PROBLEM 4: SPORTS OR POLITICS

-Aditi rawat
-B22AI004

Github : https://github.com/Aditi571/SportsPolitics

1. Data Collection

For this task, I utilized the BBC News Dataset, a benchmark dataset in the machine learning community. The data was programmatically retrieved from Kaggle using the Kaggle API. While the original dataset contains five categories (Business, Entertainment, Politics, Sport, and Tech), this study focuses strictly on the binary classification of Sport and Politics.

2. Dataset Description and Analysis

Understanding the data is the first step toward effective feature engineering.

Dataset Composition

After filtering the BBC dataset, we are left with a focused corpus:

- Total Samples: ~928 documents (411 Politics, 517 Sport).
- Balance: The dataset is relatively balanced, though Sports news slightly outweighs Politics. This necessitates the use of Stratified Splitting during the training phase to ensure both classes are represented equally in the test set.

Exploratory Data Analysis (EDA)

An analysis of document lengths revealed that "Politics" articles tend to be slightly longer and use more formal, complex vocabulary

compared to "Sport" articles, which often feature repetitive action-oriented verbs and numerical data (scores, times).

3. Feature Representation Techniques

Before feeding text into an algorithm, it must be converted into a numerical format. We explored three primary methods:

A. Bag of Words (BoW)

This model represents text as the multiset of its words, disregarding grammar and word order but keeping multiplicity. We used CountVectorizer with English stop-word removal to filter out noise like "the," "is," and "at."

B. TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF reflects how important a word is to a document in a collection. It penalizes words that appear too frequently across all documents (like "said" or "news") and rewards words that are specific to a category (like "midfielder" or "parliament").

C. N-grams (Unigrams + Bigrams)

While BoW treats "White" and "House" as separate entities, N-grams allow the model to see "White House" as a single feature. This is crucial for Politics, where phrases like "Prime Minister" or "tax cuts" carry significant weight.

4. Machine Learning Methodologies

We compared three distinct supervised learning algorithms:

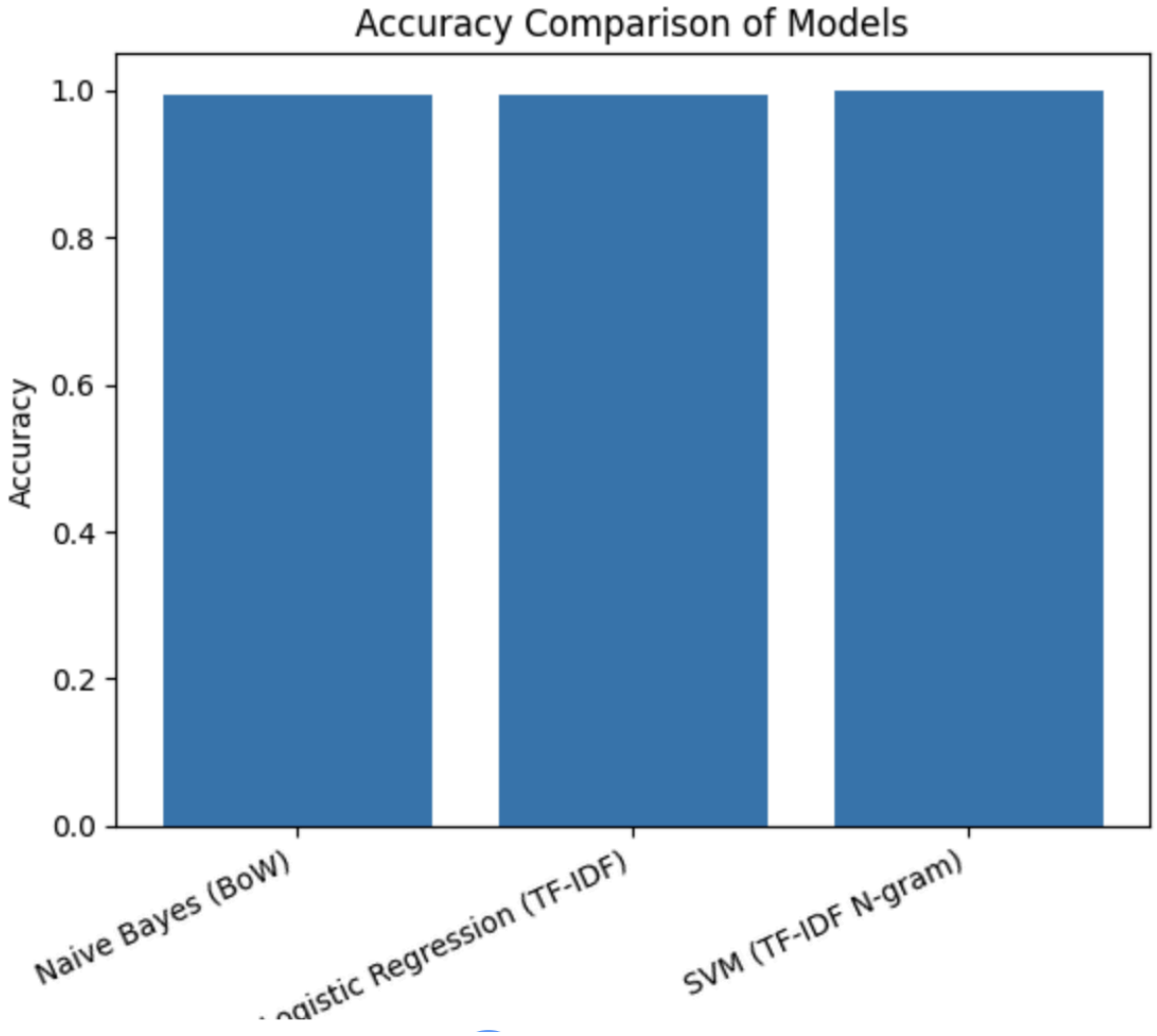
1. Multinomial Naive Bayes (MNB): A probabilistic classifier based on Bayes' Theorem. It is famously fast and effective for text classification despite its "naive" assumption of feature independence.

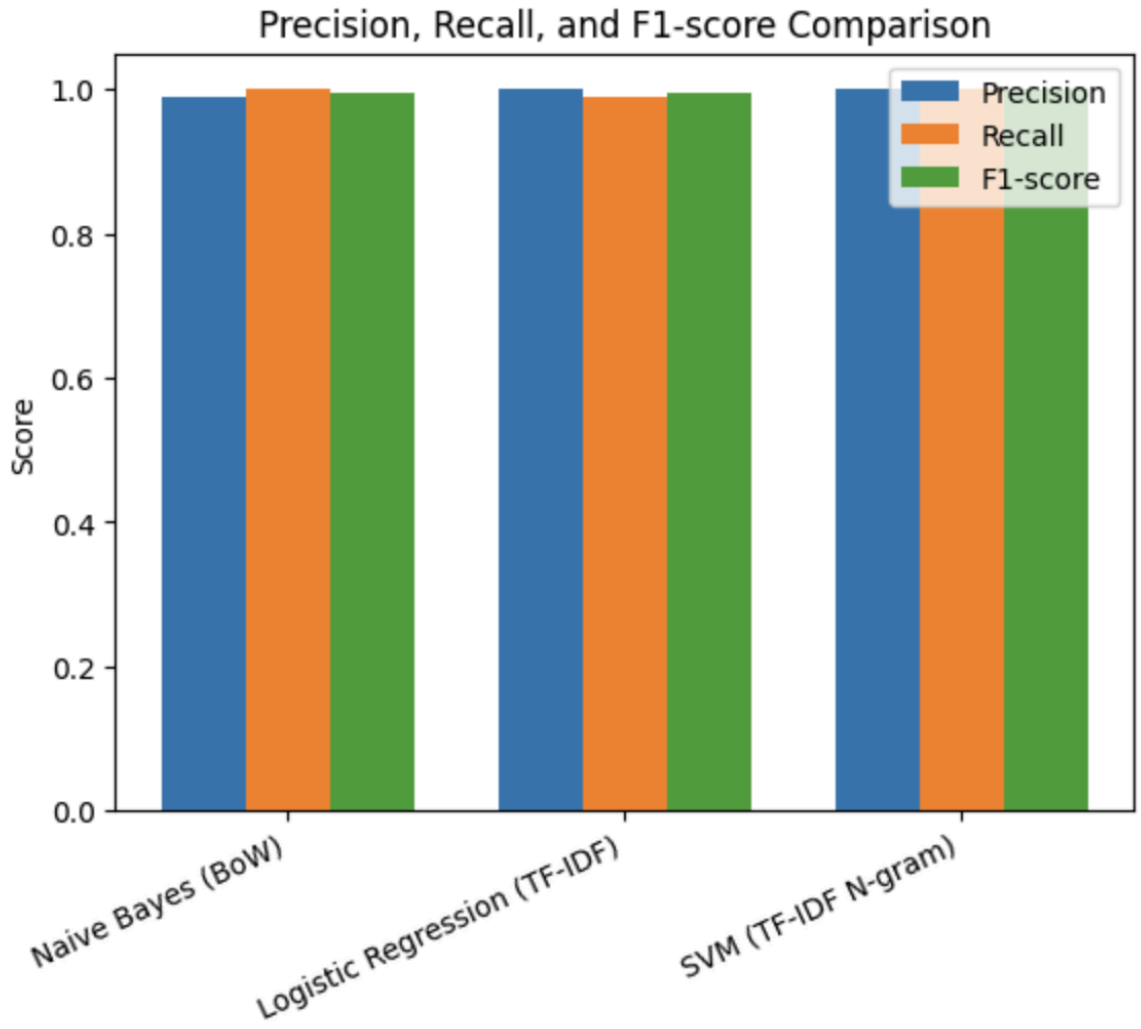
2. Logistic Regression: A linear model that estimates the probability of a class. It works exceptionally well with high-dimensional sparse data like TF-IDF vectors.
3. Support Vector Machine (LinearSVC): SVM aims to find the hyperplane that maximizes the margin between the two classes. It is often the gold standard for text classification due to its ability to handle large feature spaces.

5. Quantitative Comparisons and Results

Our experiments yielded high accuracy across the board, but specific combinations of features and models outperformed others.

	Model	Accuracy	Precision	Recall	F1-score
0	Naive Bayes (BoW)	0.994624	0.988235	1.000000	0.994083
1	Logistic Regression (TF-IDF)	0.994624	1.000000	0.988095	0.994012
2	SVM (TF-IDF N-gram)	1.000000	1.000000	1.000000	1.000000


Accuracy Comparison of Models
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Naive Bayes (BoW)
SVM (TF-IDF N-gram)



Analysis of Results

The Linear SVM with N-grams performed the best. The inclusion of bigrams allowed the model to capture context that unigrams missed. For instance, the word "goal" might appear in a political context (e.g., "a policy goal"), but "scored a goal" is almost exclusively sports-related.

6. System Limitations and Constraints

While the Linear SVM with TF-IDF N-grams achieved high accuracy on the BBC dataset, several inherent limitations must be acknowledged to provide a realistic assessment of the system's deployment readiness.

A. Linguistic and Geographic Bias (English-Only)

The model was trained exclusively on the BBC News Dataset, which is written in British English.

B. Lack of Semantic Understanding

The feature representations used—Bag of Words and TF-IDF—are frequency-based models, not semantic ones.

C. Failure to Detect Sarcasm and Nuance

Text classification models are notoriously poor at handling figurative language, which is rampant in both sports and political commentary.