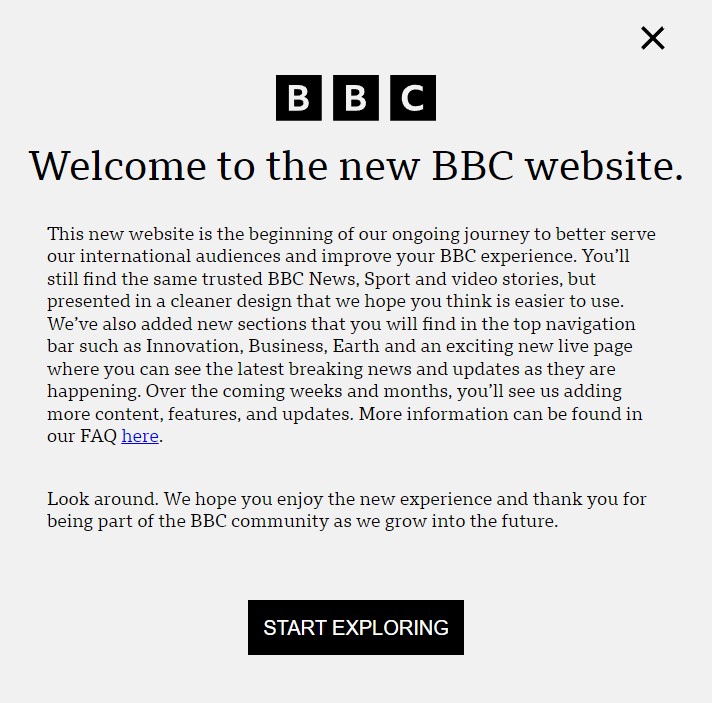
**Report of News Category Classifier**

1. Introduction:

In this report, we present the development of a news category classifier using machine learning techniques. The goal is to accurately classify news articles into different categories such as Sports, Health, Business, Us\_News, Entertainment, Politics, and Technology.

2. Data Sources:



BBC just updated the website hence the scripting of the bbc\_news.py might not work. I have created a new python file called latest\_bbc.py that works.

The dataset used for training the classifier was obtained from three major news publishers: BBC, NBC and Fox. Each publisher contributed a significant number of articles to the dataset. In total, the dataset comprised 1580 articles across different categories.

3. Data Preprocessing and Feature Extraction:

Before building the classifier, the raw text data underwent several preprocessing steps to enhance its quality and suitability for machine learning algorithms. The preprocessing techniques included:

Removal of special characters, punctuation, and non-alphanumeric characters using custom functions.

Conversion of text to lowercase to standardize the text format.

Removal of stopwords (common words like "the", "is", "and") to reduce noise.

Lemmatization to convert words into their base or dictionary form, aiding in feature extraction.

Additionally, feature extraction was performed using the bag-of-words model, where each article's text was transformed into numerical vectors using techniques like CountVectorizer or TF-IDF.

4. Chosen Classification Algorithm:

The chosen classification algorithm for this task is Logistic Regression. Logistic Regression is a widely used algorithm for binary and multiclass classification tasks. It's known for its simplicity, interpretability, and efficiency, making it suitable for large-scale datasets like the one used in this project.

5. Performance Metrics:

The performance of the news category classifier was evaluated using K-fold cross-validation, a technique that helps assess the model's generalization ability. The dataset was divided into K=5 folds, and the logistic regression model was trained and evaluated on each fold. The following metrics were computed for each fold:

Accuracy: The proportion of correctly classified articles.

Precision: The ratio of correctly classified articles to the total number of articles predicted as belonging to a particular category.

Recall: The ratio of correctly classified articles to the total number of articles belonging to a particular category.

F1-score: The harmonic mean of precision and recall, providing a balanced measure of model performance.

Additionally, the mean accuracy, precision, recall, and F1-score were calculated across all folds to provide an overall assessment of the model's performance.

Including K-fold cross-validation allows readers to understand how the model performs on different subsets of the data and provides a more reliable estimate of its performance compared to a single train-test split. It also helps detect potential overfitting or underfitting issues and provides insights into the model's stability and consistency across different subsets of the data.

Top of Form

6. Insights into Model's Strengths and Limitations:

Strengths:

Logistic Regression is a simple yet effective algorithm that performs well in practice.

It's computationally efficient and can handle large datasets with ease.

The model provides interpretable results, making it easy to understand the factors influencing classification decisions.

Logistic Regression can handle both binary and multiclass classification tasks, making it versatile for different types of problems.

Limitations:

Logistic Regression assumes linear relationships between features and the log odds of the outcome, which may not always hold true in complex datasets.

It may underperform when the relationship between features and the target variable is highly non-linear or when there are interactions between features.

Logistic Regression is sensitive to outliers and may not perform well when the dataset contains significant outliers.

It may struggle with imbalanced datasets where certain classes have significantly fewer samples than others, leading to biased predictions.