Report on Mountain Name Recognition Project with Potential Improvements

Overview:

The project uses a BERT-based model for named entity recognition (NER) to identify mountain names in text. The model classifies tokens using the BIO tagging scheme to recognize mountain names in the sentences.

Current Workflow:

1. **Dataset Creation & Augmentation:** The dataset is augmented with random mountain names inserted into predefined sentences, and labels are aligned using tokenization.
2. **Model Training:** The model is trained using a BERT-based architecture, applying multi-label classification and weighted loss functions to manage class imbalances.
3. **Inference:** The trained model predicts mountain names from input text.

Potential Improvements:

1. **Data Augmentation:**
    - Generate more domain-specific sentences (e.g., hiking, geography) and diversify sentence structures for better generalization.
2. **Handling Complex Mountain Names:**
    - Improve tokenization for compound mountain names (e.g., "Mount McKinley") and use post-processing to merge related tokens.
3. **Model Evaluation:**
    - Implement cross-validation and use performance metrics like F1 score, precision, and recall.
4. **Special Cases:**
    - Enhance handling of common terms like "Mount" or "Mountain" and add a known mountain name dictionary to handle edge cases.
5. **Handling Geographical Names (Countries vs. Mountains):**
    - The model often misclassifies country names (e.g., "South Africa," "South America," "Canada") as mountain names, which suggests a lack of contextual differentiation. This can be addressed by:
        - Implementing a post-processing step to filter out non-mountain geographical names.
        - Introducing a geographical entity recognition step before mountain name extraction to distinguish between countries, regions, and mountains more effectively.

Conclusion:

The model is capable of recognizing mountain names but has difficulty distinguishing between mountains and geographical locations like countries. By refining data augmentation and adding context-based post-processing, the model can be enhanced to provide more accurate predictions.