NLU ASSIGNMENT (QUESTION NO.4)

Name: **Kunal Kumar Mishra**
Roll: M25CSA036
Github Link:https://github.com/Givhel/NLU-Assignment-Problem-4

Answer:

1. Introduction

Natural Language Processing (NLP) is an important area of Artificial Intelligence that focuses on enabling machines to understand, analyze, and generate human language. One of the most widely studied problems in NLP is **text classification**, where a system automatically assigns predefined categories to textual documents. Text classification has numerous real-world applications such as news categorization, sentiment analysis, spam detection, and information retrieval.

With the rapid growth of digital news platforms, large volumes of textual data are generated every day. Manually organizing and classifying this information is inefficient and time-consuming. Therefore, automated text classification systems are essential for managing and retrieving information efficiently. Machine learning techniques, combined with appropriate feature extraction methods, have proven to be highly effective for such tasks.

This project focuses on building a **Sports vs Politics text classification system** using machine learning approaches. Sports and Politics are two distinct news categories that often exhibit different vocabulary usage, writing styles, and contextual patterns. By exploiting these differences, machine learning models can be trained to accurately distinguish between the two classes.

In this assignment, multiple machine learning classifiers are implemented and compared to analyze their performance on a real-world dataset. The study aims to demonstrate how different algorithms behave on the same classification task when combined with a common feature representation technique. The results of this comparison help in understanding the strengths and limitations of each model.

The remainder of this report is organized as follows: the problem statement is described next, followed by details about the dataset collection and preparation. Subsequently, the feature extraction method and machine learning models used in the study are explained. Finally, the experimental results, analysis, limitations, and conclusions are presented.

2. Problem Statement

The objective of this assignment is to **design and implement a text classification system** that reads a given text document and classifies it into one of two categories: **Sports** or **Politics**. This task is a binary text classification problem, which is a fundamental application of Natural Language Processing and Machine Learning.

As specified in the problem statement, the classifier must use **machine learning techniques** and appropriate **feature representation methods** such as Bag of Words, n-grams, or TF-IDF. Additionally, the assignment requires a **comparative analysis of at least three different machine learning models** in order to evaluate their effectiveness for this task.

The problem also emphasizes the importance of a **detailed experimental study**, including how the dataset was collected, how it was analyzed, and how different models performed quantitatively. Therefore, this project not only focuses on building a working classifier, but also on understanding the behavior and performance of different algorithms under the same experimental conditions.

To fulfill all the requirements of the assignment, this work implements three widely used machine learning classifiers and evaluates them using standard performance metrics. The complete methodology, experimental setup, results, and limitations of the system are documented in this report, along with a publicly accessible GitHub repository containing the implementation details.

3. Dataset Collection, Description, and Analysis:

In accordance with the requirements of the assignment, a dataset was collected to train and evaluate a classifier that distinguishes between Sports and Politics text documents. For this task, a publicly available and widely used news dataset was selected to ensure the reliability and relevance of the experimental results.

3.1 Dataset Collection:

The dataset used in this study is derived from the **BBC News dataset**, which contains news articles collected from the British Broadcasting Corporation (BBC). This dataset is commonly used in Natural Language Processing research and academic assignments for text classification tasks. The dataset was obtained from a public repository and downloaded in the form of a compressed archive.

After extraction, the dataset was organized into multiple category-specific folders such as *business*, *entertainment*, *technology*, *sports*, and *politics*. In order to meet the specific requirements of this assignment, only the **Sports** and **Politics** categories were selected. This resulted in a binary classification problem, as required in the problem statement.

Each article in the dataset is stored as a separate text file. All the text files from the *sport* folder were labeled as **Sport**, while all the text files from the *politics* folder were labeled as **Politics**. This folder-based labeling approach provides a clear and structured way to associate documents with their corresponding classes.

3.2 Dataset Description:

The final dataset used for experimentation consists of **928 text documents**, out of which:

- sports documents: 510
- Politcis documents: 418

Each document represents a complete news article written in English. The articles vary in length and writing style, which helps in capturing diverse linguistic patterns within each category. The dataset covers a wide range of topics within sports, such as matches, tournaments, players, and teams, as well as political topics including elections, government policies, parliamentary debates, and international relations.

The dataset is relatively balanced between the two classes, which helps prevent bias toward a particular category during model training. The availability of real-world news articles makes this dataset suitable for evaluating machine learning classifiers in a realistic text classification scenario.

3.3 Dataset Analysis:

A preliminary analysis of the dataset reveals that sports and politics articles exhibit distinct vocabulary and contextual patterns. Sports-related documents frequently contain terms related to games, scores, players, and competitions, whereas political articles commonly include terminology associated with governance, elections, legislation, and diplomacy.

Such differences in word usage make this dataset well-suited for machine learning–based classification approaches. Since the dataset contains natural, unstructured text, it presents realistic challenges such as varying document lengths, different writing styles, and domain-specific vocabulary. These characteristics allow for a meaningful evaluation of different machine learning models and feature representation techniques.

The dataset was further divided into training and testing subsets to evaluate the generalization capability of the classifiers. A stratified train-test split was used to maintain the class distribution in both subsets, ensuring a fair and reliable evaluation of model performance.

To further analyze model robustness, additional experiments were conducted under constrained and noisy conditions, which are discussed in later sections.

4. Text Preprocessing and Feature Representation

Before applying machine learning algorithms, the raw text documents must be converted into a suitable numerical format. Text data is inherently unstructured, and therefore several preprocessing and feature extraction steps are required to make it usable for machine learning models.

In addition to standard preprocessing, a controlled noise mechanism was applied in certain experiments by randomly removing a fraction of words from each document. This was done to simulate real-world scenarios where text data may be incomplete or noisy, thereby making the classification task more challenging.

4.1 Text Preprocessing:

The dataset consists of raw news articles containing natural language text. Minimal but effective preprocessing steps were applied to preserve meaningful information while reducing noise. Each document was read as a complete text file and processed as a single input sample. All text was handled in a consistent encoding format to avoid character-related issues.
Common preprocessing techniques such as stop-word removal were applied to eliminate frequently occurring but less informative words (for example, “the”, “is”, “and”). This helps in reducing dimensionality and improving model performance. No aggressive text normalization techniques such as stemming or lemmatization were applied, as the TF-IDF representation is capable of handling lexical variations effectively.

4.2 Feature Representation:

To transform the preprocessed text into numerical features, the **Term Frequency–Inverse Document Frequency (TF-IDF)** technique was used. TF-IDF is a widely used feature representation method in text classification tasks because it reflects the importance of a word in a document relative to the entire corpus.

In this work, **n-gram features** were incorporated by using both **unigrams and bigrams**. Unigrams capture individual word occurrences, while bigrams capture short contextual information by considering pairs of consecutive words. This combination allows the model to learn both basic vocabulary and limited contextual patterns.

TF-IDF weighting reduces the influence of very common words that appear across many documents while emphasizing terms that are more discriminative for a particular class. This makes it especially effective for distinguishing between sports and politics articles, which often differ in domain-specific terminology.

The TF-IDF representation was generated using a standard vectorization approach, resulting in a high-dimensional but sparse feature space. This feature space is well-suited for linear machine learning models such as Naive Bayes, Logistic Regression, and Support Vector Machines, which were used in this study.

5. Machine Learning Techniques Used:

In order to satisfy the requirement of comparing multiple machine learning approaches, three widely used text classification algorithms were implemented in this study. These models were selected because of their effectiveness in handling high-dimensional text data and their frequent use in Natural Language Processing applications.

All the models were trained using the same dataset and feature representation to ensure a fair and consistent comparison.

5.1 Multinomial Naive Bayes

Multinomial Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem. It assumes that features are conditionally independent given the class label. Despite this simplifying assumption, Naive Bayes has been shown to perform remarkably well for text classification tasks.

In this study, the Multinomial Naive Bayes classifier was used because it is well-suited for word frequency–based features such as TF-IDF. The model computes the probability of a document belonging to a class based on the distribution of words observed in the training data. Its simplicity and computational efficiency make it a strong baseline for comparison.

5.2 Logistic Regression:

Logistic Regression is a discriminative machine learning model commonly used for binary classification tasks. It models the relationship between input features and class labels using a linear decision boundary combined with a logistic function.

For text classification, Logistic Regression performs well when used with TF-IDF features. It learns weights for each feature that indicate the importance of words or n-grams in predicting the class label. In this project, Logistic Regression was selected due to its robustness, interpretability, and strong performance in high-dimensional feature spaces.

5.3 Linear Support Vector Machine (SVM):

Support Vector Machines are powerful classifiers that aim to find an optimal decision boundary that maximizes the margin between different classes. In this work, a **Linear SVM** was used because it is computationally efficient and particularly effective for text classification problems involving sparse and high-dimensional data.

The Linear SVM classifier focuses on identifying the most informative features that contribute to separating the two classes. Its ability to generalize well even with a large number of features makes it a popular choice for document classification tasks. This model often achieves high accuracy when combined with TF-IDF representations.

6. Experimental Setup:

The experimental setup was designed to evaluate the robustness of machine learning models under constrained and noisy conditions. While the same dataset and classification task were used, additional modifications were introduced to make the evaluation more realistic.

The dataset was split such that **60% of the data was used for training** and the remaining **40% was used for testing**. This reduced training size increases the difficulty of the task and allows for a more realistic assessment of generalization performance.

To further challenge the models, **controlled noise** was introduced by randomly removing a portion of words from each document during preprocessing. This simulates real-world text imperfections such as missing information or transcription errors.

Feature extraction was performed using TF-IDF with unigrams only, and the feature space was limited to a fixed number of features. All models were trained and evaluated using the same setup to ensure fair comparison. Performance was evaluated using accuracy, precision, recall, and F1-score.

6.1 Data Splitting:

The complete dataset was divided into **training and testing subsets** using a stratified train-test split. Specifically, **80% of the data was used for training** the models, while the remaining **20% was reserved for testing**. Stratification was applied to preserve the original class distribution of sports and politics articles in both subsets.

A fixed random seed was used during the splitting process to ensure reproducibility of the experimental results. This approach helps evaluate how well the trained models generalize to unseen data.

6.2 Model Training:

All three machine learning models—Multinomial Naive Bayes, Logistic Regression, and Linear Support Vector Machine—were trained using the same training dataset and TF-IDF feature representation. This uniform setup ensures that differences in performance can be attributed to the learning algorithms themselves rather than variations in data preprocessing or feature extraction.

Default hyperparameters were used for most models, while Logistic Regression was configured with an increased maximum number of iterations to ensure convergence. Each model was trained independently and evaluated using the same test dataset.

6.3 Evaluation Metrics:

All three machine learning models—Multinomial Naive Bayes, Logistic Regression, and Linear Support Vector Machine—were trained using the same training dataset and TF-IDF feature representation. This uniform setup ensures that differences in performance can be attributed to the learning algorithms themselves rather than variations in data preprocessing or feature extraction.

Default hyperparameters were used for most models, while Logistic Regression was configured with an increased maximum number of iterations to ensure convergence. Each model was trained independently and evaluated using the same test dataset.

7. Results and Quantitative Comparison:

This section presents the experimental results obtained from applying three different machine learning classifiers to the Sports vs Politics text classification task. The performance of each model was evaluated on the same test dataset using identical preprocessing and feature extraction methods, ensuring a fair comparison.

7.1 Overall Results:

The experimental evaluation was conducted on a robustness-oriented setup using the BBC News dataset, where controlled noise and reduced training data were introduced to obtain realistic performance estimates. The test set consisted of **372 documents**, containing both sports and politics articles.

Overall, all three machine learning models demonstrated **strong and consistent performance**, achieving classification accuracies close to **99%**. This indicates that even under constrained and noisy conditions, the classifiers were able to effectively distinguish between sports and politics news articles.

The **Multinomial Naive Bayes** classifier achieved the highest accuracy of approximately **99.46%**, showing excellent probabilistic modeling of word distributions despite the introduced noise. The **Linear Support Vector Machine** followed closely with an accuracy of approximately **98.92%**, benefiting from its ability to learn a robust decision boundary in high-dimensional feature space. The **Logistic Regression** model achieved an accuracy of approximately **98.66%**, slightly lower than the other two models but still demonstrating strong generalization capability.

Across all models, precision, recall, and F1-score values remained consistently high for both classes, indicating balanced performance without bias toward either sports or politics. The slight variations in performance among the models reflect differences in their learning mechanisms and sensitivity to noise, which provides meaningful insights for comparative analysis.

Overall, the results confirm that TF-IDF–based text representations combined with linear machine learning models are highly effective for news classification tasks, even when evaluated under more challenging and realistic conditions.

8. Discussion:

The experimental results indicate that all three machine learning models perform effectively on the Sports vs Politics classification task, even under constrained and noisy conditions. The consistently high performance across models demonstrates the strong discriminative power of TF-IDF features when applied to news text.

Multinomial Naive Bayes achieved the highest accuracy, suggesting that probabilistic approaches can remain robust despite reduced training data and partial information loss. Logistic Regression and Linear SVM also showed competitive performance, highlighting the suitability of linear classifiers for high-dimensional text data.

The small differences in accuracy and F1-scores across models reflect their varying sensitivity to noise and feature sparsity. Overall, the results confirm that the selected feature representation and machine learning techniques are well-suited for this classification problem.

9. Limitations:

Despite strong performance, the proposed system has certain limitations. First, the classification task is limited to only two categories, which simplifies the problem and may not reflect the complexity of real-world news classification. Second, the dataset consists of well-structured news articles, and performance may degrade when applied to informal or noisy text such as social media content.

Additionally, the introduction of artificial noise does not fully capture all real-world data imperfections. The models were also evaluated on a single dataset, which may limit their generalizability to other news sources or domains.

10. Conclustion:

In this work, a machine learning–based text classification system was developed to classify news articles as Sports or Politics. Using TF-IDF feature representation and three different machine learning models, the system achieved strong and reliable performance under both standard and robustness-oriented experimental settings.

The comparative analysis demonstrates that linear classifiers combined with TF-IDF features are highly effective for news classification tasks. Although the models perform well, further improvements can be made by extending the approach to multi-class classification and evaluating on more diverse datasets. Overall, this study successfully fulfills all the requirements of the assignment.