News Article Classification – Project Part 2

1. Introduction

With the exponential growth of digital content, automating the classification of news articles into categories such as *sports, politics,* and *technology* helps streamline content organization and enhance user experience. This project aims to develop and evaluate machine learning models that accurately classify news articles based on their textual content.

2. Problem Statement

The objective is to build a robust classifier that can:

- Automatically categorize news articles into predefined categories.
- Preprocess text, extract meaningful features, and apply machine learning models.
- Evaluate performance and derive insights to optimize the classification process.

3. Dataset Description

- Source: data_news.csv
- Features Used: headline, short_description (combined into text)
- Target Variable: category
- Total Categories: 10 (e.g., Sports, Politics, Entertainment, etc.)

4. Methodology

4.1 Data Preprocessing

- Combined headline and short description.
- Applied text cleaning: lowercasing, punctuation and digit removal, stopword filtering, and lemmatization.

4.2 Feature Extraction

- Utilized **TF-IDF Vectorizer** (max 5000 features, bigrams).
- Conducted Exploratory Data Analysis (EDA) to visualize category distribution.

4.3 Model Training

- Models used:
    - **Logistic Regression**
    - **Multinomial Naive Bayes**
    - **Support Vector Machine (SVM)**
- Data split: 80% training, 20% testing.

- Used 5-fold cross-validation for SVM.

5. Evaluation Metrics

- **Accuracy**
- **F1-Score (weighted)**
- **Confusion Matrix**
- **Cross-Validation Accuracy (for SVM)**

6. Results and Findings

6.1 Performance Comparison

Model	**Accuracy**	**F1 Score**
Logistic Regression	**0.7922**	**0.7924**
Naive Bayes	0.7726	0.7732
SVM	0.7919	0.7916

- **Best Classifications:** *Sports*, *Style & Beauty*, *Food & Drink*
- **SVM Cross-Validation Accuracy:** 0.7659 ± 0.0028
- **Example Prediction:** "Apple releases latest iPhone with AI camera" → **BUSINESS**

7. Conclusion

- All models showed strong performance, with Logistic Regression having the highest F1-score.
- SVM also performed consistently across categories with stable cross-validation results.
- TF-IDF with classic ML models is effective for multi-class text classification.
- Future improvements may include deep learning models like BERT for higher accuracy.