Sports vs Politics Text Classification

Vamsi Krishna Reddy

B23CM1045

1. Introduction

Text classification is a core problem in Natural Language Processing (NLP) where the objective is to automatically assign predefined categories to textual documents. It is widely used in news categorization, spam filtering, sentiment analysis, and document organization.

In this project, a binary classifier was developed to categorize news articles into either **Sport** or **Politics**. Three different machine learning techniques were implemented and compared using different feature representation methods.

The goal was not only to achieve high accuracy but also to analyze model behavior and validate the correctness of results.

---

2. Dataset Description

The dataset used in this project is the BBC News dataset. After filtering only the required categories, the final dataset consisted of:

- **Total samples:** 928
- **Sport articles:** 511
- **Politics articles:** 417

The dataset is reasonably balanced.

Train-Test Split

An 80–20 stratified split was performed:

- **Training set:** 742 documents
- **Test set:** 186 documents

Test distribution:

- Sport: 102

- Politics: 84

Stratified sampling ensured that both classes were proportionally represented in both sets.

---

3. Preprocessing

Minimal preprocessing was applied:

- Conversion to lowercase
- Removal of numbers and punctuation using regular expressions
- No stemming or lemmatization

The aim was to keep preprocessing simple and focus on comparing feature representation techniques and classifiers.

---

4. Feature Representation

Three feature engineering approaches were used:

4.1 Bag of Words (BoW)

Each document was converted into a frequency vector representing word counts. Word order was ignored.

Used with: **Multinomial Naive Bayes**

4.2 TF-IDF

TF-IDF weights words based on their importance within a document relative to the entire corpus. It reduces the impact of common words.

Used with: **Logistic Regression**

4.3 TF-IDF with Bigrams

In addition to individual words, pairs of consecutive words were included to capture short phrase-level information.

Used with: **Linear Support Vector Machine (SVM)**

5. Model Evaluation

5.1 Random Prediction Baseline (Sanity Check)

Since very high accuracy was observed, a random prediction baseline was implemented to verify that the models were genuinely learning patterns.

Random accuracy obtained:

0.5215 (≈ 52%)

This confirms that the dataset is not trivially predictable and that the trained models significantly outperform random guessing.

5.2 Experimental Results

Model	Feature Type	Accuracy
Naive Bayes	Bag of Words	**1.0000**
Logistic Regression	TF-IDF	**0.9892**
Linear SVM	TF-IDF + Bigrams	**1.0000**

Detailed Observations

Naive Bayes (BoW)

- Accuracy: 100%
- Precision, Recall, F1-score: 1.00 for both classes

Despite being a simple probabilistic model, Naive Bayes perfectly classified all 186 test samples.

Logistic Regression (TF-IDF)

- Accuracy: 98.92%
- Minor misclassifications occurred (very few)

Logistic Regression performed extremely well, slightly below perfect accuracy.

Linear SVM (TF-IDF + Bigrams)

- Accuracy: 100%
- Perfect classification on test set

The inclusion of bigrams likely improved phrase-level discrimination.

---

6. Analysis of Results

The extremely high accuracy can be explained by the nature of the dataset.

Sports articles contain domain-specific vocabulary such as:

- match
- goal
- team
- player
- league

Political articles contain words such as:

- government
- election
- minister
- parliament
- policy

These vocabularies rarely overlap, making the classes highly separable in feature space.

The results suggest that the dataset is almost linearly separable using standard textual features.

---

7. Validation of Model Correctness

To ensure there was no data leakage or implementation error:

1. Stratified splitting was used.
2. Vectorizers were fitted only on training data.
3. A random baseline was implemented.
4. Class distributions were verified.

Since random prediction achieved only ~52% accuracy while trained models achieved ~99–100%, it confirms that the models are genuinely learning meaningful patterns.

---

8. Limitations

Despite excellent performance, several limitations exist:

1. The dataset is clean and well-separated; real-world data may be noisier.
2. The model does not understand context or semantics.
3. Mixed-topic articles could reduce accuracy.
4. No cross-validation across different datasets was performed.

---

9. Conclusion

In this project, a binary text classification system was developed to distinguish between sports and political news articles.

Three machine learning techniques were compared using different feature representations. The results demonstrated near-perfect classification performance, indicating that the dataset is highly separable due to strong domain-specific vocabulary.

The addition of a random baseline confirmed that the models were genuinely learning patterns rather than exploiting flaws in data splitting.

This project highlights the effectiveness of classical machine learning methods for structured text classification tasks.