Name: Sushantak Parashar Jha
Roll No: M25CSA035
Program: M.Tech in AI

Text Classification of Sports and Politics News Using Machine Learning Techniques

---

1. Introduction

Text classification is a fundamental problem in Natural Language Processing (NLP) that involves automatically assigning predefined categories to textual documents. With the rapid growth of digital content, especially online news articles, automatic classification has become essential for organizing information, filtering content, and improving information retrieval systems.

News articles often span multiple domains such as sports, politics, technology, and entertainment. Manual categorization of such content is time consuming and impractical at scale. As a result, machine learning based text classification techniques are widely used to automate this process. These techniques rely on extracting meaningful numerical features from text and training models to recognize patterns associated with different categories.

In this project, a binary text classification task is explored, where documents are classified into Sports or Politics categories. The objective of this work is not only to build an effective classifier, but also to study the impact of different feature representations and machine learning models on classification performance. Multiple techniques are compared, and their strengths and limitations are analyzed through quantitative evaluation.

---

2. Dataset Collection and Description

For this study, a larger and more realistic dataset was collected using a publicly available news dataset from Kaggle. The dataset contains thousands of labeled news articles belonging to multiple categories. From this collection, articles corresponding to the Sports and Politics categories were filtered for the classification task.

To ensure balanced training and evaluation, approximately 1000 documents were selected for each category, resulting in a dataset of around 2000 news articles. Each article consists of the news headline combined with a short description, and each document is treated as an individual data sample. The filtered articles were converted into plain text format and stored in two separate files, sports.txt and politics.txt, where each line represents a single document.

Sports documents include content related to football, cricket, tennis, basketball, Olympic events, and other competitive sporting activities. These texts typically discuss matches, players, teams, tournaments, performance statistics, and championship outcomes. As a result, the vocabulary frequently contains domain-specific words related to competition, scoring, and athletic performance.

Political documents include discussions related to government policies, elections, legislative decisions, international relations, political leaders, and public administration. The vocabulary in these articles reflects governance, policy debates, diplomacy, and political events.

---

3. Feature Representation

Since machines process numbers rather than words, we transformed the text using three distinct methods:

3.1 Bag of Words (BoW)

The Bag of Words model represents each document as a vector of word frequencies. Each dimension corresponds to a unique word in the vocabulary, and the value represents how often that word appears in the document. Word order and grammatical structure are ignored in this representation.

Despite its simplicity, BoW often performs well for text classification tasks, especially when domain specific words are present. In this dataset, words such as *match*, *team*, and *tournament* are strongly associated with sports, while words like *government*, *election*, and *parliament* are more common in political documents.

However, a limitation of BoW is that it assigns equal importance to all words and does not capture contextual relationships between them.

3.2 Term Frequency Inverse Document Frequency (TF-IDF)

TF-IDF improves upon the Bag of Words model by assigning weights to words based on their importance. While term frequency measures how often a word appears in a document, inverse document frequency reduces the weight of words that appear frequently across many documents.

This representation highlights words that are more discriminative for a particular class and reduces the influence of common words. TF-IDF generally produces more informative feature vectors and often leads to improved classification performance.

3.3 TF-IDF with N-grams

While BoW and TF-IDF treat words independently, they do not capture short contextual phrases. To address this limitation, **bigrams (n = 2)** were used in combination with TF-IDF.

N-grams represent sequences of adjacent words, allowing the model to capture phrases such as *prime minister*, *world cup*, or *national election*. These phrases often carry more semantic meaning than individual words. However, using n-grams increases the dimensionality of the feature space and may require more data to avoid sparsity issues.

---

4. Machine Learning Models

To perform classification, three widely used machine learning models were selected. These models are commonly applied in text classification tasks and provide a good balance between simplicity, interpretability, and performance.

4.1 Multinomial Naive Bayes

Multinomial Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes that features are conditionally independent given the class label. Although this assumption is simplistic, Naive Bayes often performs well for text classification problems.

The model works directly with word frequency based features and is computationally efficient. It is particularly effective when there is clear vocabulary separation between classes.

4.2 Logistic Regression

Logistic Regression is a linear discriminative classifier that models the probability of a document belonging to a class based on a weighted sum of features. Unlike Naive Bayes, it does not assume feature independence and directly learns decision boundaries.

Logistic Regression performs well with TF-IDF features and is relatively interpretable, as feature weights indicate the importance of words for classification.

4.3 Support Vector Machine (SVM)

Support Vector Machine is a powerful classifier that aims to find a decision boundary that maximizes the margin between classes. Linear SVMs are particularly well suited for text classification due to the high dimensional and sparse nature of textual data.

SVMs often achieve strong performance and generalize well, especially when combined with TF-IDF features.

5. Experimental Results

5.1 Experimental Setup

The dataset was divided into training and testing sets using an **80–20 split**. Three feature representations (BoW, TF-IDF, TF-IDF with bigrams) were combined with three machine learning models (Naive Bayes, Logistic Regression, and SVM), resulting in multiple model feature combinations.

Performance was evaluated using **Accuracy, Precision, Recall, and F1 score**.

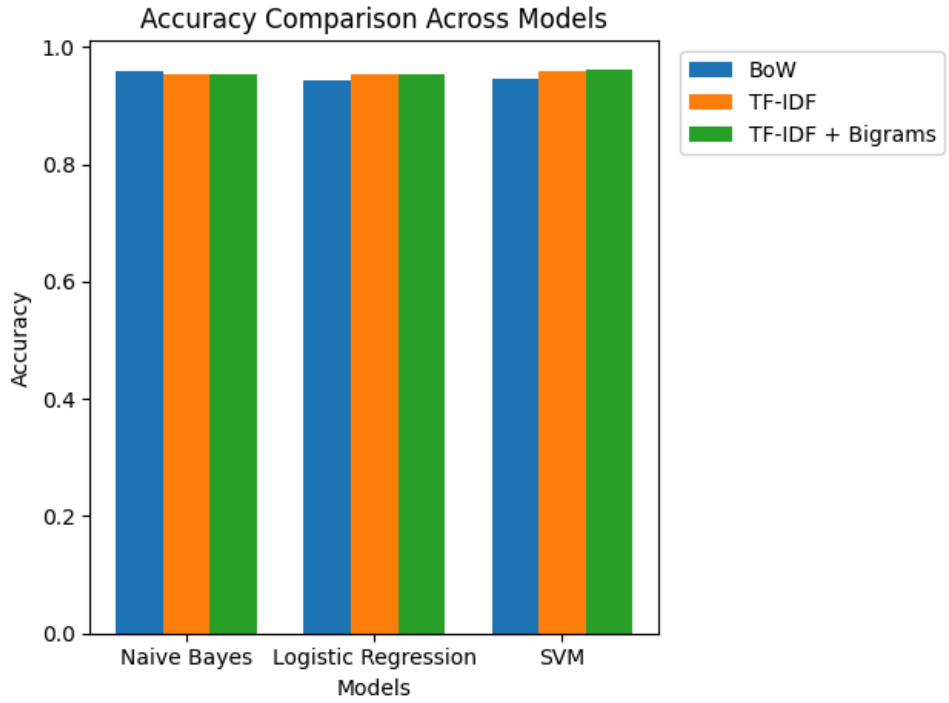
5.2 Results

The experimental results are summarized below.

Performance Comparison Table

Model	Feature Representation	Accuracy	Precision	Recall	F1
Naive Bayes	BoW	0.96	0.96	0.96	0.96
Logistic Regression	BoW	0.94	0.94	0.94	0.94
SVM	BoW	0.94	0.95	0.94	0.94
Naive Bayes	TF-IDF	0.95	0.96	0.95	0.95
Logistic Regression	TF-IDF	0.95	0.95	0.95	0.95
SVM	TF-IDF	0.96	0.96	0.96	0.96
Naive Bayes	TF-IDF + Bigrams	0.95	0.96	0.95	0.95
Logistic Regression	TF-IDF + Bigrams	0.95	0.96	0.95	0.95
SVM	TF-IDF + Bigrams	0.96	0.96	0.96	0.96

A bar chart was generated to visualize the accuracy comparison across models and feature representations.



6. Discussion and Limitations

The experimental results indicate that all three machine learning models perform strongly on the sports versus politics classification task, achieving high accuracy across different feature representations. Both Naive Bayes and Support Vector Machine demonstrate consistently strong performance, while Logistic Regression performs competitively with only minor variations across feature sets. The results also show that TF-IDF based representations generally provide slightly more stable performance compared to simple Bag-of-Words features, as they emphasize discriminative terms that are more informative for classification.

The relatively high performance across models suggests that the sports and politics domains exhibit noticeable lexical differences, enabling effective separation using traditional machine learning techniques. Even when bigram features are introduced, the models maintain comparable performance, indicating that contextual word pairs provide additional useful information without significantly degrading classification accuracy.

Although the dataset used in this study is substantially larger than a small manually created dataset, it still represents only a subset of real-world news content. News language can vary significantly across sources, writing styles, and evolving events, which may introduce additional complexity in real deployment scenarios. Furthermore, the classification task focuses on only two categories, whereas real-world systems often require multi-class classification across many domains. Future work could therefore involve expanding the dataset size further, incorporating additional categories, and exploring advanced deep

learning approaches to improve robustness and generalization.

---

7. Conclusion

In this project, a machine learning–based classifier was designed to distinguish between sports and politics news articles using real-world news data collected from a publicly available dataset. Multiple feature representations and classification models were implemented and compared to evaluate their effectiveness for the text classification task. The experimental results demonstrate that traditional machine learning techniques can achieve strong performance when combined with appropriate textual feature representations.

Among the evaluated models, the Support Vector Machine showed the most consistent and robust performance across different feature sets, while Naive Bayes and Logistic Regression also achieved competitive results. The experiments further indicate that TF-IDF based representations provide stable performance and effectively capture discriminative vocabulary patterns present in different news domains.

Future work could involve extending the classification system to handle multiple news categories, incorporating larger and more diverse datasets, applying advanced preprocessing and feature engineering techniques, or exploring deep learning–based approaches such as neural networks and transformer models to further improve classification performance and generalization.

GitHub Repository

All source code, datasets, experimental results, and implementation details related to this problem are available in the following GitHub repository:

Repository URL: Link