Sports vs Politics News Classification

B23EE1069

1 Introduction

In this project, I built a text classifier that can predict whether a news article belongs to sports or politics. This type of classification is useful in news filtering systems. The aim was to compare different machine learning models and see how they perform on text data.

2 Dataset

I used the BBC News dataset. It contains labeled news articles from different categories. I selected only sports and politics articles.

After filtering, the dataset contained 620 articles:

- Sports: 346
- Politics: 274

3 Feature Representation

Text was converted into numerical form using TF-IDF vectorization with unigram and bigram features. This method helps give importance to useful words and short phrases.

4 Models Used

I compared three models:

- Naive Bayes
- Logistic Regression
- Support Vector Machine (SVM)

5 Results

5.1 60% Test Split

Model	Accuracy
Naive Bayes	99.73%
Logistic Regression	99.19%
SVM	99.73%

5.2 75% Test Split

Model	Accuracy
Naive Bayes	99.57%
Logistic Regression	99.14%
SVM	99.57%

5.3 90% Test Split

Model	Accuracy
Naive Bayes	93.91%
Logistic Regression	82.97%
SVM	98.57%

All models performed very well because sports and politics articles use very different words. SVM and Naive Bayes usually gave the highest accuracy. Logistic Regression was slightly lower, especially when the training data was very small. Accuracy dropped when the test size increased to 90% because less data was available for training. The dataset is small and only contains two categories. The model also does not understand the meaning of text, it only learns word patterns. This project shows that simple machine learning methods with TF-IDF features can achieve very high accuracy for text classification tasks when categories are clearly different.