📝 News Category Classification: End-to-End Deep Learning Pipeline with Classical Baselines

📌 Objective

To build a robust text classification pipeline for **BBC news articles** using both classical machine learning models and deep learning architectures like **LSTM** and **Bidirectional LSTM**, with the goal of accurately predicting the category of a given article.

📂 Dataset Overview

- **Source:** BBC News Dataset
- **Attributes:**
  - text: Full news article content
  - category: Target label (e.g., business, tech, sport, etc.)

🧹 Data Preprocessing Pipeline

1. **Text Cleaning:**
   - Lowercasing
   - Removal of punctuation, numbers, and stopwords
   - Tokenization and lemmatization using NLTK
2. **Label Encoding:**
   - Each category mapped to an integer using LabelEncoder
3. **Exploratory Data Analysis (EDA):**
   - **Class Distribution:** Verified balance across categories
   - **Word Frequency Analysis:** Identified most frequent terms per class
   - **Category-wise Word Clouds:** Visual insights into dominant words per topic

🔧 Feature Engineering

1. **TF-IDF Vectorization** *(for classical models):*
   - Extracted unigram and bigram features
   - Limited to top 5000 features for efficiency
2. **Tokenizer + Padding** *(for deep learning models):*
   - Used Keras Tokenizer to convert text to sequences

    - Standardized length with pad_sequences(maxlen=200)

📊 **Classical Machine Learning Models**

Model	Accuracy
Logistic Regression	88.7%
Multinomial Naive Bayes	86.9%
Linear SVM	**89.3%**
Random Forest (TF-IDF)	83.5%

> These served as strong baselines. SVM and Logistic Regression performed particularly well on sparse TF-IDF vectors.

🤖 **Deep Learning Models**

🔁 **Model 1: LSTM**

- **Architecture:**
    - Embedding layer: 128 dimensions
    - LSTM (100 units)
    - Dense + Dropout layers
    - Output: Softmax across news categories
- **Result: Test Accuracy: ~90.56%**

🔁 **Model 2: Bidirectional LSTM**

- **Architecture:**
    - Embedding: 128 dims
    - Bidirectional(LSTM(100))
    - Dropout layers for regularization
    - Dense classifier
- **Result: Improved Test Accuracy** (e.g., ≥91%) — shows enhanced contextual learning from forward & backward sequences.

🔧 **Fine-Tuning Attempts**

- Increased LSTM units and embedding dimensions

- Tuned dropout rates (0.3–0.5)
- Adjusted learning rate via custom Adam optimizer
- Planned for pretrained embeddings (e.g., GloVe) and transformer exploration (BERT)

📈 Visualizations & Evaluation

- **Accuracy and Loss Curves** (Training vs Validation)
- **Category-wise Word Clouds**
- **Confusion Matrix** *(optional step to highlight misclassification patterns)*

🔁 Model Flow Summary

text

```
Raw Dataset → Text Cleaning → Label Encoding → EDA → Feature Engineering
  ↓                    ↓
 Classical Models (TF-IDF)      Deep Learning Models (LSTM, BiLSTM)
  ↓                    ↓
  Baseline Scores        Embedding + Sequential Modeling → Evaluation
                    ↓
                Model Comparison & Fine-Tuning
```

🧠 Conclusions

- **Classical Models** like SVM and Logistic Regression provide strong baselines with TF-IDF features.
- **LSTM architecture outperforms classical methods**, learning from sequence structure.
- **Bidirectional LSTM shows further accuracy improvement**, proving the value of contextual bidirectional flow.
- Excellent potential for further gains using pretrained embeddings (GloVe) or transformer-based models (BERT, DistilBERT).