**REPORT ON Text Summarization using NLP Pipeline**

**Submitted by Sudhanshu Patil**

**1. Data Handling: Preprocessing, Augmentation**

**Preprocessing:**

* **Cleaning:** The raw text data was cleaned to remove unnecessary characters, whitespace, and punctuation. This step ensured that the text was in a consistent format for further processing.
* **Tokenization:** The text was tokenized into individual words or tokens, which is a crucial step for both dictionary lookup and NER.
* **Normalization:** Text normalization techniques such as lowercasing and stemming were applied to standardize the text.

**Augmentation:**

* **Synonym Replacement:** Synonyms were used to augment the dataset, increasing the diversity of the text snippets.
* **Back Translation:** Text snippets were translated to another language and back to the original language to create variations in the dataset.

**2. Modelling Choices**

**Classification Approach:**

* **Rule-Based NER:** A simple rule-based approach using regular expressions was chosen for Named Entity Recognition (NER). This approach was selected due to its simplicity and effectiveness in identifying specific patterns in the text.
* **Dictionary Lookup:** A dictionary-based approach was used to identify known entities from a domain knowledge base. This method ensures that domain-specific terms are accurately captured.

**Challenges and Solutions:**

* **Challenge:** Handling variations in text and ensuring that all relevant entities are captured.
  + **Solution:** Combined rule-based NER with dictionary lookup to improve the accuracy and coverage of entity extraction.
* **Challenge:** Ensuring the system is scalable and can handle large datasets.
  + **Solution:** Implemented efficient text processing techniques and optimized the code for better performance.

**3. Performance Results**

**Classification Metrics:**

* **Precision:** The proportion of correctly identified entities among all entities identified by the system.
* **Recall:** The proportion of correctly identified entities among all actual entities present in the text.
* **F1 Score:** The harmonic means of precision and recall, providing a single metric to evaluate the model's performance.

**Entity Extraction Metrics:**

* **Partial Metrics:** Evaluated the system on a subset of the data to ensure that domain-specific terms are correctly captured.
* **Full Metrics:** Comprehensive evaluation on the entire dataset to measure the overall performance.

**Summarization:**

* **Simple Summarization:** Generated a summary based on the number of entities extracted in each category. This approach provides a quick overview of the extracted information.

**4. Error Analysis**

**Examples of Mistakes:**

* **False Positives:** Instances where non-relevant terms were incorrectly identified as entities.
* **False Negatives:** Instances where relevant entities were missed by the system.

**Confusion Among Labels:**

* **Overlap:** Some entities may belong to multiple categories, leading to confusion in classification.
* **Ambiguity:** Ambiguous terms that can be interpreted in different ways depending on the context.

**Areas of Improvement:**

* **Improved Regex Patterns:** Refining the regular expressions to reduce false positives and false negatives.
* **Contextual Understanding:** Incorporating contextual information to better disambiguate entities.

**5. Future Work**

**Proposed Ideas:**

* **Better Data Curation:** Curating a more comprehensive and diverse dataset to improve the system's robustness.
* **Bigger Models:** Leveraging larger pre-trained models such as BERT or GPT for more accurate entity extraction and summarization.
* **Advanced Fine-Tuning:** Fine-tuning the models on domain-specific data to enhance their performance.
* **Integration with External APIs:** Integrating with external APIs for real-time entity extraction and summarization.