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

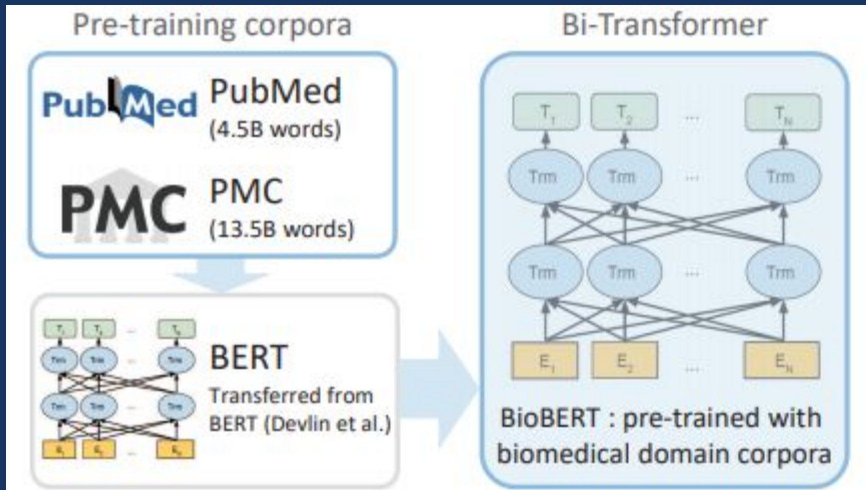
1. Sentiment classification using Drugs.com drug reviews
2. Presence of ADR classification using Twitter data
3. Name entity recognition (NER) detection of ADRs using a subset of the Twitter data

Used 8 variants of BERT:

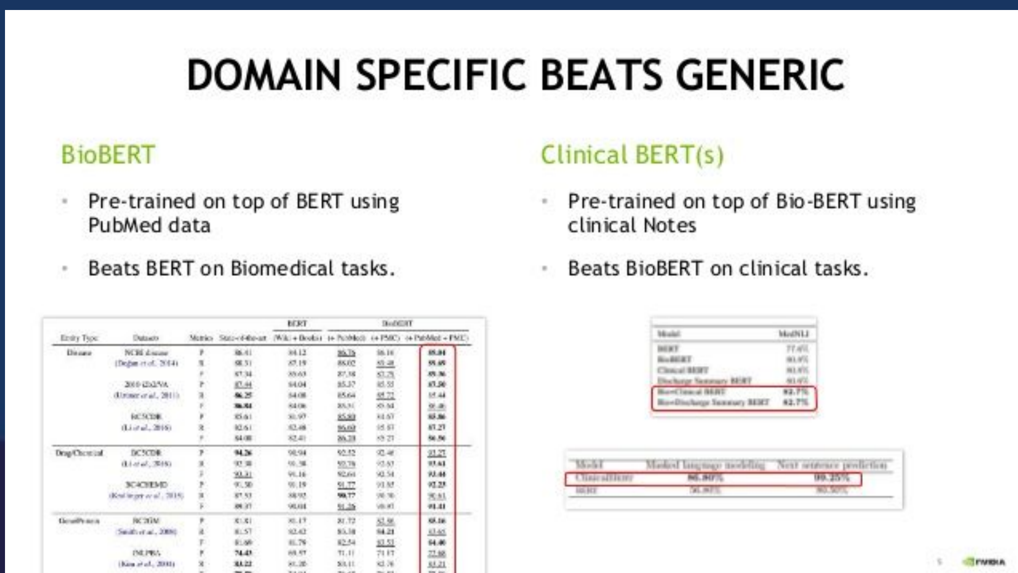
BERT Cased (B-C)	Clinical BERT All Notes (CB-A)
BERT Uncased (B-U)	Clinical BERT Discharge (CB-D)
BioBERT 1.0 (BB-1.0)	Clinical BioBERT All Notes (CBB-A)
BioBERT 1.1 (BB-1.1)	Clinical BioBERT Discharge (CBB-D)

Berkeley
UNIVERSITY OF CALIFORNIA

BERT Fine-tuning with Biomedical Corpora



BioBERT



Clinical BERT

Berkeley
UNIVERSITY OF CALIFORNIA

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

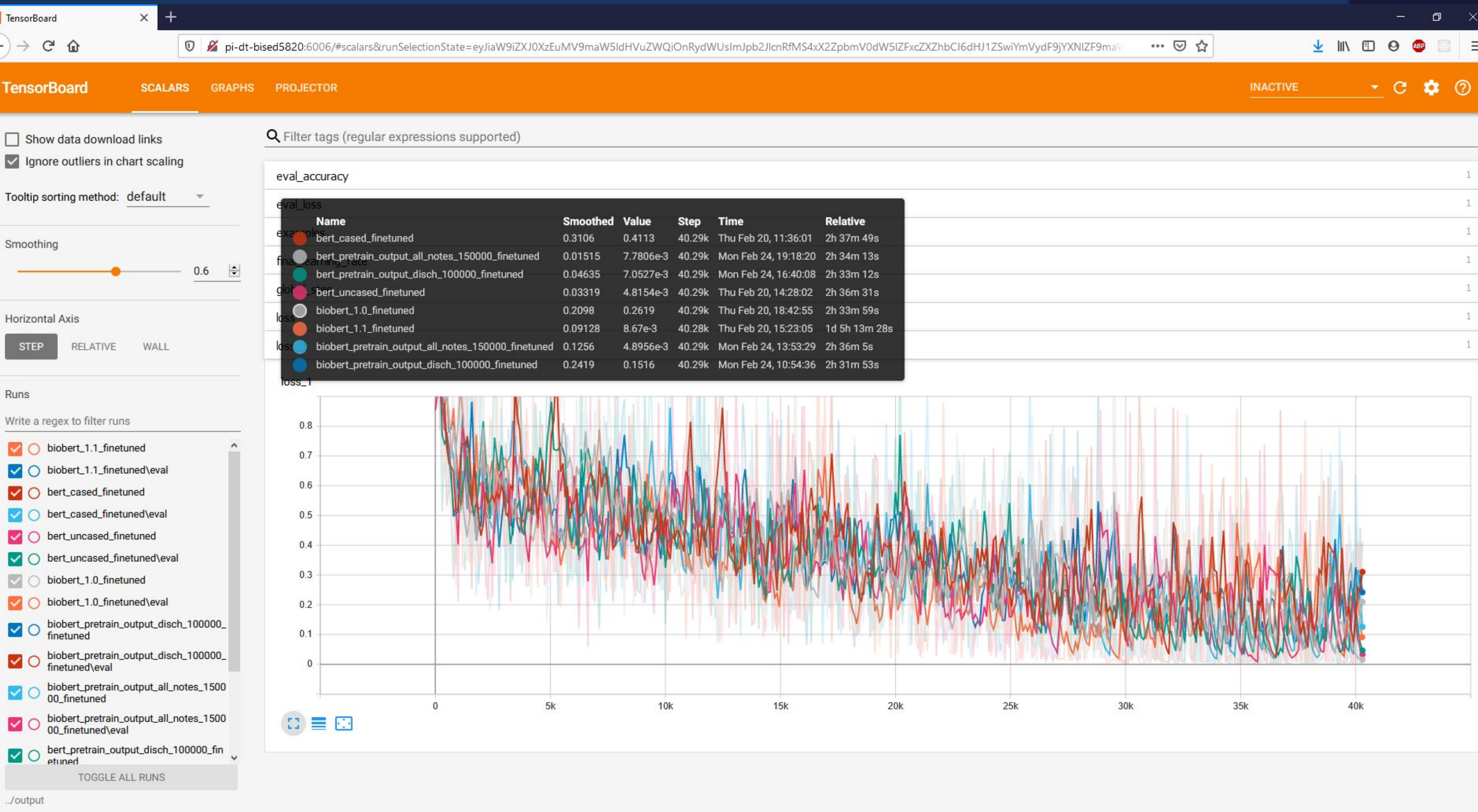
Model	Accuracy
Most Common	0.602
N-Gram + NB	0.890
ELMo + LR	0.709
Pretrained B-C + LR	0.720

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

Backup Slides

Sentiment Task Creates Positive, Negative, and Neutral Classes

Process the Training File

In [28]:
```
filename = "drugsComTrain_raw.tsv"
POSITIVE_THRESHOLD = 8
NEGATIVE_THRESHOLD = 3
```

In [29]:
```
df = pd.read_csv('../datasets/drugsCom_raw/'+filename, delimiter='\t')
```

In [30]:
```
# Remove Beginning and Ending Quotation Marks
df['review'] = df['review'].apply(lambda x: x[1:-1])
```

In [31]:
```
def makeSentimentScore(x):
    if x >= POSITIVE_THRESHOLD:
        return 1
    elif x <= NEGATIVE_THRESHOLD:
        return -1
    else:
        return 0
```

In [32]:
```
# Convert ratings to Positive and Negative Sentiment
df['rating'] = df['rating'].apply(makeSentimentScore)
```

In [39]:
```
df
```

Out[39]:

	sentence	label
0	It has no side effect, I take it in combinatio...	1
1	My son is halfway through his fourth week of I...	1
2	I used to take another oral contraceptive, whi...	0
3	This is my first time using any form of birth ...	1
4	Suboxone has completely turned my life around....	1
...	...	...
161292	I wrote my first report in Mid-October of 2014...	1
161293	I was given this in IV before surgey. I immedi...	-1
161294	Limited improvement after 4 months, developed ...	-1
161295	I've been on thyroid medication 49 years,...	1
161296	I've had chronic constipation all my adul...	1

161297 rows × 2 columns

Berkeley
UNIVERSITY OF CALIFORNIA

ADR Task Detects Presence of ADRs in Twitter Text

In [51]: df

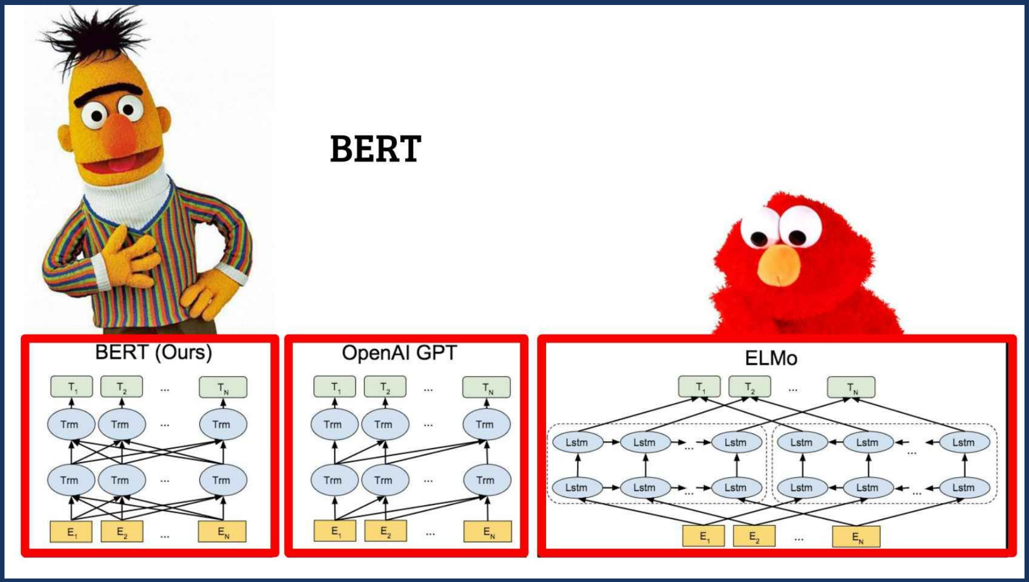
Out[51]:

	sentence	label
0	@DoctorChristian scared to start fluoxetine, w...	0
1	@IntuitiveGal1 ok, if you stopped taking the L...	0
2	Novartis announces secukinumab (AIN457) demons...	0
3	"U wailed all night; now y'r disembodied sobbi...	1
4	@irapaps you're so fucking selfish. I've got L...	0
...	...	...
7391	@RachaelHerron @marrije I haaaated Trazodone. ...	1
7392	"do you have any medication allergies?" "ASTHM...	1
7393	@ctr1945 my mom used Rivaroxaban and it gave h...	1
7394	@LilythePurr I've had too much quetiapine toda...	1
7395	Why do I always get the seroquel munchies!?	1

7396 rows × 2 columns

Berkeley
UNIVERSITY OF CALIFORNIA

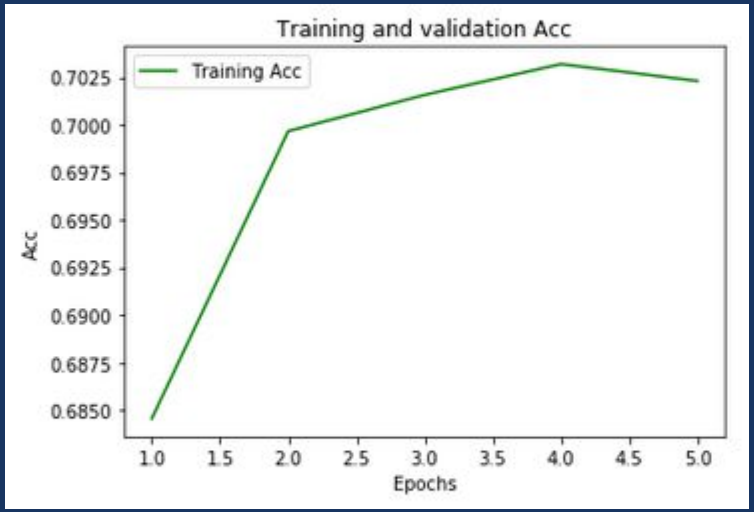
Baseline Comparison for Sentiment Task



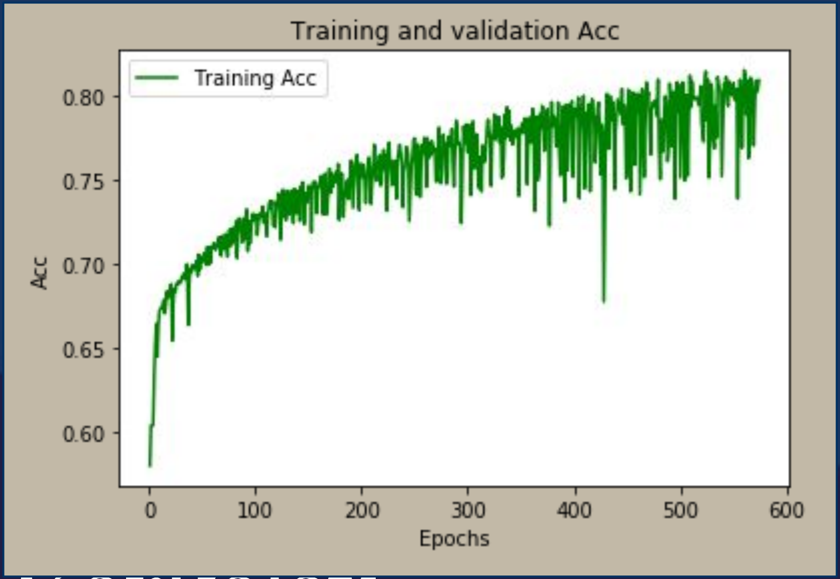
ELMo with Logistic Regression

BERT with Logistic Regression

Baseline Comparison for Sentiment Task



ELMo with Logistic Regression



BERT with Logistic Regression

Berkeley
UNIVERSITY OF CALIFORNIA

Results | Sentiment Analysis Task

Model	Test Accuracy
Most Common Class	0.602
Naive Bayes	
Elmo with Logistic Regression	0.709
Bert Cased with Logistic Reg	0.720

Baseline Comparison

Berkeley
UNIVERSITY OF CALIFORNIA

Results | Sentiment Analysis Task

Test Set Accuracy	1 Epoch	2 Epochs	3 Epochs	4 Epochs
Bert Cased	0.440	0.421	0.448	0.510
Bert Un-Cased	0.483	0.496	0.628	0.737
Biobert 1.0	0.444	0.415	0.445	0.504
Biobert 1.1	0.442	0.416	0.448	0.492
Clinical Bert (All Notes)	0.445	0.415	0.459	0.515
Clinical Bert (Discharge)	0.444	0.414	0.456	0.529
Clinical Biobert (All Notes)	0.446	0.419	0.452	0.521
Clincial Biobert (Discharge)	0.444	0.416	0.460	0.530

Berkeley
UNIVERSITY OF CALIFORNIA

Results | Sentiment Analysis Task

Test Set Loss	1 Epoch	2 Epochs	3 Epochs	4 Epochs
Bert Cased	0.824	0.851	0.876	0.888
Bert Un-Cased	0.805	0.820	0.824	0.841
Biobert 1.0	0.824	0.854	0.877	0.887
Biobert 1.1	0.824	0.854	0.877	0.877
Clinical Bert (All Notes)	0.821	0.854	0.877	0.888
Clinical Bert (Discharge)	0.824	0.855	0.874	0.889
Clinical Biobert (All Notes)	0.822	0.855	0.873	0.889
Clincial Biobert (Discharge)	0.823	0.855	0.876	0.888

Berkeley
UNIVERSITY OF CALIFORNIA

Results | ADR Task

Baseline Comparison

Model	Test Accuracy
Most Common Class	0.500
Naive Bayes	
ELMo with Logistic Regression	0.829
BERT Cased with Logistic Reg	0.827

Berkeley
UNIVERSITY OF CALIFORNIA

Results | ADR Task

Test Accuracy	1 Epoch	2 Epochs	3 Epochs	4 Epochs
BERT Cased	0.890	0.918	0.890	0.770
BERT Un-Cased	0.909	0.909	0.909	0.908
BioBERT 1.0	0.869	0.904	0.905	0.906
BioBERT 1.1	0.911	0.906	0.910	0.909
Clinical BERT (All Notes)	0.903	0.912	0.914	0.912
Clinical BERT (Discharge)	0.890	0.909	0.912	0.915
Clinical BioBERT (All Notes)	0.906	0.914	0.916	0.918
Clincial BioBERT (Discharge)	0.915	0.917	0.918	0.916

Results | ADR Task

Model	1 Epoch	2 Epochs	3 Epochs	4 Epochs
BERT Cased	0.679	0.491	0.692	0.536
BERT Un-Cased	0.348	0.558	0.662	0.678
BioBERT 1.0	0.368	0.452	0.508	0.557
BioBERT 1.1	0.427	0.583	0.724	0.771
Clinical BERT (All Notes)	0.440	0.608	0.655	0.710
Clinical BERT (Discharge)	0.331	0.429	0.502	0.552
Clinical BioBERT (All Notes)	0.419	0.595	0.729	0.751
Clincial BioBERT (Discharge)	0.357	0.574	0.631	0.699

Berkeley
UNIVERSITY OF CALIFORNIA

Results | NER Task

Baseline Comparison

Model	Test Accuracy
Most Common Class	0.938
Naive Bayes	
ELMo with Logistic Regression	0.938
BERT Cased with Logistic Reg	0.948

Berkeley
UNIVERSITY OF CALIFORNIA

Results | NER Task

Test Accuacy	1 Epoch	2 Epochs	3 Epochs	4 Epochs
BERT Cased	0.990	0.991	0.991	0.991
BERT Un-Cased	0.987	0.991	0.991	0.991
BioBERT 1.0	0.990	0.991	0.991	0.991
BioBERT 1.1	0.990	0.991	0.991	0.991
Clinical BERT (All Notes)	0.990	0.991	0.991	0.992
Clinical BERT (Discharge)	0.991	0.991	0.991	0.991
Clinical BioBERT (All Notes)	0.991	0.991	0.991	0.992
Clincial BioBERT (Discharge)	0.990	0.991	0.991	0.991

Results | NER Task

Test Loss	1 Epoch	2 Epochs	3 Epochs	4 Epochs
BERT Cased	0.056	0.038	0.041	0.029
BERT Un-Cased	0.062	0.039	0.038	0.041
BioBERT 1.0	0.049	0.037	0.037	0.031
BioBERT 1.1	0.056	0.045	0.042	0.034
Clinical BERT (All Notes)	0.053	0.036	0.037	0.025
Clinical BERT (Discharge)	0.039	0.037	0.036	0.029
Clinical BioBERT (All Notes)	0.050	0.042	0.038	0.031
Clincial BioBERT (Discharge)	0.050	0.037	0.038	0.031

Berkeley
UNIVERSITY OF CALIFORNIA