Biomedical Named Entity Recognition and Disease Clustering

Aryan Ringshia and **Neil Mankodi**
University of Michigan
aryanrr@umich.edu and nmankodi@umich.edu

1 Abstract

The large volume of biomedical literature necessitates advanced Natural Language Processing (NLP) solutions for efficient knowledge extraction. This paper presents a comprehensive approach to Biomedical Named Entity Recognition (NER) and Cluster Analysis, targeting disease-related keywords extraction and relationship unveiling within biomedical texts. Leveraging state-of-the-art NLP models, including BERT and ELECTRA, fine-tuned on biomedical data, our system achieved promising results, surpassing baseline metrics with an F1 score of 0.7. Our NER performance was promising, but our clustering analysis results were only modest improvements, indicating room for improvement in revealing hidden relationships between disease entities. This paper underscores the importance of continual refinement and exploration in tackling complex challenges within the biomedical domain.

Keywords: Biomedical Named Entity Recognition, NLP, BERT, ELECTRA, Cluster Analysis, F1 score, Disease Entities.

2 Introduction

In the realm of healthcare and medical research, the exponential growth of biomedical literature presents a formidable challenge: how to distill meaningful insights from an ever-expanding sea of text. As scientific knowledge burgeons, the demand for sophisticated tools to navigate, comprehend, and extract actionable information becomes increasingly pressing.

Our project addresses this challenge through the development of an advanced NLP solution tailored specifically for the biomedical domain. We focus on two core components: Biomedical Named Entity Recognition (NER) and Cluster Analysis.

2.1 Goals:

Our primary objective is twofold: to develop a robust Biomedical NER system capable of accurately extracting disease-related keywords from vast biomedical texts, and to conduct Cluster Analysis to unveil hidden relationships between these extracted disease entities.

2.2 Why Solving this Matters:

Central to our endeavor is the promise of expedited literature analysis, empowering healthcare professionals, researchers, and NLP practitioners to make informed decisions and drive breakthroughs in medical science. Our efforts aim to facilitate comprehensive knowledge extraction and analysis within the biomedical domain, ultimately fostering transformative advancements in medical science.

2.3 Our Approach:

To achieve our objectives, we adopted a strategy of fine-tuning existing state-of-the-art NLP models, including BERT and ELECTRA, to our specific task. By leveraging their preexisting knowledge about natural language, we aimed to tailor them to the intricacies of biomedical text, ensuring high performance in disease mention extraction. Subsequently, we employed these finetuned models to identify disease mentions and further analyzed them through Cluster Analysis to reveal insights into disease relationships and classifications.

2.4 What Others Have Done & How Our Approach Differs:

While existing approaches often focus solely on utilizing any one model, our approach distinguishes itself by the rigorous fine-tuning of multiple established language models. Moreover, while others typically employ base versions of these models, we leveraged resources like Great Lakes to explore larger versions with more parameters, resulting in enhanced performance.

2.5 How Well Did Our Approach Work:

Our models exhibited promising performance, convincingly surpassing baseline metrics. Specifically, our finetuned models achieved an impressive F1 score of 0.7 on the test data, demonstrating their efficacy in disease mention extraction.

2.6 Main Contributions / Learnings:

Through this project, we gleaned invaluable insights into the process of fine-tuning large NLP models, a skill set with broad applicability across various NLP applications and projects. Additionally, our analysis unveiled intriguing groupings within diseases, offering potential insights that could benefit the medical community at large.

3 Data

We utilized the NCBI Disease Corpus, which was introduced in the seminal paper "The NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization." This dataset has been meticulously annotated for biomedical NLP tasks. The corpus comprises 793 PubMed abstracts, containing 6,892 annotated disease mentions and 790 unique disease concepts sourced from resources like Medical Subject Headings (MeSH) and Online Mendelian Inheritance in Man (OMIM). 91% of the mentions map to a single disease concept which ensures clarity in concept normalization.

Start	End	Mention	Type
4	30	hereditary hemochromatosis	Modifier
170	197	autosomal recessive disease	DiseaseClass
198	224	hereditary hemochromatosis	SpecificDisease
2094	2120	hereditary hemochromatotic	Modifier

Table 1: Sample output of an example from the NCBI Disease Corpus.

The annotation process involved fourteen annotators, with each document reviewed by two annotators across three phases to ensure thoroughness and consistency. The dataset is partitioned into distinct subsets for training, development, and testing purposes. For preprocessing, we employed tokenization and lemmatization using SpaCy, while also removing stop words. This process aimed to transform the text into a more structured format suitable for our task. Additionally, we adopted the BIO notation for labeling the entities, marking the beginning (B), inside (I), and outside (O) of each entity. This labeling scheme provides a more granular and informative representation of disease mentions within the text. Below is a concise summary of an example from the NCBI Disease Corpus, showcasing annotated disease mentions, their abstract positions, mention types, and corresponding Disease IDs.

A summary of statistics for each dataset (train, dev, test), including the number of instances and the average length of texts is as follows:

Dataset	Instances	Average Text Length
Train	593	190
Dev	100	201
Test	100	204

Table 2: Dataset statistics

Furthermore, we present the distribution of classes in each dataset:

Classes	Train	Dev	Test
SpecificDisease	2972	412	555
Modifier	1289	214	264
DiseaseClass	769	126	121
CompositeMention	115	35	20

Table 3: Statistics of different classes in the datasets

4 Related Work

Naseem et al. (2021) have introduced a simple yet effective pre-trained language model for Biomedical Named Entity Recognition (NER). The authors have focused on the challenges faced with performing NER in the biomedical domain and gone into depth by comparing different state-of-the-art approaches (SOTAs) with their proposed model. A few of the main takeaways from this paper are the in-depth exploration of available biomedical datasets, SOTAs, and the exposure to the metrics that can be used for

analyzing the performance of the models for this specific task (precision, recall, F1). For our project, we used this paper as a reference for choosing the medical dataset, for finetuning deep learning models, and for selecting metrics for evaluation.

Lee et al. (2020) have introduced a language model that has been specifically designed for biomedical entities. The authors explore the challenges faced by standard, generalized SOTAs for biomedical NER. They focus on BERT and explore how this advanced deep learning model can be repurposed to work well with biological entities. A few of the main takeaways from this paper are the exposure to how an already established deep learning model can be refitted to work with domain-specific language, an in-depth discussion on the metrics that are useful for evaluating performance, and description of gold standard datasets that can be used to train language representation models. For our project, we used this paper as a reference for choosing the medical dataset, for finetuning deep learning models like BERT, and for selecting metrics for evaluation.

ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission" by Huang et al. (2019) introduces ClinicalBERT, a specialized language representation model designed for clinical text analysis and prediction tasks. The paper addresses the unique linguistic characteristics and challenges present in clinical notes, such as abbreviations, acronyms, and medical jargon. Our approach differs in its focus on biomedical named entity recognition (NER) and clustering, whereas "ClinicalBERT" focuses on clinical text analysis and hospital readmission prediction. "ClinicalBERT" typically uses clinical notes and electronic health records (EHRs) as its primary dataset source.

Ghosal et al. (2020) provide a comprehensive overview of data clustering techniques and their applications across various domains. The authors emphasize the significance of organizing vast volumes of data, including images, text, and genomic data, into meaningful clusters to facilitate pattern recognition and knowledge discovery. They discuss prominent clustering algorithms such as k-means, hierarchical clustering, density-based clustering, grid-based clustering, and model-based clustering, elucidating their methodologies and applications. Furthermore, the paper delves into the fields where clustering analysis has been effectively employed, with a particular focus on healthcare. In the medical domain, clustering algorithms have been instrumental in disease diagnosis, medical imaging, and personalized treatment. The review highlights the pivotal role of clustering in modernizing healthcare services, simplifying disease detection, and advancing medical research. For our project, we used this paper as a reference for implementing various clustering methods to group disease mentions.

Émilien Arnaud. et al. (2022) explores the use of pretrained BERT models to learn contextual embeddings from free-text triage notes, enabling the development of K-Means clustering models for healthcare data. Their findings suggest that pretrained transformer models, like BERT, are effective in learning contextual representations from healthcare text, which aligned with our project's goal of utilizing advanced architectures like BERT for biomedical NER. Their use of silhouette scores for evaluating the coherence of disease clusters guided our decision to use silhouette scores for clustering evaluation.

An existing project by DimasDMM (2024), serves as a baseline for disease named entity recognition (NER). The project primarily focuses on extracting disease mentions from text with a basic LSTM model and lacks hyperparameter tuning. In contrast, our approach aims to surpass this baseline by leveraging state-of-the-art architectures like BioBERT and BioALBERT which are specifically designed for biomedical data. Additionally, our project prioritizes hyperparameter tuning to optimize model performance further.

5 Methods

The outline of our methodology is as follows:

5.1 Data Preprocessing:

Initially, the data is in a raw text file format, comprising unstructured biomedical text. The data was preprocessed as follows:

1. Data Extraction - Using regular expressions (regex), we meticulously extracted pertinent information such as topics, abstracts, entities, and classes from the unstructured text data. This process ensured that only relevant infor-

mation was retained for subsequent analysis.

2. NLP Preprocessing -
   - Stop Word Removal - Common stop words, which add noise to the analysis and hold little semantic value, were systematically eliminated. This step focused the analysis on meaningful content, enhancing the quality of the results.
   - Tokenization - Text was broken down into individual words or tokens, facilitating granular-level processing and analysis. This tokenized representation served as the fundamental unit for subsequent NLP tasks.
   - Lemmatization - Words were transformed into their root form, ensuring consistency in representation, and reducing the dimensionality.
3. NER Notation - Labeled information, indicating the presence of named entities and their types, was utilized to assign BIO-based Named Entity Recognition (NER) notation to all tokens. This step prepared the data for training robust NER models capable of accurately identifying disease-related keywords within biomedical texts.
4. Feature Selection - The final dataset, comprising tokens annotated with their corresponding NER tags, was meticulously prepared for model training, validation, and evaluation. Features for model training were carefully selected based on insights gained from Hugging Face tutorials on fine-tuning models and NER tasks. This informed approach ensured that only relevant features were used, optimizing model performance and efficiency.

5.2 NER Modeling:

The deep learning models finetuned on our biomedical data to extract disease entities were as follows:

1. BERT - Bidirectional Encoder Representations from Transformers (BERT), a revolutionary language model pre-trained on large text corpora, capable of capturing intricate language nuances and contextual information. The architecture for this model can be seen in Figure 1. (Devlin et al. (2019))
2. BERT Large - A variant of BERT with an increased number of parameters, offering enhanced performance and capacity to handle complex tasks and datasets. (Devlin et al. (2019))

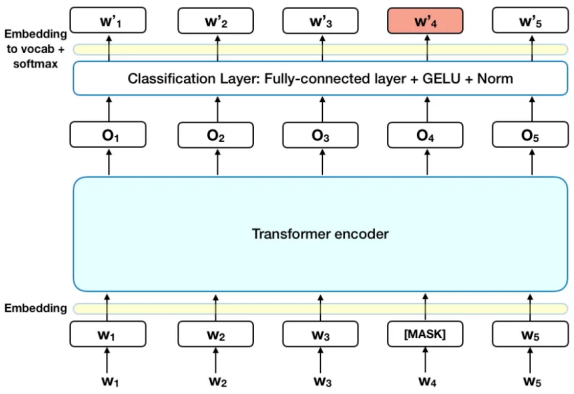

Figure 1: BERT Model Architecture.

3. ELECTRA - Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA), another transformer-based language model known for its efficiency and effectiveness in various NLP tasks. (Clark et al. (2020))
4. ELECTRA Large - A larger variant of ELECTRA, designed to tackle more demanding tasks and datasets, leveraging its increased model capacity and computational resources. (Clark et al. (2020))

5.3 Obtaining Word Representations:

Word2Vec, a popular technique for generating word embeddings, was employed to obtain numerical representations for extracted disease entities. This technique captures semantic relationships between words in a vector space, enabling meaningful comparisons and analyses. These vector representations are pivotal for clustering as they encapsulate semantic similarities between disease entities. By representing words numerically in a continuous vector space, Word2Vec facilitates the comparison of word meanings and enables clustering algorithms to identify and group related disease entities effectively.

5.4 Cluster Analysis:

The machine learning models explored to perform cluster analysis on the extracted disease entities were as follows:

1. Agglomerative - Agglomerative hierarchical

[1] https://pandas.pydata.org/
[2] https://numpy.org/
[3] https://spacy.io/
[4] https://huggingface.co/transformers/
[5] https://scikit-learn.org/stable/
[6] https://github.com/chakki-works/seqeval

clustering method that iteratively merges similar clusters based on distance metrics, progressively building a hierarchy of clusters.

2. KMeans - KMeans clustering algorithm that partitions data into K clusters based on similarity, minimizing intra-cluster variance to produce compact and well-separated clusters.
3. KMeans++ - An improved version of KMeans that selects initial cluster centers more effectively, enhancing the convergence and quality of the clustering solution.
4. GMM - Gaussian Mixture Model (GMM) clustering, which models clusters as a mixture of multivariate normal distributions, providing flexibility in capturing complex data distributions and cluster shapes.

6 Evaluation and Results

For our Named Entity Recognition (NER) task focused on identifying disease entities in biomedical texts, we employ the F1-score as our primary evaluation metric. The F1-score considers both precision and recall thus providing a balanced measure of model performance. We calculate the F1-score for disease mentions recognized by our NLP system against the gold standard annotations in the NCBI Disease Corpus.

We also establish two baseline comparisons. The first baseline involves random performance, where disease mentions are predicted without any contextual understanding. For each text, we decide the length of entities (in words) and how many entities to extract. We extract entities based on random decisions, with lengths randomly chosen between 1 to 3 words and a random number of entities between 1 to 5. Classes for entities are assigned randomly.

The second baseline uses the most frequent length of entities in the dataset. This requires minimal learning and serves as a straightforward reference. We determine the most frequent length of entities based on all available data. For each text, we extract a random number of entities from 1 to 3, and the most frequent class is assigned to all entities.

We utilized the Seqeval library to compute the scores for NER. Utilizing Seqeval ensured accurate evaluation by handling the BIO notation for entity labeling. The results are presented in the following Figure:

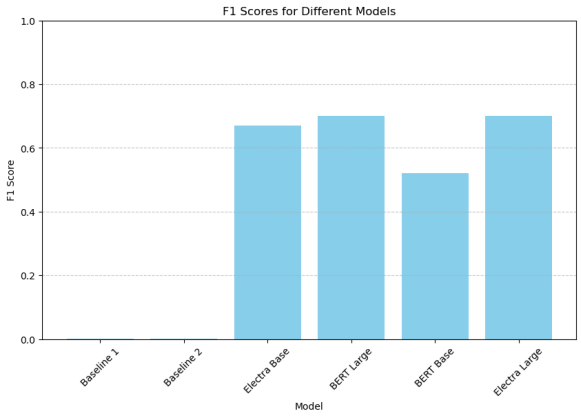

Figure 2: F1 Scores of Models.

Below is a table showing the results of our NER baseline methods and our methods:

Model	Precision	Recall	F1 Score
Baseline 1	0.002	0.002	0.002
Baseline 2	0.005	0.001	0.001
BERT-base	0.49	0.55	0.52
BERT-large	0.7	0.7	0.7
Electra-base	0.66	0.68	0.67
Electra-large	0.68	0.72	0.7

Table 4: Model Performance Comparison

For disease clustering based on shared characteristics, we use silhouette scores to evaluate clusters. This measures the cohesion and separation of generated clusters. Random clustering can be used as a simple baseline since it will not consider shared characteristics or hierarchical structure. As a simple baseline, we will randomly assign clusters to entities with the "SpecificDisease" class.

We performed clustering on SpecificDisease entities extracted from the BERT large model, as it yielded the best F1 scores on the NER task. The following plot illustrates the performance of different clustering algorithms alongside the baseline methods:

7 Discussion

7.1 Biomedical NER:

In our results, we found stark performance differences between our baseline methods and the models we employed. Baseline 1 and Baseline

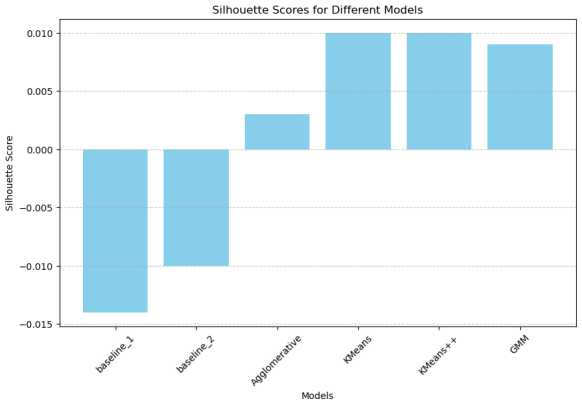

Figure 3: Silhouette Scores of Clustering Models.

2 yielded very low precision, recall, and F1 scores. This outcome was expected since these methods lack the contextual understanding necessary for accurate disease entity recognition. On the other hand, our NLP models, including BERT-base, BERT-large, Electra-base, and Electra-large, demonstrated significantly higher precision, recall, and F1 scores. This improvement highlights the effectiveness of leveraging pre-trained language models for biomedical NER tasks.

Increasing model complexity with BERT-large and Electra-large improved performance compared to their base versions, suggesting that sophisticated architectures better capture biomedical text intricacies and accurately identify disease entities. Particularly, the fine-tuned BERT-large model showed the best results among all models, highlighting the importance of fine-tuning pre-trained models on domain-specific data for optimal performance.

While our current approach demonstrates promising results, there remains ample room for refinement and exploration. Leveraging models pre-trained specifically for biomedical tasks, such as BioBERT, could potentially yield even better results. Exploring ensemble techniques to combine predictions from multiple models could enhance performance and robustness further. Another possible avenue for exploration would be delving into alternative evaluation metrics, such as entity-level metrics that capture deeper insights into model performance.

7.2 Cluster Analysis:

Our clustering models, including Agglomerative, KMeans, KMeans++, and GMM, produced Silhouette scores that outperformed both baseline approaches as expected. Despite modest improvements over the baselines, our results remain satisfactory for end-users, as they provide valuable insights into disease relationships and classifications within biomedical texts.

A few samples from each cluster are as follows:
Cluster 1:

- myalgia
- hereditary cancer
- iron overload
- protein s deficiency gene defect
- ulcerative colitis

Cluster 2:

- cleft lip palate
- sensorineural deafness
- upper respiratory infection
- mcleod syndrome
- hypomyelination

Upon examining the composition of the clusters, distinct patterns emerge, shedding light on potential relationships among disease entities. Cluster 1 encompasses a spectrum of conditions, from musculoskeletal complaints like myalgia to hereditary predispositions such as hereditary cancer and iron overload. This suggests a potential underlying genetic basis or common physiological pathways linking these diverse disorders. In contrast, Cluster 2 comprises diseases associated with developmental anomalies, such as cleft lip and palate, sensorineural deafness, and hypomyelination. The presence of these conditions within the same cluster hints at a shared developmental etiology or anatomical abnormalities, providing valuable insights for end-users seeking to discern underlying patterns and connections within disease clusters.

Our models outperform the baselines primarily due to the superior quality of our extracted disease entities and word representations. Unlike the baselines, which rely on random or most frequent approaches for entity extraction, our models leverage fine-tuned language models, resulting in higher-quality extracted entities. Additionally, our use of Word2Vec for obtaining word representa-

tions captures semantic relationships, enhancing the clustering process compared to TF-IDF-based representations used in the baselines.

Despite the improvements, our models' results remain modest. The lack of fine-tuning for Word2Vec may have impacted the quality of our vector representations, consequently affecting our silhouette scores. Furthermore, our reliance solely on word representations for clustering may overlook important contextual information, highlighting the need for additional features or contextual cues to enhance clustering performance.

8 Conclusion

In this study, we developed a tailored Natural Language Processing (NLP) solution for biomedical text analysis, focusing on Biomedical Named Entity Recognition (NER) and Cluster Analysis. Our approach, which involved fine-tuning state-of-the-art language models and exploring clustering algorithms, yielded significant insights.

Our NER models, particularly BERT Large and ELECTRA Large, exhibited impressive performance, surpassing baseline models and effectively identifying disease-related keywords within biomedical texts. This suggests the efficacy of fine-tuning large language models for domain-specific tasks, highlighting the importance of leveraging pre-existing knowledge in NLP applications.

Similarly, our exploration into Cluster Analysis revealed meaningful patterns and relationships among disease entities. While the Silhouette scores were modest, indicating room for improvement, the clustering algorithms showed promise in organizing disease entities into coherent clusters. This underscores the potential of leveraging unsupervised learning techniques to extract valuable insights from biomedical data.

Moving forward, there are several promising avenues for further exploration and refinement. Fine-tuning additional language models and incorporating domain-specific knowledge could enhance the accuracy and granularity of our NER models, facilitating more nuanced analyses of biomedical texts.

The code for this project can be found here.

9 Other Things We Tried

One avenue we pursued was prompt engineering for Named Entity Recognition (NER) using the Flan-T5-Large model. We chose Flan-T5-Large due to its pre-trained capabilities on a wide range of NLP tasks, expecting it to provide robust performance for our biomedical NER task. The prompt engineering process proved to be challenging as the model failed to properly give outputs according to the BIO (Begin, Inside, Outside) notation. Additionally, for disease clustering, we experimented with the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. However, DBSCAN clustered all points into a single cluster, indicating limitations in its applicability to our dataset.

10 What You Would Have Done Differently or Next

Reflecting on the project, if we were to revisit certain aspects with more time at our disposal, several avenues for improvement and exploration come to mind. One notable consideration is the fine-tuning of Word2Vec embeddings on our specific biomedical text corpus. While Word2Vec is a powerful tool for generating word representations, tuning it on our domain-specific data could lead to more accurate and contextually relevant embeddings. By capturing the intricate nuances of biomedical terminology, these refined word representations could potentially enhance the performance of our clustering analysis.

Additionally, an area ripe for further exploration is the investigation of a broader range of language models tailored specifically for the biomedical domain. While our project focused primarily on BERT and ELECTRA, the NLP landscape boasts a multitude of models with unique capabilities and advantages. Models such as HunFlair and scispaCy, among others, offer specialized features and pretrained embeddings that could complement our existing approach. By expanding our repertoire of language models, we could uncover new insights, address potential limitations, and further enhance the robustness of our NLP solution.

11 Group Effort

We divided responsibilities to cover a wide range of tasks efficiently in our project. One team member focused on implementing BERT Base and BERT Large models for NER. They also explored K-means and K-means++ clustering algorithms for disease clustering. On the other hand, the other team member delved into ELECTRA Base and ELECTRA Large models for NER. They also investigated Agglomerative and Gaussian Mixture Model (GMM) clustering methods to further enhance the disease clustering task. Both members fine-tuned pre-trained models and tried clustering algorithms on entities extracted from those models. Through this collaborative effort, we were able to tackle both tasks comprehensively.

References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

DimasDMM. 2024. diseases-ner: A repository for disease named entity recognition. `https://github.com/DimasDMM/diseases-ner`.

Attri Ghosal, Arunima Nandy, Amit Kumar Das, Saptarsi Goswami, and Mrityunjoy Panday. 2020. A short review on different clustering techniques and their applications. *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*, pages 69–83.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Usman Naseem, Matloob Khushi, Vinay Reddy, Sakthivel Rajendran, Imran Razzak, and Jinman Kim. 2021. Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Émilien Arnaud., Mahmoud Elbattah., Maxime Gignon., and Gilles Dequen. 2022. Learning embeddings from free-text triage notes using pretrained transformer models. In *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies - Scale-IT-up*, pages 835–841. INSTICC, SciTePress.