Disaster Tweets Prediction Using BERT

Ran Liu

Institute of Information Engineering
Chinese Academy of Sciences

November 16, 2022

Overview

Problem Definition

Disaster Tweets Prediction

- 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.

e.g.

LOOK AT THE SKY LAST NIGHT IT WAS ABLAZE!

- The author explicitly uses the word "ABLAZE" which is related to disaster.
- But it is an exaggerated expression.

Problem Description

Defn

Given a set of labeled data which we will use to train a classifier and use it to predict whether a tweet is about disaster or not.

- The training set was collected from Twitter
- It has been labeled manually.
- A binary classification problem.
- F1 score is evaluation metric.

Related Work and Challenges

Related Work - Text Classification

- Existing Methods - Rule-Based Methods
  - Rule-based methods classify text into different categories using a set of pre-defined rules.

Disadvantages

- Require a deep domain knowledge.
- Require a lot of manpower and time.
- When facing a new problem, previous rules may become useless.

Advantages

- Fast
- Easy
- Interpretable

- Existing Methods - Traditional Machine Learning (Statistical methods)
  - Naïve Bayes, Support Vector Machines, Hidden Markov Model, Random Forests...

Disadvantages

- Reliance on the handcrafted features.
- Cannot take full advantage of large training data because the features are pre-defined.

Advantages

- More accurate than rule-based methods.

Related Work - Text Classification

- Existing Methods - Deep Learning
  - Convolutional Neural Network, Long Short-Term Memory Network...

Disadvantages

- Reliance on large amount of training data.
- Weak Interpretability.

Advantages

- Capture deep contexual features.
- Greatly improve accuracy.

Challenges

- How to capture deep features?
  - Models should have the ability to capture deep features.
  - Generalization of models need to be improved.

Challenges

- Limited dataset.
  - Cannot train a model from scratch.
  - Although dataset is small, performance of model still needs to meet the requirements.

Dataset

Data Details

Table 1: Data structure

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

Data Statistics

- Number of each label in training set.

Table 2: Number of each label

label	Number
1	3721
0	3892
total	7613

Data Statistics

- Character length of text in training set.
  - max: 157 min: 7

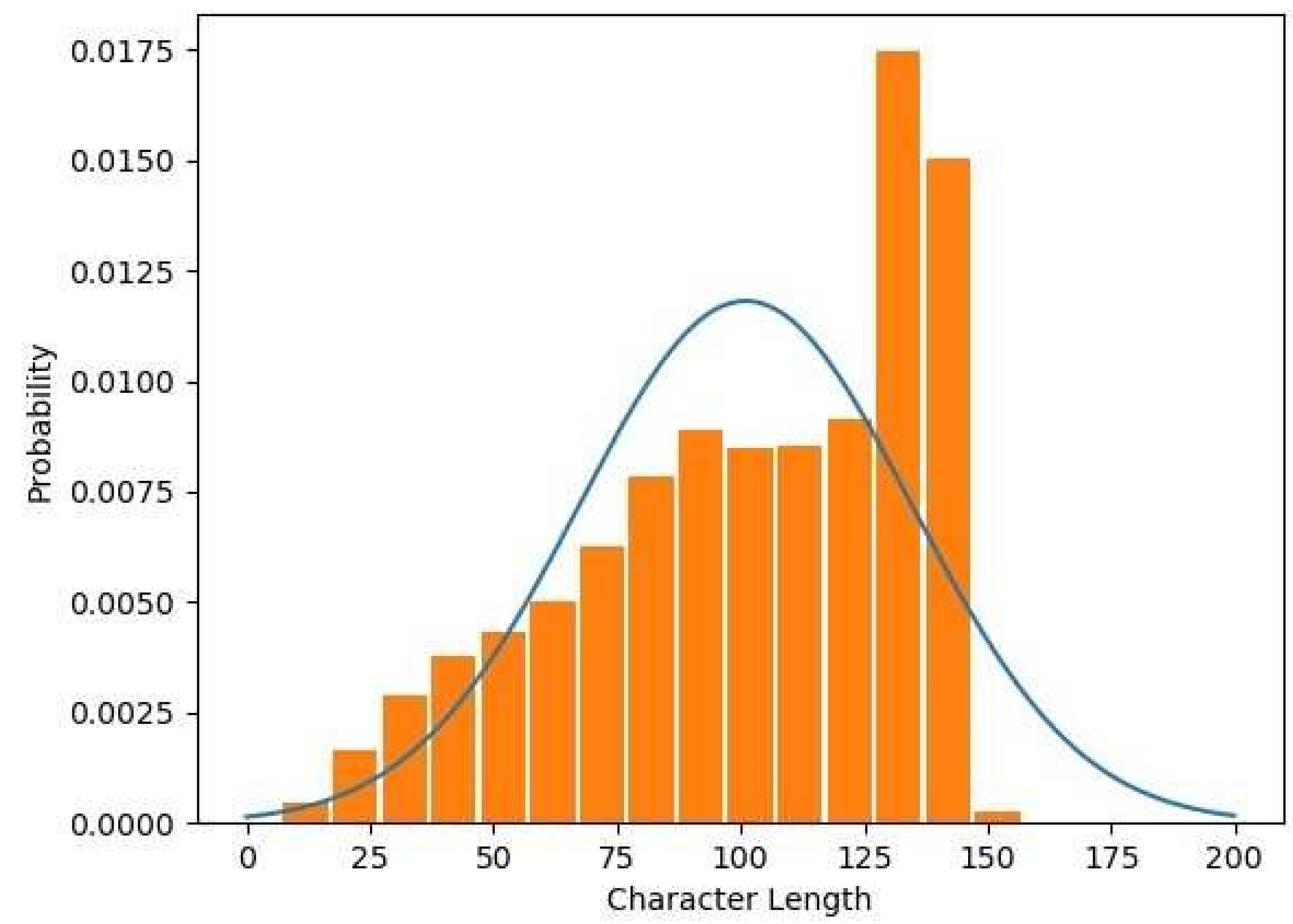

Figure 1: Distribution of Character Length

Data Statistics

- Token length of text in training set.
  - max: 84   min: 3

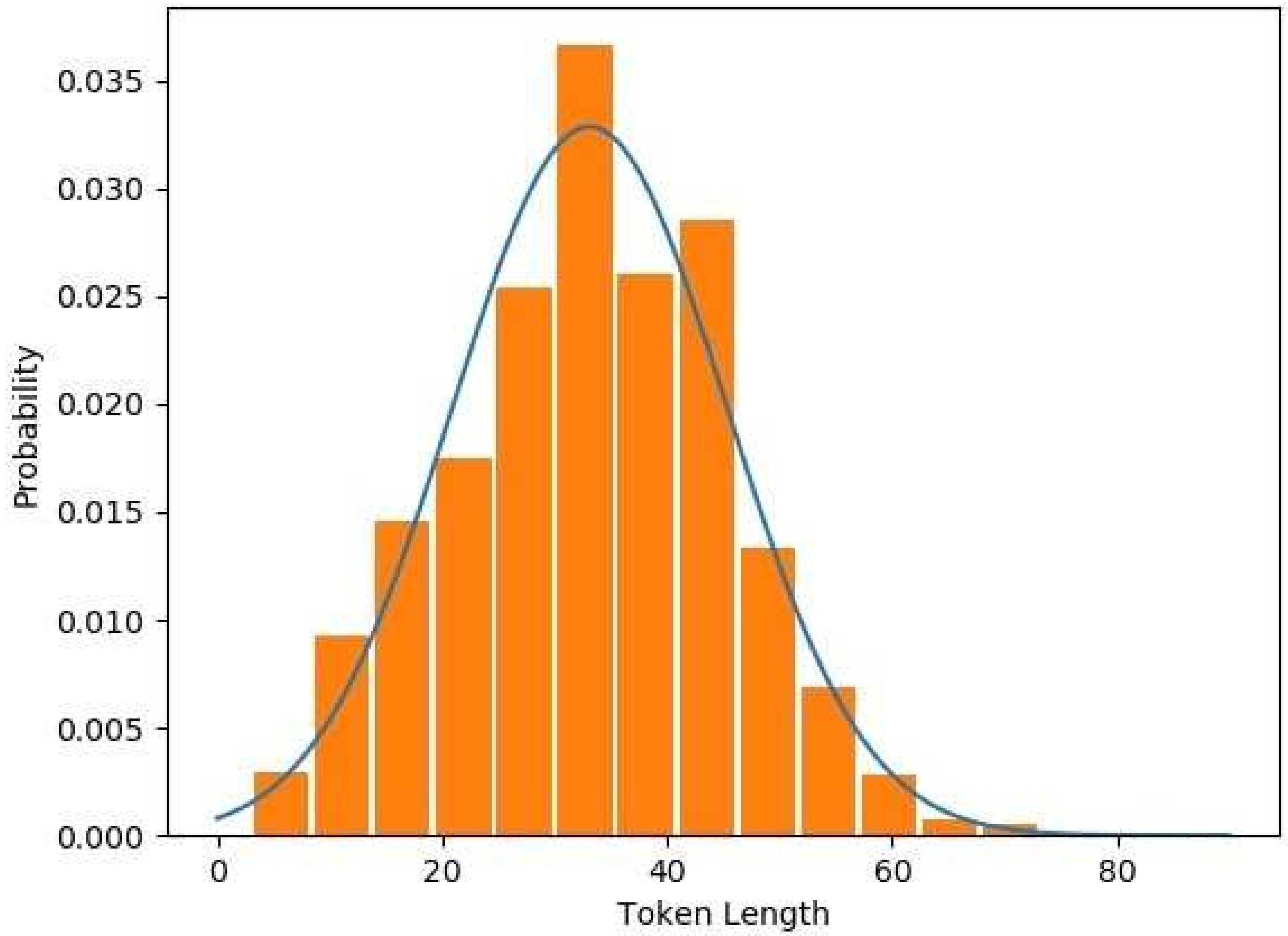

Figure 2: Distribution of Token Length

Methodology

Input Format

- Original text:
  Three people died from the heat wave so far.
- Input format:
  - token vector:
    [101, 2093, 2111, 2351, 2013, 1996, 3684, 4400, 2061, 2521, 102, 0, 0, 0, ...]
  - mask vector:
    [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, ...]
  - segment vector:
    [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]

Model Detils

Table 3: Model details

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

Model Detils

Table 4: Training setup

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

Results

Results

- $$F_1\ score = \frac{2 * precision * recall}{precision + recall}$$
- Rank: 19/870 (2.2 %)

Table 5: 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

Contact Information

Ph.D Student Ran Liu
Institute of Information Engineering
Chinese Academic of Science, Beijing, China

LIURAN@IIE.AC.CN

TEAM FOR UNIVERSAL LEARNING AND INTELLIGENT PROCESSING