First, we do the data preprocessing, delete the meaningless words like 0, is,the,are etc, and changing the characters from full width to half width.

And we found that in the training data, the amount difference between the number of label 0 and label 1 is too large.

Label 0’s amount is **17838** and Label 1‘s amount is **2562**. So we sample the 2600 data from label 0 and let the amount as much as label 1 to make training results better.

After the data preprocessing, we use tf-idf to convert the sentences into the vectors in order to input into our training model.

And about our training model, I used the **random forest** model to train the data, using the **GridSearchCV** to find suitable parameters and also using the **cross validation** (computing the cross validation value will cost some time) to evaluate the training model score.

Our precision score:

```
              precision    recall  f1-score   support

           0       0.99      0.95      0.97      3097
           1       0.61      0.89      0.73       303

    accuracy                           0.94      3400
   macro avg       0.80      0.92      0.85      3400
weighted avg       0.96      0.94      0.95      3400

(0.6141552511415526, 0.8877887788778878, 0.7260458839406209)
```

Reference:

GridSearchCV:
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html?fbclid=IwAR1KYu42yljTvDdAdoOtUiOZt_1YMsN3u6HgzidwrUSr67NzEgLFDO2MKJE

RandomForestClassifier:
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Cross-validation:
https://scikit-learn.org/stable/modules/cross_validation.html

Tfidf:
https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html