# NLP Research (Disease Detection Based on Chat History)

## Overview of State-of-the-Art NLP Models

- **BERT**: A foundational NLP model that excels in understanding context and is used in healthcare for medical text analysis and diagnosis support.

- **BioBERT**: A variant of BERT pre-trained on biomedical literature, enhancing its performance in understanding medical jargon, ideal for electronic health records (EHRs).

- **ClinicalBERT**: Further fine-tuned on clinical notes, making it adept at interpreting patient data for healthcare decision-making.

- **GPT Models**: Such as ChatGPT, which can generate human-like text for interactive diagnostic chatbots. Studies show variable accuracy across different diseases.

**Bio_ClinicalBERT**: The demo will use this model, combining BioBERT and ClinicalBERT, trained on the MIMIC-III dataset, excelling in tasks like named entity recognition and text classification.(htt8)

## Types of Data Needed

**Labeled Datasets**: Chat transcripts annotated with diseases or symptoms to help the model learn language-disease associations.

**Chat Transcripts**: A large collection of anonymized chat logs from healthcare interactions, including successful and unsuccessful diagnoses. The demo utilizes the Symptom2Disease dataset, with 1,200 data points linking symptoms to diseases.

I have used Symptom2Disease dataset in my demo that focuses on the relationship between symptoms and diseases. It consists of 1,200 data points, each representing a unique combination of a disease label and a natural language symptom description. (htt9)

## Preprocessing Strategy

Data Cleaning: Address missing annotations or incomplete logs to maintain dataset integrity.

**Tokenization:** This step breaks down raw text into smaller units that the model can understand. The text data is converted into tokens using the BertTokenizer from Hugging Face in my demo.

**Truncation and Padding**: Ensure input sequences fit model length requirements by truncating longer texts and padding shorter ones.

**Data Splitting: Maintain distribution balance of diseases in training and testing datasets.**

## Model Evaluation

To effectively evaluate a disease detection model in healthcare, a combination of metrics is essential:

**Accuracy**: Proportion of correct predictions, useful but potentially misleading in imbalanced datasets.

**Precision**: Proportion of true positive predictions among all positive predictions; high precision indicates reliability in disease predictions.

**Recall (Sensitivity)**: Proportion of actual positive cases correctly identified; high recall means most cases are detected.

**F1 Score**: Harmonic mean of precision and recall, balancing both metrics, particularly valuable for imbalanced classes.

**ROC-AUC**: Assesses the model's ability to distinguish between classes at various thresholds; higher values indicate better performance.

Additionally, a **Confusion Matrix** provides detailed insights into true/false positives and negatives. In healthcare, ROC-AUC and F1 score are favored for their balance between precision and recall, especially in imbalanced situations. However, accuracy is used in my demo due to a balanced dataset.

## Evaluating the Model in Production

Continuous monitoring is essential for maintaining model performance:

**Feedback Mechanisms**: Gather input from healthcare professionals on misclassifications.

**Regular Performance Assessment**: Evaluate model metrics against historical data to identify performance degradation.

**Clinical Validation**: Collaborate with healthcare professionals to align model predictions with clinical outcomes.

**Detailed Reporting**: Provide comprehensive reports on performance metrics and their clinical implications.

### Kaynakça

(tarih yok). https://www.jmir.org/2023/1/e47621/ adresinden alındı
(tarih yok). https://www.javatpoint.com/bert-applications adresinden alındı
(tarih yok). https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT adresinden alındı
(tarih yok). https://www.kaggle.com/datasets/niyarrbarman/symptom2disease/ adresinden alındı