The Disaster Tweets - Text Classification

CAI Shizhan, SONG Wenxin

Introduction

Twitter has become a common platform where people report emergencies they are encountering in real-time. Consequently, more agencies are now monitoring Twitter to detect disasters. However, the line between reporting real disasters and tweeting metaphorically is subtle. This project aims at building machine learning models to distinguish whether a Tweet is describing a real disaster or not.

Dataset

- Exploration

Figure 1 and 2 show the top 10 frequent keywords and locations. Most keywords are concern with disaster and corresponding emotion. Most locations are related to the US.

- Cleaning and Processing

Since

(1) *keyword* usually appear in the text as well,

(2) Around 1/3 of the *location* value is missing,

We decide not to use these two attributes in our models. We processed the text by procedures shown in Figure 3.

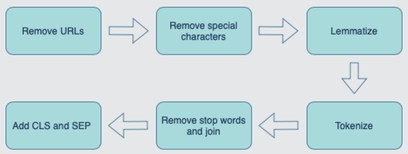

Figure 3. Flow chart of data processing

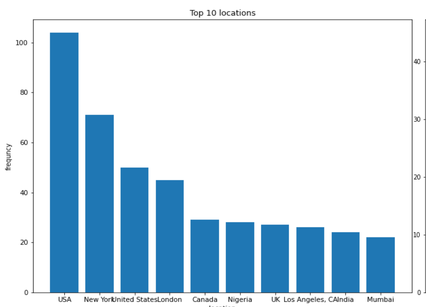

Figure 1. Top 10 frequent keywords

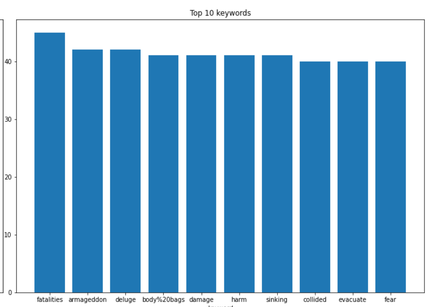

Figure 2. Top 10 frequent locations

Models

To solve the problem, we select several state-of-the-art models and compare their accuracy on the validation set.

RoBERTa

BERT is a bi-directional transformer which has high performance in language representation. Unlike some conventional models such as LSTM which can only learn from the previous word or the next word at one time, It leverages the attention mechanism and transformers to learn from words in all positions of the sentence simultaneously and accurately. RoBERTa is developed based on BERT model. It can deal with a larger dataset, has a bigger batch size and longer training time.

XLNet: Generalized Autoregressive Pretraining for Language Understanding

BERT neglects dependency between the masked positions and suffers from a pretrain-finetune discrepancy. XLNet, a generalized autoregressive pretraining method that (1) enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order and (2) overcomes the limitations of BERT thanks to its autoregressive formulation.

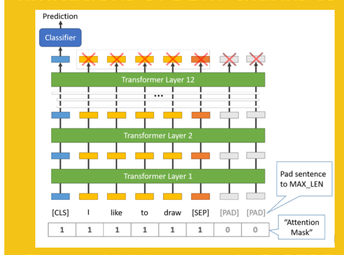

Figure 4. Model structure for RoBERTa

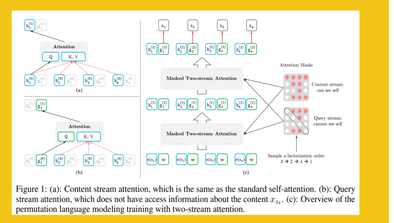

Figure 1: (a): Content stream attention, which is the same as the standard self-attention. (b): Query stream attention, which does not have access information about the content x_{z_t}. (c): Overview of the permutation language modeling training with two-stream attention.

Figure 5. Model structure for XLNet

Comparison

Bert hides a part of the words directly on the input side by introducing the "mask" mark so that these words do not play a role in the prediction. It requires the other words in the context to predict a word that was "masked".

XLNet discards "masking" the words on the input side. Some words are randomly "masked" inside the Transformer through the Attention Mask mechanism. Therefore, the words are dropped by "mask" having no effect on predicting a word.

In essence, these two methods are not much different but the location of "mask". Bert is more superficial, and XLNet hides this process inside the Transformer.

Results

[1]Roberta: Training accuracy: 80%, Kaggle score: 0.81458

[2] XLNet-Remove stop words: Training accuracy: 90.95%; Kaggle score: 0.80263

XLNet: Training accuracy: 92%; Kaggle score: 0.82041

Conclusion

BERT and XLNet have their own advantages. XLNet had increased the amount of pre-training data, using more cleaned data than BERT. For this assignment, XLNet performed better since it had achieved SOTA on the task of text classification. As for more tasks, it is necessary to do hyperparameter tuning and compare the results to determine the performance.

Limitations

1. *keyword* and *location* attributes also contain some information that might be helpful in classification. Adding them into the model may improve accuracy
2. During training the XLNet model, we had to use the basic model rather than the large one due to the limitation of GPU memory. Theoretically speaking, the result could be better using the “XLNet-Large-Cased”.

Reference

1. Building State-of-the-Art Language Models with BERT
2. RoBERTa: A Robustly Optimized BERT Pretraining Approach 1907.11692
3. XLNet: Generalized Autoregressive Pretraining for Language Understanding. https://arxiv.org/abs/1906.08237

Contribution

- CAI Shizhan : Data processing/ Modelling (XLNet models)/ Poster Making
- SONG Wenxin: Data processing/ Modelling (BERT-based models)/ Poster Making