Sports vs Politics Text Classification using Machine Learning

Roll Number: m23ma2001

February 7, 2026

1 Introduction

Text classification is one of the most fundamental tasks in Natural Language Processing (NLP) and has applications in news categorization, recommendation systems, spam detection, and content filtering. With the rapid growth of online news content, automated categorization of articles has become essential.

In this project, we design and evaluate a machine learning based classifier to categorize text documents into two domains: **Sports** and **Politics**. These domains often share vocabulary but differ semantically, making classification a challenging problem.

The main objectives of this study are:

- To build a document classification system using supervised learning techniques.
- To compare different feature representations such as Bag of Words, TF-IDF, and N-grams.
- To evaluate and compare multiple machine learning models.
- To analyze model performance and identify strengths and limitations.

2 Dataset Collection and Description

A publicly available news dataset was used for this task. The dataset consists of news articles belonging to multiple categories. For this study, only the following categories were selected:

- SPORTS
- POLITICS

Each instance consists of:

- Headline
- Short description

These fields were combined to form the final document text used for classification.

2.1 Dataset Statistics

- Total Sports documents: 5077
- Total Politics documents: 35602
- Total documents used: 40679

The dataset was divided into:

- Training set: 80%
- Testing set: 20%

3 Preprocessing

Before training, the dataset was preprocessed to improve quality and consistency:

- Conversion to lowercase
- Removal of punctuation
- Stopword removal
- Tokenization

These steps help in reducing noise and improving discriminative power of features.

4 Feature Engineering

Three feature extraction methods were used.

4.1 Bag of Words

Each document is represented as a vector of word frequencies.

$$D = [f_1, f_2, f_3, ..., f_n]$$

where f_i denotes the frequency of a word in the vocabulary.

4.2 TF-IDF

TF-IDF reduces the importance of common words and highlights informative ones.

$$TFIDF(w, d) = TF(w, d) \times IDF(w)$$

$$IDF(w) = \log\left(\frac{N}{df(w)}\right)$$

4.3 N-grams

Bigram features were used to capture contextual relationships between words, such as:

- "world cup"
- "prime minister"

5 Machine Learning Models

Three supervised machine learning algorithms were used.

5.1 Naive Bayes

Based on Bayes theorem with conditional independence assumption:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

5.2 Logistic Regression

Uses sigmoid function to model probability:

$$P(Y = 1|X) = \frac{1}{1 + e^{-w^T X}}$$

5.3 Support Vector Machine

Finds optimal separating hyperplane maximizing margin:

$$w \cdot x + b = 0$$

6 Evaluation Metrics

Performance was evaluated using:

- Accuracy
- Precision
- Recall
- F1 Score
- Confusion Matrix

7 Results and Comparison

Feature	Model	Accuracy	Precision	F1 Score
BoW	Naive Bayes	0.980	0.941	0.924
BoW	Logistic Regression	0.975	0.957	0.901
BoW	SVM	0.974	0.919	0.897
TF-IDF	Naive Bayes	0.927	0.991	0.611
TF-IDF	Logistic Regression	0.962	0.974	0.834
TF-IDF	SVM	0.977	0.952	0.909
N-grams	Naive Bayes	0.957	0.986	0.804
N-grams	Logistic Regression	0.975	0.968	0.897
N-grams	SVM	0.977	0.951	0.908

8 Analysis

The experimental results reveal several important insights:

- Bag of Words with Naive Bayes achieved the highest accuracy (98%), showing that word frequency alone is a strong indicator in domain classification.
- TF-IDF significantly improved performance for SVM and Logistic Regression by emphasizing discriminative words.
- Naive Bayes performed poorly with TF-IDF due to its probabilistic assumptions.
- N-grams improved contextual understanding, especially for political phrases.
- SVM performed consistently well across feature representations due to its ability to handle high-dimensional sparse data.

9 Limitations

- Overlapping vocabulary between sports and politics.
- Dataset imbalance (politics articles significantly more).
- Contextual ambiguity in certain headlines.
- Traditional ML models lack deep semantic understanding.

10 Conclusion

This study demonstrated the effectiveness of classical machine learning techniques for document classification. Among all combinations, Bag of Words with Naive Bayes and TF-IDF with SVM achieved strong performance. The results highlight the importance of feature engineering in text classification tasks.

Future work may include:

- Deep learning models (BERT, LSTM)
- Multi-class classification
- Domain adaptation techniques