Text Classification of News Articles Using Machine Learning

Lagudu Sai Pragathi
B23CM1021

February 15, 2026

Abstract

Text classification is a basic text processing problem in NLP with many real-world applications in information retrieval, sentiment analysis, and text filtering. This paper describes the development and evaluation of a text classification model for classifying news articles between *Sports* and *Politics*. The study employs TF-IDF features in text representation. The study also employs three supervised learning algorithms: random forests, SVM, and logistic regression. The classification accuracy of the model is evident through the experiments conducted. The paper also explores preprocessing techniques, possible data leakage issues, system limitations, and possible enhancements.

Contents

1 Introduction

Text classification is a process of classifying a set of predefined categories for a set of textual documents. With the increasing rate of growth in digital content, the need for automatic systems of text classification is felt. Machine learning approaches, using statistical models for text representation, have shown promising results for document categorization tasks.

This project focuses on distinguishing between **Sports** and **Politics** news articles. These categories are particularly suitable for classification experiments due to their distinct vocabulary and thematic structure. The objective is to design a robust classifier, evaluate multiple algorithms, and analyze performance under controlled preprocessing and feature extraction constraints.

2 Dataset Collection

The dataset used in this study is derived from the BBC news corpus, a widely used benchmark dataset in NLP research. The corpus contains professionally written news articles spanning multiple categories. For this task, only documents labeled as *Politics* (0) and *Sports* (1) were selected.

The dataset was chosen because:

- It contains clean, well-structured text
- Articles belong clearly to distinct categories
- It is commonly used for academic evaluation

3 Dataset Description and Analysis

After filtering for the two target classes, the dataset consisted of multiple documents per category. Initial exploratory analysis was conducted to ensure data quality.

3.1 Data Quality Checks

The following checks were performed:

- Missing value detection
- Duplicate document identification
- Document length analysis

Duplicate documents were removed to prevent memorization effects. Document lengths were constrained to a reasonable range (50–400 words) to avoid bias from extremely short or long articles.

3.2 Class Distribution

Balanced class distribution is essential for unbiased learning. Stratified sampling was used during train-validation splitting to preserve label proportions.

4 Data Preprocessing

Text preprocessing is crucial for reducing noise and improving model generalization. The following normalization steps were applied:

- Lowercasing text
- Removing numbers
- Removing punctuation
- Stopword removal
- Lemmatization

These operations transform raw text into a cleaner representation without introducing data leakage, as they do not depend on corpus-level statistics.

5 Feature Representation Using TF-IDF

Term Frequency–Inverse Document Frequency (TF-IDF) was used to convert text into numerical features. TF-IDF highlights informative terms while suppressing overly frequent words.

The vectorizer configuration included:

- Unigram representation
- Vocabulary size limitation
- Frequency thresholds (min_df, max_df)

The vectorizer was fitted exclusively on training data to prevent validation leakage.

6 Machine Learning Models

Three classification algorithms were evaluated:

6.1 Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees. It is robust to overfitting and capable of modeling nonlinear patterns.

6.2 Support Vector Machine (SVM)

SVM is a powerful classifier that maximizes the margin between classes. It is highly effective in high-dimensional spaces such as TF-IDF feature vectors.

6.3 Logistic Regression

Logistic Regression models class probabilities using a linear decision boundary. It is computationally efficient and interpretable.

7 Experimental Results

7.1 Accuracy Comparison

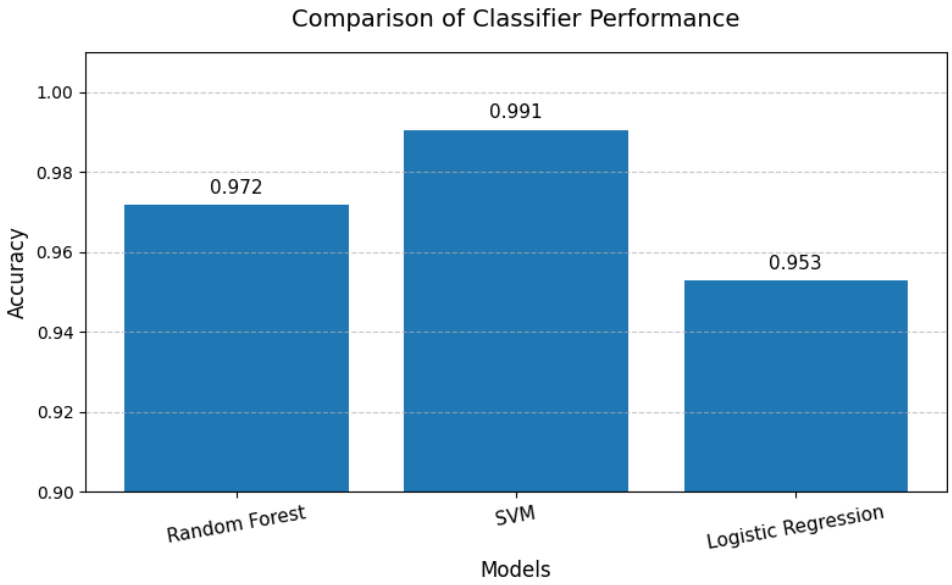

Figure 1: Accuracy Comparison of Machine Learning Models

7.2 Confusion Matrices

8 Result Analysis

The classifiers achieved very high accuracy, with minimal variation across models. Cross-validation further confirmed model stability, yielding a mean accuracy close to 99.5%.

This performance can be attributed to:

- Distinct vocabulary between categories
- Clean textual data
- Effective feature representation

Confusion matrices indicate negligible misclassification, suggesting strong separability of classes.

9 Limitations

Despite strong performance, several limitations exist:

- Dataset bias toward clean news articles
- Limited domain diversity
- Potential vocabulary dependence

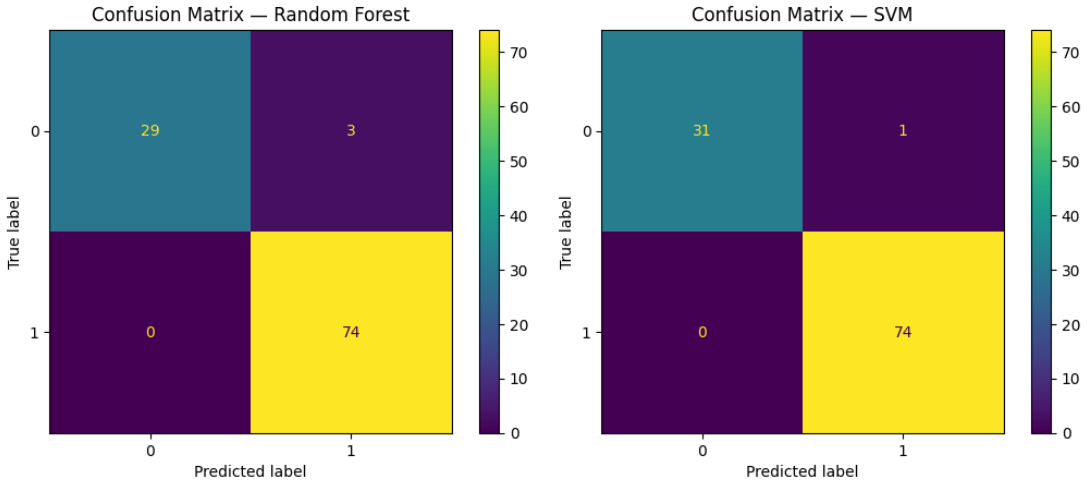

Figure 2: Confusion Matrices for Random Forest and SVM

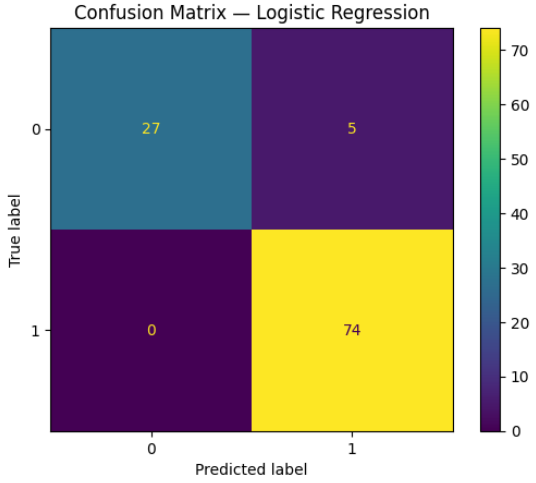

Figure 3: Confusion Matrix for Logistic Regression

Real-world data often contains noise, ambiguity, and informal language, which may reduce classifier performance.

10 Conclusion

This study demonstrates that TF-IDF combined with supervised learning algorithms can effectively classify structured news documents. Among evaluated models, all exhibited strong performance, validating the effectiveness of statistical text representations.

Future work may explore deep learning methods, contextual embeddings, and multi-class classification.