Kaggle Competition Report

林弥瑋 110062665

I. Preprocessing

Tweets are more elusive than regular text data. So I focus on cleaning and stemming.

First, I dealt with emojis. Since emoji could directly express emotion, they shouldn't be deleted. I converted them to computer-readable symbols.

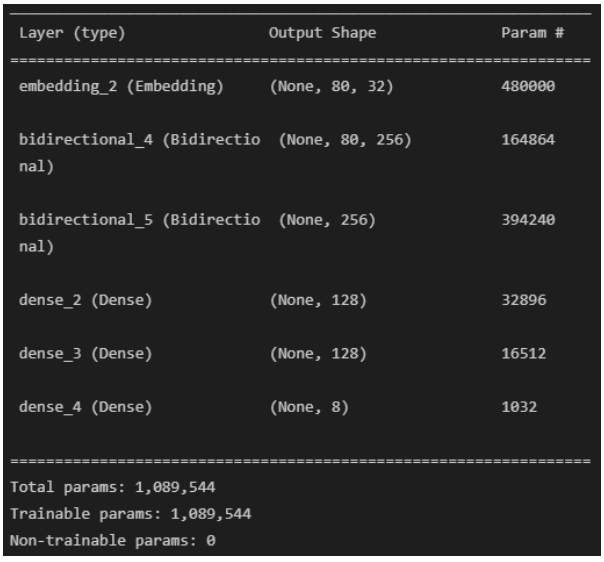
Second, I remove stopwords and punctuation. Considering tweets wouldn't follow grammar carefully, removing them might be a better choice. In addition, length of tweets is limited. Deleting stopwords and punctuation wouldn't change semantics seriously.

Third, I do stemming and lemmatization. But result is not ideal, internet slang affect stemming a lot.

Final step is tokenization and padding. Converting text to appropriate input data.

II. Model

I use bidirection LSTM as main part of my model. Then, I add 3 dense layers to associate relations between features.



Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 80, 32)	480000
bidirectional_4 (Bidirectional)	(None, 80, 256)	164864
bidirectional_5 (Bidirectional)	(None, 256)	394240
dense_2 (Dense)	(None, 128)	32896
dense_3 (Dense)	(None, 128)	16512
dense_4 (Dense)	(None, 8)	1032

Total params: 1,089,544
Trainable params: 1,089,544
Non-trainable params: 0

III. Discussions

Final score is 0.45. Strangely, result without preprocessing is better. I probably take too much information off. As hashtag, it might be an important part of sentiment.

I’ve tried BERT as my second method, but the model is too heavy to train on my own hardware. Reducing model and data is one solution. Otherwise the performance decrease a lot.