Work summary for the 7^{th} and 8^{th} week

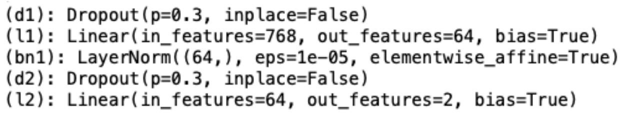
Student: Xuanyu Su
Supervisor: Isar Nejadgholi

October 26, 2020

1 RoBERTa model construction with Toxicity data

1.1 Model use and parameters

In this version of the model construction, we use the **RoBERTa** as the main model: still use the **fields** and **TabularDataset** methods to do data embedding process. Then we added **two linear layers** at the end according to the characteristics of the data to be tested in this experiment (2400 Covid data). Therefore, our model training is divided into two modules: **pretrain** and **train**. The **pretrain** is designed to initialize the parameters of the linear layers (otherwise the parameters will be initialized randomly, which is not conducive to the optimization of the final result). The **train** module is the overall training of the Toxicity data on the RoBERTa and linear layers. Set the optimizer to **AdamW**, and **weight-decay** is used to prevent overfitting.
We added two extra linear layers and there is a corresponding dropout layer for each linear layer to prevent overfitting. The added layers are as figure 1:



```
(d1): Dropout(p=0.3, inplace=False)
(l1): Linear(in_features=768, out_features=64, bias=True)
(bn1): LayerNorm((64,), eps=1e-05, elementwise_affine=True)
(d2): Dropout(p=0.3, inplace=False)
(l2): Linear(in_features=64, out_features=2, bias=True)
```

Figure 1: The structure of added layers

1.2 Results

After the RoBERTa model had been trained on Toxicity and Gab data, we test 'Gab' and 2400 Covid data on this model, then compare the results of the experiment with the previous results. The loss curve as shown in figure 2:

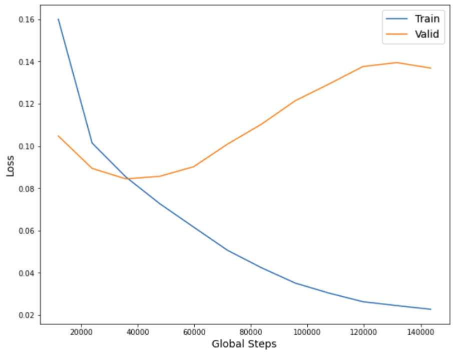
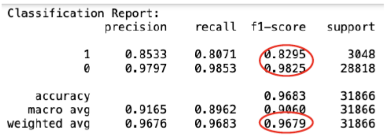
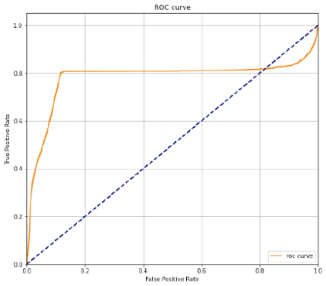


Figure 2: The loss curve of RoBERTa model constructed with toxicity and gab data

The result of in-domain test as shown in figure 3:



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

           1     0.8533    0.8071    0.8295      3048
           0     0.9797    0.9853    0.9825     28818

    accuracy                         0.9683     31866
   macro avg     0.9165    0.8962    0.9060     31866
weighted avg     0.9676    0.9683    0.9679     31866
```

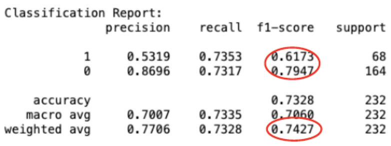


The classification report of RoBERTa model build with Toxicity data in-domain test

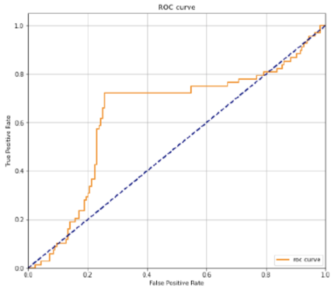
The ROC curve of RoBERTa model build with Toxicity data in-domain test

Figure 3: The classification report and ROC curve of in-domain test of roberta model constructed with toxicity and gab data.

The result of testing on 2400 Covid data as shown in figure 4:



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

           1     0.5319    0.7353    0.6173        68
           0     0.8696    0.7317    0.7947       164

    accuracy                         0.7328       232
   macro avg     0.7007    0.7335    0.7060       232
weighted avg     0.7706    0.7328    0.7427       232
```

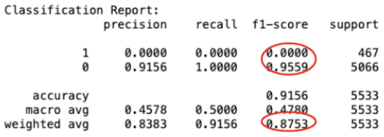


The classification report of RoBERTa model build with Toxicity data test on 2400 Covid data

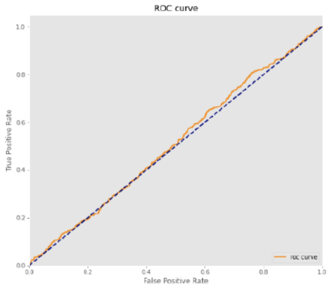
The ROC curve of RoBERTa model build with Toxicity data test on 2400 Covid data

Figure 4: The classification report and ROC curve of roberta model constructed with toxicity and gab data test on 2400 Covid data.

The result of testing on Gab data as shown in figure 5:



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

           1     0.0000    0.0000    0.0000       467
           0     0.9156    1.0000    0.9559      5066

    accuracy                         0.9156      5533
   macro avg     0.4578    0.5000    0.4780      5533
weighted avg     0.8383    0.9156    0.8753      5533
```



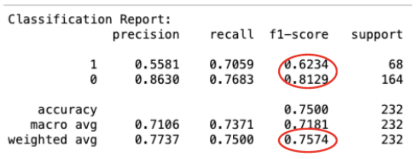
The classification report of RoBERTa model build with Toxicity data test on Gab data

The ROC curve of RoBERTa model build with Toxicity data test on Gab data

Figure 5: The classification report and ROC curve of roberta model constructed with toxicity and gab data test on gab data.

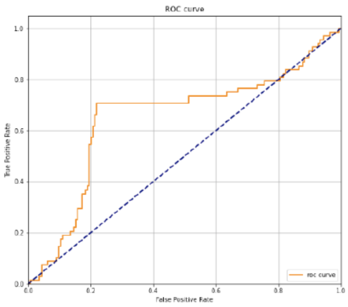
1.3 Fine tuning

After our preliminary test of the data on the model, we then fine-tuned the linear layer, and the final result of testing on 'Gab' and '2400 Covid' data shown in figure 6 and 7:



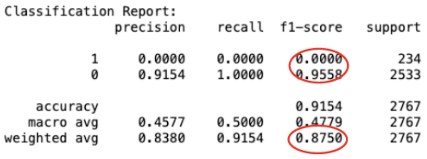
```
Classification Report:
              precision    recall  f1-score   support

           1     0.5581    0.7059    0.6234        68
           0     0.8630    0.7683    0.8129       164

    accuracy                         0.7500       232
   macro avg     0.7106    0.7371    0.7181       232
weighted avg     0.7737    0.7500    0.7574       232
```

The classification report of Fine-tuning RoBERTa model test on 2400 Covid data



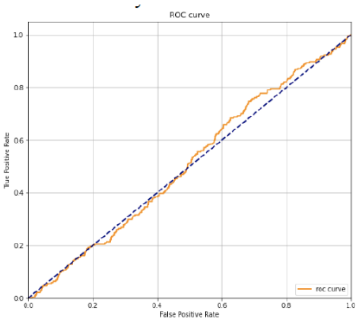
The ROC curve of Fine-tuning RoBERTa model test on 2400 Covid data

Figure 6: The classification report and ROC curve of fine tuning roberta model constructed with toxicity and gab data test on 2400 Covid data.



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

           1     0.0000    0.0000    0.0000       234
           0     0.9154    1.0000    0.9558      2533

    accuracy                         0.9154      2767
   macro avg     0.4577    0.5000    0.4779      2767
weighted avg     0.8380    0.9154    0.8750      2767
```

The classification report of Fine-tuning RoBERTa model test on Gab data



The ROC curve of Fine-tuning RoBERTa model test on Gab data

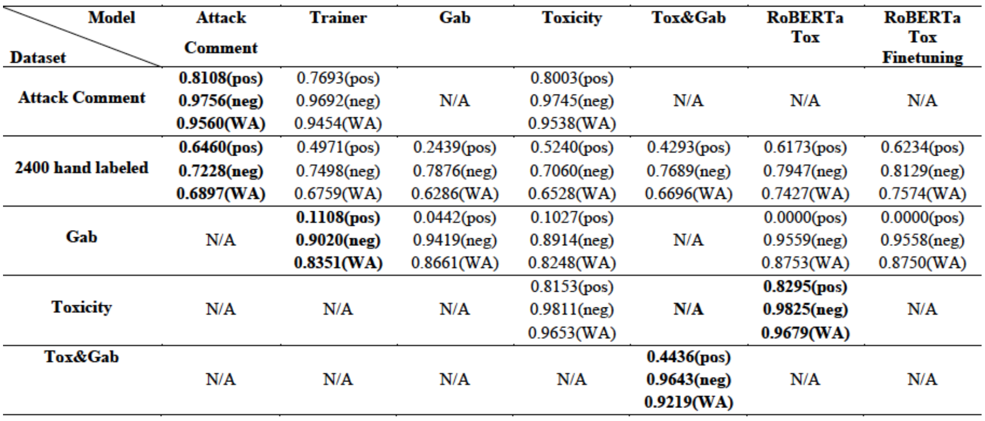
Figure 7: The classification report and ROC curve of fine tuning roberta model constructed with toxicity and gab data test on gab data.

2 Results comparison and Summary

As we can see from figure 8, after eight weeks experiments and exploration on those data sets, we currently get seven different models, for each model we tested on different data sets and make comparisons with each other. The F1 score has slightly improve on Toxicity data set when we use RoBERTa as new model.

By summarizing all the current work, we found that modifying data preprocessing can improve the results to a certain extent. But too much data intervention will interfere with model training. Therefore, our next work is to adjust the model, and even manually add some layers according to the characteristics of the test data, such as: LSTM, Linear layer, or CNN. At the same time, finetuning the parameters of these layers to try to achieve better results for the model.

Note: The final comparison of all models test on different data sets shown in the next page.



Model **Dataset**	**Attack Comment**	**Trainer**	**Gab**	**Toxicity**	**Tox&Gab**	**RoBERTa Tox**	**RoBERTa Tox Finetuning**

Attack Comment	**0.8108(pos)**
	0.9756(neg)
	0.9560(WA)
		0.7693(pos)
		0.9692(neg)
		0.9454(WA)
			N/A	0.8003(pos)
				0.9745(neg)
				0.9538(WA)
					N/A	N/A	N/A

2400 hand labeled	**0.6460(pos)**
	0.7228(neg)
	0.6897(WA)
		0.4971(pos)
		0.7498(neg)
		0.6759(WA)
			0.2439(pos)
			0.7876(neg)
			0.6286(WA)
				0.5240(pos)
				0.7060(neg)
				0.6528(WA)
					0.4293(pos)
					0.7689(neg)
					0.6696(WA)
						0.6173(pos)
						0.7947(neg)
						0.7427(WA)
							0.6234(pos)
							0.8129(neg)
							0.7574(WA)

Gab	N/A	**0.1108(pos)**
		0.9020(neg)
		0.8351(WA)
			0.0442(pos)
			0.9419(neg)
			0.8661(WA)
				0.1027(pos)
				0.8914(neg)
				0.8248(WA)
					N/A	0.0000(pos)
						0.9559(neg)
						0.8753(WA)
							0.0000(pos)
							0.9558(neg)
							0.8750(WA)

Toxicity	N/A	N/A	N/A	0.8153(pos)
				0.9811(neg)
				0.9653(WA)
					N/A	**0.8295(pos)**
						0.9825(neg)
						0.9679(WA)
							N/A

Tox&Gab	N/A	N/A	N/A	N/A	**0.4436(pos)**
					0.9643(neg)
					0.9219(WA)
						N/A	N/A

Figure 8: The comparison of results test in different models