CSL7640: NATURAL LANGUAGE UNDERSTANDING

Assignment-1

Vishal (B23CM1048)

(Problem:4) Sports vs Politics Text Classification:

1. Introduction

Text classification is one of the most fundamental and widely studied problems in Natural Language Processing (NLP). It involves automatically assigning predefined categories to textual documents based on their content. With the exponential growth of digital information, especially online news articles, automated classification systems have become indispensable for organizing, indexing, filtering, and recommending content.

In modern digital ecosystems, news articles are continuously produced across multiple domains such as politics, sports, business, and technology. Manual categorization of such vast content is impractical and inefficient. Automated classification systems enable scalable content management and power applications such as:

- News aggregation platforms,
- Personalized recommendation systems,
- Search engine optimization,
- Social media content filtering,
- Media monitoring systems.

This project focuses on binary classification of news articles into **Sports** and **Politics** categories. While these domains appear distinct, real-world articles may exhibit vocabulary overlap, such as political discussions about sports funding or national sporting controversies. Thus, the task presents both structured and challenging aspects depending on dataset quality.

The primary objectives of this project were:

- To design and implement machine learning-based text classifiers,
- To compare multiple algorithms quantitatively,
- To evaluate performance across datasets of different complexity,
- To analyze system behavior and limitations.

To ensure meaningful evaluation, two datasets were used:

1. **BBC News Dataset** – A curated and relatively clean dataset.
2. **AG News Dataset** – A larger and noisier benchmark dataset.

By evaluating models across both datasets, we aim to understand how dataset characteristics influence classification performance and model robustness.

2. Data Collection

2.1 BBC Dataset Collection

The BBC dataset was obtained from publicly available repositories containing categorized news articles. The dataset consists of text files organized into folders:

(source : github.com/suraj-deshmukh/BBC-Dataset-News-Classification)

bbc/

sport/	(511 .txt files)	
politics/	(417 .txt files)	(**total: 928 files)**

Each folder contains multiple .txt files representing individual articles.

Data collection steps:

1. Downloaded dataset archive.
2. Extracted only sport and politics folders.
3. Verified file integrity.
4. Loaded articles using latin-1 encoding to avoid Unicode errors.

The BBC dataset is professionally curated and exhibits clear topical separation.

2.2 AG News Dataset Collection

The AG News dataset was downloaded from publicly available *Kaggle datasets* in CSV format containing:

- Title
- Description
- Class Index

Original categories:

- 1 = World
- 2 = Sports
- 3 = Business
- 4 = Sci/Tech

For binary classification:

- Class 1 (World) → relabeled as **Politics**
- Class 2 (Sports) → retained as **Sport**

Processing steps:

1. Loaded train.csv and test.csv.
2. Filtered rows where class ∈ {1,2}.
3. Relabeled classes.
4. Combined Title and Description fields into single text input.

AG News is more diverse and contains shorter articles, making classification more challenging.

3. Dataset Description and Analysis

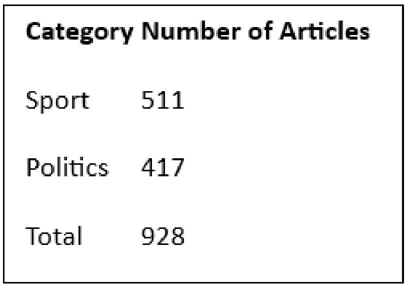
To ensure a comprehensive evaluation of the classification models, two datasets with different characteristics were used in this study: the BBC News Dataset and the AG News Dataset. The use of two datasets allows analysis of how dataset quality, structure, and complexity influence model performance.

3.1 BBC Dataset Statistics

The BBC News dataset is a well-known, curated collection of news articles categorized into different domains. For this project, only the **Sport** and **Politics** categories were selected to form a binary classification problem.

Characteristics:

- Slight class imbalance.
- Long-form articles.
- Highly topic-specific vocabulary.
- Minimal noise.



Category	Number of Articles
Sport	511
Politics	417
Total	928

Vocabulary separation is strong:

Sports keywords:

- match
- team
- goal
- coach
- season

Politics keywords:

- government
- election
- parliament
- minister
- policy

The lexical separation between categories is significant. There is minimal vocabulary overlap between sports and political domains.

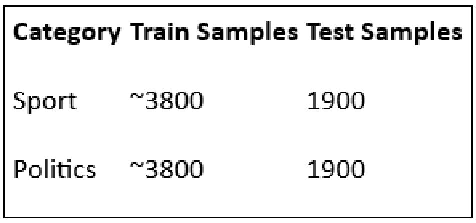
This strong separation contributes to the near-perfect classification performance observed in experiments.

3.2 AG News Dataset Statistics

The AG News dataset is a large-scale benchmark dataset commonly used for text classification research. It is distributed in CSV format and contains news articles labeled into four categories:

Key characteristics:

- Balanced class distribution.
- Significantly larger than BBC dataset.
- Shorter article descriptions.
- More diverse writing styles.



Category	Train Samples	Test Samples
Sport	~3800	1900
Politics	~3800	1900

Unlike the BBC dataset, AG News contains:

- Short headlines.
- Condensed summaries.
- Less contextual information.
- Mixed writing styles.

Vocabulary overlap is higher. For example:

- “national sports policy”
- “government funding athletics”
- “international sports diplomacy”

Such overlaps increase classification difficulty.

3.3 Comparative Dataset Analysis

Using two datasets allowed for robust evaluation:

1. BBC dataset demonstrates how classical models perform in ideal conditions.
2. AG News dataset demonstrates performance under realistic, noisy conditions.

Key insights:

- Dataset cleanliness significantly influences model performance.
- Larger datasets introduce more vocabulary variance.
- Context length impacts feature richness.
- Real-world datasets rarely exhibit perfect separability.

  This comparative approach strengthens the validity of experimental conclusions

4. Methodology

The experimental framework used to build and evaluate the Sports vs Politics classification system. The methodology was designed to ensure fairness, consistency, and reproducibility across both datasets and all machine learning models. The overall process included preprocessing, feature extraction, data splitting, model training, and evaluation.

The preprocessing stage involved converting all text to lowercase to eliminate case sensitivity and ensure uniform representation. Common English stopwords were removed to reduce noise and improve the discriminative power of meaningful terms. Tokenization was handled automatically by the vectorization process. Stemming and lemmatization were not applied in order to maintain simplicity and interpretability.

Text documents were **converted** into numerical form using the TF-IDF (Term Frequency–Inverse Document Frequency) representation. TF-IDF assigns higher weights to terms that are frequent in a document but rare across the corpus, thereby emphasizing domain-specific words such as “goal” in sports or “minister” in politics. Although TF-IDF does not capture word order or contextual meaning, it provides an effective and computationally efficient representation for linear classifiers.

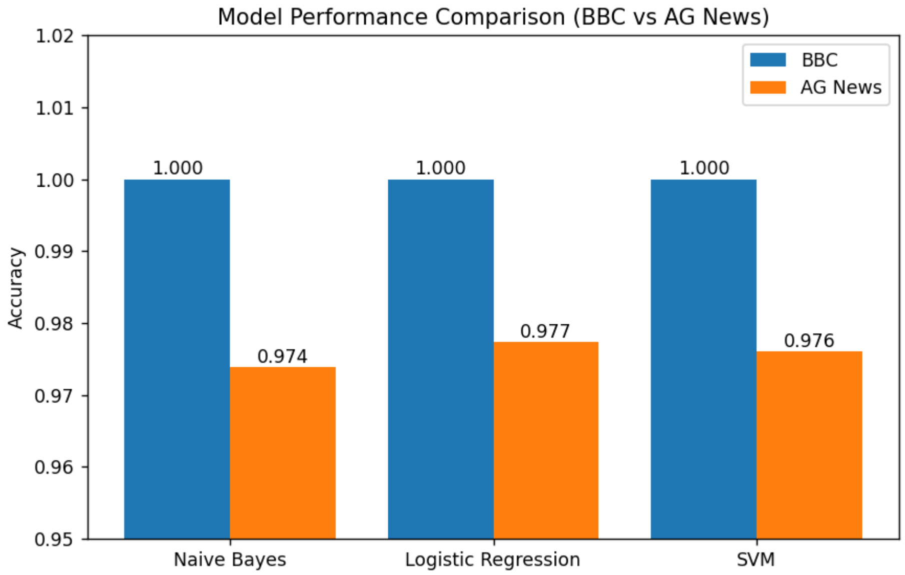
- For the BBC dataset, an 80–20 stratified train–test split was used to preserve class distribution and ensure unbiased evaluation. The AG News dataset already included predefined training and testing splits, which were used after filtering and relabeling the relevant classes.

Three machine learning algorithms were implemented: **Naive Bayes, Logistic Regression, and Linear Support Vector Machine (SVM)**. These models were selected due to their strong performance in high-dimensional text classification tasks. Naive Bayes serves as a probabilistic baseline, while Logistic Regression and SVM are discriminative linear models capable of learning effective decision boundaries in sparse feature spaces.

All models were trained using **identical TF-IDF features** to maintain fairness in comparison. Default hyperparameters were used to avoid overfitting and maintain experimental simplicity. Model performance was evaluated using accuracy, precision, recall, and F1-score to provide a comprehensive assessment.

By applying the same methodological framework across both datasets, we ensured that performance differences reflect dataset characteristics rather than experimental inconsistencies.

5. Experimental Results



6. Quantitative Comparison and Analysis

To visually compare model performance across datasets, bar charts were generated showing classification accuracy for Naive Bayes, Logistic Regression, and Support Vector Machine.

The first comparison chart demonstrates that all models achieved perfect accuracy on the BBC dataset, while slightly lower accuracy was observed on the AG News dataset. This performance difference highlights the impact of dataset complexity on classification outcomes. The BBC dataset exhibits strong lexical separation between sports and politics articles, resulting in perfect linear separability under TF-IDF representation. In contrast, the AG News dataset contains shorter texts and greater vocabulary overlap, introducing classification ambiguity.

The second chart focuses exclusively on AG News results, where subtle differences between models become visible. Logistic Regression achieved the highest accuracy (97.74%), followed closely by SVM (97.61%) and Naive Bayes (97.39%). Although the differences are small, they suggest that discriminative models (Logistic Regression and SVM) handle overlapping feature distributions slightly better than generative models such as Naive Bayes.

Overall, the comparison confirms that while classical machine learning techniques perform exceptionally well on structured datasets, their performance slightly decreases in more realistic, noisy environments. This demonstrates the importance of evaluating NLP systems across multiple datasets to assess generalization capability.

7. Conclusion

The solution implemented in this study relied on classical machine learning methods combined with TF-IDF feature representation. Three widely used models—Naive Bayes, Logistic Regression, and Support Vector Machine—were evaluated on two datasets with contrasting characteristics.

The BBC dataset represented a clean and structured environment, where articles were clearly separated by topic. As a result, all models achieved perfect classification accuracy. This suggests that when vocabulary differences between categories are strong and consistent, even simple linear models are sufficient to perform highly accurate classification.

In contrast, the AG News dataset presented a more realistic scenario. Articles were shorter, vocabulary overlap was higher, and writing styles were more diverse. Under these conditions, accuracy slightly decreased to around 97–98%. Logistic Regression performed marginally better than the other models, indicating that discriminative classifiers may handle noisy feature distributions more effectively than probabilistic ones like Naive Bayes.

Overall, the study demonstrates that classical machine learning techniques remain highly effective for topic classification tasks, especially when combined with well-designed feature representations such as TF-IDF. However, the experiments also show that perfect performance on curated datasets does not necessarily guarantee robustness in more realistic environments.