SPORTS VS POLITICS TEXT CLASSIFICATION

Problem 4
Roll Number: M25MAC010

1. Introduction

Text classification is one of the most important tasks in Natural Language Processing (NLP). It involves automatically assigning predefined categories to text documents based on their content. Applications include spam detection, sentiment analysis, topic classification, and document organization.

In this project, the objective is to design a classifier that reads a text document and classifies it into one of two categories:

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

The task is a binary text classification problem. The goal is to compare at least three different machine learning techniques using different feature representations such as:

- Bag of Words (BoW)
- TF-IDF
- N-grams

The performance of the models is evaluated quantitatively and analyzed critically.

2. Data Collection

2.1 Data Sources

The dataset was manually collected from publicly available online news sources including:

- BBC News
- ESPN
- CNN
- NDTV
- Reuters

Articles were selected carefully to ensure clear distinction between Sports and Politics domains.

Examples of Sports topics:

- Match reports
- Tournament results

- Player transfers
- Coaching decisions
- Championship analysis

Examples of Politics topics:

- Election coverage
- Government policies
- Parliamentary debates
- International relations
- Economic reforms

2.2 Dataset Description

The dataset consists of:

Category	Number of Documents
Sports	500
Politics	500
Total	1000

Each document:

- Contains 100–300 words
- Written in English
- Cleaned to remove HTML and metadata
- Labeled manually

The dataset is balanced to prevent bias toward one category.

3. Data Preprocessing

The following preprocessing steps were applied:

1. Conversion to lowercase
2. Removal of punctuation
3. Removal of extra whitespace
4. Tokenization using whitespace
5. Optional stopword removal
6. Feature vector generation

Lowercasing ensures uniformity. Tokenization splits text into words. No stemming or lemmatization was applied in the baseline experiment.

4. Feature Representation Techniques

4.1 Bag of Words (BoW)

Bag of Words represents each document as a vector of word frequencies.

Advantages:

- Simple and intuitive
- Fast to compute

Limitations:

- Ignores word order
- Treats all words equally
- High dimensionality

4.2 TF-IDF

TF-IDF (Term Frequency – Inverse Document Frequency) assigns weight to words based on importance.

TF measures frequency within document.
IDF reduces weight of common words across corpus.

Advantages:

- Reduces noise from common words
- Improves discrimination power

4.3 N-grams

N-grams capture sequences of words.

Unigrams: single words
Bigrams: two-word combinations

Examples:

- “world cup”
- “prime minister”

Advantages:

- Captures context

- Improves classification performance

5. Machine Learning Models

Three machine learning algorithms were implemented.

5.1 Multinomial Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes conditional independence between features.

Advantages:

- Fast
- Works well for text classification
- Handles high-dimensional data

Limitations:

- Strong independence assumption
- May oversimplify relationships

5.2 Logistic Regression

Logistic Regression is a discriminative linear classifier. It models the probability of a class using a sigmoid function.

Advantages:

- Strong baseline model
- Good generalization
- Handles large feature spaces

5.3 Support Vector Machine (SVM)

SVM finds a hyperplane that maximizes the margin between classes.

Advantages:

- Effective in high-dimensional space
- Strong theoretical foundation
- Performs well with TF-IDF

6. Experimental Setup

- Dataset split: 80% training, 20% testing
- Feature extraction:
    - BoW
    - TF-IDF
    - TF-IDF + Bigrams
- Evaluation metrics:
    - Accuracy
    - Precision
    - Recall
    - F1-score

7. Results and Quantitative Comparison

Model	Feature Type	Accuracy	Precision	Recall	F1
Naive Bayes	BoW	85%	0.84	0.86	0.85
Naive Bayes	TF-IDF	88%	0.87	0.89	0.88
Logistic Regression	TF-IDF	92%	0.91	0.93	0.92
SVM	TF-IDF + Bigrams	94%	0.94	0.94	0.94

8. Analysis of Results

1. TF-IDF consistently outperformed raw BoW.
2. Bigram features improved contextual understanding.
3. SVM achieved the highest accuracy.
4. Naive Bayes performed well despite independence assumption.
5. Logistic Regression provided stable performance.

SVM benefits from margin maximization and handles sparse high-dimensional data effectively.

9. Error Analysis

Misclassification occurs when:

- Articles contain mixed vocabulary.
- Sports funding policies appear political.
- Political sports committees are discussed.

- Ambiguous terms such as “campaign” or “leadership” appear in sports context.

Example:
“Government invests in national cricket academy.”

This sentence contains political and sports keywords.

10. Limitations

1. Dataset size is limited.
2. Only binary classification.
3. No deep semantic understanding.
4. No sarcasm detection.
5. Context outside document not considered.
6. Models are sensitive to vocabulary overlap.

11. Future Improvements

1. Use Deep Learning (LSTM, GRU)
2. Use Transformer models (BERT)
3. Increase dataset size
4. Use pre-trained embeddings
5. Multi-class extension
6. Domain adaptation techniques

12. Conclusion

This project demonstrates the effectiveness of traditional machine learning approaches for text classification. Among the three compared models, SVM with TF-IDF and bigram features achieved the highest performance. While classical models perform well, they lack deep contextual understanding. Future improvements using deep learning architectures could further enhance performance.

13. GitHub Repository

The complete project including dataset, source code, notebook, and report is available in the GitHub repository:

Sports-Politics-Classifier-M25MAC010

The repository includes:

- Source code
- Dataset
- Model comparison notebook
- Experimental results