**News Category Classifier Report**

**5.1 Data Sources and Collection**

1. **BBC News**: 613 articles across 7 categories (Business, Innovation, Culture, Earth, Entertainment, Arts, Travel).
2. **CNBC**: 700 articles across 7 categories (Economy, Technology, Travel, Climate, Entertainment, Media, Life).
3. **Reuters**: Attempted but scraper failed due to scraper blocked.

**Collection Method**:

* Automated scraping using **Selenium** to handle dynamic content and pagination.
* **BeautifulSoup** for parsing HTML and extracting article metadata (headlines, URLs, categories).
* Popup handling implemented for BBC/CNBC paywalls and consent forms.

**5.1 Data Preprocessing and Feature Extraction**

**Preprocessing Steps:**

1. **Cleaning:**
   * Removed duplicates and empty content entries.
   * Filled missing dates with mode ("Unknown date").
2. **Text Processing:**
   * Lowercasing, removed punctuation/numbers, tokenization.
   * Stopword removal and lemmatization using NLTK.
3. **Class Balancing:**
   * Consolidated categories (e.g., "Climate" + "Earth" → "Environment") to address imbalance.
   * Applied SMOTE to oversample minority classes.

**Feature Extraction:**

* TF-IDF Vectorization with max\_features=8000 and ngram\_range=(1,3) to capture phrases.
* Top 50 discriminative terms per category saved for interpretability (e.g., "AI" for Technology, "stocks" for Business).

**5.3 Model Selection and Performance**

**Algorithms Tested:**

1. **Logistic Regression:** Best performance with Macro F1 = 0.84.
2. **Random Forest:** Macro F1 = 0.82 (slower inference due to 200 trees).
3. **Linear SVM:** Macro F1 = 0.83 but higher class variance.

**Performance Metrics:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Macro F1** | **Weighted F1** | **Class Variance** |
| Logistic Regression | 0.80 | 0.80 | 0.086 |
| Random Forest | 0.80 | 0.80 | 0.079 |
| Linear SVM | 0.84 | 0.83 | 0.065 |

A graph of a logistic regression confusion matrix

AI-generated content may be incorrect.

A graph of a diagram

AI-generated content may be incorrect.

A diagram of a graph

AI-generated content may be incorrect.

**Confusion Matrix Insights**:

* Common misclassifications: "Lifestyle" and "Entertainment" (because of shared terms like "culture").
* "Environment" had the highest precision due to distinct keywords ("climate", "energy").

**5.4 Strengths and Limitations**

**Strengths**:

* **Robust Pipeline**: Handles dynamic content, popups, and multi-source data.
* **Interpretability**: TF-IDF features and category-specific terms provide actionable insights.
* **Scalability**: Modular code structure allows easy integration of new publishers.

**Limitations**:

* **Class Imbalance**: Despite SMOTE, rare categories (e.g., "Government") had limited samples.
* **Scraping Issues**: Reuters integration failed due to scraper blocking on website.
* **Contextual Understanding**: TF-IDF ignores word order, limiting semantic capture (e.g., sarcasm).