

Quora Question Pairs

Abstract

Quora is a place to gain and share knowledge about anything. It's a website to ask questions and connect with people who contribute unique insights and quality answers. Over 100 million people visit Quora every month, it's no surprise that many people ask similarly worded questions. But so many questions cause a lot of same questions with different word or different way to ask. These multiple questions with the same intent can cause seekers to spend more time finding the best answer to their question, and make writers feel they need to answer multiple versions of the same question. So, we try using the NLP technique to identify duplicate questions that can provide a better experience to active seekers and writers and offer more value to both groups in the long term.

Howard(Model)

Dataset for training 404289 data

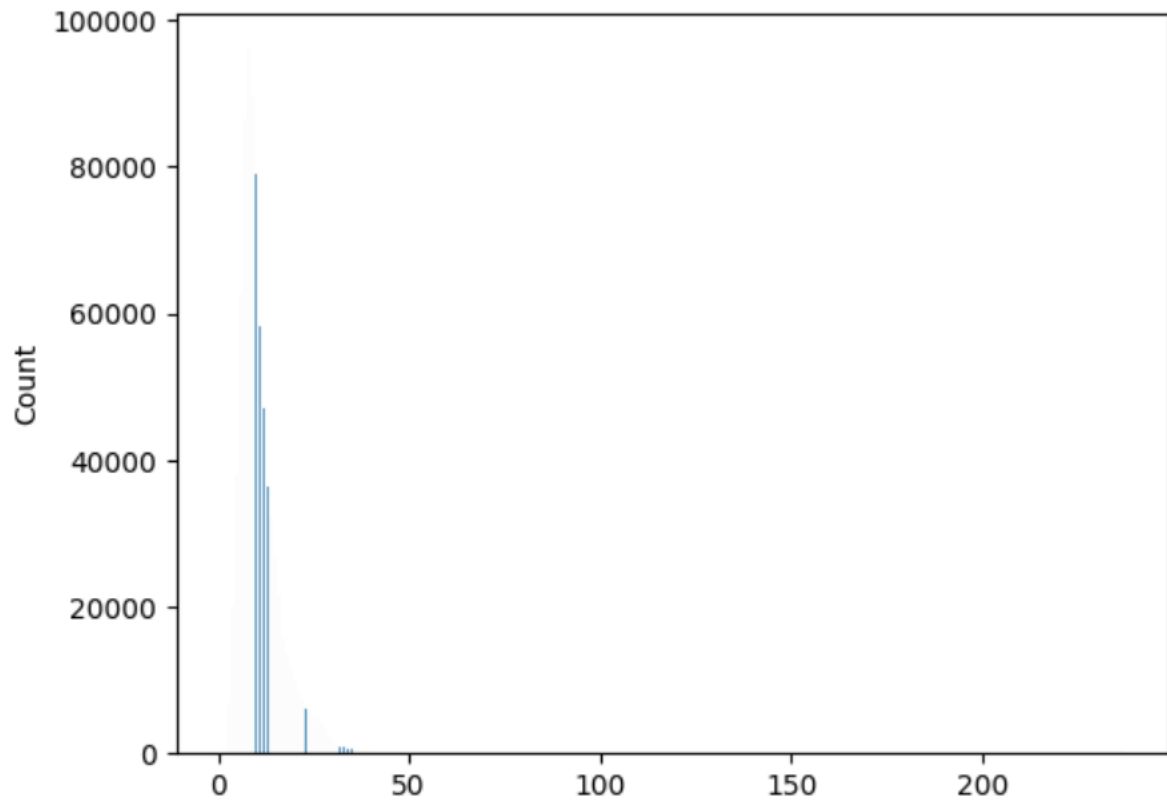
id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh... What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia... What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed of my internet co... How can Internet speed be increased by hacking...	0
3	3	7	8	Why am I mentally very lonely? How can I solve... Find the remainder when 23^{24} i...	0
4	4	9	10	Which one dissolve in water quickly sugar, salt... Which fish would survive in salt water?	0
5	5	11	12	Astrology: I am a Capricorn Sun Cap moon and c... I'm a triple Capricorn (Sun, Moon and ascendan...	1
6	6	13	14	Should I buy tiago? What keeps childern active and far from phone ...	0
7	7	15	16	How can I be a good geologist? What should I do to be a great geologist?	1
8	8	17	18	When do you use シ instead of ㇿ? When do you use "&" instead of "and"?	0
9	9	19	20	Motorola (company): Can I hack my Charter Moto... How do I hack Motorola DCX3400 for free internet?	0

Dataset for testing 3563490 data

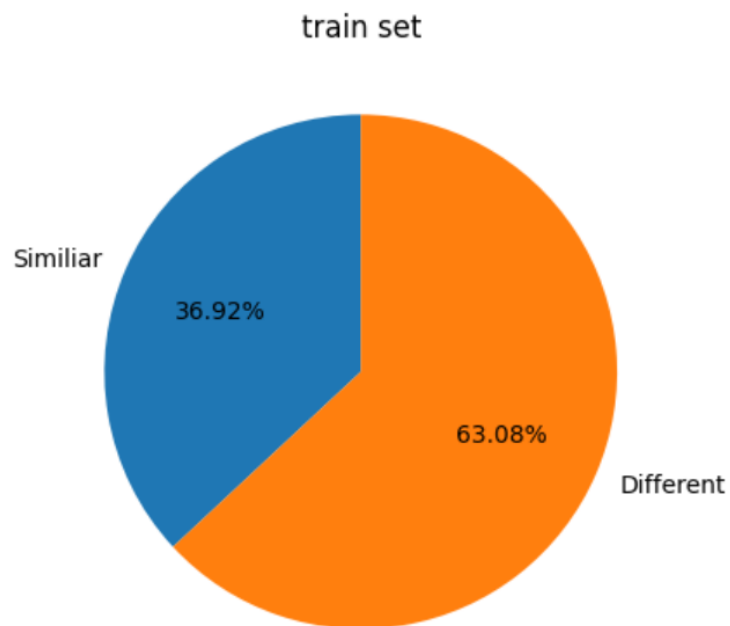
test_id	question1	question2
0	0 How does the Surface Pro himself 4 compare wit... Why did Microsoft choose core m3 and not core ...	
1	1 Should I have a hair transplant at age 24? How... How much cost does hair transplant require?	
2	2 What but is the best way to send money from Ch... What you send money to China?	
3	3 Which food not emulsifiers? What foods fibre?	
4	4 How "aberystwyth" start reading? How their can I start reading?	
5	5 How are the two wheeler insurance from Bharti ... I admire I am considering of buying insurance ...	
6	6 How can I reduce my belly fat through a diet? How can I reduce my lower belly fat in one month?	
7	7 By scrapping the 500 and 1000 rupee notes, how... How will the recent move to declare 500 and 10...	
8	8 What are the how best books of all time? What are some of the military history books of...	
9	9 After 12th years old boy and I had sex with a ... Can a 14 old guy date a 12 year old girl?	

Training data:

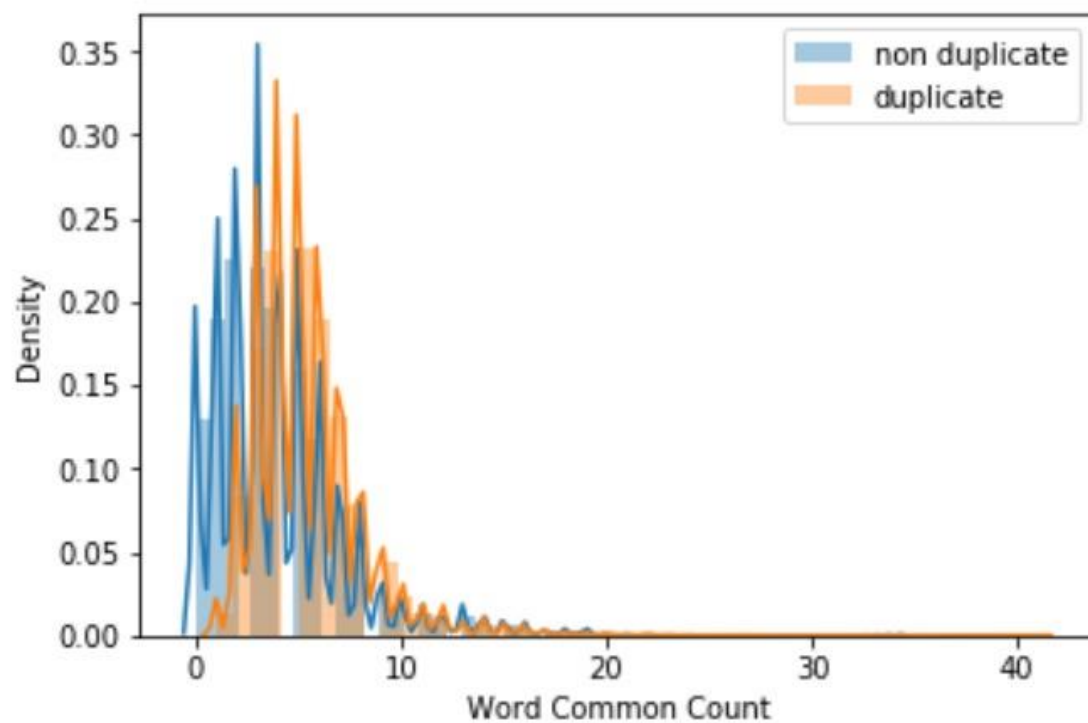
The length of the sentence in the data



The duplicate and different data comparison

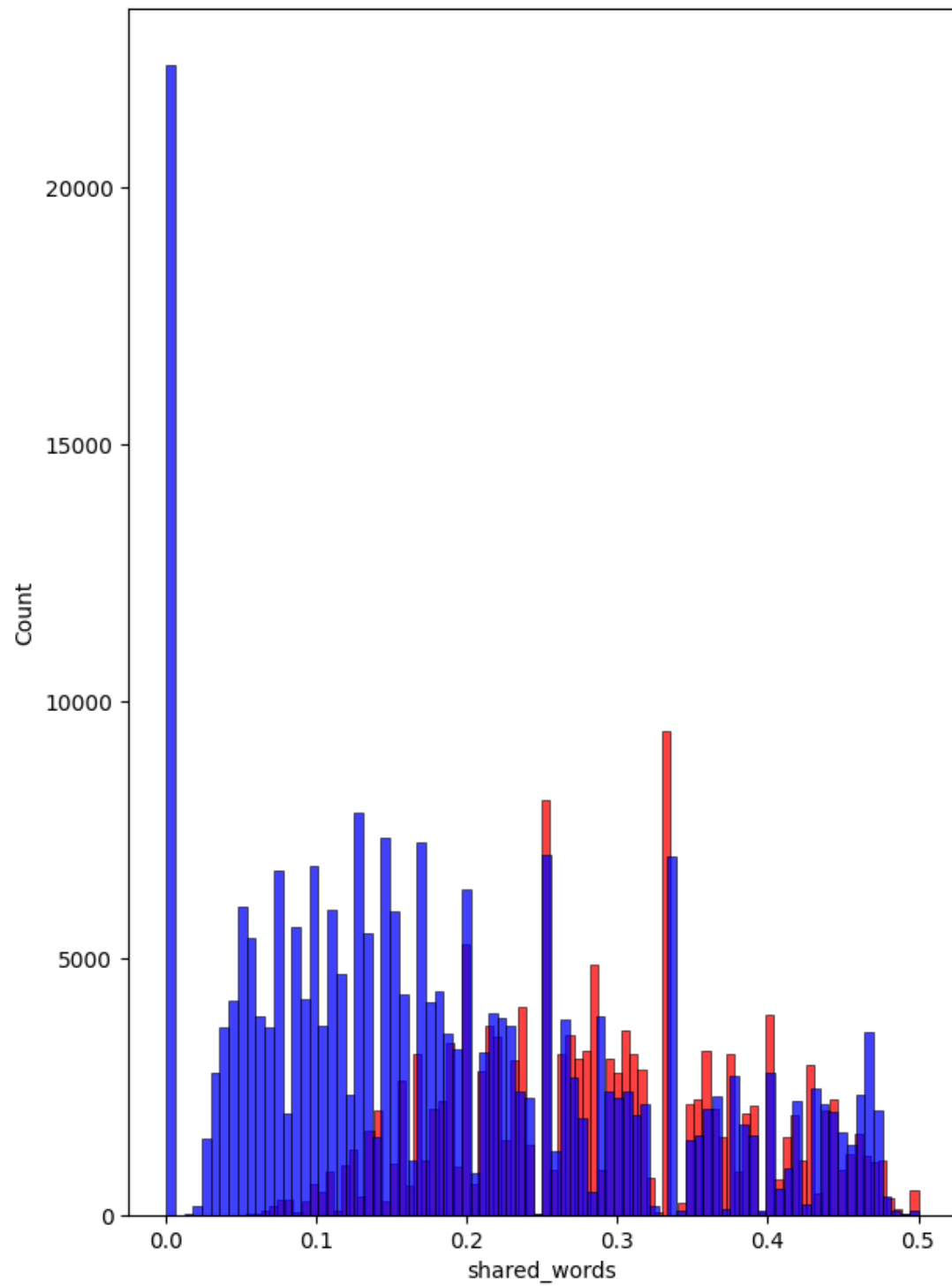


Any common words between the questions



Any shared words between the questions

Blue: is duplicate Red: Not duplicate



Bert-base-uncased model:

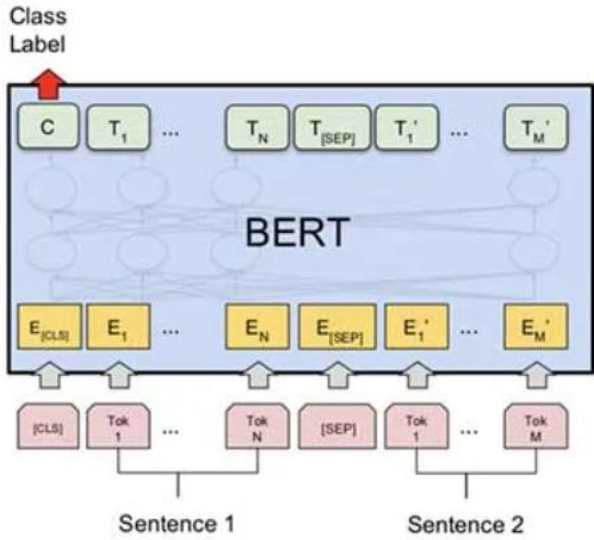
Pretrained model on English language using a masked language modeling (MLM) objective. This model is uncased: it does not make a difference between english and English.

BERT is a transformers model pretrained on a large corpus of English data which was pretrained on the raw texts only, without humans labeling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts.

Masked language modeling (MLM): taking a sentence, the model randomly masks 15% of the words in the input then run the entire masked sentence through the model and has to predict the masked words. It allows the model to learn a bidirectional representation of the sentence. The model learns an inner representation of the English language that can then be used to extract features useful for downstream task.

The inputs of the model

[CLS] Question1 [SEP] Question2 [SEP]



(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG

Transfer the data into

Batch data:

```
Input IDs: tensor([[ 101, 1045, 2215, ..., 0, 0, 0],
                  [ 101, 2339, 2003, ..., 0, 0, 0],
                  [ 101, 2129, 2079, ..., 0, 0, 0],
                  ...,
                  [ 101, 2129, 2079, ..., 0, 0, 0],
                  [ 101, 2129, 2097, ..., 0, 0, 0],
                  [ 101, 22817, 14181, ..., 0, 0, 0]])
```

```
Attention Mask: tensor([[1, 1, 1, ..., 0, 0, 0],
                        [1, 1, 1, ..., 0, 0, 0],
                        [1, 1, 1, ..., 0, 0, 0],
                        ...,
                        [1, 1, 1, ..., 0, 0, 0],
                        [1, 1, 1, ..., 0, 0, 0],
                        [1, 1, 1, ..., 0, 0, 0]])
```

```
Token Type IDs: tensor([[0, 0, 0, ..., 0, 0, 0],
                        [0, 0, 0, ..., 0, 0, 0],
                        [0, 0, 0, ..., 0, 0, 0],
                        ...,
                        [0, 0, 0, ..., 0, 0, 0],
                        [0, 0, 0, ..., 0, 0, 0],
                        [0, 0, 0, ..., 0, 0, 0]])
```

```
Targets: tensor([1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1,
                  0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
                  0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0,
                  0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1,
                  1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1,
                  0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0,
                  0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0,
                  0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0,
                  1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1,
                  1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1,
                  0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0])
```


Training

5 epoch and Learning Rate(lr) is $3e-5$ and the batch is depend on the data, it'll be 1422 in this case.

```
Epoch: 1
| Iter 100 | Avg Train Loss 0.36829567819833753 | Dev Perplexity 1.4120122640005845
| Iter 200 | Avg Train Loss 0.3381352695822716 | Dev Perplexity 1.3721972345022608
| Iter 300 | Avg Train Loss 0.3211359757184982 | Dev Perplexity 1.3528421891209148
| Iter 400 | Avg Train Loss 0.30819258719682696 | Dev Perplexity 1.348373891179349
| Iter 500 | Avg Train Loss 0.2963800723850727 | Dev Perplexity 1.3360203537585158
| Iter 600 | Avg Train Loss 0.2925376646220684 | Dev Perplexity 1.3239872664429344
| Iter 700 | Avg Train Loss 0.2911820366978645 | Dev Perplexity 1.314145187190489
| Iter 800 | Avg Train Loss 0.27900337100028993 | Dev Perplexity 1.3095773829700192
| Iter 900 | Avg Train Loss 0.2798917533457279 | Dev Perplexity 1.306145566351985
| Iter 1000 | Avg Train Loss 0.270376580953598 | Dev Perplexity 1.2979141196432822
| Iter 1100 | Avg Train Loss 0.27087289914488794 | Dev Perplexity 1.2886162475338117
| Iter 1200 | Avg Train Loss 0.26712505251169205 | Dev Perplexity 1.2872292195886537
| Iter 1300 | Avg Train Loss 0.25675598174333575 | Dev Perplexity 1.286855010220885
| Iter 1400 | Avg Train Loss 0.2632251465320587 | Dev Perplexity 1.285382129592646
```

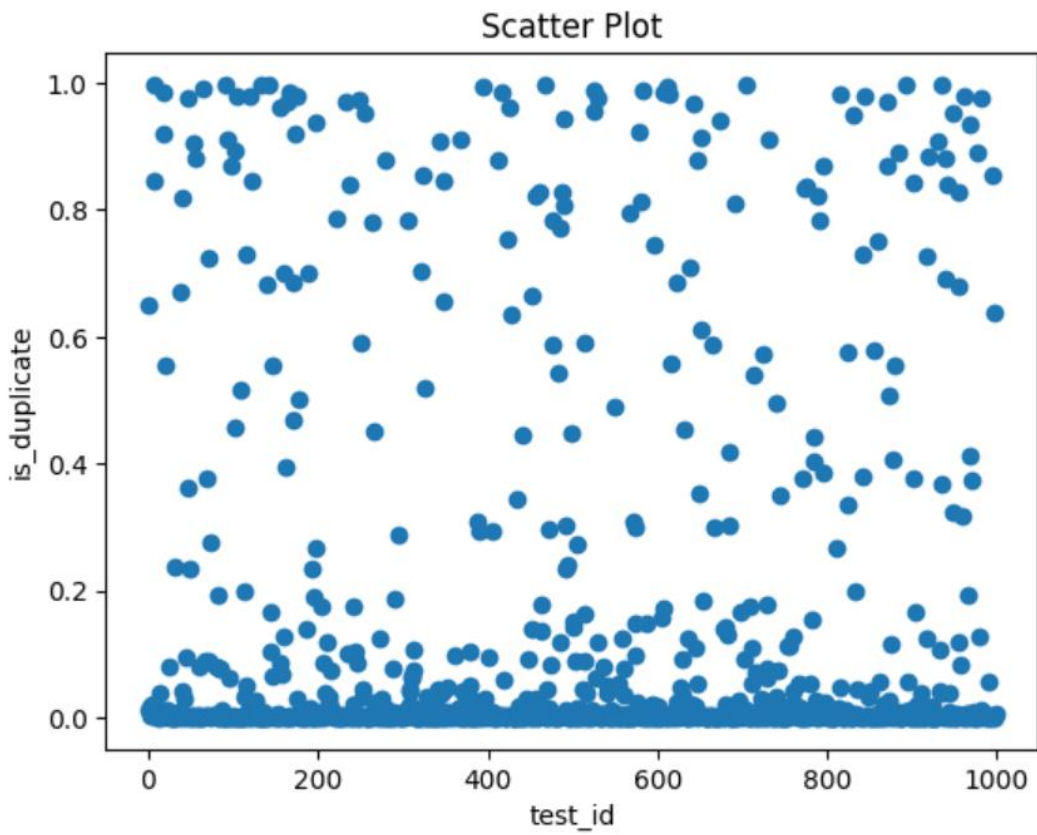
Prediction on test dataset

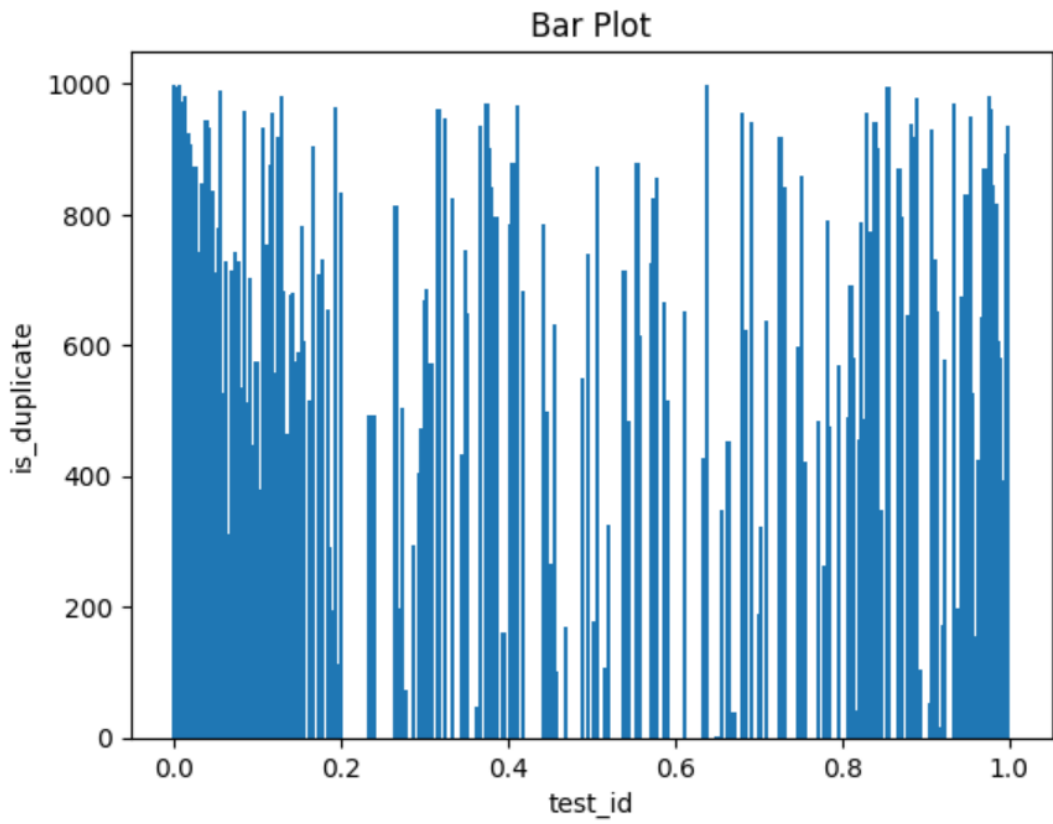
batch size as 512. And I use sigmoid to show the result of my prediction.

	test_id	question1	question2	is_duplicate
0	0	How does the Surface Pro himself 4 compare wit...	Why did Microsoft choose core m3 and not core ...	0.010535
1	1	Should I have a hair transplant at age 24? How...	How much cost does hair transplant require?	0.650432
2	2	What but is the best way to send money from Ch...	What you send money to China?	0.018178
3	3	Which food not emulsifiers?	What foods fibre?	0.004047
4	4	How "aberystwyth" start reading?	How their can I start reading?	0.020853
...
95	95	What does it mean when my husband looks at oth...	What should I do when my husband looks for oth...	0.062463
96	96	For which exam a graduate electrical student s...	What are some criteria to be called ILLEGAL im...	0.000357
97	97	How we can earn not easily?	How can I get genuine money easily?	0.869894
98	98	What are the to different symbols used by The ...	What does the nothing symbol mean ☹ ?	0.000721
99	99	What are which cannot be tamed by humans?	How did hal humans tame wild animals?	0.001602

100 rows × 4 columns

Head 1000 data of the result can show like this.





Lakshmi Jampala(Model)

RandomForestClassifier model:

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Random forests are a popular supervised machine learning algorithm. Random forests are for supervised machine learning, where there is a labeled target variable.

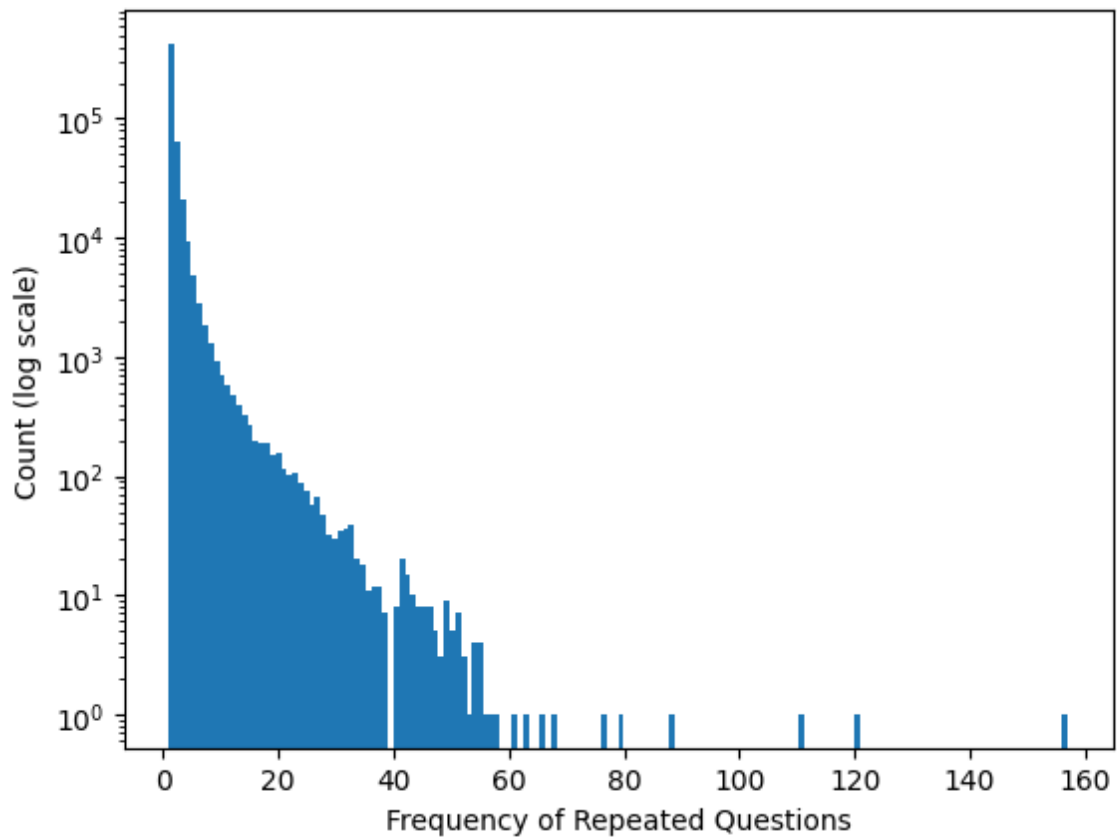
Random forests can be used for solving regression (numeric target variable) and classification (categorical target variable) problems.

Random forests are an ensemble method, meaning they combine predictions from other models.

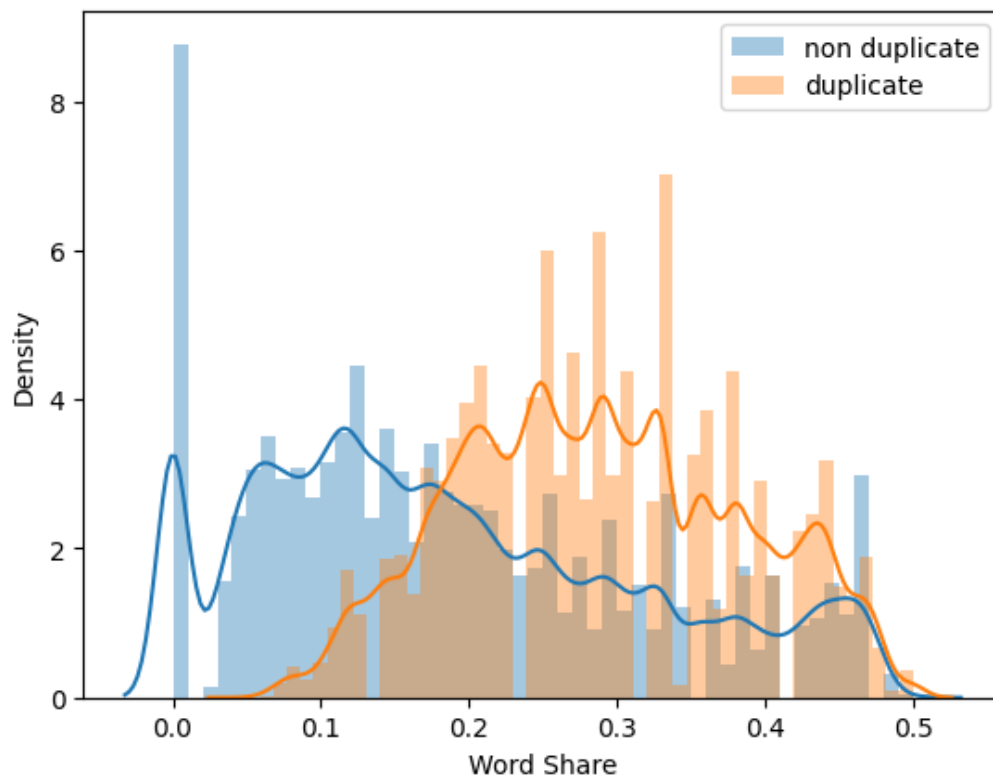
Each of the smaller models in the random forest ensemble is a decision tree.

In a random forest classification, multiple decision trees are created using different random subsets of the data and features. Each decision tree is like an expert, providing its opinion on how to classify the data. Predictions are made by calculating the prediction for each decision tree, then taking the most popular result.

Repeated questions histogram



Word sharing between the questions

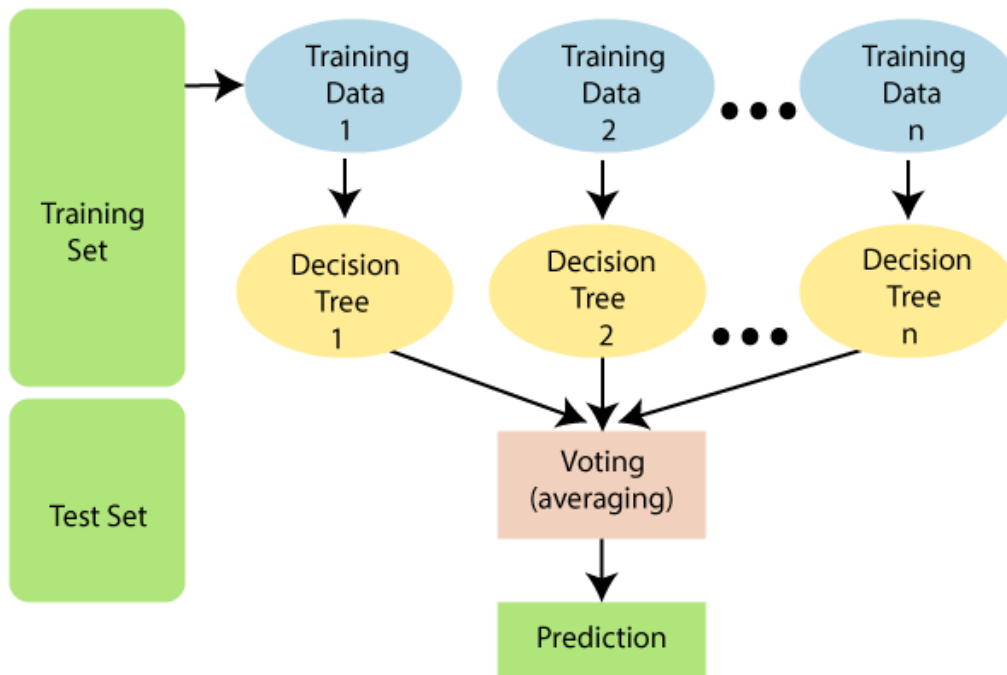


Cleaning data by removing punctuation, whitespace, numbers, stop words ...

```
Quora.head()
```

	is_duplicate	question1_data	question2_data
0	0	step step guide invest share market india	step step guide invest share market
1	0	story kohinoor koh-i-noor diamond	would happen indian government stole kohinoor ...
2	0	increase speed internet connection using vpn	internet speed increased hacking dns
3	0	mentally lonely solve	find remainder divided
4	0	one dissolve water quickly sugar salt methane c...	fish would survive salt water

Splitting the data



Applying NLP concepts to convert text data into numerical data

```
print("Vectorizing data X",X)
```

```
Vectorizing data X [[ 0.22560713  0.32502179 -0.02803988 ... -0.19613373  0.19055542
 -0.17635124]
 [ 0.16817777  0.23091618  0.04495402 ... -0.23079377  0.13712243
 -0.21380465]
 [ 0.22681373  0.32891811 -0.03921262 ... -0.18681507  0.22669766
 -0.11297099]
 ...
 [ 0.28174647  0.26994656  0.04555481 ... -0.22216995  0.1800034
 -0.10070259]
 [ 0.23694526  0.28099332 -0.04428751 ... -0.184428   0.20072742
 -0.16821643]
 [ 0.29607153  0.38105129 -0.06242954 ... -0.19255908  0.2178838
 -0.1715577 ]]
```

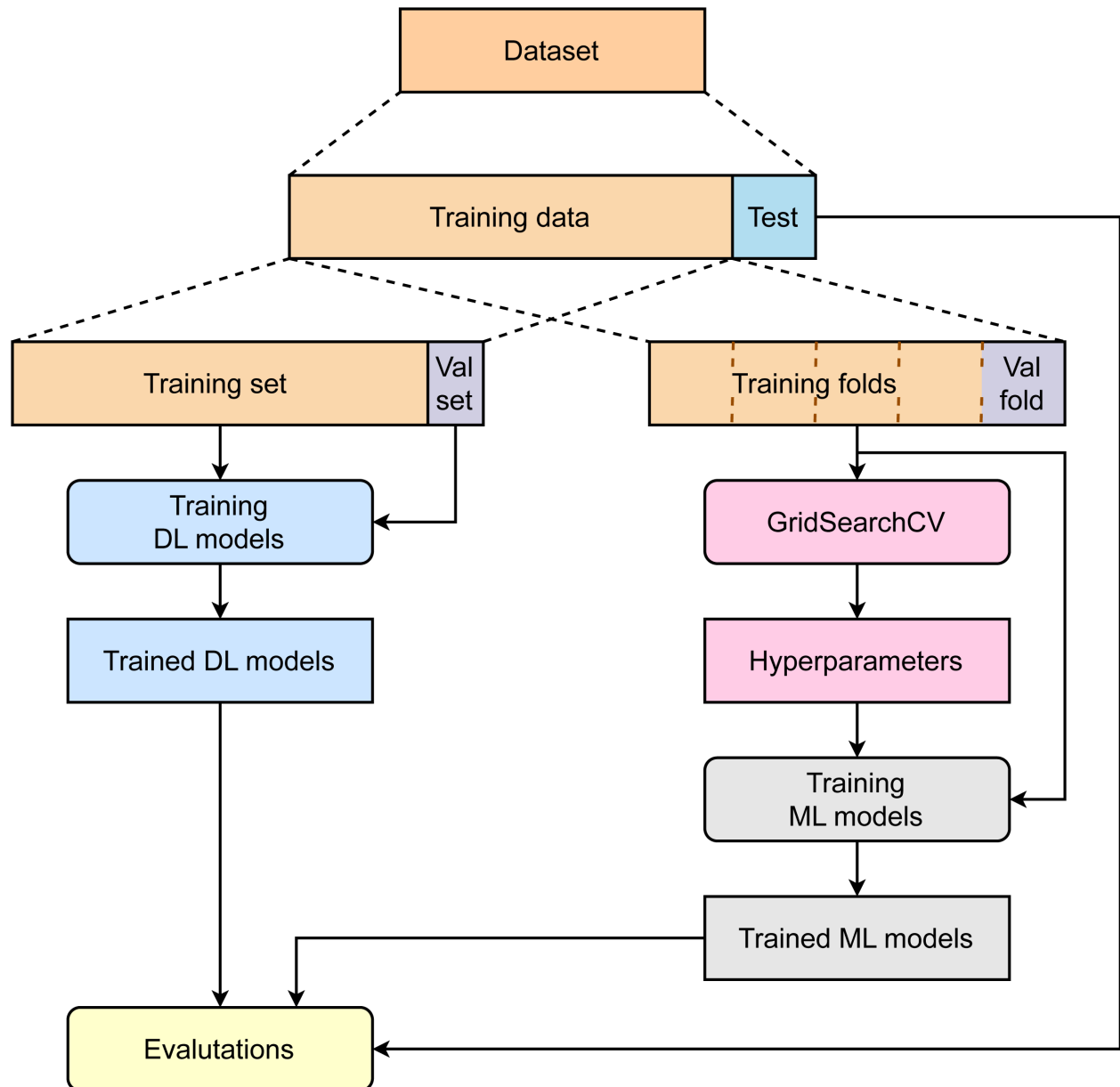
```
| print("Vectorizing data Y",Y)
```

```
Vectorizing data Y [[ 0.19455017  0.33969575 -0.04046475 ... -0.19186648  0.20476778
-0.18605424]
 [ 0.15704131  0.27530233  0.01428822 ... -0.20243841  0.15489152
-0.20057929]
 [ 0.21214438  0.34107452 -0.04463648 ... -0.18947866  0.16397035
-0.14537013]
 ...
 [ 0.2287837   0.19773758 -0.12563304 ... -0.23593432  0.12962447
-0.26771324]
 [ 0.19831685  0.30005907 -0.02401052 ... -0.14407332  0.14360739
-0.14730805]
 [ 0.29607153  0.38105129 -0.06242954 ... -0.19255908  0.2178838
-0.1715577 ]]
```

```
print("features=np.hstack((X, Y))",features)
```

```
features=np.hstack((X, Y)) [[-0.31811662  0.04843546 -0.33446787 ... -0.0806405   0.10634531
 0.30822576]
 [-0.35298337 -0.07278607 -0.35526297 ... -0.05153562  0.07325101
 0.19462576]
 [-0.36526286  0.00417036 -0.28012835 ... -0.09279771  0.07898718
 0.22555098]
 ...
 [-0.35302412 -0.05767254 -0.42713664 ...  0.01545093  0.04937567
 0.06119705]
 [-0.31443293  0.04872944 -0.2786974   ... -0.03902187  0.13474174
 0.28198695]
 [-0.35202461  0.03710565 -0.37752191 ... -0.04201102  0.11075891
 0.21190945]]
```


Randomforestclassifier model architecture:



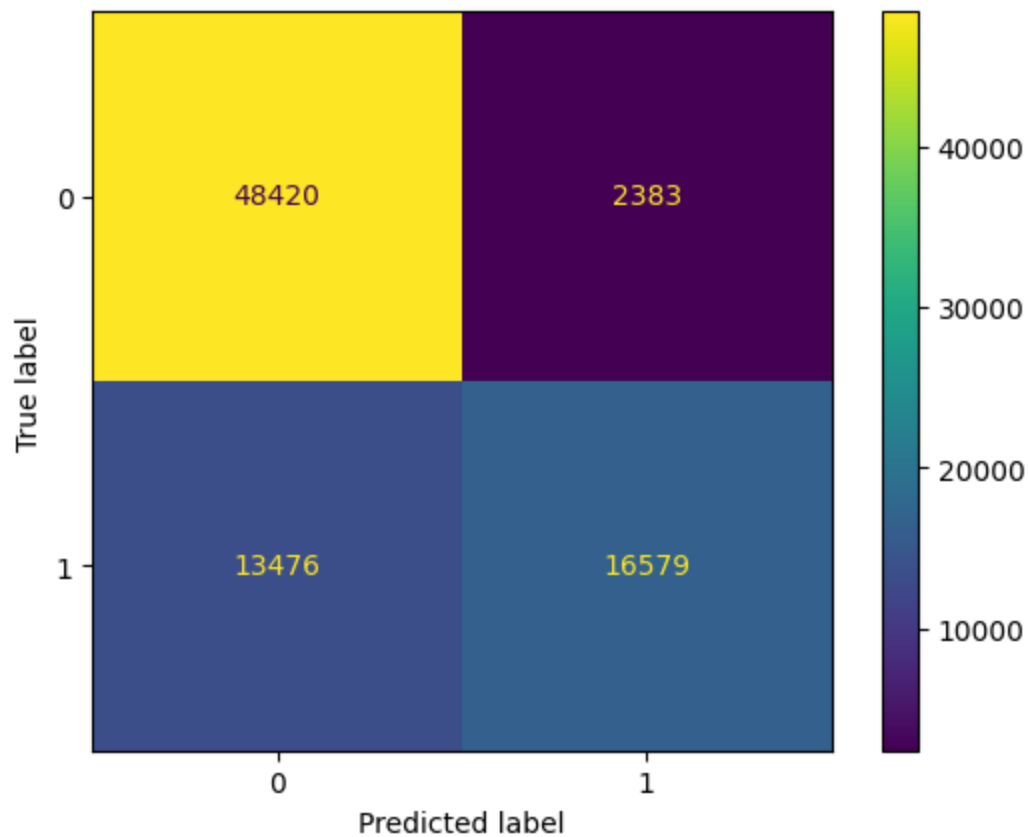
Evaluation Model

```
> Accuracy: 0.8040886492369339
      precision    recall  f1-score   support

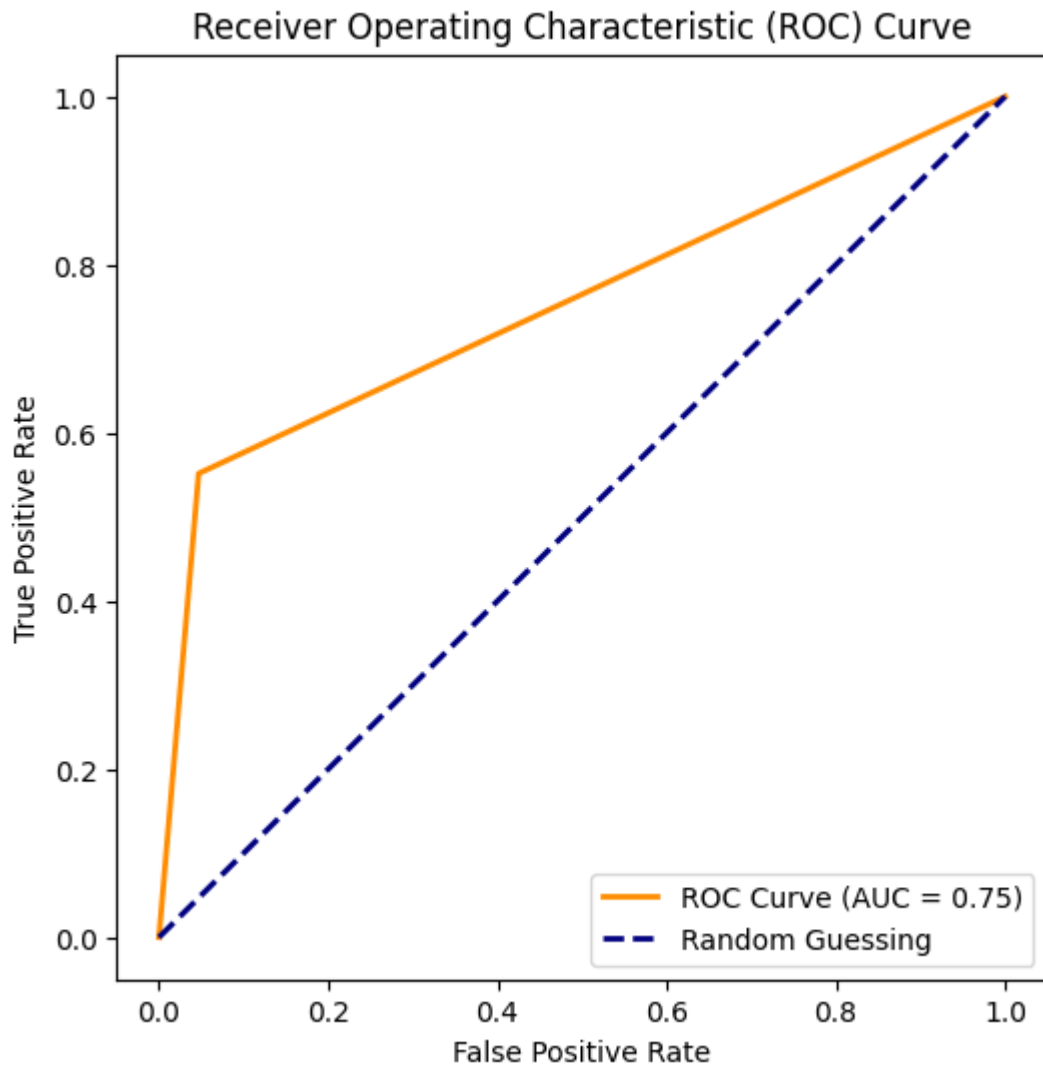
     0       0.78      0.95      0.86     50803
     1       0.88      0.55      0.68     30055

 accuracy      0.80      80858
 macro avg     0.83      0.75      0.77     80858
 weighted avg   0.82      0.80      0.79     80858
```

Confusion Matrix: Precision, Recall, and F1-Score:



Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC):



Lakshmi-References:

<https://www.nature.com/articles/s41598-022-24979-9>

<https://peerj.com/articles/cs-1570/>

<https://www.kaggle.com/competitions/quora-question-pairs/code>

<https://github.com/campusx-official/quora-question-pairs/blob/main/bow-with-basic-features.ipynb>

<https://www.javatpoint.com/machine-learning-random-forest-algorithm>

Howard-Reference:

<https://www.geeksforgeeks.org/matplotlib-pyplot-scatter-in-python/>

<https://huggingface.co/bert-base-uncased>

<https://wandb.ai/wandb/common-ml-errors/reports/How-to-Save-and-Load-Models-in-PyTorch--VmlldzozMjg0MTE>

<https://peaceful0907.medium.com/sentence-embedding-by-bert-and-sentence-similarity-759f7becbf1>

<https://www.kaggle.com/competitions/quora-question-pairs/overview>

<https://www.geeksforgeeks.org/seaborn-barplot-method-in-python/>

<https://stanford.edu/~shervine/blog/pytorch-how-to-generate-data-parallel>

GITHUB_LINK

https://github.com/ChengHao1211/NLE_project