**Resume Format**

Developed a **news topic classification** system with **97% accuracy** using the BBC News Dataset, applying robust text preprocessing, **TF-IDF with N-grams**, and a **Multinomial Naive Bayes** model.

**Project Summary: BBC News Topic Classifier**

This project developed a **news topic classifier** using the BBC News Dataset, which contains 2,225 news articles categorized into five distinct topics: business, entertainment, politics, sport, and technology. The goal was to build a model that can accurately classify unseen news articles into these categories.

The project followed a robust methodology encompassing data cleaning, preprocessing, feature extraction, and model training, ultimately achieving an impressive **97% accuracy** on the test set.

**Key Stages and Techniques:**

* **Data Loading and Initial Inspection**: The project began by loading the BBC News Dataset. Initial data inspection revealed duplicate entries in both titles and content, necessitating a comprehensive cleaning process.
* **Data Cleaning**: A meticulous cleaning phase was undertaken to address data inconsistencies:
  + Duplicate content entries were identified and processed, ensuring that even if content had multiple associated titles, the longest and most representative title was retained.
  + Crucially, it was confirmed that no content entries had conflicting categories, ruling out issues like inconsistent labeling or human error.
  + Duplicate titles with different content entries were deemed acceptable since the classification model primarily relies on content.
* **Data Preprocessing**: The news article content underwent extensive preprocessing to prepare it for machine learning:
  + Text was converted to lowercase, and only alphabetical characters were retained, removing numbers and special symbols.
  + **Tokenization** was applied to break down text into individual words.
  + **Stop words** (common words like "the," "is") were removed, along with very short tokens (less than 3 characters), to reduce noise and focus on meaningful terms.
  + **Lemmatization** was used to reduce words to their base forms (e.g., "running" to "run"), standardizing the vocabulary.
* **Feature Extraction**: **TF-IDF (Term Frequency-Inverse Document Frequency)** was utilized to convert the processed text into numerical features.
  + The TfidfVectorizer was configured to use both **unigrams and bigrams** (ngram\_range=(1, 2)) to capture single words and contextual two-word phrases.
  + **Feature dimensionality** was managed by setting max\_features to 5000, and terms were filtered based on their document frequency (min\_df=5 to ignore very rare terms and max\_df=0.7 to ignore overly common terms).
* **Model Training and Evaluation**:
  + The dataset was split into training and testing sets using a **stratified split** (stratify=cleaned\_df['category']) to ensure an even distribution of categories in both sets, preventing class imbalance issues.
  + A **Multinomial Naive Bayes** classifier was chosen for the classification task due to its effectiveness in text classification.
  + The model was trained on the processed TF-IDF features and evaluated using a **classification report** and a **confusion matrix**.
  + The model achieved a high **97% accuracy** on the test set, demonstrating strong performance. An analysis of misclassified examples revealed cases that were even challenging for human interpretation, indicating the model's robustness.

This project successfully demonstrates the power of traditional NLP techniques combined with careful data handling for building an accurate news topic classifier.