CS7650 Final Project

Identifying adverse drug reactions by using BERT models for classification and NER

Divij Mishra
Aravind Rajeev Nair
Naveen Sethuraman

Click for Presentation Video

Introduction

- Key application of NLP - auto-extracting information from medical docs
- ADR (Adverse Drug Reaction) identification - classic problem in NLP + healthcare
- Tasks:
  - 1 - Does a given sentence contain an ADR? (binary classification)
  - 2 - If so, identify both the drug and adverse reaction by name. (NER)
- Methods:
  - Fine-tune BERT, BioBERT, BioClinicalBERT.

Prior work

- Fei et al, 2021 tested multiple NLP methods on many biomedical information extraction datasets, including our ADR dataset - achieve 84% F1-score on Task 2 using BioBERT
- Cabot and Navigli, 2021 devised a novel seq-2-seq method for relation extraction tasks - achieve 82% F1-score on ADR Task 2
- Lots of work going into developing new datasets as well - Luo et al, 2022 released BioRED, a larger biomedical NER/RE dataset
- More broadly, Murphy et al, 2023 survey literature for NLP techniques employed for ADR detection

Model: BERT + Variants

- Bidirectional Encoder Representation from Transformers [Devlin et al., 2019]
  - Transformer-based pre-trained language model
  - Generates contextualized word embeddings for sequences
- Use HuggingFace libraries
- Task 1- Binary Classification
  - Add a linear layer with 2 output nodes
- Task 2 - NER
  - Add a linear layer with 5 output nodes (2 entities = 5 labels under BIO scheme)

Model: BERT variants

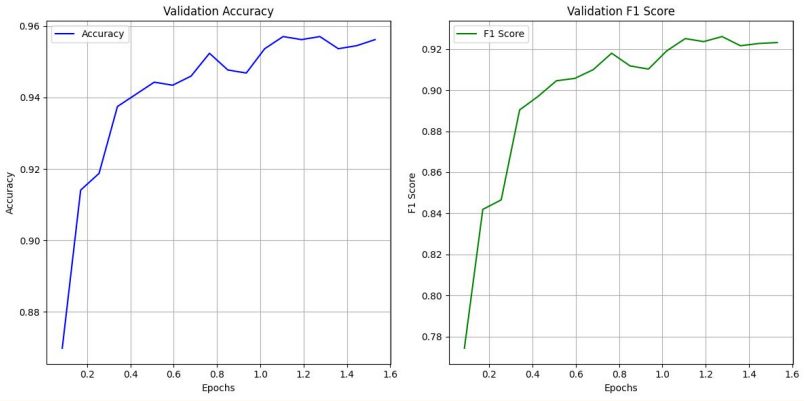
- Can improve BERT performance on specific domains by modifying pre-training/fine-tuning corpus!
- BioBERT [Lee et al, 2019] is pre-trained on a biomedical corpus.
- Bio-ClinicalBERT [Alsentzer et al, 2019] is a BERT model initialized with BioBERT weights -> fine-tuned on a clinical records corpus

Dataset

- Adverse Drug Event benchmark dataset [Gurulingappa et al., 2012]
- We use the HuggingFace version [ade_corpus_v2]
- Task 1 - sentence classification:
  - Contains 23.5k samples labelled for binary classification
- Task 2 - drug + effect identification:
  - Contains 6.8k samples
  - Given as a relation extraction problem - since each sentence is annotated for exactly one drug and one effect, can consider it to be an NER problem.
- For both tasks, used an 80-10-10 split

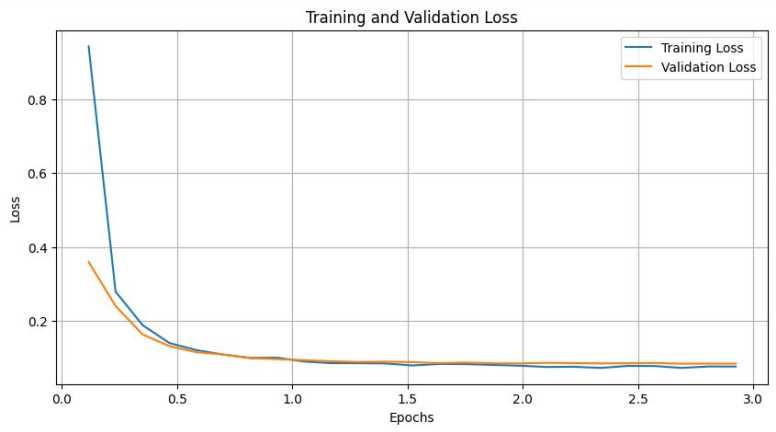
Experiments - Task 1: Binary classification

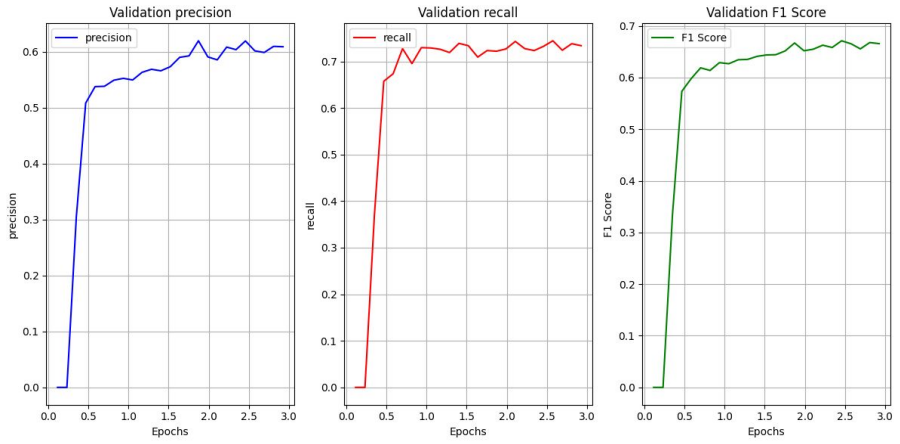
- We fine-tuned BERT, BioBERT, and BioClinicalBERT on the ADR classification task, for 2 epochs, with early stopping, batch size = 16, and evaluation steps = 100. (plots shown for BioBERT)



Experiments - Task 2: NER

- We fine-tuned BERT, BioBERT, and BioClinicalBERT on the ADR NER task, for 3 epochs, with early stopping, batch size = 32, and evaluation steps = 20. (plots shown for BioBERT)





Results

- For Task 1: Binary Classification, we saw that BioBERT performed the best (91% F1-score), with BioClinicalBERT giving comparable performance. BERT performed much worse than the other two.

Model	F1-score	Precision	Recall
BERT	0.7605	0.6939	0.8413
BioBERT	0.911	0.8947	0.927
BioClinBERT	0.8939	0.8868	0.9010

- For Task 2: NER, all three performed comparably. BioBERT performed the best (65% F1-score).

Model	F1-score	Precision	Recall
BERT	0.6332	0.5789	0.6987
BioBERT	0.6552	0.602	0.7187
BioClinBERT	0.6437	0.5804	0.7225

Conclusion

- Fine-tuned Transformer-based models show good results on ADR detection!
  - Effective, low data-requirement method
- Pre-training also matters
  - BioBERT performs significantly better than BERT.
  - However, BioClinicalBERT shows no improvement over BioBERT, surprising!
- Room for improvement - previous work hit 84% score on Task 2 - we likely need to do more hyperparameter tuning, train for longer with higher regularization.