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.

Analyze Data

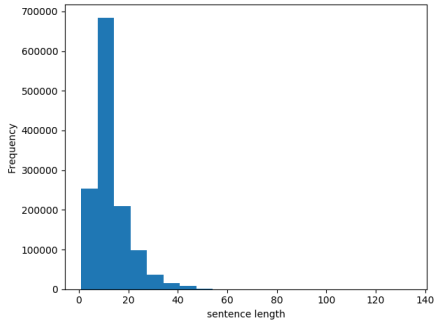
I. Text Data
   - Total unique words – lower case
   - Imbalanced/Balanced data
   - Clean data, e.g. remove punctuation

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

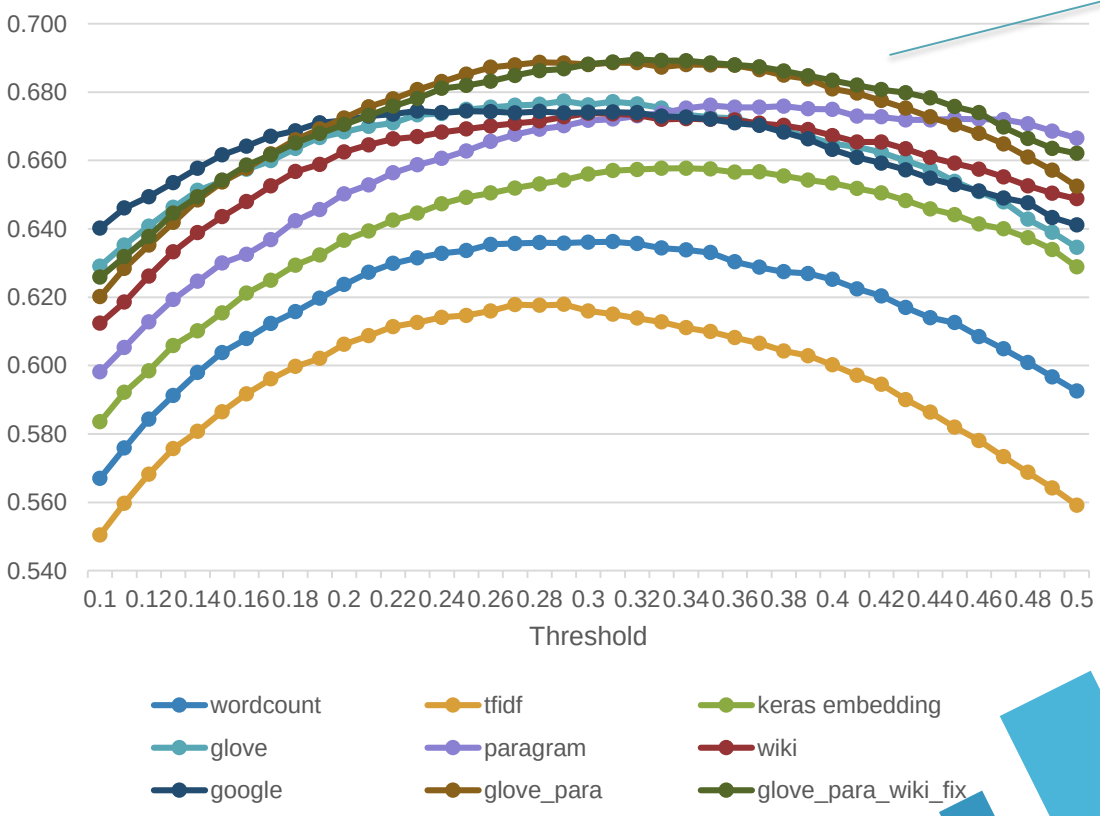
Try Different Models and Different Features

- Word Binary, Word Count, TFIDF
- Five Different Embedding
- GRU, LSTM
- Cross Entropy Loss, Focal Loss
- Train Embedding, Not Train Embedding(fix embedding)
- Pre-padding text, Post-padding text
- Small, Medium, Big NN

F1 Score Comparison

Model	F1 Score
wordcount	0.6454
wordbinary	0.6451
tfidf	0.6179
keras embedding	0.6577
glove	0.6774
paragram	0.6761
wiki	0.6738
google	0.6745
glove_lstm	0.6792
glove_lstm_focal	0.6755
glove_lstm_post_padding	0.6783
glove_medium_nn	0.6817
glove_big_nn	0.6827
glove_fix_embedding	0.6721
glove_big_nn_fix_embedding	0.6800
glove_para	***0.6887***
glove_para_fix_embedding	0.6848
glove_para_wiki_fix	***0.6896***

F1 Score in different threshold



Best models:
glove_para
glove_para_wiki

Observations

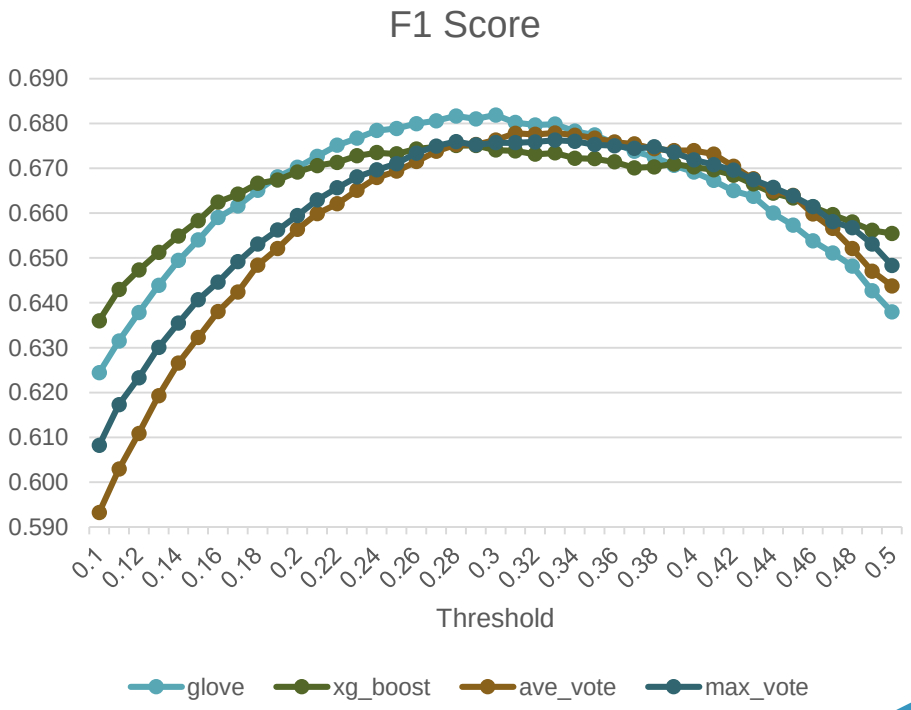
- F1(Word Binary) ≈ F1(Word Count)
- F1(Word Count) > F1(TFIDF)
- F1(Glove) > F1(Paragram) > F1(Google) > F1(Wiki) > F1(Random)
- F1(LSTM) > F1(GRU)
- F1(Cross Entropy Loss) > F1(Focal Loss)
- F1(Train Embedding) > F1(Not Train Embedding)
  time(Train embedding) > time(Not Train Embedding)
- F1(Pre-padding text) ≈ F1(Post-padding text)
- F1(Big NN) > F1(Medium NN) ≈ F1(Small NN)

**F1(xxx) mean xxx model's F1 score.*
**time(xxx) means time consumption for training xxx model.*

3.

Ensemble Different Models

Average Vote/Max Vote/ XG-Boost



Observations

- Max vote, average vote, XG-Boost are not significantly better than single classifier.
- For XG boost, further fine-tune hyper parameter are needed to achieve performance gain.
- Try other classifier:
  - Multinomial naïve bayes
  - Guassian naïve bayes
  - Support vector classifier
  - Multi-layer perceptron classifier
  - AdaBoostClassifier
  - RandomForestClassifier
  - GradientBoostingClassifier
  - LogisticRegression

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

- Clean text/data
- Embedding, better than word count, word binary
- Train pre-trained embeddings
- Cross-Entropy Loss
- LSTM, Maxpool1D, CON1D, better than GRU and simple logistic regression
- Combined embedding, more features in the same neural network

What does not work

- TFIDF
- Focal Loss
- 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.
- Build up a specific word dictionary related to religion, race, sex, and etc. and create new features according to the dictionary.
- Specific training in topics, such as religion, race, sex, and etc.
- 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