Comparative Study of Named Entity Recognition Models for Medical Transcription

Akshay Srinivasan,
300223628,
University of Ottawa

Surya Kiran Suresh,
300269939,
University Of Ottawa

Abstract

Named Entity Recognition for medical text has been in the research area for quite a while now. Performing NER on Medical transcription has many applications including making a brief summary of the conversation, listing out medicines for dispensary, identifying medical history using keywords, etc. Many researchers have tried to build an ideal NER model for the medical domain. Few such attempts resulted in the introduction of models such as BIOBERT, PubmedBERT, BlueBERT, etc that could be used for a variety of NLP tasks. At the same time, it is also observed that models that are pretrained on an English corpus and further fine-tuned on a domain-specific dataset, give good results but are subpar when compared with the former. In this study, we compare the performance of these domain-specific models pretrained on medical corpora with language models (RoBERTa, BERT, etc) pretrained on English Corpus. Both of these models are finetuned on a custom-curated dataset and the results are analysed. It is found that models pre-trained on biomedical text show a good improvement in the performance of the medical NER task and BIOBERT seemed to perform slightly better than the other models.

Introduction

Named entity recognition is the task of recognising key entities like location, person, organization, etc. from the text. Medical NER, however, focuses more specifically on identifying the key terms in the field of biomedicine such as diseases, chemicals, drugs, etc. In the field of Medicine, there are plenty of Medical Transcriptors who transcribe the audio recordings between the doctor and the patient recorded during the consultation. With the advancement of technology in the field of NLP, Automatic Speech Recognition (ASR) (Chiu C-C. et. al., 2017) systems have been adopted to perform medical transcription. However, people are still working on building a proper system for performing medical NER on these transcripts.

With the advent of pre-trained language models, word embedding models like word2vec (Mikolov et. al., 2014) and FastText (Bojanowski P et. al., 2016) have improved the performance of NLP tasks. However, these embeddings learned were mostly contextually independent and there was a need to understand the occurrences of words in the text along with the context in which they occurred. This led to the rise of context-dependent word embeddings that implemented transformers as part of their language models and would be pre-trained on a given corpus of text to learn contextualised representations of words. BERT(Bidirectional Encoder Representations from Transformers) (Devlin J et. al., 2018) was one such model that was the pioneer of a lot of transformer-based language models. Using BERT-like transformers significantly enhanced the performance of models to perform NER tasks. All these models would be initially pre-trained on an English corpus and would further be fine-tuned on the NER dataset to detect entities.

The need for NER in medical transcription arises because, in real-life scenarios, the medical transcription text has to be read by some clerk to find the diagnosis of the patient, medicines prescribed, treatments that were taken, insurance codes, etc. NER can be leveraged to identify these entities from the transcribed text and easily isolate important information from the entire text. Medical NER also gives rise to a number of other applications that include Summarization of medical text, Question and Answering, Relation Extraction, etc.

The volume of biomedical text out there continues to exponentially increase with scientists publishing pub-med articles every day. This has led to a demand for language models pre-trained on biomedical text in order to perform NLP tasks accurately. Since the task of medical NER aims at performing NER on bio-medical data, researchers have proposed language models that learn contextual word representations directly from bio-medical text rather than learning word representations of the English corpus. These models were proposed to ensure an improvement in the performance of NLP tasks like Medical NER. The goal of this study is to study the difference in the performance between a language model pre-trained on the English corpus and a model pre-trained on the biomedical text directly.

Apart from the fact that there's a need for models to understand bio-medical text, there is also an increasing demand for proper bio-medical datasets to perform NER and identify biomedical texts. Most existing corpora focus on tagging a single entity and training the model on a corpus for detecting a single entity is not efficient in the long run. We will have to train the model again and again on multiple corpora to identify different entities. Hence, an ideal dataset would be something that's been annotated with multiple entities so that the language model would be fine-tuned in one shot on this dataset. This study also aims at putting together a corpus for medical NER by utilizing data from different bio-medical corpora so that the model would identify entities like chemicals and diseases.

The study focuses on building a flexible pipeline for performing Medical NER on biomedical texts. The NER model would be detecting entities that include chemicals and diseases since these are the key terms that anyone looking at a medical transcription would want to know. The pipeline is implemented by utilizing a pre-trained transformer model like BERT to perform tokenization and further fine-tune on the dataset to identify medical entities. The transformer part of the pipeline will be implemented by 4 different transformers- BERT, ROBERTA (Liu Y et. al., 2019), BIOBERT (Lee J et. al., 2019) and Microsoft's PubMedBERT (Gu Y et. al., 2022), out of which the first 2 are trained on the English corpus and the last 2 are pre-trained on bio-medical text. The model is fine-tuned on all these transformer models separately to study and evaluate the impact on the performance of the corresponding transformers. Finally, the performance of the 4 implementations is compared to study the improvement in performance when a model pre-trained on biomedical text is used.

Related Work

Medical NER is a grossing research area in recent times with the rise of the pandemic. More and more medical data is made available for the public to build tools that can help the healthcare industry. Even though the data is available, it is in an unstructured format (Missen et.al. 2020) and to convert it into a structured one, there is a need for professionals who are trained in the industry. There have been multiple approaches proposed for the Medical NER task lately. There are a lot of challenges when it comes to preprocessing text for biomedical entities. Pubmed,WebMD and PMC articles are currently the primary source of medical text in the industry, but these articles contain a lot of ambiguity in words due to the frequent usage of abbreviations, synonyms, nested Named Entities, etc (Nayel et al., 2019). For eg. the abbreviation "CLD" can be equivalent of "Chronic Lung Disease" or "Cholesterol-lowering Drug", which falls under 2 different entities. This makes the task an active research field.

Perera et. al. summarizes the different approaches, challenges and results of the existing solutions for Named entity extraction and Relation Detection. The article mentions the usage of different algorithms for NLP tasks such as Word representational models, BERT, Rule-based models, Dictionary-based models, Deep learning methods such as LSTM, RNNs etc. Yifan Peng et. al. proposes an evaluation benchmark named Biomedical Language Understanding Evaluation (BLUE) to facilitate research in the development of pre-training language representations in the biomedicine domain. This evaluation benchmark consists of five NLP tasks and includes ten datasets. These tasks are tested on ELMo, BioBERT, a SOTA model and their own BERT model and the results are posted. They conclude that BERT models perform better than other models for NLP tasks. A hybrid approach to NER on biomedical text based on PubMed articles is proposed in R. Ramachandran et. al. They use Transfer learning along with a dictionary that stores all the entities extracted from the unlabelled text to train the hybrid model. The proposed hybrid model was able to achieve an F1 score of 73.79% for 5 entities that were proposed in the paper.

BioBERT(Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) is the state of the art NLP model in the medical domain to perform common NLP tasks. Lee J et. al. propose a BERT model pretrained on Pubmed and PMC data that can be further fine-tuned for NLP tasks such as NER, Relation Extraction, Question Answering, etc. The model is first trained on English Wikipedia and BookCorpus to learn the language representations and then further trained on the Domain-specific Pubmed and PMC articles.

All these research works have led us to the current era where domain-specific transformer models are being introduced to fulfil specific NLP tasks with higher efficiency.

Methodology

This section discusses the overall pipeline of the NER system and further talks about the principle behind the different transformers being implemented.
The overall architecture of the NER model as shown in Figure 1, consists of preparing the bio-medical dataset for training, after which it will be passed on to the data preprocessing stage. Once, the data is converted into the necessary format according to the medical-NER task, it is then ready to be passed on to the transformer. The transformer's tokenizer is first utilized to tokenize the input training data and convert it into a format that is compatible with the particular transformer in use. The tokenized input is further fed to the transformer model for fine-tuning to detect medical entities.

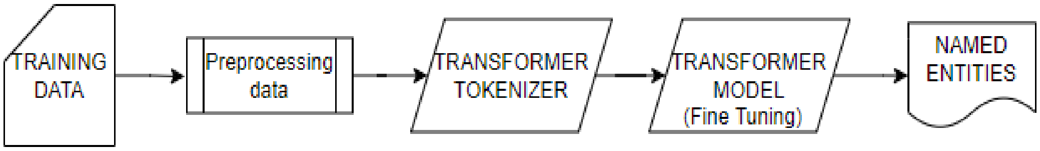

Figure 1

Data Preparation

To prepare data, the bio-medical corpus from NCBI corpus(Doğan, R et. al.) and the BC5CDR Corpus (Li Jao & Sun et. al.) were chosen. The NCBI corpus is a collection of around 1000 pubmed articles which are marked with all occurrences of diseases. The BC5CDR corpus is a collection of 1500 pubmed articles which have been annotated with chemical and disease instances separately. Data from both these corpora are combined and ensured that the occurrences of chemicals and diseases are annotated using the BIO schema of NER. Every single word from the combined corpus will be tagged with either of the following tags- B-Chemical, I-Chemical, B-Disease, I-Disease and O(out of scope).

Preprocessing

To implement a transformer-based language model like BERT, the input data must be transformed into an acceptable format, so that each sentence can be given to the model and the relevant word embedding can be obtained.Firstly, the combined corpus is split sentence-wise. The labels(tags for entities) corresponding to every single word in each sentence are extracted. The transformer model is always fed with a sentence of a specific length as input. The maximum length of a phrase is usually determined by the data we're dealing with. To make up for sentences that are shorter than this maximum length, paddings (empty tokens) will be added to the sentences. The sentences are further tokenized by passing them to the transformer tokenizer. The role of a transformer tokenizer is to convert sentences into tokens(words). The tokenizer used by these models is known as a word piece tokenizer, where a token can be a word or a part of a word. For example, the word 'surf' can be considered as a token and the word 'surfing' will be split into 2 tokens 'surf' and '##ing". The continuation of a word is represented with a double hash(##) at the beginning of the token. A unique id for identifying each token is also generated in during tokenization which will be fed to the next step along with the tokens. Once the sentences are tokenized, it is ready to be fed to the transformer model for fine-tuning.

Fine-Tuning Using Transformers

The fine-tuning section of the pipeline represented in figure 1 is implemented using 4 different transformer models which will be discussed in the upcoming sections. Each of these models will be having its own tokenizer and then will be fine-tuned. The study aims at assessing the impact of these transformer-based models on the medical NER system.

1. BERT(Bi-directional Encoder Representations from Transformers)

BERT is a pre-trained transformer model trained on the English Book Corpus and it will be incorporated in this study to serve as a baseline when comparing with the other transformer models. As the name suggests BERT utilizes Transformers (Vaswani A et. al.), which consists of attention mechanisms that strive to learn the contextual relations of words in a text. Ideally, transformers include 2 mechanisms - an encoder that generates a representation for the input text and a decoder that tries to re-construct or generate a prediction depending on the task at hand. Out of the 2 mechanisms, only the encoder technique is required because BERT's purpose is to generate a language model.

Unlike the traditional directional models that try to read through text sequentially either from right-to-left or left-to-right, Transformer’s encoder tries to scan the word sequence in one shot. This led the model to be called bi-directional, even though ‘non-directional’ would have been a more accurate term. This ability of the transformer enables it to understand the context of a particular based on its neighbouring words(surroundings).

Figure 2 outlines the working of BERT on a high level. The tokens obtained from the tokenizing stage will be embedded in the form of vectors(w1, w2, w3, w4, w5) before being processed by the neural network. The output obtained from the Encoder will be a continuous sequence of vectors denoted by size H(w’1,w’2,w’3,w’4,w’5), where each input token corresponds to each of the vectors occurring at the same index. In order to aid the model in contextual learning as well as understanding word occurrences during pre-training, BERT makes use of two training strategies:

A. **Masking**
Everytime a sequence of words is fed into BERT, 15% of words in the sequence are replaced with a ‘mask’ token i.e. the token will be hidden from the model. Now, based on the context obtained from the non-masked words, the model tries to predict the token value of these hidden words. This is done by stacking a classification layer on the encoder and calculating the prediction probabilities for the output generated by this layer.

B. **Next Sentence Prediction**
The model accepts input in pairs of sentences and tries and learns to guess whether the second sentence in a pair is the original document's next sentence. It is made sure that the second sentence in the pair is the original document’s next sentence 50% of the time, while in the remaining 50%, a random sentence is put in the pair. This random sentence would not be related to the first sentence. In the pre-training stage, the model is pre-trained on unlabelled words which were extracted from Book Corpus and the English Wikipedia corpus.

Finally, in fine-tuning, the model will be modified by adding an extra classification layer for predicting the NER label. the model will be receiving a sequence of text(bio-medical text in our case) and will be required to tag the sequence with different entities (‘chemical’ and ‘disease’) appearing in that text.

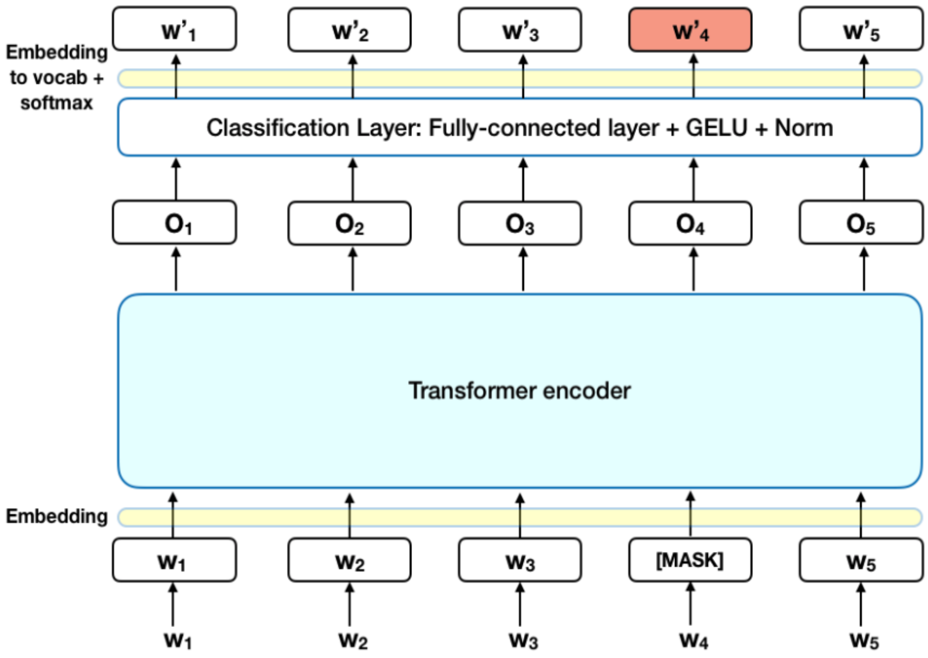
w'1
w'2
w'3
w'4
w'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

Figure 2

2. ROBERTA(Robustly Optimized BERT pretraining Approach)

As the name suggests Roberta is a much more optimized version of BERT that is proven to perform more efficiently than BERT. The authors of Roberta claim that BERT is severely under-trained and that's why they put forward some improvements in the following aspects.

A. Pre-training Data

Since the authors felt that BERT was under-trained, Roberta was trained on a much larger dataset that included English articles from the CommonCrawl News Data, Open Web-text data, and story-like data from the Common Crawl dataset.

B. Dynamic Masking

The objective of the masking stage in BERT is to mask some tokens in each sequence and predict these tokens. In BERT, masking is done only once, meaning that the pattern of masking is the same for all the steps in training. In Roberta however, the authors have replicated the training data multiples times so that the masking pattern is different for each of these replicated instances of training data. Apart from this, a dynamic masking pattern is incorporated where a new pattern for masking is created every time a word sequence is given to the model

C. Removing NSP

The authors of Roberta also proved that there's no necessity to include a Next Sentence Prediction task, because the model would perform better even without it

D. Batch Size

Finally, it was found that having large Batch Sizes during pre-training enables faster optimization and improvement in the end-task performance if fine-tuned properly.

E. Tokenization

For performing Tokenization, RoBERTa implements Byte-Pair Encoding(BPE) scheme, unlike its BERT counterpart which uses character level BPE.

RoBERTa can be fine-tuned in a manner similar to BERT by adding an additional classification layer for performing NER on medical data.

3. BIOBERT

BioBERT is a transformer language model pre-trained in the biomedical domain. The architecture of the BIOBERT is of the standard BERT model, however, it varies in the overall pre-training and training data used for the process. BERT is generally trained on English Wikipedia articles. But if we were to perform an NLP task using BERT on bio-medical text, the model won't do a good job of identifying words that are very domain specific(proper nouns and other bio-medical terms)and wouldn't be common in an English corpus. Hence, BIOBERT is introduced to pre-train on bio-medical text with the goal of achieving improved performance in the medical NER task.

BIOBERT is first initialized with the weights obtained from the vanilla BERT model which was pre-trained on the traditional English corpus. On top of this, BIOBERT is further pre-trained on bio-medical text acquired from PubMed Abstracts and Pubmed Central Articles. So this results in a language representation model that has not only been pre-trained in the English language but also in the field of bio-medicine. The tokenizer part of the BIOBERT uses the same word piece tokenizer as BERT to carry out preprocessing. BIOBERT is finally fine-tuned for identifying medical entities by adding a final layer that utilizes the representations learnt from its penultimate layer.

4. PubMedBERT

To tackle the issue of learning word representations thoroughly in the field of bio-medicine Microsoft came up with their own version of BERT. Unlike the BIOBERT model, the PubMed BERT is entirely pre-trained on bio-medical text from scratch. Microsoft wanted to prove that training solely on domain-specific data is much more beneficial than training on a general-domain text first. Since PubMed Bert has been specially trained on the bio-medical domain alone, Microsoft claims that PubMedBERT gives more priority to biomedical terms and considers them as "first-class citizens" instead of computing bandwidth and diverting attention to the irrelevant out-domain text. PubMedBERT is pre-trained entirely on the data obtained from PubMed abstracts and Pubmed Central articles. However, it was found that using an in-domain training approach will produce better results for bio-medical applications rather than following the mixed-domain approach.

Experimentation

Experimental Setup

The implementation of the Medical NER pipeline was adapted from the github repository(lcampillos, 2020) to perform the following experimentation. Four different Medical NER systems were constructed with each system consisting of a pre-trained version of the BERT-base cased model, Roberta base model, BIOBERT base cased model and the Pubmed BERT base uncased model respectively. The BERT and RoBERTa models were representative of a language model pre-trained on a general domain text, whereas BIOBERT and PubMed BERT were representatives of a language model trained on in-domain text(Bio-medical articles). These 2 classes of models each representative of general-domain and in-domain text can be considered to be categories 1 and 2 respectively. The Huggingface library was used to implement the transformers and the pipeline is made using PyTorch. The following experiments were conducted on Google Colab which uses 12gb of RAM and a Tesla K80 GPU.

Dataset Creation

The dataset used for fine-tuning the models was aggregated using 2 corpora. The first corpus is the NCBI dataset that contains the disease mentions. The NCBI dataset is then converted into IOB format. The second corpus is the BC5CDR dataset that contains disease and chemical entities. These entities are tagged in IOB format separately for the same text. These separately tagged files are then combined into one single dataset by applying conditional column intersection. Finally, NCBI and BC5CDR data are combined into a single file for easier training.

FineTuning & Evaluation

The dataset created from the NCBI and BC5CDR corpora was then fed to the pre-trained transformer model for fine-tuning. The training data was split into 2 parts out of which one part(10% of training data) was set aside for validation. For performing fine-tuning, Adam was used as the optimizer, the batch size was set to 32 and the learning rate to $3 * 10^{-5}$.

Each of these 4 models was fine-tuned on the same dataset under similar conditions for about 8 epochs before moving on to evaluation. For every epoch, the average training loss, validation loss and the F1-score on the validation data were calculated and used as the evaluation metrics.

The validation loss and training loss for each of the models across 8 epochs were plotted. Figures 3-6 illustrate the convergence of the loss curves for the 4 models. For the BERT model that is trained on the language corpus, we find that the training loss drops just below 0.2 in the 8th epoch whereas validation loss stays constant from the 4th epoch as 0.4. This shows that the model is not able to learn new representations even when the epochs are increased. The model is trying to get overfit to the training data. Roberta on the other hand shows a nice drop in the loss values both for training and validation data. The training curve saturates around epoch 6 and tries to stay constant thereon. There is a good drop in the validation loss till the 8th epoch. This shows that the model is finetuned well and it is able to learn the entities well. We record training and validation to be lesser than 0.1 in the final epoch. The training curve of BIOBERT is in contrast to the validation curve. The training curve has a decrease in loss over time but the validation curve has an increase in the loss. Even though the increase in loss is not very significant, it shows us that the model is not able to learn much past the 3rd epoch. The graph of Pubmed BERT is pretty similar to what we saw with BIOBERT. We find that the Validation loss starts to increase after the 2nd epoch whereas the training loss is decreasing. This again means that the model is not able to capture any new representations.

On the whole, we find that the loss values of the models that have been pretrained on the medical corpus have a lesser loss than the language models that are trained on the English corpus. Pubmed BERT produces the least loss values of all the models but this might indicate a possibility of overfitting. The comparison of F-scores lets us analyse the performance of these models in depth.

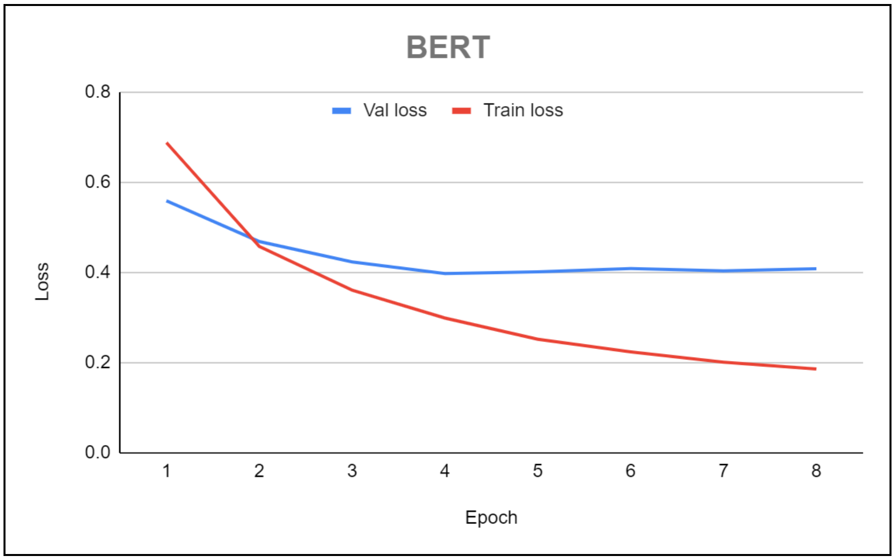

Figure 3 - Training and validation loss curve for BERT model

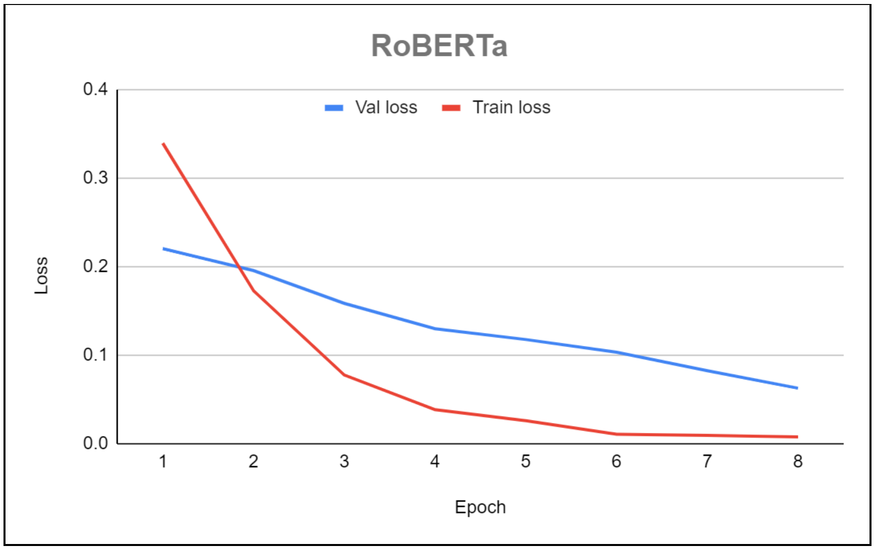

Figure 4 - Training and validation loss curve for RoBERTa model

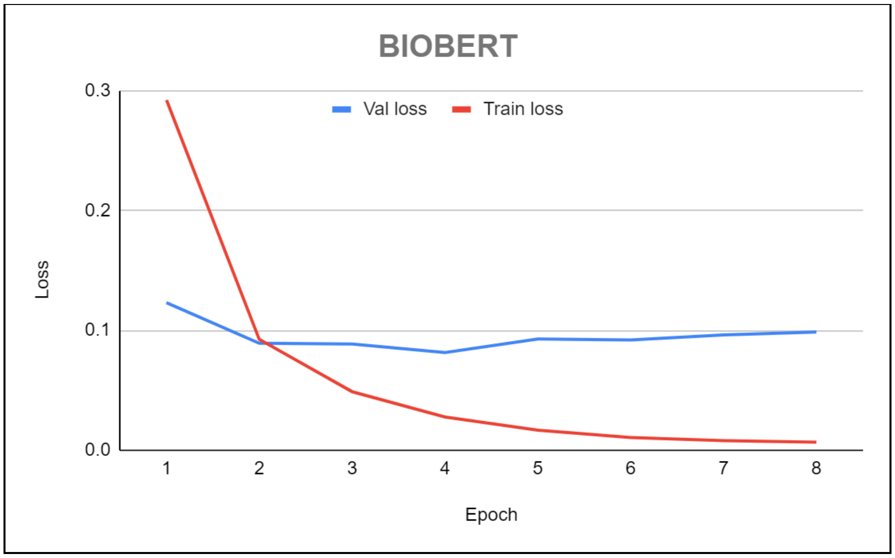

Figure 5 - Training and validation loss curve for BIOBERT model

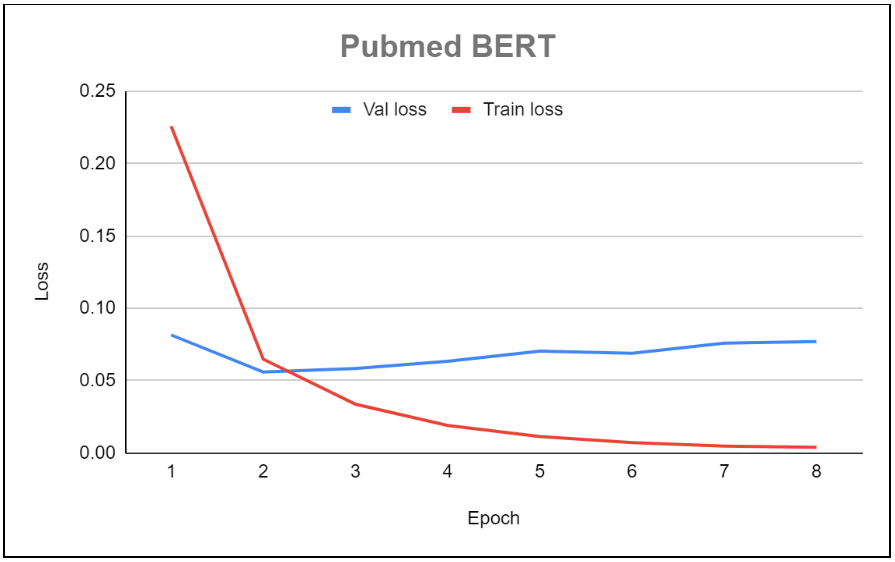

Figure 6 - Training and validation loss curve for Pubmed BERT model

Table 1 summarizes the performance of all the 4 models by displaying their corresponding F1-scores on validation data. Category 1 consists of models that were trained on the English corpus and category 2 consists of models on the bio-medical text. It can be observed that there is an overall performance improvement when going from category 1 to 2. In category 1, since RoBERTa is known to be a more optimized model, there's a slight improvement in the F-score when compared to BERT. But the F-score shoots up to 0.94 and 0.95 in category 2 indicating that the models in category 2 have in fact shown a greater improvement in performance when the language models were pretrained directly on the bio-medical domain.

	Model Name	**F1-score**
Category 1	BERT	0.8052
	RoBERTa	0.8504
Category 2	BIOBERT	0.9506
	PubMed BERT	0.9447

Table 1 - F Scores on validation data for 4 models

Statistical Analysis

From the results obtained and illustrated in table 1, it is clear that the BERT models in category 2 performed way better than the other language models in category 1. If we were to pick the best model out of the two categories, RoBERTa from category 1 and BIOBERT from category 2 would be the best performing models. To further justify the improvement in the performance between the 2 categories, Paired T-test(Kim, 2015) is performed and the statistical significance is studied. For performing T-test, the F-scores of BIOBERT and RoBERTa are taken over 5 different sets of validation data(existing validation data is split into 5 portions) and this is represented in table 2.

The null hypothesis (Ho) and the alternate hypothesis (Ho) are used in the T-test (H1). The purpose of this test is to determine which hypothesis is most likely to succeed. The null hypothesis states that all of the models are equivalent and similar in performance. The alternative hypothesis, which is the contradiction of the null hypothesis, argues that the models do not perform similarly, implying a significant difference in performance. The p-value for the experiment was found to be 8.55 * 10^{-3} after completing a T-test. For a 95% confidence level, the critical value would be 0.05. When the p-value is compared to the critical value, it becomes clear that the p-value is smaller than the critical value. As a result, we can conclude that the null hypothesis will be rejected, whereas the alternative hypothesis will be successful. Hence, the performance between the 2 models differs significantly.

Validation Data	**RoBERTa**	**BIOBERT**
Set 1	0.8489	0.9764
Set 2	0.8644	0.9518
Set 3	0.8584	0.9685
Set 4	0.8596	0.9667
Set 5	0.8495	0.9492

Table 2 - F-scores over 5 sets of validation data

From the experiments that were conducted, it can be clearly stated that the models in category 2, which have been pre-trained to understand the words in biological context perform way better than the models which been generally trained in the english corpus. This observation was also further proven by performing the paired T-test. Models in category 2 ensure an improvement in the results of the medical NER task under the same conditions without requiring any additional pre-processing or computation power.

Conclusion

The study was successful in implementing transformer-based language models to perform Medical NER on Bio-medical text. A proper NER dataset tagged with biomedical entities was constructed by fetching data from the existing biomedical corpora. BERT, RoBERTa, BIOBERT and PubMed BERT were the four models that were implemented in the pipeline to perform medical NER. The performance of all 4 models was evaluated and compared using validation data. Results showed that there was a huge improvement in the performance of the NER model when BIOBERT or PubMed BERT was used to detect entities on medical data. This proved that there's a huge difference in the performance in the language model depending on the domain specificity of pre-training. Upon further analysis of the results it was found that, though PubMedBERT had lesser loss values, the F-score of BIOBERT was slightly higher.

The study was worthwhile because we know now that the domain in which the language model has been pre-trained, plays a huge role in the performance of the model for any NLP task. Any transformer-based language model can be fine-tuned for a particular NLP task if it has been pre-trained in the appropriate domain. The task of Medical NER could be further enhanced by creating a system that is able to detect more medical entities like 'microbe' that cause the disease, 'drugs' and 'dosages' administered to the patient and so on. This would make the process of extracting information from medical transcriptions easier, but it would require a sufficient amount of dataset that has been annotated with the appropriate entities for pre-training.

References

1. Chiu C-C, Tripathi A, Chou K, Co C, Jaitly N, Jaunzeikare D, Kannan A, Nguyen P, Sak H, Sankar A, Tansuwan J, Wan N, Wu Y, Zhang X (2017) Speech recognition for medical conversations. https://doi.org/10.48550/arXiv.1711.07274

2. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient Estimation of Word Representations in Vector Space. https://doi.org/10.48550/arXiv.1301.3781

3. Bojanowski P, Grave E, Joulin A, Mikolov T (2016) Enriching Word Vectors with Subword Information. https://doi.org/10.48550/arXiv.1607.04606

4. Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805

5. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/arXiv.1907.11692

6. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2019) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz682

7. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H (2022) Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Transactions on Computing for Healthcare 3:1–23. https://doi.org/10.1145/3458754

8. Missen, Malik Muhammad Saad & Naeem, Aqsa & Asmat, Hina & Salamat, Nadeem & Akhtar, Nadeem & Coustaty, Mickaël & Prasath, Surya. (2021). Improving seller–customer communication process using word embeddings. Journal of Ambient Intelligence and Humanized Computing. 12. 10.1007/s12652-020-02323-1.

9. Nayel HA, Shashirekha HL, Shindo H, Matsumoto Y (2019) Improving Multi-Word Entity Recognition for Biomedical Texts. https://doi.org/10.48550/ARXIV.1908.05691

10. Perera, N., Dehmer, M., & Emmert-Streib, F. (2020). Named Entity Recognition and Relation Detection for Biomedical Information Extraction. Frontiers in cell and developmental biology, 8, 673. https://doi.org/10.3389/fcell.2020.00673

11. Peng Y, Yan S, Lu Z (2019) Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In: Proceedings of the 18th BioNLP Workshop and Shared Task. Association for Computational Linguistics, Florence, Italy, pp 58–65

12. Ramachandran, R., Arutchelvan, K. Named entity recognition on bio-medical literature documents using hybrid based approach. J Ambient Intell Human Comput (2021). https://doi.org/10.1007/s12652-021-03078-z

13. Doğan, R. I., Leaman, R., & Lu, Z. (2014). NCBI disease corpus: a resource for disease name recognition and concept normalization. Journal of biomedical informatics, 47, 1–10. https://doi.org/10.1016/j.jbi.2013.12.006

14. Li, Jiao & Sun, Yueping & Johnson, Robin & Sciaky, Daniela & Wei, Chih-Hsuan & Leaman, Robert & Davis, Allan Peter & Mattingly, Carolyn & Wiegers, Thomas & lu, Zhiyong. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database. 2016. baw068. 10.1093/database/baw068.

15. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention Is All You Need. https://doi.org/10.48550/ARXIV.1706.03762

16. Kim, T. (2015). T test as a parametric statistic. Korean Journal of Anesthesiology.

17. Icampillos, 2020, https://github.com/lcampillos/Medical-NER/blob/master/bert_ner.ipynb