

# Enhancing Pharmacovigilance with Drug Reviews and Social Media

Brent Biseda  
University of California, Berkeley  
[brentbiseda@berkeley.edu](mailto:brentbiseda@berkeley.edu)

Katie Mo  
University of California, Berkeley  
[kmo@berkeley.edu](mailto:kmo@berkeley.edu)

# 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.

# 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?

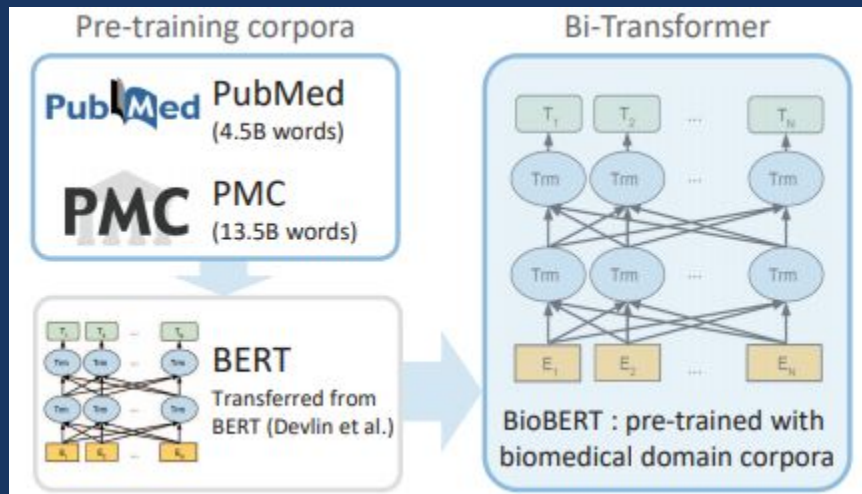
# Tasks and Approach

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

### DOMAIN SPECIFIC BEATS GENERIC

#### BioBERT

- Pre-trained on top of BERT using PubMed data
- Beats BERT on Biomedical tasks.

#### Clinical BERT(s)

- Pre-trained on top of Bio-BERT using clinical Notes
- Beats BioBERT on clinical tasks.

Entity Type	Dataset	Metrics	BERT				BioBERT			
			Macro	State-of-the-art	(Wiki + BioRxiv)	(+ PubMed + PMC)	Macro	State-of-the-art	(Wiki + BioRxiv)	(+ PubMed + PMC)
Disease	NCBI disease (Duggan et al., 2014)	F	86.51	84.12	85.19	85.28	86.84	86.84	86.84	86.84
		R	87.51	85.19	86.02	86.28	87.01	86.84	86.84	86.84
		P	87.51	85.19	86.02	86.28	87.01	86.84	86.84	86.84
	2019 G2A7A (Liu et al., 2019)	F	85.28	84.04	85.27	85.52	85.50	85.50	85.50	85.50
		R	86.28	84.08	85.64	85.72	85.74	85.74	85.74	85.74
ICD9CM		F	86.81	84.08	85.51	85.52	86.81	86.81	86.81	86.81
		R	85.81	83.97	85.82	85.87	85.86	85.86	85.86	85.86
		P	85.81	83.97	85.82	85.87	85.86	85.86	85.86	85.86
		F	85.81	83.97	85.82	85.87	85.86	85.86	85.86	85.86
		R	85.81	83.97	85.82	85.87	85.86	85.86	85.86	85.86
Drug/Chemical	DCDDB (Li et al., 2018)	F	94.26	90.94	92.32	92.46	94.27	94.27	94.27	94.27
		R	92.38	90.38	92.38	92.38	93.43	93.43	93.43	93.43
		P	92.38	90.38	92.38	92.38	93.43	93.43	93.43	93.43
	SCADDBD (Kohler et al., 2015)	F	92.38	90.38	92.38	92.38	93.43	93.43	93.43	93.43
		R	92.38	90.38	92.38	92.38	93.43	93.43	93.43	93.43
Genetics	BC2000 (Smith et al., 2008)	F	81.41	80.17	81.72	81.72	81.41	81.41	81.41	81.41
		R	81.57	80.42	81.58	81.58	81.55	81.55	81.55	81.55
		P	81.58	80.42	81.58	81.58	81.55	81.55	81.55	81.55
	NSLPPS (Kuo et al., 2010)	F	74.49	68.57	71.11	71.11	74.49	74.49	74.49	74.49
		R	84.22	81.20	84.11	84.11	84.21	84.21	84.21	84.21

Model	Macro F1
BERT	77.0%
BioBERT	80.0%
Clinical BERT	80.0%
Discharge Summary BERT	80.0%
Bio-Clinical BERT	82.7%
Bio-Discharge Summary BERT	82.7%

Model	Medical language modeling	Next sentence prediction
ClinicalBERT	86.80%	88.25%
BERT	76.80%	86.50%

## 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)

# Results | Sentiment Classification

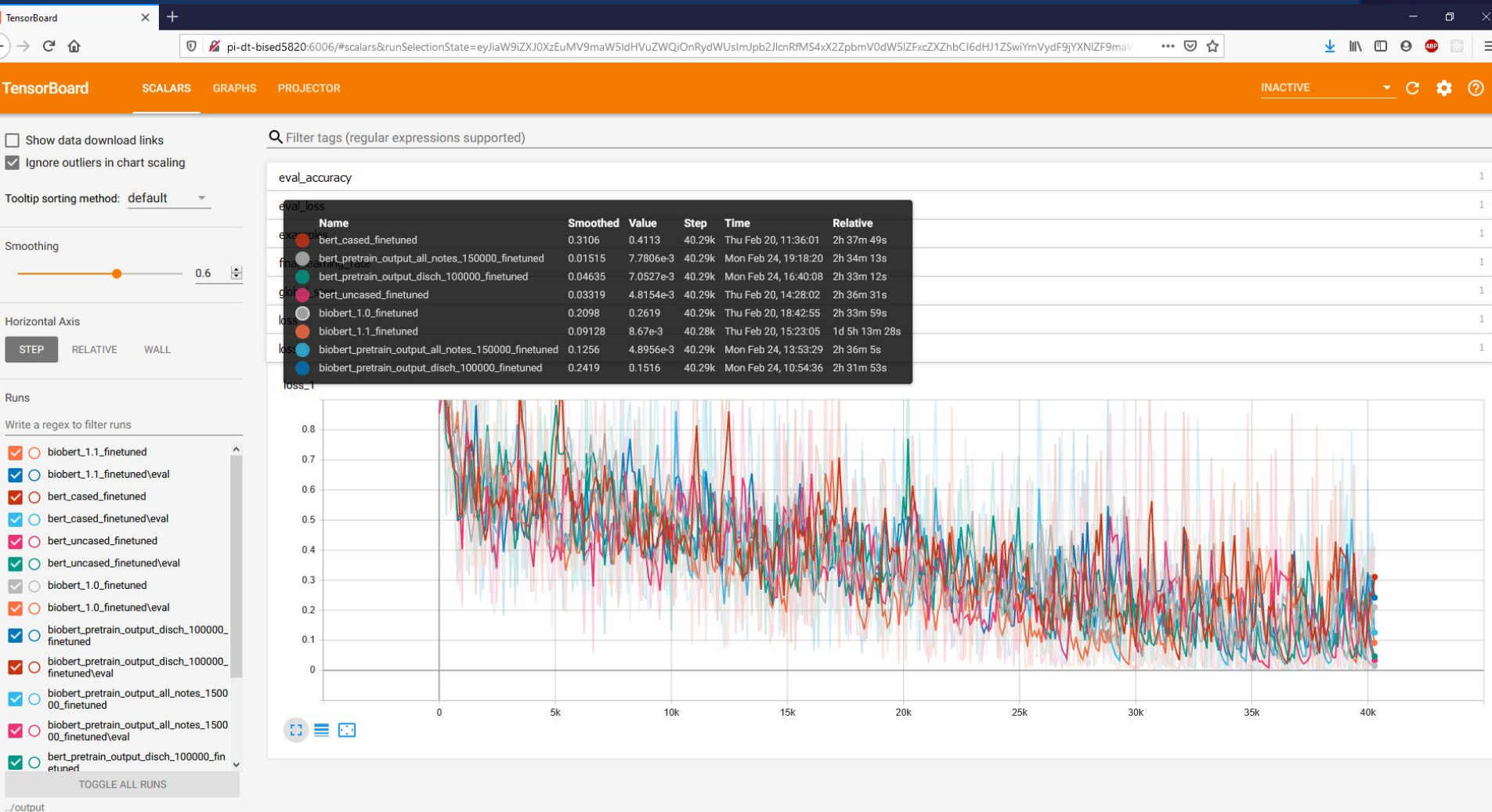
## Baseline Test Accuracies

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	-

# 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

# 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

# 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!