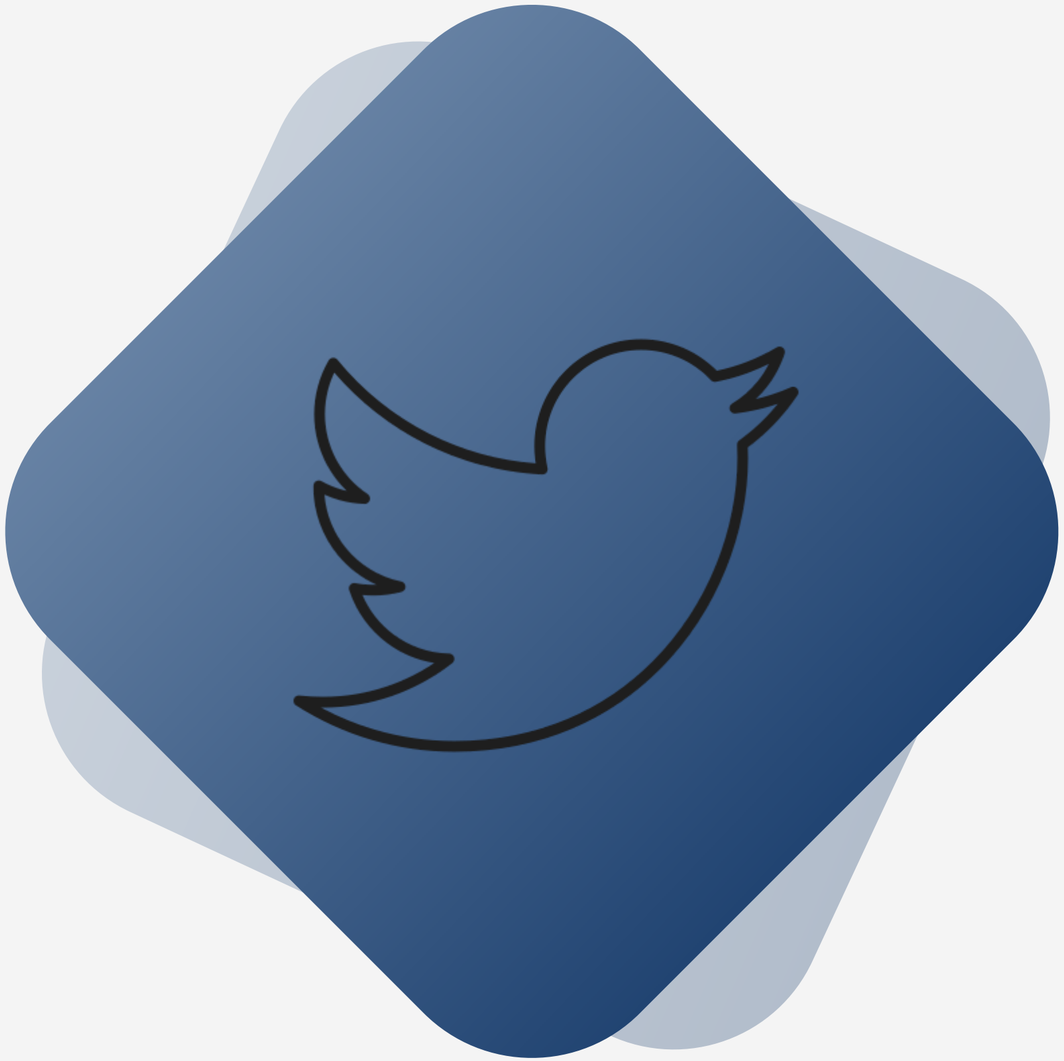
Sentiment Analysis

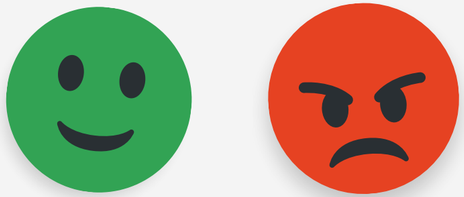
Group 10

İrem Aydın
Aksanur Konuk
Elife Kocabey
Zeynep Yılmaz
İrem Kıranmezar

Abstract

- Explores sentiment analysis on tweets.
- Uses Sentiment140 dataset with 1.6M tweets.
- Focus on preprocessing, feature engineering, and model evaluation.
- Subset of 10,000 balanced tweets used for analysis





Data Preprocessing

Read → **Clean** → **Tokenization** → **Future Engineering**

head(), tail(), info()
Check null and duplicate value: isnull().sum() and duplicated.sum()

Retaining Relevant Columns : Target and Text
Text Preprocessing:
-Convert lowercase
-Remove some patterns (urls, @username etc.)

-Tokenization for splits each tweet (NLTKlibrary)
-Remove stop words, such as "the" "is" and "and"

-Lemmatization
-Whitespace, Rare Word, Duplicate Row Removal
-Target Variable Adjustment (Target values from 4 to 1)
-Feature Vectorization(TF-IDF)

```
Model: DECISION TREE (GINI)
Accuracy Score: 0.67553
Confusion MAtrix:
[[702 268]
 [354 593]]
Classification Report:
              precision    recall  f1-score   support

           0       0.66      0.72      0.69       970
           1       0.69      0.63      0.66       947

    accuracy                           0.68      1917
   macro avg       0.68      0.67      0.67      1917
weighted avg       0.68      0.68      0.67      1917

Cross-Validation Scores:
fit_time          10.85459
score_time         0.00855
test_accuracy      0.66971
test_precision     0.68103
test_recall        0.63440
test_f1            0.65648
dtype: object
```

Decision Tree - Gini

Accuracy = 0.67553

```
Model: DECISION TREE (GAIN)
Accuracy Score: 0.67345
Confusion MAtrix:
[[702 268]
 [358 589]]
Classification Report:
              precision    recall  f1-score   support

           0       0.66      0.72      0.69       970
           1       0.69      0.62      0.65       947

    accuracy                           0.67      1917
   macro avg       0.67      0.67      0.67      1917
weighted avg       0.67      0.67      0.67      1917

Cross-Validation Scores:
fit_time          10.32016
score_time         0.00578
test_accuracy      0.66997
test_precision     0.67687
test_recall        0.64591
test_f1            0.66051
dtype: object
```

Decision Tree - Gain

Accuracy = 0.67345

```
Model: NAIVE BAYES
Accuracy Score: 0.59416
Confusion MAtrix:
[[766 204]
 [574 373]]
Classification Report:
              precision    recall  f1-score   support

           0       0.57      0.79      0.66       970
           1       0.65      0.39      0.49       947

    accuracy                           0.59      1917
   macro avg       0.61      0.59      0.58      1917
weighted avg       0.61      0.59      0.58      1917

Cross-Validation Scores:
fit_time          0.16465
score_time        0.01901
test_accuracy     0.59105
test_precision    0.65253
test_recall       0.38524
test_f1           0.48423
dtype: object
```

Naive Bayes

Accuracy = 0.59416

```
Model: K-NEAREST NEIGHBORS (KNN)
Accuracy Score: 0.67919
Confusion MAtrix:
[[638 332]
 [283 664]]
Classification Report:
              precision    recall  f1-score   support

           0       0.69      0.66      0.67       970
           1       0.67      0.70      0.68       947

    accuracy                           0.68      1917
   macro avg       0.68      0.68      0.68      1917
weighted avg       0.68      0.68      0.68      1917

Cross-Validation Scores:
fit_time          0.03492
score_time        0.21386
test_accuracy     0.64310
test_precision    0.64797
test_recall       0.64590
test_f1           0.63807
dtype: object
```

K-Nearest Neighbor (KNN)

Accuracy = 0.67919

```
Model: LOGISTIC REGRESSION
Accuracy Score: 0.75483
Confusion MAtrix:
[[702 268]
 [202 745]]
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.72      0.75       970
           1       0.74      0.79      0.76       947

    accuracy                           0.75      1917
   macro avg       0.76      0.76      0.75      1917
weighted avg       0.76      0.75      0.75      1917

Cross-Validation Scores:
fit_time          0.43347
score_time        0.00405
test_accuracy     0.75372
test_precision    0.73780
test_recall       0.78513
test_f1           0.76054
dtype: object
```

Logistic Regression

Accuracy = 0.75483

```
Model: RANDOM FOREST CLASSIFIER
Accuracy Score: 0.72822
Confusion MAtrix:
[[716 254]
 [267 680]]
Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.74      0.73       970
           1       0.73      0.72      0.72       947

    accuracy                           0.73      1917
   macro avg       0.73      0.73      0.73      1917
weighted avg       0.73      0.73      0.73      1917

Cross-Validation Scores:
fit_time         11.85068
score_time        0.05030
test_accuracy     0.73650
test_precision    0.73070
test_recall       0.74719
test_f1           0.73864
dtype: object
```

Random Forest Classifier

Accuracy = 0.72822

```
Model: SUPPORT VECTOR CLASSIFIER
Accuracy Score: 0.75535
Confusion MAtrix:
[[698 272]
 [197 750]]
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.72      0.75       970
           1       0.73      0.79      0.76       947

    accuracy                           0.76      1917
   macro avg       0.76      0.76      0.76      1917
weighted avg       0.76      0.76      0.76      1917

Cross-Validation Scores:
fit_time          45.16633
score_time         6.82055
test_accuracy      0.75515
test_precision     0.73721
test_recall        0.79089
test_f1            0.76298
dtype: object
```

Support Vector Classifier

Accuracy = 0.75535

```
Model: ARTIFICIAL NEURAL NETWORK (ANN)
Accuracy Score: 0.67971
Confusion MAtrix:
[[671 299]
 [315 632]]
Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.69      0.69       970
           1       0.68      0.67      0.67       947

    accuracy                           0.68      1917
   macro avg       0.68      0.68      0.68      1917
weighted avg       0.68      0.68      0.68      1917

Cross-Validation Scores:
fit_time          36.27629
score_time         0.00703
test_accuracy      0.68641
test_precision     0.68825
test_recall        0.67782
test_f1            0.68277
dtype: object
```

Artificial Neural Network (ANN)

Accuracy = 0.67971

Optimization

Objective

Improve Logistic Regression performance via hyperparameter tuning.

Method

Used RandomizedSearchCV with 100 iterations and 5-fold cross-validation.

Parameters Tuned

- C: 0.1, 1, 10
- Solver: 'liblinear', 'saga'
- Max Iterations: 100, 200, 300

Results

- Best parameters: solver='saga', max_iter=100, C=1
- Retrained and evaluated on the test set using accuracy, classification report, and confusion matrix.
- Outcome: Improved prediction accuracy with optimized parameters.

Optimization

```
Fitting 5 folds for each of 18 candidates, totalling 90 fits
Best Parameters: {'solver': 'saga', 'max_iter': 300, 'C': 1}
Best Score: 0.7568953824564957
Accuracy:  0.7552
Classification Report:

              precision    recall  f1-score   support

           0       0.78      0.72      0.75       970
           1       0.74      0.78      0.76       947

    accuracy                           0.75      1917
   macro avg       0.76      0.75      0.75      1917
weighted avg       0.76      0.75      0.75      1917
```

Libraries

NLTK

kaggle

scikit learn

pandas

seaborn

matplotlib

Thank you...