For annotation, I used the model I trained on the previous class to label the competition data. The logic is the same as my college, so I would highlight some differences.

Preprocessing:

- using the TweetTokenizer from nltk
- using the twitter **GloVe** embedding dictionary to replace_token_with_index

https://nlp.stanford.edu/projects/glove/

The CNN model summary:

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
features (Embedding)         (None, 30, 25)            3117200
_________________________________________________________________
conv1d (Conv1D)              (None, 29, 64)            3264
_________________________________________________________________
global_max_pooling1d (Global (None, 64)                0
_________________________________________________________________
dense (Dense)                (None, 100)               6500
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 101
=================================================================
Total params: 3,127,065
Trainable params: 3,127,065
Non-trainable params: 0
```

The performance of this model:

Classification Report

```
print(classification_report(y_test_1d, y_pred_1d))
```

```
              precision    recall  f1-score   support

    NEGATIVE       0.78      0.78      0.78    159494
    POSITIVE       0.78      0.78      0.78    160506

    accuracy                           0.78    320000
   macro avg       0.78      0.78      0.78    320000
weighted avg       0.78      0.78      0.78    320000
```

Accuracy Score

```
accuracy_score(y_test_1d, y_pred_1d)
```

```
0.78078125
```

Summary:

- In practice, GloVe performs almost the same with Word2Vec
- Accuracy score is almost the same with LSTM

Output:

	text	label	score	elapsed_time
0	hey swissborg like article coindesk simple fai...	POSITIVE	0.887344	0.028207
1	global insight survey finding highnetworth ind...	NEUTRAL	0.566071	0.047597
2	digital evolution wealth management emerging t...	NEUTRAL	0.610986	0.066294
3	rise roboadvisers uae lowcost platform targeti...	POSITIVE	0.830619	0.082792
4	never get second chance brand	NEGATIVE	0.398817	0.098139
...	...	...	...	...
11803	secure money advisor premier retirement planni...	NEUTRAL	0.552806	189.579415
11804	million woman could saving adequately retireme...	NEGATIVE	0.136557	189.597675
11805	million already investing w cbinsights mikequi...	NEUTRAL	0.556318	189.615027
11806	million already investing w cbinsights mikequi...	NEUTRAL	0.556318	189.631894
11807	wealth private client investment manager offer...	NEUTRAL	0.477278	189.649198

11808 rows × 4 columns

Rough look, neutral tweet is easier to classify positive or negative, compared LSTM.

Next Step:

- Correct Annotation: compared with the LSTM model, to find out the different annotations, and then manually annotate tweet.
- Optimize the model: because the model training take a lot of time, I have not finished gird researching.