Work summary of 12th week

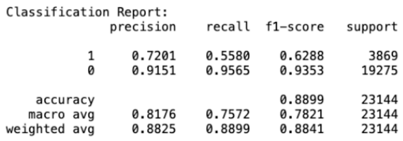
Student: Xuanyu Su
Supervisor: Isar Nejadgholi

November 18, 2020

1 RoBERTa model construction with Attack Comment data

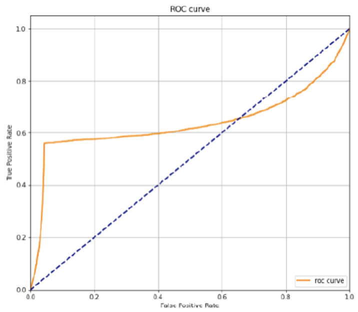
By observing the results of the previously established model, we found that the Attack Comment data has better results on multiple databases, and by combining the results in week 9, we found that the accuracy of the RoBERTa model built using Toxicity data have better results than the model built using BERT. Therefore, we guess that using Attack Comment data may achieve better results than Toxicity data when we train on the RoBERTa model. In order to verify our conjecture, in this module, we use Attack Comment data to establish a RoBERTa model, and the parameters still use the previous settings. The only difference is that in this version, we removed the last fine-tuning module (in view of the previous model, after adding the fine-tuning module, the model did not get better improvement).

1.1 In-domain test results



```
Classification Report:
              precision    recall  f1-score   support

           1     0.7201    0.5580    0.6288      3869
           0     0.9151    0.9565    0.9353     19275

    accuracy                         0.8899     23144
   macro avg     0.8176    0.7572    0.7821     23144
weighted avg     0.8825    0.8899    0.8841     23144
```

The classification report of in-domain test on Attack Comment data

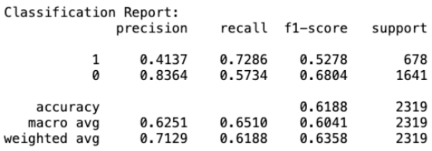


The ROC curve of in-domain test on Attack Comment data

Figure 1: The classification report and ROC curve of in-domain test on Attack Comment data

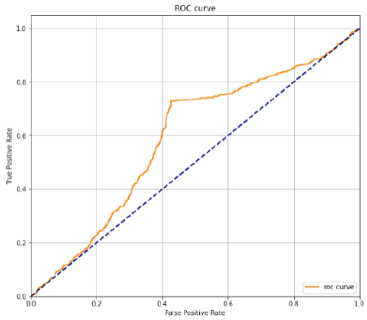
1.2 Cross test results

1. Test on 2400 Covid data.



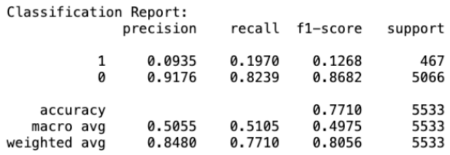
```
Classification Report:
              precision    recall  f1-score   support

           1     0.4137    0.7286    0.5278       678
           0     0.8364    0.5734    0.6804      1641

    accuracy                         0.6188      2319
   macro avg     0.6251    0.6510    0.6041      2319
weighted avg     0.7129    0.6188    0.6358      2319
```

The classification report of testing on 2400 Covid data



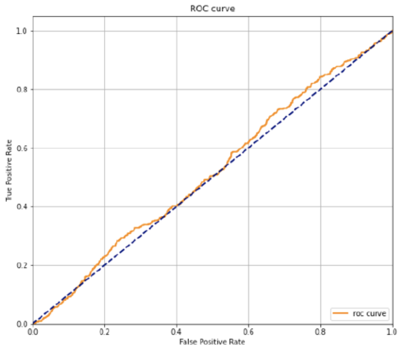
The ROC curve of testing on 2400 Covid data

Figure 2: The classification report and ROC curve of testing on 2400 Covid data



```
Classification Report:
              precision    recall  f1-score   support

           1     0.0935    0.1970    0.1268       467
           0     0.9176    0.8239    0.8682      5066

    accuracy                         0.7710      5533
   macro avg     0.5055    0.5105    0.4975      5533
weighted avg     0.8480    0.7710    0.8056      5533
```

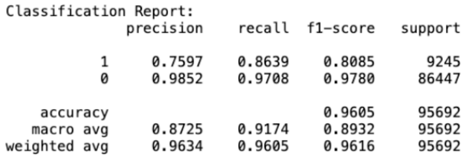
The classification report of testing on Gab data



The ROC curve of testing on Gab data

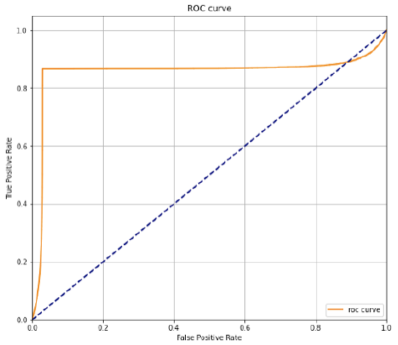
Figure 3: The classification report and ROC curve of testing on gab data

2. Test on Gab data.
3. Test on Toxicity data.



```
Classification Report:
              precision    recall  f1-score   support

           1     0.7597    0.8639    0.8085      9245
           0     0.9852    0.9708    0.9780     86447

    accuracy                         0.9605     95692
   macro avg     0.8725    0.9174    0.8932     95692
weighted avg     0.9634    0.9605    0.9616     95692
```

The classification report of testing on Toxicity data



The ROC curve of testing on Toxicity data

Figure 4: The classification report and ROC curve of testing on toxicity data

1.3 Summary

Through the observation of the above results, we found that after training the model with the attack comment data, only the result of **Gab** data has been improved slightly, the results of the remaining databases did not exceed the results tested on the previous model. (final results comparison table see in the appendix.)

2 Results comparison

In this module we make comparison among three distance graph(RoBERTa model trained on Toxicity, Gab and Attack Comment data). Figure 5,6,7 for distance comparison, figure 8,9,10 after normalization.

3 Conslusion

We can observe from the above six distance graphs: in a same model, the distribution trend of the different test data sets tend to be similar to the data of the training model. For example, in figure 5, because the model is a RoBERTa model trained with Toxicity data, the distribution trend of the Toxicity data is leftward, thus we can observe in this picture that the other three databases all show similar distribution trends, but Due to the different distribution ranges of positive and negative data pairs, the results are superior and inferior.

In the same way, we can observe from figure 7 that since the model is trained through Attack Comment data, and the dataset has a steeper distribution trend, in this version of the distance comparison, other data sets have the similar distribution trends.

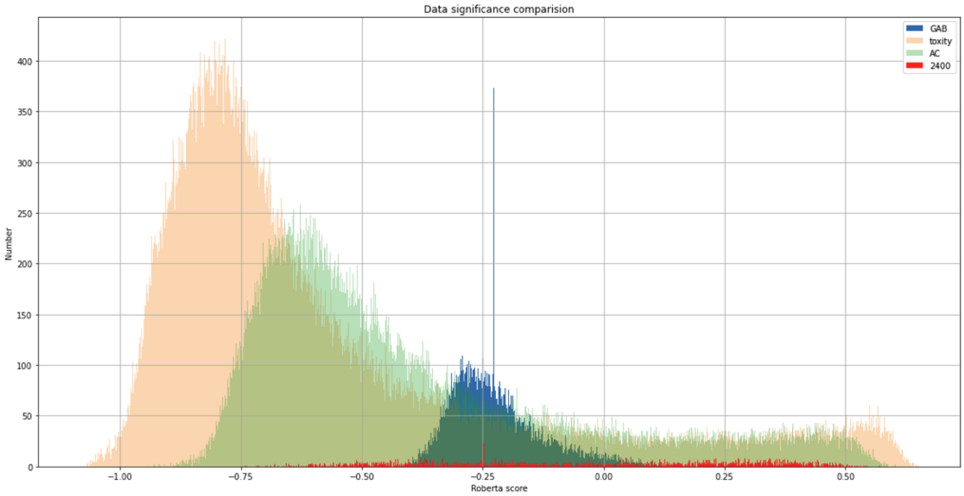


Figure 5: The distance comparison of RoBERTa model trained with Toxcitiy data

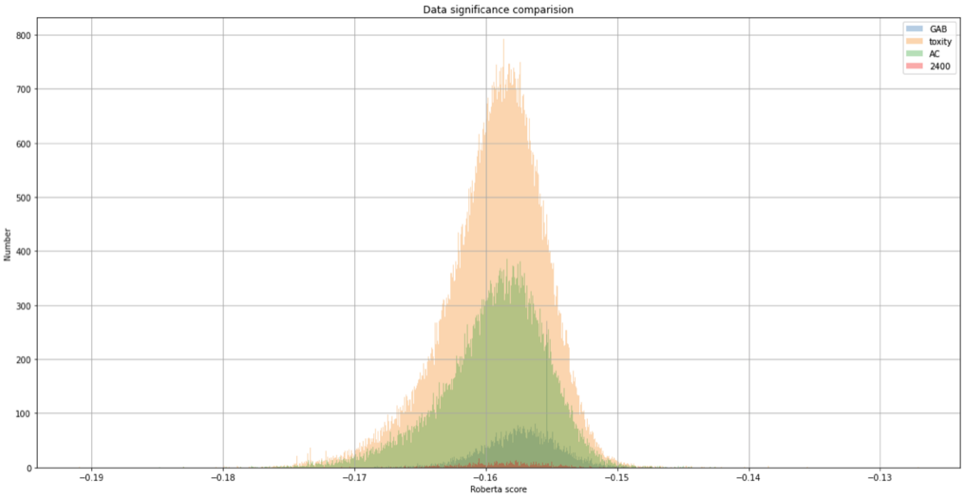


Figure 6: The distance comparison of RoBERTa model trained with Gab data

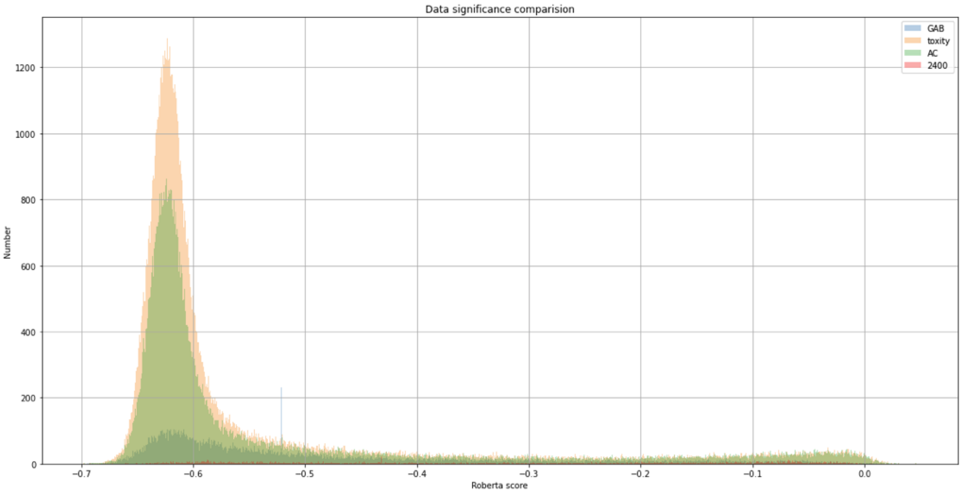


Figure 7: The distance comparison of RoBERTa model trained with Attack Comment data

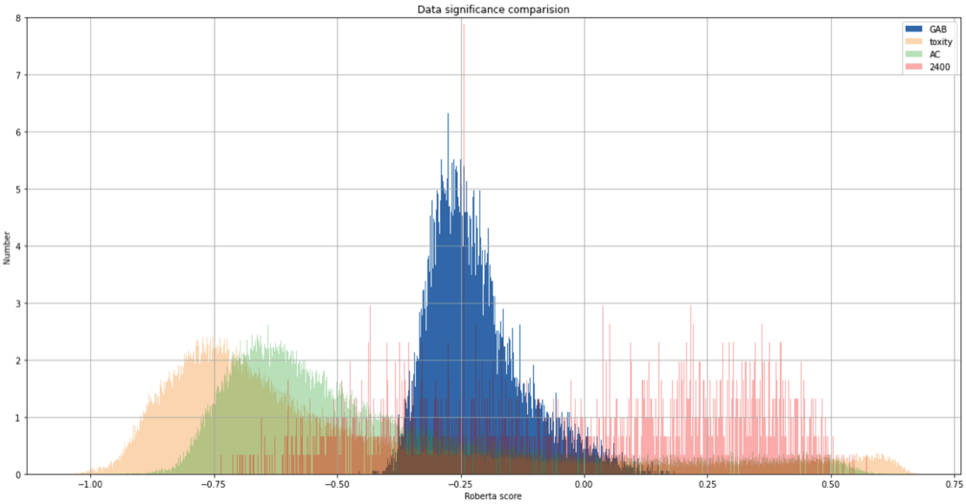


Figure 8: The distance comparison of RoBERTa model trained with Toxcitiy data after normalization

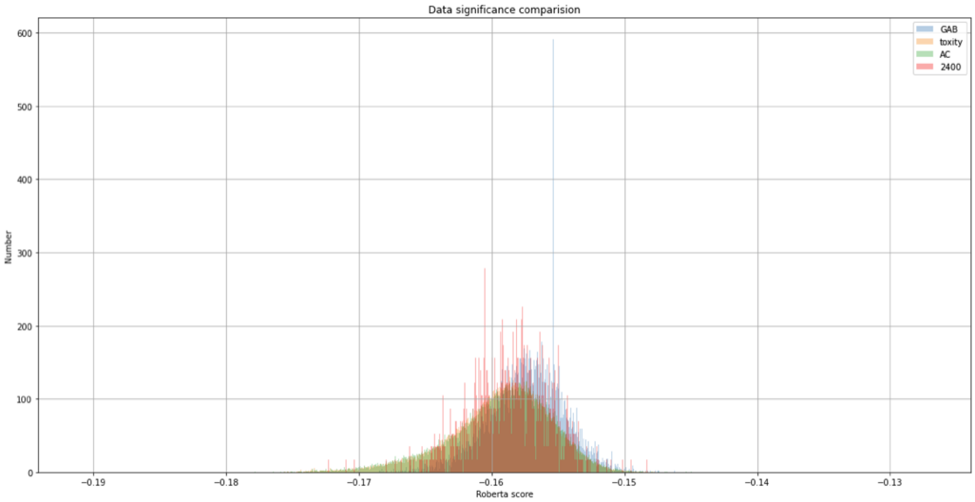


Figure 9: The distance comparison of RoBERTa model trained with Gab data after normalization

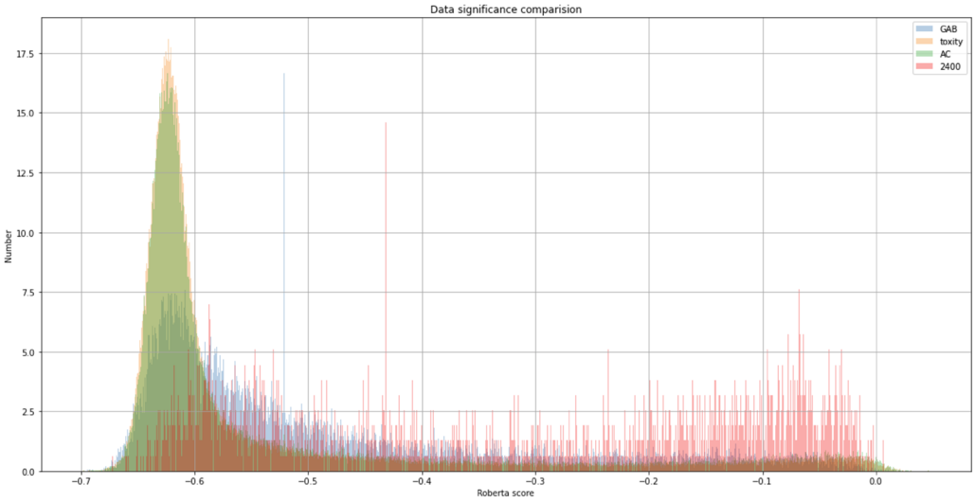
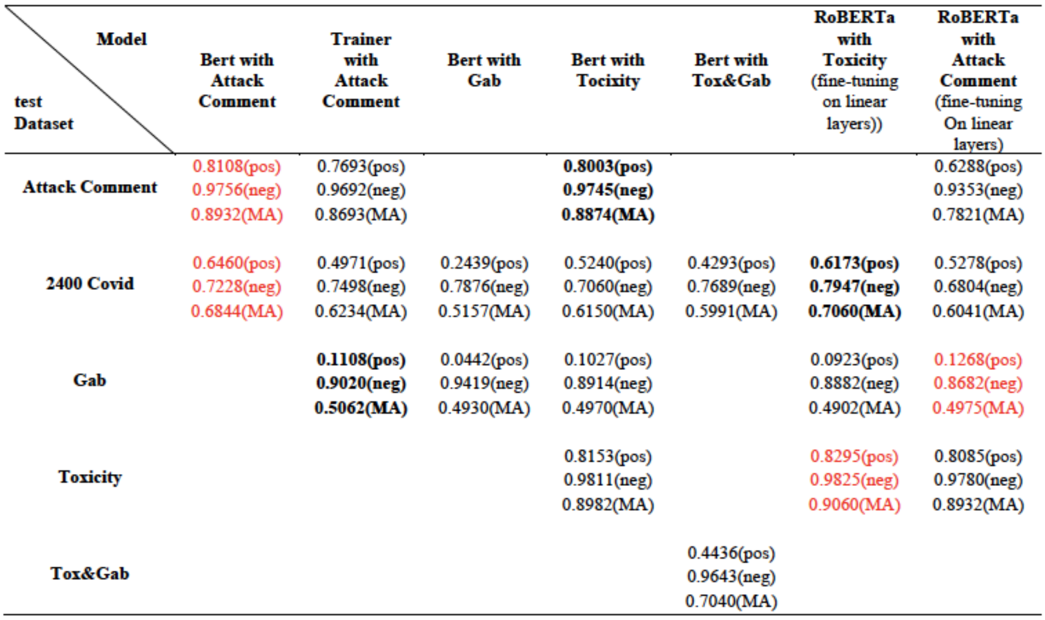


Figure 10: The distance comparison of RoBERTa model trained with Attack Comment data after normalization

4 Appendix

4.1 The results of testing different data sets on different models



Model / test Dataset	Bert with Attack Comment	Trainer with Attack Comment	Bert with Gab	Bert with Tocixity	Bert with Tox&Gab	RoBERTa with Toxicity (fine-tuning on linear layers))	RoBERTa with Attack Comment (fine-tuning On linear layers)

Attack Comment	0.8108(pos)
	0.9756(neg)
	0.8932(MA)
		0.7693(pos)
		0.9692(neg)
		0.8693(MA)
				0.8003(pos)
				0.9745(neg)
				0.8874(MA)
							0.6288(pos)
							0.9353(neg)
							0.7821(MA)

2400 Covid	0.6460(pos)
	0.7228(neg)
	0.6844(MA)
		0.4971(pos)
		0.7498(neg)
		0.6234(MA)
			0.2439(pos)
			0.7876(neg)
			0.5157(MA)
				0.5240(pos)
				0.7060(neg)
				0.6150(MA)
					0.4293(pos)
					0.7689(neg)
					0.5991(MA)
						0.6173(pos)
						0.7947(neg)
						0.7060(MA)
							0.5278(pos)
							0.6804(neg)
							0.6041(MA)

Gab		**0.1108(pos)**
		0.9020(neg)
		0.5062(MA)
			0.0442(pos)
			0.9419(neg)
			0.4930(MA)
				0.1027(pos)
				0.8914(neg)
				0.4970(MA)
						0.0923(pos)
						0.8882(neg)
						0.4902(MA)
							0.1268(pos)
							0.8682(neg)
							0.4975(MA)

Toxicity				0.8153(pos)
				0.9811(neg)
				0.8982(MA)
						0.8295(pos)
						0.9825(neg)
						0.9060(MA)
							0.8085(pos)
							0.9780(neg)
							0.8932(MA)

Tox&Gab					0.4436(pos)
					0.9643(neg)
					0.7040(MA)		

Figure 11: Comparison of the results different databases on different models

Note: The red font indicates: the best result of one database among different models, the bold font indicates the second best result.

4.2 Analyse of the table

Combining the results of different databases tested on different models, we guess that the reason **why Attack Comment data performs better on the Bert model** is that although Attack Comment is similar with Toxicity database, since Toxicity database has longer sentence length and more extensive range of data, when testing other databases (2400 and Gab), it is less possible to predict the sub-intervals more accurately. On the contrary, Attack Comment data has a smaller data distribution range and shorter sentences length than Toxicity, thus it could be more accurately adapt and predict other databases. (The above conjecture is limited to Attack Comment and Toxicity data sets, since the scope of their two databases is broader than other databases.)

But why the performance of tox on roberta is better? Through our previous comparison of the Bert and RoBERTa models, we found that the RoBERTa model allows larger data input, which means that the RoBERTa model is better for larger databases. From the figure5 above, we could know that the data scale of the Toxicity database is the largest data in all databases, which also confirms our conjecture from another aspect.