**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

Text classification or assigning fixed categories to textual documents is one of the fundamental problems in Natural Language Processing (NLP). The automatic classification has become essential in organizing information, filtering the contents and improving the systems used in finding information, especially on online news articles because of the phenomenal increase in digital information especially online news articles.

News stories often feature a wide range of issues, such as politics, sporting activities, technology, and entertainment. These contents are time-consuming and impractical to be categorised manually in large scale. Machine learning-based methods of text classification are thus commonly used to automate this process. These techniques are based on the training models to find the patterns associated with different categories and extract meaningful numerical characteristics of text.

This is a project that explores a binary text classification whereby documents are classified as either politics or sports. Besides the creation of an effective classifier, the purpose of this work is to explore the impact of different feature representations and machine learning models on the classification performance rates. With the help of quantitative assessment, various methods are compared and their pros and cons are discussed.

## 2. Dataset Collection and Description

To collect more realistic and larger data, this study collected an open dataset on Kaggle news which is openly available. The dataset contains thousands of news items that are labeled according to different categories. In the classification task, an article that belonged to the Sports and Politics categories in this collection was picked.

Each category was selected to contain approximately 1000 documents to ensure equal training and evaluation, and it produced approximately 2000 news articles. Each article consists of the news headline and a short description, and one document is managed as another data sample. Each line of the two files, sports.txt and politics.txt has one document each, but the articles in these files are filtered and in plain text format.

Sports documents cover football, cricket, tennis, basketball, the Olympics and any other competitive sports. These texts are often covered with matches, players, teams, competitions, performance data, and championship results. Due to this fact, Domain specific terms, which are related to the competition, scoring and athletic performance are usually present in the vocabulary.

Political documents embrace government policies, elections, legislative decisions, international relations, political leaders and the administration of the people. The words used in these articles are political, policy debate, 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.

However, the disadvantage of BoW is that it assigns equal weights to all words and disregards their contextual relationship.

### 3.2 Term Frequency Inverse Document Frequency (TF-IDF)

TF-IDF improves the Bag of Words model by assigning the words weights based on their importance. Inverse document frequency reduces the weight of those words which are present in many documents, whereas term frequency is the frequency of words in a document. This representation dilutes the effect of frequent words and puts more focus on words which are more discriminatory against a specific class. TF-IDF usually leads to improved classification and more informative features.

### 3.3 TF-IDF with N-grams

Although TF-IDF and BoW treat words independently, they cannot identify short contextual phrases. This constraint was overcome by means of combining bigrams (n = 2) with the TF-IDF. The use of n-grams (sets of neighbouring words) helps the model to identify phrases such as prime minister, world cup and national election. The semantic meaning of such phrases is often larger than the semantic meaning of separate words. However, the use of n-grams increases the dimensionality of the feature space and it may require more data to eliminate sparsity issues.

---

## 4. Machine Learning Models

The three popular machine learning models were selected to perform classification. The models are a good balance between performance, interpretability and simplicity and are commonly applied in situations where one is doing their text classification.

### 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.

TF-IDF features work well with logistic Regression, which is comparatively interpretable because the weights of features, which are the weights of words, show how important they are to classification.

### 4.3 Support Vector Machine (SVM)

Support Vector machine is an effective classifier which seeks to discover a boundary between the classes that maximizes the margin between the classes. Linear SVMs are best suited where text classification is involved as text data is so high dimensional and sparse.

When used together with TF-IDF features, SVMs tend to perform highly and have a good generalization ability.

## 5. Experimental Results

### 5.1 Experimental Setup

The dataset was split into 80 and 20 sets to train and test respectively. Three feature representations (BoW, TF-IDF, TF-IDF with bigrams) were integrated with three machine learning models (Naive Bayes, Logistic Regression and SVM) which led to the combination of various models having distinct feature representations. **Accuracy, Precision, Recall,** and **F1 score** were used to measure performance.

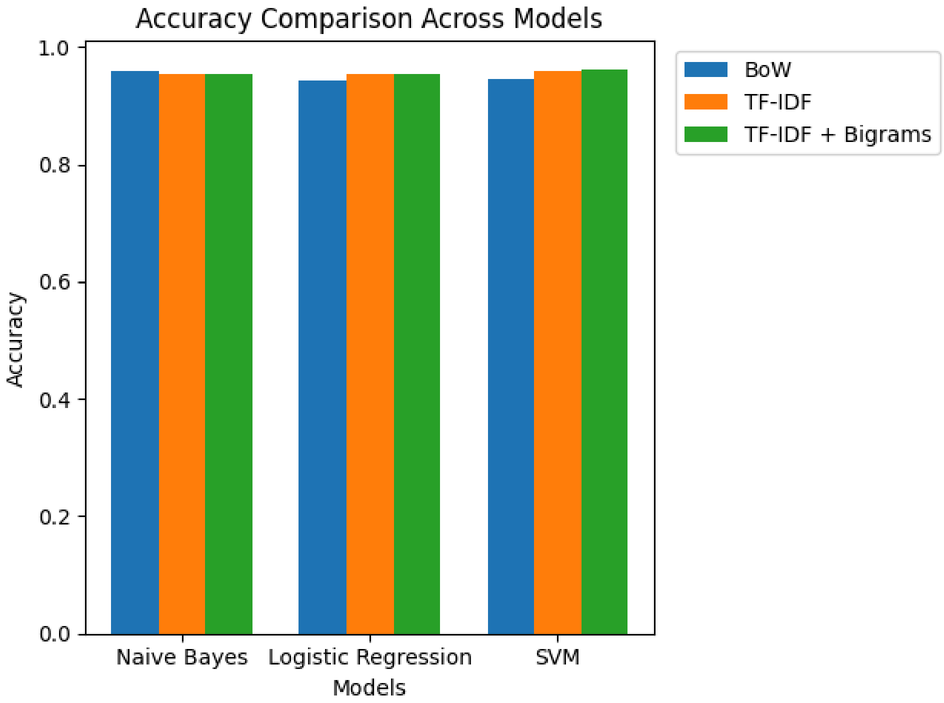
### 5.2 Results

The table below summarizes the results of the experiment:

## 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 |

To compare the accuracy of the models and feature representations, a bar chart was created.



## 6. Discussion and Limitations

The findings of the experiment demonstrate that all the three machine learning models are highly accurate using various feature representations and also do well on the sports and politics classification task. Although the Logistic Regression is also quite competitive with minor differences in the execution that arise with different sets of features, Naive Bayes and the Support Vector machine always perform quite well. Moreover, the findings indicate that since representations based on TF-IDF place more emphasis on the discriminative terms that are more informative in regards to the classification, they tend to have a little bit more stable results when compared to the basic Bag-of-Words features.

The overall high performance of the model is an indicator that the process of separating the political and sports domains has lexical distinctions that can be effectively separated through the use of standard machine learning methods. The models do not change in their performance with the addition of bigram features implying that contextual pairs of words provide more useful information without significantly reducing the classification accuracy.

The data considered in this research remains a partial collection of the real news material, even though this is much larger than a small manually created dataset. The broad differences in language of news, styles of writing and evolving events in real deployment scenarios can make them more complex. Moreover, the task of classification only takes into account two classes, but it is often required to use multi-class classification and a variety of domains in real-life systems. Hence, additional work might require the expansion of the dataset further, the inclusion of new categories, and exploration of state-of-the-art methods in deep learning in order to become stronger and more generalised.

## 7. Conclusion

This project has been based on real world news content, gathered in a publicly accessible data set, to develop a machine learning based classifier that can differentiate between political and sport related news articles. Their application and comparison exposed the performance of the various representations of features and classifier models in the text classification exercise.

The results of the experiments show the machine learning methods, which are considered as traditional, can be applied to achieve good outcomes when they are accompanied with the suitable textual feature representations. Despite the results of Naive Bayes and Logistic regression models being also appealing, the Support Vector machine model was most reliable and consistent to various sets of features compared to other models which were tested.

Moreover, we can also note that the performance of TF-IDF-based representations is equal, and it can capture discriminative vocabulary patterns that occur in various domains of news. To enhance the level of classification performance and generalisation, future studies may employ more sophisticated preprocessing and feature manipulation techniques, apply the classification algorithm to a wider range of news categories, apply larger and more diverse data, or adopt the deep learning-based systems like neural networks or transformer models.