Work summary of 13th week

Student: Xuanyu Su
Supervisor: Isar Nejadgholi

November 25, 2020

Through the observation of the 12-week results, we found that the distribution of the test results on the RoBERTa model trained by Toxicity and Attack Comment data in different databases seems to have a certain regular pattern, thus we tried to synthesize the results of different models to create an evaluate value, which is similar to the F1 scores. By using this value, the user can make a preliminary judgment on the training data set. When this value is lower than a certain threshold, the database will be regarded as a database with poor performance (or less relative with the subject of the experiment), which can reduce the time wasted blindly training the data set to some extent. Therefore, in this week's work, we mainly analyze and summarize the three different database distribution images obtained last week.

1 Mistake correction

1.1 Phenomenon analysis

By observing the three data distribution graphs of last week, we found that: by observing the distribution graphs on the two models trained with Attack Comment and Toxicity data, especially the distribution of Attack comment. We can notice that in Figure 1, the distribution of data is almost concentrated in the negative part of x-axis (roberta score), and only a few data are distributed in the positive category, which means there are very few data in the positive class, the results of prediction will be very low for the positive class accordingly. However, by combining the F1 scores of this model, we found that the score is 0.6288 for the positive class, which shows that there are some errors in this data distribution map. In order to solve these errors, we made the following assumptions and analysis.

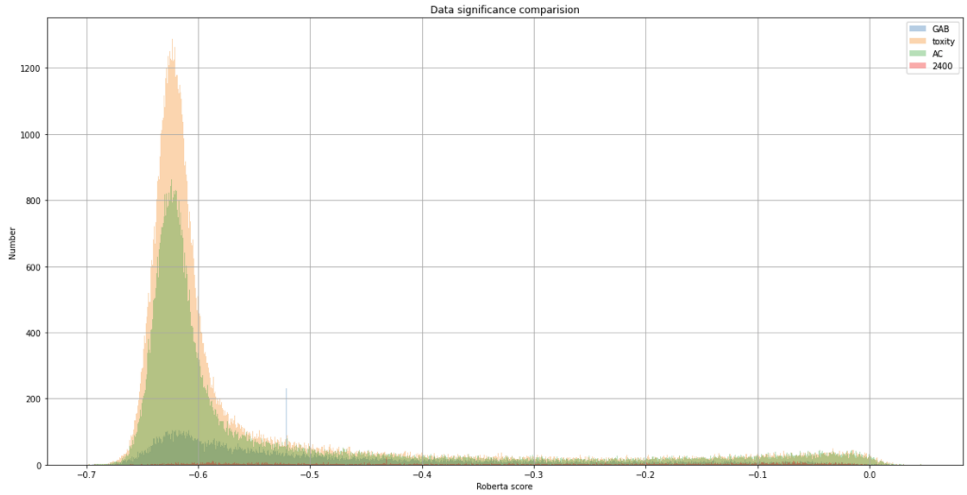

Figure 1: The distance comparison of different data sets test on the RoBERTa model trained with Attack Comment data

1.2 Possible reasons

1. The numbers of epoch.
2. The step of fine-tuning the test data in the extra linear layers.
After reviewing the experimental codes and steps, we found that when drawing the distribution image on the model built with Toxicity data, we set the number of epochs to **12**, and at the same time, divided the whole data to be

tested on this model into train, val and test sets before testing. Then we use the train and val sets to pretrain the extra linear layers. The distribution after we did the two operations are shown in the figure 2:

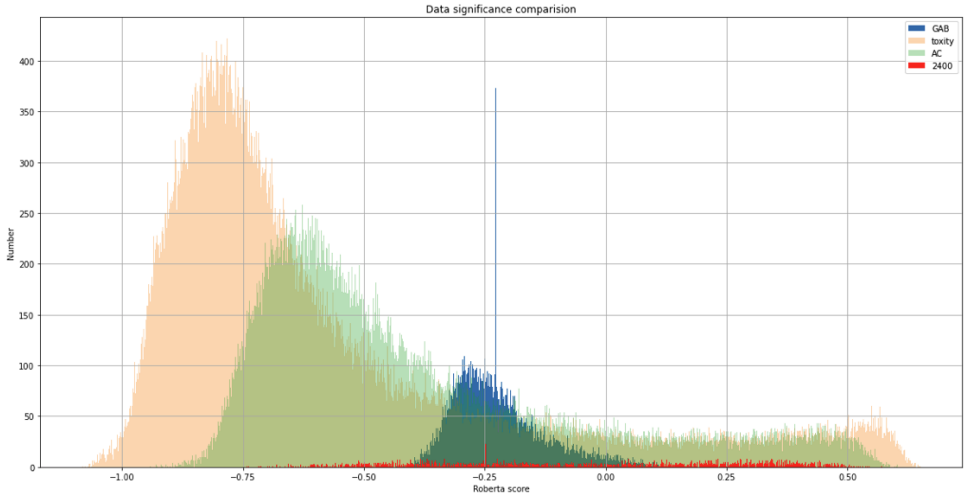

Figure 2: The distance comparison of different data sets test on the RoBERTa model trained with Toxicity data

Through the comparison between Figure 2 and Figure 1, we can find that the distribution interval of different data in Figure 2 is larger, and the distribution is more scattered, the distribution on the x-axis is also more uniform.

1.3 RoBERTa model with Toxicity data(3 epochs and no fine-tuning)

In order to verify the above hypothesis, we retrain the RoBERTa model with Toxicity data, but set the number of **epoch into 3**, and no **fine-tuning on the extra linear layers**, the distribution shown in the figure 3:

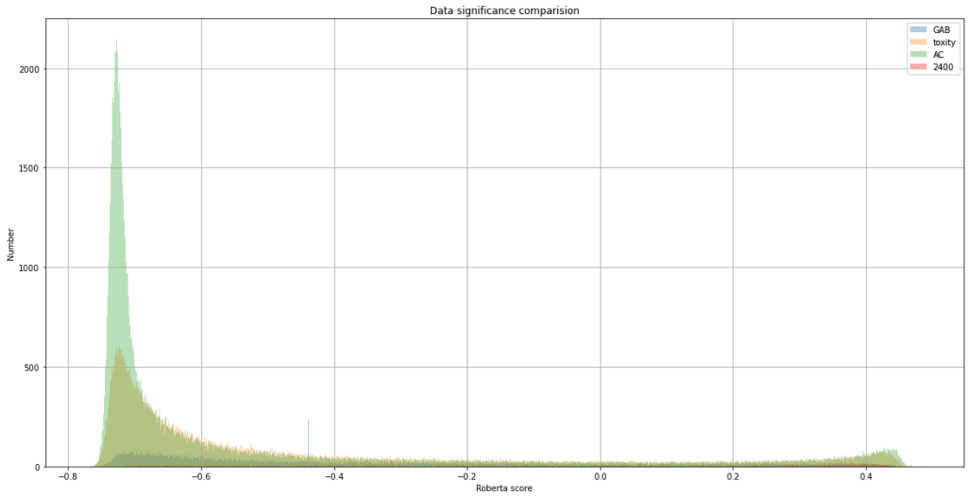

Figure 3: The distance comparison of different data sets test on the RoBERTa model trained with Toxicity data after modified the parameters

1.4 RoBERTa model with Toxcity data (3 epochs and fine-tuning on linear layers)

In order to confirm whether the fine-tuning step will make the data distribution more dispersed or not, we pretrain the test data on the linear layer,s the distribution shown in the figure 4:

By comparing figure 3 and figure 4, we can find that under the premise that the same numbers of epoch , by pretraining the test data on the linear layers, the corresponding database distribution could be more dispersed. Through the comparison of figure 4 and figure 2, we observe that after fine-tuning of the test data is performed, by increasing the number of epochs, the data distribution can be more dense and precise.

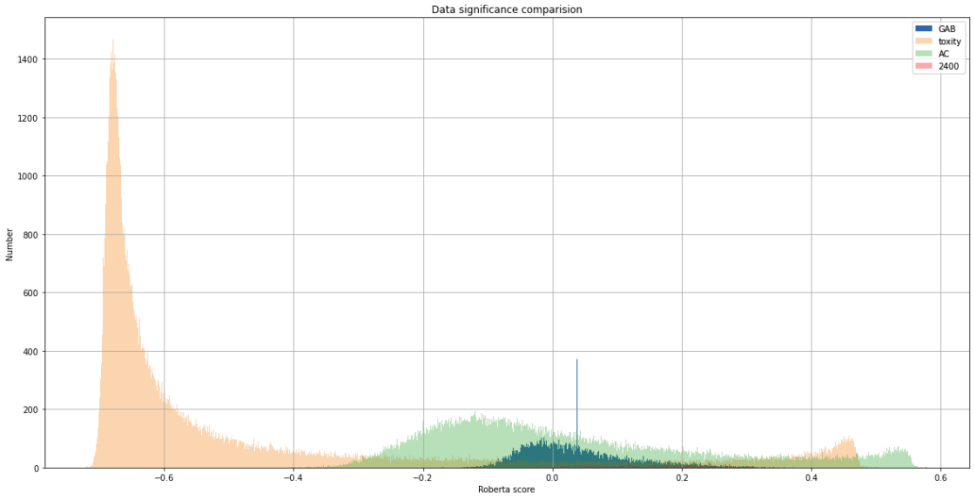

Figure 4: The distance comparison of different data sets test on the RoBERTa model trained with Toxicity data after modified the parameters

1.5 Future work

Through the analysis of the above phenomena and the discussion of the reasons, we could find that the number of epochs and fine-tuning do play a sufficient role in the data distribution. Therefore, in the next work, we hope to retrain the three models, at the same time, increase the number of epochs and analyze the results again.

1.6 Difficulties

In view of our final goal is to derive the corresponding consensus through the distribution of the database to be tested, then make a preliminary judgment on the quality of the database (without the participation of data labels). However, through the analysis of the above reasons, if **fine-tuning** is the main reason that causes the distribution of different databases to be more dispersed, this contradicts the unsupervised prediction we expect, since when the test data is pretrained on the linear layers, this operation requires the participation of data labels.

2 Derivation of Database Distribution Difference Formula

Although the above results have some shortcomings, we have conducted an analysis based on the results of last week's data distribution and preliminary deduce the possible determinants.

2.1 Possible influencing factors

Through the observation of the previous results, we finally deduce the three major factors that determine the consensus. They are:

1. **Degree of overlap:** the degree of overlap between different data sets (each image is generated by training a database to obtain a model, and different databases are used to test on this model, therefore, for each test data, there will be at least one compare data set, that is, the data set we use to train the model. Therefore, the degree of overlap between the test data and the data set of the training model could be used as a factor of judgment).

2. **Data distribution range:** by calculating the size of the distribution range of the database on the **x-axis**, we can have a preliminary understanding the scale of the positive and negative data in the database. On the basis of the original model training data, by calculating the percentage of the new test data, this also could be used as a factor of judgment).

3. **Degree of deviation(skewness):** by analyzing the degree of left and right offset of the database, we can know the distribution of positive and negative data in the database. The calculation method of skewness degree is as follows:

    - **Pearson's first skewness coefficient (mode skewness):** The Pearson mode skewness, or first skewness coefficient, is defined as $\frac{mean-mode}{standarddeviation}$

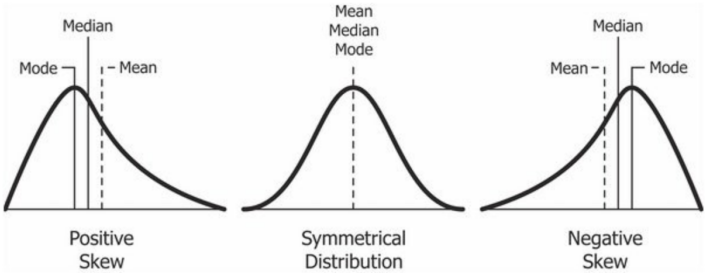

Figure 5: A general relationship of mean and median under differently skewed unimodal distribution

- **Pearson's second skewness coefficient (median skewness):** The Pearson median skewness, or second skewness coefficient is defined as $\frac{3(mean-median)}{standarddeviation}$

2.2 Summary

The judgment of the above influence factors is only a preliminary analysis, and the calculation of specific parameters will not be carried out until a new round of experimental results are obtained. At the same time, in the next work, we will not only analyze and discover the influence factors, but also analyze the weights of different factors.

3 Appendix

Model / test Dataset	Bert with Attack Comment	Trainer with Attack Comment	Bert with Gab	Bert with Tocixity	RoBERTa with Toxicity (fine-tuning on linear layers))	RoBERTa with Attack Comment (fine-tuning On linear layers)
Attack Comment	**0.8108(pos)**	0.7693(pos)		**0.8003(pos)**	0.8269(pos)	0.6288(pos)
	0.9756(neg)	0.9692(neg)		**0.9745(neg)**	0.9809(neg)	0.9353(neg)
	0.8932(MA)	0.8693(MA)		**0.8874(MA)**	0.9039(WA)	0.7821(MA)
2400 Covid	0.6460(pos)	0.4971(pos)	0.2439(pos)	0.5240(pos)	**0.6173(pos)**	0.5278(pos)
	0.7228(neg)	0.7498(neg)	0.7876(neg)	0.7060(neg)	**0.7947(neg)**	0.6804(neg)
	0.6844(MA)	0.6234(MA)	0.5157(MA)	0.6150(MA)	**0.7060(MA)**	0.6041(MA)
Gab		**0.1108(pos)**	0.0442(pos)	0.1027(pos)	0.0923(pos)	0.1268(pos)
		0.9020(neg)	0.9419(neg)	0.8914(neg)	0.8882(neg)	0.8682(neg)
		0.5062(MA)	0.4930(MA)	0.4970(MA)	0.4902(MA)	0.4975(MA)
Toxicity				0.8153(pos)	0.8295(pos)	0.8085(pos)
				0.9811(neg)	0.9825(neg)	0.9780(neg)
				0.8982(MA)	0.9060(MA)	0.8932(MA)

Figure 6: Comparison of the results different databases on different models

After we complete the result in the RoBERTa with Toxicity column, the model that performed best on the Attack Comment dataset has changed from Bert with Attack Comment to RoBERTa with Toxicity data.