****

**Contextual Entity Extraction and Classification in Clinical Notes Using Transformer Model**

**By**

**Tanvi Mahajan - tm101****4**

**A research project submitted in partial fulfillment of the requirements for the Degree of Master of Science in Data Science Specialization in Statistics at Rutgers, the State University of New Jersey**

**May 2025**

## **Abstract:**

This project focuses on the critical task of contextual entity extraction and classification in clinical notes using transformer models, with a primary emphasis on the Bio\_ClinicalBERT model. Clinical notes contain a wealth of medical information; however, the unstructured nature of this data presents significant challenges for traditional analysis and decision support systems. Named Entity Recognition (NER) helps bridge this gap by identifying and categorizing key medical entities such as diseases, medications, and procedures within the text.

In this study, traditional machine learning models including Logistic Regression, Naive Bayes, and Conditional Random Fields (CRF) were evaluated alongside the advanced transformer-based Bio\_ClinicalBERT model. The MACCROBAT 2018-2020 dataset was employed for training and testing purposes, offering a rich source of annotated clinical entities across 42 distinct categories. Data preprocessing steps included normalization of clinical abbreviations, label alignment, and careful handling of missing annotations to ensure robust model training.

Performance evaluation based on precision, recall, F1-score, and accuracy demonstrated that the Bio\_ClinicalBERT model significantly outperformed the traditional methods. The model achieved an impressive F1-score of 0.83 and an accuracy of 0.91, showcasing its ability to capture the complexities and contextual dependencies inherent in clinical text. Fine-tuning strategies, such as optimizing learning rates, batch sizes, and incorporating dropout for regularization, further contributed to model improvements.

Challenges such as class imbalance, memory limitations during model training, and the need for efficient hyperparameter tuning were encountered and addressed. A detailed analysis of entity-specific performance revealed that the model performed exceptionally well on frequently occurring entity types but faced difficulties with underrepresented classes.

The findings underscore the transformative potential of leveraging deep learning, particularly domain-specific pre-trained models, in healthcare applications. Enhanced entity extraction can drive better information retrieval, aid clinical decision-making, and ultimately contribute to more efficient and effective healthcare delivery. This project highlights not only the current capabilities but also future directions for improving clinical NLP through continued advancements in transformer-based models and dataset curation.

5.4 Table of Contents:

[Abstract: 2](#_Toc197107973)

[Chapter 1. Introduction: 4](#_Toc197107974)

[Chapter 2. Objective: 5](#_Toc197107975)

[Chapter 3. Background and Research on Models 6](#_Toc197107976)

[**3.1 Logistic Regression** 6](#_Toc197107977)

[**3.2 Naive Bayes** 6](#_Toc197107978)

[**3.3 Conditional Random Fields (CRF)** 7](#_Toc197107979)

[**3.4 Bio\_ClinicalBERT Model** 7](#_Toc197107980)

[Chapter 4. Data Collection 7](#_Toc197107981)

[**4.1 Data Source** 7](#_Toc197107982)

[**4.2 Dataset Characteristics** 8](#_Toc197107983)

[**4.3 Data Preprocessing** 9](#_Toc197107984)

[Chapter 5. Experiments 10](#_Toc197107985)

[**5.1 Model Training** 10](#_Toc197107986)

[**5.2 Training the Data and Performance Evaluation** 11](#_Toc197107987)

[**5.3 Results** 12](#_Toc197107988)

[Chapter 6. Conclusion 14](#_Toc197107989)

**5.5 List of Tables:**

[**Table 1: Performance Metrics Comparison** 12](#_Toc197107693)

[**Table 2: Performance metrics of top-performing classes** 13](#_Toc197107694)

[**Table 3: Performance metrics of low-performing classes** 14](#_Toc197107695)

**5.6 List of Illustrations:**

[**Figure 1: Structure of the MACCROBAT 2018\_2020 Dataset.** 8](#_Toc197107561)

[**Figure 2: Distinct entities in the dataset** 8](#_Toc197107562)

[**Figure 3: Structure of the Dataset after adding the numeric tags column** 9](#_Toc197107563)

[**Figure 4: Code snippet for data processing to tokenize and align labels** 10](#_Toc197107564)

[**Figure 5: Training and Validation Loss over Epochs** 13](#_Toc197107565)

[**Figure 6: Precision, Recall, and F1 Score over Epochs** 13](#_Toc197107566)

## **Chapter 1. Introduction:**

In recent years, the healthcare industry has witnessed a significant shift towards the digitalization of clinical data, with electronic health records (EHRs) becoming an integral part of patient care. These records often contain vast amounts of unstructured textual data, including clinical notes written by healthcare professionals. While rich in information, this unstructured data poses significant challenges for automated systems to interpret, extract, and analyze effectively. As a result, there is a growing need for advanced natural language processing (NLP) techniques to process clinical texts and extract valuable insights for medical decision-making, research, and patient care.

Named Entity Recognition (NER) is one of the foundational tasks in clinical NLP, aimed at identifying and classifying specific entities in clinical text, such as diseases, medications, symptoms, procedures, and temporal expressions. Accurate entity extraction is critical in clinical settings because it enables various downstream tasks, such as information retrieval, clinical decision support, and automated medical record summarization. Traditional machine learning models, such as Logistic Regression and Naive Bayes, have been used for this purpose, but their performance in handling complex, domain-specific clinical data often falls short due to limited contextual understanding.

In this project, the application of state-of-the-art transformer-based models, specifically **Bio\_ClinicalBERT**, in performing NER tasks in clinical notes is explored. Transformer models, particularly those pre-trained on large biomedical and clinical corpora, have demonstrated exceptional capabilities in understanding context and semantic relationships in text, making them particularly suitable for clinical NLP tasks. Fine-tuning these models for the specific task of clinical NER holds the potential to significantly improve the accuracy and effectiveness of entity extraction, especially in complex and diverse clinical contexts.

**NLP Problem:** The problem addressed in this project is **Named Entity Recognition (NER)** in clinical notes. NER is a key task in Natural Language Processing (NLP), where the goal is to identify and classify predefined entities from unstructured text. In clinical notes, these entities can include diseases, medications, procedures, symptoms, and other important aspects of patient information. The challenge in clinical NER lies in the domain-specific nature of medical terminology, complex sentence structures, and the relationships between entities (such as time expressions and causal links).

## **Chapter 2. Objective:**

The primary goal of this project is to develop and evaluate a system for contextual entity extraction and classification in clinical notes using advanced transformer models. Specifically, the project aims to:

1. Apply the MACCROBAT 2018-2020 dataset to train and test NER models, capitalizing on its diverse and structured clinical annotations across 42 entity types, including diseases, procedures, medications, symptoms, time expressions, and patient demographics.
2. Compare different machine learning approaches to NER in clinical text, including traditional models such as Logistic Regression, Naive Bayes, and Conditional Random Fields (CRF), as well as transformer-based models like Bio\_ClinicalBERT.
3. Assess model performance using standard evaluation metrics such as precision, recall, F1-score, and accuracy to understand the strengths and weaknesses of each approach and identify the most effective models for clinical NER tasks.
4. Investigate the potential for fine-tuning transformer-based models and exploring hyperparameter optimization, model selection, and error analysis to improve NER performance.

**Approach to Solve the Problem:**

To solve this problem, deep learning techniques were leveraged, focusing on transformer-based models like **Bio\_ClinicalBERT**, a domain-specific variant of BERT pre-trained on clinical texts. The core of our approach is to fine-tune this transformer model for the clinical NER task to achieve more accurate entity extraction in medical notes. To benchmark the performance, traditional machine learning models like **Logistic Regression**, **Naive Bayes**, and **Conditional Random Fields (CRF) were also explored**.

**Multinomial Classification:** This is a **multinomial classification** problem, as the aim is to classify text into multiple entity categories, such as diseases, medications, symptoms, etc.

**Baseline:** The baseline models used for comparison were **Logistic Regression** and **Naive Bayes**. These were implemented with simple token-level features to evaluate their performance against more advanced techniques.

**Machine Learning Architectures Used:**

* **Logistic Regression** (for token classification)
* **Naive Bayes** (for token classification)
* **Conditional Random Fields (CRF)** (for sequence labeling)
* **Bio\_ClinicalBERT** (fine-tuned for clinical NER)

## **Chapter 3. Background and Research on Models**

### **3.1 Logistic Regression**

**Logistic Regression** is a simple, yet powerful linear model used for binary classification tasks. In the context of NER, each token in the text can be treated as an independent feature and predict the corresponding entity label. Although Logistic Regression is fast and easy to implement, it does not consider the complex relationships between words in the text, which limits its performance on tasks with more intricate dependencies, such as NER in clinical notes.

### **3.2 Naive Bayes**

**Naive Bayes** is another traditional classifier based on Bayes' theorem, if the features (tokens) are conditionally independent given the class. Like Logistic Regression, Naive Bayes also lacks the ability to capture the sequential or contextual dependencies in text, making it less suitable for tasks like NER that require an understanding of token relationships.

### **3.3 Conditional Random Fields (CRF)**

**Conditional Random Fields (CRF)** are a type of probabilistic graphical model that can be used for sequence labeling tasks, making them particularly suitable for NER. Unlike Logistic Regression and Naive Bayes, CRFs consider the context of neighboring tokens to make predictions. This allows CRFs to model the dependencies between labels in a sequence, which is essential for tasks like NER, where consecutive tokens often share relationships (e.g., a disease followed by its medication).

### **3.4 Bio\_ClinicalBERT Model**

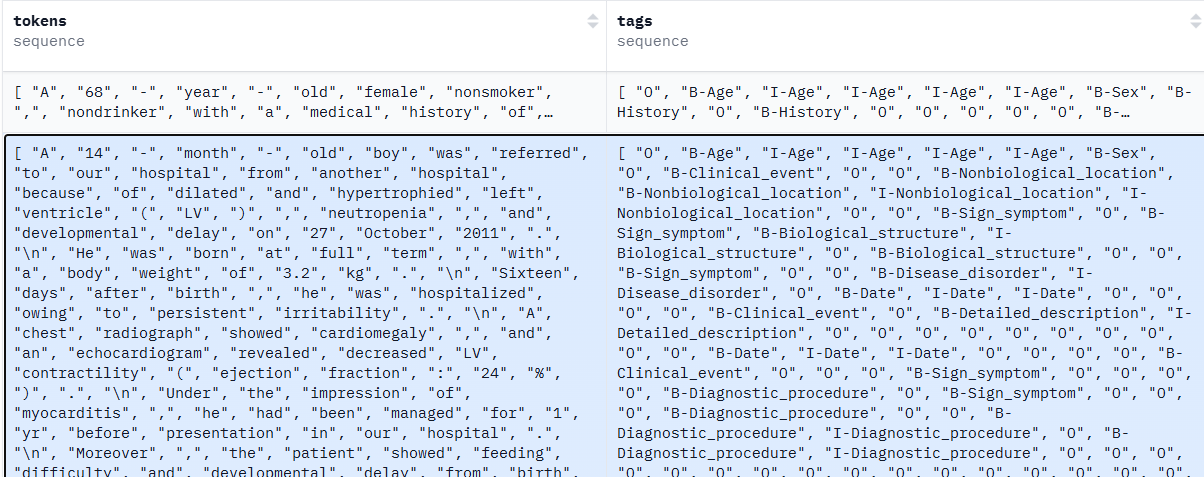
**Bio\_ClinicalBERT** is a transformer-based model pre-trained specifically on clinical texts. Based on the original BERT architecture, it has been fine-tuned on large-scale biomedical and clinical corpora to capture domain-specific terminology and context. BERT and its variants have revolutionized NLP by effectively capturing both short- and long-range dependencies in text using self-attention mechanisms. This makes transformer models like Bio\_ClinicalBERT highly effective for tasks like NER, where understanding context and relationships between entities is critical.

In this project, **Bio\_ClinicalBERT** is fine-tuned on the clinical NER task, taking advantage of its pre-training on medical texts and its ability to understand the contextual dependencies between entities within clinical notes.

## **Chapter 4. Data Collection**

### **4.1 Data Source**

The data used in this project comes from the **MACCROBAT 2018\_2020 dataset**, available on **Hugging Face**. The dataset includes a rich set of annotations for Named Entity Recognition (NER) in clinical notes and spans 42 distinct entity categories such as diseases, medications, procedures, symptoms, and time expressions. The clinical notes data span the years 2018 to 2020, providing a recent and relevant dataset for modern biomedical NLP tasks. The text for tokens in data are extracted from PubMed Central full-text documents but have been edited to include only clinical case report details. The data exhibits an imbalanced distribution across entity classes, with common entities such as DISEASE\_DISORDER and SIGN\_SYMPTOM appearing more frequently than less common categories like DEVICE or ANATOMICAL\_SITE. This imbalance was accounted for during model evaluation by using weighted metrics.



**Figure 1: Structure of the MACCROBAT 2018\_2020 Dataset.**

A table of medical information

AI-generated content may be incorrect.

**Figure 2: Distinct entities in the dataset**

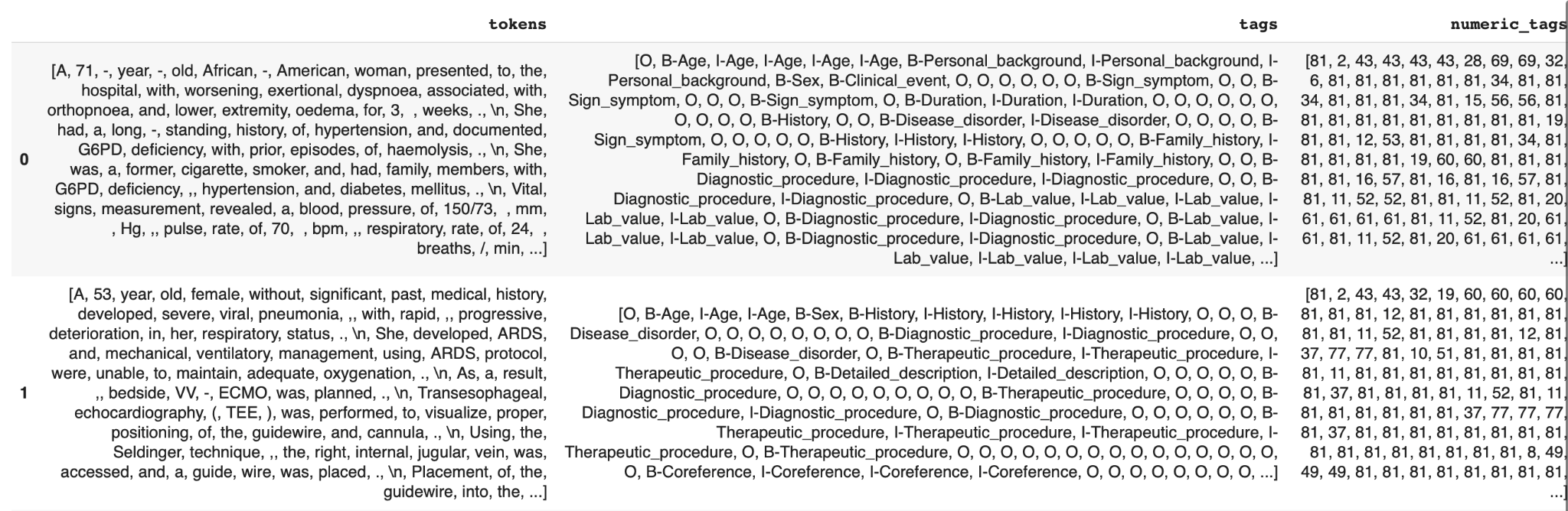
### **4.2 Dataset Characteristics**

* **Size**: The dataset consists of 7,350 **sentences** with over **118,000 annotations**.
* **Entity Types**: It includes 42 different entity types, covering a broad spectrum of clinical information.
* **Split**: The dataset was split into 70 % - Train Set, 15% - Validation Set and 15% - Test Set

### **4.3 Data Preprocessing**

The dataset was already tokenized, reducing the need for complex text preprocessing. However, it was still needed to:

* **Handle Missing Data**: Any missing annotations were excluded or filled based on context.
* **Normalize Abbreviations**: Some clinical abbreviations were normalized (e.g., “HTN” to “Hypertension”).
* **Numeric Tags:** Numeric tags were added alongside the existing BIO label tagging to simplify processing.
* **Entity Alignment**: Numeric labels were aligned with tokenized word IDs to ensure proper mapping during training.



**Figure 3: Structure of the Dataset after adding the numeric tags column**



**Figure 4: Code snippet for data processing to tokenize and align labels**

## **Chapter 5. Experiments**

### **5.1 Model Training**

This project employs both baseline machine learning models and a fine-tuned pre-trained transformer-based model to solve the entity extraction and classification problem.

Four different models were trained and evaluated for clinical entity extraction:

1. **Logistic Regression**: A simple baseline model that uses token-level features derived from the text.
2. **Naive Bayes**: Another baseline model that utilizes a probabilistic approach to classify entities based on token likelihood.
3. **Conditional Random Fields (CRF)**: A sequence-based model that captures dependencies between adjacent tokens.
4. **Bio\_ClinicalBERT**: The advanced approach utilizes a pre-trained transformer-based BERT model fine-tuned for the NER task, with inputs tokenized using the Hugging Face tokenizer. The model was sourced from Hugging Face at [Bio\_ClinicalBERT](https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT). The Bio\_Clinical BERT model was initially pre-trained on clinical notes from the MIMIC III NOTE EVENTS table, which is a freely available database comprising de-identified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Medical Center between 2001 and 2012 totaling approximately 880 million words. Notes were sectioned using a rule-based splitter and tokenized with SciSpacy. The pre-training employed a duplication factor of 5, masking 15% of input tokens with a maximum of 20 masked tokens per sequence.

The architecture of Bio\_ClinicalBERT includes the following components:

* **Embedding Layer**: Incorporates word embeddings (vocabulary size: 28,996; embedding dimension: 768), positional embeddings (maximum sequence length: 512), and token type embeddings.
* **Transformer Encoder**: Comprises 12 layers, each featuring multi-head self-attention, GELU activation, and feed-forward networks with 3,072 hidden units.
* **Dropout Layers**: Dropout (p=0.1) is applied after embedding and intermediate layers to reduce overfitting.
* **Classifier**: A dense layer with SoftMax activation for multi-class token classification. It maps the contextualized token embeddings (768 dimensions) to 82 output classes corresponding to BIO entity labels.

**Inputs:** Tokenized clinical notes, where each note was split into sequences of tokens using the **Auto Tokenizer** associated with Bio\_ClinicalBERT. Special tokens such as [CLS] and [SEP] were added to the sequences to indicate the start and end of each input text.

* Input embeddings: Pre-trained word embeddings from Bio\_ClinicalBERT (768 dimensions per token).

**Outputs:** A sequence of predictions corresponding to each input token, assigning one of the 82 BIO entity labels (e.g., B-Disease, O, I-Medication).

**Fine-Tuning and Regularization:** The key hyperparameters, including the learning rate, batch size, and training epochs, to optimize performance were adjusted. For hyperparameter tuning, a learning rate of 2×10-5 was used, a value small enough to prevent drastic updates that could disrupt the pre-trained weights while still allowing the model to learn gradually. To further refine optimization, a **cosine decay strategy** was employed, which started with a higher learning rate and reduced it smoothly over time, improving convergence as training progressed.

To address memory limitations, a **batch size of 8** was used**.** For training, I experimented with different numbers of epochs (5, 10, 15, and 20), ultimately achieving the best results with **20 epochs**. Regularization was applied using a **weight decay of 0.01** to discourage overly large weights and a **dropout rate of 0.1** to promote robust learning by introducing randomness in the training process.

Finally, the model’s performance was evaluated using the **F1 score**, which balances precision and recall—critical for tasks like clinical entity extraction and classification. These hyperparameters collectively optimized the model for our task, balancing learning efficiency, regularization, and overall performance.

### **5.2 Training the Data and Performance Evaluation**

It was observed that training and validation loss steadily decreased across epochs, converging after ~15 epochs which indicates that the model generalizes well without significant overfitting. **Precision, recall, and F1 scores** were monitored for validation performance, with the best F1 score achieved at the 20th epoch. The **precision** and **recall** for individual entity types were also calculated to gain deeper insights into the model's performance. This analysis helped identify the **top-performing entities**, where the model demonstrated high accuracy and coverage, as well as the **least-performing entities**, which faced challenges due to factors such as data imbalance or ambiguity in the clinical text. This entity-specific evaluation provided valuable information for understanding the model's strengths and areas for improvement.

### 

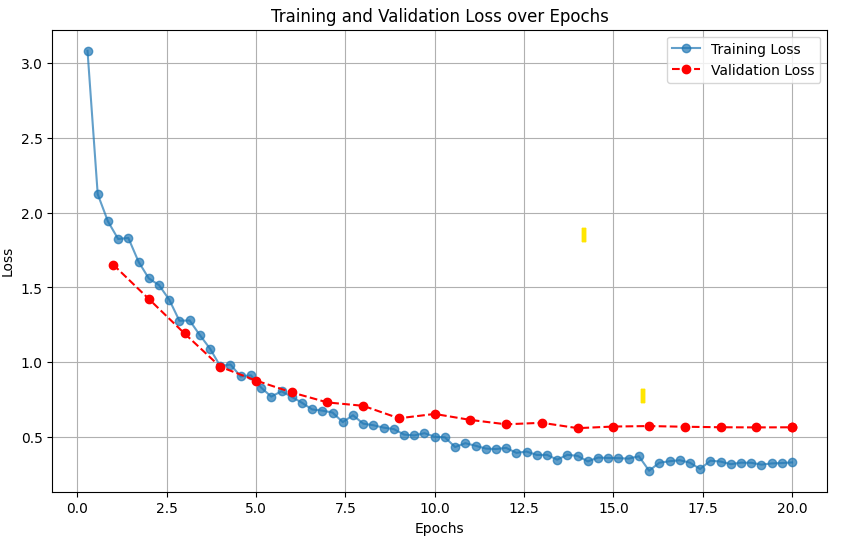
### **5.3 Results**

The results for the performance of models were as follows:

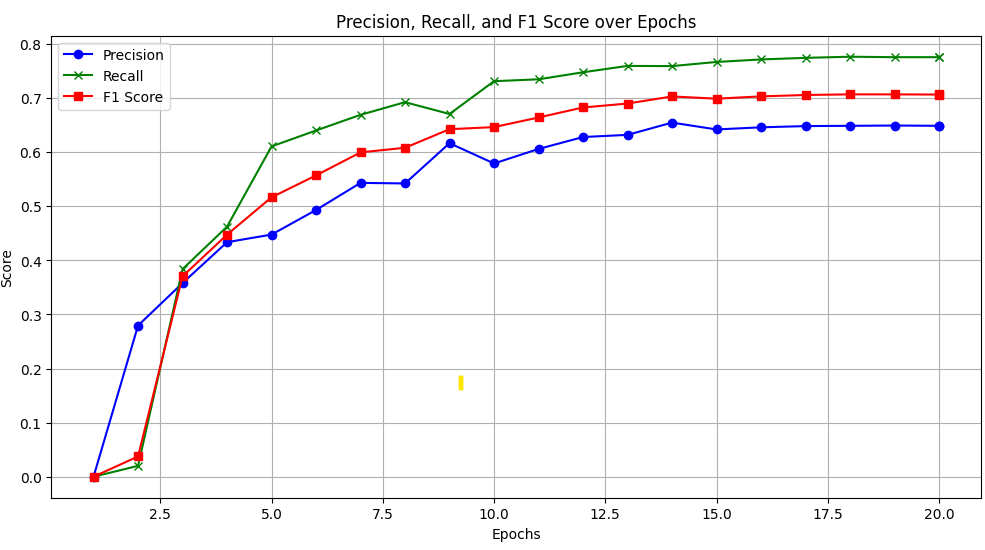
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Naive Bayes** | **CRF Model** | **Bio\_Clinical\_BERT** |
| **Precision** | 0.75 | 0.75 | 0.79 |
| **Recall** | 0.66 | 0.78 | 0.87 |
| **F1-score** | 0.67 | 0.76 | 0.83 |
| **Accuracy** | 0.75 | 0.78 | 0.91 |

**Table 1: Performance Metrics Comparison**

The **Bio\_ClinicalBERT** model significantly outperformed the traditional models in all evaluation metrics, particularly in recall (0.87) and F1-score (0.83). This indicates that transformer-based models, particularly domain-specific ones like Bio\_ClinicalBERT, are well-suited for clinical NER tasks.



**Figure 5: Training and Validation Loss over Epochs**



**Figure 6: Precision, Recall, and F1 Score over Epochs**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Age** | **Sex** | **Medication** | **Diagnostic\_Procedure** | **Therapeutic-Procedure** | **Lab\_Value** |
| **Precision** | 0.94 | 0.94 | 0.91 | 0.72 | 0.80 | 0.75 |
| **Recall** | 0.94 | 0.98 | 0.97 | 0.90 | 0.91 | 0.86 |
| **f1 Score** | 0.94 | 0.96 | 0.94 | 0.80 | 0.85 | 0.80 |

**Table 2: Performance metrics of top-performing classes**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Volume** | **Color** | **Subject** | **Frequency** | **Weight** |
| **Precision** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **Recall** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **f1 Score** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

**Table 3: Performance metrics of low-performing classes**

**5.4 Challenges**

* **Data Imbalance:** Some entity types (e.g., medical devices, adverse events) were underrepresented in the dataset, making it difficult for the model to generalize effectively for these categories.
* **Memory Constraints:** The batch size was reduced to 8 and leveraged gradient accumulation to simulate larger batches.
* **Convergence Issues:** Early stopping criteria and validation loss monitoring were implemented to prevent overfitting.

## **Chapter 6. Conclusion**

This project demonstrated that transformer-based models, particularly Bio\_ClinicalBERT, provide a significant performance improvement over traditional machine learning techniques such as Logistic Regression, Naive Bayes, and Conditional Random Fields (CRF) in the task of clinical Named Entity Recognition (NER). By leveraging a domain-specific pre-trained model and fine-tuning it on the MACCROBAT 2018-2020 dataset, impressive performance metrics were also achieved, with a particular emphasis on improved recall and F1-scores. These results confirm the importance of context-aware models in extracting valuable information from complex, unstructured clinical notes.

The findings also highlight critical challenges, including handling class imbalance, managing computational resource limitations, and optimizing hyperparameters. Despite these challenges, the project showcases the transformative potential of deep learning models in revolutionizing healthcare data analysis. Future work could involve the integration of larger and more diverse datasets, exploration of more sophisticated fine-tuning techniques, and development of hybrid models that combine rule-based and machine learning approaches for even greater accuracy.

In conclusion, the success of Bio\_ClinicalBERT in this project reaffirms the importance of continuous advancements in domain-specific language models to drive innovations in clinical NLP. With further research and refinement, such models can play a pivotal role in enhancing patient care, supporting medical research, and improving healthcare outcomes on a broader scale.

**5.8 References:**

**Research Paper**

**Title:** Publicly Available Clinical BERT Embeddings

**Authors:** Emily Alsentzer, John Murphy, William Boag, et al.

**Year:** 2019

**DOI:** [10.48550/arXiv.1904.05342](https://arxiv.org/pdf/1904.03323)