Name: Sushantak Parashar Jha
Roll No: M25CSA035
Program: M.Tech in AI
GitHub Repository and Pages Link: Link

Text Classification of Sports and Politics News Using Machine Learning Techniques

1. Introduction

One of the core issues in Natural Language Processing (NLP) is text classification, which is the automatic assignment of predetermined categories to textual documents. Automatic classification has become crucial for information organization, content filtering, and enhancing information retrieval systems due to the explosive growth of digital content, particularly online news articles.

News stories frequently cover a variety of topics, including politics, sports, technology, and entertainment. Such content is time-consuming and impractical to manually categorise on a large scale. Text classification methods based on machine learning are therefore frequently employed to automate this procedure. These methods rely on training models to identify patterns linked to various categories and extracting significant numerical features from text.

This project investigates a binary text classification task in which documents are categorised as either politics or sports. In addition to developing an efficient classifier, the goal of this work is to investigate how various feature representations and machine learning models affect classification performance. Through quantitative evaluation, several approaches are contrasted, and their advantages and disadvantages are examined.

2. Dataset Collection and Description

A larger and more realistic dataset was gathered for this study using a Kaggle news dataset that is openly accessible. Thousands of news articles with labels from various categories are included in the dataset. For the classification task, articles that fit into the Sports and Politics categories were selected from this collection.

About 1000 documents were chosen for each category to guarantee balanced training and evaluation, yielding a dataset of about 2000 news articles. The news headline and a brief description make up each article, and each document is handled as a separate data sample. A single document is represented by each line in the two distinct files, sports.txt and politics.txt, which contain the filtered articles in plain text format.

Football, cricket, tennis, basketball, the Olympics, and other competitive sports are all covered in sports documents. Matches, players, teams, competitions, performance data, and championship results are frequently covered in these texts. Because of this, domain-specific terms pertaining to competition, scoring, and athletic performance are commonly found in the vocabulary.

Government policies, elections, legislative decisions, international relations, political leaders, and public administration are all covered in political documents. These articles' vocabulary reflects political events, policy discussions, diplomacy, and governance.

3. Feature Representation

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

3.1 Bag of Words (BoW)

Every document is represented as a vector of word frequencies by the Bag of Words model. The value indicates how frequently a particular word occurs in the document, and each dimension corresponds to a distinct word in the vocabulary. This representation disregards grammatical structure and word order.

BoW frequently does well on text classification tasks despite its simplicity, particularly when domain-specific words are present. While terms like government, election, and parliament are more frequently found in political documents, terms like match, team, and tournament are strongly linked to sports in this dataset.

Nevertheless, a drawback of BoW is that it gives every word the same weight and ignores the contextual connections between them.

3.2 Term Frequency Inverse Document Frequency (TF-IDF)

By giving words weights according to their significance, TF-IDF enhances the Bag of Words model. Inverse document frequency lessens the weight of words that occur frequently across numerous documents, whereas term frequency counts the number of times a word appears in a document.

This representation lessens the impact of common words and emphasises words that are more discriminatory for a given class. TF-IDF typically results in better classification performance and more informative feature vectors.

3.3 TF-IDF with N-grams

Although TF-IDF and BoW handle words separately, they are unable to recognise brief contextual phrases. Bigrams (n = 2) were combined with TF-IDF to overcome this constraint.

N-grams, which are collections of neighbouring words, enable the model to recognise phrases like "prime minister," "world cup," and "national election." The semantic meaning of these phrases is frequently greater than that of individual words. Nevertheless, employing n-grams raises the feature space's dimensionality and might necessitate additional data in order to prevent sparsity problems.

---

4. Machine Learning Models

Three popular machine learning models were chosen to carry out classification. These models offer a good balance between performance, interpretability, and simplicity and are frequently used in text classification tasks.

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**.

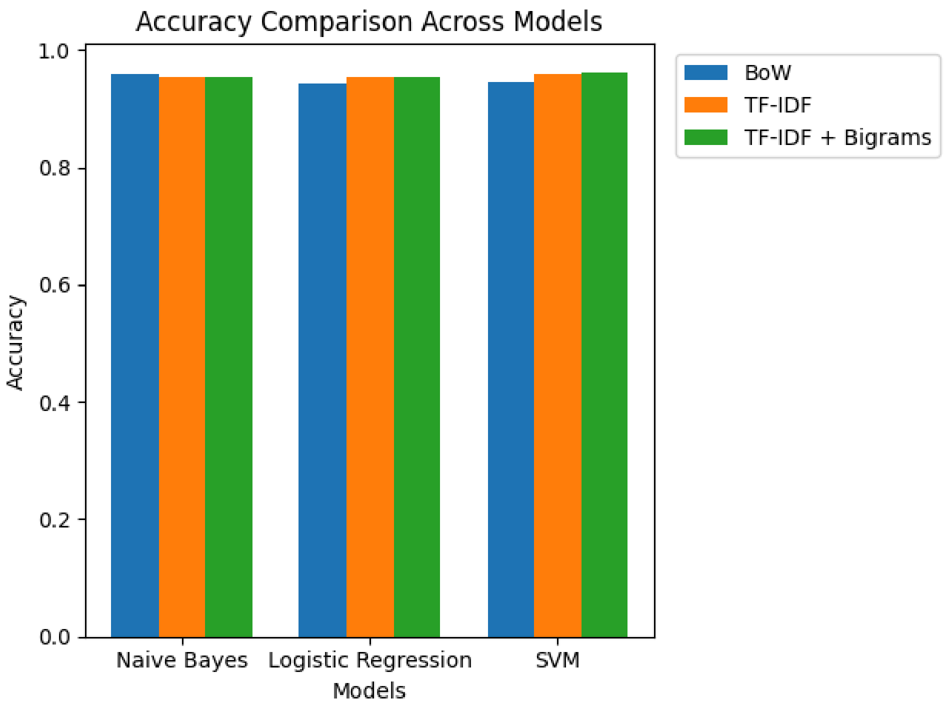
5.2 Results

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 results of the experiment show that all three machine learning models achieve high accuracy across a variety of feature representations and perform well on the sports versus politics classification task. While Logistic Regression performs competitively with only slight variations across feature sets, Naive Bayes and Support Vector Machine consistently exhibit excellent performance. Furthermore, the results demonstrate that because TF-IDF-based representations prioritise discriminative terms that are more instructive for classification, they typically offer slightly more stable performance than simple Bag-of-Words features.

The relatively high model performance suggests that the political and sports domains have recognisable lexical differences, enabling effective separation using standard machine learning techniques. The models continue to perform similarly even after bigram features are added, suggesting that contextual word pairs offer more helpful information without appreciably lowering classification accuracy.

The dataset used in this study is still only a subset of real-world news content, despite being significantly larger than a small manually created dataset. Real deployment scenarios may become more complex due to the wide variations in news language across sources, writing styles, and changing events. Additionally, the classification task only considers two categories, whereas multi-class classification across numerous domains is frequently necessary for real-world systems. Therefore, to increase robustness and generalisation, future work may entail enlarging the dataset even more, adding new categories, and investigating cutting-edge deep learning techniques.

7. Conclusion

This project used real-world news data gathered from a publicly accessible dataset to create a machine learning-based classifier that can differentiate between news articles about politics and sports. The efficacy of various feature representations and classification models for the text classification task was assessed through implementation and comparison. The experimental results show that when combined with suitable textual feature representations, conventional machine learning techniques can achieve strong performance.

While Naive Bayes and Logistic Regression also produced competitive results, the Support Vector Machine model performed the most reliably and consistently across various feature sets among the models that were evaluated. Additionally, the experiments show that representations based on TF-IDF offer consistent performance and successfully capture discriminative vocabulary patterns found in various news domains.

To further enhance classification performance and generalisation, future work might apply sophisticated preprocessing and feature engineering techniques, expand the classification system to handle multiple news categories, incorporate larger and more varied datasets, or investigate deep learning-based approaches like neural networks and transformer models.