News Article Classification using NLP and Machine Learning

1. Overview
In today's digital world, classifying articles into specific categories like sports, politics, and technology can help improve content management and recommendation systems. This project aims to develop a machine learning model to classify news articles into predefined categories based on their content.

2. Problem Statement
The goal is to build a robust classifier to automatically categorize news articles into various categories. The tasks include preprocessing text, extracting features, training classification models, and evaluating performance.

3. Data Collection and Preprocessing
The dataset used contains labeled news articles with fields such as category, headline, link, short_description, and keywords. The text data was cleaned by removing missing values, lowercasing, removing punctuation, and stopword removal. The cleaned data was then ready for feature extraction.

4. Feature Extraction and EDA
TF-IDF and Bag-of-Words (BoW) methods were applied to convert text into numerical features. Each feature matrix had a shape of (50000, 5000). Exploratory data analysis showed balanced distribution across categories such as Sports, Politics, and Food & Drink.

5. Model Development and Training
Three classification models were developed: Logistic Regression, Naive Bayes, and Support Vector Machine (SVM). Data was split into training and testing sets using stratified sampling. Hyperparameter tuning and 5-fold cross-validation were applied to ensure robustness.

6. Model Evaluation
Model performance was evaluated using Accuracy, Precision, Recall, and F1 Score.

Logistic Regression:
Accuracy: 0.7882, Precision: 0.7893, Recall: 0.7882, F1 Score: 0.7885

Naive Bayes:
Accuracy: 0.7750, Precision: 0.7778, Recall: 0.7750, F1 Score: 0.7756

Support Vector Machine (SVM):
Accuracy: 0.7914, Precision: 0.7916, Recall: 0.7914, F1 Score: 0.7912

7. Conclusion
Among the models evaluated, SVM showed the best performance with an accuracy of 79.14%. It performed consistently well across all categories. This model can be effectively deployed to categorize large volumes of news articles in real-world applications.