Fine–Tuning Named Entity Recognition For Clinical Tag Extraction Using Pretrained Language Models

Martin Deutsch, Alberto Moreno, Eva Sousa, Claire Cashmore, Johann Schmidt,

Adeyemi Victor Gbadamosi

JANUARY 2024

ABSTRACT

Background

With the increasing popularity of evidence-based medicine among physicians, there is a strong need to improve the techniques used to extract data from medical records. As of right now, the standard procedure calls for medical practitioners to thoroughly examine a wide range of information sources, which is a laborious process.

Methods

This study refines state-of-the-art (SOTA) pretrained language models (PLMs), such as BioElectra, BioLinkBert, PubMedBert, and Clinical Longformer, specifically for named entity recognition of medical entities. The resulting model facilitates the extraction of predefined entities from medical articles, streamlining information retrieval and elevating the quality of evidence-based medicine. Performance analysis involves a meticulous examination of the F1 score for each named entity tag across various learning rates.

Results

PubMedBert emerged as the frontrunner among the models, exhibiting an average F1 score of 0.64 and an output range of 0.3. This performance was achieved after training with a learning rate and a weight decay of 5e-5 and 0.03 respectively, employing default HuggingFace hyperparameters.

Conclusion

This research showcases the potential of fine-tuned language models in significantly improving the extraction of medical entities from articles. The demonstrated success, as indicated by the average F1 scores, underscores the efficacy of the finetuning Pretrained Language Models in advancing evidence-based medicine. The optimized model not only expedites information retrieval but also enhances the overall quality of extracted data, marking a valuable contribution to the evolving landscape of medical research and practice.

Introduction

In the realm of modern medicine, information is not just power; it is a lifeline. To deliver high–quality healthcare quickly and accurately, it is essential to be able to recognize and classify medical entities, such as diseases, drugs, patient data, and clinical processes. A branch of natural language processing called Entity Recognition (NER) is crucial to this procedure. Over the years, the intersection of healthcare and technology has seen significant breakthroughs, and one of the driving forces behind this change has been the quick development of PLMs (Shah et al., 2020; Poniszewska–Marańda et al., 2023).

With these recent achievements in pre–trained language models, it is no surprise that it is been applied to several industries with many more use cases being discovered(Chang et al., 2023). The healthcare and medical industry could also benefit from the application of PLM and several models

have been explored to address this (Wang et al., 2021). The ability of PLMs to automate the extraction of important entities—such as disease names, drug mentions, and medical procedures—from complex medical texts is essential in an era where the volume of medical literature and patient data is growing exponentially.

Evidence–based medicine is essential for advancing medical research and exploring new drugs effectiveness or side–effects but managing the vast amount of data generated can be daunting. For this, medical researchers synthesize knowledge by combining information from multiple journals to identify trends, patterns, and inconsistencies. Named Entity Recognition (NER) can play a crucial role in this process. NER is a powerful tool that can automate the identification and categorization of specific entities within clinical trial documents and medical journals, such as drug names, adverse events, and patient demographics. This technology streamlines data extraction and analysis, expediting the research process and quality improvement. By accurately identifying and categorizing key information, NER could ensure that researchers and medical professionals can make informed decisions, ultimately improving the efficiency and safety of clinical trials. However, we still have a lot of work to be done especially in the area of misinformation which underpins any advantage PLM could bring to healthcare(Yun et al., 2023).

In this investigation, we undertake named entity extraction from both the methods and abstract sections of published medical journals. Our approach employs state-of-the-art (SOTA) architectures, namely BioElectra (raj Kanakarajan et al., 2021), PubMedBert (Gu et al., 2021), BioLinkBert (Yasunaga et al., 2022), and Clinical Longformer (Li et al., 2022) all of which have been fine-tuned on medical literature. Notably, the models utilized in our experiments are adaptations specifically tailored for medical applications.

Building on the findings from a comprehensive survey conducted by (Wang et al., 2021), our chosen models demonstrated superior metrics compared to other medically fine-tuned pretrained large language models.

Furthermore, we incorporate Clinical Longformer into our investigation due to its unique advantage of not being constrained by the 512-token limitation observed in Bert models. This characteristic expands the scope and depth of our analysis, enhancing the model's capacity to capture and interpret extensive medical text.

Related works

Efforts of others who have worked in this field was looked at to find parallels and contrasts(Wen et al., 2021) labelled medical texts and using a dictionary of medical entities and finetuned a BiLSTM–CRF tagging model on Chinese electronics medical records. Using the Facebook research data and the CHIA dataset, (Tian et al., 2021) developed four transformers based pretrained models to accomplish named entity recognition. The aim was to develop a model that could be applied to automated electronic trial eligibility assessment. (Lei et al., 2014) used maximum entropy, conditional random fields, structural support vector machines, and support vector machines to conduct named entity recognition on notes that were randomly picked from discharge summaries including four entities. The outcome was to contribute to the use of Chinese named entity recognition systems. (Macri et al., 2022)annotated free text noted collated from outpatients' ophthalmology visits for NER. The aim was to provide a way to extract data from free text electronic notes.

Methodology

Datasets

Although there are many datasets which can be used for this task, a bespoke dataset extracted from 1549 records was created. A record consists of an extraction of a publication of a randomized controlled clinical trial with either only the abstract or abstract and full methods part. The records were annotated using the SWT tool software designed by Tut–All Software GmbH as described in figure 1. The SWT is a tool to automatically select, assess data of clinical publications (records) and extract variables out of them. A basic workflow can be seen in Figure 1.

The annotation process resulted in the identification and labelling of significant entities across 2816 tags. These tags represent various entity types within the SWT tool, such as (condition, design, subject). Subsequently, the abstracts and methods of each record were collated with their corresponding tags and organized into a Json file. To enhance data preparation, careful consideration was given to the context in which these entities appeared.

Further refining the dataset involved identifying and tagging entities within the abstracts and methods based on the context, resulting in the creation of two initial datasets. Recognizing the limitations of BERT-based models, which can only handle 512 tokens, one of the datasets underwent additional segmentation, leading to the generation of two additional datasets with varying total text lengths.

Abstract only – this contains the abstract of the record and the entity in the abstract.

Abstract with method – here the abstract from the record was merged with the methods increasing the total length of the text.

Abstract with method split into segments of 50 words (50 AM) – the abstract and method dataset was split into chunks of 50 words by counting the words.

Abstract with method split into segments of 512 chunks (512 AM) – With this the length of text in the dataset was maxed to 512 words.

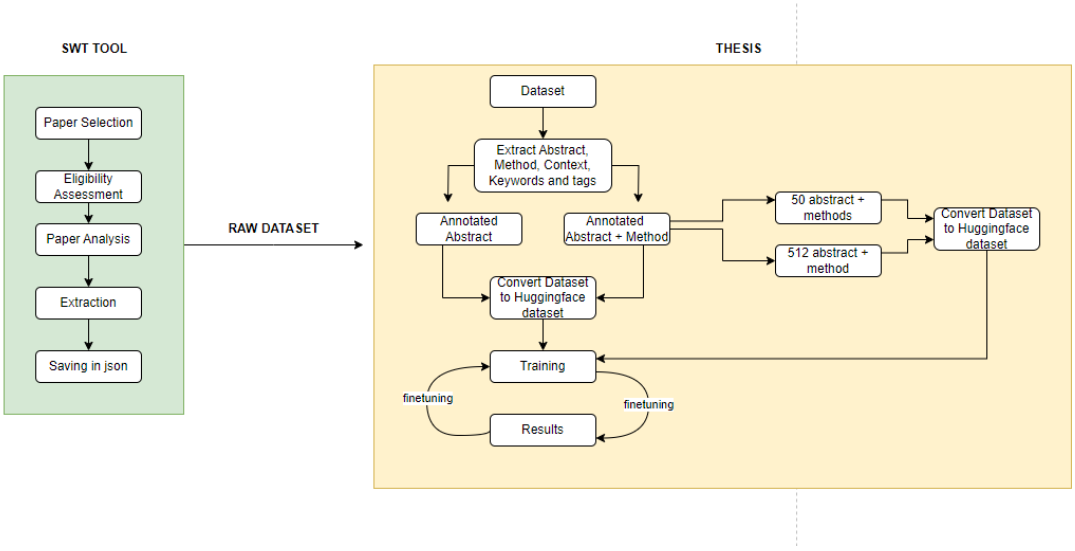

Figure 1. showing the project flow from the paper selection to the training and results using the PLMs

Table 1. The base description of the dataset after processing the raw data from the SWT tool in preparation for training and finetuning. The abstract with method data has 5 times the text and 3 times more unique texts in the data.

	Abstract only	abstract with methods
number of records	1549	1549
max length of tokens	731	5385
total number of texts	443716	2029783
total number of unique texts	15121	48916
total number of unique entity	5047	5047
all annotations	51154	51154

Table 2. final datasets used for the 80% training, 10% validation and 10% test.

	Abstract only	Abstract with method	50 AM	512 AM
total	1549	1549	50046	5563
training	1239	1239	40036	4450
validation	155	155	5005	557
test	155	155	5005	556

How we label the texts.

The raw Json file from the SWT tool contained 1549 dictionaries with 4 keys.

PMID– This is the PubMed reference number for records indexed in PubMed.

PAPER– This is the title of the record.

TEXT– this is the abstract or method depending on what was extracted by the SWT tool.

CONTEXT – this is the annotation with the surrounding words from the text.

ANNOTATIONS–this entities and their corresponding entity type.

To create a labelled dataset for training from the Json dictionaries, the context was the backbone, this ensured only the right entity in the text was matched. Following table 3, an empty label with equal length to the full–length text was created first. After which, the index of the context in the text was determined by matching the context in the full–length text. This served as the bounding range for the entity type, any other reoccurrence of the entity outside this range is ignored. But wherever no context was found or provided we matched every occurrence of the entities in the text.

Finally, we matched the entity index in the context with the empty labels. We removed entity types with very little representations but allowed for some imbalance to mimic real world cases as seen in table 4.

Table 3. Overview of the labelling methodology. starting from the empty labels to the labelled output with the correct tags.

empty label	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
actual text	This	was	a	controlled	study.	Primary	efficacy	measures	in	the	controlled	study
index of words	0	1	2	3	4	5	6	7	8	9	10	11
context	This	was	a	controlled	study.	Primary	efficacy					
tags				controlled	study							
labelled output	0	0	0	B–DES	I–DES	0	0	0	0	0	0	0


```
{
 "pmid": "15590953",
 "ner_tags": [7, 0, 1, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 "tokens": ["rivastigmine", "for", "dementia", "associated", "with", "parkinson
```


Figure 2. Final tag representations to be used in training.

Table 4. the distribution of entity types in the 4 datasets.

	Abstract only	Abstract + methods	50 AM	512 AM
B– SUBJECTS	1853	3661	3661	3661
B–CONDITION	1384	1949	1949	1949
B–DESIGN	826	1393	1393	1393
B–GROUP A	3300	5498	5498	5498
B–GROUP B	2531	4730	4730	4730
B–GROUP C	562	1330	1330	1330
B–GROUP D	167	395	395	395
I– CONDITION	3689	5648	5648	5648
I–DESIGN	3176	5423	5423	5423
I–GROUP A	2098	3030	3030	3030
I–GROUP B	1553	2610	2610	2610
I–GROUP C	548	1303	1303	1303
I–GROUP D	230	348	348	348
I–SUBJECTS	1418	3681	3681	3681
O	538946	2416829	2416829	2416829

Experiments

Google colab was used in the training and finetuning process with V100 and T4 GPU. Preliminary experiments had been done with BioElectra on this project, so naturally this model was selected as the base model for this thesis. The 4 datasets were split into training validation and test of 80%, 10%, 10% respectively and a maximum token length of 512. BioElectra was trained with the datasets and the learning rate was varied to get the best convergence f1 score as in table 6. The other models were trained using the best performing hyperparameters.

The f1 score was computed for each entity type in the dataset during the evaluation phase of the training. So, to determine the best model we compared the average f1 score for each experiment.

The ability of the model to learn evenly across all tags was looked at using the f1 score range (the difference between the maximum and minimum f1 score gotten for each tag).

Due to the ability of clinical Longformer to accept longer sequences of text, it was trained on a max token length of 4096. Amongst our datasets, abstract with method consistently had tokens which are above this, so it was the only dataset trained on clinical Longformer.

Hyperparameters

The hyperparameters used was from the HuggingFace trainer arguments. The epochs, learning rates and weight decay was varied see table 6.

A self-defined hyperparameter is the compute metric which was used to evaluate the model in training and testing. A custom metric function using seqeval was created which computes the f1score for each entity type in the model. The f1 score combined with the accuracy and loss gave a better oversight into the models' performance because it combined the precision and recall. The model's overall performance is then determined by calculating the average f1 score. The advantage of this method over using the default f1score is that the model performance on each tag can be observed.

Using a HuggingFace function to calculate the number of training parameters in the pretrained language models. BioElectra and PubMedBert has 108903183 training parameters while BioLinkBert and clinical Longformer has 107653647 and 148080399 respectively.

Table 5.hyperparameters used for the training.

hyperparameters	values
Optimizer	adamw_torch
Epoch	Varied
evaluation strategy	Epoch
learning rate	Varied
weight decay	Varied
Max token length	4096(clinical Longformer), 512 (BioElectra, PubMedBert, BioLinkBert)
Tokenizer	Fast tokenizer

.

Table 6. All trainings done and the different learning rate and weight decay used for the different models.

model	learning rate	weight decay	tags	data	punctuation marks
BioElectra	5e-5, 1e-5, 1e-6, 7.5e-4, 5e-4	0.003	15	abstract	yes
BioElectra	5e-5, 1e-5, 1e-6,7.5e-5, 2.5e-5	0.003	15	abstract with method	yes
BioElectra	5e-5, 1e-5, 1e-6, 5e-6, 1e-7, 1.25e-6, 7.5e-7, 2.5e-6	0.003	15	50 AM	yes
BioElectra	5e-5, 1e-5, 1e-6, 1e-4, 2.5e-5, 5e-6, 5e-4, 3e-5	0.003	15	512 AM	yes
BioElectra	5e-4	0.004	15	512 AM	yes
BioLinkBert	7.5e-5, 5e-4, 5e-5	0.003	15	abstract	yes
BioLinkBert	7.5e-5, 2.5e-5, 1e-5, 5e-5	0.003	15	abstract with method	yes
BioLinkBert	5e-6, 1e-6, 5e-5	0.003	15	50 AM	yes
BioLinkBert	2.5e-5, 3e-5, 5e-5	0.003	15	512 AM	yes
PubMedBert	7.5e-5, 5e-4, 5e-5	0.003	15	abstract	yes
PubMedBert	7.5e-5, 2.5e-5, 1e-5, 5e-5	0.003	15	abstract with method	yes
PubMedBert	5e-6, 1e-6, 5e-5	0.003	15	50 AM	yes
PubMedBert	2.5e-5, 3e-5, 5e-5	0.003	15	512 AM	yes
clinical longformer	8.5e-5, 2.5e-5, 5e-5, 5e-4, 1.25e-5, 1e-5	0.003	15	abstract with method	yes
clinical longformer	5e-5	0.001	15	abstract with method	yes
clinical longformer	5e-5	0.005	15	abstract with method	yes
clinical longformer	5e-5	0.0005	15	abstract with method	yes

Results

Our investigation into NER models yielded insightful findings. Specifically, BioElectra demonstrated optimal performance within a learning rate range of 5e–6 to 1e–4. For PubMedBert and BioLinkBert, the most effective learning rate was 5e–5, showcasing their robust performance within the 5e–6 to 5e–4 range. Notably, Clinical Longformer exhibited its best performance at a learning rate of 5e–05; however, its average F1 score fell below that of PubMedBert, BioLinkBert, and BioElectra. In our attempts to enhance results, various hyperparameters was experimented with, identifying a weight decay of 1e–4 as optimal across all experiments.

When the number of tags increased by 50 percent while keeping other factors constant, all models experienced a slight performance drop. Remarkably, among the datasets, the 50 AM consistently produced great results across all models. The removal of punctuation marks from certain datasets did not significantly impact model output.

In general, PubMedBert outshone other models with an average F1 score of 0.64, closely followed by BioLinkBert at 0.6. Although the differences in results were marginal, the decisive metric was the range, representing the disparity between the maximum and minimum F1 scores for entity types. This range serves as a statistical measure of a model's ability to learn across both majority and minority entity types. PubMedBert exhibited a range of 0.3, while BioLinkBert's range was notably higher at 0.76, as illustrated in Table 7.

The disparity in range is crucial, highlighting PubMedBert's superior capacity to learn minority tags. A compelling example is evident in Table 7, where PubMedBert successfully predicted entity type "Group D," while BioLinkBert produced blank predictions for the same category. This underlines PubMedBert's effectiveness in capturing nuances within the data, showcasing its potential for improved performance in handling minority entity types.

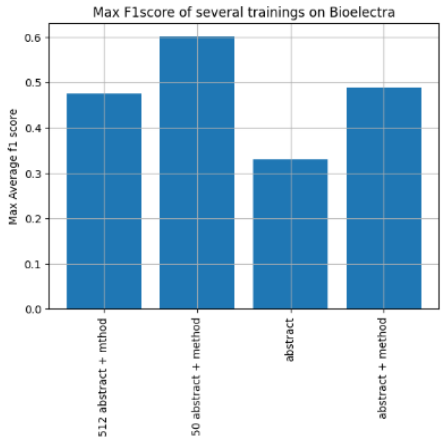

Figure 3. BioElectra performed best on the 50 abstract + methods dataset with an average f1 score of 0.6.

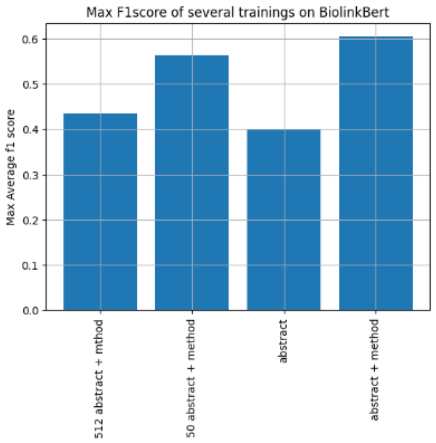

Figure 4. BioLinkBert performed best on the 50 abstract + methods dataset with an average f1 score of 0.6.

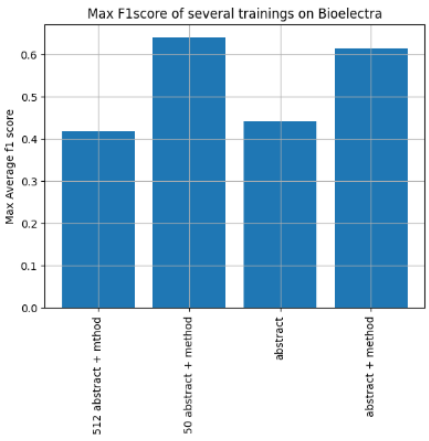

Figure 5 PubMedBert performed best on the 50 abstract + methods dataset with an average f1 score of 0.64.

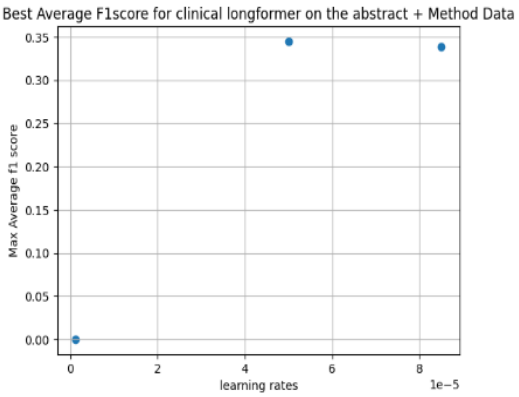

Figure 6. Clinical Longformer was only trained on abstract + method and the best average f1 score is 0.35.

Table 7.Best performing metrics for respective model and the PubMedBert has a short range and better average f1 score.

model	BioElectra	BioLinkBert	PubMedBert	clinical Longformer
F1 B–cond	0.550409	0.651489	0.591656	0.350158
F1 I–cond	0.641374	0.627758	0.591304	0.439058
F1 B–des	0.618337	0.652361	0.627767	0.525458
F1 I–des	0.732379	0.759283	0.743383	0.587174
F1 B–subj	0.585276	0.664596	0.553165	0.333884
F1 I–subj	0.535414	0.605505	0.502674	0.379399
F1 B–group A	0.724379	0.688406	0.744641	0.456688
F1 I–group A	0.601415	0.547434	0.572997	0.477845
F1 B–group B	0.672773	0.556943	0.685285	0.361809
F1 I–group B	0.481178	0.52669	0.49505	0.226744
F1 B–group C	0.644444	0.612378	0.718686	0
F1 I–group C	0.533784	0.358621	0.602941	0
F1 B–group D	0.709677	0	0.807175	0
F1 I–group D	0.375	0	0.710744	0
average f1score	0.600417071	0.604288667	0.639104857	0.344851417

Discussions

The modern era of PLMs commenced with the introduction of word embedding models, such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). These models formed the basis for computers to encode the relationships and context between words in a text. Over the years, the field has witnessed significant progress, from the transformer archictecture (Vaswani et al., 2017) to techniques like Embedding from Language Models (ELMo) (Peters et al., 2018) and Universal Language Model Fine-Tuning (ULMFiT) (Howard & Ruder, 2018). Transformer models have particularly revolutionized natural language processing by demonstrating improved ability to manage dependencies between words in a sentence(Naseem et al., 2021), leading to breakthroughs in Named Entity Recognition (NER) that align well with Evidence-Based Medicine.

The adopted approach involved the use of the IOB format, popularized by the CoNLL NER task (Sang & De Meulder, 2003) and initially proposed by Ramshaw & Marcus in 1995. In this format, "I" designates words inside an entity, "O" denotes non-entity words, and "B" marks the beginning word of an entity(Ramshaw & Marcus, 1999). (Cho et al., 2013) demonstrated that the IOB format can yield excellent results in named entity recognition.

In real-world scenarios, data imbalances are common. The model's performance on skewed data, specifically its ability to produce results for minority tags, was observed. PubMedBert demonstrated that achieving this is possible without utilizing data balancing techniques, showcasing the best F1 scores for the minority entity type Group D in comparison to others. Methods proposed by (Dor et al., 2020; Shaikh et al., 2021) aim to enhance the ability of pretrained language models to learn from minority entity types.

A surprising outcome was the clinical Longformer's approximately 50 percent lower average F1 score compared to other models, despite its capability to process longer text sequences. While (Lei et al.,

2014)suggested that increased text input should lead to better learning, our results contradicted this assumption. Another unexpected finding was the 50 AM's performance, outperforming datasets with longer texts and closely matching the abstract with methods dataset, on all model training.

While F1 scores in this work were lower compared to other studies, this could be attributed to the bespoke dataset used. The results underscore the importance of dataset format, quality, and labelling processes. Many NER works rely on publicly available datasets that have been refined and labelled to suit the task, resulting in high F1 scores. However, for applications with numerous, complex, and interrelated entity types, bespoke datasets become necessary. Additionally, the methods for tagging datasets for training need to be fine-tuned and standardized. The possibility of improving results by segmenting the dataset with a focus on using complete sentences for better learning is also considered.

To address overfitting and enhance model generalization, increasing the number of datasets is essential. Exploring text data augmentation works and suggestions (Wei & Zou, 2019) is one way to significantly increase the dataset size and balance tags. Finetuning the clinical Longformer could benefit from warm-up steps, as described in(Beltagy et al., 2020; Su et al., 2021), potentially improving performance and enabling evaluation of clinical text with extended token lengths. Varying epochs during training helped alleviate overfitting, but this remained a persistent issue. The adjustment of learning rates and weight decay was crucial for achieving optimal convergence.

Conclusion and future works

To fully utilize the biomedical pretrained language models in clinical applications, the performance through refining the dataset is necessary. Future annotations with more data have the potential to improve the model. The current methods are to label it once but ignore it in every other occurrence however this could potentially confuse the model if they are used in different contexts. (Gu et al., 2021) has shown that the tagging methodology and dummification could have different results in PubMedBert, future works could explore creating other bespoke datasets with other methods. In our experiments, the dataset was chunked by linearly counting a set number of tokens, this can be improved by chunking with full stops or other sentence terminating punctuations. This will ensure the chunks the models are learning from are complete sentences or phrases. Also, augmentation techniques could be used to improve the dataset or more records could be explored using the SWT tool to increase the dataset for future works.

The trustworthiness of the results is an important factor that could determine if finetuned PLMs can be used in the field.

BIBLIOGRAPHY

Beltagy, I., Peters, M. E. & Cohan, A. (2020) Longformer: The long–document transformer. *arXiv preprint arXiv:2004.05150*.

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C. & Wang, Y. (2023) A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Clark, K., Luong, M.–T., Le, Q. V. & Manning, C. D. (2020) Electra: Pre–training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. & Kuksa, P. (2011) Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE), 2493− 2537.

Devlin, J., Chang, M.–W., Lee, K. & Toutanova, K. (2018) Bert: Pre–training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J. & Poon, H. (2021) Domain–specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1), 1–23.

Habibi, M., Weber, L., Neves, M., Wiegandt, D. L. & Leser, U. (2017) Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14), i37–i48.

Howard, J. & Ruder, S. (2018) Universal Language Model Fine–tuning for Text Classification Melbourne, Australia, July. Association for Computational Linguistics.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. & Dyer, C. (2016) Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Lei, J., Tang, B., Lu, X., Gao, K., Jiang, M. & Xu, H. (2014) A comprehensive study of named entity recognition in Chinese clinical text. *Journal of the American Medical Informatics Association*, 21(5), 808–814.

Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H. & Luo, Y. (2022) Clinical–Longformer and clinical–bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*.

Macri, C., Teoh, I., Bacchi, S., Sun, M., Selva, D., Casson, R. & Chan, W. (2022) Automated Identification of Clinical Procedures in Free–Text Electronic Clinical Records with a Low–Code Named Entity Recognition Workflow. *Methods of Information in Medicine*, 61(03/04), 084–089.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013) Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nadeau, D. & Sekine, S. (2007) A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1), 3–26.

Naseem, U., Razzak, I., Khan, S. K. & Prasad, M. (2021) A comprehensive survey on word representation models: From classical to state–of–the–art word representation language models. *Transactions on Asian and Low–Resource Language Information Processing*, 20(5), 1–35.

Pennington, J., Socher, R. & Manning, C. D. (2014) Glove: Global vectors for word representation, *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018) Deep Contextualized Word Representations New Orleans, Louisiana, June. Association for Computational Linguistics.

Poniszewska–Marańda, A., Vynogradnyk, E. & Marańda, W. (2023) Medical Data Transformations in Healthcare Systems with the Use of Natural Language Processing Algorithms. *Applied Sciences*, 13(2), 682.

raj Kanakarajan, K., Kundumani, B. & Sankarasubbu, M. (2021) BioElectra: pretrained biomedical text encoder using discriminators, *Proceedings of the 20th Workshop on Biomedical Language Processing*.

Ramshaw, L. A. & Marcus, M. P. (1999) Text chunking using transformation–based learning, *Natural language processing using very large corpora*Springer, 157–176.

Sang, E. F. & De Meulder, F. (2003) Introduction to the CoNLL–2003 shared task: Language–independent named entity recognition. *arXiv preprint cs/0306050*.

Shah, D., Dixit, R., Shah, A., Shah, P. & Shah, M. (2020) A comprehensive analysis regarding several breakthroughs based on computer intelligence targeting various syndromes. *Augmented Human Research*, 5, 1–12.

Su, X., Miller, T., Ding, X., Afshar, M. & Dligach, D. (2021) Classifying long clinical documents with pre–trained transformers. *arXiv preprint arXiv:2105.06752*.

Tian, S., Erdengasileng, A., Yang, X., Guo, Y., Wu, Y., Zhang, J., Bian, J. & He, Z. (2021) Transformer–based named entity recognition for parsing clinical trial eligibility criteria, *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017) Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, B., Xie, Q., Pei, J., Chen, Z., Tiwari, P., Li, Z. & Fu, J. (2023) Pre–trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, 56(3), 1–52.

Wang, B., Xie, Q., Pei, J., Tiwari, P. & Li, Z. (2021) Pre–trained language models in biomedical domain: A systematic survey. *arXiv preprint arXiv:2110.05006*.

Wei, J. & Zou, K. (2019) Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Yasunaga, M., Leskovec, J. & Liang, P. (2022) Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*.

Yun, H. S., Marshall, I. J., Trikalinos, T. & Wallace, B. C. (2023) Appraising the Potential Uses and Harms of LLMs for Medical Systematic Reviews. *arXiv preprint arXiv:2305.11828*.

Beltagy, I., Peters, M. E. & Cohan, A. (2020) Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Cho, H.-C., Okazaki, N., Miwa, M. & Tsujii, J. i. (2013) Named entity recognition with multiple segment representations. *Information Processing & Management*, 49(4), 954-965.

Dor, L. E., Halfon, A., Gera, A., Shnarch, E., Dankin, L., Choshen, L., Danilevsky, M., Aharonov, R., Katz, Y. & Slonim, N. (2020) Active learning for BERT: an empirical study, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J. & Poon, H. (2021) Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1), 1-23.

Howard, J. & Ruder, S. (2018) Universal Language Model Fine-tuning for Text Classification Melbourne, Australia, July. Association for Computational Linguistics.

Lei, J., Tang, B., Lu, X., Gao, K., Jiang, M. & Xu, H. (2014) A comprehensive study of named entity recognition in Chinese clinical text. *Journal of the American Medical Informatics Association*, 21(5), 808-814.

Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H. & Luo, Y. (2022) Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013) Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Naseem, U., Razzak, I., Khan, S. K. & Prasad, M. (2021) A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5), 1-35.

Pennington, J., Socher, R. & Manning, C. D. (2014) Glove: Global vectors for word representation, *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018) Deep Contextualized Word Representations New Orleans, Louisiana, June. Association for Computational Linguistics.

raj Kanakarajan, K., Kundumani, B. & Sankarasubbu, M. (2021) BioElectra: pretrained biomedical text encoder using discriminators, *Proceedings of the 20th Workshop on Biomedical Language Processing*.

Ramshaw, L. A. & Marcus, M. P. (1999) Text chunking using transformation-based learning, *Natural language processing using very large corpora*Springer, 157-176.

Sang, E. F. & De Meulder, F. (2003) Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Shaikh, S., Daudpota, S. M., Imran, A. S. & Kastrati, Z. (2021) Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models. *Applied Sciences*, 11(2), 869.

Su, X., Miller, T., Ding, X., Afshar, M. & Dligach, D. (2021) Classifying long clinical documents with pre-trained transformers. *arXiv preprint arXiv:2105.06752*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017) Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, B., Xie, Q., Pei, J., Tiwari, P. & Li, Z. (2021) Pre-trained language models in biomedical domain: A systematic survey. *arXiv preprint arXiv:2110.05006*.

Wei, J. & Zou, K. (2019) Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Wen, C., Chen, T., Jia, X. & Zhu, J. (2021) Medical named entity recognition from un-labelled medical records based on pre-trained language models and domain dictionary. *Data Intelligence*, 3(3), 402-417.

Yasunaga, M., Leskovec, J. & Liang, P. (2022) Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*.