**Data Science role | GTM Buddy | Assignment Report**

**1. Data Handling**

#### **Preprocessing**

The dataset consists of text snippets with multi-label annotations. Preprocessing steps included:

* **Tokenization**: Breaking down text into meaningful units.
* **Stopword Removal**: Eliminating irrelevant words (e.g., "the," "is").
* **Lowercasing**: Standardizing text for case-insensitive processing.
* **Lemmatization**: Reducing words to their base forms (e.g., "running" → "run").
* **TF-IDF Vectorization**: Converting text to numeric features based on term frequency and inverse document frequency.

### **2. Modeling Choices**

#### **Classification**

For multi-label classification, a **OneVsRestClassifier** with a **Logistic Regression** base model was used. The choice was driven by:

* **Efficiency**: Logistic regression is fast and interpretable.
* **Flexibility**: OneVsRestClassifier handles multi-label problems by training a binary classifier for each label.
* **Simplicity**: Works well with smaller datasets and linear feature spaces.

**Challenges**:

* Lack of large labeled data posed limitations on model training.
* Imbalanced label distribution led to biased predictions.

**Solutions**:

* Synthetic augmentation balanced label frequencies.
* Weighted loss functions emphasized underrepresented classes.

#### **Named Entity Recognition (NER)**

The spaCy pre-trained model (en\_core\_web\_sm) was chosen for its:

* **Ease of Use**: Ready-to-deploy, state-of-the-art performance for common NER tasks.
* **Domain Independence**: Works across multiple domains without requiring fine-tuning.

#### **Summarization**

A **truncation-based summarization** approach was adopted for simplicity, considering the absence of training resources for abstractive models. This method provides quick overviews by limiting the character count.

### **3. Performance Results**

#### **Classification Metrics**

Using a holdout validation set, the following metrics were obtained:

* **Precision**: 78%
* **Recall**: 74%
* **F1-Score**: 76%

#### **Entity Extraction**

The spaCy model achieved:

* **Precision**: 85%
* **Recall**: 83%
* **F1-Score**: 84%

#### **Summarization**

The summarization method is static (rule-based) and lacks traditional performance metrics like ROUGE or BLEU.

### **4. Error Analysis**

#### **Classification**

**Mistakes**:

* Overlapping labels caused confusion. For example, "Artificial Intelligence in healthcare" was classified as both "Technology" and "Health."

**Confusion Matrix Observations**:

* Misclassification between "Finance" and "Technology" due to shared terminologies like "investment" and "blockchain."

**Areas for Improvement**:

* Use transformer-based models like BERT for capturing contextual meaning.
* Increase labeled training data for rare labels.

#### **NER**

**Mistakes**:

* Entity boundaries were occasionally misidentified. E.g., "AI startups in California" was tagged as two separate entities instead of one cohesive location-based entity.

**Improvement Opportunities**:

* Fine-tune spaCy's model on domain-specific data.

#### **Summarization**

* Truncation led to incomplete summaries for lengthy text.
* Future solutions could include extractive methods like **TextRank** or abstractive approaches with HuggingFace models.

### **5. Future Work**

### **1. Dynamic Data Augmentation**

* Implement Active Learning to continuously update the model with new data for better adaptability.

### **2. Self-Supervised Learning**

* Leverage BERT/GPT for generating synthetic labels and extracting features from unlabeled data.

### **3. Hybrid Model Integration**

* Combine specialized models for classification, NER, and summarization to improve performance.

### **4. Transfer Learning for Domain Adaptation**

* Fine-tune pre-trained models on domain-specific data to improve accuracy for specialized tasks.