Sports vs Politics Text Classification Using Machine Learning

Name: Yerra Nithin Manoj
Roll Number: B22CS066
Course: Natural Language Understanding
Assignment: Assignment 1 — Problem 4

Introduction

Text classification is one of the core tasks in Natural Language Processing where the goal is to automatically assign categories to text based on its content. With the huge amount of digital information available today, automatic classification systems help organize data efficiently and make information easier to search and analyze.

In this problemt, the goal is to build a classifier that can distinguish between Sports and Politics related text. These two domains were chosen because they have different styles of writing and vocabulary, which makes them suitable for testing classification models. The project also focuses on comparing different machine learning techniques to understand how they perform on the same dataset and which approach works best for this kind of task.

Dataset Collection

For this project, text samples were collected from the BBC News dataset, which is commonly used for text classification research. The dataset contains real news articles categorized into different topics such as sports, politics, business, and technology. From this dataset, only sports and politics related content was selected.

Instead of using entire long articles, the text was divided into smaller sentence-level samples. This helps increase the number of training examples and allows the models to learn patterns more effectively. Using real news data makes the classification task more realistic and ensures that the vocabulary reflects real-world language usage.

Dataset link

Dataset Analysis

After preparing the dataset, it was observed that sports related samples frequently contain words such as player, match, medal, training, and competition. On the other hand, politics samples include words such as government, policy, parliament, election, and minister.

The sentences in the dataset vary in length, usually between 8 and 20 words. The dataset is balanced between the two categories, which is important because it prevents the model from becoming biased toward one class. Overall, the vocabulary differences between the two domains make it easier for machine learning models to identify patterns.

Feature Representation

Before training machine learning models, text data needs to be converted into numerical form. There are several ways to represent text numerically, including Bag of Words, n-grams, and TF-IDF. In this project, TF-IDF was used because it provides a good balance between simplicity and effectiveness.

TF-IDF works by assigning higher importance to words that appear frequently in a document but are less common across other documents. This helps the model focus on meaningful words instead of very common ones like “the” or “is”. Because of this, TF-IDF is widely used in text classification tasks and works well for datasets with clear vocabulary differences.

Machine Learning Models

Three different machine learning models were implemented and compared in this project.

The first model is **Naive Bayes,** which is a probabilistic classifier based on Bayes’ theorem. It assumes that features are independent, which is not always true in real life, but surprisingly it works very well for text classification problems because word frequencies provide strong signals.

The second model is **Logistic Regression**, which is a linear classification algorithm that estimates probabilities using a logistic function. It is simple but effective and often performs well when the relationship between features and classes is linear.

The third model is **Support Vector Machine (SVM)**, which is a more advanced algorithm that tries to find the best boundary separating different classes. It is particularly useful when dealing with high-dimensional data like text features.

Results and Comparison

The performance of the models was evaluated using accuracy as the main metric. The results showed that Naive Bayes achieved an accuracy of 1.00, while Logistic Regression and SVM both achieved an accuracy of 0.83.

Model	Accuracy
Naive Bayes	1.00
Logistic Regression	0.83
SVM	0.83

The high performance of Naive Bayes can be explained by the fact that it works very well with small datasets and is able to capture word distribution effectively. Logistic Regression and SVM also performed well but were slightly affected by the limited amount of training data.

Limitations

One limitation of this project is the relatively small dataset size. Although sentence-level samples were created, the dataset may not fully represent

the diversity of language used in real-world news articles. Some words could also appear in both domains, which may lead to occasional misclassification.

Another limitation is that the models rely only on surface-level text features and do not consider deeper contextual meaning. More advanced approaches such as deep learning models could potentially improve performance by capturing semantic relationships between words.

Conclusion

This project demonstrated how machine learning techniques can be used to classify text into sports and politics categories. The results show that even simple models like Naive Bayes can perform extremely well when the dataset has clear vocabulary differences. Logistic Regression and SVM also showed strong performance, highlighting their effectiveness for text classification tasks.

Overall, this problem helped in understanding the importance of feature representation, dataset preparation, and model comparison. Future improvements could include using a larger dataset, experimenting with n-gram features, and exploring more advanced natural language processing techniques.