SCORING PATIENT NOTE TAKING USING NLP

J COMPONENT PROJECT REPORT

Winter 2021-2022

Submitted by

Ritika Singh (19BCI0166)

Prajjwal Gupta (19BCI0171)

Saurabh Singh (19BCI0184)

in partial fulfilment for the award of the degree of

B. Tech

in

Computer Science and Engineering

and Information Security

Vellore-632014, Tamil Nadu, India

School of Computer Science and Engineering

April 2022

CONTENTS

TITLE: SCORING PATIENT/ NOTE TAKING USING NLP

Abstract

Whenever we go to a doctor or physician we have seen that they take notes of our symptoms, complaints, and our history to determine an accurate diagnosis. These doctors have various years of experience in note-taking and analyzing them. Assessing the notes taken by doctors requires feedback from other doctors when they are practicing this. Every medical student has to go through an exam conducted by NBME which requires them to write patient notes. The process of assessing the notes for every candidate manually is very time-consuming as well as labor intensive for trained physicians. With the advancements in Natural Language Processing, the manual job of a physician to analyze every patient's notes thoroughly to ensure a proper diagnosis of the symptoms, complaints of the patients, and their medical history can be improved. So, we propose a methodology for the National Board of Medical Examiners (NBME), which assesses the skills of writing patient notes for the Medical Licensing Examination. Using NLP, the task of identifying clinical concepts in patients' notes following the exam rubric will be done.

Almost 90% of the 2.5 quintillion bytes of data that is being produced each day is unlabelled and unuseful because there is no efficient way to use the data automatically. One of the main tasks is to design a technique which can classify and understand the data in a better way. This is a real problem when it comes to analyzing the patient's notes and clinical records of these candidates manually by the trained physicians. This requires significant time along with human and financial resources.

Using NLP models like BERT, ALBERTA, DEBERTA, and ROBERTA we will be showing the result of the input given by the trained physicians to analyze the patient notes for all the candidates.

Introduction

Patient notes is one of the areas in medical where the large amount of data get wasted daily. One of the aim is to study the data and visualize it to understand the quality and use of these large amount of data. These data can be used to classify certain diagnosis, can help in identifying medicines, diseases based on genetic history etc.

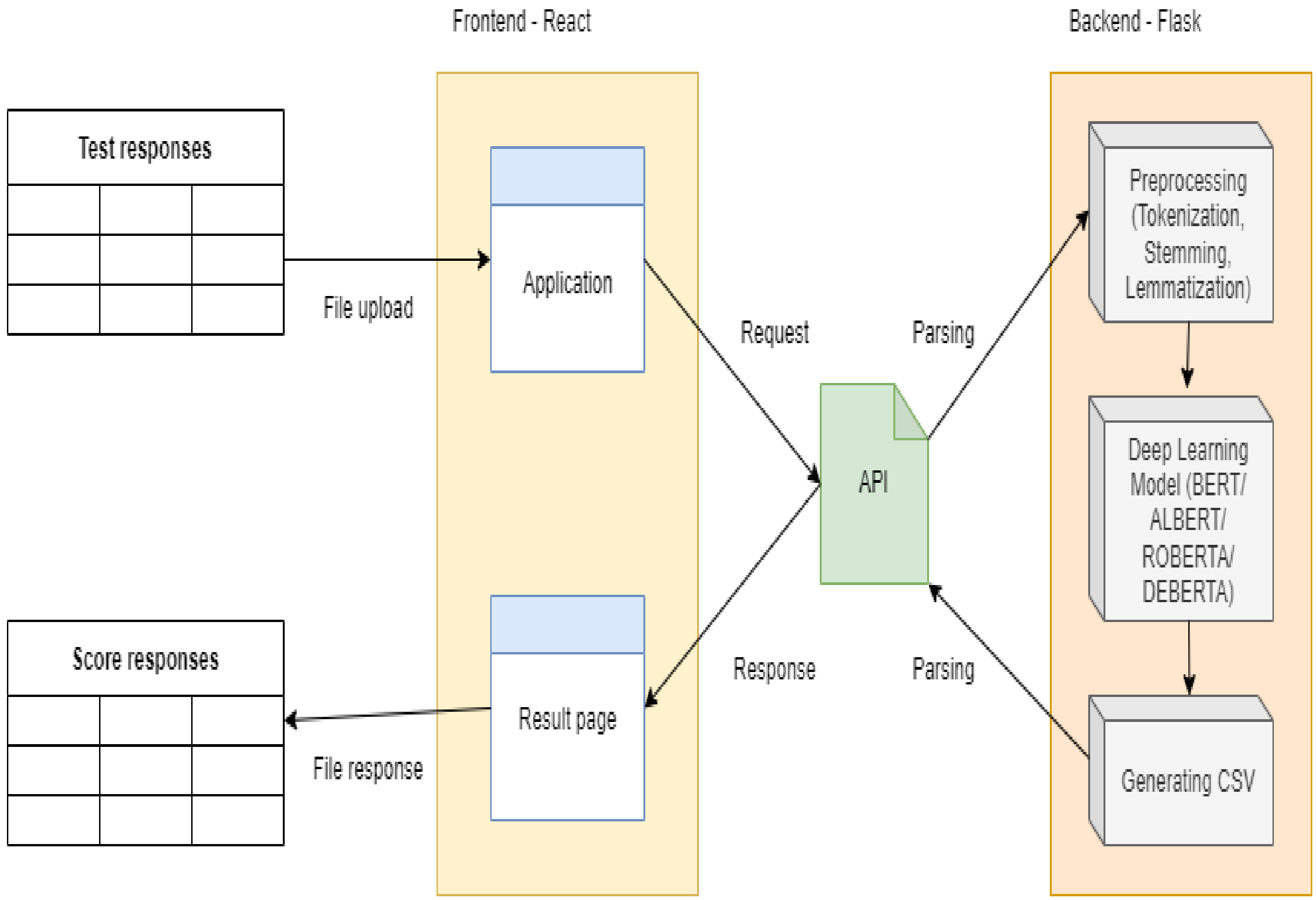
NBME take this exam where students attempt to write patient notes by real interviewing a person or from a video. Then, afterwards the notes of every candidate is sent to other physicians who are experienced in this to score them which takes a lot of time and resources.

The objective of our proposed methodology is to make the manual task of trained physicians to analyze all the candidates notes to correctly map the features or diseases with the patients symptoms, problems and medical history using NLP models.

Some of the statements like "quitting job" and "no longer interested on working " referring to the same feature/ problem have to be mapped correctly according to the exam rubrics. Another objective is combining multiple text segments or sentences having ambiguous meanings which basically correspond to a particular feature. Our aim is to create an automated approach for mapping clinical concepts from an exam rubric to various expressions of these concepts in clinical patient notes submitted by medical students.

We will be developing a full software solution in which the input will be in the form of a csv file uploaded by the trained physicians and the output will be the mapped feature and the particular locations of the part of the notes implying the annotations for scoring the candidates. As a result of this, medical practitioners will be able to explore the full potential of patient notes and analyze them to use it in better ways.

Architecture diagram



Background study

[1] Karami, Amir & Gangopadhyay, Aryya & Zhou, Bin & Kharrazi, Hadi. (2017). Fuzzy Approach Topic Modeling for Health and Medical Corpora. International Journal of Fuzzy Systems. 20. 10.1007/s40815-017-0327-9.

This paper proposes a method based on Fuzzy Latent Semantic Analysis method which is based on topic modeling which is a type of unsupervised classification.The problem statement that this paper aims to solve is to make the huge amount of medical records and documents useful by automating the whole process.
Their proposed method uses topic modeling combined with a fuzzy logic. FLSA addresses the issue of redundancy in the health and medical corpora and introduces a new method for estimating the number of subjects. The quantitative analysis reveal that FLSA outperforms and outperforms the most prevalent topic model, latent Dirichlet allocation.

Dataset:
5 publicly available health datasets have been used which includes the M-dataset, N-dataset and O-dataset.

FLSA has seven steps using Local Term Weighting, Global
Term Weighting , and Fuzzy Clustering. Various methods for performance analysis which were used include F-measure, MCC etc. Their new approach is tested well with both discrete and continuous data and can estimate the number of topics efficiently. Their future work includes applying this fuzzy approach to social media data.

[2] Ding, Liangping, Zhang, Zhixiong, Liu, Huan, Li, Jie and Yu, Gaihong. "Automatic Keyphrase Extraction from Scientific Chinese Medical Abstracts Based on Character-Level Sequence Labeling" Journal of Data and Information Science, vol.6, no.3, 2021, pp.35-57. https://doi.org/10.2478/jdis-2021-0013

To find the main points and extract important and topical phrases of text Automatic Keyphrase Extraction is extremely helpful. In this paper the author proposes an autonomous key extraction model for Chinese Scientific research to find Medical abstracts that combines the benefits of sequence labeling formulation with a pretrained language model.

Dataset: https://github.com/possible1402/nlp_chinese_corpus
Author used data from the Chinese Science Citation Database to created a large-scale dataset in the medical sector, with 100,000 abstracts as the training set, 6,000 abstracts as the development

set, and 3,094 abstracts as the test set. The dataset was labeled using Inside–Outside–Beginning tagging scheme.

To distinguish between English and Chinese, we employed Unicode Coding, which treats each English word as an elementary unit and each Chinese character as an elementary unit.

Challenges faced for Chinese Keyphrase Extraction:
1- Lack of publicly available dataset
2- Relying on word segmentation tools
3- BERT can only accept input with a maximum length of 512 characters. As a result of this limitation, some source input text will be abbreviated, posing the risk that the model will incorrectly forecast a single letter as a key. Tagging scheme: side–Outside–Beginning

Contributions:

1- Fine tuning the Bert parameters for the large-scale keyphrase extraction dataset.
2- This approach is independent of Chinese tokenizer.
3- Different experiments are designed using both supervised(CRF) and unsupervised(TF-IDF) models such as different ML models and BERT model to compare word level and character level sequence learning approaches.
4- Automatic key extraction was discussed here with experiments as a character-level sequence labeling problem rather than a word-level sequence labeling task, fine-tuned our key extraction model on scientific Chinese medical papers using the pretrained language model BERT.

[3] Rasmy, L., Xiang, Y., Xie, Z. et al. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digit. Med. 4, 86 (2021). https://doi.org/10.1038/s41746-021-00455-y

BERT models show a very high accuracy in this problem domain, so the authors of this paper propose an alternative model which they named MED-BERT which combines BERT and its derivatives which are known to be excellent with structured EHR.
Performance metrics and experiments show that the accuracy increases substantially when compared to earlier depp learning based models.

Dataset which was used was the EHR dataset on large records. BERT framework is used because it is known to be highly accurate with EHR data. They mainly extracted their data from Cerner Health Facts and Truven Health MarketScan

The main contributions of the authors of this paper are-
1)They did a demonstration of the BERT on EHR dataset with a high accuracy difference visible.
2)They observe the increased performance on various datasets, different sample sizes of data etc.
3)Identifies contextual semantics accurately among all kinds of EHR data.
4)Implemented this model with a user interface to check the dependency semantics.

[4] Wu H, Ji J, Tian H, Chen Y, Ge W, Zhang H, Yu F, Zou J, Nakamura M, Liao J
Chinese-Named Entity Recognition From Adverse Drug Event Records: Radical Embedding-Combined Dynamic Embedding–Based BERT in a Bidirectional Long Short-term Conditional Random Field (Bi-LSTM-CRF) Model
JMIR Med Inform 2021;9(12):e26407
URL: https://medinform.jmir.org/2021/12/e26407
DOI: 10.2196/26407

A large number of Adverse drug events are increasingly reported year by year in Electronic Medical Records which are source of infor for Adverse drug reactions. In this paper the author discusses how to extract ADR information from EHRs and medical reports.
BERT-Bi-LSTM-CRF-Radical model was used here which takes radical features and token features as input and appropriately identified the sentence's target entities
Experiments were designed between Man-Machine approaches to find the accuracy and effectiveness.

Dataset : ADEs recorded by the Drum Tower Hospital from 2016 to 2019. The author personally annotated 24,890 examples from Chinese ADR reports' free-text part.

BBC Radical Method consists of:

1. Bidirectional Encoder Representations from Transformers (BERT)
2. Bidirectional long short-term memory (bi-LSTM)
3. Conditional random field (CRF)

BERT and Bi-LSTM-CRF were two new named entity recognition (NER) models that integrated these token and radical features.

The proposed model performed well in extracting ADR-related data from ADE reports, and the results show that utilizing our method to extract ADR-related data will improve the quality of ADR reports and postmarketing drug safety reviews.
The application of NER in NLP technology may extract the target entity from free text automatically, and the extracted information can then be employed in statistical analysis, such as knowledge base–building tasks.
This work used a domain-specific NER task in Chinese ADE records, which could help with ADR evaluation and postmarketing medication safety evaluation.

[5] Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition Yun He1 , Ziwei Zhu1 , Yin Zhang1 , Qin Chen2 , James Caverlee1 https://arxiv.org/abs/2010.03746

There are various information related to health data which can be very useful when it comes to analyzing important features related to diseases and treatment. There are various applications to this problem statement which includes medical question answering, named entity recognition etc. The authors proposed a model in which they integrate the BERT model with pre-trained disease knowledge which again improves the accuracy.
All the various derivatives of BERT like BlueBERT, ALBERT, SciBERT, ROBERTA were compared and contrasted and the results show increased accuracy in all cases.

They train the BERT model and then add them with disease knowledge to infer the disease and symptoms and train it then.
Given a paragraph, BERT can get the title for the paragraph and the sub articles.

They test this model with 3 tasks which included medical question answering for which MEDIQA and TRECQA datasets are used, medical language inference for which MEDNLI dataset is used and disease name recognition fro which BC5CDR dataset is used.

[6] Xiong Y, Chen S, Chen Q, Yan J, Tang B
Using Character-Level and Entity-Level Representations to Enhance Bidirectional Encoder Representation From Transformers-Based Clinical Semantic Textual Similarity Model: ClinicalSTS Modeling Study
JMIR Med Inform 2020;8(12):e23357
URL: https://medinform.jmir.org/2020/12/e23357
DOI: 10.2196/23357

Electronic health record (EHR) systems are widely utilized in hospitals around the world for storing, sharing, and exchanging health information. The author in this paper proposes a system based on BERT and two other types of representation which re character level, sentence level and entity level module.

1- Tackling the out-of-vocabulary (OOV) problem in natural language processing (NLP) using character-level representation based on CNN.
2- Model clinical entity information in clinical text snippets using entity-level representation. Uses two methods

A- Entity 1: First encodes entities in a text snippet by the corresponding entity label sequence
B - Entity 2: It encodes entities with their representation on Mesh(knowledge graph) in medical domain.
3- Sentence level Representation module to encode clinical text snippet pairs using a pretrained language model bidirectional encoder representation from transformers (BERT).

Finally to aggregate all of them modules, we concatenate them together. Then the author used Multi layer perceptron (MLP) to predict the STS score. We found that domain-specific pretrained BERT models are better than general pretrained BERT models.

Preprocessing:
1- Converting clinical text snippets into lowercase
2- Tokenize clinical text snippets using special symbols

Corpus : 2019 n2c2/OHNLP track on ClinicalSTS. For the ClinicalSTS task, the n2c2/OHNLP organizers manually annotated a total of 2055 clinical text snippet pairs by two medical professionals, with 1643 pairs serving as the training set and 412 serving as the test set.

[7] Yao L, Jin Z, Mao C, Zhang Y, Luo Y. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. *J Am Med Inform Assoc*. 2019;26(12):1632-1636. doi:10.1093/jamia/ocz164

This paper deals with the classification of medicines into disease categories. After research they found that BERT achieves higher accuracy that other methods. Further before training the model, fine tuning of the BERT model is done using an unlabelled medical dataset.Their models achieved an accuracy of around 90%, when tested with clinical data.

They also propose a new methodology in which they combine the BERT model with a clinical corpora. The input is in the form of clinical text data, which after various preprocessing stages like segmentation achieves a higher accuracy than any other model.

They did a study of deep learning models for TCM clinical text classification which is also compared to traditional machine learning approaches.They propose a fine-tuning pretrained deep language models with large-scale unlabeled TCM clinical data, which currently yields the best results.

Their methodology is divided into 3 steps which includes the pre-training of the BERT model, followed by step 2, which is using these weights for initialization and step3 uses the weights after fine tuning for initialization.

[8] Li F, Jin Y, Liu W, Rawat B, Cai P, Yu H Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)–Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study

JMIR Med Inform 2019;7(3):e14830

URL: https://medinform.jmir.org/2019/3/e14830

DOI: 10.2196/14830

BERT model has gained various success in the area of question answering and named entity recognition. In this paper the author proposed BERT model in the area of Named Entity Normalization. The aim is to check the effectiveness of BERT model on biomedical and clinical entity normalization.

Deep representation-learning approaches learn word representations from a large amount of unannotated data, which makes them more generalizable than models trained on small amounts of annotated data.

DATASET : 1.5 million EHR notes (EhrBERT) available on Github. We used the MADE (Medication, Indication, and Adverse Drug Events) corpus [which was created in response to the MADE 1.0 challenge. The corpus contains 1089 EHR notes, which were separated into 876 training notes and 213 testing notes.

Our data consisted of 1.5 million notes from electronic health records (EHRs) that had not been tagged.

BioBERT was fine-tuned on this big set of unlabeled EHR notes first.As a result, we developed a BERT-based model based on 1.5 million electronic health record notes (EhrBERT).

Then different BERT models were used on clinical and bio records and compared to find the effectiveness using MetaMap and disease name normalization.

Methodology:

1. First, the authors' off-the-shelf models for the MetaMap and DNorm were used to run them on test sets once. We believe these models have been fine-tuned to get the greatest results as strong baselines.
2. Second, the BERT, BioBERT, and EhrBERT studies were repeated three times. We used a new random seed to initialise the model for each run. The model was ran on the test set after training to achieve precision, recall, and F1.
3. Finally, based on the outcomes of these runs, the t test was used to see if the performances of two models were significantly different.

[9] Faiza Khan Khattak, Serena Jeblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, Frank Rudzicz,

A survey of word embeddings for clinical text,Journal of Biomedical Informatics,Volume 100,Supplement,2019,100057,ISSN 1532-0464,https://doi.org/10.1016/j.yjbinx.2019.100057.

(https://www.sciencedirect.com/science/article/pii/S2590177X19300563)

This paper discusses the different types of clinical corpora, word representations, pre-trained clinical word-vector embeddings, evaluation, applications and limitations of each.

They use a popular method, word embeddings for representing the semantics of the text data and focus on heo it can be applied to represent text data. Their described method is-

1)Choose the training data.

2)Pre-process the data

3)Choosing the algorithm for word embeddings.

4)Train these word embeddings on BERT model.

5)Apply the trained model to the target data.

They proposed a method to represent the medical data including the semantics of the data into word embeddings.

Recognizing these word embeddings from this medical data includes various steps like data preprocessing, model fitting and evaluation.Preprocessing includes various steps like tokenization, removing stop words, stemming and lemmatization etc.

After pre processing training of the neural network is done followed by tuning the hyperparameters**.** The **DATASET** used was the medical data from PubMed abstracts.

[10] Z. Dai, X. Wang, P. Ni, Y. Li, G. Li and X. Bai, "Named Entity Recognition Using BERT BiLSTM CRF for Chinese Electronic Health Records," 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2019, pp. 1-5, doi: 10.1109/CISP-BMEI48845.2019.8965823.

This paper presents the neural network approaches to Natural Language Processing for Clinical Health Records using Named Entity Recognition. It compares various models like BiLSTM and 2 pre-trained models including word2vec and BERT. The results show that the BERT model showed the highest accuracy in extracting the valuable medical information.

The main aim is to identify and classify the medical records which assigns a label to every part of the text.

Furthermore, this work uses the BIO tagging system to identify entities, with a token being labeled as B-label if it is the start of a named entity. If the token is inside a named object but not the first, use I-label, otherwise, it's an O-label.

The phases in the named entity recognition process utilizing the BiLSTM plus CRF model are as follows:

1) In the model's first layer, the look-up layer, each word in the sentence is mapped in the vector using the pre-trained word embedding.

2) Automatically extract the sentence features using the model's second layer, the Bidirectional Long Short-Term Memory (BiLSTM) layer.

3) Sequence labeling is performed between phrases in the model's third layer,which is the Conditional Random Field (CRF) layer.

4) The [previous steps are repeated until all of the data has been labeled.

[11] Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, Jian Wang, An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition, *Bioinformatics*, Volume 34, Issue 8, 15 April 2018, Pages 1381–1388, https://doi.org/10.1093/bioinformatics/btx761

This paper proposes a neural network methodology which combines attention based BiLSTM and Conditional random field layer for chemical Named Entity Recognition.

It uses the information obtained using attention mechanism to enforce tagging consistency for multiple tokens in the same document. Features used in traditional NER methods like POS tagging are used along with these neural network models which have shown an increase in the accuracy.

Their work focuses on feature engineering, neural network model ,and the tagging inconsistency of the NER.

The main contributions of their work includes-

1)Capturing similar entities and constantly increasing the accuracy with each round.

2)NER methods liek POS tagging, dictionary features are used.

Using above these 2 contributions, the accuracy of this model has increased to more than 90%.

[12] Agrawal, Ankit & Tripathi, Sarsij & Vardhan, Manu & Sihag, Vikas & Choudhary, Gaurav & Dragoni, Nicola. (2022). BERT-Based Transfer-Learning Approach for Nested Named-Entity Recognition Using Joint Labeling. Applied Sciences. 12. 976;. 10.3390/app12030976.

Most of the current state-of-the-art models deal with the problem of embedded/nested entity recognition with very complex neural network architectures. The author proposed to solve the problem of nested named-entity recognition using the transfer-learning approach. Different variants of fine-tuned, pretrained, BERT-based language models were used for the problem.

Two different datasets were used for four and two levels of annotations.The conclusion author got was that this fine tuned Bert model was significantly better than other models like Bi-LSTM-CRF.

They performed various experiments on the dataset to arrive at their approach. Various metrics were used to check the performance measure of the proposed methodology.

The performance of their model was also compared and contrasted with other models like Bi-LSTM. The best BERT model achieved an accuracy of about 90% with the **JNLPBA DATASET**.

[13] DocBERT: BERT for Document Classification
Ashutosh Adhikari,   Achyudh Ram,1   Raphael Tang,
Equal contribution.
Jimmy Lin
David R. Cheriton School of Computer Science
University of Waterloo

The author in this paper discusses fine tuned BERT model for document classification. It contains few nuances: First documents frequently have multiple labels over dozens of classes, which is unusual for the tasks investigated by BERT. Second, when it comes to document classification, modeling syntactic structure is less important than it is for other challenges like natural language inference and sentiment classification.
Starting with BERT base and BERT large models and to adapt BERT for doc classification introducing a fully connected layer over final hidden state.
Optimizing the entire model end to end during fine tuning with additional softmax parameters to minimize the cross entropy and binary cross entropy.
BERT large achieves state-of-the-art results on all four datasets, followed by BERT base. The author here varied different parameters to check the efficiency of different values on 4 datasets given.

DATASETS:

- Reuters-21578 (Reuters; Apté et al., 1994)
- IMDB reviews
- arXiv Academic Paper dataset (AAPD; Yang et al., 2018)
- Yelp 2014 reviews

[14] BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision Chen Liang*, Yue Yu*, Haoming Jiang*, Siawpeng Er, Ruijia Wang, Tuo Zhao, Chao Zhang Georgia Institute of Technology, Atlanta, GA, USA {cliang73,yueyu,jianghm,ser8,rwang,tourzhao,chaozhang}@gatech.edu

The author here studied open doman NER problem under distant supervision and came to conclusion that it doesn,t provide better solution so came to a new framework BOND, which improves the prediction performance of NER models by leveraging the power of pre-trained language models (e.g., BERT and RoBERTa).

Two stage training algorithm is proposed:
1- We apply the pre-trained language model to NER tasks utilising remote labels, which improves recall and precision dramatically.
2- We abandon the remote labels in favour of a self-training technique to increase the model's performance.

Challenges:
The labels produced through remote supervision are frequently sloppy and incomplete. The open-domain datasets' remote labels have significantly lower precision and recall. This makes training accurate NER models extremely difficult.

BERT models with their variants such as ALBERTA, ROBERTA etc are essentially huge neural networks based on bi-directional transformer designs that are trained fully unsupervised utilising open-domain data. When applying pre-trained language models to downstream tasks, all that is required is a minor modification and adaptation of the model using efficient and scalable stochastic gradient-type techniques.

[15] Extracting Family History of Patients From Clinical Narratives: Exploring an End-to-End Solution With Deep Learning Models

This paper highlights the importance of family history in the diagnosis of patients. However, this type of data is largely unstructured and the paper highlighted the unavailability of efficient solutions for the same. The authors used a bidirectional encoder representation from transformers(BERT) based approach for the task of information retrieval and extraction from corpus of data. The authors also used a voting based ensembling method to improve the results obtained.

[16] Extracting Drug Names and Associated Attributes From Discharge Summaries: Text Mining Study

Dataset: 2018 n2c2 shared task data

This paper highlights the problems associated with the highly unstructured nature of language corpus in medical context. The authors highlight the limitations of the existing approaches a perform an extensive error analysis of different methods.

The authors propose a novel solution DrugEx which consists of consists of a named entity recognizer (NER) to identify drugs and associated attributes and a relation extraction (RE) method to identify the relations between them. For the NER task the authors explore various models like bidirectional long-short term memory with conditional random fields [BiLSTM-CRFs] and combined them with various embedding strategies like word embedding, character embedding [CE], and semantic-feature embedding.

The experiments showed that the best model (BiLSTM-CRFs with pretrained word embeddings [PWE] and CE) achieved lenient micro F-scores of 0.921 for NER, 0.927 for RE, and 0.855 for the end-to-end system.

[17] A Hybrid Model for Family History Information Identification and Relation Extraction: Development and Evaluation of an End-to-End Information Extraction System

Dataset: BioCreative/OHNLP 2018 corpus, Medical Information Mart for Intensive Care (MIMIC-III) clinical database, 2019 National Natural Language Processing Clinical Challenge (n2c2)

This paper highlights the unstructured nature and the complexities associated with the information extraction task related to family history in a medical corpus. The authors develop an end-to-end hybrid approach combining machine-learning and rule-based learning approaches. For entity recognition, they trained bidirectional long short-term memory deep learning models.

They created a voting ensemble that combined the predictions of all individual models and incorporated static word embeddings and context-dependent embeddings. For the relation extraction task, the authors trained 2 relation extraction models. The first model determined the living status of each family member. The second model identified observations associated with each family member.

[18] Named entity recognition for question answering

This paper highlights the common structure of question answering systems which usually contain a named entity recognizer (NER). The authors highlight that no formal assessment of the performance of NERs in the context of question answering has been done. The authors present a NER that aims at higher recall by allowing multiple entity labels to strings.

In their proposed approach, the NER is embedded in a question answering system and the overall QA system performance is compared to that of one with a traditional variation of the NER that only allows single entity labels.

Methodology:

1. Information flow

Sample interaction of the users can be described using the following steps:

Step 1: A patient is identified who is to be assessed. The features (symptoms and conditions) evident and present in the patient are noted down.

Step 2: The doctor interacts with the patient by asking them questions related to the conditions and making notes of the same.

Step 3: The doctor's notes along with the identified features in the patient are entered into an excel sheet (.csv file)

Step 4: The prepared file is uploaded on our web portal where it is processed and the trained models carry out their intended tasks.

Step 5: The final results returned by the model are displayed on the webpage.

2. NLP problem formulation

The given problem statement can be posed as a combination of test segmentation and question answering tasks.

Text segmentation: The process of splitting written text into meaningful components, such as words, sentences, or subjects, is known as text segmentation. Natural language processing refers to both the mental processes that people utilise while reading text and the artificial processes that are implemented in computers. In our project, the main output is the text segments in the doctor's notes which have identified particular features present in the patient.

Question Answering: Question answering is a major NLP challenge as well as a long-standing AI milestone. A user may ask a question in plain language and receive an immediate and concise response using QA technologies. The capacity to read a piece of literature and then answer questions about it is known as reading comprehension. Reading comprehension is challenging for computers because it necessitates a combination of natural language comprehension and global knowledge. In our project, the context is provided by the doctor's notes. The questions queried are the features present in the particular patient and the answer expected is the text segment in which the doctor has identified the query feature.

3. Proposed solution

In this project, we create a web portal for the intended task of question answering based on text segmentation where the file containing the doctor's note and patient features is uploaded.

The front end of the application is developed using HTML, CSS and JavaScript. The backend mainly involves an API hosted on the cloud. The trained model is stored in a .h5 file which is stored in a cloud container. When a request on the predicted route is initiated, the saved model is loaded and the predictions made are returned.

The major preprocessing involved was tokenization. We used the transformers library which contains the AutoTokenizer utility. The fields tokenized with truncation are: "feature_text" and "pn_history".

The data was also converted into lowercase form to remove any case related ambiguity.

The "-OR-" token was replaced by ";-" and the "-" token was replaced by " ".

4. Dataset

The dataset used is Clinical Patient notes by NBME (National Board of Medical Examiners). The dataset contains the following files:

è patient notes.csv - A file containing about 40,000 Patient Note history segments. Only a few of them have characteristics that have been annotated. On the notes without comments, you might want to use unsupervised learning approaches. The public version of this file does not include the patient notes from the test set.

· pn num - Each patient note's unique identification.

· case num - The patient note's unique identification for the clinical case it reflects.

· pn history - The encounter's text as written down by the test taker.

è features.csv - Each clinical case's feature (or important idea) rubric.

· feature num - Each feature's unique identification.

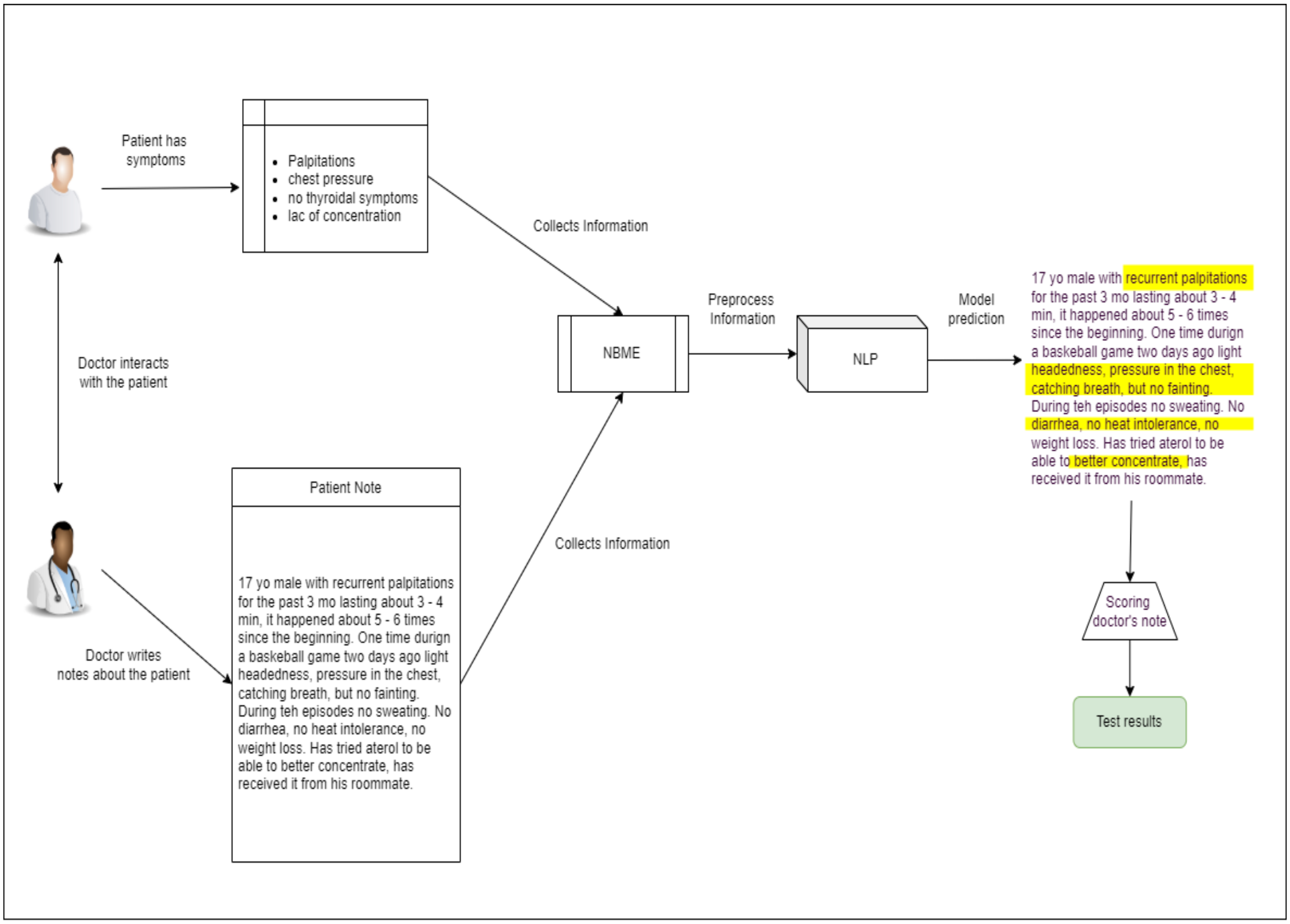
· case num - Each case's unique identity.

- feature text - The feature's description.

è train.csv - Feature annotations for 1000 of the patient notes, 100 for each of ten cases.

- id - Unique identifier for each patient note/feature pair.
- pn_num - The patient note is annotated in this row.
- feature_num - The feature annotated in this row.
- case_num - The case to which this patient note belongs.
- annotation - The text(s) within a patient note indicating a feature. A feature may be indicated multiple times within a single note.
- location - Character spans indicating the location of each annotation within the note. Multiple spans may be needed to represent an annotation, in which case the spans are delimited by a semicolon;

Proposed Model:



17 yo male with recurrent palpitations for the past 3 mo lasting about 3 - 4 min, it happened about 5 - 6 times since the beginning. One time durign a basketball game two days ago light headedness, pressure in the chest, catching breath, but no fainting. During teh episodes no sweating. No diarrhea, no heat intolerance, no weight loss. Has tried aterol to be able to better concentrate, has received it from his roommate.

17 yo male with recurrent palpitations for the past 3 mo lasting about 3 - 4 min, it happened about 5 - 6 times since the beginning. One time durign a basketball game two days ago light headedness, pressure in the chest, catching breath, but no fainting. During teh episodes no sweating. No diarrhea, no heat intolerance, no weight loss. Has tried aterol to be able to better concentrate, has received it from his roommate.

The proposed methodology is built around the Roberta model which is a derivative of BERT, a widely used and adopted model in NLP applications. Roberta utilizes a Robustly Optimized BERT Pretraining Approach which results in a significant decrease in training and convergence time and outperforms BERT on extrinsic evaluation on various tasks.

The authors used mainly 3 strategies to achieve improvement of ROBERTa over BERT:

- Removing the Next Sentence Prediction (NSP) goal: The model is trained to predict whether the observed document segments come from the same or other documents using an auxiliary Next Sentence Prediction (NSP) loss in the next sentence prediction. The authors tested multiple versions with and without NSP loss and found that eliminating the NSP loss matches or slightly improves downstream job performance.

- Masking pattern dynamically changing: In the BERT architecture, masking is done just once during data preparation, resulting in a single static mask. To avoid utilising a single static mask, the training data is replicated and masked 10 times throughout 40 epochs, each time using a different mask strategy, resulting in four epochs with the same mask. This method is in contrast to dynamic masking, which uses different masking each time data is input into the model.

- Training with larger batch sizes and longer sequences: BERT was originally trained for 1M steps and 256 sequences in a batch size of 256. The authors used 125 steps of 2K sequences and 31 steps of 8k sequences to train the model in this article. The big batches enhance perplexity on the masked language modelling objective as well as end-task accuracy, which offers two benefits. Large batches can also be parallelized more easily using distributed parallel training.

The architecture used in our project is based on transfer learning over a pretrained model provided by hugging face. Transfer learning is a machine learning technique in which a model created for one job is utilized as the basis for a model on a different task. We added 3 fully connected dense layers of 512, 512 and 1 unit over the top of the ROBERTa model with dropout layers in between. The loss criterion used was BCE (Binary Cross Entropy) with Logits Loss. The model was implemented, trained and inferred using PyTorch library.

The hyperparameters related to the preprocessing and training of the model are as follows:

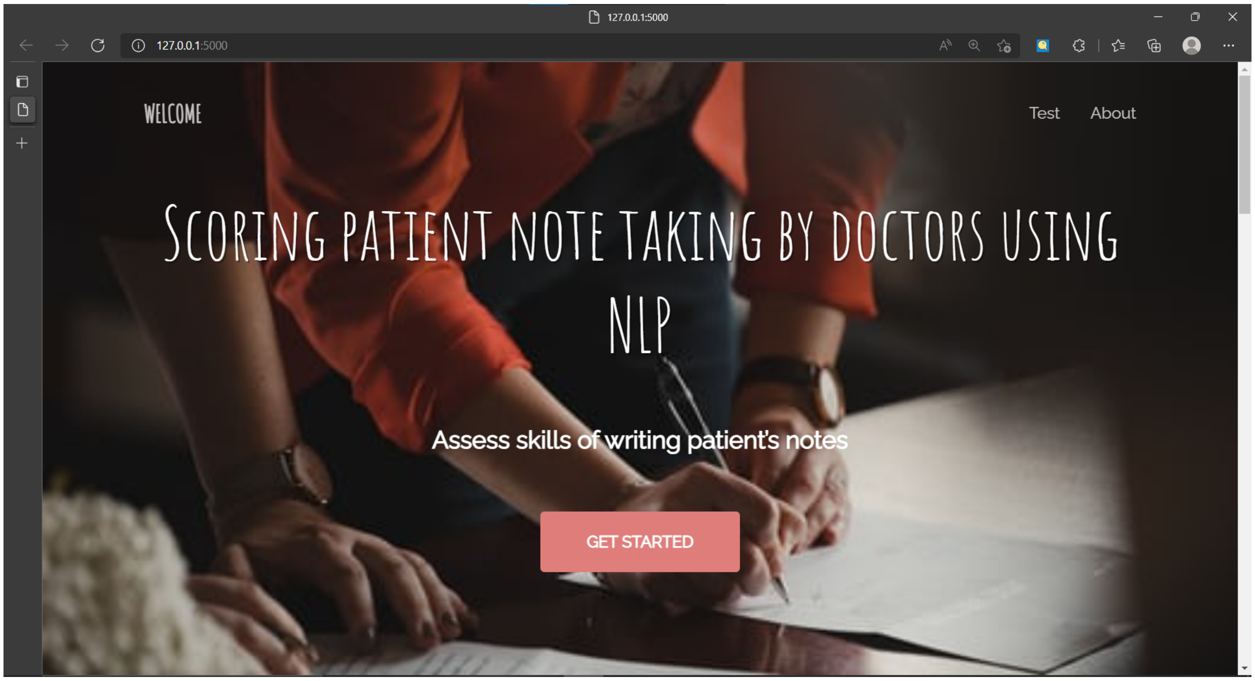
Hyperparameter	**Value**
Sequence max length	416
Padding	Max length
Dropout rate	0.2
Learning rate	1e-5
Batch size	8

Results and Discussions

WEB APPLICATION

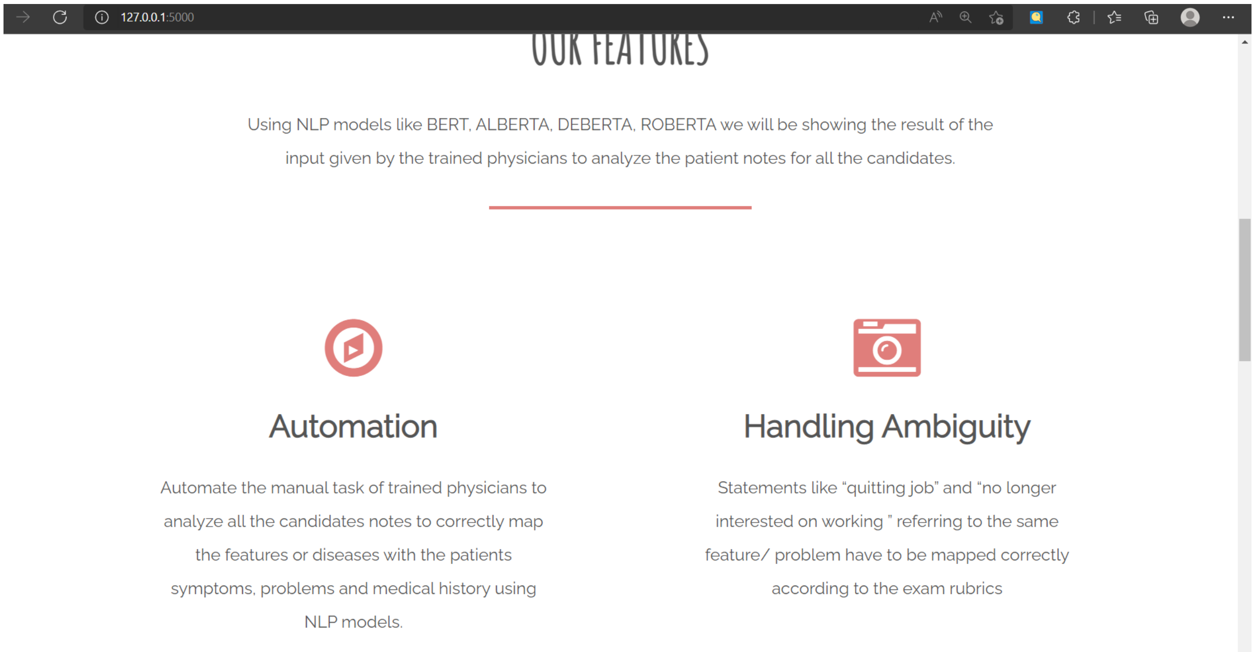
Frontend-

LANDING PAGE-It includes all the features and information of our website.



SCORING PATIENT NOTE TAKING BY DOCTORS USING NLP

Assess skills of writing patient's notes



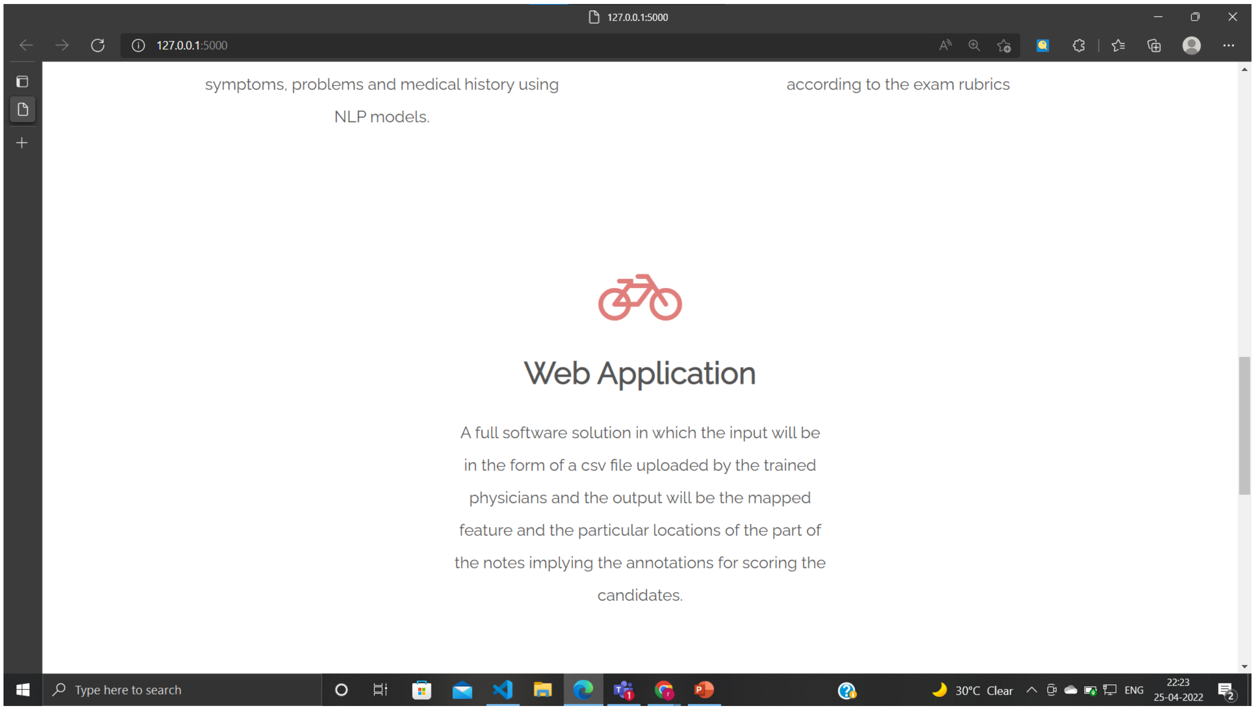
Using NLP models like BERT, ALBERTA, DEBERTA, ROBERTA we will be showing the result of the input given by the trained physicians to analyze the patient notes for all the candidates.

Automation

Automate the manual task of trained physicians to analyze all the candidates notes to correctly map the features or diseases with the patients symptoms, problems and medical history using NLP models.

Handling Ambiguity

Statements like "quitting job" and "no longer interested on working " referring to the same feature/ problem have to be mapped correctly according to the exam rubrics

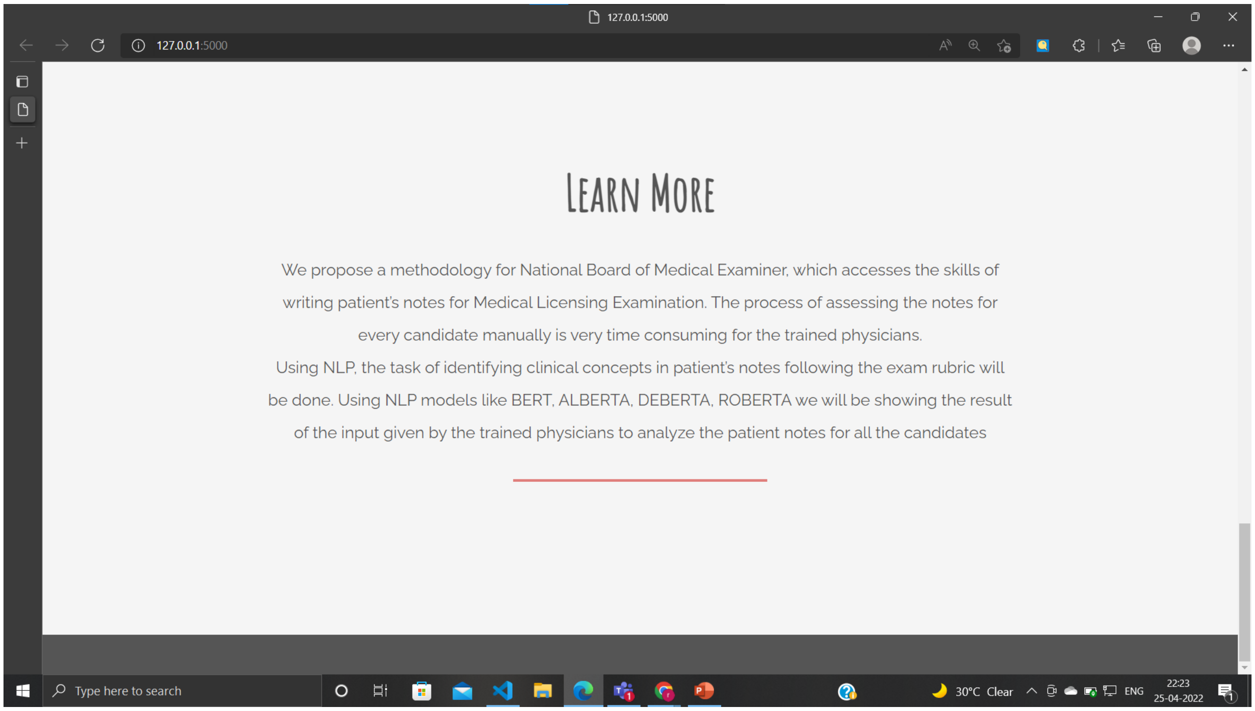


symptoms, problems and medical history using NLP models.

according to the exam rubrics

Web Application

A full software solution in which the input will be in the form of a csv file uploaded by the trained physicians and the output will be the mapped feature and the particular locations of the part of the notes implying the annotations for scoring the candidates.

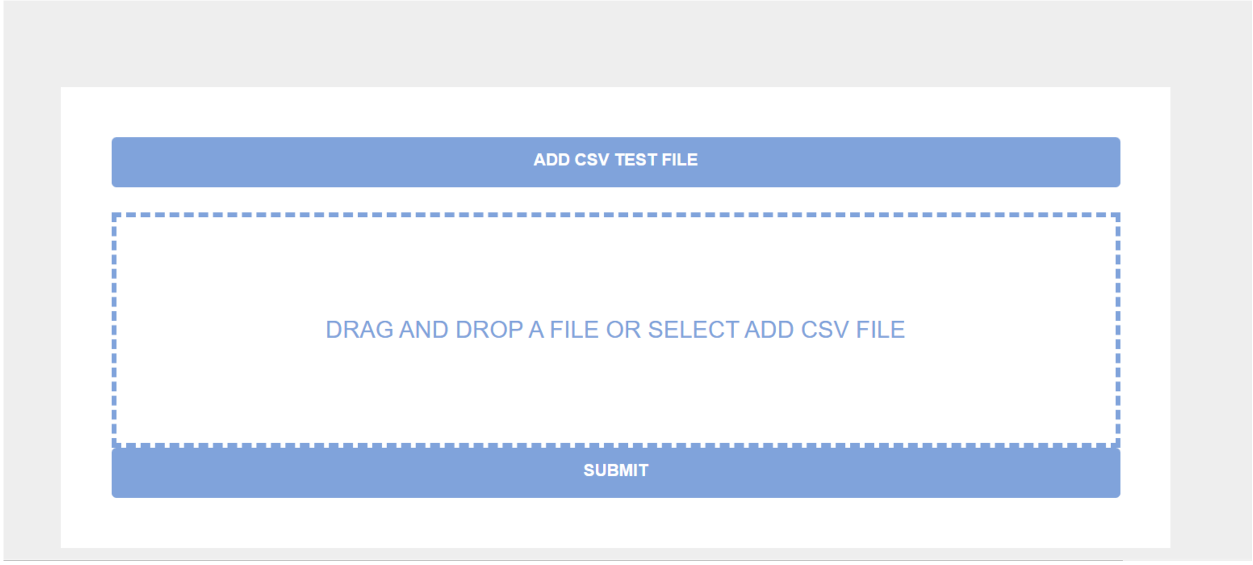


Learn More

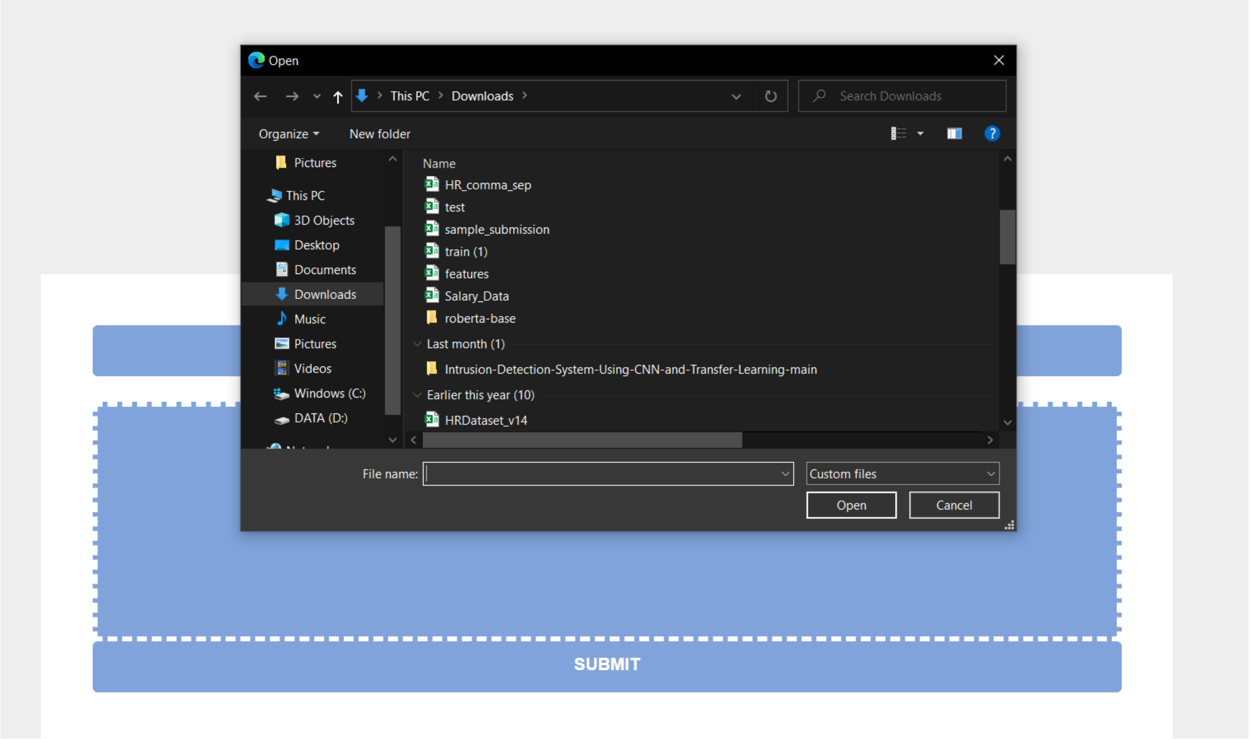
We propose a methodology for National Board of Medical Examiner, which accesses the skills of writing patient's notes for Medical Licensing Examination. The process of assessing the notes for every candidate manually is very time consuming for the trained physicians.
Using NLP, the task of identifying clinical concepts in patient's notes following the exam rubric will be done. Using NLP models like BERT, ALBERTA, DEBERTA, ROBERTA we will be showing the result of the input given by the trained physicians to analyze the patient notes for all the candidates

TESTING PAGE- (CSV FILE UPLOAD PAGE)

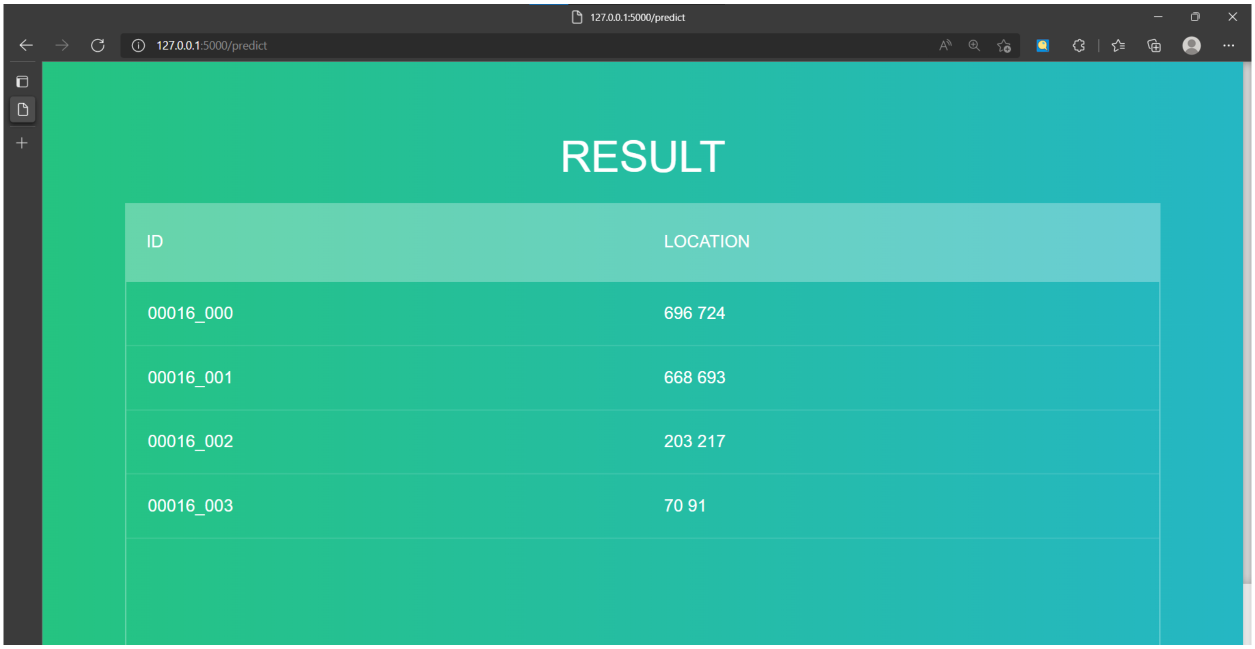
The trained physicians can upload the answers of the doctors taking the examinations from here to get automated results!



They have to choose a CSV file.



RESULT PAGE-



RESULT

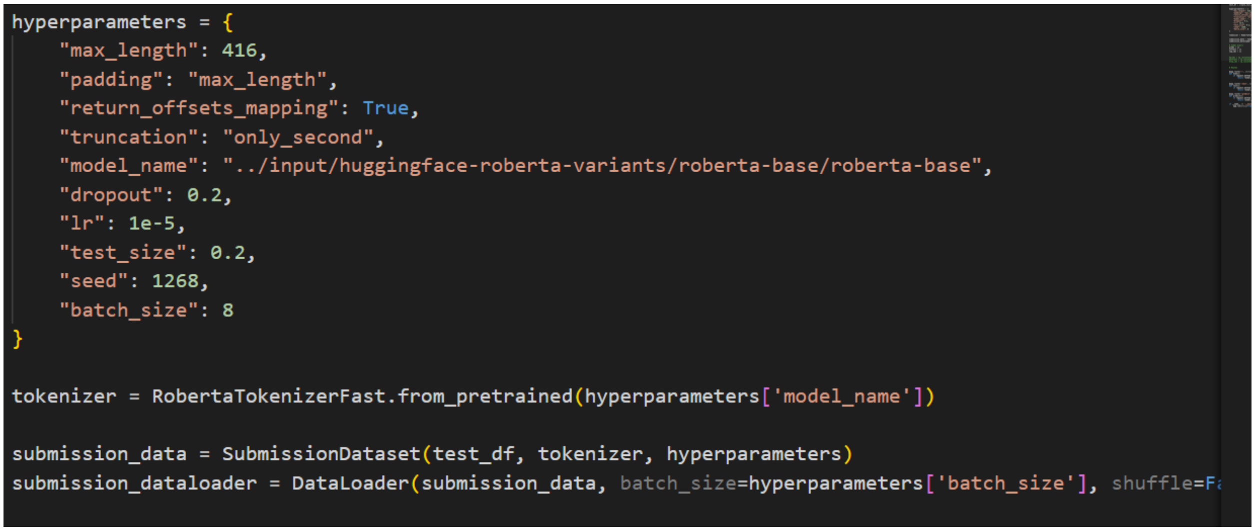
ID	LOCATION
00016_000	696 724
00016_001	668 693
00016_002	203 217
00016_003	70 91

Here, we can see that the locations of the annotations describing the symptoms of the patients in the patient notes taken by the examinees are generated which can be used to calculate the results of the examinations based on how many symptoms have been correctly identified.

For eg: for ID 00016_000 location 696 to 724 includes “dad with a recent heart attack”

SAMPLE CODE-

ROBERTA TOKENIZER

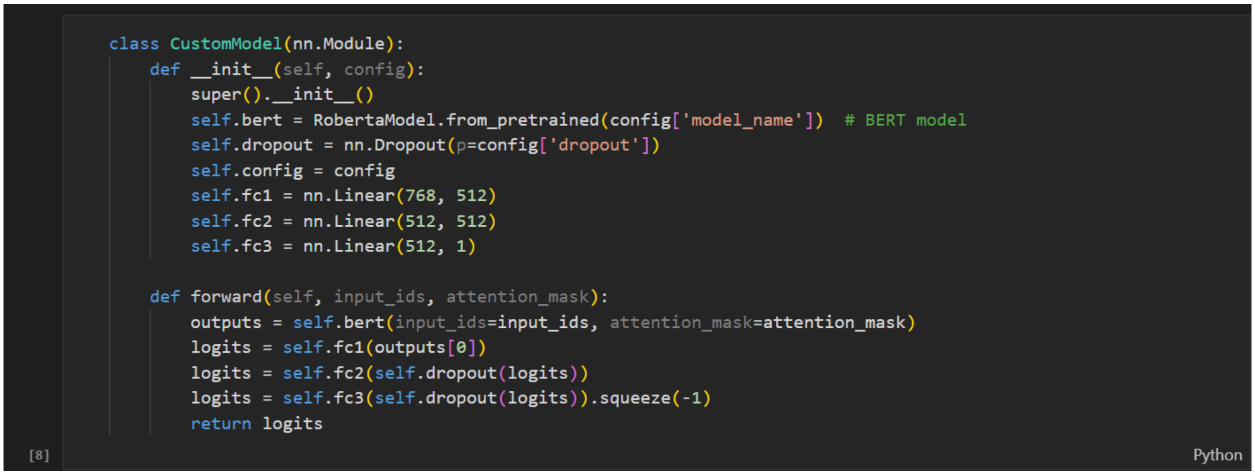


```
hyperparameters = {
    "max_length": 416,
    "padding": "max_length",
    "return_offsets_mapping": True,
    "truncation": "only_second",
    "model_name": "../input/huggingface-roberta-variants/roberta-base/roberta-base",
    "dropout": 0.2,
    "lr": 1e-5,
    "test_size": 0.2,
    "seed": 1268,
    "batch_size": 8
}

tokenizer = RobertaTokenizerFast.from_pretrained(hyperparameters['model_name'])

submission_data = SubmissionDataset(test_df, tokenizer, hyperparameters)
submission_dataloader = DataLoader(submission_data, batch_size=hyperparameters['batch_size'], shuffle=Fa
```

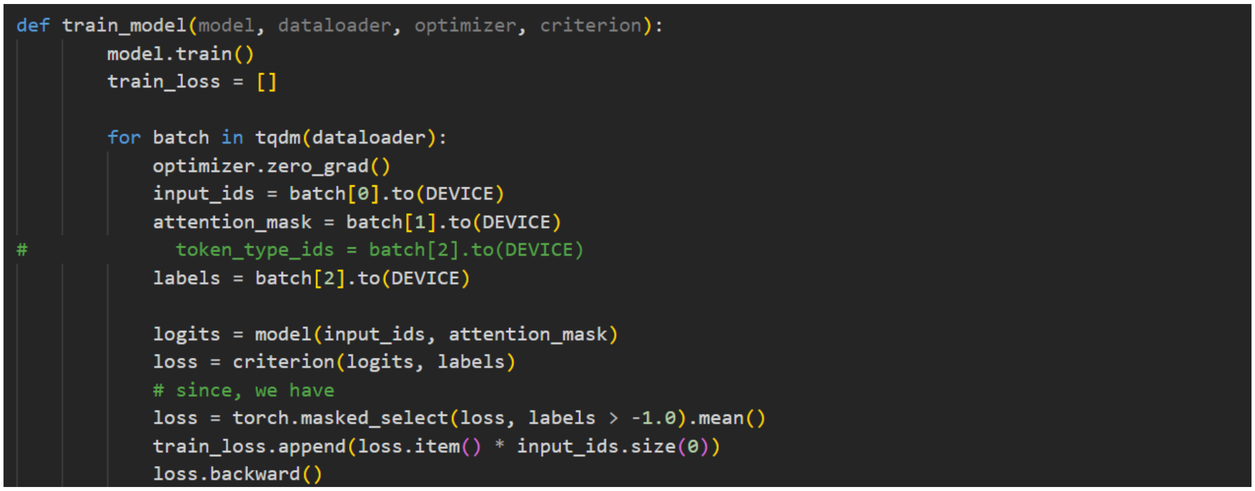
ROBERTA MODEL-



```python
class CustomModel(nn.Module):
    def __init__(self, config):
        super().__init__()
        self.bert = RobertaModel.from_pretrained(config['model_name'])  # BERT model
        self.dropout = nn.Dropout(p=config['dropout'])
        self.config = config
        self.fc1 = nn.Linear(768, 512)
        self.fc2 = nn.Linear(512, 512)
        self.fc3 = nn.Linear(512, 1)

    def forward(self, input_ids, attention_mask):
        outputs = self.bert(input_ids=input_ids, attention_mask=attention_mask)
        logits = self.fc1(outputs[0])
        logits = self.fc2(self.dropout(logits))
        logits = self.fc3(self.dropout(logits)).squeeze(-1)
        return logits
```

MODEL TRAINING-

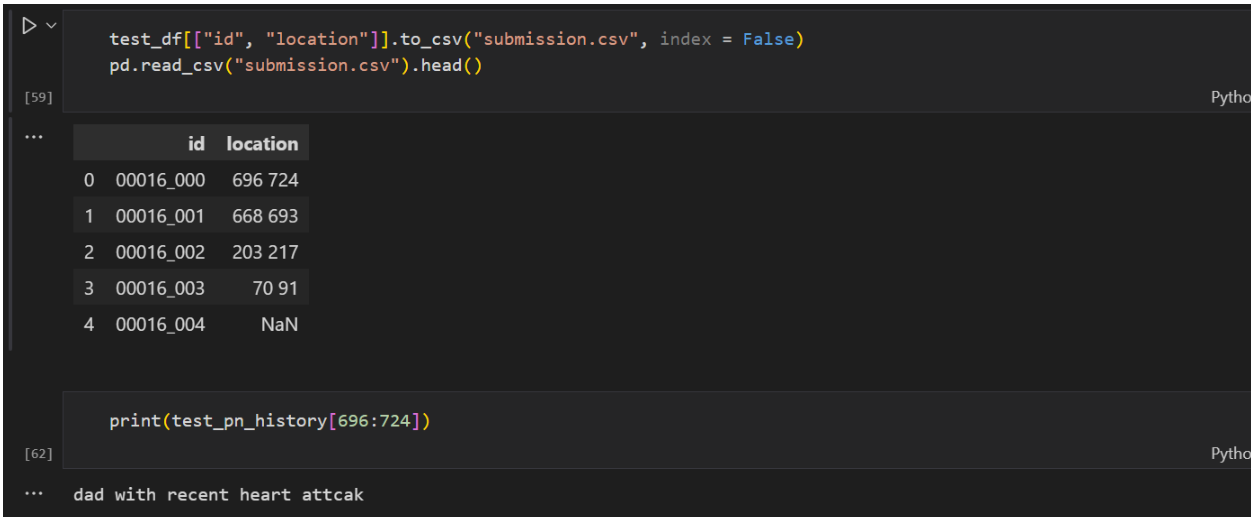


```python
def train_model(model, dataloader, optimizer, criterion):
        model.train()
        train_loss = []

        for batch in tqdm(dataloader):
            optimizer.zero_grad()
            input_ids = batch[0].to(DEVICE)
            attention_mask = batch[1].to(DEVICE)
#             token_type_ids = batch[2].to(DEVICE)
            labels = batch[2].to(DEVICE)

            logits = model(input_ids, attention_mask)
            loss = criterion(logits, labels)
            # since, we have
            loss = torch.masked_select(loss, labels > -1.0).mean()
            train_loss.append(loss.item() * input_ids.size(0))
            loss.backward()
```

GETTING THE LOCATIONS AND ANNOTATIONS-



```
test_df[["id", "location"]].to_csv("submission.csv", index = False)
pd.read_csv("submission.csv").head()
```

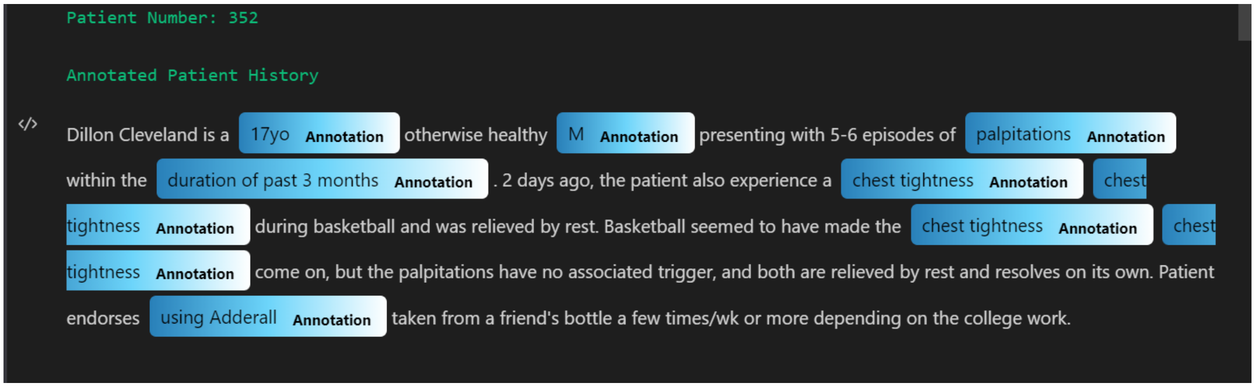
```
         id  location
0  00016_000  696 724
1  00016_001  668 693
2  00016_002  203 217
3  00016_003    70 91
4  00016_004      NaN
```

```
print(test_pn_history[696:724])
```

```
dad with recent heart attcak
```

VISUALIZING THE MODEL-

ANNOTATIONS-

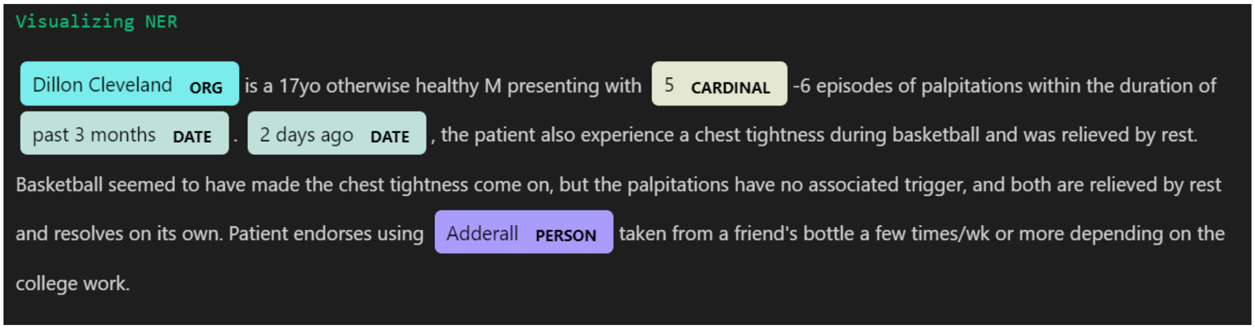


Patient Number: 352

Annotated Patient History

Dillon Cleveland is a 17yo Annotation otherwise healthy M Annotation presenting with 5-6 episodes of palpitations Annotation within the duration of past 3 months Annotation . 2 days ago, the patient also experience a chest tightness Annotation chest tightness Annotation during basketball and was relieved by rest. Basketball seemed to have made the chest tightness Annotation chest tightness Annotation come on, but the palpitations have no associated trigger, and both are relieved by rest and resolves on its own. Patient endorses using Adderall Annotation taken from a friend's bottle a few times/wk or more depending on the college work.

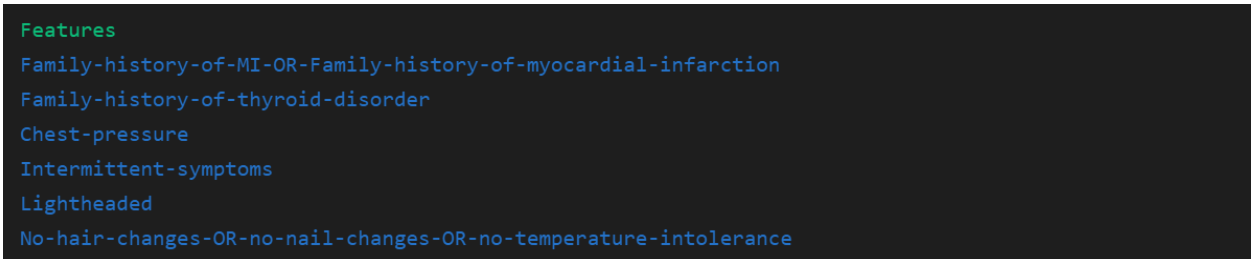
VISUALIZING NER-



Visualizing NER

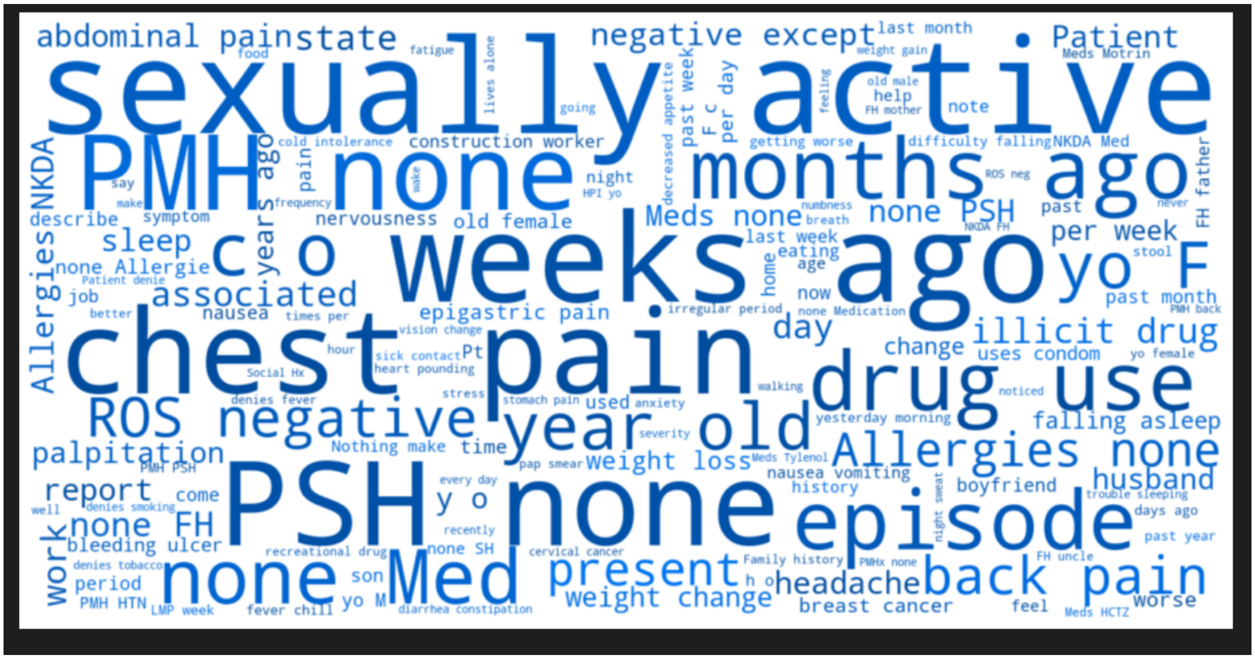
Dillon Cleveland ORG is a 17yo otherwise healthy M presenting with 5 CARDINAL -6 episodes of palpitations within the duration of past 3 months DATE . 2 days ago DATE , the patient also experience a chest tightness during basketball and was relieved by rest. Basketball seemed to have made the chest tightness come on, but the palpitations have no associated trigger, and both are relieved by rest and resolves on its own. Patient endorses using Adderall PERSON taken from a friend's bottle a few times/wk or more depending on the college work.

FEATURES-



```
Features
Family-history-of-MI-OR-Family-history-of-myocardial-infarction
Family-history-of-thyroid-disorder
Chest-pressure
Intermittent-symptoms
Lightheaded
No-hair-changes-OR-no-nail-changes-OR-no-temperature-intolerance
```

GENERATING WORDCLOUD FOR THE FEATURES/ANNOTATIONS



Conclusion:

So, we can conclude that checking examination papers for patient note-taking can be made a lot simpler by automating the whole process. Using various Natural Language processing visualization techniques, we can visualize the annotations in an interactive way. We can even mark the NER in the patient notes to get the features and make wordclouds for them.

Using various NLP techniques like tokenization we are preprocessing the dataset. Using the ROBERTA model, we are able to accurately mark the locations of the annotations i.e the symptoms of the patients who have been taken down by the doctors who are taking the exam.

So, the trained physicians who earlier had to manually check the notes can now use the web application to upload the examination notes, and get the locations of the annotations and whether they are present or not and accordingly mark the candidates hence making the whole process a lot simpler. While testing we can see that the locations of the right annotations have been correctly marked.

References:

[1] Karami, Amir & Gangopadhyay, Aryya & Zhou, Bin & Kharrazi, Hadi. (2017). Fuzzy Approach Topic Modeling for Health and Medical Corpora. International Journal of Fuzzy Systems. 20. 10.1007/s40815-017-0327-9.

[2] Ding, Liangping, Zhang, Zhixiong, Liu, Huan, Li, Jie and Yu, Gaihong. "Automatic Keyphrase Extraction from Scientific Chinese Medical Abstracts Based on Character-Level Sequence Labeling" Journal of Data and Information Science, vol.6, no.3, 2021, pp.35-57. https://doi.org/10.2478/jdis-2021-0013

[3] Rasmy, L., Xiang, Y., Xie, Z. et al. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digit. Med. 4, 86 (2021). https://doi.org/10.1038/s41746-021-00455-y

[4] Wu H, Ji J, Tian H, Chen Y, Ge W, Zhang H, Yu F, Zou J, Nakamura M, Liao J
Chinese-Named Entity Recognition From Adverse Drug Event Records: Radical Embedding-Combined Dynamic Embedding–Based BERT in a Bidirectional Long Short-term Conditional Random Field (Bi-LSTM-CRF) Model
JMIR Med Inform 2021;9(12):e26407
URL: https://medinform.jmir.org/2021/12/e26407
DOI: 10.2196/26407

[5] Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition Yun He1 , Ziwei Zhu1 , Yin Zhang1 , Qin Chen2 , James Caverlee1 https://arxiv.org/abs/2010.03746

[6] Xiong Y, Chen S, Chen Q, Yan J, Tang B
Using Character-Level and Entity-Level Representations to Enhance Bidirectional Encoder Representation From Transformers-Based Clinical Semantic Textual Similarity Model: ClinicalSTS Modeling Study
JMIR Med Inform 2020;8(12):e23357
URL: https://medinform.jmir.org/2020/12/e23357
DOI: 10.2196/23357

[7] Yao L, Jin Z, Mao C, Zhang Y, Luo Y. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. *J Am Med Inform Assoc*. 2019;26(12):1632-1636. doi:10.1093/jamia/ocz164

[8] Li F, Jin Y, Liu W, Rawat B, Cai P, Yu H Fine-Tuning Bidirectional Encoder Representations From Transformers (BERT)–Based Models on Large-Scale Electronic Health Record Notes: An Empirical Study

JMIR Med Inform 2019;7(3):e14830

URL: https://medinform.jmir.org/2019/3/e14830

DOI: 10.2196/14830

[9] Faiza Khan Khattak, Serena Jeblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, Frank Rudzicz,

A survey of word embeddings for clinical text,Journal of Biomedical Informatics,Volume 100,Supplement,2019,100057,ISSN 1532-0464,https://doi.org/10.1016/j.yjbinx.2019.100057.

(https://www.sciencedirect.com/science/article/pii/S2590177X19300563)

[10] Z. Dai, X. Wang, P. Ni, Y. Li, G. Li and X. Bai, "Named Entity Recognition Using BERT BiLSTM CRF for Chinese Electronic Health Records," 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2019, pp. 1-5, doi: 10.1109/CISP-BMEI48845.2019.8965823.

[11] Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, Jian Wang, An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition,

Bioinformatics, Volume 34, Issue 8, 15 April 2018, Pages 1381–1388, https://doi.org/10.1093/bioinformatics/btx761

[12] Agrawal, Ankit & Tripathi, Sarsij & Vardhan, Manu & Sihag, Vikas & Choudhary, Gaurav & Dragoni, Nicola. (2022). BERT-Based Transfer-Learning Approach for Nested Named-Entity Recognition Using Joint Labeling. Applied Sciences. 12. 976;. 10.3390/app12030976.

[13] DocBERT: BERT for Document Classification
Ashutosh Adhikari, Achyudh Ram,1 Raphael Tang,
Equal contribution.
Jimmy Lin
David R. Cheriton School of Computer Science
University of Waterloo

[14] BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision
Chen Liang*, Yue Yu*, Haoming Jiang*, Siawpeng Er, Ruijia Wang, Tuo Zhao, Chao Zhang
Georgia Institute of Technology, Atlanta, GA, USA
{cliang73,yueyu,jianghm,ser8,rwang,tourzhao,chaozhang}@gatech.edu

[15] Extracting Family History of Patients From Clinical Narratives: Exploring an End-to-End Solution With Deep Learning Models

[16] Extracting Drug Names and Associated Attributes From Discharge Summaries: Text Mining Study

[17] A Hybrid Model for Family History Information Identification and Relation Extraction: Development and Evaluation of an End-to-End Information Extraction System

[18] Named entity recognition for question answering