Sports vs Politics Text Classification

CSL 7640 – Natural Language Understanding

Assignment 1 – Problem 4

1 Introduction

Text classification is one of the most fundamental and widely studied problems in Natural Language Processing (NLP). It involves assigning predefined categories to textual documents based on their content. Applications of text classification include spam detection, sentiment analysis, topic labeling, content filtering, recommendation systems, and automated journalism categorization.

In this project, we focus on a binary classification problem: categorizing news articles into two domains – **Sports** and **Politics**. These two domains are semantically distinct and typically contain domain-specific vocabulary. The objective of this work is to design a classification system using classical machine learning techniques and compare the performance of multiple models under the same feature representation.

The main goals of this project are:

- To construct a dataset containing sports and politics news articles.
- To preprocess textual data using appropriate feature extraction techniques.
- To implement and compare at least three machine learning classifiers.
- To evaluate model performance using quantitative metrics.
- To analyze the strengths and limitations of the proposed system.

2 Dataset Description

The dataset used in this project is derived from the BBC Full-Text News Dataset. This dataset contains news articles categorized into multiple domains. For this experiment, only two categories were selected:

- **Sports**
- **Politics**

After filtering and preprocessing, a total of 928 documents were used for the experiment. The dataset was divided into:

- 80% Training Set
- 20% Testing Set

The sports category typically contains vocabulary such as *match, team, tournament, coach, league, championship*, whereas the politics category includes words such as *government, minister, parliament, election, policy, party.* This lexical distinction contributes significantly to model separability.

3 Data Preprocessing

Textual data requires careful preprocessing before applying machine learning algorithms. The following preprocessing steps were applied:

- Lowercasing all text
- Removal of formatting characters
- Tokenization through vectorization
- Feature extraction using TF-IDF

No aggressive cleaning or domain-specific filtering was applied in order to preserve realistic document structure.

4 Feature Representation

The textual documents were transformed into numerical feature vectors using **Term Frequency - Inverse Document Frequency (TF-IDF)** representation.

TF-IDF is defined as:

$$TF\text{-}IDF(t,d) = TF(t,d) \times \log\left(\frac{N}{df(t)}\right)$$

where:

- $TF(t,d)$ is the frequency of term t in document d,

- N is the total number of documents,
- $df(t)$ is the number of documents containing term t.

TF-IDF reduces the weight of frequently occurring words and increases the importance of discriminative terms. Only unigram features were used in the final model configuration.

5 Machine Learning Models

Three supervised classification models were implemented and compared.

5.1 Multinomial Naive Bayes

Multinomial Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes conditional independence between features given the class label. The posterior probability is computed as:

$$P(C|D) \propto P(C) \prod_{i=1}^{n} P(w_i|C)$$

where C is the class and w_i are words in the document.

Naive Bayes is computationally efficient and performs well in text classification due to the high dimensional sparse feature space.

5.2 Logistic Regression

Logistic Regression is a linear classifier that models the probability of class membership using the sigmoid function:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

It learns a linear decision boundary in feature space. Regularization helps prevent overfitting.

5.3 Support Vector Machine (SVM)

Support Vector Machine aims to find the optimal separating hyperplane that maximizes the margin between classes. For linearly separable data, the decision function is:

$$f(x) = w^T x + b$$

SVM is particularly effective for high-dimensional text classification problems.

6 Experimental Results

The performance of all three models was evaluated using accuracy, precision, recall, and F1-score.

Model	**Accuracy**
Multinomial Naive Bayes	1.00
Logistic Regression	0.9946
Linear SVM	1.00

Table 1: Model Performance Comparison

The classification report for SVM indicated perfect precision, recall, and F1-score for both categories.

The high performance can be attributed to strong lexical separability between the two classes. The vocabulary used in sports articles differs significantly from political discourse, making the classification task linearly separable.

7 Analysis and Discussion

All three classifiers performed exceptionally well. Naive Bayes and SVM achieved perfect accuracy, while Logistic Regression performed marginally lower. The slight variation may be due to regularization and optimization behavior.

The experiment demonstrates that traditional machine learning models remain highly competitive for structured topic classification tasks. The linear decision boundary learned by SVM effectively captures the domain-specific vocabulary differences.

8 Limitations

Despite strong results, several limitations must be acknowledged:

- The dataset exhibits high lexical separability, which simplifies the classification task.
- Real-world news data may contain overlapping vocabulary between domains.
- The model does not account for semantic meaning or contextual nuances.
- Sarcasm and implicit political references are not handled.
- Generalization to unseen domains is uncertain.

9 Future Work

Future improvements may include:

- Incorporating bigram and trigram features
- Applying deep learning models such as LSTMs
- Using transformer-based models like BERT
- Evaluating cross-domain generalization

10 Conclusion

This project presented a comparative study of three classical machine learning models for sports versus politics text classification. Using TF-IDF feature representation and linear classifiers, extremely high accuracy was achieved.

The results confirm that well-structured news data with distinct domain vocabulary can be effectively classified using traditional machine learning techniques. While the dataset used in this study is relatively clean and separable, future work may focus on more challenging and diverse corpora.