Comparison of Bag-of-Words and TF-IDF on IMDb Dataset

Introduction:

Text classification is a core task in Natural Language Processing (NLP) where the objective is to categorize text documents into predefined classes. To perform this, textual data must be converted into numerical features. Two popular techniques for this transformation are Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF). This report evaluates and compares the performance of BoW and TF-IDF on a sentiment classification task using a movie review dataset.

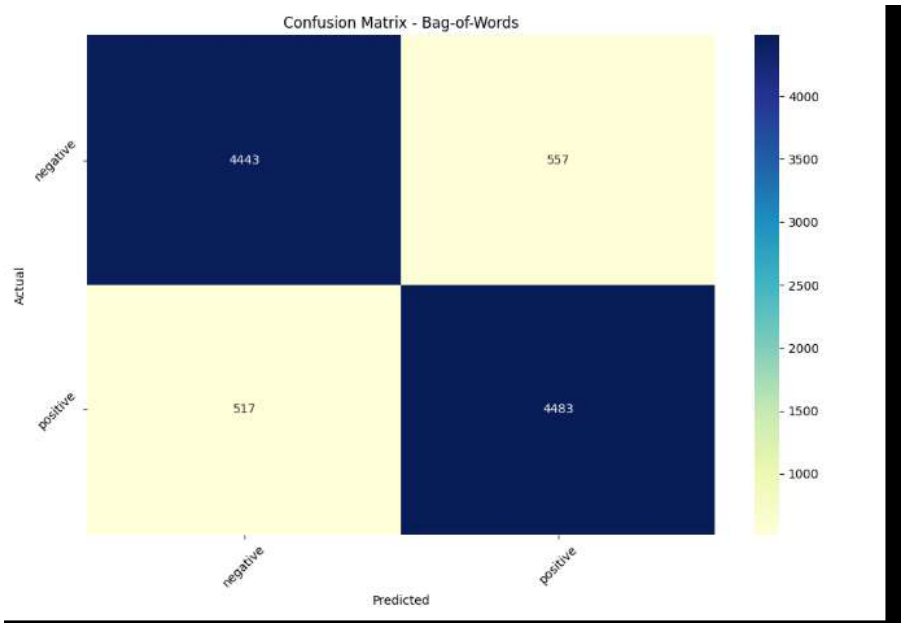
About Dataset

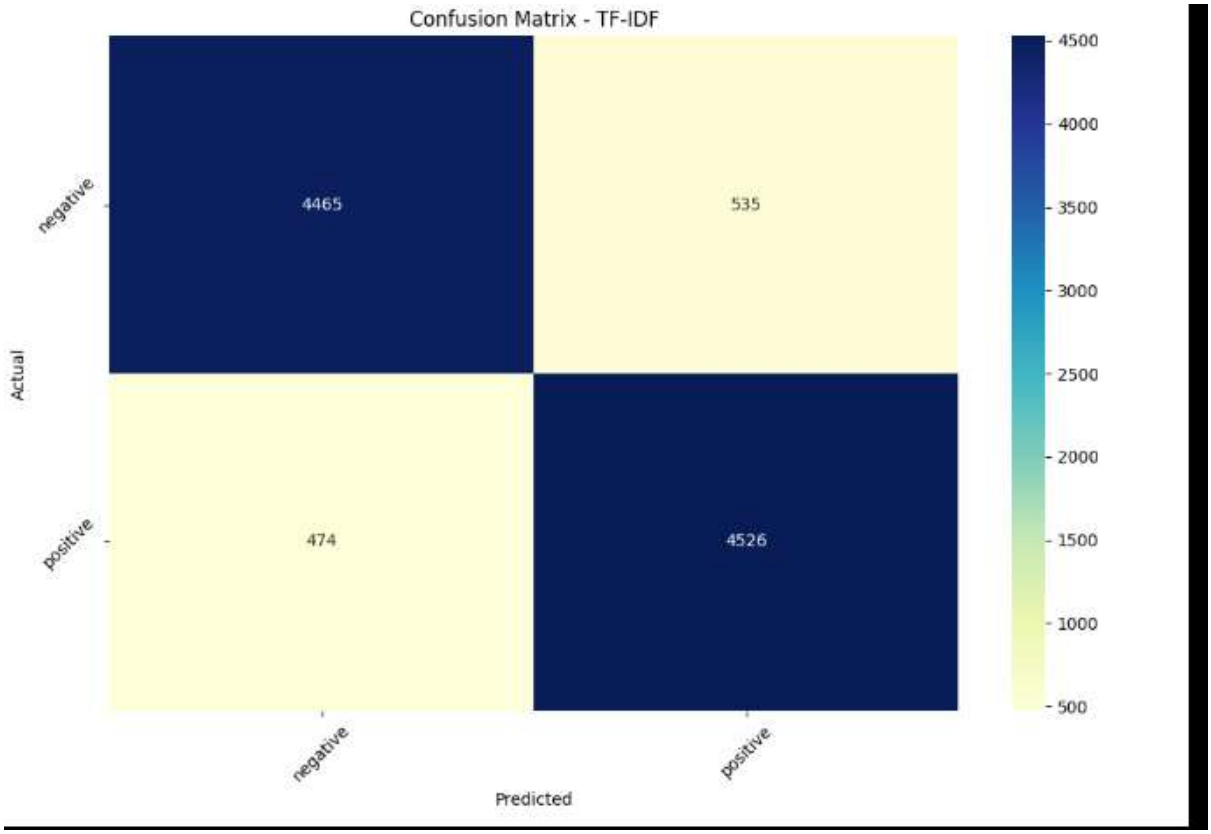
I have chosen the IMDb Movie Reviews dataset from Kaggle, which 500 labeled movie reviews. Each review is classified into either positive or negative sentiment categories.

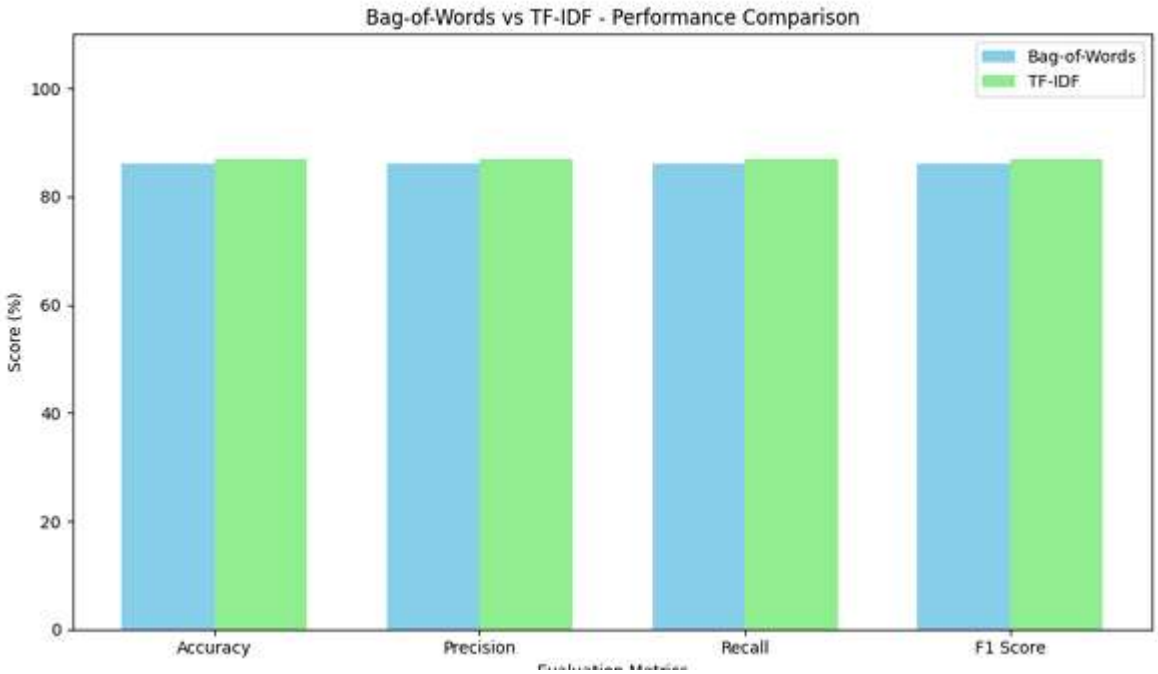
Results

Performance Comparison Table:

Metric	Bag-of-Words	TF-IDF
Accuracy	73.00%	75.00%
Precision	73.48%	77.07%
Recall	73.00%	75.00%
F1 Score	72.63%	74.18%




Confusion Matrix - TF-IDF
4465
535
474
4526
negative
positive
Actual
negative
positive
Predicted
4500
4000
3500
3000
2500
2000
1500
1000
500


Bag-of-Words vs TF-IDF - Performance Comparison
Bag-of-Words
TF-IDF
Score (%)
100
80
60
40
20
0
Accuracy
Precision
Recall
F1 Score