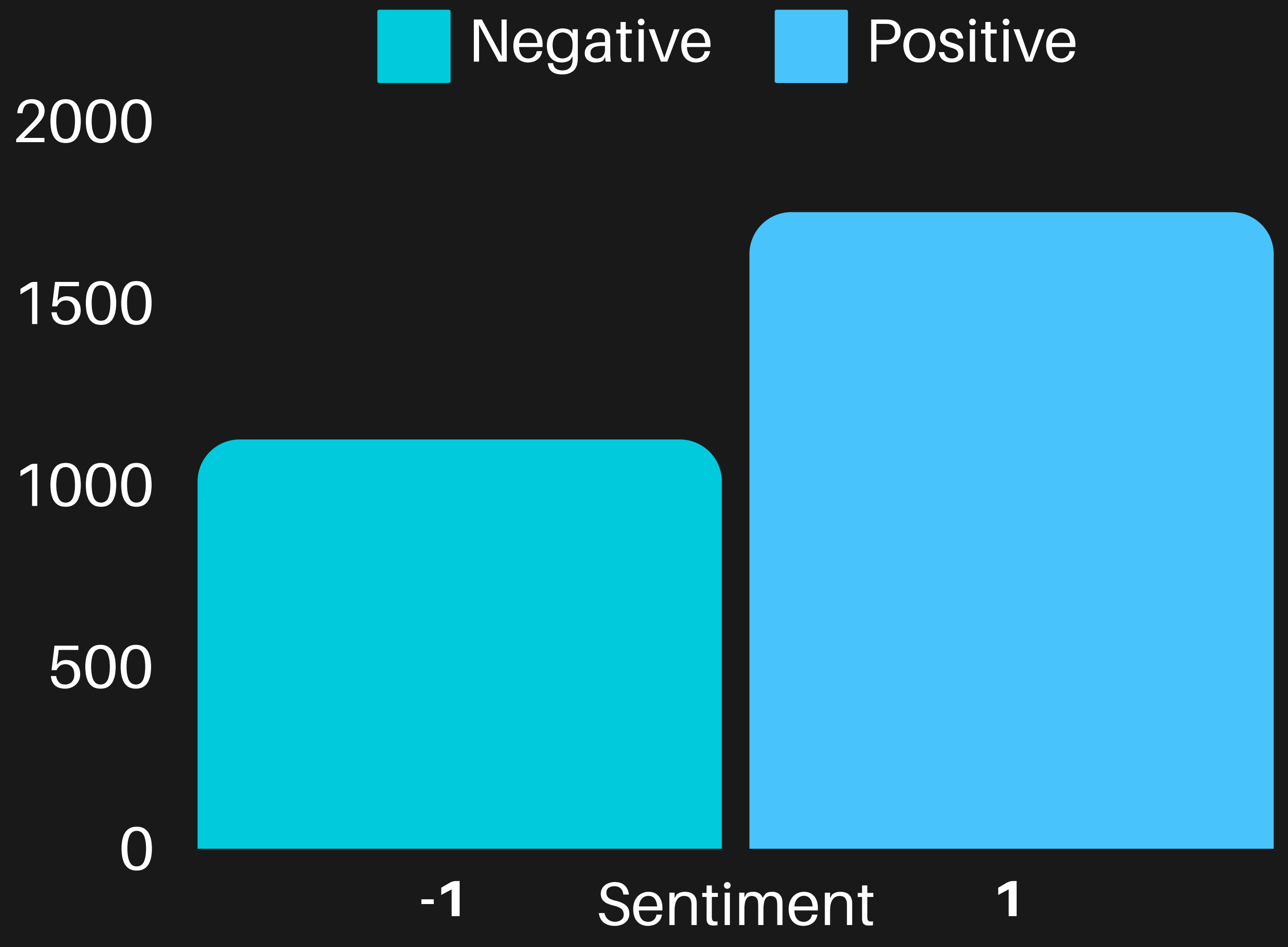


COMPLETE DATASET LABEL DISTRIBUTION





TRAINING SET LABEL DISTRIBUTION



TEST SET LABEL DISTRIBUTION

IMPLEMENTATION DETAILS

- **Initialization**:
    - Description: Initialize data structures for word counts and vocabulary.
    - Implementation: Utilize **defaultdict** and set in Python for efficient storage.
- **Training (fit) Method**:
    - Description: Count word occurrences and calculate probabilities.
    - Implementation: Loop over training data, update counts, and apply add-1 smoothing.
- **Prediction Method**:
    - Description: Predict labels for test documents and calculate class probabilities.
    - Implementation: Iterate over test documents, update scores, and determine the predicted label.
- **Prior Probability Calculation**:
    - Description: Calculate prior probabilities for each class.
    - Implementation: Compute class prior probabilities based on document counts.

IMPLEMENTATION DETAILS

- Smoothing Technique:
    - Description: Use of add-1 smoothing to handle unseen words.
- Dependencies:
    - Description: Reliance on Python's defaultdict and numpy library.
- Parameters:
    - Description: Adjustment of the smoothing parameter alpha.
- Performance:
    - Description: Scalability and efficiency for large datasets and text classification tasks.

METRICS

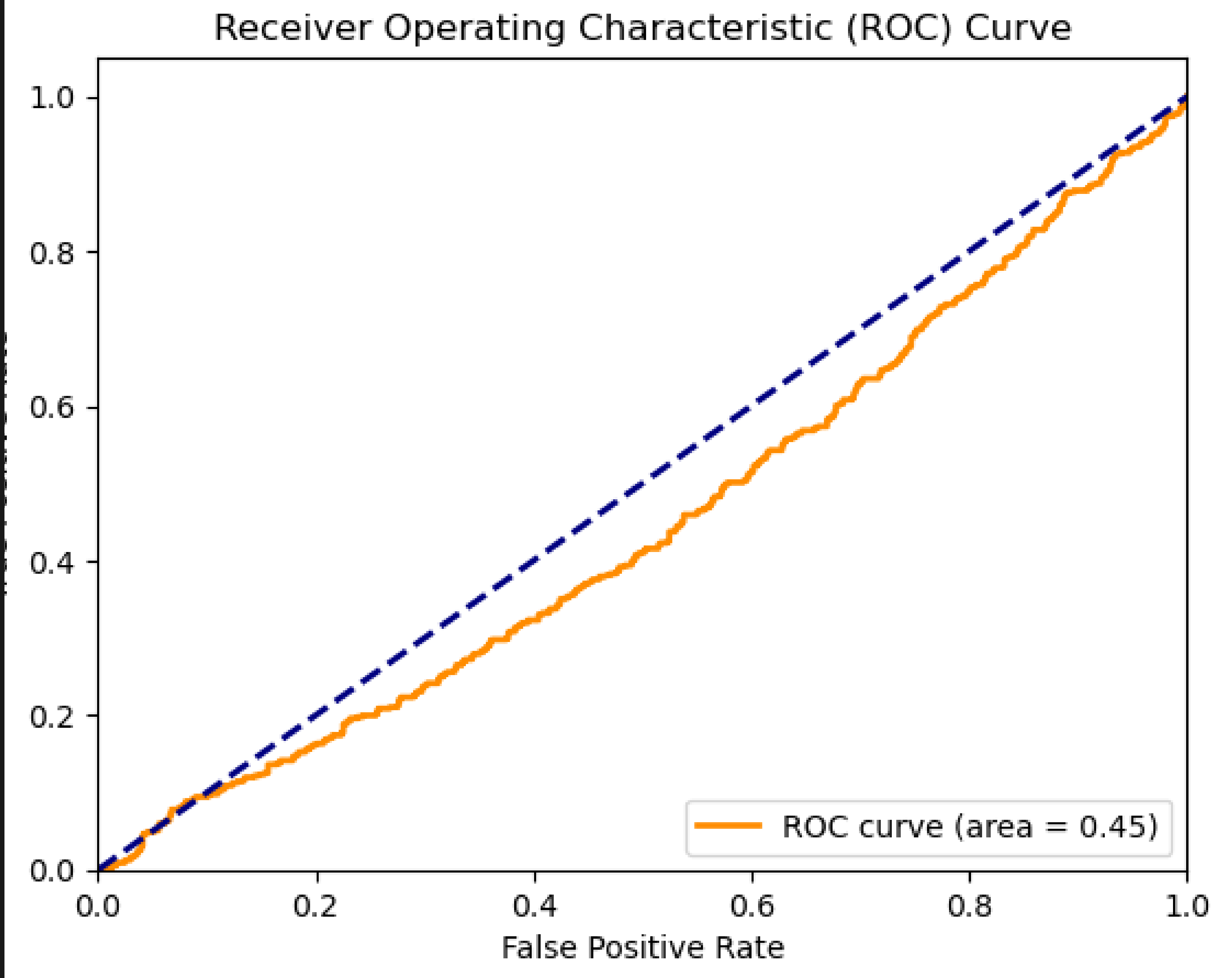
MODEL 1

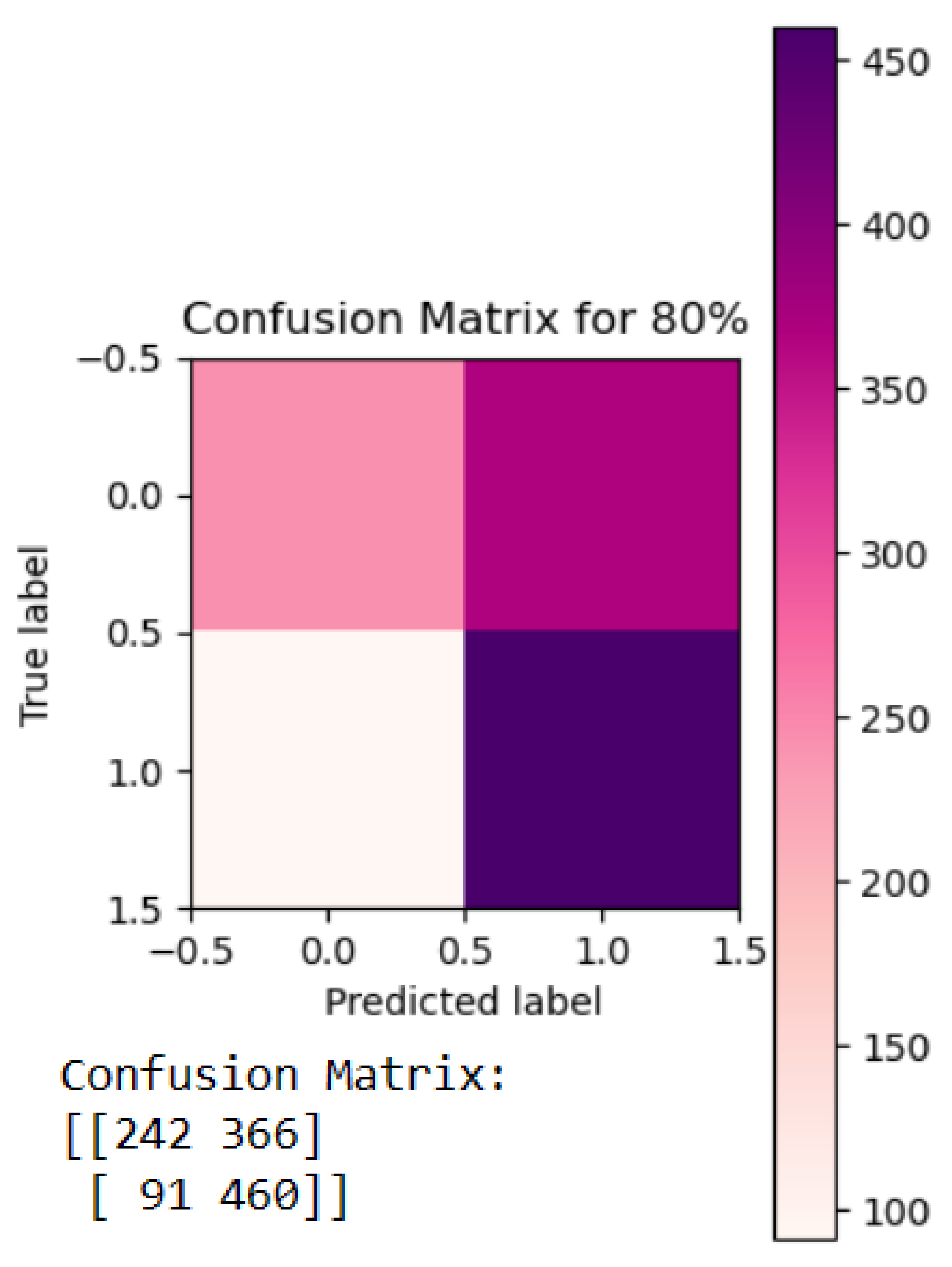
```
TRAINING SIZE: 80 %
TRAINING CLASSIFIER...
TESTING CLASSIFIER...
TEST RESULTS / METRICS:
NUMBER OF TRUE POSITIVES: 460
NUMBER OF TRUE NEGATIVES: 242
NUMBER OF FALSE POSITIVES: 366
NUMBER OF FALSE NEGATIVES: 91
SENSITIVITY (RECALL):
0.8348457350272233
SPECIFICITY: 0.3980263157894737
PRECISION: 0.5569007263922519
NEGATIVE PREDICTIVE VALUE:
0.7267267267267268
ACCURACY: 0.6056945642795514
F-SCORE: 0.6681190994916486
```

MODEL 2

```
TRAINING SIZE: 60 %
TRAINING CLASSIFIER...
TESTING CLASSIFIER...
TEST RESULTS / METRICS:
NUMBER OF TRUE POSITIVES: 447
NUMBER OF TRUE NEGATIVES: 222
NUMBER OF FALSE POSITIVES: 386
NUMBER OF FALSE NEGATIVES: 104
SENSITIVITY (RECALL):
0.8112522686025408
SPECIFICITY: 0.3651315789473684
PRECISION: 0.5366146458583433
NEGATIVE PREDICTIVE VALUE:
0.6809815950920245
ACCURACY: 0.5772217428817946
F-SCORE: 0.6459537572254335
```

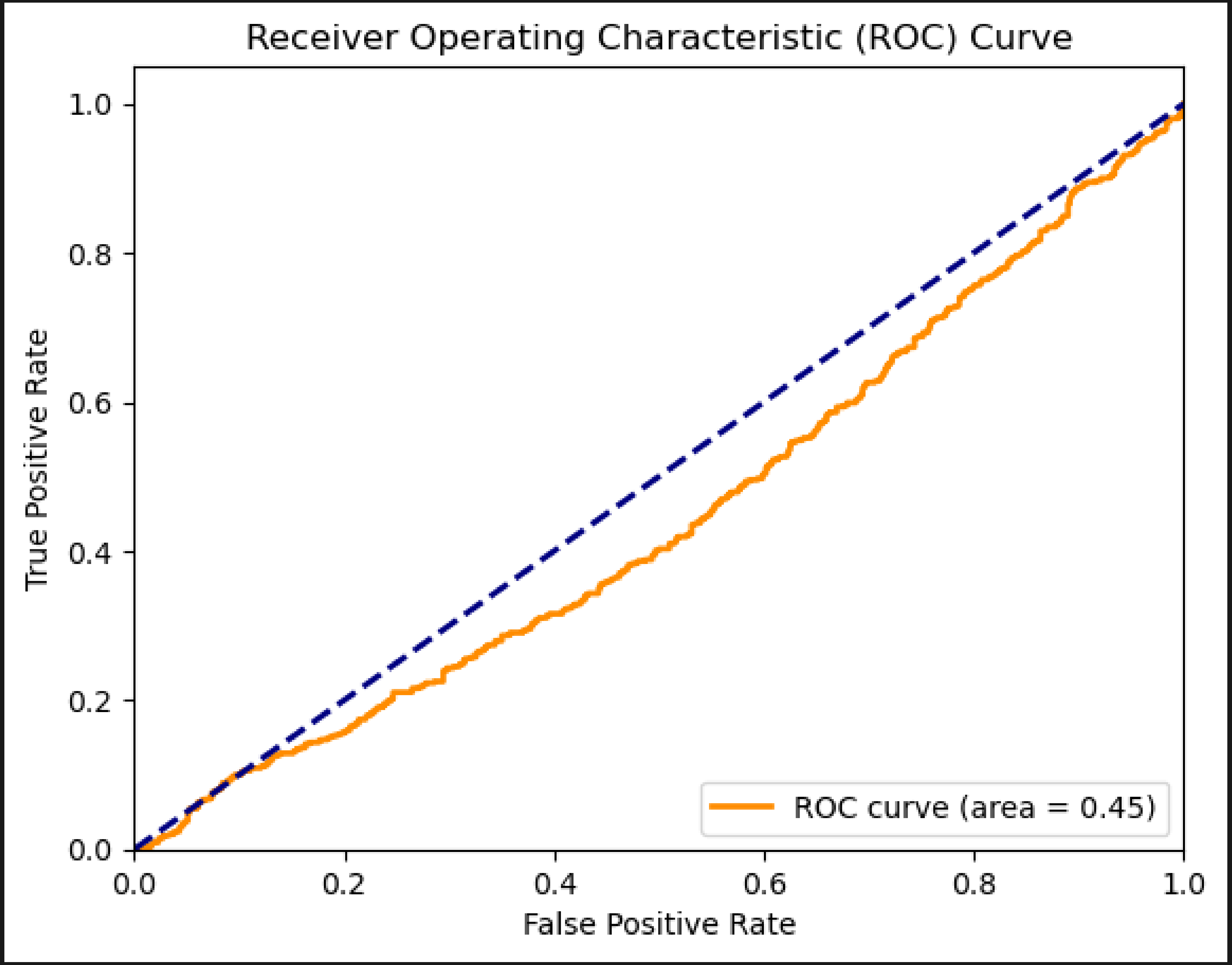
EVALUATION OF MODEL 1

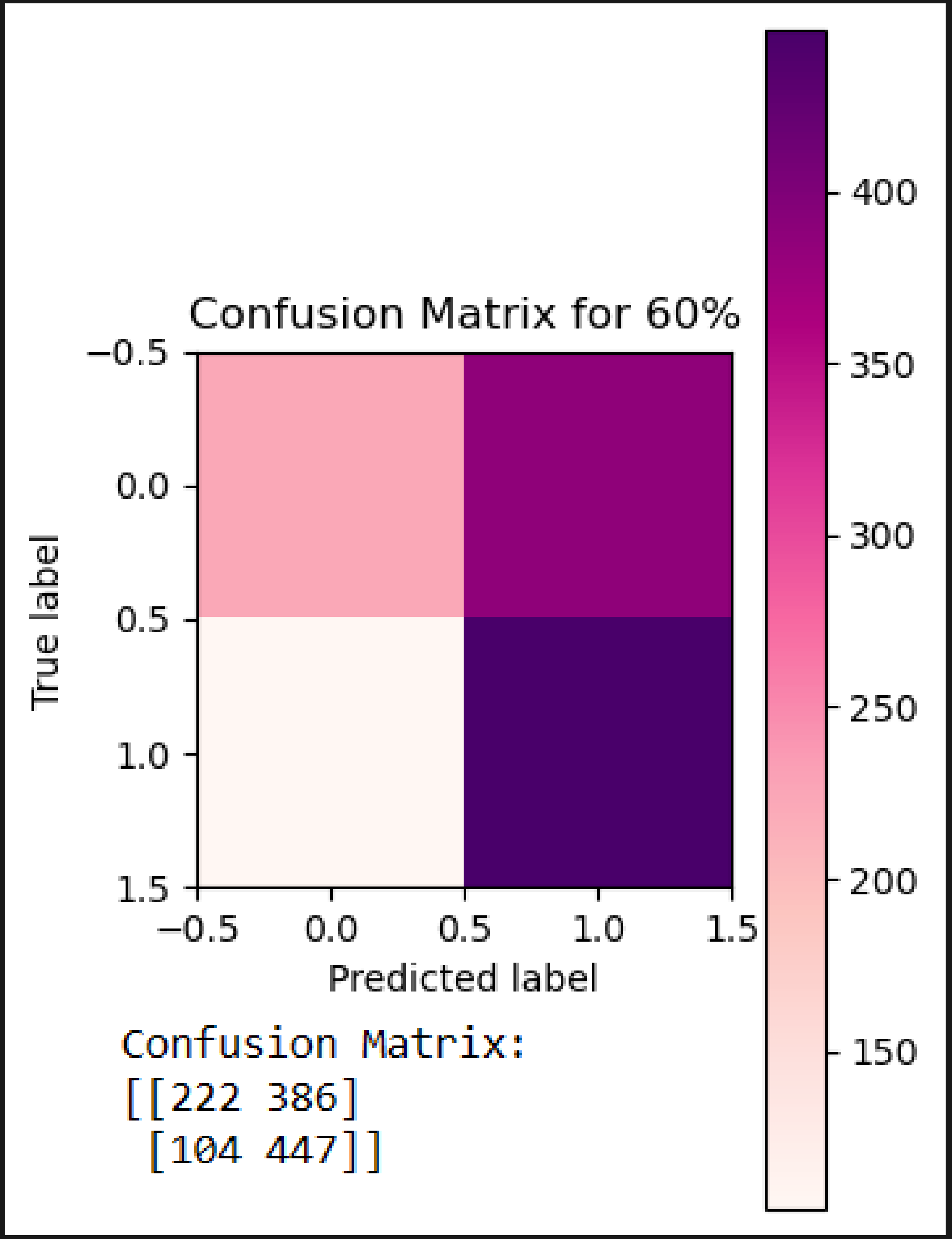




```
Confusion Matrix:
[[242 366]
 [ 91 460]]
```

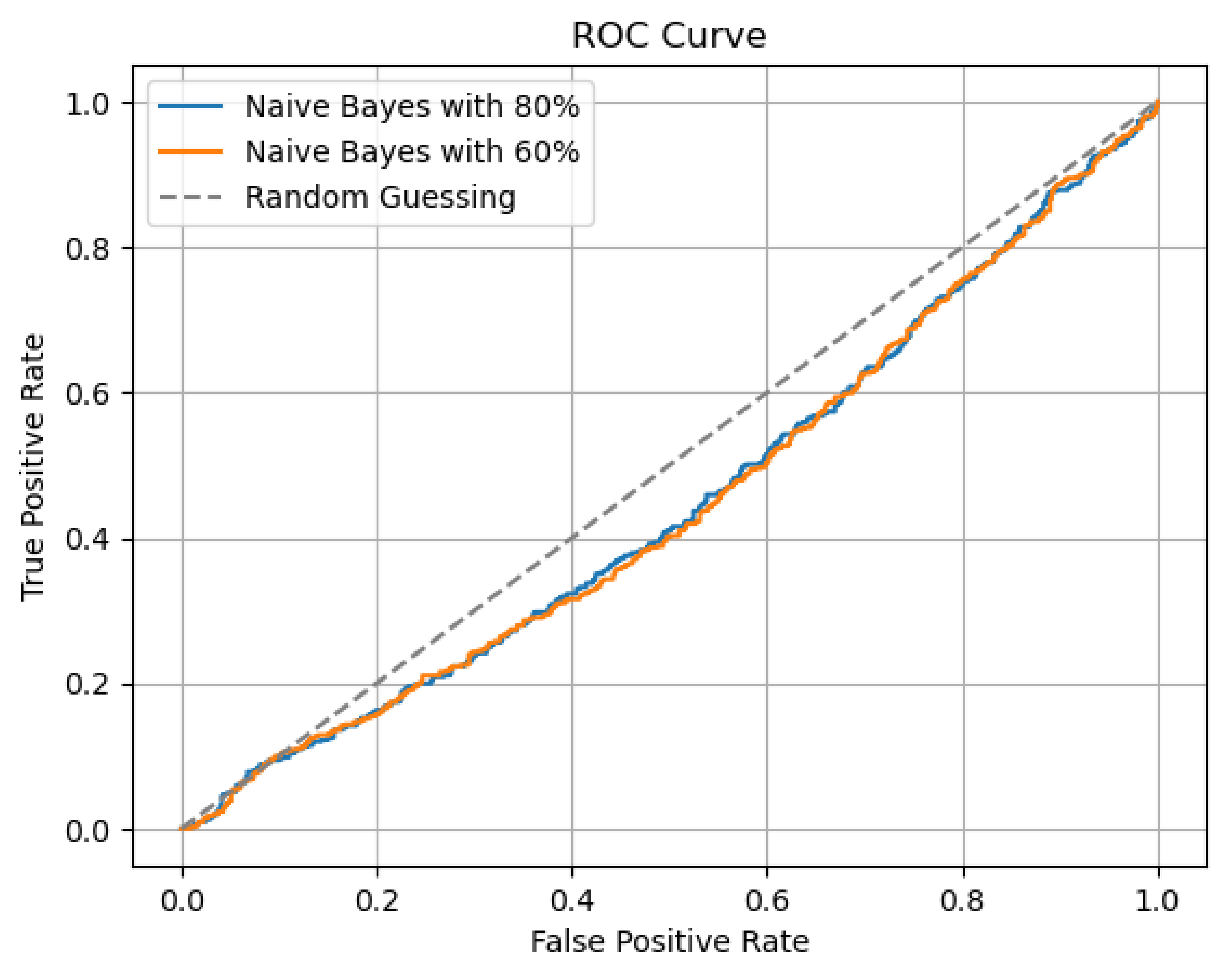
EVALUATION OF MODEL 2





```
Confusion Matrix:
[[222 386]
 [104 447]]
```

ROC COMPARISON OF 1 & 2



SUMMARY

CHALLENGES AND ISSUES :

- POTENTIAL CHALLENGE REGARDING EFFICIENCY,ESPECIALLY FOR LARGE DATASETS OR VOCABULARIES, DUE TO ITERATIVE COMPUTATIONS.
- BIASED PREDICTIONS: THE CLASSIFIER MAY EXHIBIT A BIAS TOWARDS PREDICTING THE MAJORITY CLASS, RESULTING IN LOWER ACCURACY, PRECISION, AND RECALL FOR MINORITY CLASSES.

IMPROVEMENTS :

- ADVANCED LANGUAGE MODELS
  - EXPLORE INTEGRATION OF N-GRAMS OR WORD EMBEDDINGS.
- SMOOTHING TECHNIQUES
  - INCLUDE ADD-K SMOOTHING OR GOOD-TURING SMOOTHING OPTIONS.
- EFFICIENCY OPTIMIZATION
  - UTILIZE VECTORIZED OPERATIONS OR SPARSE MATRIX REPRESENTATIONS.