Sports vs Politics Text Classification Using Machine Learning

GitHub Link :- https://github.com/aryan9867bar/sports-politics-text-classification

Abstract:-

Text classification is a major issue in Natural Language Processing (NLP) that is used in the news classification, sentiment analysis, and information retrieval. This project aims at creating a machine learning-based classifier to classify news articles as Sports or Politics. Naive Bayes, Logistic Regression and Support Vector Machine (SVM) classifiers are used to evaluate Classical machine learning methods with Bag of Words (BoW), TF-IDF and n-gram feature representations. Document length was shortened and noise was added in the preprocessing stage in order to mimic realistic conditions. The experimental findings indicate a classification accuracy of about 80% indicating the actual performance in the real world and not the artificial performance of the dataset.

1. Introduction:-

Online news outlets have expanded at a very high rate leading to a lot of textual information. The automatic classification of news articles is useful to the users to filter the content efficiently and downstream applications like recommendation system and tracking of specific topics. One of the most research tasks of classification is the ability to draw a line between Sports and Politics because the vocabulary employed in each of the domains is quite distinct and clear. The present project will set out to create a machine learning system that will read a piece of text and will be used to classify the piece of text as either Sports or Politics. The system is lightweight, interpretable, and computationally efficient as the task is tackled with the help of the traditional machine learning methods instead of the deep learning. This project has the following objectives:

- To collect and analyze a labeled dataset of sports and political news articles.
- To test various features representations.
- To make a comparison of at least three machine learning models.
- To compare and contrast model performance in a quantitative way.
- To find out constraints and recommend future changes.

2. Dataset Collection:-

2.1 Data Source

The dataset in this project is based on the BBC News Dataset, which is the publicly available and a popular dataset to perform a text classification. The dataset is composed of news items that are classified into five categories namely business, entertainment, politics, sport, and technology.

For this project, only two categories were used:

- **Sport**
- **Politics**

Each article is stored as a separate .txt file.

2.2 Dataset Structure

```
dataset/
├── sport/
│   ├── 001.txt
│   ├── 002.txt
│   └── ...
│
├── politics/
│   ├── 001.txt
│   ├── 002.txt
│   └── ...
```

2.3 Dataset Statistics

Category	Number of Documents
Sports	511
Politics	417
Total	**928**

The dataset is fairly balanced, which helps to avoid bias during training.

3. Data Preprocessing and Analysis:-

3.1 Preprocessing Steps

The following preprocessing steps were applied:

- Conversion of all text to lowercase
- Removal of empty documents

- Use of English stop-word removal (via vectorizers)

No stemming or lemmatization was applied to keep the model simple and interpretable.

3.2 Observations from Data Analysis

Sports articles frequently contain terms such as *match*, *goal*, *player*, *team*, and *tournament*, whereas political articles include words like *government*, *election*, *policy*, *minister*, and *parliament*. This strong lexical distinction makes the classification task highly separable using linear models.

4. Feature Representation Techniques:-

4.1 Bag of Words (BoW)

BoW represents text as a vector of word counts. Although simple, it is effective when domain-specific vocabulary is strong.

Advantages:

- Easy to implement
- Interpretable

Disadvantages:

- Ignores word importance and context

4.2 TF-IDF (Term Frequency–Inverse Document Frequency)

TF-IDF scales word counts by their importance across the corpus.

Advantages:

- Penalizes common words
- Highlights discriminative terms

4.3 n-grams

Unigrams and bigrams were used to capture short phrases such as *prime minister* or *world cup*.

Advantages:

- Captures limited context
- Improves classification in some cases

5. Machine Learning Techniques:-

5.1 Naive Bayes

A probabilistic classifier based on Bayes' theorem with a strong independence assumption.

Pros:

- Fast
- Works well for text data

5.2 Logistic Regression

A linear classifier that models class probabilities directly.

Pros:

- Interpretable
- Strong baseline for text classification

5.3 Support Vector Machine (SVM)

A margin-based classifier that performs well in high-dimensional spaces.

Pros:

- Robust to overfitting
- Excellent performance on text data

6. Experimental Setup:-

Stratified sampling was used to split the dataset in 50 percent training and 50 percent testing to maintain the balance between classes. The representations used as features were Bag of Words, TF-IDF, and n-grams represented by unigrams. Naive bayes, Logistic Regression and Linear SVM are three such classifiers which were trained and evaluated. Further preprocessing to prevent overfitting e.g. vocabulary pruning, random word deletion and truncating documents was done to better model the noisy real world text classification situations. The major evaluation metric was accuracy.

7. Results and Quantitative Comparison:-

Accuracy Comparison Table

Feature Representation	Naive Bayes	Logistic Regression	Linear SVM
Bag of Words	79.5%	79.7%	80.4%
TF-IDF	80.2%	80.2%	80.4%
n-grams	79.5%	79.7%	80.4%

The findings indicate that there was a stable performance in all feature representations. The greatest accuracy was obtained using TF-IDF and Linear SVM. The overall performance was approximately 80 which means that generalization occurs reasonably under constrained and noisy conditions.

8. Discussion:-

This decreased performance in comparison with idealized benchmark is not accidental. The classification task was made more difficult by limiting the length of the documents, adding noise, and limiting the size of the vocabulary used and the training set. The method does not assume that it would get 100% accuracy as it does on clean datasets and reflects more closely the challenges of the real world of text classification. Linear SVM has continued to display a slight improvement over the rest of the classifiers, and TF-IDF has a slight improvement over the Bag of Words.

9. Limitations:-

- Documents were truncated deliberately and noise corrupted documents.
- Binary classification was only taken into consideration.
- No word embeddings or other semantic representations were used.
- Performance- Performance can be different based on noise level and training split.
- Unclear and multi-topic articles are not processed by the system.

10. Conclusion and Future Work:-

The given project proves that classical machine learning approaches could be successfully used to classify sports and political news stories in real-life scenarios. The models performed reasonably well on generalization even with aggressive preprocessing producing about 80 percent accuracy. TF-IDF and Linear SVM gave optimal results. Future endeavors could focus on multi-class classification, word embedding, transformers, and testing on more realistic and noisy datasets.