Sports vs Politics Text Classification

CSL7640 – Natural Language Understanding

Your Name
Roll Number

February 15, 2026

1 Introduction

Text classification is a fundamental task in Natural Language Processing (NLP) that involves assigning predefined categories to textual documents. In this project, a machine learning based classifier was developed to distinguish between two domains: **Sports** and **Politics**.

Automatic classification of news articles is important for content organization, recommendation systems, and information retrieval. The goal of this work is to design, implement, and compare multiple machine learning models for this binary classification task.

2 Dataset Collection

A small experimental dataset was created manually for demonstration purposes.

2.1 Sports Samples

- India won the cricket match
- The football team scored three goals
- Olympic games start next month
- The athlete broke the world record
- Hockey tournament begins tomorrow

2.2 Politics Samples

- The election results were announced
- Parliament passed a new bill
- Government launched a new policy

- The president gave a speech
- The minister held a press conference

Each document was labeled as either SPORTS or POLITICS.

3 Preprocessing and Feature Extraction

Before training the models, text preprocessing was performed:

- Lowercasing all words
- Removing punctuation
- Tokenization
- Converting text into numerical features

For feature representation, **TF-IDF (Term Frequency – Inverse Document Frequency)** was used. TF-IDF captures the importance of words in a document relative to the corpus.

4 Machine Learning Techniques

Three machine learning algorithms were implemented and compared.

4.1 Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes conditional independence between features. It is fast and works well for text data.

4.2 Logistic Regression

Logistic Regression is a linear classification model that estimates the probability of a document belonging to a class using a sigmoid function.

4.3 Support Vector Machine (SVM)

SVM finds an optimal hyperplane that maximizes the margin between classes. It performs well on high-dimensional sparse data such as text.

5 Training and Testing

The dataset was split into training and testing sets using a 70:30 ratio. Models were trained on the training data and evaluated on unseen test data.

6 Results and Evaluation

6.1 Accuracy Comparison

Model	Accuracy
Naive Bayes	83.33%
Logistic Regression	100%
SVM	100%

6.2 Confusion Matrix – Naive Bayes

	Predicted Sports	Predicted Politics
Actual Sports	2	1
Actual Politics	0	3

6.3 Confusion Matrix – Logistic Regression

	Predicted Sports	Predicted Politics
Actual Sports	3	0
Actual Politics	0	3

6.4 Confusion Matrix – SVM

	Predicted Sports	Predicted Politics
Actual Sports	3	0
Actual Politics	0	3

7 Analysis

From the results, Logistic Regression and SVM achieved perfect classification accuracy on the test dataset. Naive Bayes performed slightly lower due to its strong independence assumption.

SVM showed robust performance because it handles sparse TF-IDF vectors effectively and finds an optimal decision boundary.

8 Limitations

- Small dataset size
- Limited vocabulary diversity
- Domain overlap possible in real news

- No deep semantic understanding

9 Future Work

Future improvements may include:

- Larger real-world dataset
- Deep learning models (LSTM, BERT)
- Multi-class news classification
- Real-time news scraping

10 Conclusion

This project demonstrated how machine learning techniques can be applied to classify text documents into Sports and Politics categories. TF-IDF feature extraction combined with SVM and Logistic Regression produced the best results. The study highlights the effectiveness of classical ML approaches for text classification tasks.