**COMPLEX**

­­

**COMPUTING**

**PROBLEM**

**CT- 485**

**DATASET:**

Mrs. May, a dedicated patient seeking comprehensive healthcare, visited Leeds General Infirmary due to persistent respiratory issues. Under the expert care of Dr. Ray Johnson, a pulmonologist, her medical journey unfolded with a thorough examination of her history, including a prior consultation with Dr. Parker in May 2019. Recognizing the severity of her condition, Dr. Johnson recommended a range of diagnostic procedures, from spirometry and chest X-rays to electrocardiograms and bone density scans. The prescribed treatment plan encompassed a multidimensional approach, incorporating inhalers, bronchodilators, and medications to address COPD, occasional joint pain, and anemia. Dr. Johnson's holistic care extended beyond traditional medicine, encompassing lifestyle modifications, nutritional advice, and mental health support. Mrs. May, actively engaged in her well-being, participated in workshops, support groups, and complementary therapies, such as yoga and meditation.

As part of her ongoing care, Dr. Johnson emphasized the importance of regular health screenings, including lipid panels and diabetes screenings, to manage potential comorbidities. Mrs. May, appreciating the thorough approach, followed personalized dietary plans to support her overall health and alleviate occasional digestive discomfort. The holistic care plan also included referrals to specialists such as a gastroenterologist for an esophagogastroduodeoscopy (EGD) procedure to address swallowing difficulties and a neurologist to investigate occasional headaches.

In addition to managing her respiratory health, Dr. Johnson addressed Mrs. May's concerns about potential osteoporosis due to long-term corticosteroid use, prescribing calcium and vitamin D supplementation alongside bone density scans. Recognizing the potential impact of stress on her immune system, Dr. Johnson recommended mindfulness-based interventions and referred Mrs. May to a psychologist for cognitive-behavioral therapy to address anxiety related to her chronic health conditions.

Acknowledging Mrs. May's commitment to quitting smoking, Dr. Johnson enrolled her in a specialized smoking cessation program, combining counseling, behavioral therapy, and pharmacotherapy. The program aimed to enhance her chances of success in maintaining a smoke-free lifestyle. Mrs. May, grateful for the support, explored alternative therapies such as aromatherapy and essential oils to promote relaxation.

As part of her comprehensive care, Dr. Johnson discussed the potential benefits of influenza and pneumonia vaccinations, emphasizing the importance of preventive measures in managing respiratory health. Mrs. May, recognizing the value of vaccinations, also engaged in discussions about potential genetic factors in her respiratory condition through genetic counseling, exploring any familial predispositions and assessing her risk factors.

Given Mrs. May's history of allergies, Dr. Johnson recommended regular cleaning of her living space and hypoallergenic bedding. The proactive measures aimed to minimize potential allergic triggers and prevent exacerbations of her respiratory condition. Dr. Johnson also discussed the importance of maintaining a smoke-free environment at home to further support Mrs. May's respiratory health, considering the impact of environmental factors.

Addressing Mrs. May's concerns about potential medication interactions with her daily coffee consumption, Dr. Johnson provided guidance on the timing of medications to minimize any adverse effects. The discussion extended to dietary habits, with Dr. Johnson recommending moderation in caffeine intake due to the sensitivity of certain medications to caffeine.

As part of the ongoing monitoring process, Dr. Johnson equipped Mrs. May with a peak flow meter to track her respiratory function at home. The device allowed her to report any significant changes, fostering a collaborative approach to managing her condition. Dr. Johnson also emphasized the importance of open communication, encouraging Mrs. May to inform all healthcare providers about her complete medication list to ensure coordinated and safe care.

Recognizing the potential impact of weather changes on Mrs. May's respiratory symptoms, Dr. Johnson discussed preventive measures and adjustments to her treatment plan during seasonal transitions. The proactive approach aimed to minimize the risk of exacerbations and enhance Mrs. May's respiratory well-being.

Mrs. May, expressing interest in complementary therapies, attended hospital-sponsored workshops on respiratory yoga to learn breathing exercises and relaxation techniques. Dr. Johnson supported her exploration of alternative therapies, acknowledging the potential benefits of integrative approaches in managing chronic conditions.

In response to Mrs. May's inquiries about pulmonary rehabilitation exercises, Dr. Johnson provided a customized exercise plan. The plan incorporated aerobic exercises and strength training to improve Mrs. May's overall fitness and enhance her exercise capacity, contributing to better respiratory health.

Dr. Johnson, recognizing the potential impact of hormonal changes on respiratory symptoms, discussed the role of hormonal fluctuations in women. He recommended monitoring any patterns and symptoms related to hormonal changes, contributing to a more nuanced understanding of Mrs. May's condition.

Mrs. May, curious about the potential benefits of meditation, participated in a hospital-sponsored mindfulness meditation session. The session aimed to explore the impact of mindfulness on stress reduction and overall well-being, complementing her medical treatment.

Dr. Johnson, considering the potential impact of respiratory conditions on exercise tolerance, prescribed a short-acting bronchodilator for use before physical activities. The prescription aimed to enhance Mrs. May's exercise capacity and support her in maintaining an active lifestyle.

Acknowledging the potential impact of air pollution on respiratory health, Dr. Johnson recommended wearing a mask during outdoor activities and discussed the benefits of air purifiers for indoor spaces with Mrs. May. The measures aimed to minimize exposure to environmental triggers and enhance respiratory well-being.

Mrs. May, expressing concern about potential weight loss, was referred to a nutritionist for a personalized dietary plan. The plan focused on maintaining optimal nutritional intake, addressing potential weight loss concerns, and supporting her overall well-being.

Dr. Johnson, addressing the impact of chronic stress on the immune system, recommended mindfulness-based interventions to enhance Mrs. May's resilience and immune function. The approach aimed to support her psychological well-being and contribute to a holistic understanding of health.

Mrs. May, inquiring about potential genetic factors in her respiratory condition, underwent genetic counseling. The counseling process explored any familial predispositions and assessed her risk factors, contributing to a more comprehensive understanding of her health.

As part of her ongoing care, Mrs. May received a pedometer to track her daily steps. The initiative aimed to encourage an active lifestyle, aligning with Dr. Johnson's emphasis on the importance of regular physical activity in respiratory health.

Dr. Johnson, recognizing the importance of vaccination in preventing respiratory infections, recommended an annual flu vaccine and a pneumonia vaccine booster for Mrs. May. The preventive measures aimed to reduce the risk of respiratory infections and support her overall health.

Mrs. May, experiencing occasional difficulty swallowing, underwent a thorough evaluation by a gastroenterologist. The evaluation included an esophagogastroduodenoscopy (EGD) procedure, revealing mild esophageal irritation. Dietary modifications, such as avoiding spicy foods, were recommended to alleviate her swallowing difficulties.

**Baseline NER Application Design by ChatGpt**



**1. Define Named Entity Types:**

Specify the medical entities relevant to your application, such as diseases, medications, anatomical entities, etc.

**2. Collect and Prepare Data:**

Gather a labeled dataset containing medical texts annotated with named entities.

**3. Preprocessing:**

* Tokenize the medical text.
* Perform part-of-speech tagging and lemmatization.
* Prepare the data in a format suitable for training, including the input text and corresponding medical entity labels.

**4. Split Data:**

Split the dataset into training and testing sets to evaluate the model's performance.

**5. Feature Extraction:**

Represent the words or tokens as features that the model can use for training. This could be embeddings, word vectors, or any other relevant representation.

**6. Choose a Model:**

Select a baseline NER model, such as a Conditional Random Field (CRF) or a machine learning classifier.

**7. Train the Model:**

Train the chosen model on the medical training dataset.

**8. Evaluate the Model:**

Assess the model's performance on the testing dataset using metrics relevant to the medical domain, considering factors like precision, recall, and F1 score.

**9. Tune Hyperparameters (Optional):**

 Adjust the hyperparameters for optimal performance.

**10. Apply the Model:**

Use the trained model to recognize named entities in new, unseen medical text data.

**11. Post-processing (Optional):**

Perform post-processing to refine results, such as handling abbreviations or merging adjacent entities.

**12. Deploy the Model:**

Deploy the model in a production environment for real-time or batch processing of medical texts.

**13. Monitor and Update:**

Regularly monitor the model's performance in a medical context and update it as needed with new labelled medical data.

**Improved Baseline NER Application Design by:**



**Nashra Ghaffar (CT-32)**



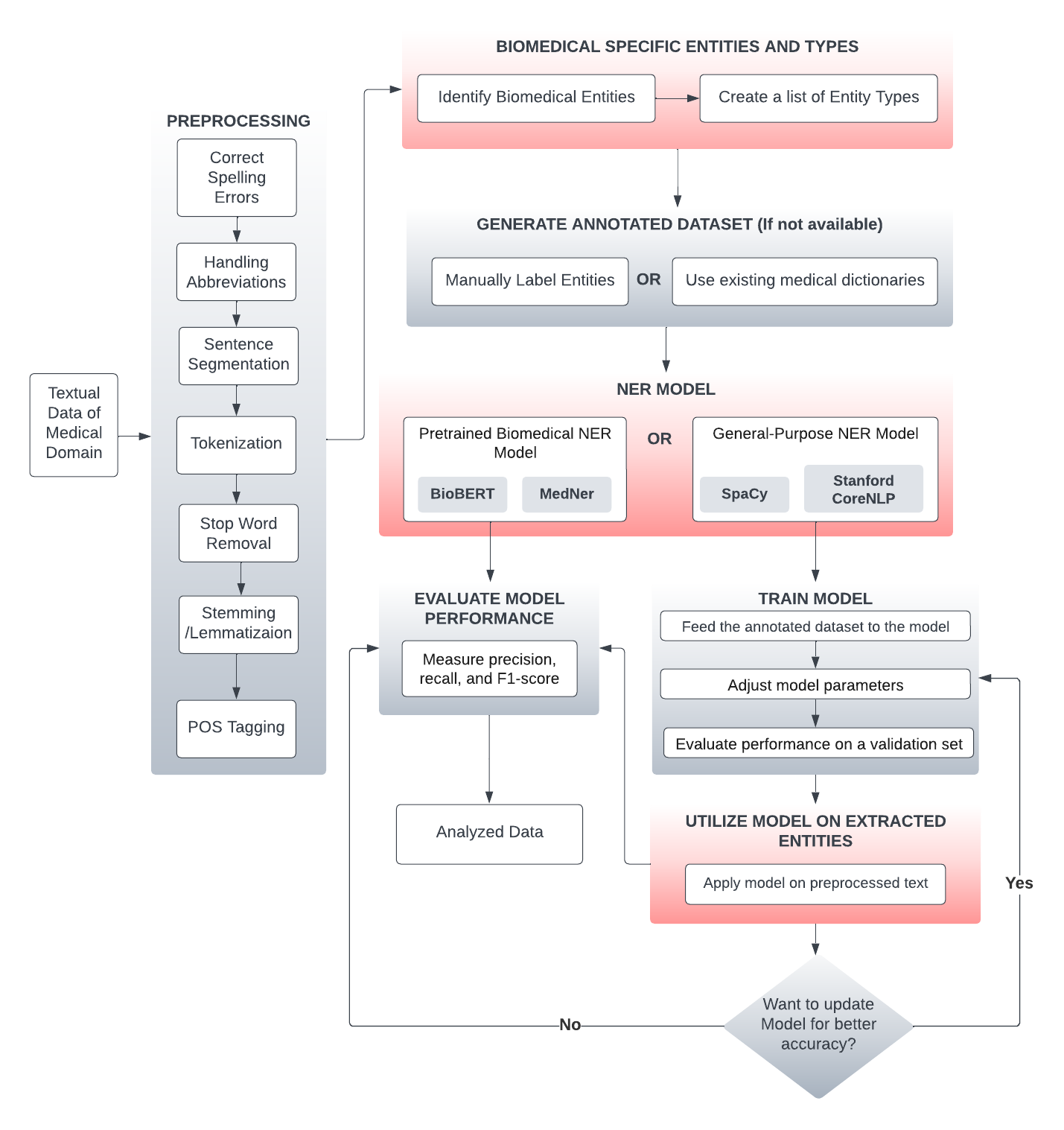
**Mehak Hussain (CT-30)**



**Mahshid Fatima (CT-23)**



**M. Abdul Rafay (CT-12)**

****

**1. Preprocessing of Text:**

**i) Correct spelling errors and typos:**

Correct spelling ensures your message is accurately conveyed and avoids confusion. Typos can change the meaning of words or sentences, leading to misunderstandings and misinterpretations.

**For example:**

* Mrs. May underwent an esophagogastrodudeoscopy (EGD) procedure to address swallowing difficulties.

Here, “esophagogastrodudeoscopy” should be “esophagogastroduodenoscopy”.

**ii) Handling Abbreviations:**

Expand abbreviations to their full forms if needed. This can help the model understand and recognize entities more accurately. Here, the specific medical abbreviation lookups will be used to avoid the ambiguity of the similar abbreviations.

**For example:**

In our dataset, abbreviations like "X-ray," "COPD," and others have been expanded to their full forms for clarity and completeness.

Here, we can use Medical Ontologies or Dictionaries for finding the correct abbreviation because we know that one abbreviation can have many full forms.

**iii) Sentence Segmentation:**

Split the text into sentences. NER often involves analysing context within sentences. So we can easily tokenize the text in sentences.  
  
 **Example of first paragaph:**

1. Mrs. May, a dedicated patient seeking comprehensive healthcare, visited Leeds General Infirmary due to persistent respiratory issues.
2. Under the expert care of Dr. Ray Johnson, a pulmonologist, her medical journey unfolded with a thorough examination of her history, including a prior consultation with Dr. Parker in May 2019.
3. Recognizing the severity of her condition, Dr. Johnson recommended a range of diagnostic procedures, from spirometry and chest X-rays to electrocardiograms and bone density scans.
4. The prescribed treatment plan encompassed a multidimensional approach, incorporating inhalers, bronchodilators, and medications to address COPD, occasional joint pain, and anemia.
5. Dr. Johnson's holistic care extended beyond traditional medicine, encompassing lifestyle modifications, nutritional advice, and mental health support.
6. Mrs. May, actively engaged in her well-being, participated in workshops, support groups, and complementary therapies, such as yoga and meditation.

**iv) Tokenization:**

Text must first be split into smaller splices that the NER system can process. These splices can be as small as single words or as large as whole sentences.

**For example:**

“Mrs. May, a dedicated patient seeking comprehensive healthcare” may be split into the following tokens: “Mrs”., “May”, “a”, “dedicated”, “patient”, “seeking”, “comprehensive”, “healthcare"

**v) Stopword Removal:**

Remove common words (stopwords) that don't carry significant meaning for NER. This can include words like "the," "and," "is," etc.  
  
**For example:**

Mrs. May, dedicated patient seeking comprehensive healthcare, visited Leeds General Infirmary due to persistent respiratory issues. Under expert care Dr. Ray Johnson, pulmonologist, medical journey unfolded thorough examination history, including prior consultation Dr. Parker May 2019. Recognizing severity condition, Dr. Johnson recommended range diagnostic procedures, spirometry chest X-rays electrocardiograms bone density scans. Prescribed treatment plan encompassed multidimensional approach, incorporating inhalers, bronchodilators, medications address COPD, occasional joint pain, anemia. Dr. Johnson's holistic care extended beyond traditional medicine, encompassing lifestyle modifications, nutritional advice, mental health support. Mrs. May, actively engaged well-being, participated workshops, support groups, complementary therapies, yoga meditation.

Here you can see that all the common words i.e. of, her, the, a, in, etc have been removed.

**vi) Stemming or Lemmatization:**

Reduce words to their base or root form. This step helps in reducing dimensionality and improving the model's ability to generalise. Lemmatization is preferred over stemming as it considers the context and returns a valid word.

**For example:**

Mrs. May, dedicate patient seek comprehensive healthcare, visit Leeds General Infirmary due persistent respiratory issues. Under expert care Dr. Ray Johnson, pulmonologist, medical journey unfold thorough examination history, include prior consultation Dr. Parker May 2019. Recognize severity condition, Dr. Johnson recommend range diagnostic procedures, spirometry chest X-ray electrocardiogram bone density scan. Prescribe treatment plan encompass multidimensional approach, incorporate inhaler, bronchodilator, medication address COPD, occasional joint pain, anemia. Dr. Johnson's holistic care extend beyond traditional medicine, encompass lifestyle modification, nutritional advice, mental health support. Mrs. May, actively engage well-being, participate workshop, support group, complementary therapy, yoga meditation.

Here we can see that all the words are converted into their base form.

i.e.

Dedicated —> Dedicate

Seeking —> Seek

Visited —> visit

**vi) Part-Of-Speech Tagging:**

It will identify the grammatical role of each word, which can aid entity recognition.

**For example:**

Mrs./NNP

May/NNP,/,

a/DT

dedicated/JJ

patient/NN

seeking/VBG

comprehensive/JJ

healthcare/NN …

**2. Define Biomedical-Specific Entities and Types:**

Models that are used for NER are: Natural Language Toolkit (NLTK) and SpaCy

**i) Identify Biomedical Entities:**

Focus on entities relevant to biomedical and healthcare domains. Some examples include:

* **Medical conditions:** Respiratory issues, COPD, osteoporosis, anxiety.
* **Medications:** Inhalers, bronchodilators, calcium supplements, vitamin D.
* **Procedures:** EGD, spirometry, chest X-ray, genetic counseling.
* **Body parts:** Lungs, bones, immune system, nervous system.
* **Healthcare professionals:** Pulmonologist, gastroenterologist, neurologist, psychologist.

**ii) Create a list of Entity Types:**

Categorise your identified entities for clearer recognition and classification.

**Example categories:**

DOCTORS, MEDICATIONS, PROCEDURES, CONDITIONS, BODY\_PARTS, etc.

**3. Generate an Annotated Dataset:**

If the dataset is not available then we have to generate an annotated dataset for training the model.

* **Manually Label Entities:** Annotate each entity occurrence in the text with its corresponding type. This requires expert knowledge and can be time-consuming.

So here’s the another method which will save our time:

* **Consider Semi-supervised Techniques:** Utilize an existing medical ontologies or dictionaries to partially automate annotation will be more beneficial for annotation.

**4. Named Entity Recognition Model:**

**i) Choose methods of NER:**

* **Pre-trained Biomedical NER Model**: Explore models specifically designed for biomedical text, like ***BioBERT*** or ***MedNer***. These models are pre-trained on large biomedical datasets and can quickly adapt to our domain.

BERT (Bidirectional Encoder Representations from Transformers), can enhance the model's ability to understand the context and accurately recognize medical entities.

* **General-purpose NER Models:** Consider models like ***SpaCy*** or ***Stanford CoreNLP*** if domain-specific options are limited. These models might require fine-tuning with your annotated dataset for optimal performance.

The Model will train on the basis of following points:

1. Contextual analysis is crucial to disambiguate terms based on surrounding words and phrases. Machine learning models can be trained to understand the context of specific medical terms.

**Example:**  The term "cold" could refer to a respiratory infection or a low temperature, and disambiguating based on context is crucial.

1. We can use sequence labeling models like Conditional Random Fields (CRF) or Recurrent Neural Networks (RNNs), so it will handle entity overlapping by considering the entire sequence of words.

**Example**: Recognizing both "type 2 diabetes" and "insulin resistance" when they appear closely in a sentence.

1. Medical information is available in multiple languages, requiring the NER system to support multilingual entity recognition.So utilizing multilingual embeddings and training models on diverse multilingual datasets can enhance the system's ability to recognize medical entities across languages.

**Example**: Recognizing medical entities in a document that contains both English and French medical terms, like "asthma" and "asthme."

**REFERENCE**:

BioBERT :

<https://academic.oup.com/bioinformatics/article/36/4/1234/5566506>

BioNer:

<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04994-3>

**ii) General-Purpose NER Model: (Train Model)**

* **Feed the annotated dataset to the model:** Provide examples of text with labelled entities for the model to learn from.
* **Adjust Model Parameters:** Fine-tune hyperparameters like learning rate and optimizer to improve accuracy.
* **Evaluate Performance on a Validation set: (Test Model)** Assess the model's ability to identify and classify entities before applying it to unseen data.

**5. Apply the Train Model to Extracted Entities:**

* Run the model on your preprocessed and tokenized text.
* The model will predict the most likely entity type for each token.
* Combine the predictions to identify and extract complete entity mentions.

**Example:**

**Persons:** Mrs. May, Dr. Ray Johnson, Dr. Parker

**Locations:** Leeds General Infirmary

**Time:** May 2019

**Medical Terms:** Spirometry, Chest X-rays, Electrocardiograms, Bone density scans, COPD         (Chronic Obstructive Pulmonary Disease), Inhalers, Bronchodilators, Joint pain,  Anemia

**Treatment Components:** Inhalers, Bronchodilators, Medications

**Activities and Therapies:** Workshops, Support groups, Yoga, Meditation

**Concepts:** Comprehensive healthcare, Respiratory issues, Diagnostic procedures, Multidimensional approach, Lifestyle modifications, Nutritional advice, Mental health support

(Example of first paragraph of Dataset)

**6. Evaluate Model Performance:**

Evaluating Named Entity Recognition (NER) models for the medical domain involves assessing their ability to accurately identify and classify entities, such as diseases, medications, procedures, and symptoms, within medical texts. Here are several common evaluation metrics and approaches for NER in the medical domain:

**Precision, Recall, and F1 Score:**

* **Precision (P)**: It measures the accuracy of the model by calculating the ratio of correctly identified entities to the total entities predicted by the model.
* **Recall (R)**: It measures the model's ability to capture all relevant entities by calculating the ratio of correctly identified entities to the total entities present in the dataset.
* **F1 Score:** The harmonic mean of precision and recall, providing a balanced measure of a model's overall performance.

**Entity-Level Evaluation:**

Assessing the model's performance at the individual entity level. It involves comparing the model's predictions against the ground truth annotations for each entity type (e.g., disease, medication) separately.

**Token-Level Evaluation:**

Evaluating the model's performance at the token level, considering each individual word in the text. This approach accounts for partial matches and helps assess the boundary accuracy of identified entities.

**Confusion Matrix:**

Constructing a confusion matrix to visualize the model's performance, showing true positives, true negatives, false positives, and false negatives.

**Domain-Specific Metrics:**

Considering domain-specific metrics tailored to the medical field, such as the recognition of rare diseases or entities. These metrics can address the unique challenges of medical text.

**Relation Extraction Evaluation:**

For certain applications, evaluating not only the recognition of individual entities but also the extraction of relationships between entities. This is relevant in scenarios where understanding the interactions between medical entities is crucial.

**Error Analysis:**

Conducting a thorough error analysis to understand common patterns of mistakes made by the model. This can guide improvements in the training data or model architecture.

**Comparative Studies:**

Comparing the performance of different NER models, including state-of-the-art models and baseline approaches, to benchmark the system against existing solutions.



**ANALYSIS**

**Analysis of Both Solutions**

1. **Spelling Errors and Typos:**

* *Baseline:* The initial response doesn't mention the importance of correcting spelling errors and typos in medical terms.
* *Improved Solution:* Recognizes the significance of accurate spelling for medical terms, ensuring precise interpretation by the NER model.

1. **Handling Abbreviations:**

* *Baseline:* Doesn't emphasise the expansion of abbreviations for improved model understanding.
* *Improved Solution:* Recommends expanding abbreviations and introduces the use of medical ontologies or dictionaries for clarity.

1. **Sentence Segmentation:**

* *Baseline:* Lacks explicit mention of the importance of sentence segmentation for contextual analysis in NER.
* *Improved Solution:* Recognizes the significance of sentence segmentation for improved tokenization and contextual analysis.

1. **Tokenization:**

* *Baseline:* Briefly mentions tokenization without emphasising its importance in preprocessing.
* *Improved Solution:* Explicitly includes tokenization as a crucial step in preprocessing, ensuring a clearer understanding.

1. **Stopword Removal:**

* *Baseline:* Doesn't discuss the removal of common words (stopwords) to focus on more meaningful entities.
* *Improved Solution:* Introduces the removal of stopwords for streamlined input data, emphasising words with more significant meaning for NER.

1. **Stemming or Lemmatization:**

* *Baseline:* Doesn't address the importance of stemming or lemmatization for reducing dimensionality.
* *Improved Solution:* Highlights the significance of stemming or lemmatization to improve the model's ability to generalise.

1. **Biomedical-Specific Entities and Types:**

* *Baseline:* Lacks a clear identification and categorization of entities specific to the biomedical domain.
* *Improved Solution:* Clearly identifies biomedical entities and categorises them for better recognition and classification.

1. **Named Entity Recognition Model:**

* *Baseline:* Briefly mentions choosing a baseline NER model without delving into specific considerations or options.
* *Improved Solution:* Provides detailed considerations, including pre-trained biomedical NER models, BERT, and general-purpose NER models, showcasing a deeper understanding.

1. **Entity Extraction Example:**

* *Baseline:* Doesn't include an example of entity extraction from the model.
* *Improved Solution:* Demonstrates an example of extracted entities, offering a practical application of the trained model.

In summary, the improved solution not only adds crucial elements that were missing in the baseline response but also provides a more comprehensive, detailed, and practical guide for developing a state-of-the-art NER model in the medical domain.