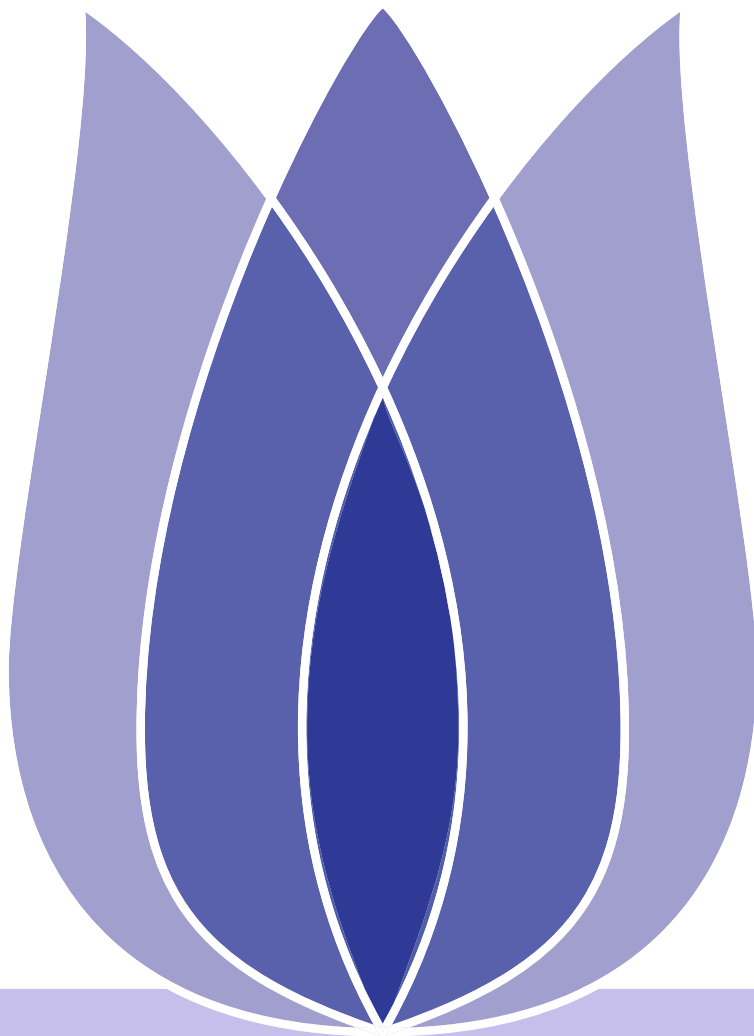


Disaster Tweets Prediction Using BERT

Ran Liu

Institute of Information Engineering
Chinese Academy of Sciences

November 8, 2022





Overview

- [Problem Definition](#)
- [Related Work and Challenges](#)
- [Dataset](#)
- [Methodology](#)
- [Results](#)

Problem Definition

- Disaster Tweets Prediction
- Problem Description

Related Work and Challenges

- Related Work - Text Classification
- Challenges

Dataset

- Data Details
- Data Statistics

Methodology

- Input Format
- Model Detils

Results



Problem Definition

Disaster Tweets Prediction

Problem Description

Related Work and Challenges

Dataset

Methodology

Results

Problem Definition



Disaster Tweets Prediction

Problem Definition
Disaster Tweets Prediction
Problem Description
Related Work and Challenges
Dataset
Methodology
Results

- 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.



TULIP

Team for Universal Learning and Intelligent Processing



Problem Description

- Problem Definition
- Disaster Tweets Prediction
- Problem Description
- Related Work and Challenges
- Dataset
- Methodology
- Results

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.



[Problem Definition](#)

[Related Work and Challenges](#)

[Related Work - Text Classification](#)

[Challenges](#)

[Dataset](#)

[Methodology](#)

[Results](#)

Related Work and Challenges



Related Work - Text Classification

Problem Definition
Related Work and Challenges
Related Work - Text Classification
Challenges
Dataset
Methodology
Results

■ 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





Related Work - Text Classification

- [Problem Definition](#)
- [Related Work and Challenges](#)
- [Related Work - Text Classification](#)
- [Challenges](#)
- [Dataset](#)
- [Methodology](#)
- [Results](#)

- 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

Problem Definition
Related Work and Challenges
Related Work - Text Classification
Challenges
Dataset
Methodology
Results

■ Existing Methods - Deep Learning

◆ Convolutional Neural Network, Long Short-Term Memory...

Disadvantages

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

Advantages

- ◆ Capture deep contextual features.
- ◆ Greatly improved accuracy.





Challenges

- [Problem Definition](#)
- [Related Work and Challenges](#)
- [Related Work - Text Classification](#)
- [Challenges](#)**
- [Dataset](#)
- [Methodology](#)
- [Results](#)

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



Challenges

Problem Definition
Related Work and Challenges
Related Work - Text Classification
Challenges
Dataset
Methodology
Results

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





[Problem Definition](#)

[Related Work and Challenges](#)

[Dataset](#)

[Data Details](#)

[Data Statistics](#)

[Methodology](#)

[Results](#)

Dataset



- [Problem Definition](#)
- [Related Work and Challenges](#)
- [Dataset](#)
- [Data Details](#)**
- [Data Statistics](#)
- [Methodology](#)
- [Results](#)

Table 1: Data structure

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



- [Problem Definition](#)
- [Related Work and Challenges](#)
- [Dataset](#)
- [Data Details](#)
- [Data Statistics](#)
- [Methodology](#)
- [Results](#)

- Number of each label in training set.

Table 2: Number of each label

label	Number
1	3721
0	3892
total	7613



- [Problem Definition](#)
- [Related Work and Challenges](#)
- [Dataset](#)
- [Data Details](#)
- [Data Statistics](#)**
- [Methodology](#)
- [Results](#)

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

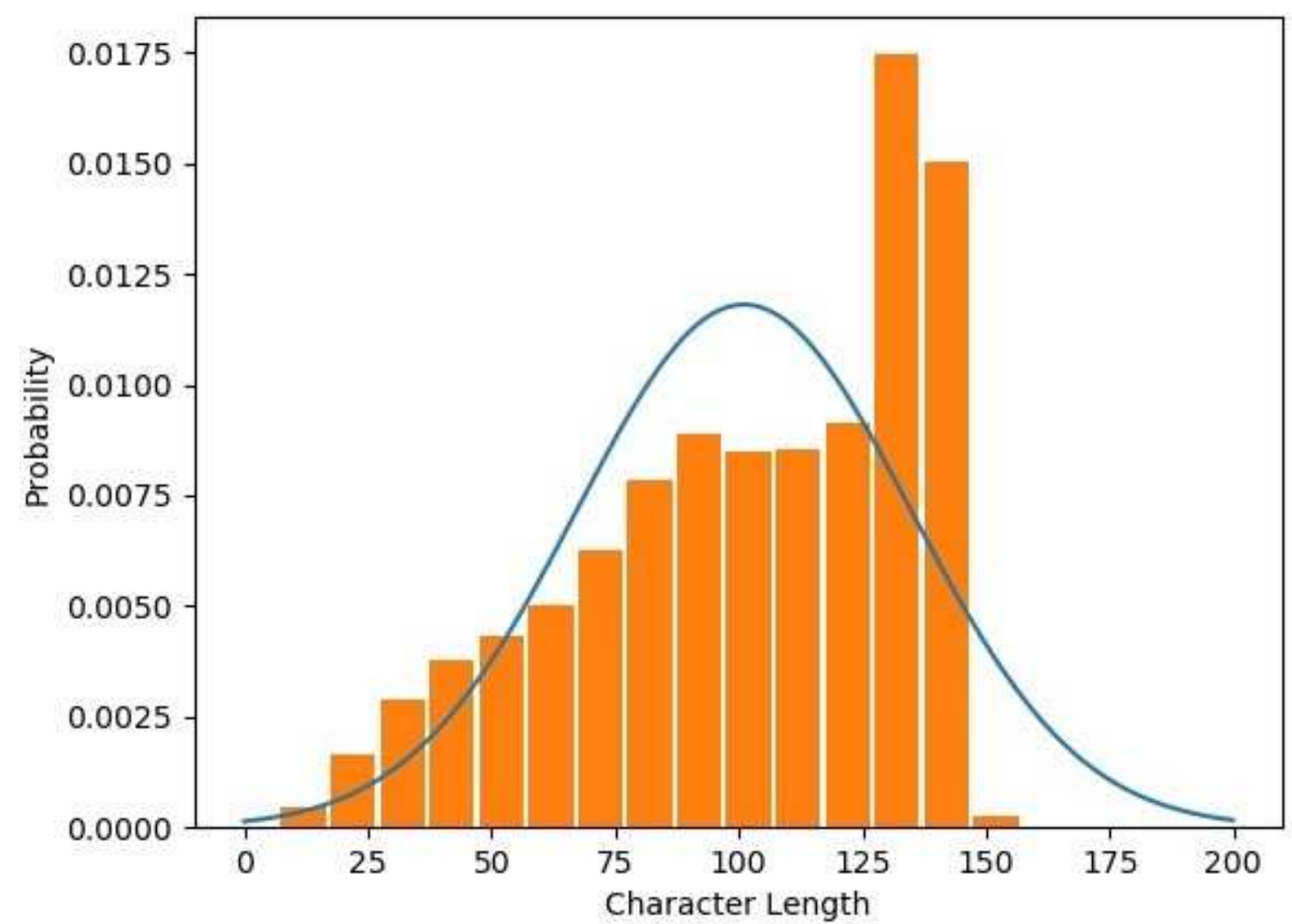


Figure 1: Distribution of Character Length

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

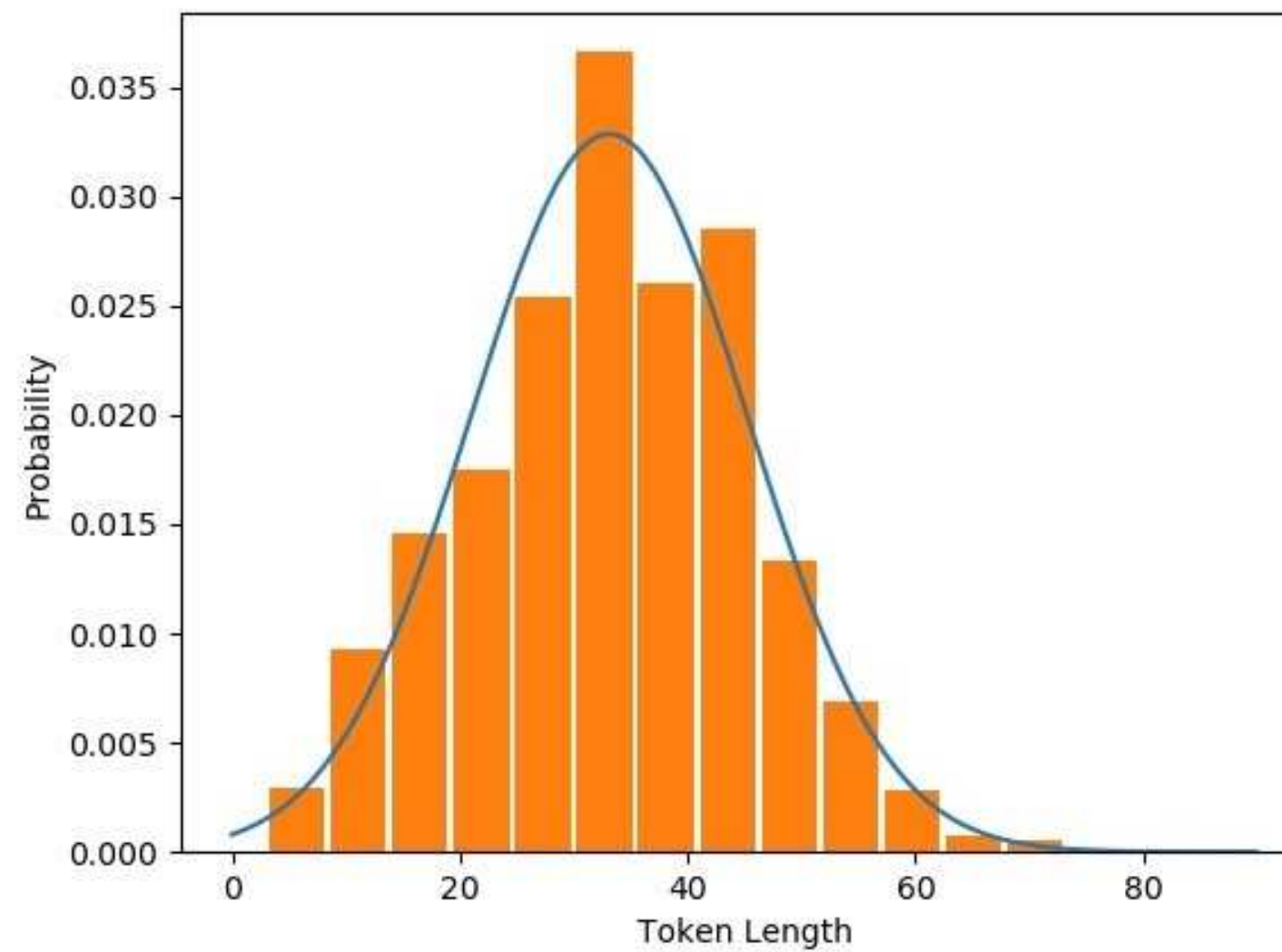


Figure 2: Distribution of Token Length





[Problem Definition](#)

[Related Work and Challenges](#)

[Dataset](#)

[Methodology](#)

[Input Format](#)

[Model Detils](#)

[Results](#)

Methodology



Input Format

- [Problem Definition](#)
- [Related Work and Challenges](#)
- [Dataset](#)
- [Methodology](#)
- [Input Format](#)**
- [Model Detils](#)
- [Results](#)

- 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, ...]



- Problem Definition
- Related Work and Challenges
- Dataset
- Methodology
- Input Format
- Model Detils
- Results

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



- [Problem Definition](#)
- [Related Work and Challenges](#)
- [Dataset](#)
- [Methodology](#)
- [Input Format](#)
- [Model Detils](#)**
- [Results](#)

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



- [Problem Definition](#)
- [Related Work and Challenges](#)
- [Dataset](#)
- [Methodology](#)
- [Results](#)

Results



- [Problem Definition](#)
- [Related Work and Challenges](#)
- [Dataset](#)
- [Methodology](#)
- [Results](#)

- $F_1 \text{ score} = \frac{2 * \textit{precision} * \textit{recall}}{\textit{precision} + \textit{recall}}$
- Rank: 21/887 (2.3 %)

Table 5: Results

Model	$F_1 \text{ score}$
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

