Sentiment Analysis of Movie Reviews

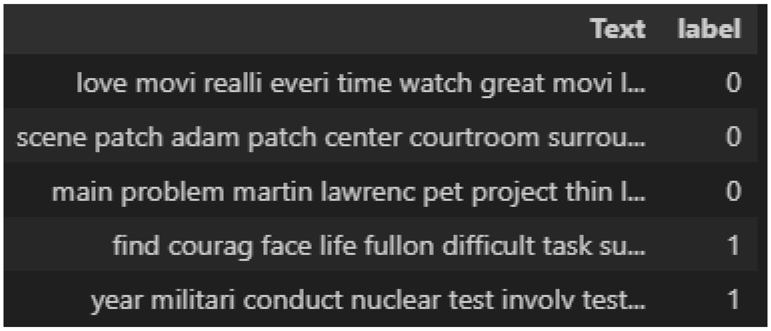
Team# 4

ID	Name
2022170385	محمد متولي عبدالحميد عوض محمد
2022170375	محمد عادل علي حسن
2022170389	محمد منير تاج الدين منصور
2022170373	محمد طارق الحسين محمد منصور العراقي
2022170456	مينا باسم نادي

Preprocessing

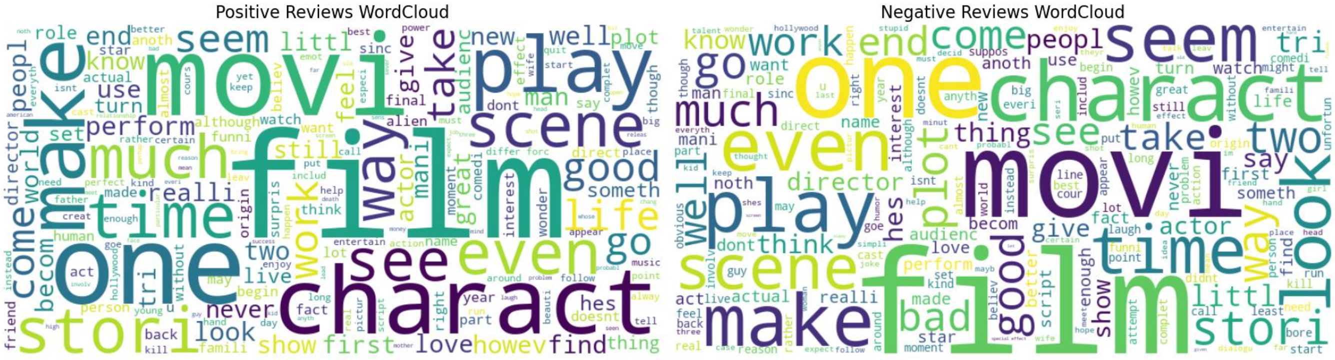
- **Applied common techniques:**
  - ✓ Lowercasing
  - ✓ Punctuation removal
  - ✓ Tokenization
  - ✓ Stopwords removal
  - ✓ SnowballStemmer (after testing lemmatizer & various stemmers)

- **Now, our preprocessed data looks like:**

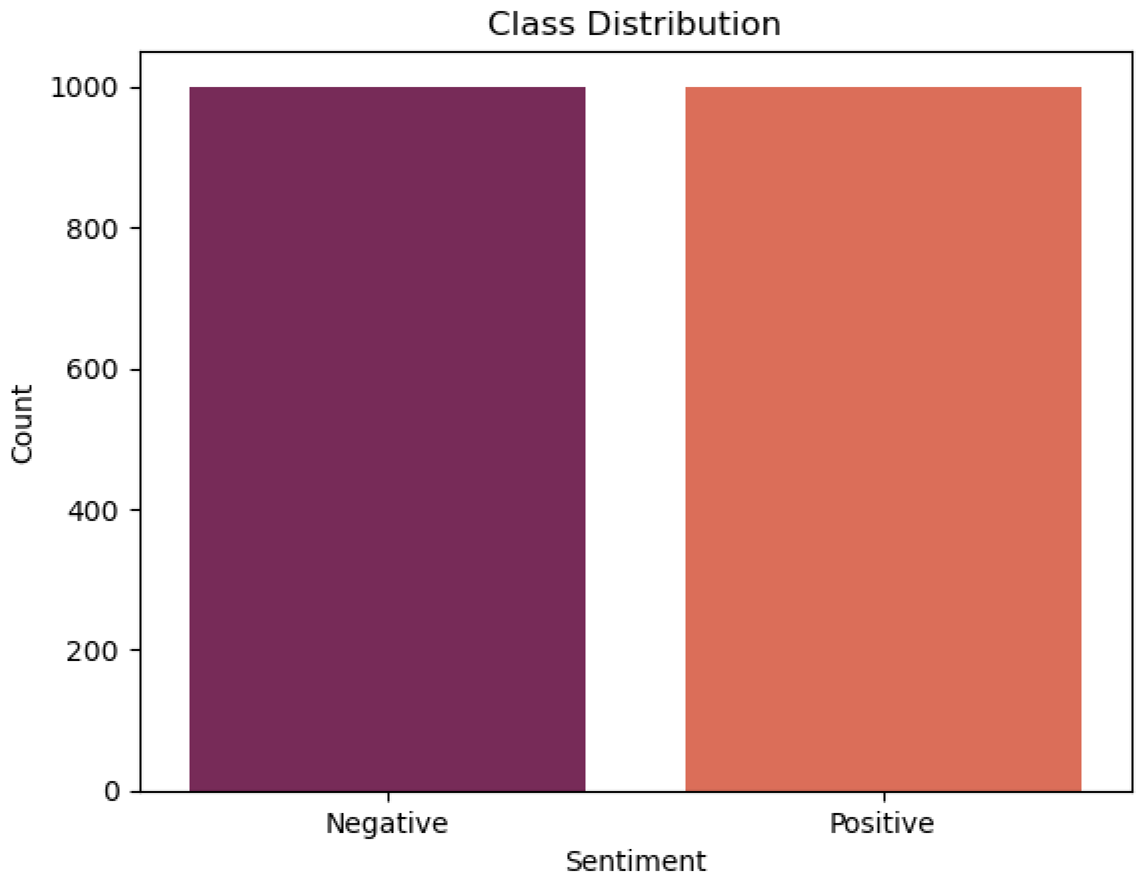


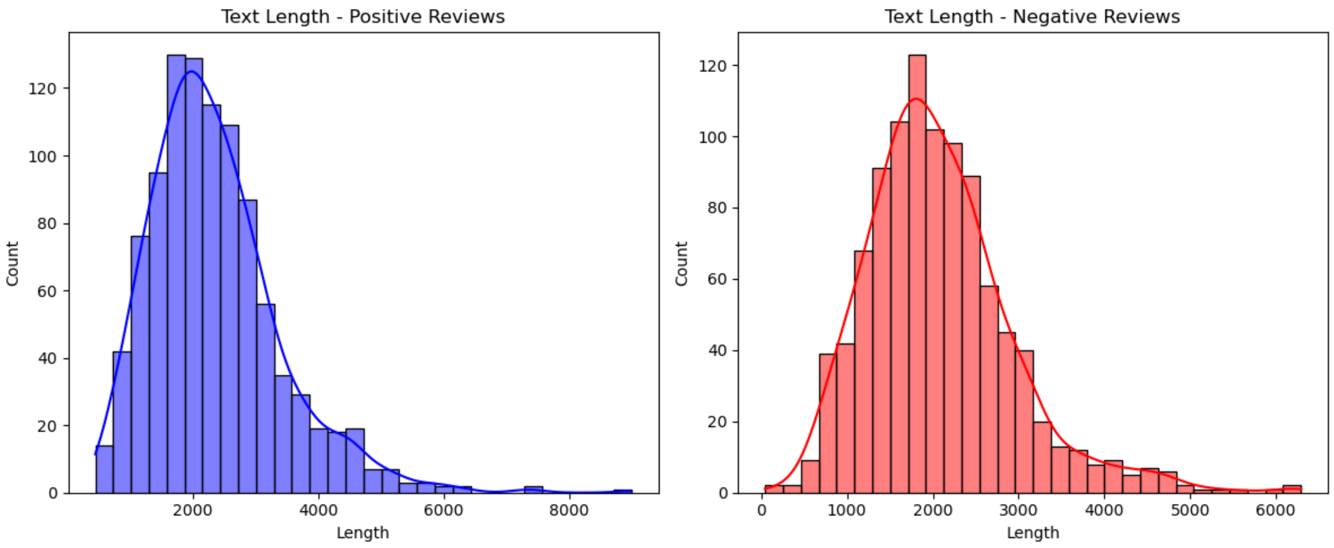
Text	label
love movi realli everi time watch great movi l...	0
scene patch adam patch center courtroom surrou...	0
main problem martin lawrenc pet project thin l...	0
find courag face life fullon difficult task su...	1
year militari conduct nuclear test involv test...	1

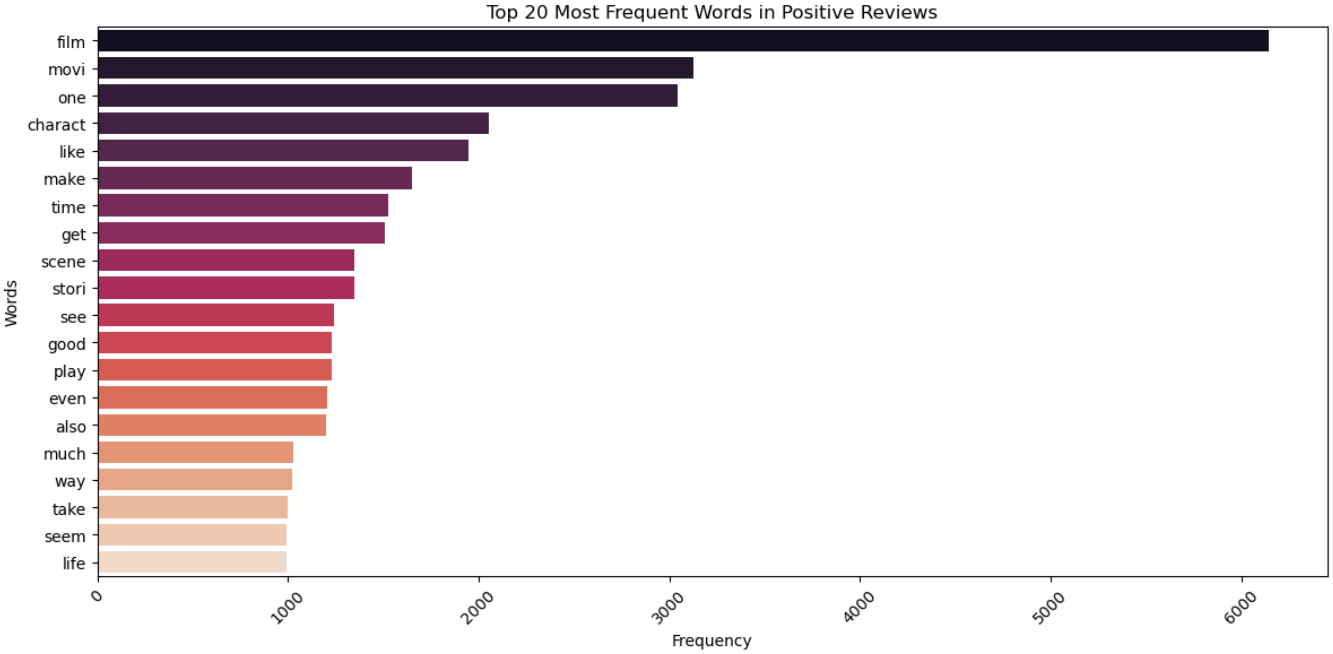
EDA

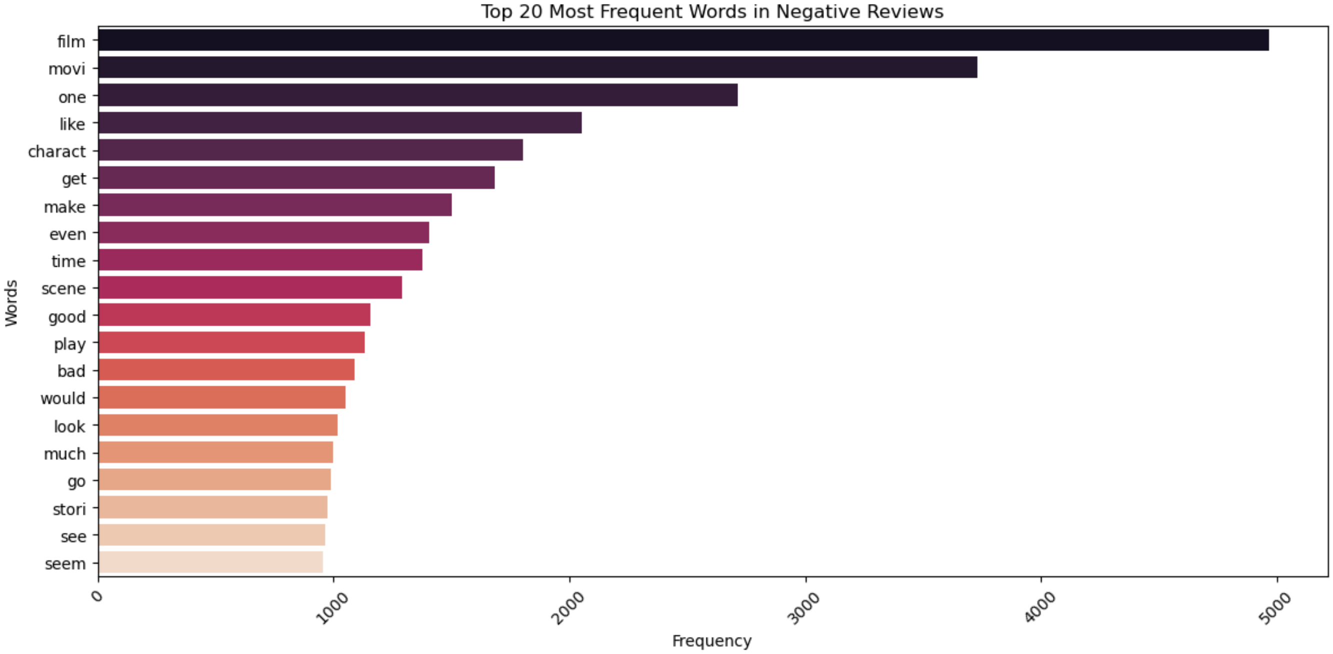

Positive Reviews WordCloud

Negative Reviews WordCloud


Class Distribution
Count
0
200
400
600
800
1000
Negative
Positive
Sentiment


Text Length - Positive Reviews
Count
0
20
40
60
80
100
120
2000
4000
6000
8000
Length
Text Length - Negative Reviews
Count
0
20
40
60
80
100
120
0
1000
2000
3000
4000
5000
6000
Length


Top 20 Most Frequent Words in Positive Reviews
film
movi
one
charact
like
make
time
get
scene
stori
see
good
play
even
also
much
way
take
seem
life
Words
0
1000
2000
3000
4000
5000
6000
Frequency


Top 20 Most Frequent Words in Negative Reviews
film
movi
one
like
charact
get
make
even
time
scene
good
play
bad
would
look
much
go
stori
see
seem
Words
0
1000
2000
3000
4000
5000
Frequency

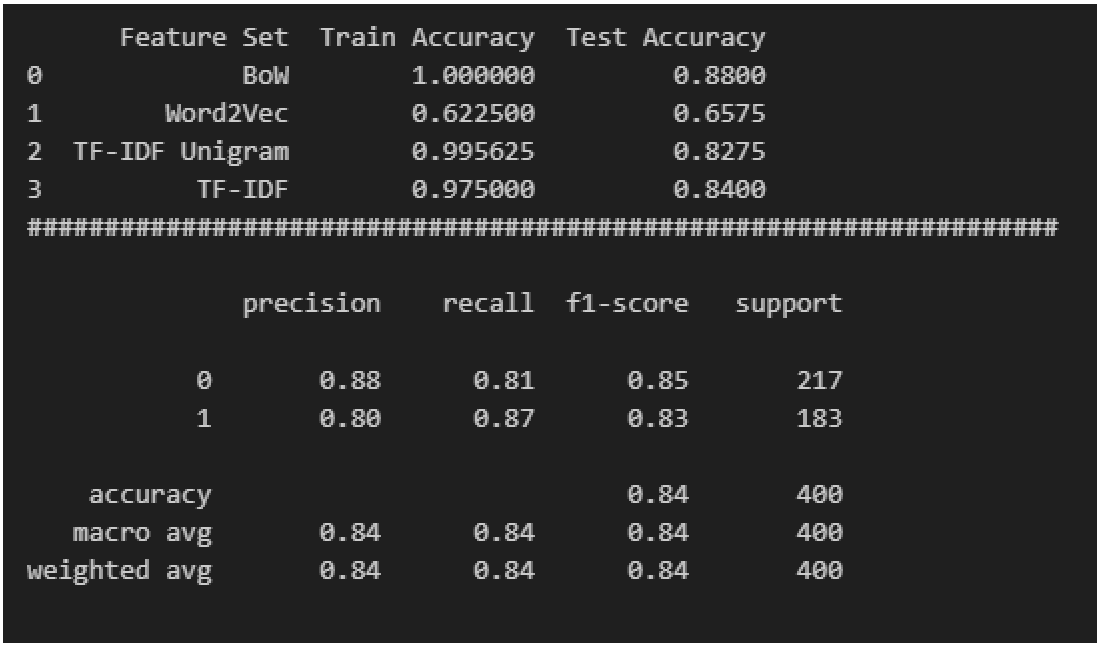
Feature Extraction

- **Separately applied:**
  - ✓ BOW
  - ✓ TF-IDF (Unigrams)
  - ✓ TF-IDF (Unigrams + Bigrams)
  - ✓ Word2Vec Embeddings

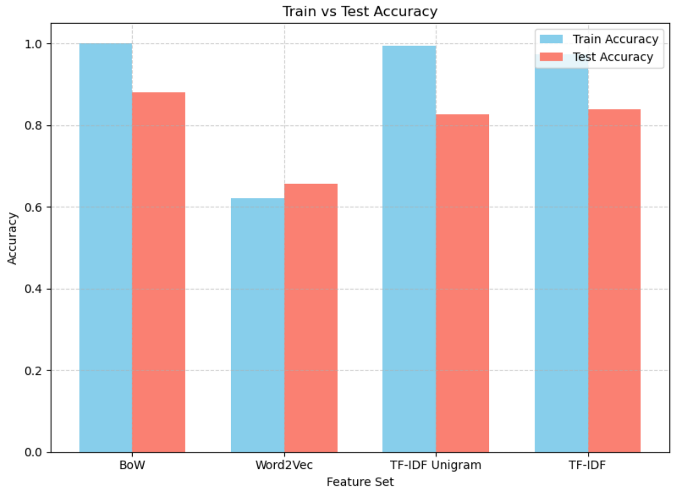
- Evaluated every model with every feature extraction technique to find the best technique combination between Feature Extraction and Evaluation.

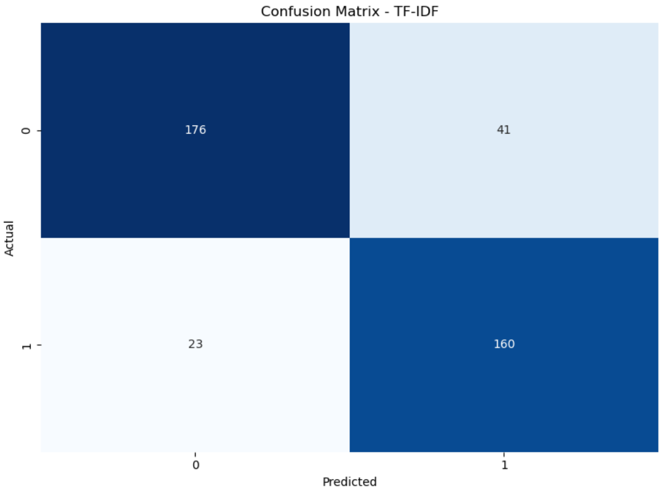
Model Training/Testing

Logistic Regression:

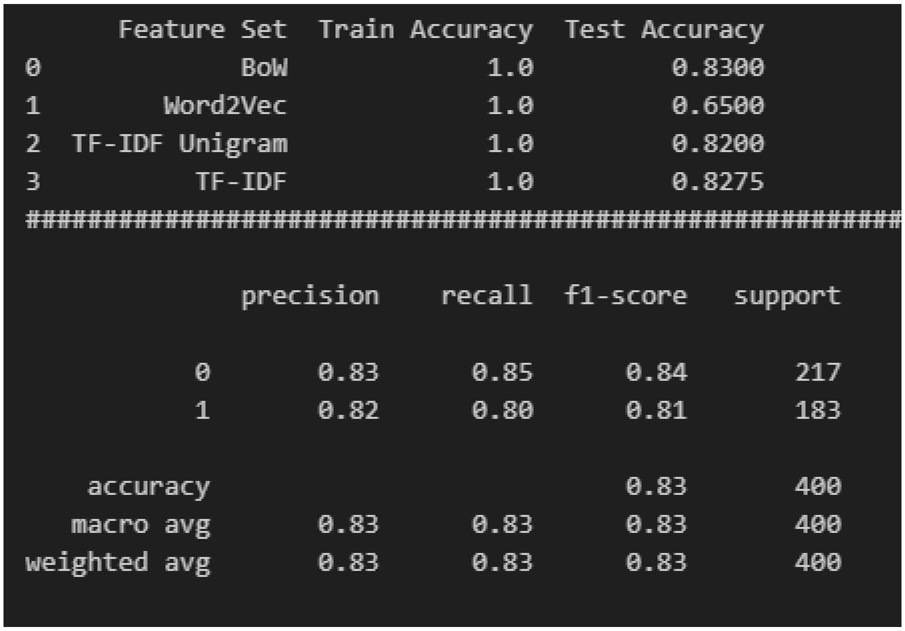


```
      Feature Set  Train Accuracy  Test Accuracy
0             BoW        1.000000         0.8800
1        Word2Vec        0.622500         0.6575
2  TF-IDF Unigram        0.995625         0.8275
3          TF-IDF        0.975000         0.8400
###################################################################

              precision    recall  f1-score   support

           0       0.88      0.81      0.85       217
           1       0.80      0.87      0.83       183

    accuracy                           0.84       400
   macro avg       0.84      0.84      0.84       400
weighted avg       0.84      0.84      0.84       400
```

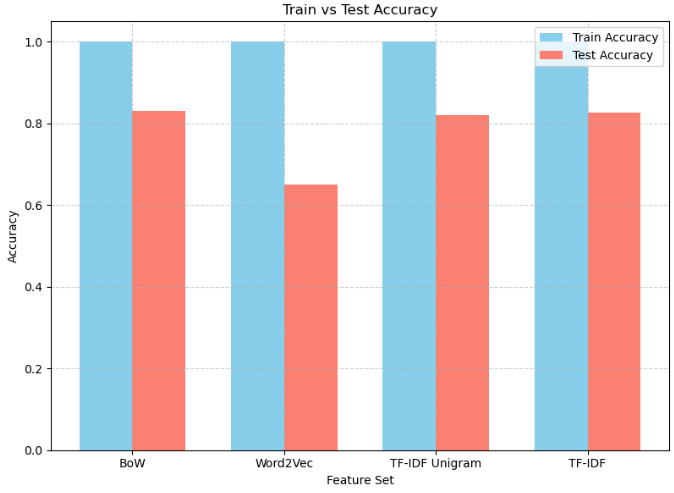


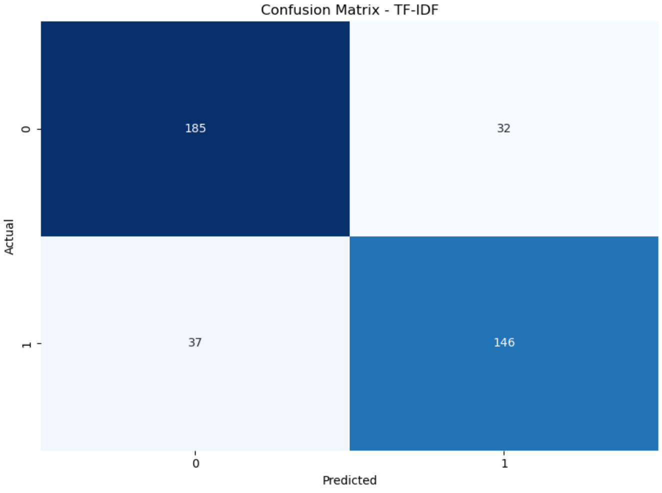


Random Forest:

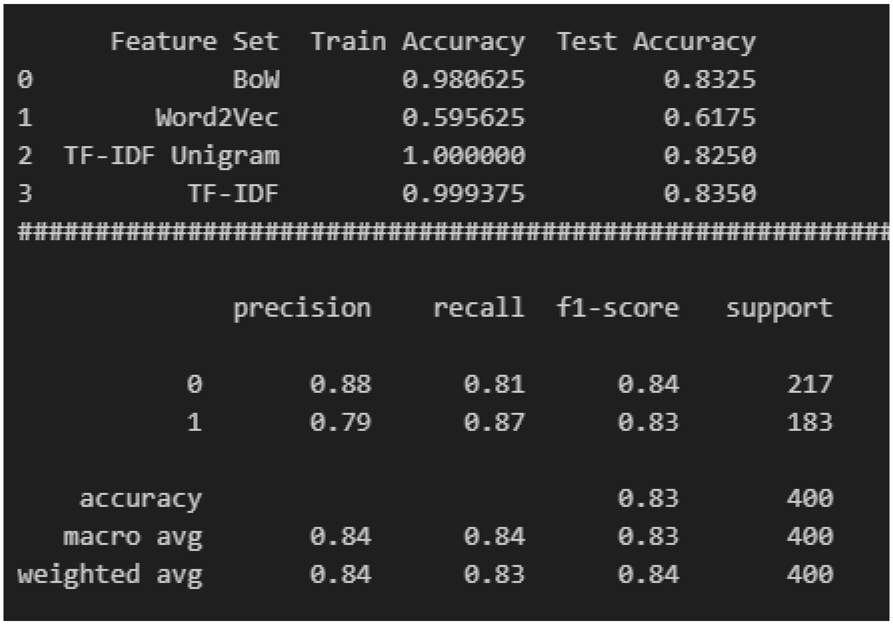


```
      Feature Set  Train Accuracy  Test Accuracy
0             BoW             1.0         0.8300
1        Word2Vec             1.0         0.6500
2  TF-IDF Unigram             1.0         0.8200
3          TF-IDF             1.0         0.8275
#########################################################

              precision    recall  f1-score   support

           0       0.83      0.85      0.84       217
           1       0.82      0.80      0.81       183

    accuracy                           0.83       400
   macro avg       0.83      0.83      0.83       400
weighted avg       0.83      0.83      0.83       400
```

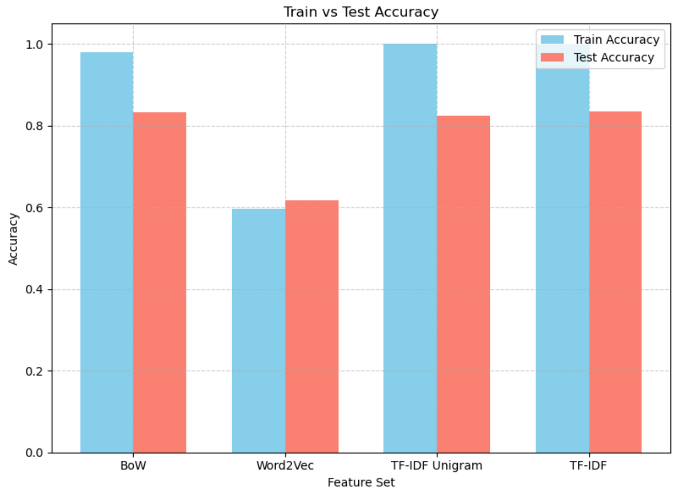


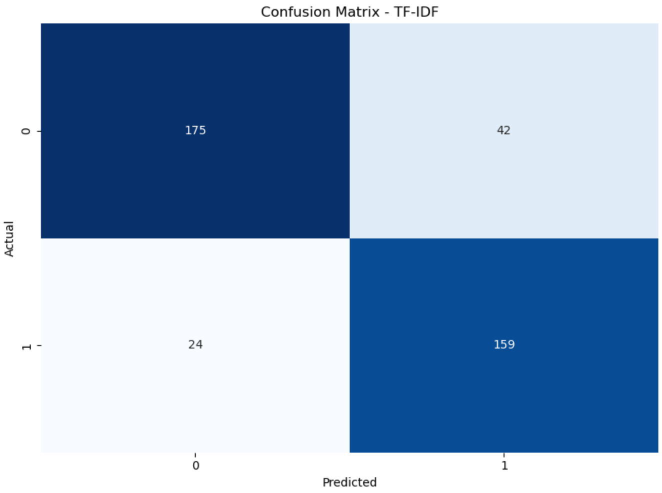


SVM:

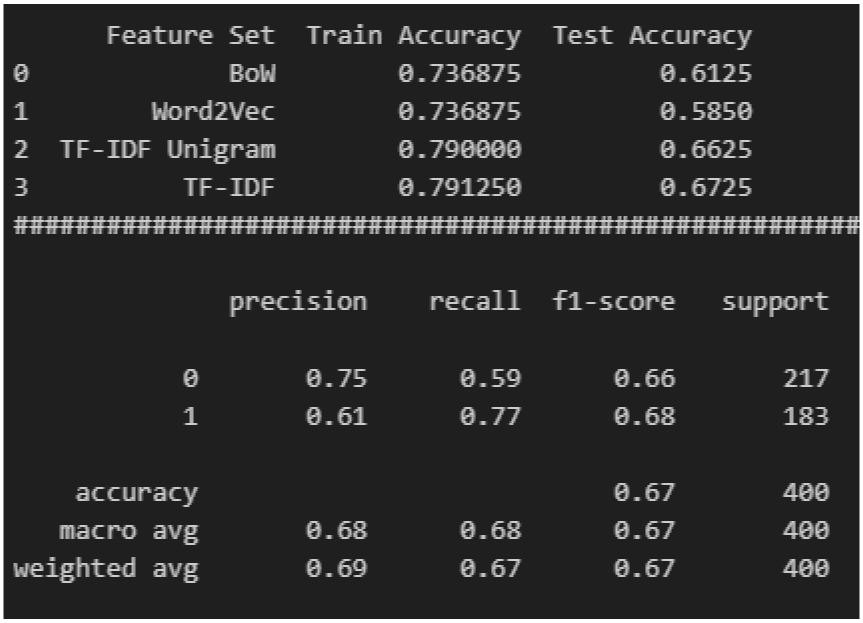


```
     Feature Set  Train Accuracy  Test Accuracy
0            BoW        0.980625         0.8325
1       Word2Vec        0.595625         0.6175
2  TF-IDF Unigram       1.000000         0.8250
3          TF-IDF       0.999375         0.8350
#########################################################

              precision    recall  f1-score   support

           0       0.88      0.81      0.84       217
           1       0.79      0.87      0.83       183

    accuracy                           0.83       400
   macro avg       0.84      0.84      0.83       400
weighted avg       0.84      0.83      0.84       400
```

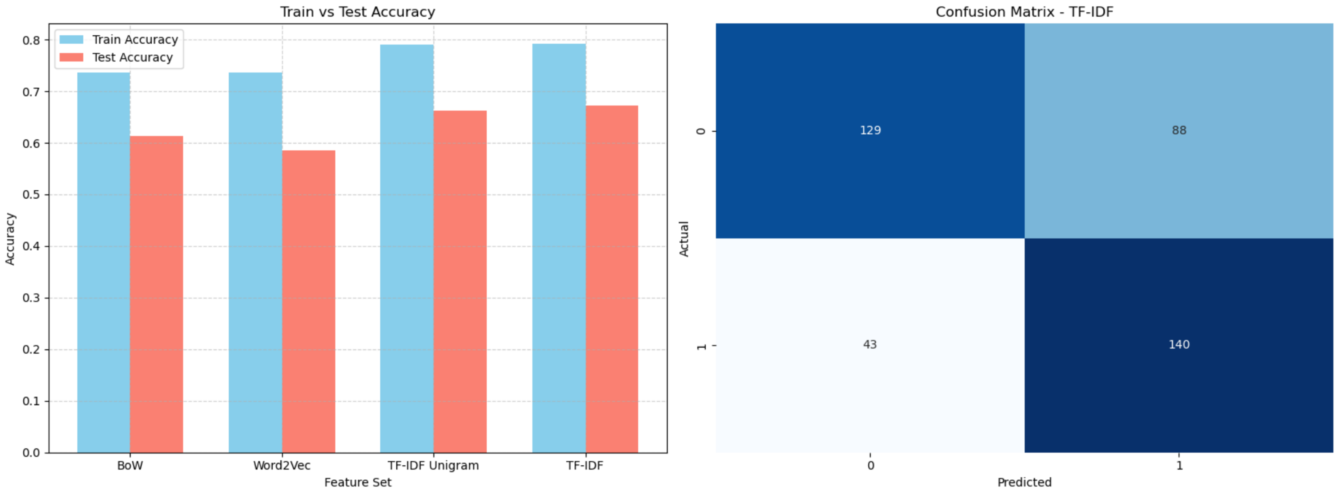




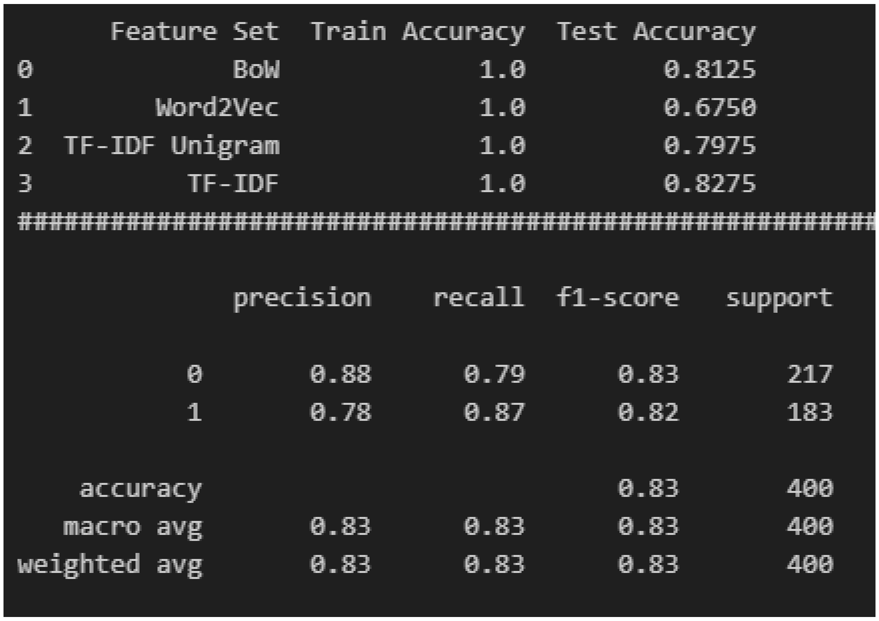
KNN:



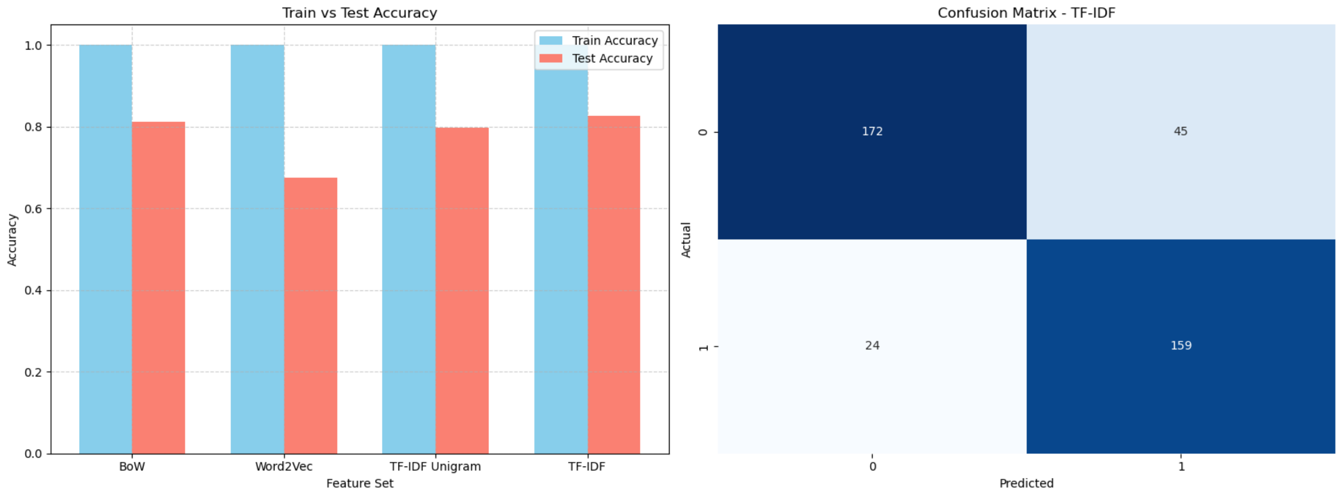
```
      Feature Set  Train Accuracy  Test Accuracy
0             BoW        0.736875         0.6125
1        Word2Vec        0.736875         0.5850
2  TF-IDF Unigram        0.790000         0.6625
3          TF-IDF        0.791250         0.6725
#######################################################

              precision    recall  f1-score   support

           0       0.75      0.59      0.66       217
           1       0.61      0.77      0.68       183

    accuracy                           0.67       400
   macro avg       0.68      0.68      0.67       400
weighted avg       0.69      0.67      0.67       400
```



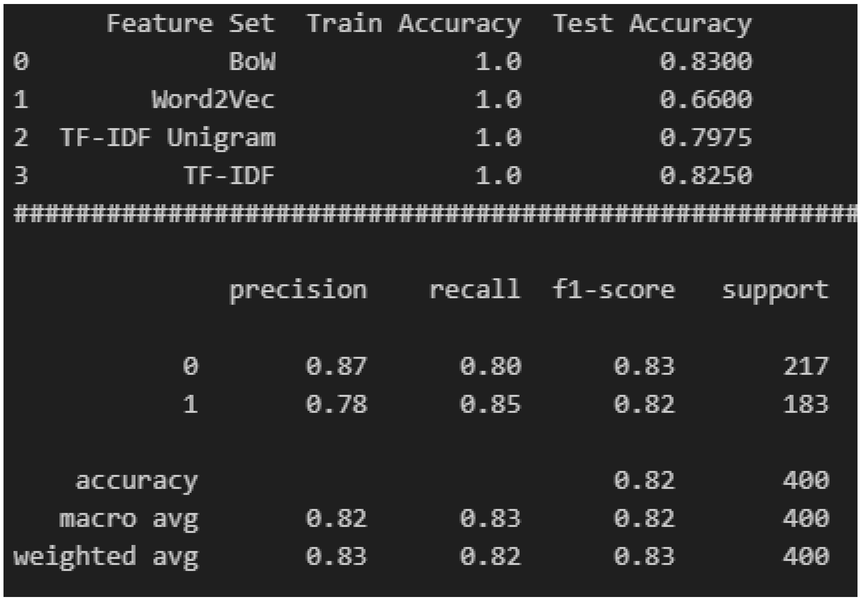
XGBoost:



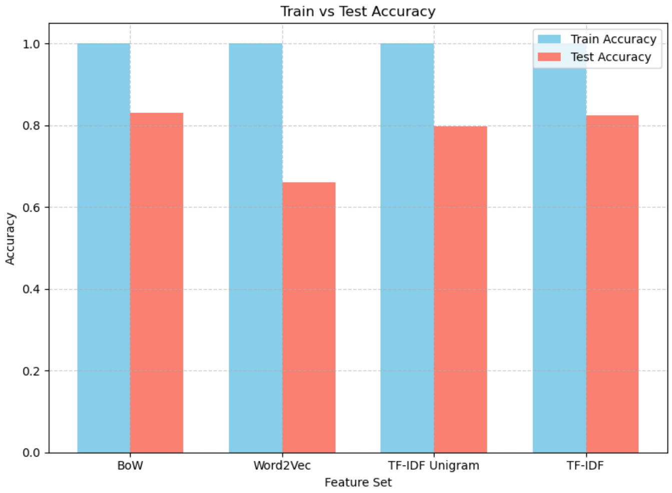
```
       Feature Set  Train Accuracy  Test Accuracy
0              BoW             1.0         0.8125
1         Word2Vec             1.0         0.6750
2   TF-IDF Unigram             1.0         0.7975
3           TF-IDF             1.0         0.8275
########################################################

              precision    recall  f1-score   support

           0       0.88      0.79      0.83       217
           1       0.78      0.87      0.82       183

    accuracy                           0.83       400
   macro avg       0.83      0.83      0.83       400
weighted avg       0.83      0.83      0.83       400
```

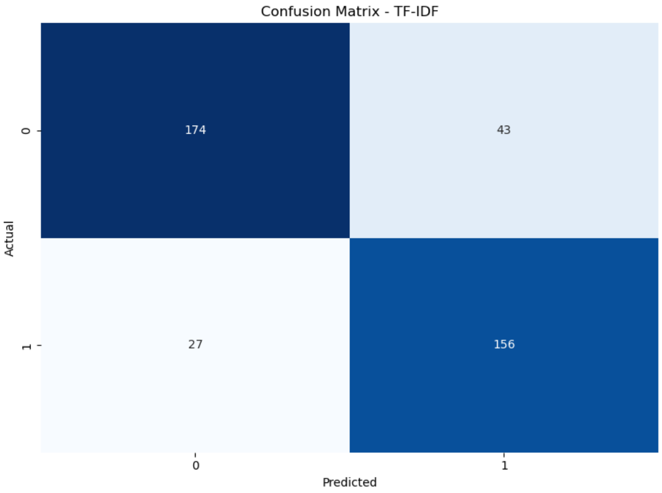


LGBM:



```
      Feature Set  Train Accuracy  Test Accuracy
0             BoW             1.0         0.8300
1        Word2Vec             1.0         0.6600
2  TF-IDF Unigram             1.0         0.7975
3          TF-IDF             1.0         0.8250
#######################################################

              precision    recall  f1-score   support

           0       0.87      0.80      0.83       217
           1       0.78      0.85      0.82       183

    accuracy                           0.82       400
   macro avg       0.82      0.83      0.82       400
weighted avg       0.83      0.82      0.83       400
```

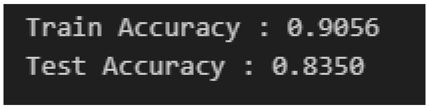




Grid Search & PCA

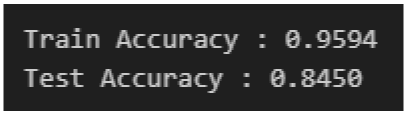
After trying both **GridSearch** and **PCA** with the models with the highest accuracy (**SVM**, **Logistic**), we took the best parameters and used them in the two models for prediction.

Logistic:



```
Train Accuracy : 0.9056
Test Accuracy : 0.8350
```

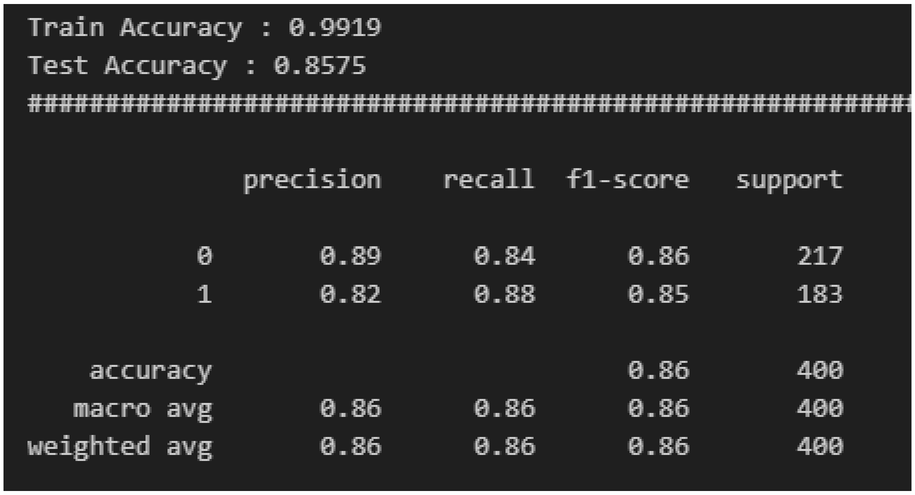
SVM:



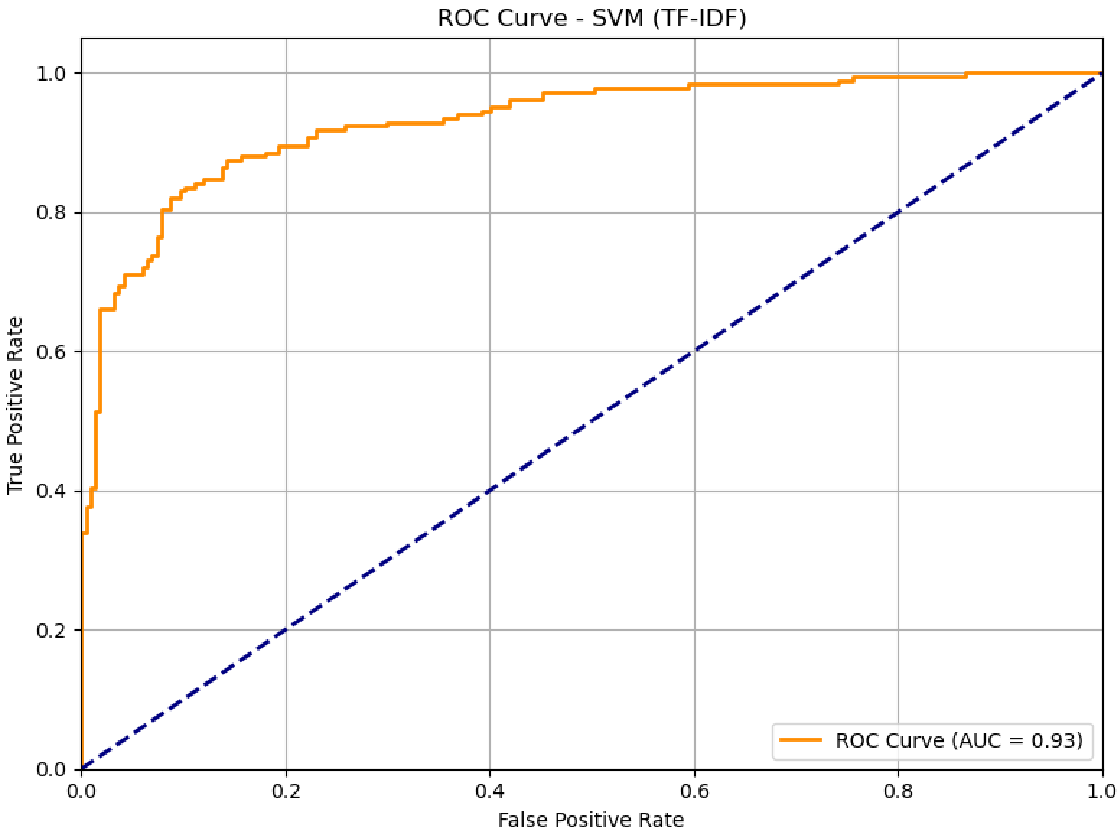
```
Train Accuracy : 0.9594
Test Accuracy : 0.8450
```

Conclusion

The best model was **SVM** with **TF-IDF** and is the one used in **deployment**.



```
Train Accuracy : 0.9919
Test Accuracy : 0.8575
#########################################################

              precision    recall  f1-score   support

           0       0.89      0.84      0.86       217
           1       0.82      0.88      0.85       183

    accuracy                           0.86       400
   macro avg       0.86      0.86      0.86       400
weighted avg       0.86      0.86      0.86       400
```


ROC Curve - SVM (TF-IDF)
True Positive Rate
False Positive Rate
ROC Curve (AUC = 0.93)