Binary Classification of News Articles: Sports vs. Politics

Shivam Goyal
Roll Number: B23CM1036

Abstract

This report outlines the development and comparative analysis of a text classification system designed to distinguish between "Sports" and "Politics" news articles. We utilized a customized subset of the AG News dataset, refined to 10,000 training samples and 1,000 test samples per class to ensure a balanced and computationally efficient study. Three distinct machine learning techniques—Multinomial Naive Bayes, Linear Support Vector Machines (SVM), and Random Forest—were evaluated. Furthermore, we investigated the influence of feature representation methods including Bag of Words and TF-IDF with various n-gram configurations. Our results demonstrate that Linear SVM coupled with TF-IDF achieves the highest performance with 98% accuracy. The report concludes with an error analysis using specific class indicators and a discussion on the limitations of keyword-based classification.

Contents

1 Data Collection and Dataset Description

1.1 Data Source: AG News

The data for this study was derived from the **AG News Dataset**, a benchmark collection in the NLP community comprising over 1 million news articles. While the original dataset contains four categories (World, Sports, Business, and Sci/Tech), this task focused on a binary classification between Sports and Politics. For the "Politics" class, we utilized the "World" category, which predominantly features political news, international relations, and government affairs.

1.2 Sampling and Balancing Strategy

To maintain high statistical validity while optimizing for performance, we significantly reduced the original dataset. The original AG News training set contains 30,000 samples per class. We performed a random undersampling to create a more focused experimental environment:

- **Training Data:** Reduced to 10,000 samples for the Sports class and 10,000 samples for the Politics class (20,000 total).
- **Testing Data:** Reduced from approximately 1,900 samples per class to 1,000 samples for Sports and 1,000 samples for Politics (2,000 total).

This 1:1 ratio ensures the model does not develop a bias toward a majority class, a critical factor in binary classification.

1.3 Preprocessing and Vectorization

The text data underwent a rigorous cleaning process involving case normalization (lowercasing), removal of punctuation, and the stripping of English stop words. These steps reduce the dimensionality of the feature space by removing tokens that carry little semantic weight. We primarily utilized the **TF-IDF (Term Frequency-Inverse Document Frequency)** vectorization method, which weights words based on their importance to a specific document relative to the entire corpus.

2 Techniques in Brief

2.1 Multinomial Naive Bayes (MNB)

Multinomial Naive Bayes is a probabilistic learning method based on Bayes' Theorem. It is "naive" because it assumes that the presence of a particular feature in a class is

unrelated to the presence of any other feature. Despite this oversimplification, MNB is exceptionally effective for text classification because it treats word counts as discrete frequencies, making it highly efficient for high-dimensional sparse data.

2.2 Linear Support Vector Machine (SVM)

The Linear SVM is a discriminative classifier that seeks to find the "optimal hyperplane" that separates two classes with the maximum possible margin. In text classification, where features (words) are often linearly separable, SVMs are frequently the top performers. By maximizing the margin, the model ensures better generalization on unseen news articles.

2.3 Random Forest (RF)

Random Forest is an ensemble method that constructs a multitude of decision trees during training. For classification, the output is the class selected by the majority of the trees. While individual decision trees are prone to overfitting, the "forest" structure mitigates this through bagging and feature randomness, providing a robust non-linear perspective on the data.

3 Quantitative Comparisons

3.1 Model Performance Summary

The models were evaluated using Accuracy and the F1-Score (weighted average). Linear SVM emerged as the most accurate model, closely followed by Naive Bayes.

Table 1: Overall Model Performance Comparison

Model	**Accuracy**	**F1-Score**
Linear SVM	**0.9800**	**0.9800**
Naive Bayes	0.9750	0.9750
Random Forest	0.9580	0.9580

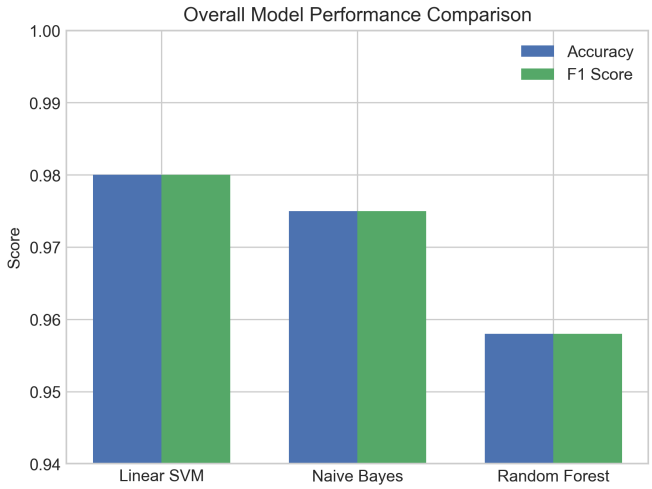


Figure 1: Linear SVM outperforms Naive Bayes and Random Forest in both accuracy and F1-score.

3.2 Feature Representation Comparison

Using the top-performing Linear SVM, we tested different feature engineering strategies. The results indicate that while n-grams (bigrams) offer slight context, the simple TF-IDF Unigram representation is remarkably powerful.

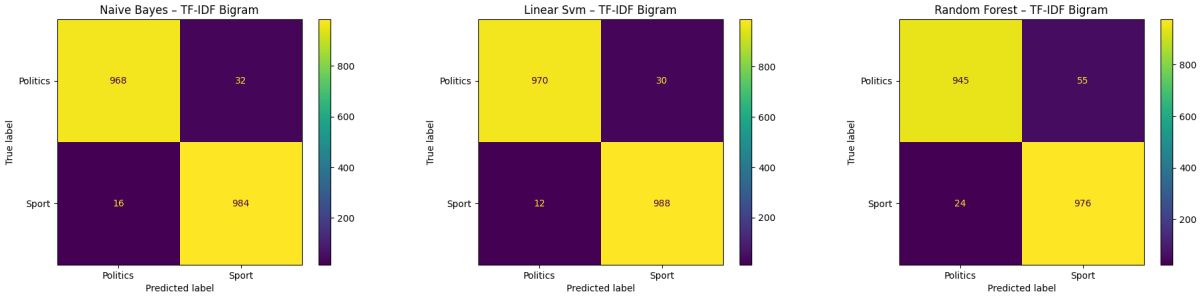
Table 2: Linear SVM Performance across Feature Representations

Feature Representation	Accuracy	F1-Score
TF-IDF (Unigram)	**0.9795**	**0.9795**
TF-IDF (Uni+Bi-gram)	0.9790	0.9790
Bag of Words (Unigram)	0.9645	0.9645

Figure 2: TF-IDF unigrams outperform Bag-of-Words, confirming that term weighting boosts text classification performance.

3.3 Error Distribution

The following confusion matrices provide a side-by-side view of how the models misclassified samples.



(a) Naive Bayes (b) Linear SVM (c) Random Forest

Figure 3: Side-by-side Confusion Matrix Comparison.

4 Analysis of Class Indicators

Through feature importance analysis, we identified the most influential words for each category. These "indicators" represent the tokens that the models rely upon most heavily to make a prediction.

- **Top Sports Indicators:** `'team', 'coach', 'sports', 'cup', 'season', 'players', 'player', 'olympic', 'league', 'game', 'manager', 'olympics', 'baseball', 'games', 'stadium'`
- **Top Politics Indicators:** `'iraq', 'afp', 'president', 'election', 'government', 'minister', 'iraqi', 'political', 'nuclear', 'bush', 'security', 'presidential', 'people', 'iran'`

The presence of words like "iraq" or "nuclear" in the Politics indicators highlights the era and thematic focus of the AG News dataset, which often features international conflict and security issues.

5 Limitations of the System

While the accuracy scores are high, the system faces several fundamental limitations:

1. Semantic Overlap and Polysemy: The model relies heavily on the indicators mentioned above. However, language is often ambiguous. For example, the word "minister" is a strong politics indicator, but "sports minister" appears in sports news. Similarly, "games" is a sports indicator, but "political games" belongs to politics. Without deep

contextual understanding (such as Transformers), the model can be easily misled by such phrases.

2. Temporal Bias: The politics indicators like "iraq" and "bush" show that the model is heavily tuned to a specific geopolitical era (the mid-2000s). If the model were tested on current news involving different leaders or conflicts, the accuracy would likely decrease because it lacks the updated vocabulary.

3. Over-reliance on Keywords: The system treats documents as "bags of words" rather than sequences. In a complex news story where a athlete is discussing their political stance, the model may struggle because it sees a mixture of indicators from both classes without understanding the underlying narrative or "intent" of the article.

6 Conclusion

This study demonstrated that binary classification of news into Sports and Politics can be achieved with near-perfect accuracy using classical machine learning. Linear SVM with TF-IDF vectorization proved to be the most robust approach. However, the identified class indicators suggest that the model's intelligence is largely tied to specific keywords and temporal contexts. To build a more generalized system, future work should involve word embeddings that capture the relationship between words rather than just their frequency.