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

Student: Xuanyu Su
Supervisor: Isar Nejadgholi

October 26, 2020

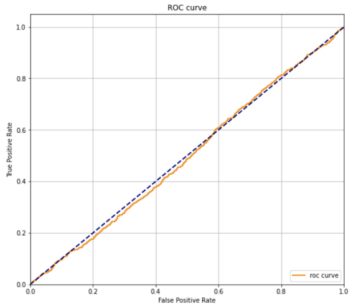
1 Augment 'Toxicity' data with 'Gab' data

In the last week's work, to get better results, we carried out deeper preprocessing operations on the data. To expand the data volume, we integrated part of the 'Attack Comment' data with the 'Gab' data. The results are shown in the figure below:

Classification Report:

	precision	recall	f1-score	support
1	0.0840	0.0300	0.0442	701
0	0.9155	0.9699	0.9419	7599
accuracy			0.8905	8300
macro avg	0.4998	0.4999	0.4930	8300
weighted avg	0.8453	0.8905	0.8661	8300

The classification report of Gab model2.0 on test dataset



The Roc curve of Gab model2.0 on test dataset

Figure 1: The classification report and ROC curve of Gab data

It can be seen from the above results that the model does not work well after the data expansion operations. By comparing the accuracy of different models on different databases (as shown in figure 2), we can find that the 'Attack Comment' model has achieved the best results. However, in our newly acquired database 'Toxicity', it contains more comprehensive data than 'Attack Comment'. Hence we decided to use the 'Toxicity' and 'Gab' data sets, and then retrieve relevant data through the manually created 'racism-word' table extract. Constructing a model by using the above data.

Model / Dataset	Attack Comment	Trainer	Gab
Attack Comment	0.8108(pos) 0.9756(neg) 0.9560(WA)	0.7693(pos) 0.9692(neg) 0.9454(WA)	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)
Gab	N/A	0.1108(pos) 0.9020(neg) 0.8351(WA)	0.0442(pos) 0.9419(neg) 0.8661(WA)

Figure 2: The comparison of three model test on three data sets

1.1 Data preprocessing

we did the same data preprocessing operations on this 'Toxicity' and 'Gab' data sets as in Attack Comment model construction. We got **5860** racism related data from 'Gab' data set among **27665** data in total. Then applied the same method to process the three Toxicity tables(toxicity-train, toxicity-val and toxicity-test), and finally got **41560** racism related data from four tables.
Then we divided the processed data according to the ratio of 8:1:1 (train:val:test) and started to train the model.

1.2 Model used and results

In this version of the model construction, we still use the same model framework as the previous '**Attack Comment**' model: use the **fields** and **TabularDataset** methods to do data embedding process, and then imported data into the **bert-base-uncased** model, set the criterion to **BCELoss()**, the optimizer to **AdamW**, and **weight-decay** is used to prevent overfitting. The results of in-domain test and cross-domain tests shown in the below:

1. model loss:

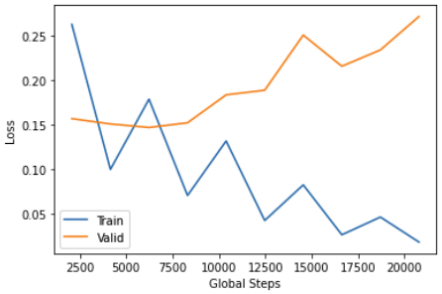
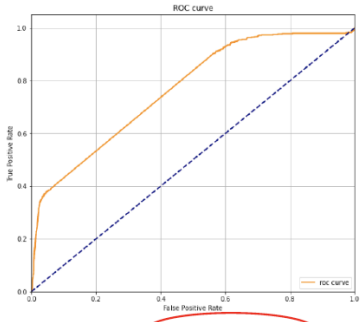


Figure 3: The loss curve of model constructed with toxicity and gab data

2. in-domain test:

	precision	recall	f1-score	support
1	0.6852	0.3279	0.4436	677
0	0.9430	0.9866	0.9643	7635
accuracy			0.9330	8312
macro avg	0.8141	0.6573	0.7040	8312
weighted avg	0.9220	0.9330	0.9219	8312

The classification report of toxicity & gab model in- domain test



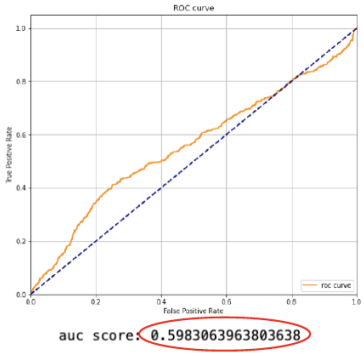
The ROC curve of toxicity & gab model in- domain test

Figure 4: The classification report and ROC curve of in-domain test of model constructed with toxicity and gab data

3. test on 2400 data:

	precision	recall	f1-score	support
1	0.4355	0.4233	0.4293	678
0	0.7645	0.7733	0.7689	1641
accuracy			0.6710	2319
macro avg	0.6000	0.5983	0.5991	2319
weighted avg	0.6683	0.6710	0.6696	2319

The classification report of toxicity & gab model test on 2400 Covid data



The ROC curve of toxicity & gab model test on 2400 Covid data

Figure 5: The classification report and ROC curve of model constructed with toxicity and gab data test on 2400 Covid data.

1.3 Summary

The comparison table of 'Tox and Gab' model testing on different data sets as shown in figure 6. After testing the above two databases('Gab' and 2400 Covid data sets), the results compared with using toxicity data alone, the effect of the model extracted from Gab and Toxicity data is not improved, and the accuracy rate and F1 score is much lower. Therefore, through the analysis of the above results, we conclude that data expansion cannot play a good role in improving the model effect. Next we hope to try to improve the accuracy of the results by changing the use of the model. And we use the test accuracy rate of 2400 Covid data as the judgment standard for the accuracy of the final model.

Dataset \ Model	Attack Comment	Trainer	Gab	Toxicity	Tox&Gab
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
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)
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
Toxicity	N/A	N/A	N/A	0.8153(pos) 0.9811(neg) 0.9653(WA)	**N/A**
Tox&Gab	N/A	N/A	N/A	N/A	**0.4436(pos)** **0.9643(neg)** **0.9219(WA)**

Figure 6: The comparison of test results of different models on different databases.