DISASTER TWEETS PREDICTION USING BERT

RAN LIU

Abstract. Twitter has become an important communication channel in times of emergency. The ubiquitousness of smartphones enables people to announce an emergency they're observing in real-time. Because of this, more agencies are interested in programatically monitoring Twitter (i.e. disaster relief organizations and news agencies). But, it's not always clear whether a person's words are actually announcing a disaster or just an exaggerated expression. To handle this problem, we use BERT as classifier to identify whether a tweet is about disaster or not. Our method has achieved good results, ranking top 2.2% in Kaggle competition.

Contents

Date: 2022-11-16.

Key words and phrases. Deep Learning, Pretrained Language Model, Natural Language Processing.

1. Introduction

Twitter has become an important communication channel in times of emergency. But it's not always clear whether a person's words are actually announcing a disaster. For instance, "*Look at the sky last night it was ablaze!* " In this tweet, The author explicitly uses the word "*ablaze*" which is related to disaster, but actually it is an exaggerated expression. Our goal is to resolve this problem.

Concretely, given a set of labeled data, we will use them to train a text classifier and use it to predict whether a tweet is about disaster or not.

In this paper, we first analyze the training set, including calculate distribution of characters and tokens of texts, then convert texts into standard input format which BERT can process. After that, we fine-tune BERT to adapt to disaster tweets prediction task. Our method has achieved good results, ranking top 2.2% in Kaggle competition.

2. Related Work

Rule-Based methods classify texts into different categories using a set of pre-defined rules. This kind of methods are easy to implement and fast when running, also have good interpretability, while require a lot of manpower and time. What's worse, when facing a new problem, previous rules may become useless.

Statistical methods, such as Naïve Bayes, support vector machines, hidden Markov model, and random forests, are more accurate than rule-based methods. On the other hand, statistical methods cannot take full advantage of large training data because the features are pre-defined.

Deep learning which is represented by Convolutional Neural Network and Long Short-Term Memory Network, is the current mainstream method. It has strong ability to capture deep contexual features and can improve performance obviously. But weak interpretability and extreme reliance on large amount of training data are its main drawbacks.

As for disaster tweets prediction task, issues to be resolved are how to capture deep features and improve generalization of model. What's more, because of limited labeled data, we cannot train a model from scratch. Although dataset is small, performance of model still needs to meet the requirements.

3. Dataset

The training set has a total of 7631 tweets, consisting of 3721 tweets about disaster and 3892 tweets which are not about disaster. Structure of data can be seen in table 1

The maximum character length of text in training set is 157 and minimun is 7, with an average of 101 characters.

Since BERT takes token as word vector unit, we also calculate token statistics of training set. The maximum token length of text is 84 and minimun is 3, with an average of 33 tokens.

Distribution of character and token length are shown in fig. 1 and fig. 2.

TABLE 1. Data structure

Term	Example
id	210
keyword	airplain accident
location	Eagle Pass, Texas
text	A Cessna airplane accident in Mexico...
label	1

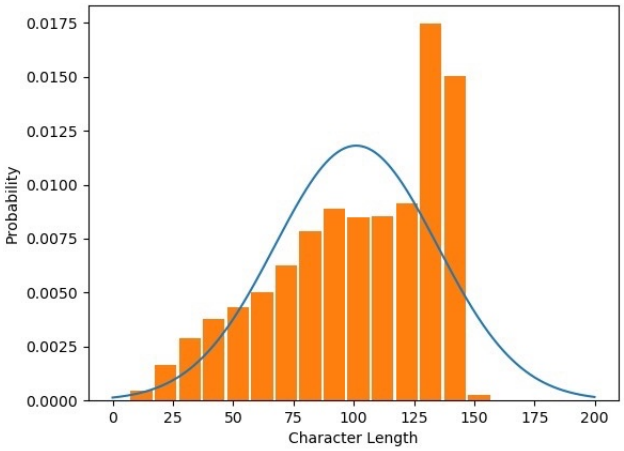

FIGURE 1. Distribution of Character Length

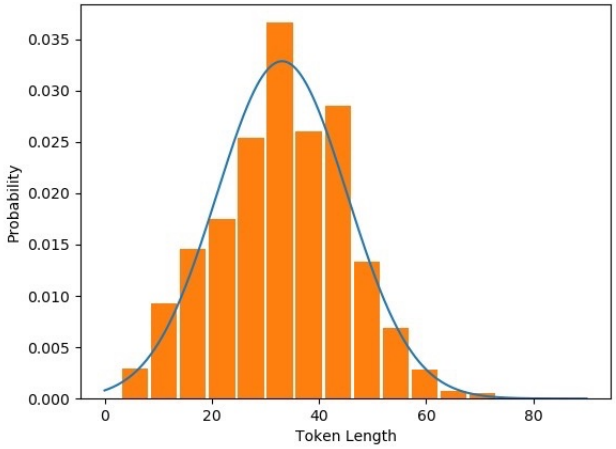

FIGURE 2. Distribution of Character Length

TABLE 2. Model details

Model	Layer	Hidden	Attention	Mask
bert-base-cased	12	768	12	Token
bert-base-uncased	12	768	12	Token
bert-large-cased	24	1024	16	Token
bert-large-uncased	24	1024	16	Token
bert-large-wwm-cased	24	1024	16	Span
bert-large-wwm-uncased	24	1024	16	Span

TABLE 3. Training set

Name	Value
Token length	256
Dropout rate	0.1
Train : Validation	8 : 2
Batch size	16
Number of epochs	3
Optimizer	Adam
β_1	0.9
β_2	0.999
Learning rate	5e-5, 3e-5, 2e-5

4. METHODOLOGY

4.1. **Input format.** In order to convert texts into vectors that BERT can process, we should transform each tweet text into three vector, which are token vector, mask vector, segment vector, respectively.

- **Token vector** represents index of each token according to the vocabulary, the rest is padded with 0.
- **Mask vector** is used to calculate attention score without considering the meaningless part which is padded with 0.
- **Segment vector** is used to split two sentences. For the case there is only one sentence, segment vector is a zero vector.

4.2. **Model details.** We use bert-base and bert-large as our classification model, each of which has cased and uncased versions [2]. Furthermore, we compare with bert that uses whole word mask [1] to verify the effectiveness of this method. Model details are shown in table 2.

5. EXPERIMENTS AND RESULTS

5.1. **Training setup.** Hyparameters and other training setup are the same as specified in [2]. Setup details can be seen in table 3.

TABLE 4. Results

Model	F_1 *score*
bert-base-cased	0.825
bert-base-uncased	0.831
bert-large-cased	0.830
bert-large-uncased	**0.848**
bert-large-wwm-cased	0.828
bert-large-wwm-uncased	0.825

5.2. **Evaluation metric.** We use F_1 *score* as evaluation metric, which is defined in eq. (5.1).

$$F_1 \ score = \frac{2 \ * \ precision \ * \ recall}{precision \ + \ recall} \tag{5.1}$$

5.3. **Results.** In all models used in this paper, the highest F_1 *score* is 0.848 which is obtained by bert-large-uncased. Detailed experimental results are shown in table 4.

5.4. **Rank.** 19/870 (top 2.2%)

6. CONCLUSION

In this paper, we use different versions of BERT as classification models to predict whether a tweet is about disaster or just an exaggerated expression.

Limited by computing resources, models are not fully trained, and no other pre-trained language models are used to compare with BERT.

References

[1] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514, 2021.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100084, China

Email address: liuran@iie.ac.cn