Medical Concept Extraction and Relationship Classification from Patient Records

April 2023

Team Name: ConceptMiners
Team Members:

- Meet Gandhi; pennkey: mgandhi5; Email: `mgandhi5@seas.upenn.edu`
- Rohan Saraogi; pennkey: rsaraogi; Email: `rsaraogi@seas.upenn.edu`
- Tarun Raheja; pennkey: traheja; Email: `traheja@seas.upenn.edu`

Abstract

In this project, we compare the performance of deep learning models, specifically transformers and Long Short-Term Memory (LSTM) networks, against traditional machine learning baselines in a healthcare context. Our investigation explores the impact of data augmentation and importance weight adjustments on model performance, particularly in addressing class imbalance and improving prediction accuracy. Results demonstrate the superior performance of deep learning models over machine learning baselines across various performance metrics. Notably, transformers exhibit promising results with fewer training epochs compared to LSTMs. Furthermore, the study reveals distinct sensitivities of transformers and LSTMs to changes in importance weights, highlighting the need for further research into model architectures and optimization techniques. The findings contribute to a broader understanding of artificial intelligence applications in healthcare, emphasizing the potential and challenges of AI-driven decision-making in medical settings.

1 Introduction

The motivation for our project lies in the increasing need for accurate and efficient extraction of medical concepts and relationship classification from patient records. This task is critical for improving clinical decision-making, patient care, and facilitating medical research. The problem formulation stems from the challenge of accurately identifying and classifying medical concepts and relationships within the vast and complex landscape of electronic health records (EHRs). Our contributions to the field include the development of ontology-aware deep learning models that leverage the Unified Medical Language System (UMLS) to enhance the performance of medical concept extraction and relationship classification.

In the realm of medical concept extraction, various methods have been employed, with token-level machine learning (ML) models serving as the baseline. However, there remains a gap in the utilization of ontologies and the semantic information they contain, which can contribute to improved model performance, in terms of recall. Our paper focuses on closing this gap by incorporating the UMLS, a comprehensive collection of medical terminologies, into our BiLSTM-CRF and Transformer-based models. By doing so, we aim to demonstrate the potential benefits of integrating ontological information into deep learning models for the task of medical concept extraction and relationship classification. The reader should be interested in our project as it tackles a vital problem in the medical domain, with the potential to greatly enhance healthcare and research by enabling more effective use of EHR data.

In this paper, we present a detailed exploration of our approach, which consists of developing and comparing BiLSTM-CRF and Transformer models that are ontology-aware, utilizing UMLS to improve medical concept extraction and relationship classification from patient records. Our results demonstrate the effectiveness of incorporating ontological information into deep learning models, which results in higher recall for class-imbalanced entities important for the medical domain. The reader can expect a comprehensive examination of our methodology, a discussion of related work in the field, and a thorough evaluation of our models' performance in comparison to the token-level ML baseline. By the end of the paper, the reader will gain a clear understanding of the potential benefits of integrating ontological information into deep learning models for medical concept extraction and relationship classification, thereby contributing to the existing literature in the field.

2 Related Work

In this summary of related work, we highlight prominent studies that have significantly contributed to the field of medical concept extraction and relationship classification from patient records.

One of the seminal works in medical concept extraction is the cTAKES (clinical Text Analysis and Knowledge Extraction System) developed by [Sav+10]. cTAKES is a natural language processing (NLP) system that identifies and maps medical concepts from clinical narratives to standard terminologies, such as UMLS. This system has since been widely adopted in various clinical applications and research endeavors.

A more recent approach to medical concept extraction is the application of deep learning models. In particular, the BiLSTM-CRF model, introduced by [HXY15], has demonstrated significant improvements in named entity recognition (NER) tasks, including those in the medical domain. This model combines bidirectional long short-term memory (BiLSTM) neural networks with conditional random fields (CRFs) to capture contextual information and model label dependencies in a sequential data.

In the context of relationship classification, the BioBERT model proposed by [Lee+19] stands out as a prominent development. BioBERT is a pre-trained biomedical language representation model based on the original BERT architecture by [Dev+18]. By leveraging large-scale biomedical text corpora for pre-training, BioBERT has shown remarkable performance improvements in various biomedical NLP tasks, including relation extraction and entity recognition.

Our project stands out as innovative compared to prior work in several significant ways. First, while previous research in medical concept extraction and relationship classification has demonstrated the effectiveness of deep learning models, such as BiLSTM-CRF and BioBERT, our project takes a step further by incorporating ontological information from the Unified Medical Language System (UMLS) into these models. By leveraging the wealth of semantic information available in UMLS, we enhance the capabilities of our models to recognize and classify medical concepts and relationships, going beyond the existing approaches in the field.

Second, our project addresses the challenge of class imbalance in medical concept extraction and relationship classification tasks. Previous work, although effective in many aspects, has not extensively focused on improving recall for the underrepresented classes. Our ontology-aware deep learning models, by utilizing the rich knowledge embedded in UMLS, are better equipped to recognize and classify rare medical concepts and relationships. This improvement in recall is particularly valuable in the medical domain, where accurate identification of rare concepts can be crucial for clinical decision-making, patient care, and medical research.

3 Dataset and Features

3.1 Dataset

We are using Harvard Medical School's "n2c2 adverse drug events (ADE) and medication extraction in the electronic health records" dataset for the project. The dataset has 303 de-identified medical records for training and 202 for testing from the MIMIC-III database.

For every medical record text file the dataset contains an annotation file with domain expert annotated entity tags including drug, strength, dosage, duration, frequency, form, route, reason, and ADE tags. For example the line ”T3 ADE 11270 11293 Abdominal wall hematoma” in the annotation file indicates that the entity with ID ”T3” is of type ”ADE”, occurs from index 11270 to index 11293 in the associated medical record text file, and has text ”Abdominal wall hematoma”.

In addition to the entity annotations the annotation files also contain domain expert annotated relationship tags including strength-drug, dosage-drug, duration-drug, frequency-drug, form-drug, route-drug, reason-drug, and ADE-drug relationships tags. For example, the line ”R1 ADE-Drug Arg1:T3 Arg2:T4” in the annotation file indicates that the relationship with ID ”R1” is of type ”ADE-Drug” and is between entities with IDs ”T3” and ”T4”.

3.2 Parsing Approach

We identify text spans in patient records that meet the following criteria: they include at least two entities and a relation between those entities. The start and end of these spans are determined by locating the nearest full-stop or line break before the first entity and after the last entity. If an entity does not have a relationship with any other entities in the span, we extract a text span that includes only that single entity. These text spans containing entities and relations are used for both Named Entity Recognition (NER) and relation extraction tasks, while spans with only entities are used solely for the NER task. Additionally, we merge spans that contain different relationships but share identical text due to the inclusion of more than two entities.

Below we can see the class distributions of the entity and relationship tags. As can be seen, there is a class imbalance issue, with the ADE and Duration entities/relationships being among the least represented.

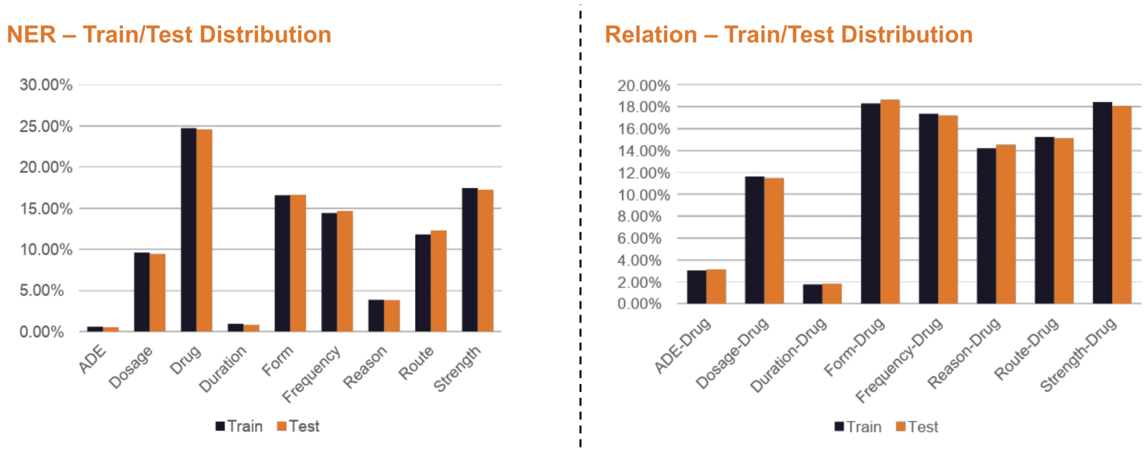


Figure 1: Class Distributions

Finally, in our deep learning models for concept extraction and relationship classification, we are using context-insensitive and aware word embeddings as inputs. The details of these embeddings are explained in detail in the methodology section.

4 Methodology

4.1 Machine Learning Baseline

In our baseline methodology, we employ token-level features, such as part-of-speech (POS) tags and case information, to capture contextual information within a window of two tokens for medical concept extraction. A token-level classification model is developed using these features to identify and classify medical concepts within the text. For relationship classification, we extract features using the Term Frequency-Inverse Document Frequency (TF-IDF) representation of the shortest dependency paths between the identified entities. Both the medical concept extraction and relationship classification tasks utilize Logistic Regression as the baseline model, leveraging the selected features to perform classification tasks. This baseline approach provides a foundation for comparing the performance of our proposed ontology-aware deep learning models in medical concept extraction and relationship classification tasks.

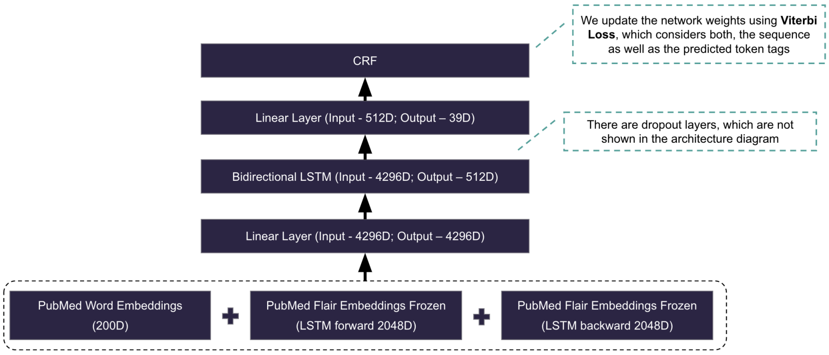
Feature	**Description**
pos tag	Part-of-speech tag of the token
istitle	True if the token is in title case, otherwise False
isupper	True if the token is in uppercase, otherwise False
isalpha	True if the token consists only of alphabetic characters, otherwise False
isnumeric	True if the token consists only of numeric characters, otherwise False
containsnumbers	True if the token contains any numeric characters, otherwise False

Table 1: Token-level features and their descriptions

4.2 Deep Learning Models for Named Entity Recognition

4.2.1 Bidirectional LSTM-CRF

The Bidirectional LSTM with Conditional Random Fields (BiLSTM-CRF) model has proven to be effective for Named Entity Recognition (NER) tasks, particularly in medical concept extraction. This model combines the strengths of both BiLSTMs and CRFs to effectively capture sequential patterns and dependencies in textual data. BiLSTMs consist of two separate LSTM layers, one processing the input sequence in the forward direction and the other in reverse, enabling the model to capture context from both past and future tokens. The CRF layer, added on top of the BiLSTM layer, models dependencies between labels in the output sequence to improve prediction accuracy. In the context of our project, the BiLSTM-CRF model is employed for medical concept extraction from patient records, taking as input the combined context-insensitive PubMed word embeddings and context-aware PubMed Flair embeddings. The BiLSTM layer captures contextual information, while the CRF layer models dependencies between medical concept labels. As a result, the BiLSTM-CRF model demonstrates high performance in identifying and classifying medical concepts, making it an appropriate choice for our NER task. The model architecture is shown below.


CRF
Linear Layer (Input - 512D; Output – 39D)
Bidirectional LSTM (Input - 4296D; Output – 512D)
Linear Layer (Input - 4296D; Output – 4296D)
PubMed Word Embeddings (200D)
+
PubMed Flair Embeddings Frozen (LSTM forward 2048D)
+
PubMed Flair Embeddings Frozen (LSTM backward 2048D)

We update the network weights using **Viterbi Loss**, which considers both, the sequence as well as the predicted token tags

There are dropout layers, which are not shown in the architecture diagram

Figure 2: NER LSTM-CRF Model

4.2.2 Transformer-CRF

The pretrained Transformer with Conditional Random Fields (Transformer-CRF) model is an alternative approach we explore for Named Entity Recognition (NER) in medical concept extraction. Transformers, a type of neural network architecture, excel in modeling long-range dependencies and capturing context in textual data using self-attention mechanisms. By pretraining the Transformer model on a large corpus of clinical notes, we obtain a strong language representation well-suited for medical concept extraction tasks. CRFs, added on top of the pretrained Transformer layer, model dependencies between labels in the output sequence, enhancing overall prediction accuracy.

In our project, the pretrained Transformer-CRF model is employed for medical concept extraction from patient records. The pretrained Transformer layer captures contextual information from the input text, while the CRF layer models dependencies between medical concept labels. As a result, the Transformer-CRF model (upon sufficient training), demonstrates high performance in identifying and classifying medical concepts, making it a valuable alternative for the NER task in our project. The model architecture is shown below.

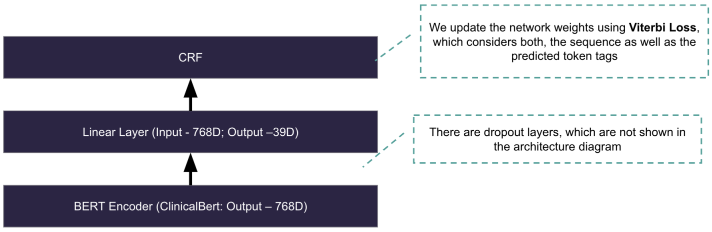

CRF
Linear Layer (Input - 768D; Output –39D)
BERT Encoder (ClinicalBert: Output – 768D)
We update the network weights using **Viterbi Loss**, which considers both, the sequence as well as the predicted token tags
There are dropout layers, which are not shown in the architecture diagram

Figure 3: NER Transformer Model

4.3 Deep Learning Model for Relationship Extraction

In this section, we describe our approach for relationship extraction using a linear layer on top of Flair embeddings. Flair embeddings, generated by the state-of-the-art Flair language modeling framework, have proven effective in capturing context-dependent information in textual data. By leveraging the contextualized information provided by Flair embeddings, we obtain a rich representation of the input text, crucial for accurately identifying relationships between medical concepts. We opted not to experiment with Transformer models in this task, as the Flair embeddings already demonstrated great performance.

Our relationship extraction approach feeds the Flair embeddings into a linear layer, designed to learn and model the relationships between medical concepts present in the text. The linear layer, also known as a fully connected layer, acts as a classifier that maps input Flair embeddings to relationship classes. By training the linear layer on a dataset of annotated relationships, it learns to identify and classify relationships in unseen text. In summary, our approach combines a linear layer with context-aware Flair embeddings for effective extraction of relationships between medical concepts, contributing to a comprehensive understanding of patient records.

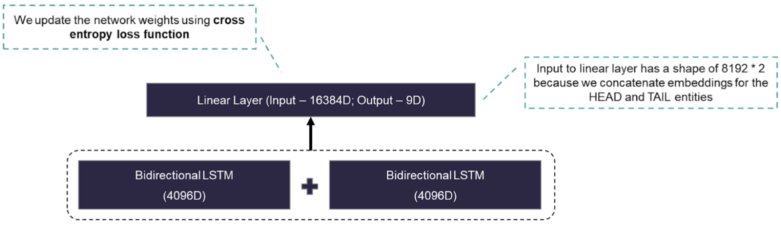

We update the network weights using **cross entropy loss function**
Linear Layer (Input – 16384D; Output – 9D)
Input to linear layer has a shape of 8192 * 2 because we concatenate embeddings for the HEAD and TAIL entities
Bidirectional LSTM (4096D)
Bidirectional LSTM (4096D)

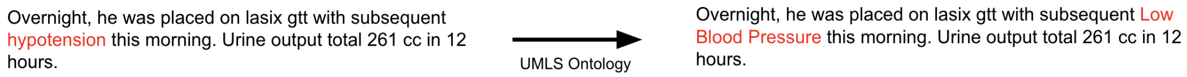
Figure 4: Relationship Extraction Model

4.4 UMLS Augmentation

The integration of Unified Medical Language System (UMLS) augmentation enhances the ontology-awareness of our models for medical concept and relationship extraction tasks. UMLS is a comprehensive knowledge source containing a wide range of biomedical concepts, terms, and relationships. By incorporating UMLS information, we improve the models' understanding of biomedical entities and relationships, leading to better extraction performance.

To achieve this, we map medical concepts identified in the text to their corresponding UMLS Concept Unique Identifiers (CUIs) and leverage the structured knowledge available in the UMLS Metathesaurus, such as preferred terms, synonyms, and semantic types. This additional information improves concept and relationship identification by handling variations in medical terminology. Furthermore, we integrate information from the UMLS Semantic Network, providing additional context and constraints to the relationship extraction model, making it more sensitive to the underlying semantics of the relationships. In summary, UMLS augmentation serves as a valuable technique to enhance the ontology-awareness of our models, resulting in improved performance in the extraction tasks.

We generate augmented data by taking a subset of the original data with ADE entities, and doubling it in size using UMLS augmentation. We specifically focus on ADE entities as we feel the other entities already have reasonable Precision/Recall/F1-scores whereas the ADE entities have poor scores.



Overnight, he was placed on lasix gtt with subsequent hypotension this morning. Urine output total 261 cc in 12 hours.

Overnight, he was placed on lasix gtt with subsequent Low Blood Pressure this morning. Urine output total 261 cc in 12 hours.

Figure 5: Simple example of ADE augmentation. ‘hypotension‘ in the left sentence is replaced with ‘Low Blood Pressure‘ in the right sentence to create an augmented sample.

5 Results

We aim to extract entities and their relationships from unstructured clinical textual data. For NER we trained two deep learning models and ran some experiments on them to see if they capture more detail than the baseline ML models. For Relationship Extraction we trained a deep learning model to observe if it captures more relationships and intricacies than a baseline ML method.

In the following subsections we display the training loss and metric tables for each of our models.

5.1 Loss function

We have utilized the following loss functions

1. Cross-entropy loss: This has been used for the NER baseline model, REL baseline model, and REL LSTM model

2. Viterbi loss: For the NER models we have optimized Viterbi loss. Viterbi loss is a commonly used loss function in named entity recognition (NER) models, including those implemented in Flair NLP. It is named after the Viterbi algorithm, which is a dynamic programming algorithm used to find the most likely sequence of hidden states in a hidden Markov model.

   In the context of NER, the Viterbi loss is a variant of the cross-entropy loss that takes into account not only the correctness of the predicted entity label for each token in the sequence but also the coherence of the predicted entity labels across the entire sequence. This is achieved by calculating the score of the most likely entity label sequence using the Viterbi algorithm, and penalizing the model for deviating from this sequence.

   The Viterbi loss is useful in NER models because it encourages the model to make consistent and plausible predictions for the entire input sequence, rather than simply optimizing for local correctness at each individual token. This can lead to better performance on tasks that require accurate labeling of longer entity spans, such as identifying multi-word names or phrases.

5.2 Loss Function Scaling

In our typical deep learning experiments, we have utilized a scaling factor for the loss function that is based on the inverse of entity counts. However, we also conducted a separate series of experiments where we specifically targeted the underrepresented ADE class in combination with the Reason class. This was motivated by the fact that, in medical text, both ADE and Reason often refer to diseases, making it difficult to distinguish between them without proper contextualization.

By selectively scaling the loss for only the ADE and Reason classes, we aimed to encourage the model to more effectively leverage the surrounding context and improve its ability to differentiate between these two types of entities. In particular, we gave the ADE and Reason tags 5 times the weight compared to the other tags.

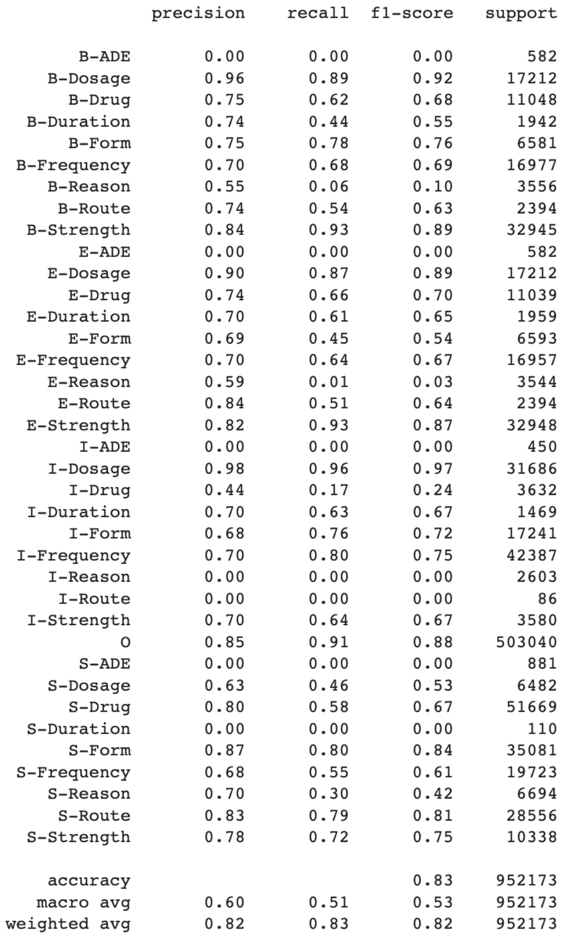
5.3 Performance Metrics

We looked at measures that are informative under class imbalance - Precision, Recall and F1 Score. However we mainly focused on Recall since it is important to ensure that valid entities and relationships are detected (perhaps at the cost of false-positives).

5.4 NER experiments

5.4.1 Baseline

The following table shows the test performance metrics for our baseline model on the NER task. Note that in this case each of the tags are split into their BIOES sub parts in line with the BIOES tagging scheme for NER.

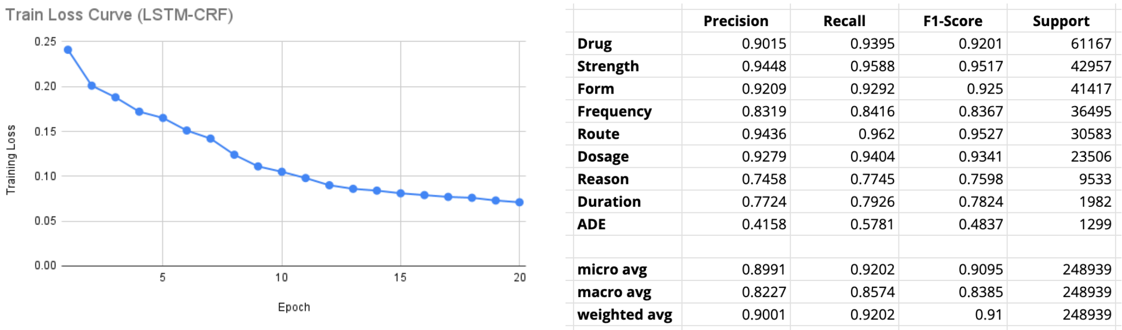


```
              precision    recall  f1-score   support

       B-ADE       0.00      0.00      0.00       582
    B-Dosage       0.96      0.89      0.92     17212
      B-Drug       0.75      0.62      0.68     11048
  B-Duration       0.74      0.44      0.55      1942
      B-Form       0.75      0.78      0.76      6581
 B-Frequency       0.70      0.68      0.69     16977
    B-Reason       0.55      0.06      0.10      3556
     B-Route       0.74      0.54      0.63      2394
  B-Strength       0.84      0.93      0.89     32945
       E-ADE       0.00      0.00      0.00       582
    E-Dosage       0.90      0.87      0.89     17212
      E-Drug       0.74      0.66      0.70     11039
  E-Duration       0.70      0.61      0.65      1959
      E-Form       0.69      0.45      0.54      6593
 E-Frequency       0.70      0.64      0.67     16957
    E-Reason       0.59      0.01      0.03      3544
     E-Route       0.84      0.51      0.64      2394
  E-Strength       0.82      0.93      0.87     32948
       I-ADE       0.00      0.00      0.00       450
    I-Dosage       0.98      0.96      0.97     31686
      I-Drug       0.44      0.17      0.24      3632
  I-Duration       0.70      0.63      0.67      1469
      I-Form       0.68      0.76      0.72     17241
 I-Frequency       0.70      0.80      0.75     42387
    I-Reason       0.00      0.00      0.00      2603
     I-Route       0.00      0.00      0.00        86
  I-Strength       0.70      0.64      0.67      3580
           O       0.85      0.91      0.88    503040
       S-ADE       0.00      0.00      0.00       881
    S-Dosage       0.63      0.46      0.53      6482
      S-Drug       0.80      0.58      0.67     51669
  S-Duration       0.00      0.00      0.00       110
      S-Form       0.87      0.80      0.84     35081
 S-Frequency       0.68      0.55      0.61     19723
    S-Reason       0.70      0.30      0.42      6694
     S-Route       0.83      0.79      0.81     28556
  S-Strength       0.78      0.72      0.75     10338

    accuracy                           0.83    952173
   macro avg       0.60      0.51      0.53    952173
weighted avg       0.82      0.83      0.82    952173
```

Figure 6: Table for NER

5.4.2 LSTM model

The following train-loss curve and test performance metrics show the results of training our LSTM model on the NER task.



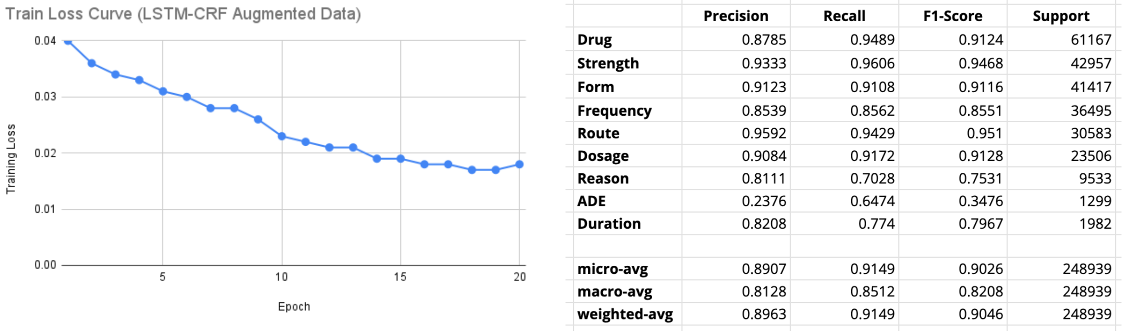
(a) Training Loss Curve

	Precision	Recall	F1-Score	Support
Drug	0.9015	0.9395	0.9201	61167
Strength	0.9448	0.9588	0.9517	42957
Form	0.9209	0.9292	0.925	41417
Frequency	0.8319	0.8416	0.8367	36495
Route	0.9436	0.962	0.9527	30583
Dosage	0.9279	0.9404	0.9341	23506
Reason	0.7458	0.7745	0.7598	9533
Duration	0.7724	0.7926	0.7824	1982
ADE	0.4158	0.5781	0.4837	1299
micro avg	0.8991	0.9202	0.9095	248939
macro avg	0.8227	0.8574	0.8385	248939
weighted avg	0.9001	0.9202	0.91	248939

(b) Test performance metrics

Figure 7: LSTM-CRF plot and table for NER

The following train-loss curve and test performance metrics show the results of fine-tuning our LSTM model (from above) for an additional 20 epochs on the augmented data.



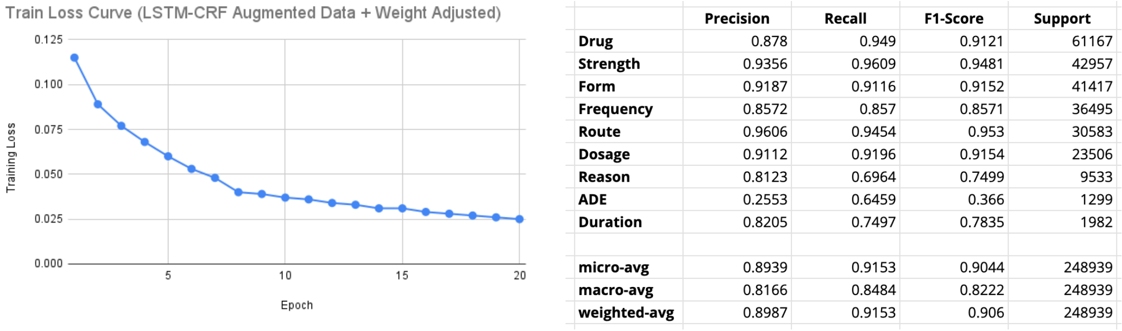
(a) Training Loss Curve

	Precision	Recall	F1-Score	Support
Drug	0.8785	0.9489	0.9124	61167
Strength	0.9333	0.9606	0.9468	42957
Form	0.9123	0.9108	0.9116	41417
Frequency	0.8539	0.8562	0.8551	36495
Route	0.9592	0.9429	0.951	30583
Dosage	0.9084	0.9172	0.9128	23506
Reason	0.8111	0.7028	0.7531	9533
ADE	0.2376	0.6474	0.3476	1299
Duration	0.8208	0.774	0.7967	1982
micro-avg	0.8907	0.9149	0.9026	248939
macro-avg	0.8128	0.8512	0.8208	248939
weighted-avg	0.8963	0.9149	0.9046	248939

(b) Test performance metrics

Figure 8: LSTM-CRF-Augmented plot and table for NER

The following train-loss curve and test performance metrics show the results of fine-tuning our LSTM model on the augmented data with additional weight adjustments.



(a) Training Loss Curve

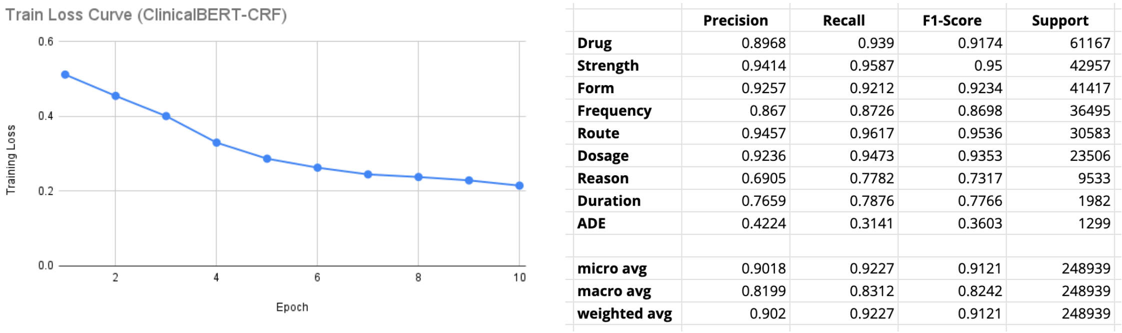
	Precision	Recall	F1-Score	Support
Drug	0.878	0.949	0.9121	61167
Strength	0.9356	0.9609	0.9481	42957
Form	0.9187	0.9116	0.9152	41417
Frequency	0.8572	0.857	0.8571	36495
Route	0.9606	0.9454	0.953	30583
Dosage	0.9112	0.9196	0.9154	23506
Reason	0.8123	0.6964	0.7499	9533
ADE	0.2553	0.6459	0.366	1299
Duration	0.8205	0.7497	0.7835	1982
micro-avg	0.8939	0.9153	0.9044	248939
macro-avg	0.8166	0.8484	0.8222	248939
weighted-avg	0.8987	0.9153	0.906	248939

(b) Test performance metrics

Figure 9: LSTM-CRF-Augmented-Weighted plot and table for NER

5.4.3 Transformer model

The following train-loss curve and test performance metrics show the results of training our transformer model on the NER task.

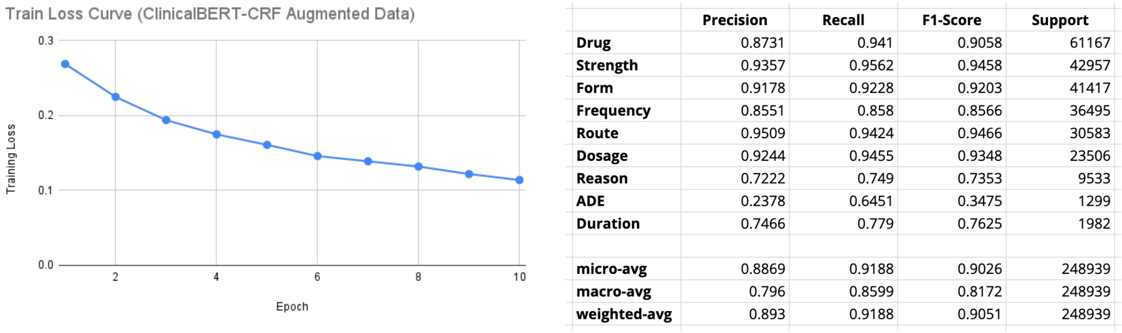


	Precision	Recall	F1-Score	Support
Drug	0.8968	0.939	0.9174	61167
Strength	0.9414	0.9587	0.95	42957
Form	0.9257	0.9212	0.9234	41417
Frequency	0.867	0.8726	0.8698	36495
Route	0.9457	0.9617	0.9536	30583
Dosage	0.9236	0.9473	0.9353	23506
Reason	0.6905	0.7782	0.7317	9533
Duration	0.7659	0.7876	0.7766	1982
ADE	0.4224	0.3141	0.3603	1299
micro avg	0.9018	0.9227	0.9121	248939
macro avg	0.8199	0.8312	0.8242	248939
weighted avg	0.902	0.9227	0.9121	248939

(a) Training Loss Curve (b) Test performance metrics

Figure 10: Transformer-CRF plot and table for NER

The following train-loss curve and test performance metrics show the results of fine-tuning our transformer model (from above) for an additional 10 epochs on the augmented data.

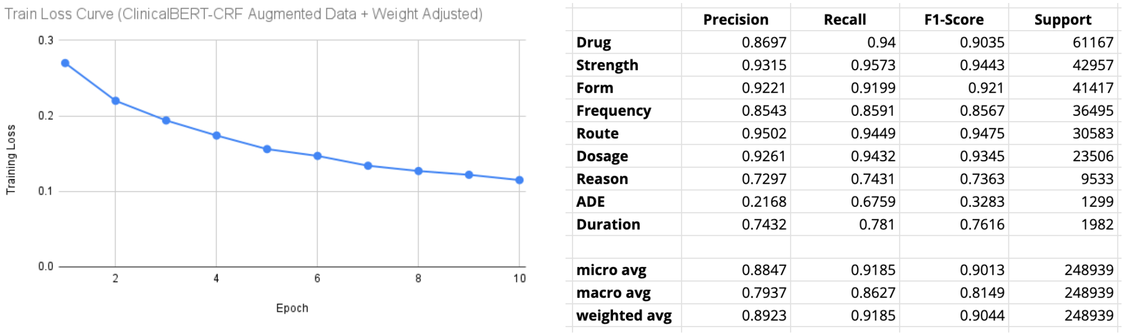


	Precision	Recall	F1-Score	Support
Drug	0.8731	0.941	0.9058	61167
Strength	0.9357	0.9562	0.9458	42957
Form	0.9178	0.9228	0.9203	41417
Frequency	0.8551	0.858	0.8566	36495
Route	0.9509	0.9424	0.9466	30583
Dosage	0.9244	0.9455	0.9348	23506
Reason	0.7222	0.749	0.7353	9533
ADE	0.2378	0.6451	0.3475	1299
Duration	0.7466	0.779	0.7625	1982
micro-avg	0.8869	0.9188	0.9026	248939
macro-avg	0.796	0.8599	0.8172	248939
weighted-avg	0.893	0.9188	0.9051	248939

(a) Training Loss Curve (b) Test performance metrics

Figure 11: Transformer-CRF-Augmented plot and table for NER

The following train-loss curve and test performance metrics show the results of fine-tuning our transformer model for an additional 10 epochs on the augmented data with additional weight adjustments.



	Precision	Recall	F1-Score	Support
Drug	0.8697	0.94	0.9035	61167
Strength	0.9315	0.9573	0.9443	42957
Form	0.9221	0.9199	0.921	41417
Frequency	0.8543	0.8591	0.8567	36495
Route	0.9502	0.9449	0.9475	30583
Dosage	0.9261	0.9432	0.9345	23506
Reason	0.7297	0.7431	0.7363	9533
ADE	0.2168	0.6759	0.3283	1299
Duration	0.7432	0.781	0.7616	1982
micro avg	0.8847	0.9185	0.9013	248939
macro avg	0.7937	0.8627	0.8149	248939
weighted avg	0.8923	0.9185	0.9044	248939

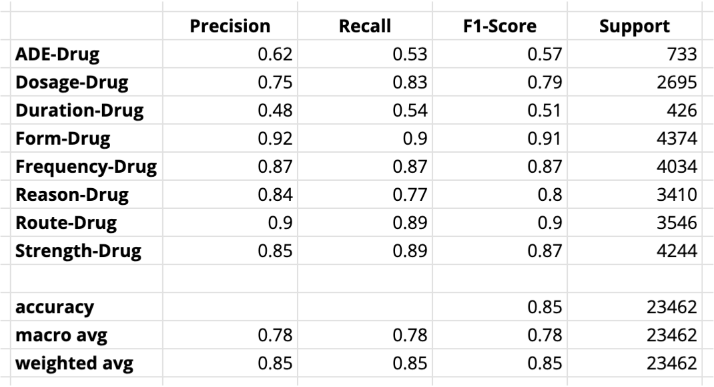
(a) Training Loss Curve (b) Test performance metrics

Figure 12: Transformer-CRF-Augmented-Weighted plot and table for NER

5.5 Relationship Extraction experiments

5.5.1 Baseline

The following train-loss curve and test performance metrics show the results of running our baseline model on the Relation Extraction task.

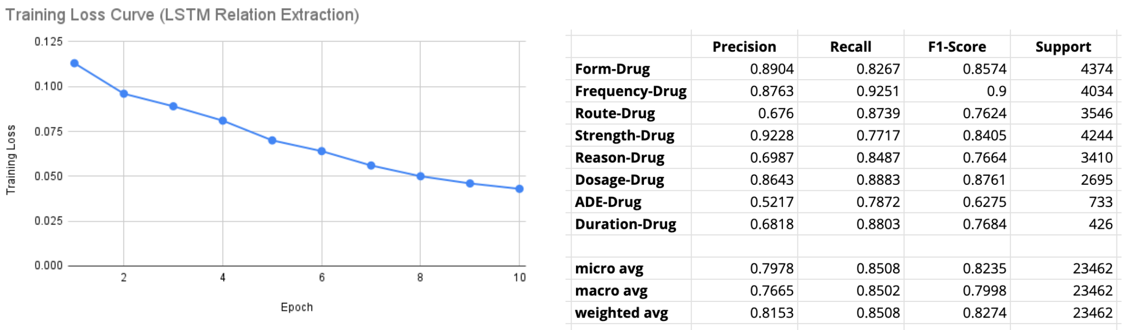


	Precision	Recall	F1-Score	Support
ADE-Drug	0.62	0.53	0.57	733
Dosage-Drug	0.75	0.83	0.79	2695
Duration-Drug	0.48	0.54	0.51	426
Form-Drug	0.92	0.9	0.91	4374
Frequency-Drug	0.87	0.87	0.87	4034
Reason-Drug	0.84	0.77	0.8	3410
Route-Drug	0.9	0.89	0.9	3546
Strength-Drug	0.85	0.89	0.87	4244
accuracy			0.85	23462
macro avg	0.78	0.78	0.78	23462
weighted avg	0.85	0.85	0.85	23462

Figure 13: Baseline table for Relation Extraction

5.5.2 LSTM model

The following train-loss curve and test performance metrics show the results of running our LSTM model on the Relation Extraction task.



	Precision	Recall	F1-Score	Support
Form-Drug	0.8904	0.8267	0.8574	4374
Frequency-Drug	0.8763	0.9251	0.9	4034
Route-Drug	0.676	0.8739	0.7624	3546
Strength-Drug	0.9228	0.7717	0.8405	4244
Reason-Drug	0.6987	0.8487	0.7664	3410
Dosage-Drug	0.8643	0.8883	0.8761	2695
ADE-Drug	0.5217	0.7872	0.6275	733
Duration-Drug	0.6818	0.8803	0.7684	426
micro avg	0.7978	0.8508	0.8235	23462
macro avg	0.7665	0.8502	0.7998	23462
weighted avg	0.8153	0.8508	0.8274	23462

(a) Training Loss Curve

(b) Test performance metrics

Figure 14: LSTM model plot and table for Relationship Extraction

6 Discussion

To evaluate the performance of our named entity recognition (NER) models, we conducted three experiments while keeping the architecture consistent. We compared the results of each experiment with the baseline. The first experiment utilized the entire raw data, while in the second experiment, we fine-tuned the model on ADE augmented data. In the third experiment, we focused on enhancing the model's ability to use context by giving more weight to reasons being predicted as ADEs.

For Relationship Extraction, we conducted two experiments. Our initial deep learning model provided satisfactory results, and we concentrated on extracting medical concepts from the UMLS ontology.

6.1 Findings

- All deep learning models demonstrate a significant improvement over machine learning baselines across all performance metrics, showcasing their superiority in the given tasks.
- Remarkably, the transformer model surpasses the LSTM model with 20 epochs in terms of micro and weighted average F1 scores, even after being trained for only 10 epochs. However, both models exhibit similar macro F1 scores, indicating comparable performance in capturing class-specific details.

- Our initial expectation was that data augmentation would improve the F1 scores of underrepresented classes. While we observed an increase in recall, it came at the expense of reduced precision, leading to an overall decrease in F1 score. Interestingly, the recall for reasons diminished, suggesting that the model might be confusing adverse drug events (ADEs) with reasons. Although this is a limitation, it should be noted that our model is not intended to serve as the sole decision-making authority in medical cases.

- To address the aforementioned issue, we adjusted the importance weights of ADEs and reasons in order to encourage the model to pay more attention to context. The outcome was not entirely as anticipated; neither deep learning model showed any improved performance.

- We also did not observe an overall performance improvement across both models with the data augmentation, ostensibly because we only augmented ADE data.

6.2 Significance

This work holds significant importance in the context of current research, as it contributes to the ongoing advancements in natural language processing and deep learning, particularly within the healthcare domain. The results of this study offer valuable insights into the performance of deep learning models, such as transformers and LSTMs, when compared to traditional machine learning techniques. By demonstrating the superiority of deep learning models in various performance metrics, this work underscores their potential for practical applications.

Furthermore, the study explores the effects of data augmentation and the adjustment of importance weights in addressing class imbalance and improving model performance. These findings can guide researchers and practitioners in refining their models for more accurate and reliable predictions, ultimately leading to better-informed decisions in medical settings.

The observed differences between transformer and LSTM models, as well as their varying sensitivity to changes in importance weights, highlight the need for further investigation into model architectures and optimization techniques. By identifying the limitations and strengths of each model, this work contributes to the broader understanding of how different models can be leveraged to address specific challenges in the healthcare domain. Ultimately, this research paves the way for future studies aiming to improve the quality and effectiveness of artificial intelligence applications in medicine and beyond.

6.3 Limitations and Ethical Considerations

In this study, we explore the limitations and shortcomings of ontology-aware deep learning models for medical concept extraction and relationship classification, including error cases, generalization concerns, and ethical considerations. Our models may face challenges generalizing to different data sources or medical sub-domains due to varying electronic health record (EHR) formats, language, and content. Additionally, the models might struggle with ambiguous or misspelled medical terms, complex syntactic structures, or novel abbreviations, leading to errors in concept extraction or relationship classification.

Ethical considerations include potential biases in training and testing data, which may impact clinical decision-making and patient care if unaddressed. Ensuring diverse, balanced, and representative data is crucial for mitigating these biases. Moreover, secure and privacy-preserving environments are necessary to prevent unauthorized access to sensitive medical information and maintain compliance with healthcare regulations. Lastly, our models may be susceptible to adversarial attacks, requiring the development of robust defenses to ensure their reliability and security in real-world applications.

6.4 Future Research Directions

In future research, we plan to explore different methods of incorporating Unified Medical Language System (UMLS) ontologies into our models. Leveraging the rich knowledge provided by UMLS could lead to enhanced understanding of medical concepts and improved extraction of relationships among them. This may involve experimenting with novel embedding techniques, knowledge graph integration, or other advanced techniques that can better utilize the information contained within the UMLS ontologies. Additionally, refining data parsing methods could significantly improve our model performance by more accurately and efficiently processing complex medical records and reducing the impact of noise or inconsistencies in the data.

Another direction for future work includes the joint training of Named Entity Recognition (NER) and Relation Extraction models, which could enable better end-to-end learning and facilitate the discovery of meaningful relationships among medical concepts. Furthermore, advanced loss functions could be explored to handle class imbalance more effectively and enhance the overall performance of underrepresented classes. Techniques such as focal loss, cost-sensitive learning, or adaptive loss weighting may prove beneficial in addressing the challenges posed by imbalanced datasets, thereby improving the model's accuracy and generalizability in medical concept extraction and relationship classification tasks.

7 Conclusions

In conclusion, this study demonstrates the superiority of deep learning models, specifically transformers and LSTMs, over traditional machine learning baselines across various performance metrics in the healthcare domain. The research also investigates the effects of data augmentation and importance weight adjustments on model performance, providing valuable insights for addressing class imbalance and improving model accuracy. Additionally, the observed differences between the transformer and LSTM models, as well as their distinct sensitivity to changes in importance weights, emphasize the need for further exploration into model architectures and optimization techniques. This work contributes to a broader understanding of the potential and challenges of artificial intelligence applications in medicine, paving the way for future studies aiming to enhance the quality and effectiveness of AI-driven decision-making in healthcare settings.

8 References

References

[Dev+18] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[HXY15] Zhiheng Huang, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF models for sequence tagging". In: *arXiv preprint arXiv:1508.01991* (2015).

[Lee+19] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4 (2019), pp. 1234–1240.

[Sav+10] Guergana K Savova et al. "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications". In: *Journal of the American Medical Informatics Association* 17.5 (2010), pp. 507–513.