Work summary of 10th and 11th week

Student: Xuanyu Su
Supervisor: Isar Nejadgholi

November 12, 2020

1 Modifications to the graph of distance comparison

1.1 Explanation of the summing of 'x' and 'y'

We can see from figure1(The results of predictions about 'x' and 'y') that two values returned after model prediction (one is positive and the other is negative), in the last layer of the model, we use the **argmax** function to compare the absolute values of the two numbers, and keep the max one for final prediction. Hence, we sum up x and y for the same reason, this could also be treated as a practice the principle of argmax in another way: by adding two numbers, the sign of the number depends on which of the two numbers has the greater absolute value.(e.g., in the first row of figure1, the absolute value of the negative numbers is greater than the absolute value of positive numbers, therefore, the sum should be negative in this case. And the same for the others.)

```
prediction_list_GAB[0:8]
```

```
[[1.4558666944503784, -1.7571042776107788],
 [1.6604719161987305, -2.0417697429656982],
 [1.3461217880249023, -1.6179455518722534],
 [1.2649874687194824, -1.5062655210494995],
 [0.8044998049736023, -0.9145079255104065],
 [1.032060980796814, -1.205680012702942],
 [1.4560240507125854, -1.7718833684921265],
 [1.0455304384231567, -1.2274339199066162]]
```

Figure 1: The results of prediction about 'x' and 'y'

1.2 Modification on the 'y-axis'

The distance comparison results 1.0 as shown in the figure2: we use **roberta score** as the horizontal axis of the figure, and the **length of each piece of data** as the vertical axis.
We can observe that the length of a single data of '2400 Covid' dataset is much smaller than that of other databases (too small to be recognized). Therefore, we modified the coordinate axis of the picture. Since the x-axis still represents the possibility that the data belongs to either the positive or the negative categories. Here, we only modify the y-axis: change the y-axis from the original **length of a single sentence** through **normalization** the length of each sentence. Then change it to the percentages, so that the data can be evenly distributed on the picture, which is also convenient for comparison and analysis.

The figure of distance comparison after modifying the y-axis as shown in the figure3:

2 RoBERTa model construction with 'East Asia prejudice' data (multi-label classification)

After reading the paper '**Detecting East Asian Prejudice on Social Media** , we tried to follow the author's method mentioned in the paper to process the data and reproduce the model. In the paper, the author mainly analyzes the prejudiced comments about East Asians on Twitter under the outbreak of Covid-19 and the preprocessing of related data.

In this module, we conducted three versions of experiments and adjustments. For the first version of model, we used the same parameters as mentioned in the paper for constructing and testing model. In the second version of

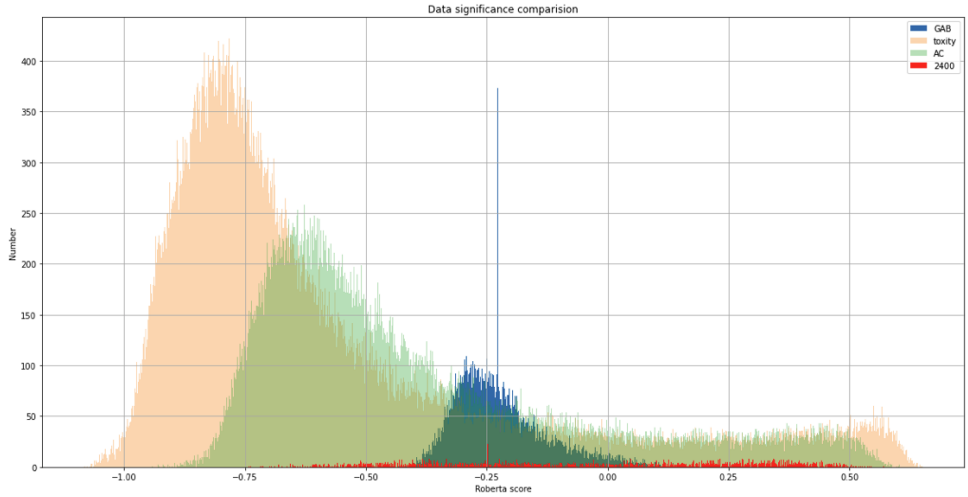

Figure 2: The results of distance comparison among different datasets

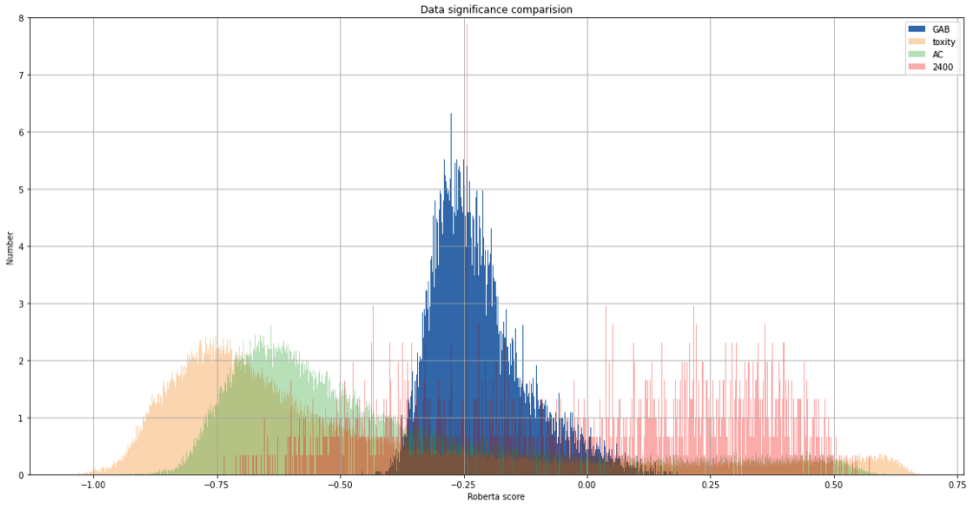

Figure 3: The results of distance comparison among different datasets after modification

model, we fine-tuned the model based on the first version. For the third version, we fine-tuned the model according to the database to be tested.

2.1 RoBERTa model 1.0

2.1.1 Data label processing

In the paper, the author manually categorized and labeled the 20k data. Two annotators independently annotated the data. When opinions diverged, the final results were evaluated by experts and the results were divided into the following five categories:

After observing the data, it is found that there are only a small number of **counter-speech** classes. Therefore, the

```
expert
counter_speech                          116
discussion_of_eastasian_prejudice      1029
entity_directed_criticism              1433
entity_directed_hostility              3898
none_of_the_above                     13524
dtype: int64
```

Figure 4: The types and numbers of data labels

author combined **counter-speech** and **discussion-of-eastasian-prejudice** in the data processing stage.

2.1.2 Hashtag processing

In the paper, there are more than **1,000** high-frequency hashtags related to both EA(east asian) and Covid. In order to classify these hashtags, the author also classified these hashtags through annotators, then it was divided into the following four categories:

```
hashtags.annotator1
hashtags_needed_to_identify_both_themes                4835
hashtags_not_used_at_all_to_identify_themes           10769
hashtags_only_used_to_identify_covid_relevance         4239
hashtags_only_used_to_identify_east_asian_relevance     157
dtype: int64
```

Figure 5: The types and numbers of hashtags

2.1.3 Data label hierarchy

A small number of tweets contained more **than one label category** but each tweet can only be assigned to one category in this taxonomy as they are mutually exclusive. To address this, the author established a **hierarchy of label categories**:

1. entity-directed hostility;
2. entity-directed criticism;
3. counter speech;
4. discussions of East Asian prejudice.

For example, if a single tweet contained both **counter speech** and **hostility** (e.g. It's not fair to blame Chinese scientists, blame their lying government") then it was annotated as **hostility**.

2.1.4 Data preprocessing

In this module we were doing the same operations as we did before: remove URLs, remove symbols, remove emojis, transfer uppercase to lowercase and remove unrelated hashtags.

2.1.5 Model Construction

In this version of model construction we use the same RoBERTa base model as well as the same model parameters like we did in our previous work.
In the final layer of our model, instead of doing binary classification we followed the operations in the paper, we still use argmax but doing a multi-label classification, in our case there are four labels: entity-directed hostility, entity-directed criticism, counter speech and discussions of East Asian prejudice and None.

Here we use a **one-hot** method to generate a **20k * 4** empty matrix, and then according to the result of argmax, the position corresponding to the largest absolute value is changed to 1, and the following label result is obtained.

	comment	label
0	no doubt a china female eastasia	[0, 0, 0, 1]
1	the eastasia is happening behind the live str...	[0, 0, 0, 1]
2	afraid	[0, 0, 0, 1]
3	everybody should wear masks	[0, 0, 0, 1]
4	this makes me remember the sad days in 2003 c...	[0, 0, 1, 0]

Figure 6: The data table after data preprocessing and multi-label transfermation

2.1.6 Results

The overall result of this model **Accuracy: 0.8067857142857143 F1 Score (Micro) = 0.8067857142857143**.

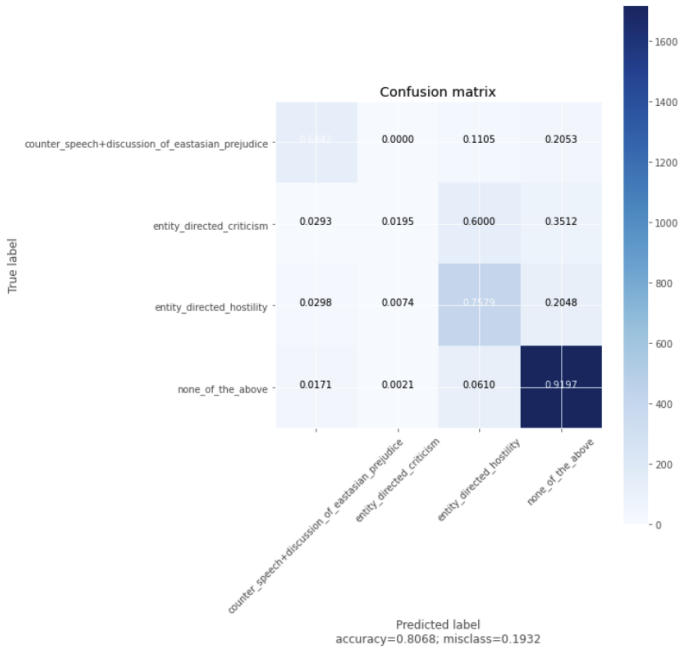

Figure 7: The confusion matrix of RoBERTa model in-domain test on 'EA' dataset

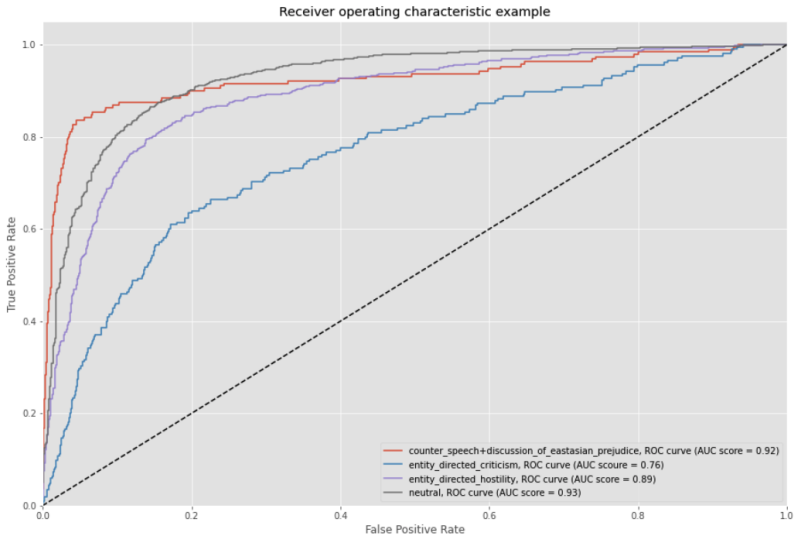

Figure 8: The ROC curve of RoBERTa model in-domain test on 'EA' dataset

Classification Report:

	precision	recall	f1-score	support
counter_speech+discussion_of_eastasian_prejudice	0.7065	0.6842	0.6952	190
entity_directed_criticism	0.3333	0.0195	0.0369	205
entity_directed_hostility	0.6120	0.7579	0.6772	537
none_of_the_above	0.8860	0.9197	0.9025	1868
accuracy			0.8068	2800
macro avg	0.6345	0.5953	0.5780	2800
weighted avg	0.7808	0.8068	0.7819	2800

Figure 9: The classification report of RoBERTa model in-domain test on 'EA' dataset

2.2 Fine-tuning model 1.0

In this version of the model establishment, we did the same data processing and model parameter adjustments as the first version. The only difference is: here we **freezed the parameters in RoBERTa model**, and trained the subsequent linear layers separately, so that they get initialized parameters (rather than randomly).

2.2.1 Results of in-domain test

Classification Report:

	precision	recall	f1-score	support
counter_speech+discussion_of_eastasian_prejudice	0.7151	0.6737	0.6938	190
entity_directed_criticism	0.5385	0.0341	0.0642	205
entity_directed_hostility	0.6233	0.7579	0.6840	537
none_of_the_above	0.8859	0.9272	0.9061	1868
accuracy			0.8121	2800
macro avg	0.6907	0.5982	0.5870	2800
weighted avg	0.7985	0.8121	0.7875	2800

Figure 10: The classification report of RoBERTa model in-domain test on 'EA' dataset after fine tuning

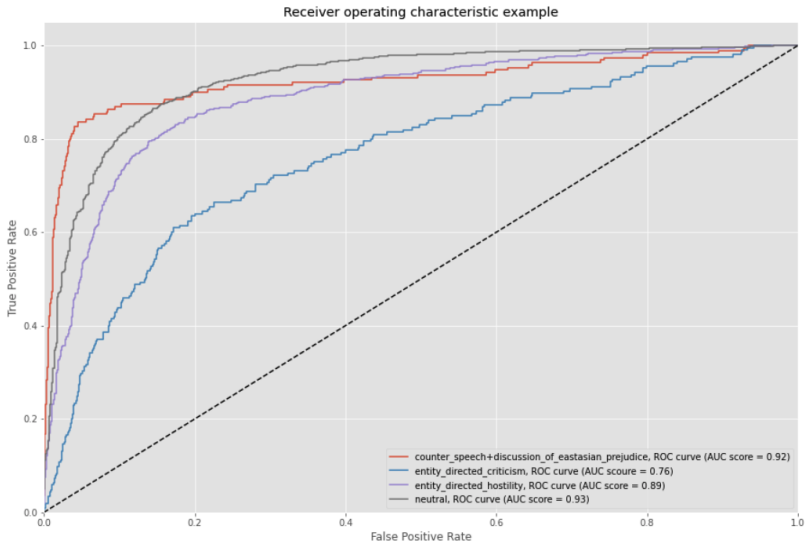

Figure 11: The ROC curve of RoBERTa model in-domain test on 'EA' dataset after fine tuning

2.2.2 Results of testing on 2400 Covid dataset

Classification Report:

	precision	recall	f1-score	support
1	0.5434	0.6932	0.6092	678
0	0.8569	0.7593	0.8052	1641
accuracy			0.7400	2319
macro avg	0.7001	0.7263	0.7072	2319
weighted avg	0.7653	0.7400	0.7479	2319

Figure 12: The classification report of RoBERTa model cross test on 2400 Covid dataset after fine tuning

2.3 Fine-tuning model 2.0

The difference between this version of the model and the previous one is: on the basis of the second version, we divide the test data into three parts: train, val and test. Then use train and val dataset to pre-train the linear layers. The

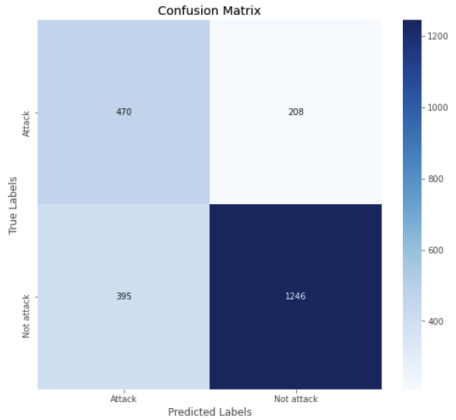

Figure 13: The confusion matrix of RoBERTa model cross test on 2400 Covid dataset after fine tuning

final linear layers of the RoBERTa model can be well adapted to the data to be tested(2400 Covid in this case). The test results are as follows:

2.3.1 Results of testing on 2400 Covid dataset

Classification Report:

	precision	recall	f1-score	support
1	0.5412	0.6174	0.5768	149
0	0.8202	0.7692	0.7939	338
accuracy			0.7228	487
macro avg	0.6807	0.6933	0.6853	487
weighted avg	0.7348	0.7228	0.7275	487

Figure 14: The classification report of RoBERTa model cross test on 2400 Covid dataset after fine tuning 2.0

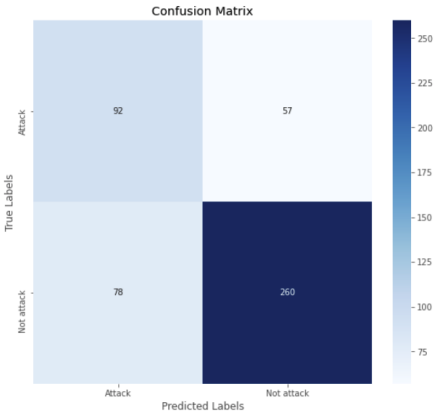

Figure 15: The confusion matrix of RoBERTa model cross test on 2400 Covid dataset after fine tuning 2.0

2.4 Conclusion and Summary

At the beginning of the experiment, we did not process and merge the data labels, so the initial accuracy was only **0.03**. When we reread the paper and merge the minor types of labels, the accuracy of the model has been significantly improved. We found that in this experiment, through the preprocessing of the data and labels, and the use of multi-class prediction models, the results and accuracy of the experiment can be improved to varying degrees. Therefore, we guess that in the later data processing, we focus on 'theme categories' of hashtags, and at the same time make multi-class predictions where feasible.

3 Appendix

3.1 The comparison of the effect of the 2400 Covid dataset on different models

Test Dataset \ Model	**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 Toxicity** (fine-tuning on pretrain of linear layer on test data)	**RoBERTa with EA** (fine-tuning on linear layers)	**RoBERTa with EA** (fine-tuning on pretrain of linear layer on test data)
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)	**0.6234(pos)** **0.8129(neg)** **0.7181(MA)**	**0.6092(pos)** **0.8052(neg)** **0.7072(MA)**	0.5768(pos) 0.7939(neg) 0.6853(MA)

Figure 16: Comparison of the results of 2400 Covid database on different models