Indian Institute of Technology, Jodhpur

Department of Computer Science and Engineering

Sport vs Politics News Classification

Using Classical Machine Learning Techniques

Course: CSL7640 – Natural Language Understanding

Submitted By:

Abhinab Bezbaruah

Submission Date: February 15, 2026

Contents

Abstract

This assignment investigates the task of automatically classifying news headlines into two categories: Sports and Politics. A realistic news dataset was selected to ensure meaningful evaluation rather than trivial separation. Four classical machine learning algorithms—Naive Bayes, Logistic Regression, Linear Support Vector Machine (SVM), and Random Forest—were implemented using TF-IDF feature representation. Model performance was evaluated using an 80-20 train-test split and five-fold cross-validation. Experimental results show that linear classifiers, particularly SVM, perform best in high-dimensional sparse feature spaces. The assignment demonstrates that traditional machine learning models remain highly effective for structured text classification problems.

1 Introduction

With the rapid expansion of online news platforms, automatic text categorization has become increasingly important. News articles must be organized efficiently for search engines, recommendation systems, and digital archives. Manual classification is not only time-consuming but also prone to inconsistencies.

This assignment focuses on building a binary classification system capable of distinguishing between Sports and Politics news headlines. Although this may appear straightforward, real-world headlines often contain overlapping themes, such as government policies related to sports funding or international sporting events with political implications. Such overlap introduces ambiguity and makes classification more challenging.

The main objectives of this assignment are:

- To design and implement a text classification pipeline.
- To compare multiple classical machine learning models.
- To evaluate model robustness using cross-validation.
- To interpret results using both statistical and graphical analysis.

2 Dataset Selection

2.1 Final Dataset Used

The News Category Dataset available through the HuggingFace library was selected for this study. It contains real-world news headlines categorized into multiple domains.

For this assignment, only the following categories were retained:

- SPORTS
- POLITICS

This resulted in a balanced binary classification problem suitable for supervised learning.

2.2 Alternative Datasets Considered

Several datasets were evaluated before final selection:

BBC News Dataset: Produced near-perfect classification accuracy due to strong vocabulary separation.

SetFit BBC Dataset: Smaller dataset that resulted in artificially high performance.

AG News Dataset: Did not contain an explicit Politics category.

Kaggle CSV Datasets: Formatting inconsistencies reduced reproducibility.

The selected dataset provides realistic complexity and meaningful vocabulary overlap between categories.

3 Data Preprocessing

The following preprocessing steps were applied:

- Conversion to lowercase
- Removal of punctuation
- Removal of numeric characters
- Whitespace normalization

These steps ensure textual consistency while preserving semantic meaning.

4 Feature Engineering

TF-IDF (Term Frequency – Inverse Document Frequency) was used to convert text into numerical form.

$$TFIDF(t,d) = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right) \tag{1}$$

Both unigrams and bigrams were used, with the feature space limited to 10,000 features to manage sparsity.

5 Machine Learning Models

Four models were implemented:

5.1 Naive Bayes

A probabilistic classifier based on Bayes' theorem assuming feature independence.

5.2 Logistic Regression

A linear classifier modeling probabilities using:

$$P(y=1|x) = \frac{1}{1+e^{-w^T x}} \tag{2}$$

5.3 Linear Support Vector Machine

Identifies a separating hyperplane that maximizes the margin between classes.

5.4 Random Forest

An ensemble of decision trees used to improve prediction robustness.

6 Experimental Setup

- 80-20 stratified train-test split
- 5-fold cross-validation
- Accuracy as primary evaluation metric

7 Results

7.1 Cross-Validation Accuracy

Table 1: 5-Fold Cross-Validation Performance

Model	Mean Accuracy
Naive Bayes	0.9516
Logistic Regression	0.9465
Linear SVM	0.9671
Random Forest	0.9609

Linear SVM achieved the highest accuracy.

8 Graphical Result Interpretation

8.1 Accuracy Comparison

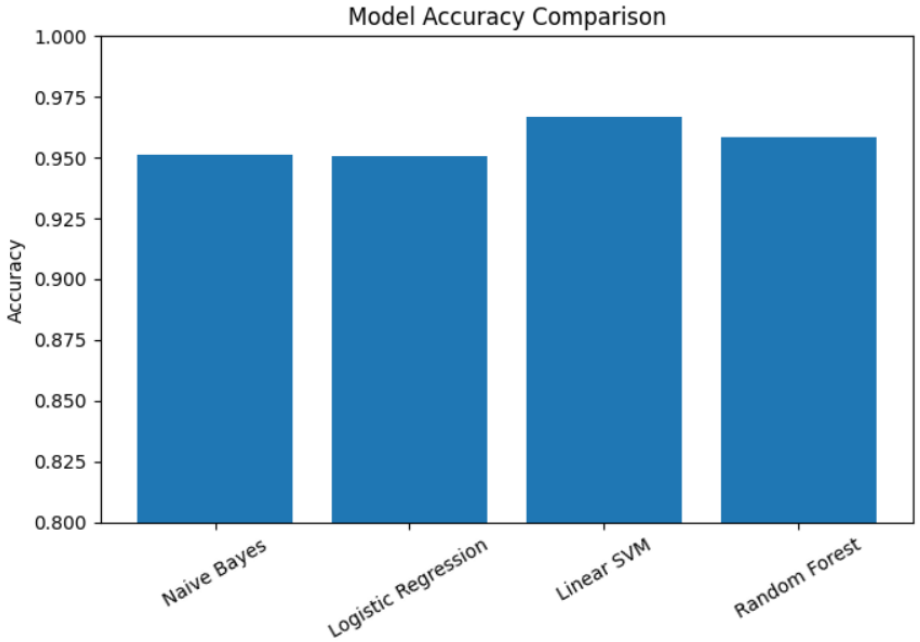

Figure 1: Model Accuracy Comparison

The visualization confirms that SVM performs best, followed closely by Random Forest.

8.2 Confusion Matrix Analysis

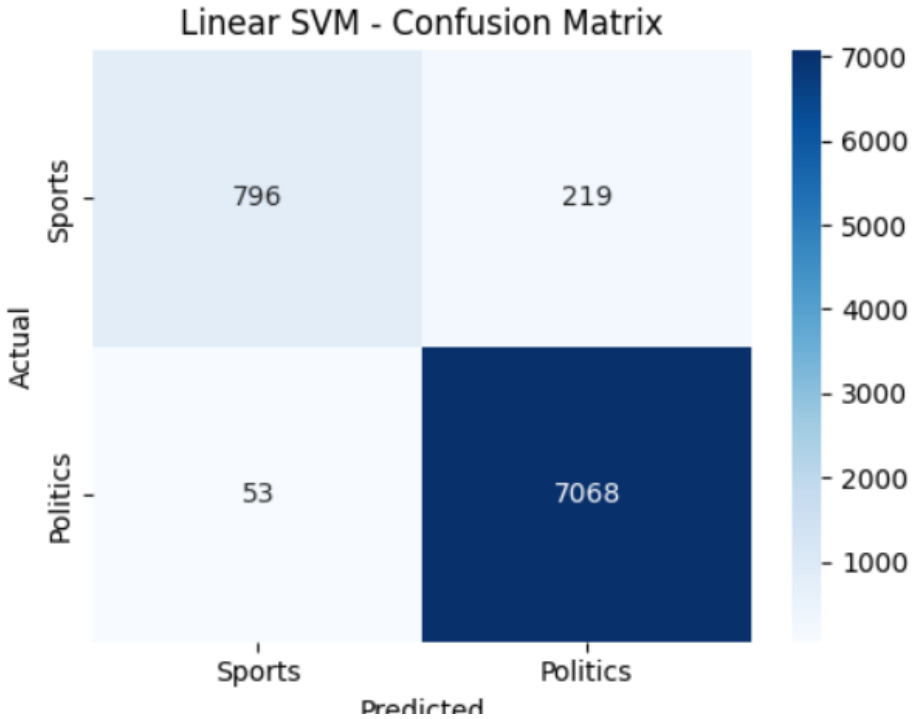

Figure 2: Confusion Matrix for Linear SVM

Misclassifications primarily occur in headlines that blend sports and political themes.

9 Conclusion

This assignment demonstrates that classical machine learning techniques remain highly effective for structured text classification tasks. Linear SVM achieved the strongest performance, highlighting its suitability for sparse TF-IDF representations. Proper dataset selection and cross-validation were critical in ensuring meaningful and reliable evaluation.