Features Capturing in Clinical Patient Notes

Ahmed Emad Eldeen
ahmedemadeldeen9112@gmail.com

Micheal Hany
michaelhany510@icloud.com

Mohamed Mostafa
mohamed.mostafa2774@gmail.com

Mohab Ali
ghobmohab@gmail.com

***Abstract*—Scoring patient notes that medical students write during their clinical skills assessment can be a very time consuming and overwhelming task for professional doctors. Those notes include very long sentences and due to the large number of notes, medical professors find themselves spending hours and a significant amount of effort to correctly interpret and analyze them. Consequently, we developed a Named Entity Recognition method using Natural Language Processing and Transformers to extract only the most important and representative sentences and words out of the given patient notes. We tried two different approaches and compared their results to identify the best solution, which appeared to be the second one using the Deberta Model with a precision of 0.84 , recall of 0.89 and f1-score of 0.866 .**

Keywords—Named Entity Recognition, Natural Language Processing, Transformers

Introduction

Assessing physicians' proficiency in documenting patients' symptoms and history is vital in practical medical exams. However, evaluating these patient notes traditionally involves time-consuming human assessment. The integration of deep learning offers a promising avenue for more efficient evaluation methods.

The Step 2 Clinical Skills examination of the USMLE relied on manual evaluation of patient notes, demanding substantial resources. Despite advancements in natural language processing, computational scoring of patient notes remains challenging due to varied expressions of clinical features and complex linguistic nuances.

This research aims to develop an automated method for identifying and mapping clinical concepts in patient notes. It is an NER (Named Entity Recognition) problem where we want to map the different ways of expressing some terminology to this terminology. So, simply our inputs are a punch of sentences, and outputs are a classification for each word so that we can determine to which class every word belongs, it can belong to a certain terminology or it can be a non-deterministic word that is referred to as 'others' in our model. This initiative seeks to revolutionize patient note evaluation, enhancing the time needed to assess patient notes.

I.Related Work.

I .Traditional Rule-Based NER Approaches

A. Approach
Early systems in biomedical Named Entity Recognition [4] (NER) largely hinged on rule-based methods. These systems were engineered with manually crafted rules or dictionaries specifically designed to identify entities such as diseases, chemicals, or genes.

B. Strengths
Rule-based approaches are characterized by their high precision within specifically defined domains and clear interpretability of rules

C. Weaknesses
They struggled to recognize new terms or different ways of saying the same thing. They needed a lot of work from experts to create and update these rules, and they often missed other relevant terms.

D. Relation to Current approach
Distinct from DeBERTa and BERT. Rule-based relies on predefined rules, while DeBERTa and BERT learn from data.
Provides more flexibility and scalability without needing extensive rules.

State of the art involves more nuanced rules, hybrid systems combining machine learning, and domain-specific enhancements for better precision in specialized areas.

II. Transfer Learning and Semi-Supervised Learning in NER

A. Approach
Transfer learning is about teaching a model with a lot of data on one task and then tweaking it to do well on a related but different task. Semi-supervised learning mixes a little bit of data where we know the answers (labeled) with a lot of data where we don't (unlabeled).

B. Strengths
These methods [4] are great because they don't need as much labeled data, which can be hard and expensive to get.

They let us use information from similar tasks or areas, and they're really good for situations where we don't have a lot of resources.

C. Weaknesses

The success of transfer learning depends a lot on how similar the two tasks are. If they're too different, it might not work well, or it could even make things worse (negative transfer).

D. Relation to Current approach

Central to DeBERTa and BERT's methodology, enabling the use of pre-trained knowledge and reducing the need for extensive labeled data, thereby facilitating efficient and precise patient note analysis.

State of the art includes advanced models like BERT and DeBERTa for broader adaptability, cross-lingual and multi-domain applications, and methods aiming at reducing data needs with few-shot and zero-shot learning.

I. Dataset & Features

A. Dataset Description

The dataset encompasses approximately 40,000 patient notes pertaining to around 100 patients, with feature annotations available for a subset of approximately 1000 notes so we splitted our Data into Train/Test/Validation 80%,10%,10% respectively. Within the patient_notes.csv file, each entry is uniquely identified by a patient note identifier and a corresponding clinical case identifier, accompanied by the recorded encounter text as documented by the test taker. The features.csv file provides an exhaustive list of key concepts for each clinical case, identified by a unique feature identifier, case identifier, and a descriptive feature text. The train.csv file is predominantly dedicated to housing feature annotations for 1000 patient notes, distributed across ten cases, with each entry comprising a unique identifier for the patient note/feature pair, the associated patient note and feature identifiers, the relevant case information, the text indicative of the annotated feature (which may occur multiple times within a single note), and character spans specifying the location of each annotation within the note. In instances where multiple spans are necessary, they are appropriately delimited by semicolons. This meticulously structured dataset is tailored to facilitate a nuanced exploration of clinical case data and annotation details .

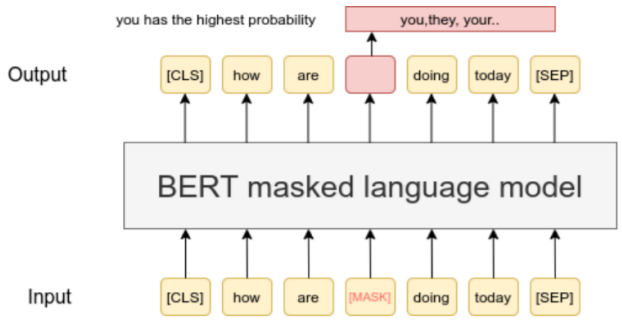
B. Dataset Preprocessing

In the pre-processing phase, our primary emphasis was on ensuring the cleanliness and uniformity of the dataset. To achieve this, we employed various techniques, including the removal of HTML tags, special characters, and single characters. Additionally, we performed the conversion of all data to lowercase for standardization purposes. Subsequently, meticulous attention was directed towards rectifying annotation errors and addressing spelling discrepancies within the dataset. This involved the application of both advanced and manual correction techniques to ensure a consistent and accurate representation of information. Following data cleansing, we proceeded with tokenization and labeling, tailoring our approach to the specific requirements of each subsequent analytical method, which will be expounded upon in later sections. This comprehensive pre-processing strategy underscores our commitment to cultivating a robust and refined dataset for subsequent analyses.

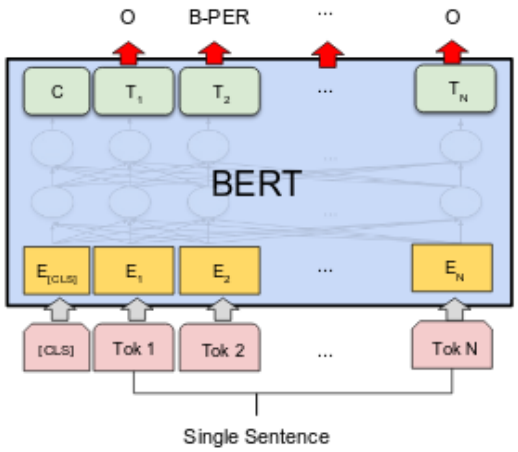
II. Methods

Upon extensive review of our problem domain, We attempted two methods to solve this issue. Both methods are based on BERT architecture**[2]**, known for its effectiveness in understanding the semantics of sentences. The initial strategy involves leveraging Named Entity Recognition (NER), a well-established method in natural language processing. The second approach introduces a hybrid methodology combining NER with Question Answering (QA), presenting a unique and challenging endeavor. This hybrid model aims to harness the strengths of both techniques, promising a nuanced and comprehensive solution to our task. This strategic selection reflects a thoughtful consideration of the intricacies inherent in our dataset, with the ultimate goal of enhancing the model's efficacy in extracting pertinent information from patient notes.

A. NER Method using BERT.



[7] In the first method, we used a BERT architecture called distilbert-base-uncased. We pre-trained this architecture using the unsupervised learning paradigm using all the patient_notes data. In this unsupervised training, we hide some words and challenge the model to predict them, known widely as MLM. This approach helped the architecture understand the meanings of words better. We chose this method to utilize all our data because only 1000 out of 43,000 patient notes were labeled.



[8] By pretraining on the unlabeled data, we created a model that can extract features effectively. After that, we fine-tuned this model using the labeled data to predict word classes. This approach allowed us to train the model to recognize specific categories while benefiting from the broader understanding obtained from the unlabeled data. This way gives results comparable to a supervised model trained on 3x-8x the amount of labeled data.[6]

B. QA/NER Hybrid Method based on DeBerta

Model we are going to fine tune to be able to make it work on our task is DeBerta so what is Deberta ?

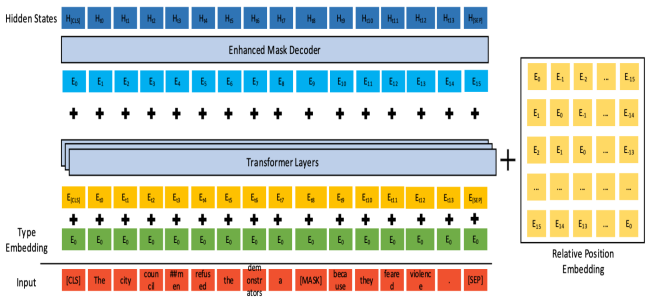


Fig 1. The model architecture of DeBERTa

[1] DeBERTa , short for Decoding-Enhanced BERT with Disentangled Attention, represents a significant advancement over previous transformer models like BERT and RoBERTa. The innovation lies in its disentangled attention mechanism and an enhanced masked decoder, which collectively contribute to improved natural language understanding.

DeBERTa introduces a disentangled attention mechanism where each input word is represented by two separate vectors encoding content and position. Unlike traditional models, these vectors are treated separately throughout the network. Disentangled attention is achieved by using separate matrices for the transformation of content and positional embeddings, enabling a more nuanced understanding of the relationships between words.

.

DeBERTa proves beneficial for long sequences, as it allows tokens to attend directly to a larger context, theoretically enabling longer sequences. The model outperforms RoBERTa in various downstream NLP tasks, demonstrating its effectiveness in tasks like text classification.

The proposed methodology represents a novel QA/NER hybrid approach [5] , combining aspects of Question Answering (QA) and Named Entity Recognition (NER). In essence, this approach integrates characteristics of NER, as each token undergoes classification. Simultaneously, it incorporates elements to QA by strategically placing the feature text at the beginning of the input,as if it was a question in front of a context. The key conceptualization revolves around the classifier discerning whether a specific token holds significance for the provided feature text within the input. The inclusion of feature text is designed to guide the model in the relevant spans within the text. This strategy is particularly relevant in the context of multi-span QA, as the conventional extractive QA methods excel primarily in single-span predictions. This innovative approach underscores our commitment to adapting and synthesizing methodologies, aiming for heightened precision and relevance in navigating the complexities inherent in our task.

The model is trained to predict the location of relevant information within the patient notes, aligning with the provided feature text. This involves classifying each token in the input text as either relevant or non-relevant to the specified feature and the Binary Cross Entropy with Logits.

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}(Y_i \cdot \log \hat{Y}_i + (1 - Y_i) \cdot \log (1 - \hat{Y}_i))$$

This loss function is commonly used for binary classification problems. It combines the sigmoid activation function with the binary cross entropy loss, making it suitable for tasks where each instance has two possible outcomes as we mentioned before which is our task actually.

IV. Experiments/Results

A. NER Method using BERT.

- For the Pre-training Phase:

A) Hyperparameters: A **batch size** of **32** for training, we chose this after multiple experiments and found it to be the most reasonable balance between training speed, loss fluctuations and memory consumption**. A learning rate of 1e-4** as it achieves the minimum training loss in less number of epochs. smaller LRs were very slow and higher LRs kept the training loss large. **The number of epochs is 6**, as we found this number to provide the lowest training loss without overfitting, combined with our chosen LR.

B) Primary metrics: At this phase, we were only evaluating the **training loss** as our metric.

C) Results: We achieved a final training loss of **0.971**

Step	Training Loss	Step	Training Loss
500	5.777400	50000	0.977600
1000	4.811900	50500	0.969700
1500	4.383600	51000	0.955500
2000	3.951100	51500	0.972000
2500	3.547200	52000	0.971100

- For the Fine-tuning Phase:

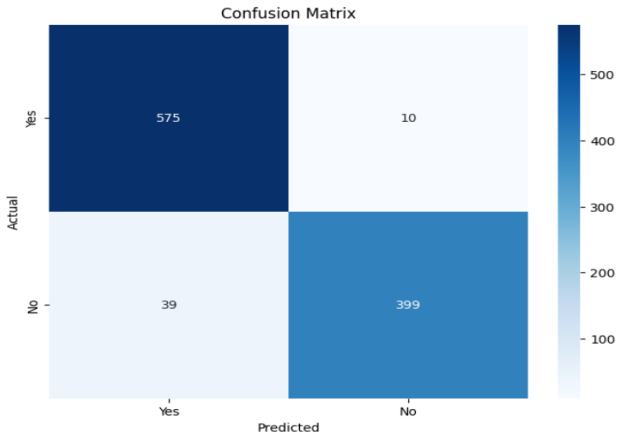
A)Hyperparameters: A **batch size** of **16** for training, we chose this after multiple experiments and found it to be the most reasonable balance between training speed, loss fluctuations and memory consumption**. A learning rate of 1e-3** as it achieves the minimum training and validation loss in less number of epochs. smaller LRs were very slow and higher LRs kept the training loss large. **The number of epochs is 15**, as we found this number to provide the lowest training and validation loss without overfitting, combined with our chosen LR. **The number of folds is 5**, as for our small number of labeled data, we found this to be the most suitable number to evaluate the model with a reasonable amount of test data.

B) Primary metrics: Precision, Recall and F1 score.

C) Results: We reached those scores averaged over the 5 folds: **Precision = 0.85, Recall = 0.89,** and **F1-Score = 0.87**.

B . QA/NER Hybrid Method based on DeBerta.

1) **DeBERTa hyperparameters:** A **batch size** of 8 for both training and evaluation the choice of 8 is a reasonable balance between training speed and memory consumption**. A gradient accumulation step** we set to 2 means that gradients are accumulated over 2 steps before updating the model's weights. This is useful when because of limited GPU memory **A learning rate of 2e-5** as it achieves best performance .**A weight decay of 0.01** is a regularization technique helps prevent overfitting by encouraging smaller weights **num_train_epochs are** 5 epochs is relatively common and provides a reasonable balance between training time and model convergence
2) **DeBERTa Primary metrics:** Precision, Recall and F1 score.
3) **Confusion matrix:**



4) **Results:**

Epoch	Training Loss	Validation Loss	Precision	Recall	F1
1	0.021000	0.018828	0.723450	0.837672	0.776382
2	0.014400	0.014451	0.776129	0.908771	0.837229
3	0.011300	0.012851	0.852415	0.872918	0.862544
4	0.008300	0.012668	0.828704	0.900927	0.863308
5	0.007200	0.013098	0.842898	0.891330	0.866438

“ Quantitative Results of Deberta”

CONCLUSION

Both approaches provided a performance close to each other, but we decided to use the later one. As the BERT model may suffer from instability due to the imbalanced labels and the very small size of test data, the later method seemed to be a better and more stable approach so The second method is better in terms of accuracy and reliability. The high precision and recall indicate that it is effective in correctly identifying and categorizing entities within clinical texts. This suggests that Deberta is a promising tool for automating and improving the extraction of medical information from unstructured text, thereby aiding in better healthcare data management and analysis. The findings underscore the importance and effectiveness of using advanced NLP models like Deberta for clinical and medical informatics tasks.

Future

We need to Expand Dataset and Domain Specificity: Broaden the dataset to include a more diverse set of patient notes from different medical fields and demographics. This could improve the model's robustness and applicability across various

medical conditions and patient populations. Multilingual and Dialect Support: Extend the model's capabilities to understand and process patient notes in multiple languages and dialects. This would make the tool more versatile and useful in a global healthcare setting. Integration with Electronic Health Records (EHRs): Work on seamless integration of this NLP technology with existing EHR systems to enhance real-time data extraction, interpretation, and decision-making in clinical settings. Improve Interpretability and Explainability: Despite high performance, many advanced NLP models remain somewhat "black-box" in nature. Enhancing the interpretability of these models will increase trust and adoption among healthcare practitioners. Real-time Feedback and Learning: Implement a system where the model can receive real-time feedback from healthcare professionals to learn and improve continuously. This could include a mechanism for users to correct or validate the extracted information.

References

1. *DEBERTA DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION (2021)*
2. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019)*
3. Attention Is All You Need (2017)
4. Recent Trends in Named Entity Recognition (NER) (2021)
5. QaNER: Prompting Question Answering Models for Few-shot Named Entity Recognition (2022)
6. Doe, J. (2021). A pre-training and self-training approach for biomedical named entity recognition
7. MLM Photo: MLM — Sentence-Transformers documentation (sbert.net)AAAAAQAw
8. Bert Training Photo: Named entity recognition with Bert (depends-on-the-definition.com)

Contribution

Names	Research	Coding
Ahmed Emad Eldeen	DEBERTA :QA/NER	Artitecture building , Training Model
Mohab Ali	DEBERTA :QA/NER	Pre-Processing s , Testing Model,
Mohamed Mostafa	NER : BERT	Pre-Processing s , Testing Model,
Michael Hany	NER : BERT	Artitecture building , Training Model