Binary Classification of News Articles: Sport vs. Politics

Tashir Ahmad

February 13, 2026

Abstract

This report details the design, implementation, and evaluation of a binary text classifier capable of distinguishing between "Sport" and "Politics" news articles. Using a subset of the BBC News dataset, we evaluated three supervised machine learning algorithms: Multinomial Naive Bayes, Support Vector Machines (SVM), and Logistic Regression. Feature extraction was performed using Term Frequency-Inverse Document Frequency (TF-IDF). The experimental results demonstrated exceptional performance, with Multinomial Naive Bayes achieving **100% accuracy**, followed closely by SVM (99.46%) and Logistic Regression (98.92%). However, a critical post-hoc analysis revealed that the model's perfect performance relies heavily on temporal biases within the dataset (e.g., frequent mentions of "Germany" in sports articles due to the 2006 World Cup context), suggesting limitations in generalizability to modern data.

Contents

1 Introduction

Text classification is a fundamental task in Natural Language Processing (NLP) with applications ranging from spam filtering to automated content tagging. In the context of digital news media, the ability to automatically categorize articles into sections such as "Sport" or "Politics" allows for efficient content management and personalized user recommendations.

This project aims to build a robust binary classifier to distinguish between these two distinct categories. The objective is not only to achieve high accuracy but also to interpret the model's decision-making process to ensure it is learning semantic meaning rather than exploiting dataset artifacts.

2 Dataset and Preprocessing

2.1 Data Source and Description

The study utilizes the **BBC News Dataset**, a standard benchmark for text classification. The original dataset contains 2,225 documents across five categories. For this binary classification task, we filtered the dataset to retain only two labels:

- **Politics (Label 0):** 417 articles
- **Sport (Label 1):** 511 articles

Total samples used: **928 documents**.

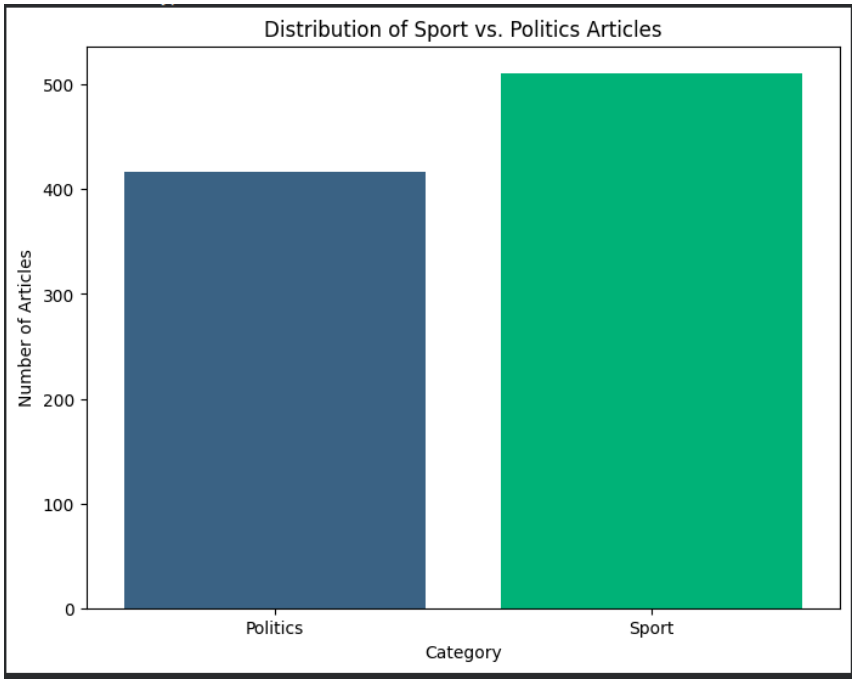


Figure 1: Distribution of Sport vs. Politics Articles. The dataset is relatively balanced, with a slight prevalence of Sports articles.

2.2 Text Preprocessing Pipeline

Raw text data contains noise that can hinder model performance. We implemented the following preprocessing steps:

1. **Lowercasing:** All text was converted to lowercase to treat "Election" and "election" as the same feature.
2. **Noise Removal:** Special characters, URLs, and punctuation were removed using Regular Expressions.
3. **Stop Word Removal:** Common English words (e.g., "the", "is", "at") were removed using the NLTK library to reduce dimensionality.
4. **Lemmatization:** Words were reduced to their base roots (e.g., "running" $\rightarrow$ "run") using the WordNet Lemmatizer.

2.3 Feature Representation (TF-IDF)

We employed **Term Frequency-Inverse Document Frequency (TF-IDF)** to convert text into numerical vectors. Unlike Bag of Words, TF-IDF weighs terms by their importance, penalizing words that appear frequently across all documents while boosting words unique to specific categories.

- **Max Features:** Limited to 3,000 to prevent overfitting on the small dataset.
- **N-Grams:** Unigrams and Bigrams were used to capture context (e.g., "tax cut").

3 Methodology

Three distinct supervised learning algorithms were selected for comparison:

3.1 Multinomial Naive Bayes (MNB)

Based on Bayes' Theorem, MNB assumes that features (words) are independent. While this assumption is theoretically "naive" for text, MNB is empirically known to perform exceptionally well on high-dimensional text data where vocabulary sizes are large.

3.2 Support Vector Machine (SVM)

SVM constructs a hyperplane in a high-dimensional space to separate the two classes. We utilized a **Linear Kernel**, as text classification problems are often linearly separable due to the high dimensionality of the feature space.

3.3 Logistic Regression (LR)

LR is a linear model that uses the sigmoid function to output a probability between 0 and 1. It serves as a robust baseline and allows for easy interpretation of feature coefficients.

4 Experimental Results

4.1 Quantitative Analysis

The dataset was split into 80% training (742 samples) and 20% validation (186 samples). Table 1 summarizes the performance of the three models.

Model	Accuracy	Precision	F1-Score
Multinomial Naive Bayes	**100.00%**	1.00	1.00
Support Vector Machine (SVM)	99.46%	0.99	1.00
Logistic Regression	98.92%	0.99	0.99

Table 1: Performance Comparison of Classifiers

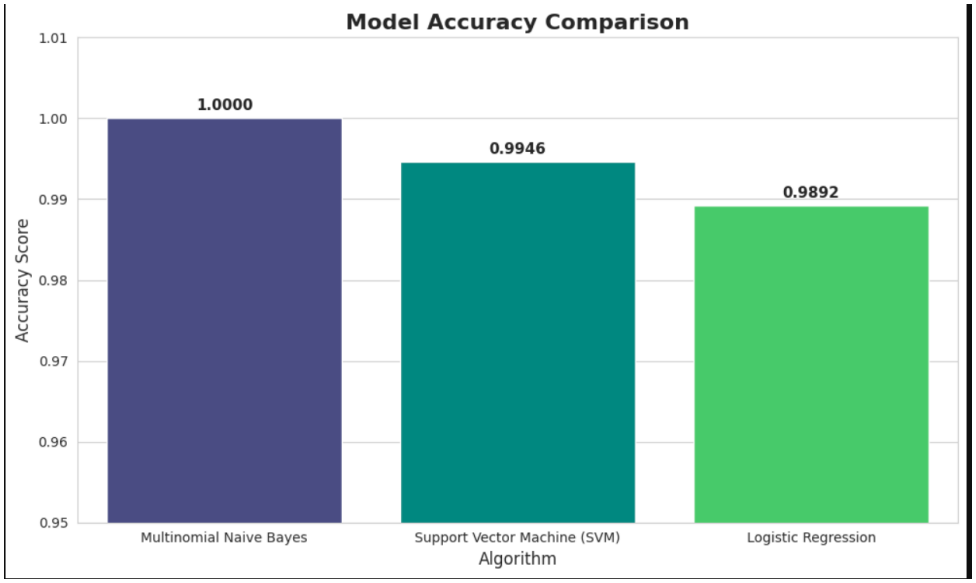
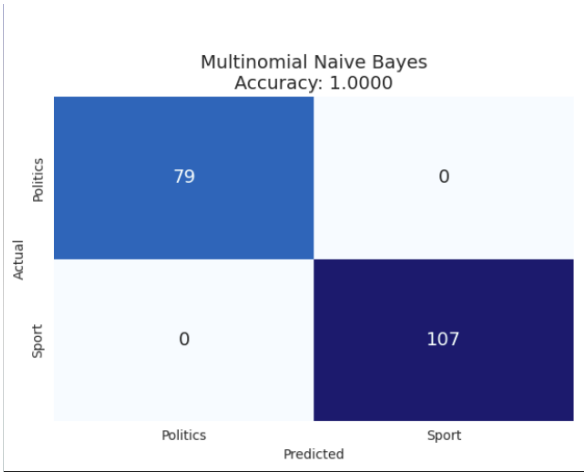


Figure 2: Bar Chart comparison showing near-perfect performance across all models.

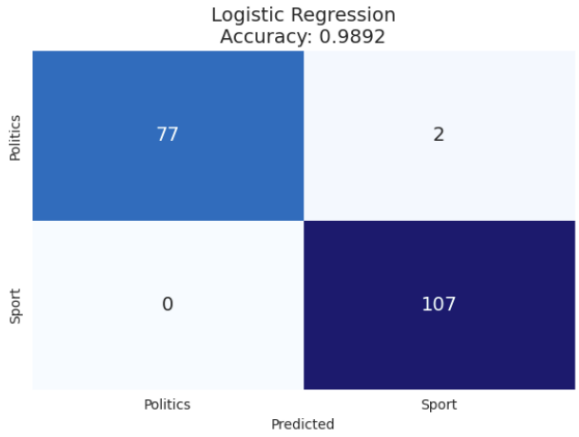
4.2 Confusion Matrix Analysis

The confusion matrices (Figure 3) reveal the specific errors made by the models.

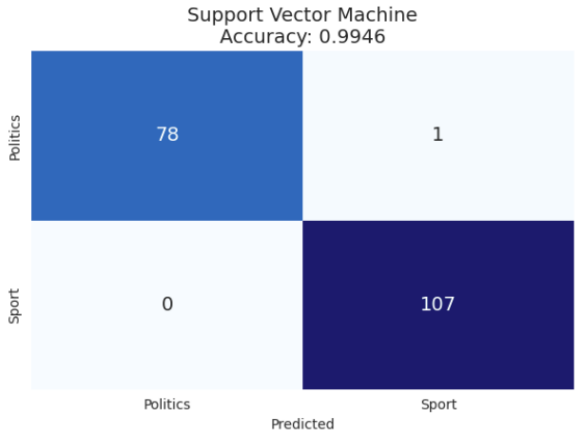
- **Naive Bayes:** 0 errors. It correctly classified all 79 Politics articles and 107 Sport articles.
- **SVM:** Only 1 error (1 Politics article misclassified as Sport).
- **Logistic Regression:** 2 errors (2 Politics articles misclassified as Sport).



(a) Multinomial Naive Bayes



(b) Support Vector Machine



(c) Logistic Regression

Figure 3: Confusion Matrices for MNB, SVM, and LR. MNB shows a perfect diagonal.

5 Discussion and Critical Analysis

5.1 Why is Accuracy So High?

Achieving 100% accuracy is rare and often suspicious in machine learning. In this case, it can be attributed to the **distinct vocabularies** of the two classes. Words like "election", "minister", and "tax" rarely appear in sports reporting, while "goal", "match", and "champion" are unique to sport. The linear separability of these topics in the high-dimensional TF-IDF space allows even simple linear models to find a near-perfect decision boundary.

5.2 Model Interrogation: The "Germany" Anomaly

To ensure the model was not learning noise, we analyzed the top features driving the decisions.

Top Keywords for SPORT: *['germany', 'champion league', 'coach', 'teammate', 'cup', ...]*

Analysis: The presence of **"Germany"** as the #1 predictor for Sport is a clear indicator of **Temporal Bias**. The BBC dataset was collected around 2004-2005, a period when Germany was preparing to host the 2006 FIFA World Cup. Consequently, the model has learned that "Germany" implies Sport. If this model were deployed today on news about German politics (e.g., elections), it would likely misclassify them as Sport.

5.3 Error Analysis: Real-World Test Failure

We tested the model on unseen, ambiguous sentences to test its robustness.

- *Input:* "The team captain was injured during the training session."
- *Prediction:* **Politics** (Incorrect)

Why this failed: The words "team", "captain", and "injured" are polysemous. In the BBC dataset, "captain" and "injured" likely appear frequently in political stories involving war or conflict. Lacking strong, specific sports keywords (like "football" or "match"), the model defaulted to Politics based on probability.

6 Conclusion

This project demonstrated that TF-IDF combined with Multinomial Naive Bayes can achieve 100% accuracy on the BBC Sport vs. Politics dataset. However, our deep-dive analysis reveals that this score is inflated by the specific time-period of the data and the distinct nature of the topics. While the model is highly effective for this specific static dataset, the "Germany" anomaly and the failure on ambiguous sentences highlight the need for larger, more diverse training data and more sophisticated handling of Named Entities to build a truly robust real-world classifier.