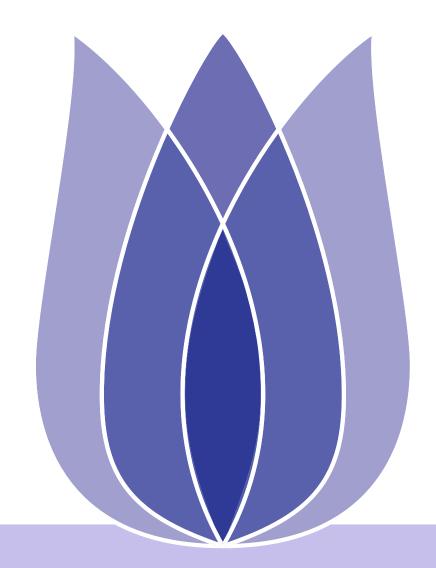
# **FLIP01 Final Assessment**

Cong Ma

QingDao Technological University

2021-01-07





### **Overview**

**Problem Definition** 

Data Visualization

Build The Model

Conclusion

**Problem Definition** 

**Problem Description** 

**Data Visualization** 

Data Visualization

**Build The Model** 

Model:MLP

Model:roBERTa

Conclusion





#### Problem Definition

**Problem Description** 

Data Visualization

**Build The Model** 

Conclusion

# **Problem Definition**





## **Problem Description**

**Problem Definition** 

Problem Description

Data Visualization

**Build The Model** 

Conclusion

With all of the tweets circulating every second it is hard to tell whether the sentiment behind a specific tweet will impact a company, or a person's, brand for being viral (positive), or devastate profit because it strikes a negative tone.

- What's the Sentiment of this tweet.
- What's the part of the tweet (word or phrase) that reflects the sentiment.

ID	text	selected_text	sentiment
cb774db0d1	Uh oh, I am sunburned	I am sunburned	negative
549e992a42	We saw that the baddie's the best	best	positive
f84b89a828	Sounds like me	Sounds like me	neutral





Problem Definition

#### Data Visualization

Data Visualization

**Build The Model** 

Conclusion

# **Data Visualization**





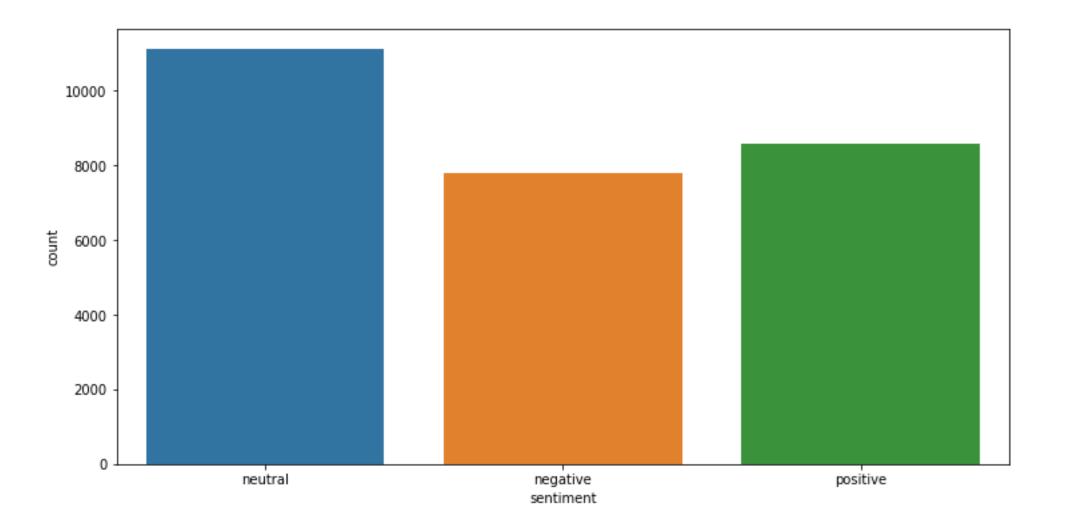
**Problem Definition** 

Data Visualization

#### Data Visualization

**Build The Model** 

- First, check the data. The training set contains 27482 data.
  - ◆ Take a look at the proportion of different types of text in the training set
  - It can be seen that the number of three kinds of data is relatively average. In addition, there are more neutral texts.







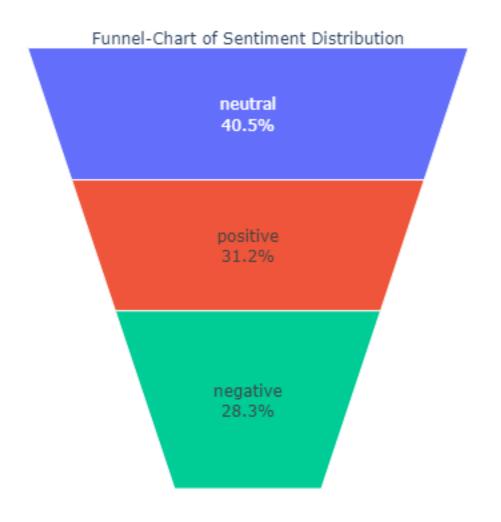
Problem Definition

Data Visualization

Data Visualization

Build The Model

- Let's take a look at the proportion of each category of text in another visualization method.
  - ◆ Most of them are neutral emotions, has 40.5 present. Positive are 31.2 persent. Negetive are 28.3 persent.







**Problem Definition** 

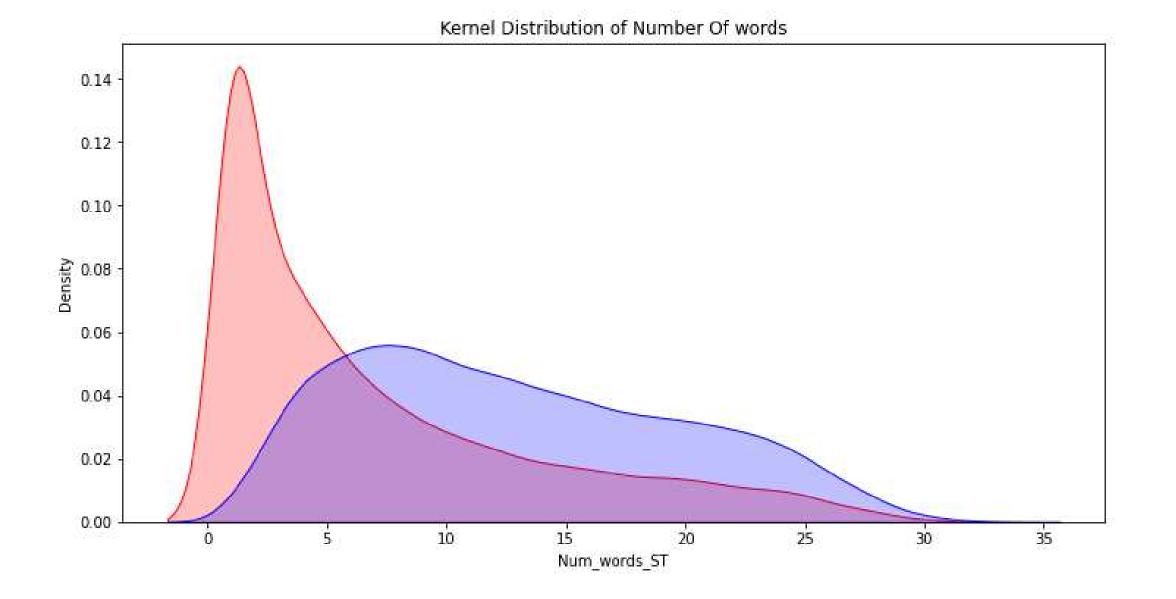
Data Visualization

Data Visualization

Build The Model

Conclusion

■ Count the distribution interval of the length of the given text and the selected text.







**Problem Definition** 

Data Visualization

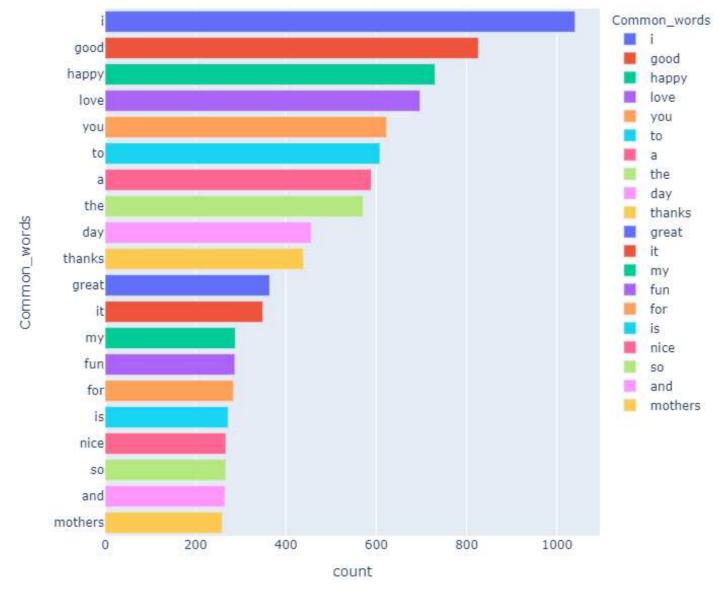
Data Visualization

Build The Model

Conclusion

■ Statistics of positive emotions were selected in the text of the highest frequency of the first few words.









**Problem Definition** 

Data Visualization

Data Visualization

**Build The Model** 

Conclusion

■ The statistical results will be generated word cloud to more intuitive look at the frequency of words.

#### WordCloud of Postive Tweets

```
friends Story
Ta NameLength

really
Wow morning
night Online
fun new
Ghost free
Yay goodbecame
break
Playing cooler Th Sfillin
followers
feelhehe feedingsinteresting
```



**Problem Definition** 

Data Visualization

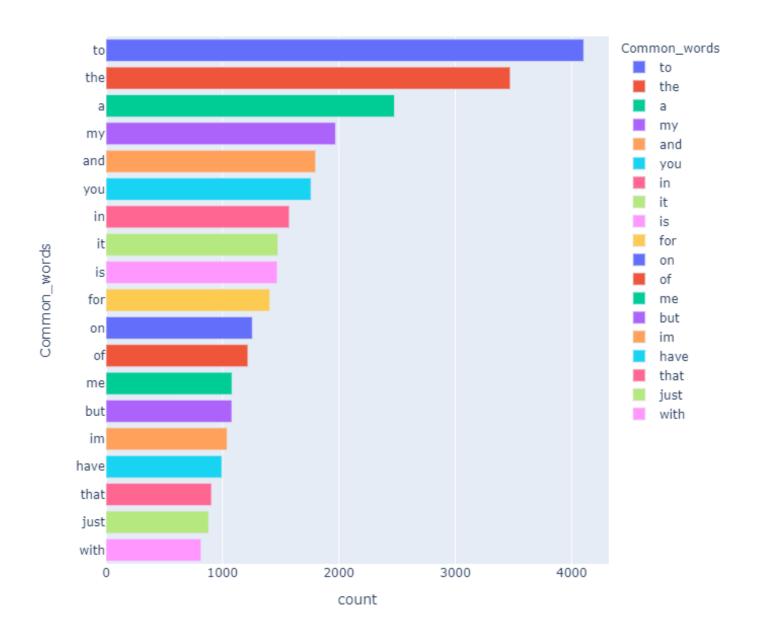
Data Visualization

Build The Model

Conclusion

■ Statistics of neutral emotions were selected in the text of the highest frequency of the first few words.









**Problem Definition** 

Data Visualization

Data Visualization

Build The Model

Conclusion

■ The statistical results will be generated word cloud to more intuitive look at the frequency of words.

#### WordCloud of Neutral Tweets

```
SOOOO
respond didnt
going
hopeful
grilled
till much
back olives
yesterday
back olives
yesterday
high BH spoke
smf high BH spoke
smf all text love
defying shameles days
reckon responded thring
mushroomslength gravity cheese
```



**Problem Definition** 

Data Visualization

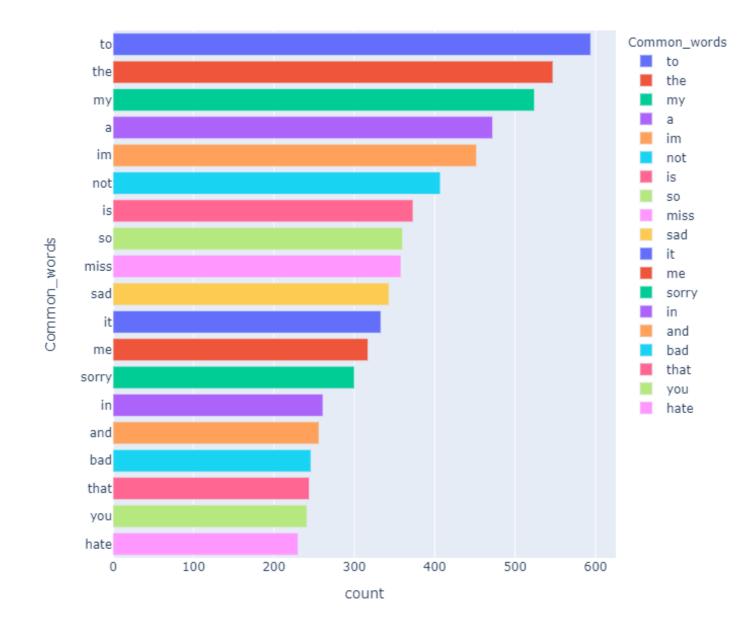
Data Visualization

Build The Model

Conclusion

■ Statistics of negative emotions were selected in the text of the highest frequency of the first few words.









Problem Definition

Data Visualization

Data Visualization

Build The Model

Conclusion

■ The statistical results will be generated word cloud to more intuitive look at the frequency of words.

# WordCloud of negative Tweets

```
Tall

Solvength

See

Id Will

Will

Sons boss

Louddn

Sons boss

Diego

DANGERously

Same

Bullying

Wanted

Interview

SAD

Sons boss

Diego

Shullying

SAD

San biect

San
```





Problem Definition

Data Visualization

Data Visualization

Build The Model

- It can be seen that our previous statistical text contains some words without emotional tendency.
- After we delete these words, we count the frequency of each word.





**Problem Definition** 

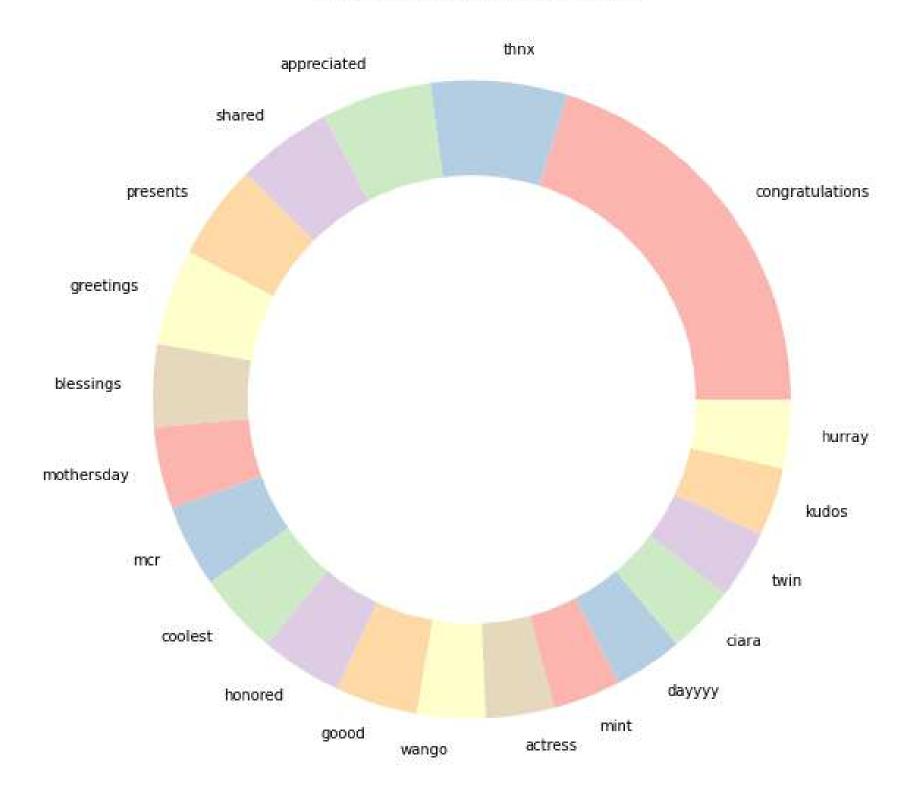
Data Visualization

#### Data Visualization

Build The Model

Conclusion

#### DoNut Plot Of Unique Positive Words







Problem Definition

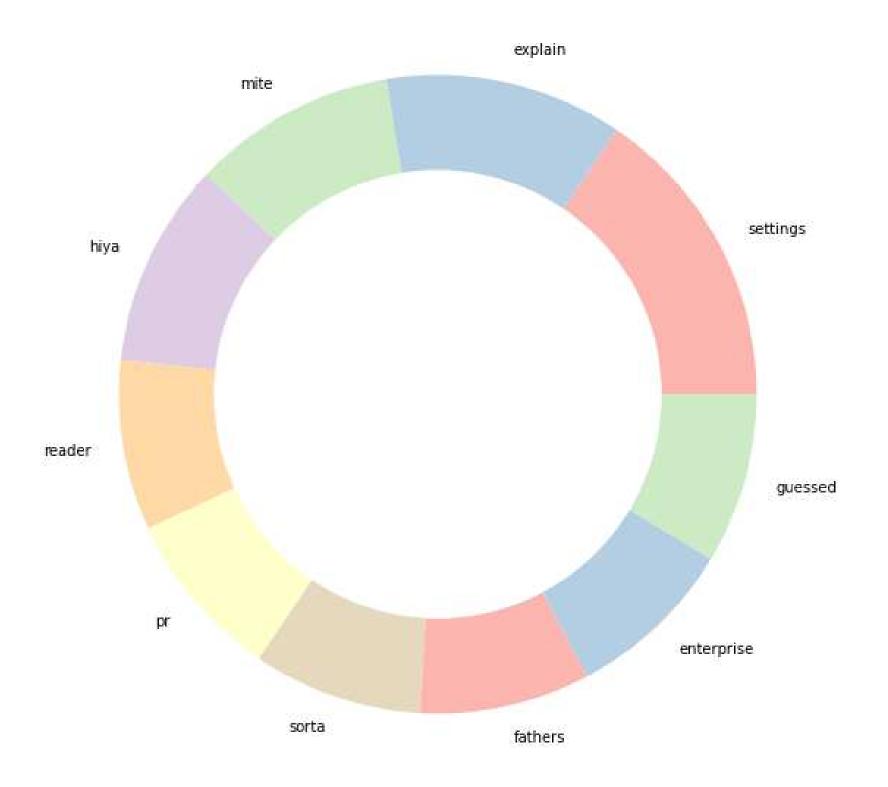
Data Visualization

Data Visualization

Build The Model

Conclusion

#### DoNut Plot Of Unique Neutral Words





Problem Definition

Data Visualization

Data Visualization

Build The Model

Conclusion

#### DoNut Plot Of Unique Negative Words







Problem Definition

Data Visualization

#### Build The Model

Model:MLP

Model:roBERTa

Conclusion

# **Build The Model**





### **Model:MLP**

**Problem Definition** 

Data Visualization

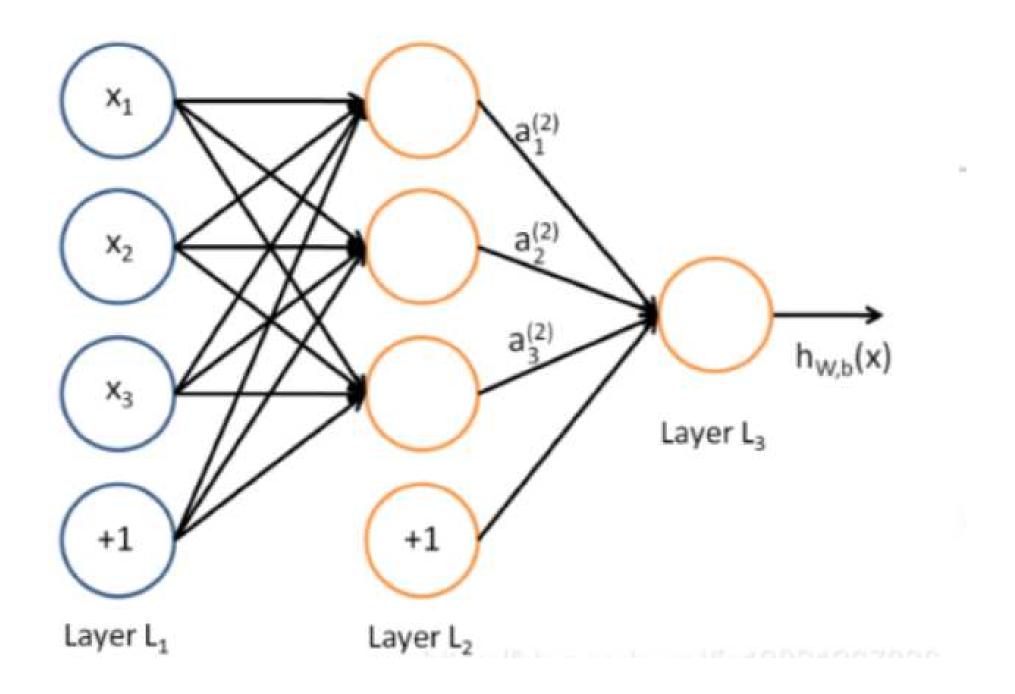
Build The Model

#### Model:MLP

Model:roBERTa

Conclusion

■ First, the MLP model with simple structure is used.







#### **Model:MLP**

**Problem Definition** 

Data Visualization

Build The Model

#### Model·MLP

Model:roBERTa

- Based on our previous data visualization. We set the MAX\_LEN = 48.
- The learning rate is 0.8.
- Activation function is "Relu".
- The output is one dimension and the convolution kernel size is 1 \* 1.
- The optimizer is "SGD".
- The loss function is "categorical\_crossentropy".
- $\blacksquare$  Epochs = 10.
  - ◆ Finally, the loss rate of the trained model is 0.5393.





#### Model:roBERTa

**Problem Definition** 

Data Visualization

Build The Model

Model:MLP

Model:roBERTa

Conclusion

In order to obtain higher accuracy, I choose the widely used model named roBERTa.

- Roberta: a robust method to optimize the pre training of Bert.
- Roberta is an improved algorithm of bert.
  - With bigger batch and more data, let the model train longer.
  - ◆ Removed the NSP (next sense prediction) task.
  - Train on a longer sequence.
  - Mask mechanism for dynamically modifying training data.





#### Model:roBERTa

**Problem Definition** 

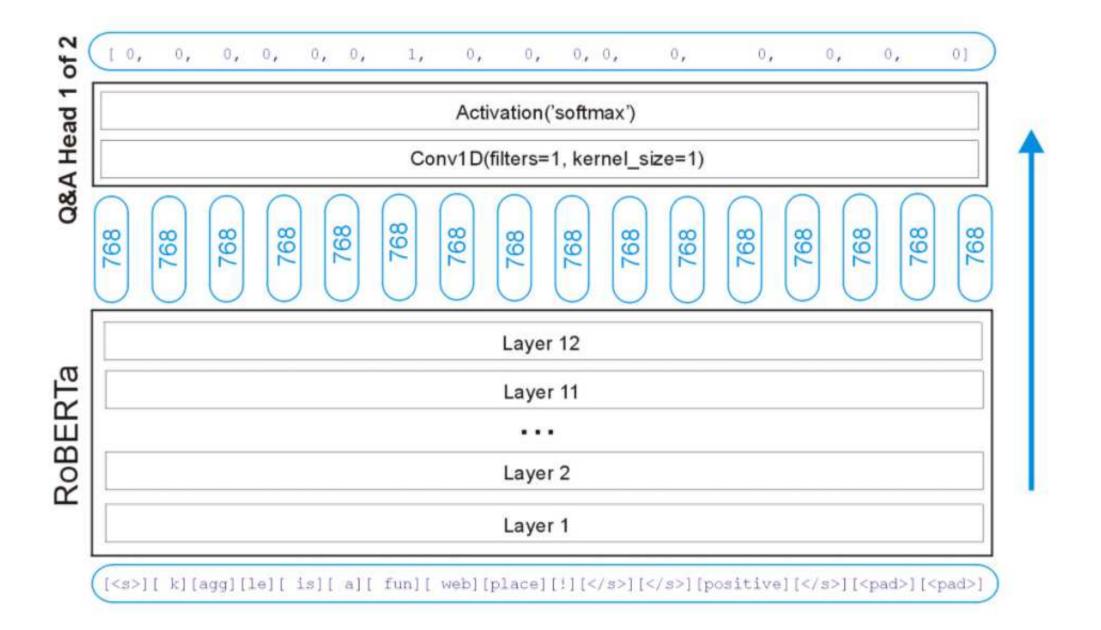
Data Visualization

Build The Model

Model:MLP

Model:roBERTa

- Activation function is "softmax".
- The output is one dimension and the convolution kernel size is 1 \* 1.





#### Model:roBERTa

**Problem Definition** 

Data Visualization

Build The Model

Model:MLP

Model:roBERTa

- Based on our previous data visualization. We set the MAX\_LEN = 48.
- The learning rate is 0.9.
- The optimizer is "Adam".
- The loss function is "categorical\_crossentropy".
- Using k-fold cross validation, it is divided into five parts. Train five times.
- Finally, the loss rate of the trained model is 0.7.





Problem Definition

Data Visualization

Build The Model

Conclusion

Conclusion





### Conclusion

Problem Definition

Data Visualization

Build The Model

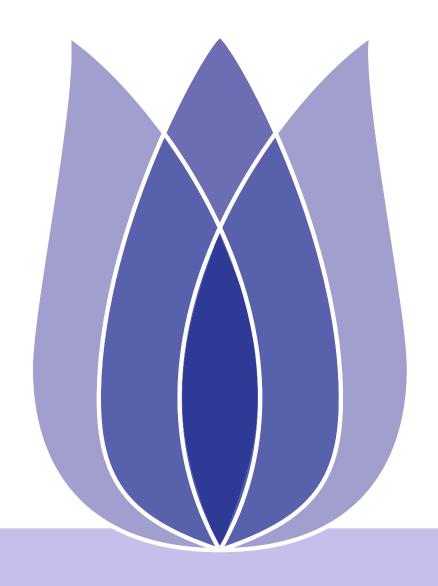
Conclusion

- Formalize the problem of *Group Outlying Aspects Mining* by extending outlying aspects mining;
- Propose a novel method GOAM algorithm to solve the *Group Outlying Aspects Mining* problem;
- Utilize the pruning strategies to reduce time complexity.





# **Contact Information**



Made By Cong Ma QingDao Technological University