Kaggle
Quora-insincere-questions-classification

HELLO!

I am Julie

I am here because I love Kaggle Competition.

You can find me at yanjialie@gmail.com

1.

Analyize Data

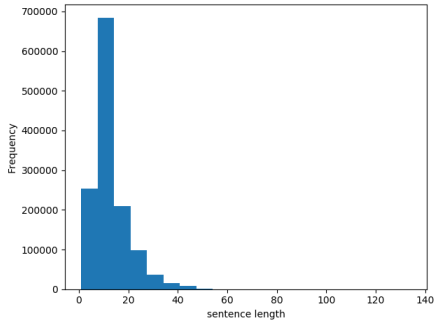
I. Text Data
   - Total unique words – lower case
   - Imbalanced/Balanced data
   - Clean data, e.g. remove punctuation

II. Four different embedding provided
   - Coverage

- Total unique words after cleaning: 52075
- Imbalanced data:

Ones	6.1870%
Zeros	93.8130%

- Sentence length distribution



- Embedding coverage

```
glove  vocab coverage 51.4034%
glove  words coverage 98.8647%
parag  vocab coverage 60.6852%
parag  words coverage 99.1889%
wiki  vocab coverage 38.8647%
wiki  words coverage 98.1467%
google  vocab coverage 31.2295%
google  words coverage 88.0577%
```

2.

Construct Neural Network

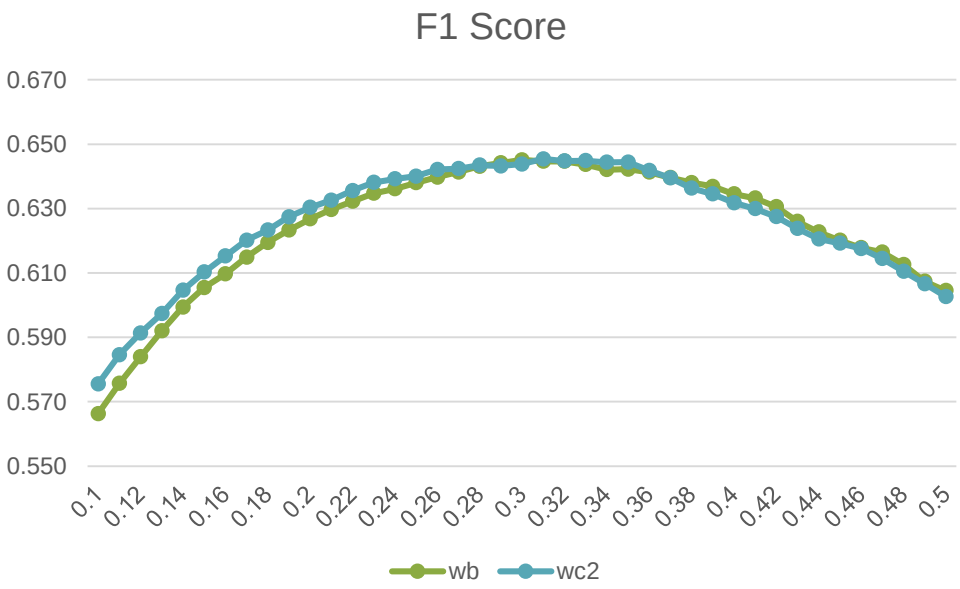
Models

- Logistic Regression
- Simple RNN
- Attention
- LSTM
- GRU

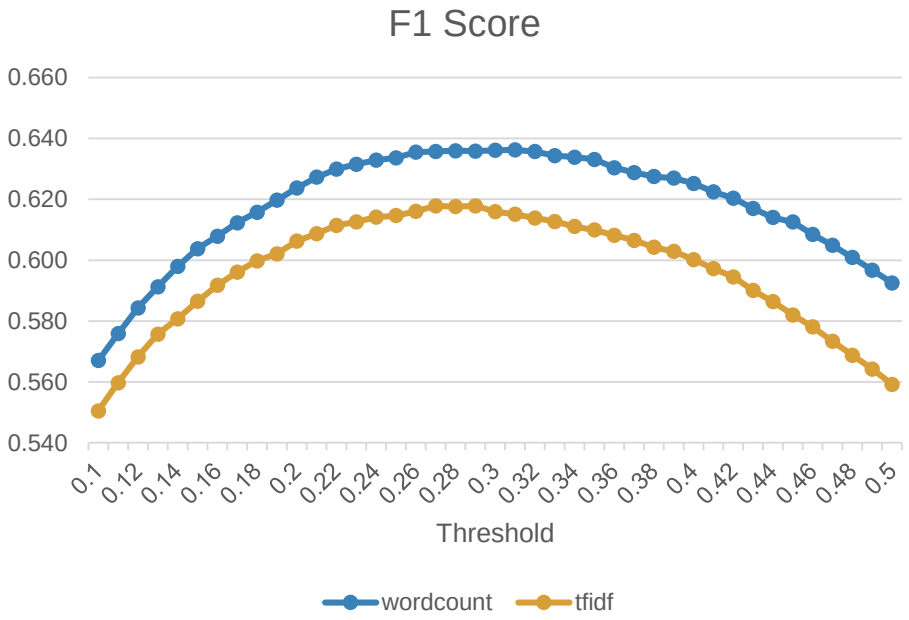
Features

- TF-IDF
- Word count
- Glove
- GoogleNews
- Paragram
- WikiNews
- Heuristic features

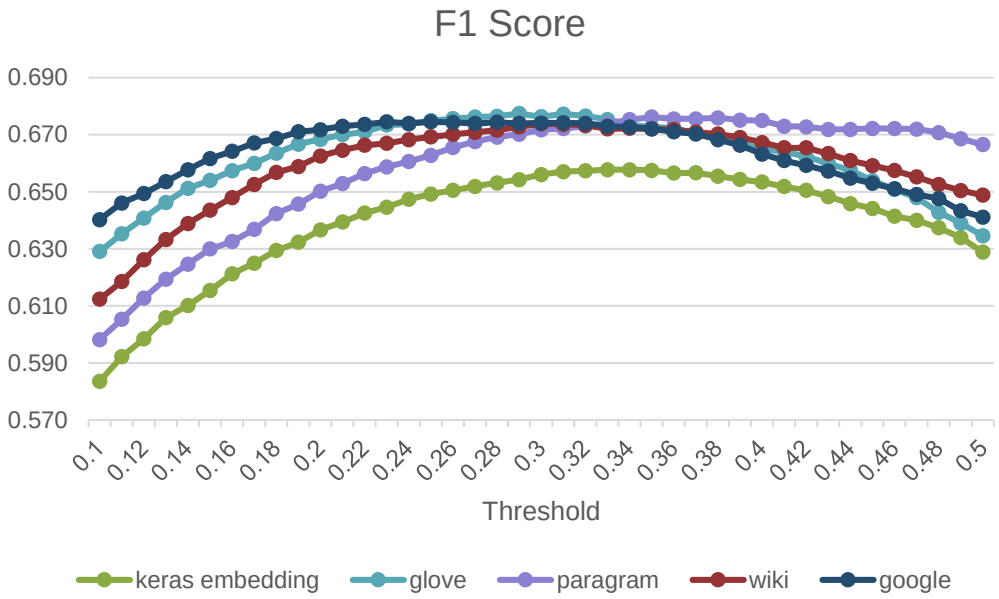
Word Binary V.S. Word Count



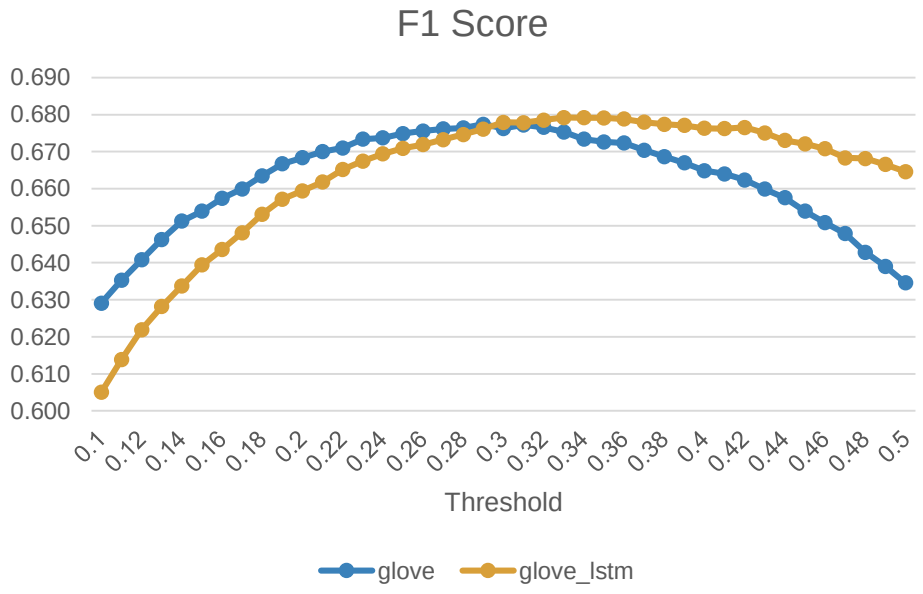
Word Count V.S. TFIDF



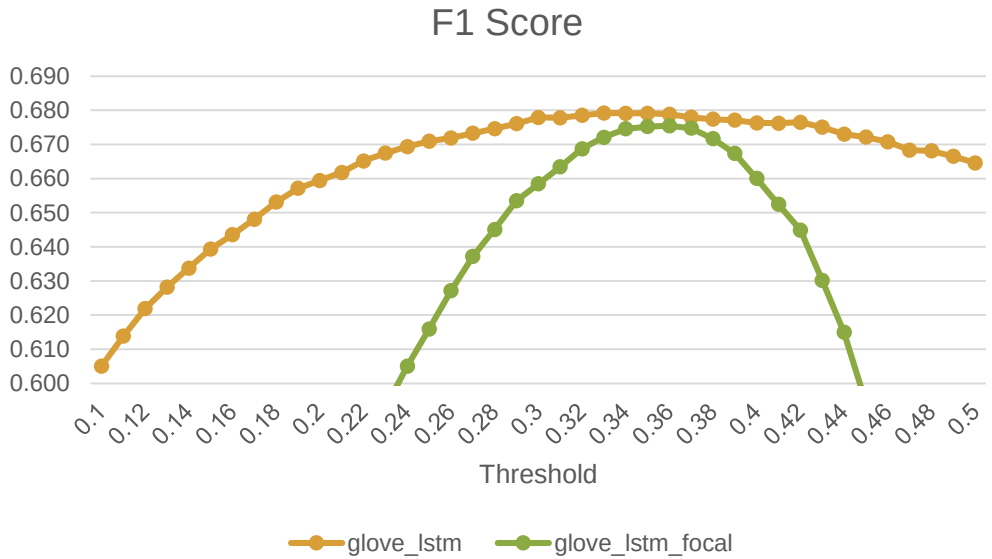
Five Embedding Comparison



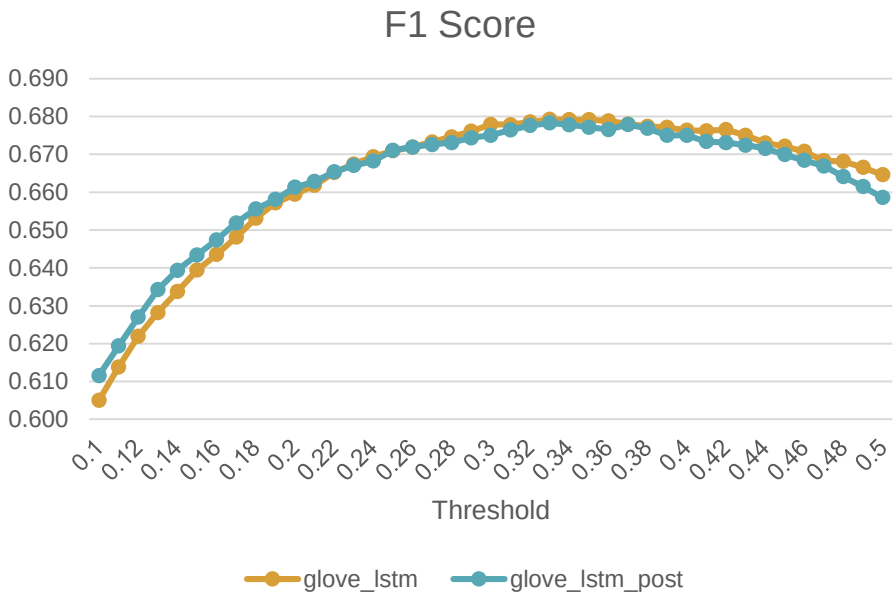
GRU V.S. LSTM


F1 Score
0.690
0.680
0.670
0.660
0.650
0.640
0.630
0.620
0.610
0.600
0.1
0.12
0.14
0.16
0.18
0.2
0.22
0.24
0.26
0.28
0.3
0.32
0.34
0.36
0.38
0.4
0.42
0.44
0.46
0.48
0.5
Threshold
glove
glove_lstm

Cross Entropy Loss V.S. Focal Loss

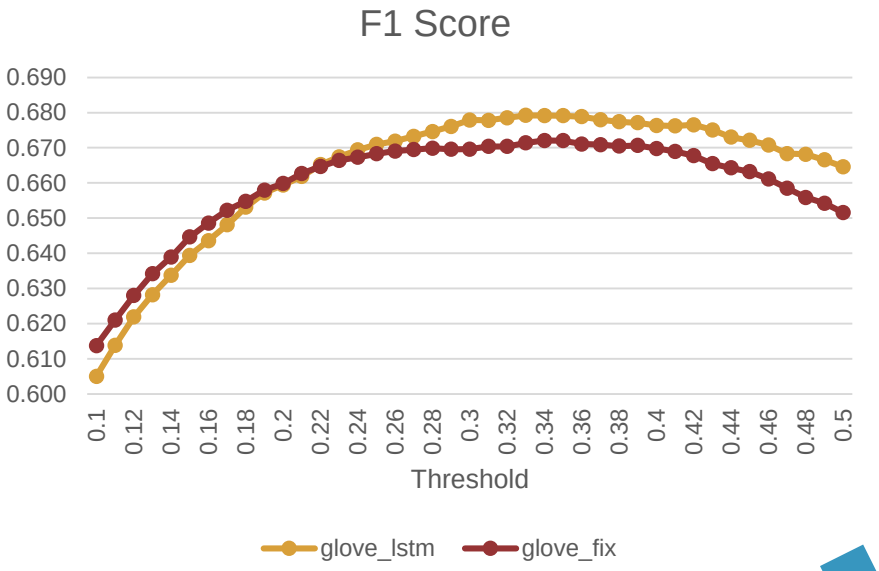


Pre-padding V.S. Post-padding

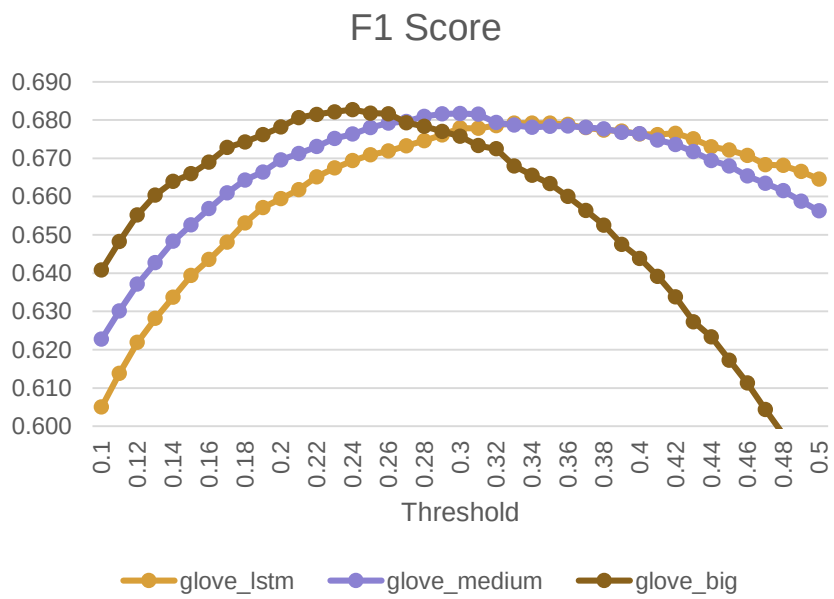


Train Embedding V.S. Not Train Embedding

- Not train embedding - much faster – less accuracy



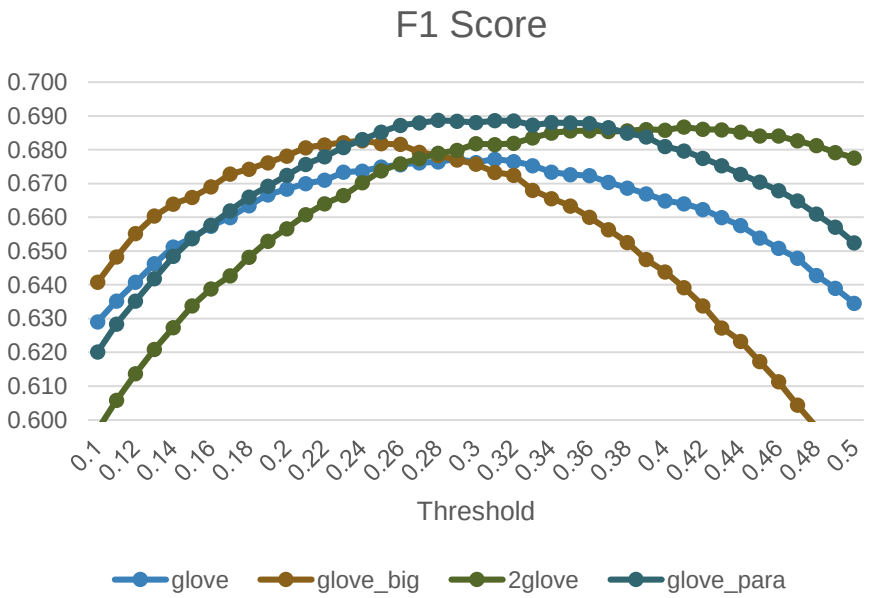
Small, Medium, Big NN




```
inp = Input(shape=(maxlen,))
x = Embedding(max_features, embed_size,
weights=[embedding_matrix])(inp)
x = Bidirectional(CuDNNLSTM(64, return_seq
x = GlobalMaxPool1D()(x)
x = Dense(16, activation="relu")(x)
x = Dropout(0.1)(x)
x = Dense(1, activation="sigmoid")(x)
```

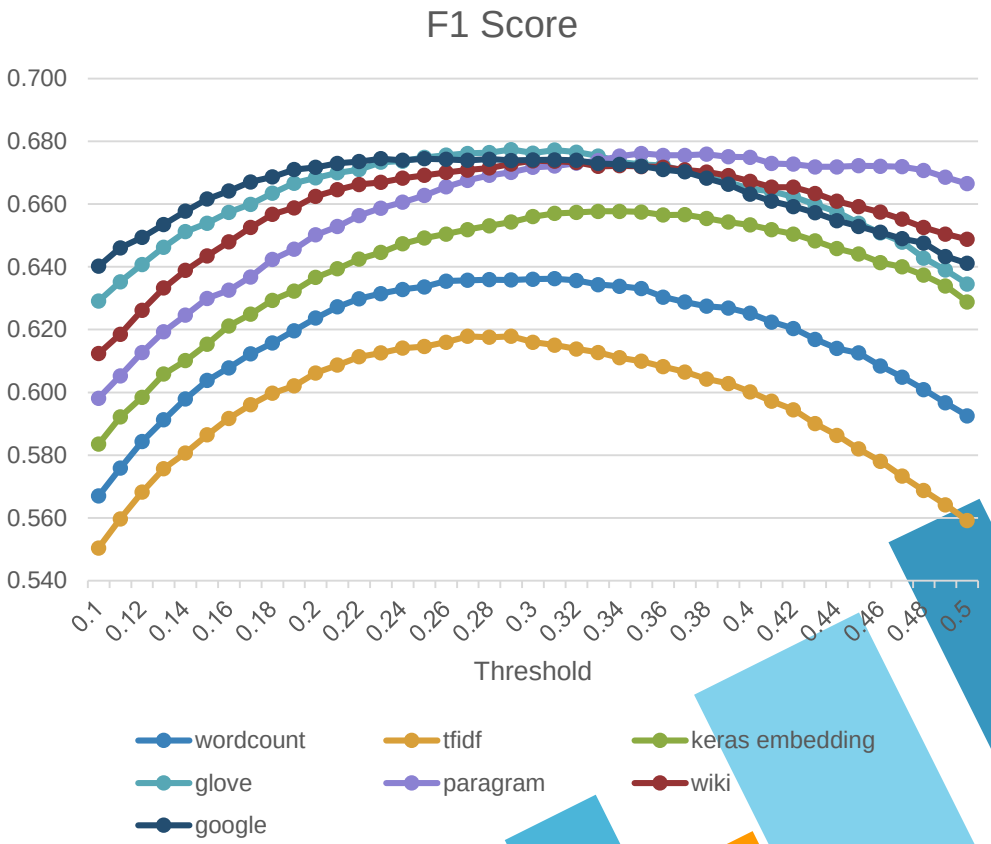
```
inp = Input(shape=(maxlen,))
x = Embedding(max_features, embed_size,
weights=[embedding_matrix])(inp)
x = Bidirectional(CuDNNLSTM(128, return_se
x = Conv1D(64,kernel_size=3, padding='same
x = GlobalMaxPool1D()(x)
x = Dense(16, activation="relu")(x)
x = Dropout(0.1)(x)
x = Dense(1, activation="sigmoid")(x)
```

Glove + Paragram 600 feature as input



Model Comparison

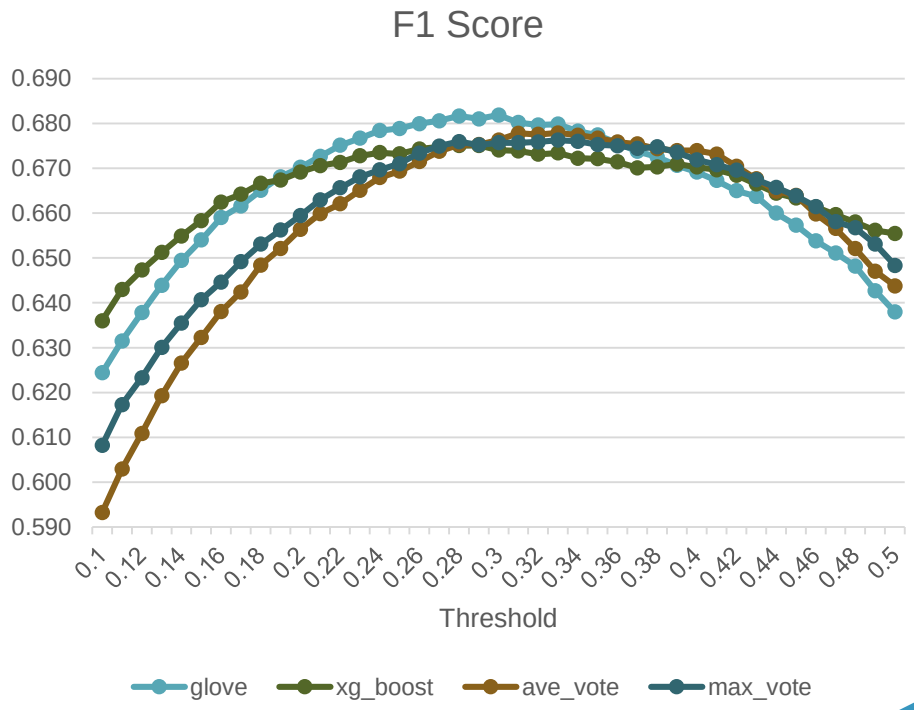
- Word count + Logistic regression
- TFIDF + Logistic regression
- Keras Embedding + Bidirectional GRU
- Glove Embedding + Bidirectional GRU
- GoogleNews Embedding + Bidirectional GRU
- Paragram Embedding + Bidirectional GRU
- WikiNews Embedding + Bidirectional GRU



3.

Ensemble Different Models

Average Vote/Max Vote/ XG-Boost



4.

Error Analysis

Most Errors in the following topics

- Religions
- Politics, government, country
- Race
- Sex
- History, war

question_text	target
My IQ is over 160, why don't my classmates tre...	1
A very nice Chinese couple was interested in a...	1
Why did Hitler executed Jews with gas, ovens, ...	0
What criminal activities has Vladimir Putin ta...	1
How come hormone treatment changes sexual orie...	1
Is there a term for those who "hope for a verd...	1
What did Trump learn from Hitler?	1
How should liberals respond to lies being told...	0
What is stopping the Hindu Brahmins from conve...	0
Isn't more likely that Iran didn't attack Isra...	0
Who thinks they can trump the Trump?	0
Do ribbed condoms clean your sinuses better th...	1
Why do Jains have rules which sometimes have n...	1
Let's say that 99% of British Muslims condemn ...	0
How would you survive on a deserted island wit...	1
Didn't God order us to have and enforce only H...	1
Why don't Trump supporters stop purchasing cor...	0
Why do Spain, Portugal, some Latin countries, ...	0
Is it true that to become a member of Quora ma...	1
I am depressed .Why has the professor to treat...	1
Can you create a horse using a Christmas tree,...	1
Will Prince Charles have to abdicate the thron...	1
Why don't we start investing on transfering th...	0
Why do so many women are being abused every day?	0
Why is my vegan friend angry at me after his g...	1
Why are terrorists attacking the UK?	0
I want to make this a post. I think everyone o...	0
Are there more witnesses for aliens than for t...	0

5.

Summary

What works	What does not work
Clean text/data	Raw data
Embedding, better than word count, word binary	TFIDF
Train pre-trained embeddings	
Cross-Entropy Loss	Focal Loss
LSTM, Maxpool1D, CON1D, better than GRU and simple logistic regression	
Combined embedding, more features in the same neural network	Complex and bigger neural network, more layers about GRU and LSTM.
	Ensemble model, XG-boost, max vote, avg vote only improves a little
	Extra statistical features # the number of words # the number of unique words # the number of characters # the number of upper characters Only improve a little.

Have not tried

- Upsampling, downsampling
- Concatenate all four embeddings together as input.
- Training in specific the topics, such as religion, where most errors are.
- More epochs, only two epochs applied.
- Train all data.

THANKS!

Any questions?

You can find me at

- yanjialie@gmail.com

Reference

- https://www.kaggle.com/c/quora-insincere-questions-classification/
- https://www.kaggle.com/c/quora-insincere-questions-classification/discussion
- https://www.kaggle.com/c/quora-insincere-questions-classification/kernels