Enhancing Pharmacovigilance with Drug Reviews and Social Media

Brent Biseda
University of California, Berkeley
brentbiseda@berkeley.edu

Katie Mo
University of California, Berkeley
kmo@berkeley.edu

Berkeley
UNIVERSITY OF CALIFORNIA

Definition

phar·ma·co·vig·i·lance
noun

The practice of monitoring the effects of medical drugs after they have been licensed for use, especially in order to identify and evaluate previously unreported adverse reactions.

Berkeley
UNIVERSITY OF CALIFORNIA

Introduction

Relying on current mechanisms alone can result in underreporting of adverse drug reactions (ADRs)

Can we leverage the quantity and expediency of drug reviews and social media to enhance pharmacovigilance?

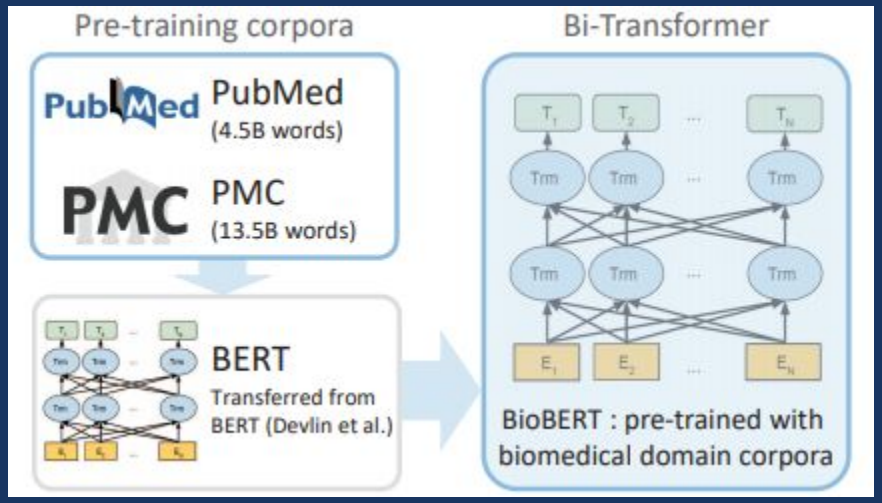
Tasks and Approach

1. Sentiment classification using Drugs.com drug reviews
2. Presence of ADR classification using Twitter data
3. Name entity recognition (NER) detection of ADRs using a subset of the Twitter data

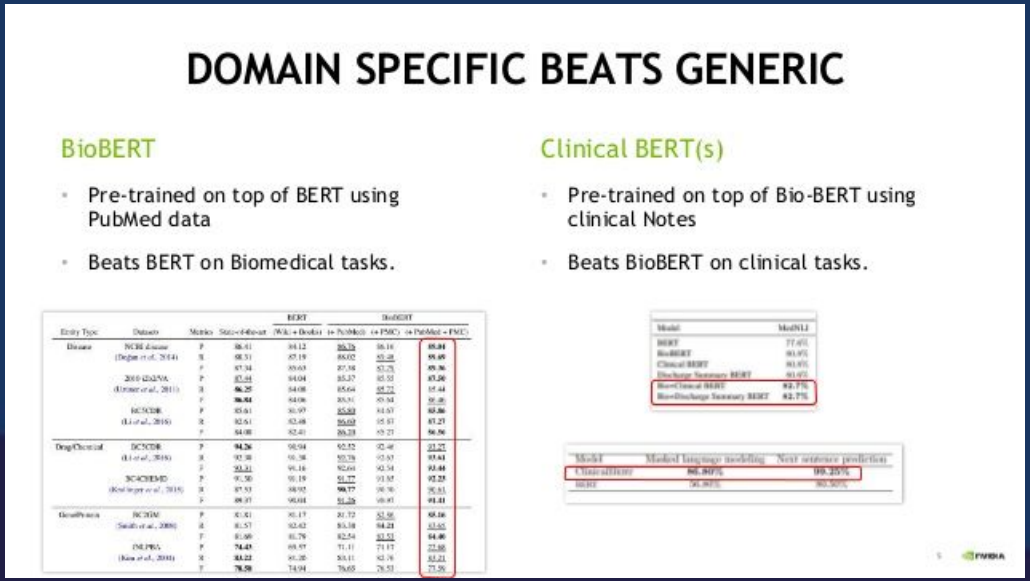
Used 8 variants of BERT:

BERT Cased (B-C)	Clinical BERT All Notes (CB-A)
BERT Uncased (B-U)	Clinical BERT Discharge (CB-D)
BioBERT 1.0 (BB-1.0)	Clinical BioBERT All Notes (CBB-A)
BioBERT 1.1 (BB-1.1)	Clinical BioBERT Discharge (CBB-D)

BERT Fine-tuning with Biomedical Corpora



BioBERT



Clinical BERT

Datasets

Drug reviews
N = 215,063

User provided a rating of the drug from 1–10

Binned into 3 classes:
60% negative (rating 1–3)
18% neutral (rating 4–7)
22% positive (rating 8–10)

Twitter
N = 4,169

Annotations included whether the tweet had mentioned an ADR or not

Highly imbalanced dataset with only 11% positives
Created oversampled and undersampled datasets

A subset was used for NER of ADRs with BIO labels (N = 965)

Berkeley
UNIVERSITY OF CALIFORNIA

Results | Sentiment Classification

Baseline Test Accuracies

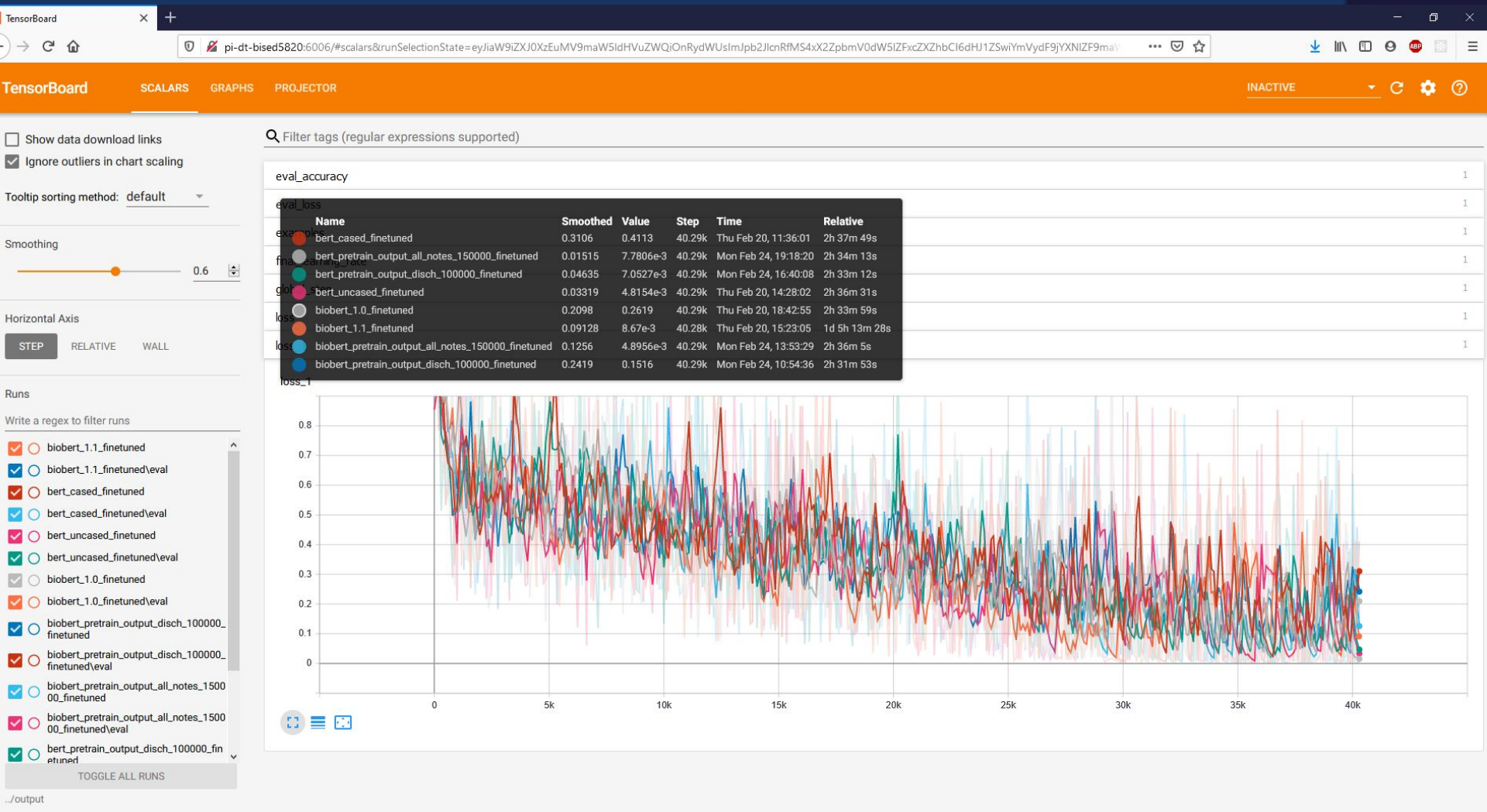
Model	Accuracy
Most Common	0.602
N-Gram + NB	0.890
ELMo + LR	0.709
Pretrained B-C + LR	0.720

BERT Models Test Accuracies

Model	1 Epoch	2 Epochs	3 Epochs	4 Epochs	10 Epochs
B-C	0.824	0.851	0.876	0.888	-
B-U	0.805	0.820	0.824	0.841	-
BB-1.0	0.824	0.854	0.877	0.887	-
BB-1.1	0.824	0.854	0.877	0.877	-
CB-A	0.821	0.854	0.877	0.888	-
CB-D	0.824	0.855	0.874	**0.889**	**0.906**
CBB-A	0.822	0.855	0.873	**0.889**	-
CBB-D	0.823	0.855	0.876	0.888	-

Berkeley
UNIVERSITY OF CALIFORNIA

BERT Fine-tuning Tensorboard



Results | ADR Classification

Test F-Scores

Model	Imbalanced	Oversampled	Undersampled
Most Common	0	0	0
N-Gram + NB	0.197	0.324	0.408
B-C	0.570	0.464	0.487
B-U	**0.590**	0.476	0.546
CBB-D	0.544	0.523	0.510
B-U features + LR	0.562	0.463	0.786
B-U features + CNN	0.655	0.720	0.951
B-U features + LSTM	0.978	**0.995**	**0.995**

Berkeley
UNIVERSITY OF CALIFORNIA

Results | NER of ADRs

Baseline Test F-Scores

Model	F-Score
Most Common Class	0.324
CRF	0.502

BERT Models Test F-Scores

Model	3 Epochs	5 Epochs	10 Epochs
B-C	0.652	0.687	0.687
B-U	0.549	0.684	**0.720**
BB-1.0	0.489	0.663	0.696
BB-1.1	0.521	0.661	0.652
CB-A	0.546	0.641	0.662
CB-D	0.546	0.685	0.681
CBB-A	0.602	0.619	0.646
CBB-D	0.556	0.647	0.649

Berkeley
UNIVERSITY OF CALIFORNIA

Conclusions

- No superior performance from BioBERT or Clinical BERT in comparison to regular BERT
- Fine-tuning for a larger number of epochs had better performance for sentiment classification and NER
- Use of an additional classifier on top of BERT extracted features improved performance, especially when the dataset is limited in size for ADR classification

Questions or Comments?

Thank You!

Berkeley
UNIVERSITY OF CALIFORNIA