Medical Concept Extraction and Relationship Classification from Patient Records

Team Name – ConceptMiners

Team Members:
- Meet Gandhi
- Rohan Saraogi
- Tarun Raheja

Contents

SECTION ONE

Problem Statement & Motivation

Problem Statement

To help clinicians by extracting medical concepts like drug, strength, dosage, duration, frequency, etc., from patient records and establish relations between them.

Overall, we are solving the task of Named Entity Recognition and Relation extraction from medical notes

Motivation

- There is a need for accurate and efficient extraction of medical concepts and relationship classification from patient records.
- Accurately identifying and classifying medical concepts and relationships within electronic health records is a challenge due to the lack of context and their complex nature.
- Our project aims to develop ontology-aware deep learning models that leverage the Unified Medical Language System (UMLS) to enhance the performance of medical concept extraction and relationship classification.
- The contributions of our project have the potential to improve clinical decision-making, patient care, and facilitate medical research.

Our goal is to maximize Recall for underrepresented entities like adverse drug events (ADEs) in our dataset, as our models are intended to provide support to clinicians rather than make final decisions.

SECTION TWO

Dataset

Dataset

- We are using Harvard Medical School's "n2c2 adverse drug events (ADE) and medication extraction in the electronic health records" dataset for our project.
- The dataset includes 303 de-identified medical records for training and 202 for testing from the MIMIC-III database.
- The dataset contains annotation files with domain expert annotated entity tags and relationship tags for every medical record text file.
- The entity tags include drug, strength, dosage, duration, frequency, form, route, reason, and ADE tags, while the relationship tags include strength-drug, dosage-drug, duration-drug, frequency-drug, form-drug, route-drug, reason-drug, and ADE-drug relationships tags.

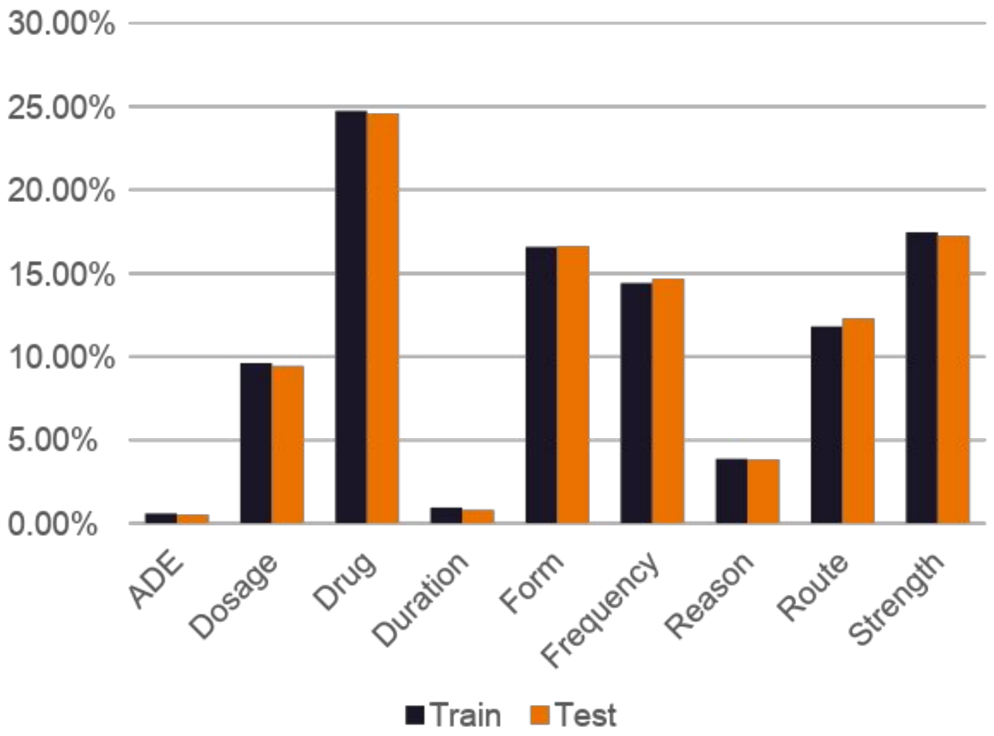
The raw text files provided have very long patient records, which must be parsed in a suitable format for our deep-learning models

Data Parsing Approach

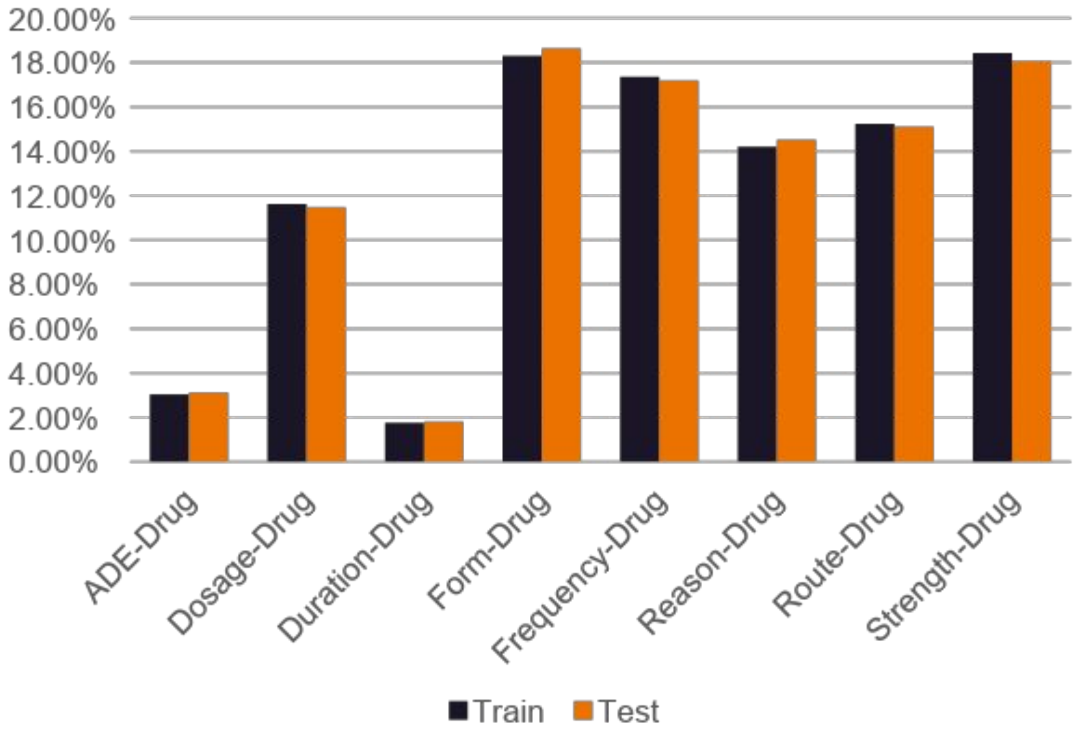
- We extract text spans from patient records that contain at least two entities and an accompanying relation between the entity pairs.
- The start and end of the spans are determined by the nearest full-stop or line break before the first entity and after the last entity.
- For entities that do not have a relationship with other entities, we extract text spans that cover that single entity.
- We use the spans containing entities and relations for both NER and relation extraction tasks.
- Spans with only entities are solely utilized for the NER task.
- Finally, we identify spans containing different relationships but identical text due to the presence of more than two entities and merge them.

Class distribution

NER – Train/Test Distribution



Relation – Train/Test Distribution



SECTION THREE

Methodology

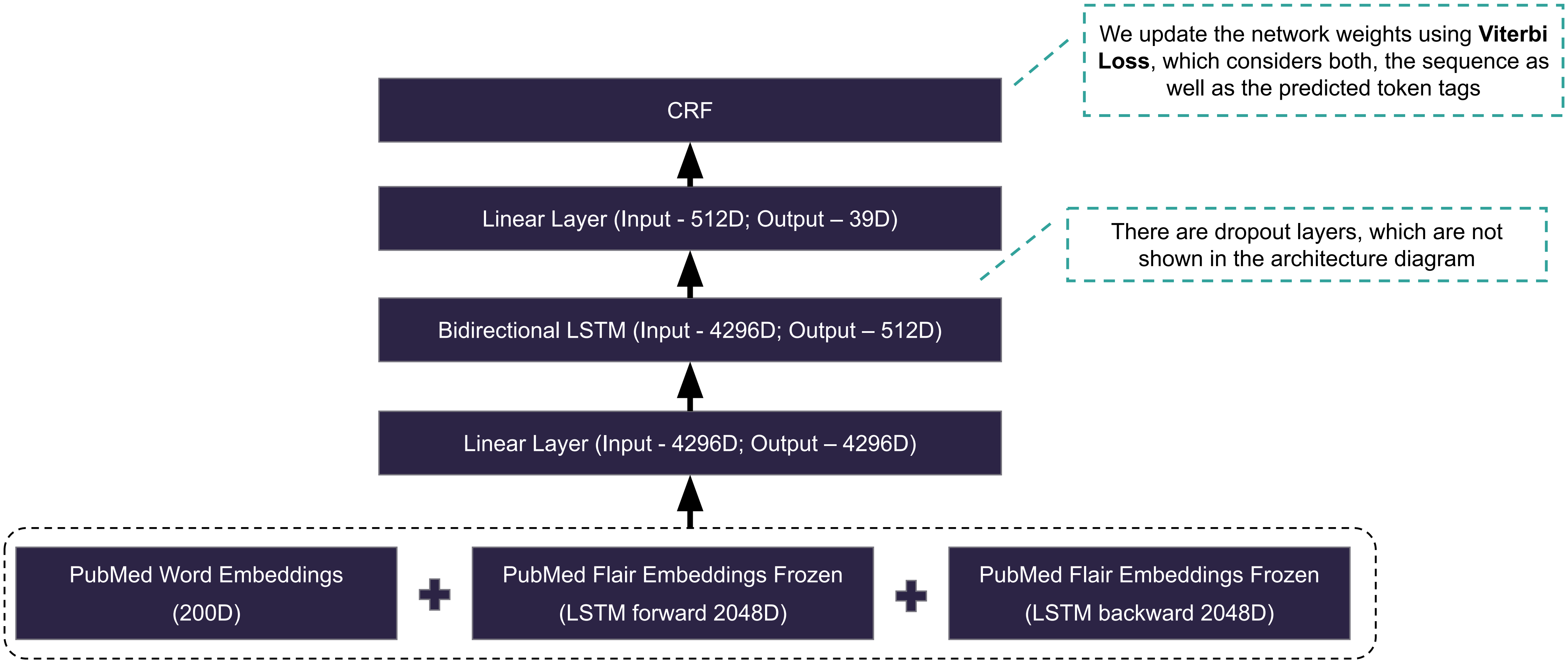
Named Entity Recognition (NER): Baseline

- Our baseline NER model employs token-level features, such as POS tags and case information, to capture contextual information within a window of two tokens for medical concept extraction.
- Then we fit a Logistic Regression model on these features for token-level classification to identify the medical concepts.

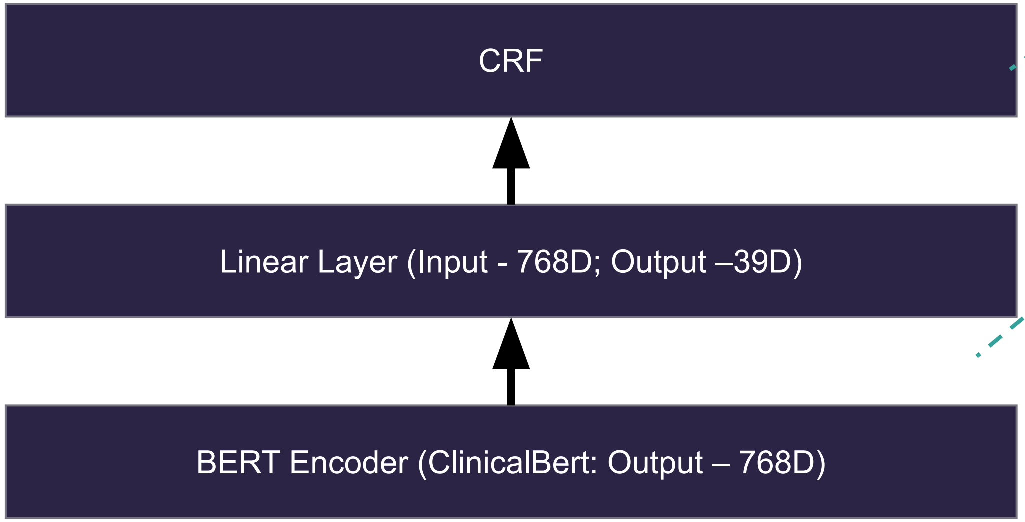
Feature	**Description**
pos tag	Part-of-speech tag of the token
istitle	True if the token is in title case, otherwise False
isupper	True if the token is in uppercase, otherwise False
isalpha	True if the token consists only of alphabetic characters, otherwise False
isnumeric	True if the token consists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

Named Entity Recognition (NER): LSTM-CRF



Named Entity Recognition (NER): Transformer-CRF



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

UMLS Data Augmentation for NER: ADE focused

For a given entity,

- Selected top Concept Unique Identifier (CUI) (assuming it is above a score threshold)
- Selected an alias for the CUI based on certain conditions
- Replaced the entity with the alias in the text

Augmentation Example

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

→ UMLS Ontology

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

Relation Extraction: Baseline

- For relationship classification, we extract features using the TF-IDF representation of the shortest dependency paths between the entities.
- Then we fit a Logistic Regression on these representation to classify relationship between the entity pairs.

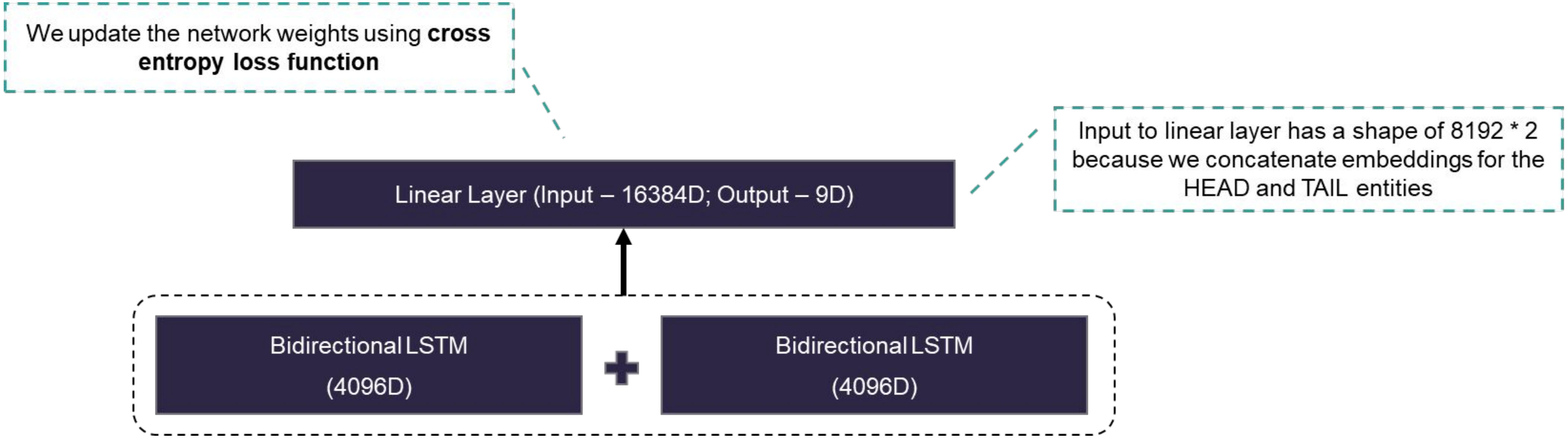
Shortest Dependency Path Example

She was started on prophylactic Oxacillin to cover skin flora , and Dermatology was consulted along with Neurology and Ophthalmology for the ophthalmic involvement

→ Dependency Parsed

ophthalmic for consulted started on Oxacillin

Relation Extraction: LSTM Model



The raw text is modified with special tokens to help the model understand the relationships better. For example:
⟨S:PERSON⟩Bill⟨/S:PERSON⟩was born in⟨O:CITY⟩Seattle⟨/O:CITY⟩

SECTION FOUR

Results

NER Results: Trained on raw data

Baseline

	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

LSTM

	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

Transformer

	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

NER Results: Trained on raw data + finetuned on UMLS augmented data for ADE

LSTM

	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

Transformer

	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

NER Results: Trained on raw data + finetuned on UMLS augmented data for ADE + Weights adjusted

LSTM

	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

Transformer

	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

Relation Extraction Results

Baseline

	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

LSTM

	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

SECTION FIVE

Conclusion

Conclusion

- Deep learning models significantly outperform machine learning baselines in all performance metrics.
- Transformer model outshines LSTM (20 epochs) with only 10 epochs of training in micro and weighted average F1 scores; macro F1 scores are similar for both models.
- Data augmentation increases recall for underrepresented classes but reduces precision, leading to a decrease in overall F1 scores; model potentially confuses ADEs with reasons.
- Forcing model to focus more on context by adjusting importance weights of ADEs and reasons doesn't yield anticipated improvement in deep learning models' performance.
- No overall performance improvement observed with data augmentation, possibly due to augmenting only ADE data.

SECTION SIX

Future Work, Limitations & Ethical Considerations

Limitations & Ethical Consideration

- Our models may have limitations in generalizing to different data sources or medical sub-domains.
- We could not train our deep learning models due to compute constraints on Google Colab.
- Ensuring secure and privacy-preserving environments is critical to prevent unauthorized access to sensitive medical information.
- Developing robust defenses against adversarial attacks is essential to ensure the reliability and security of our models in real-world applications.

Future Work

- Explore different ways of incorporating UMLS ontologies.
- Better data parsing could implement our model performance.
- Joint training of NER and Relation Extraction models
- Advanced loss functions could be explored to handle class imbalance and improve overall performance of underrepresented classes

Thank you!