**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:**

**Logistic Regression**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Precision | Recall | F1-Score | Support |
| Category | **0** | 0.92 | 0.81 | 0.86 | 27 |
| **1** | 0.88 | 0.51 | 0.65 | 41 |
| **2** | 1.00 | 0.81 | 0.90 | 37 |
| **3** | 0.71 | 0.96 | 0.82 | 70 |
| **4** | 0.77 | 0.83 | 0.80 | 36 |

**Random Forest**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Precision | Recall | F1-Score | Support |
| Category | **0** | 0.79 | 0.81 | 0.80 | 27 |
| **1** | 0.92 | 0.54 | 0.68 | 41 |
| **2** | 1.00 | 0.86 | 0.93 | 37 |
| **3** | 0.72 | 0.91 | 0.81 | 70 |
| **4** | 0.79 | 0.83 | 0.81 | 36 |

**Linear SVM**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Precision | Recall | F1-Score | Support |
| Category | **0** | 0.82 | 0.81 | 0.85 | 27 |
| **1** | 0.88 | 0.66 | 0.73 | 41 |
| **2** | 1.00 | 0.86 | 0.93 | 37 |
| **3** | 0.80 | 0.93 | 0.86 | 70 |
| **4** | 0.78 | 0.86 | 0.82 | 36 |

**Comparative 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 |

**Confusion Matrices**

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**:

**Key Observations**

1. **Technology Class Challenges**:
   * **Low Recall**: All models struggle with Technology (51-66% recall). Misclassifications often go to Environment (e.g., 17 in Logistic Regression, 11 in SVM), suggesting overlapping keywords (e.g., "green tech").
   * **Precision**: Highest in Random Forest (91.7%) but lowest recall. SVM balances this with 65.9% recall and 81.8% precision.
2. **Environment Class Dominance**:
   * **High Performance**: Highest recall (92.9% in SVM) and precision (80-95%), likely due to distinct vocabulary. However, it attracts misclassifications from other classes (e.g., Tech, Business).
3. **Lifestyle Class Precision**:
   * **Perfect Precision**: Random Forest and SVM achieve 100% precision (no false positives). However, recall is ~86.5%, indicating some Lifestyle articles are misclassified as Business/Entertainment.
4. **Entertainment Improvements**:
   * **SVM Boosts Recall**: 86.1% recall in SVM vs 83.3% in others. Misclassifications reduced compared to Logistic Regression.
5. **Business Class Consistency**:
   * Stable performance across models (~81% recall, 78-88% precision). Misclassifications spread across Tech/Environment/Entertainment.

**Model Comparison**

* **Logistic Regression**: Simplest but underperforms, especially for Technology (51% recall).
* **Random Forest**: Strong precision for Technology (91.7%) and Lifestyle (100%) but lower recall for Tech (53.7%).
* **Linear SVM**: Best balance—higher recall for Tech (65.9%) and Environment (92.9%) with competitive precision.

**5.4 Strengths and Limitations**

**Strengths:**

1. **High Performance on Distinct Classes:**
   * **Environment:** All models achieved high recall (92–95%) and precision (80–95%) due to unique vocabulary (e.g., "sustainability," "climate").
   * **Lifestyle:** Random Forest and SVM achieved 100% precision, indicating no false positives for this class.
2. **Balanced Metrics in SVM:**
   * Linear SVM showed the best trade-off between recall and precision, especially for challenging classes like Technology (66% recall, 82% precision) and Entertainment (86% recall).
3. **Robustness to Overfitting (Random Forest):**
   * Random Forest achieved high precision for Technology (91.7%) and Lifestyle (100%), suggesting strong feature discrimination despite smaller class sizes.
4. **Interpretability (Logistic Regression):**
   * Logistic Regression, while simpler, provided a baseline for understanding feature importance, useful for debugging misclassifications.

**Limitations:**

1. **Class Imbalance Bias:**
   * The Environment class (67–70 samples) dominated predictions, causing models to over-predict it for ambiguous cases (e.g., Technology articles misclassified as Environment).
2. **Technology Class Challenges:**
   * Low recall (51–66%) across models, likely due to overlapping terms with Environment (e.g., "tech innovation" vs. "green tech").
3. **Entertainment Misclassifications:**
   * Articles were often confused with Business or Environment (e.g., Logistic Regression misclassified 5 Entertainment cases as Environment).
4. **Model-Specific Weaknesses:**
   * **Logistic Regression**: Struggled with complex patterns (e.g., 17 Technology articles misclassified as Environment).
   * **Random Forest:** Lower recall for Technology (53.7%) due to conservative predictions.
   * **SVM:** Slightly higher false positives for Business (3 FP) compared to other models.
5. **Feature Sensitivity:**
   * All models relied heavily on keyword overlap, failing to capture nuanced context (e.g., sarcasm in Entertainment or domain-specific jargon in Technology).