DISASTER TWEETS PREDICTION USING BERT

~~GANG LI~~RAN LIU

ABSTRACT. ~~The abstract will be put here, ...~~ 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.

~~At a high level, what is the problem area you are working in and why is it important? It is important to set the larger context here. Why is the problemof interest and importance to the larger community?~~

~~This paragraph narrows down the topic area of the paper. In the first paragraph you have established general context and importance. Here you establish specific context and background.~~

~~"~~In this paper, we ~~show that ...". This is the key paragraph in the intro - you summarize, in one paragraph, what are the main contributions of your paper given the context you have established in paragraphs 1 and 2. What is the general approach taken? Why are the specific results significant? This paragraph must be really good.~~ 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.

~~You should think about how to structure these one or two paragraph summaries of what your paper is all about. If there are two or three main results, then you might consider itemizing them with bullets or in test.~~

- ~~e. g., First . ..~~
- ~~e.g., Second . ..~~
- ~~e.g., Third . ..~~

~~If the results fall broadly into two categories, you can bring out that distinction here. For example, "Our results are both theoretical and applied in nature. (two sentences follow, one each on theory and application)"~~

2. Related Work

~~Keep this at a high level, you can refer to a future section where specific details and differences will be given. But it is important for the reader to know at a high level, what is new about this work compared to other work in the area.~~ **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.

~~"The remainder of this paper is structured as follows..." Give the reader a roadmap for the rest of the paper. Avoid redundant phrasing, "In Section 2, In section 3, ... In Section 4, ... " etc.~~ **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.

TABLE 1. Data structure

Term	Example
id	210
keyword	airplain accident
location	Eagle Pass, Texas
text	A Cessna airplane accident in Mexico.~~, , , and We have , , the range: . 1/2~~..
label	1

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

~~Test citation [?]. and [?] or ?].~~ 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 **??**

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

~~Number:~~

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 **??** and **??**.

~~For , as shown below:~~

$$a = b \times \sqrt{ab}$$

4. METHODOLOGY

5. ~~PRELIMINARIES~~

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.

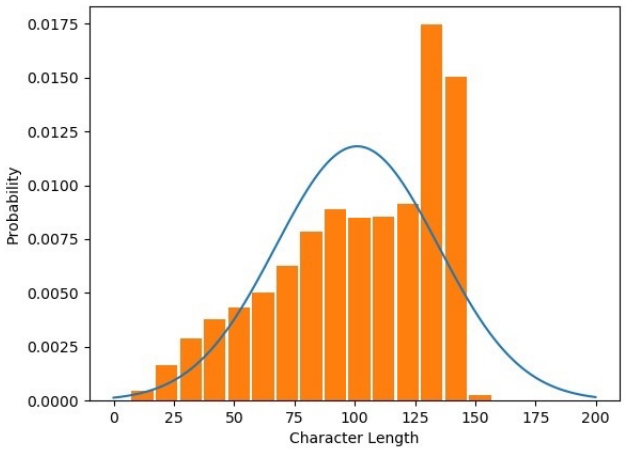

FIGURE 1. Distribution of Character Length

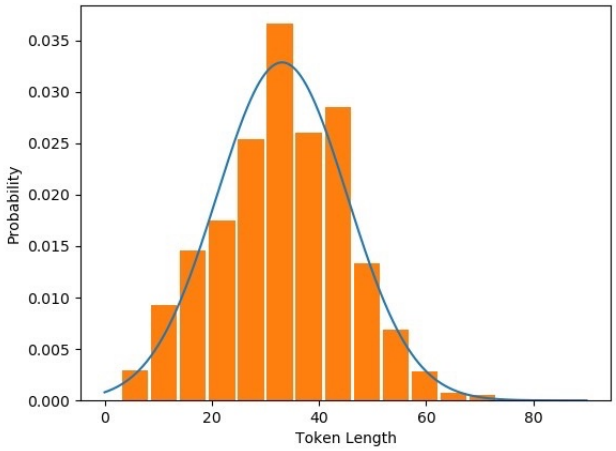

FIGURE 2. Distribution of Character Length

- **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 [**?**]. Furthermore, we compare with bert that uses whole word mask [**?**] to verify the effectiveness of this method. Model details are shown in **??**.

5. ~~METHOD~~

5. EXPERIMENTS AND RESULTS

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. ~~Precision Comparison on Event Detection Methods~~Training set

Name	~~OR Event Detection AC Event Detection TC Event Detection~~ Value
~~precision~~ Token length	~~0.83~~ 256
Dropout rate	~~0.69~~ 0.1
Train : Validation	~~0.46~~ 8 : 2
~~recall~~ Batch size	~~0.68~~ 16
Number of epochs	~~0.48~~ 3
Optimizer	~~0.36~~ Adam
~~F-score~~ β_1	~~0.747~~ 0.9
β_2	~~0.57~~ 0.999
Learning rate	~~0.4~~ 5e-5, 3e-5, 2e-5

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

6. ~~EXPERIMENT AND ANALYSIS~~

6. ~~CONCLUSIONS~~

5.1. **Evaluation metric.** We use F_1 *score* as evaluation metric, which is defined in **??**.

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

5.2. **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 **??**.

~~ACKNOWLEDGEMENT~~

5.3. **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.

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

~~The authors would like to thank ...~~Limited by computing resources, models are not fully trained, and no other pre-trained language models are used to compare with BERT.

Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100084, China
Email address: liuran@iie.ac.cn