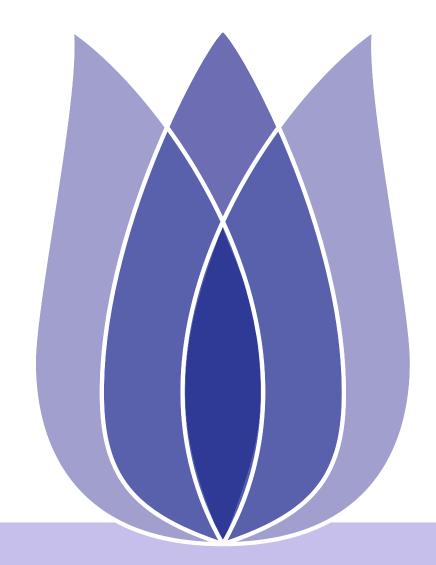
FLIP01 Final Assessment

Cong Ma

QingDao Technological University

2021-01-15





Overview

Problem Definition

Data Visualization

Data Process

Build The Model

Conclusion

Problem Definition

Problem Description

Data Visualization

Data Visualization

Data Process

Build The Model

Model:CRF+LSTM

Model:roBERTa

Conclusion

Contact Information





Problem Definition

Problem Description

Data Visualization

Data Process

Build The Model

Conclusion

Problem Definition





Problem Description

Problem Definition

Problem Description

Data Visualization

Data Process

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

Data Process

Build The Model

Conclusion

Data Visualization





Problem Definition

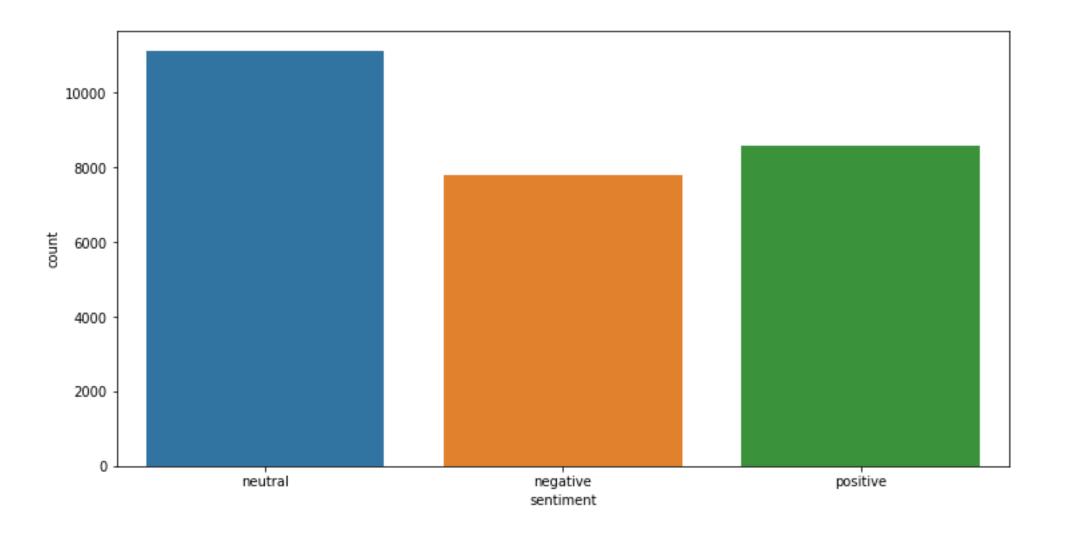
Data Visualization

Data Visualization

Data Process

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

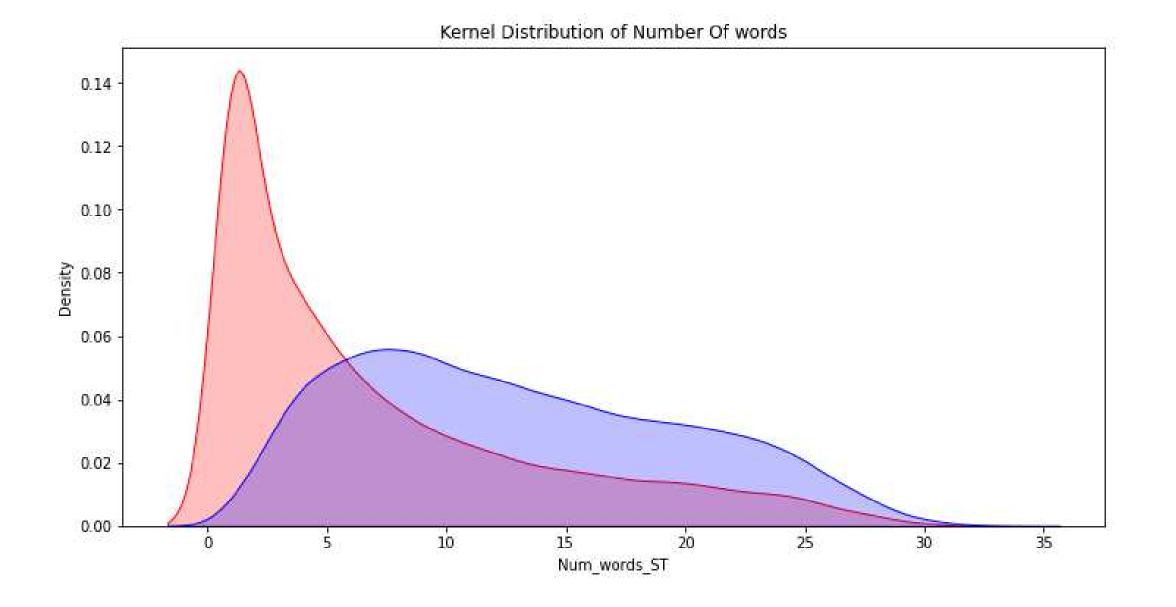
Data Visualization

Data Process

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

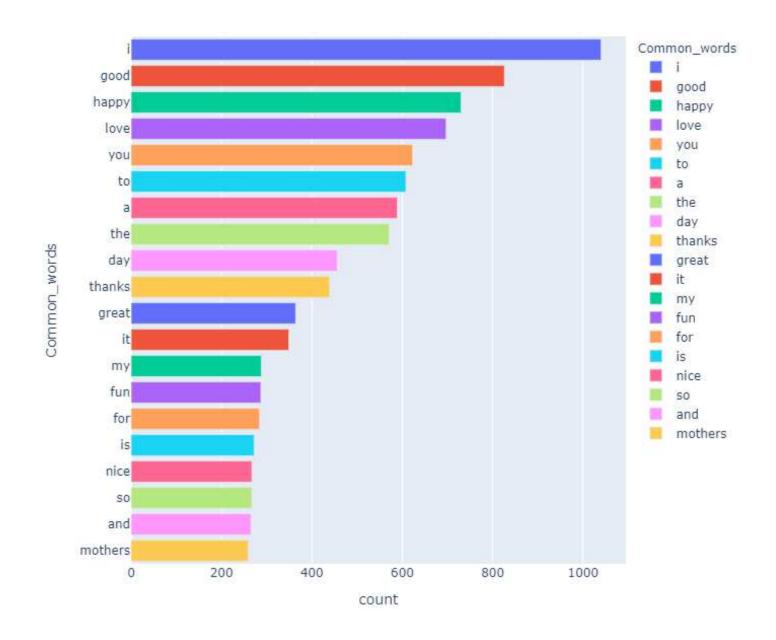
Data Process

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

Data Process

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 to the fillin
feelhehe feedingsinteresting
```





Problem Definition

Data Visualization

Data Visualization

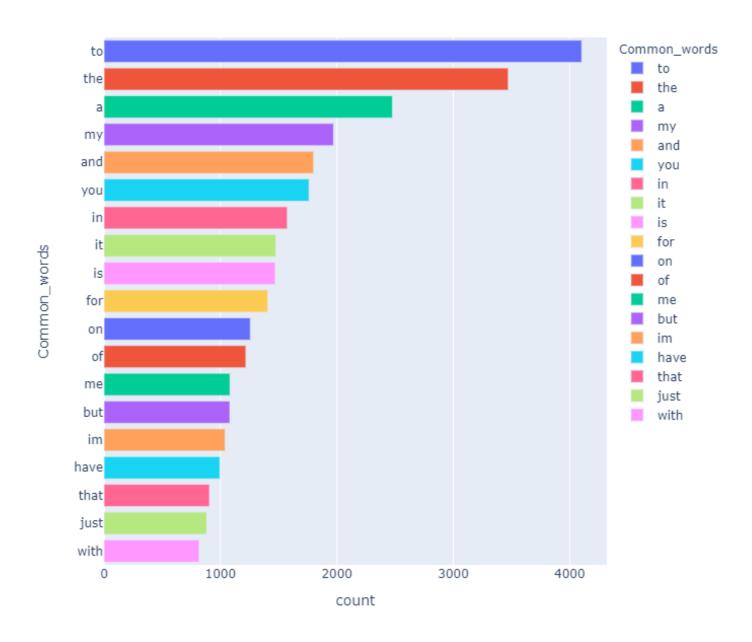
Data Process

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

Data Process

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
object d dtype
back olives
yesterday
back olives
yesterday
high BH spoke
smf high BH spoke
smf all text love
defying shameles days
reckon responded firting
mushroomsLength gravity cheese
```





Problem Definition

Data Visualization

Data Visualization

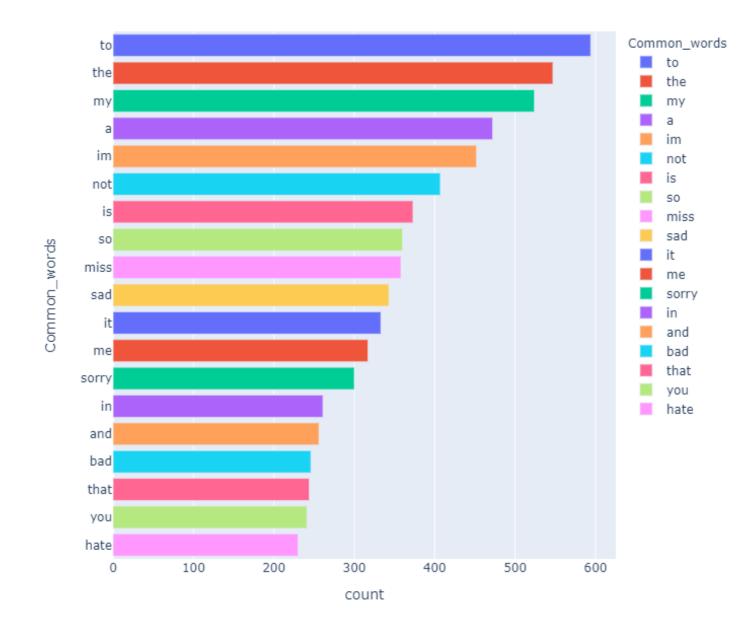
Data Process

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

Data Process

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

Wanted

Sons boss

Diego

Sons boss

Sons boss

Diego

Sons boss

Diego
```





Problem Definition

Data Visualization

Data Visualization

Data Process

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

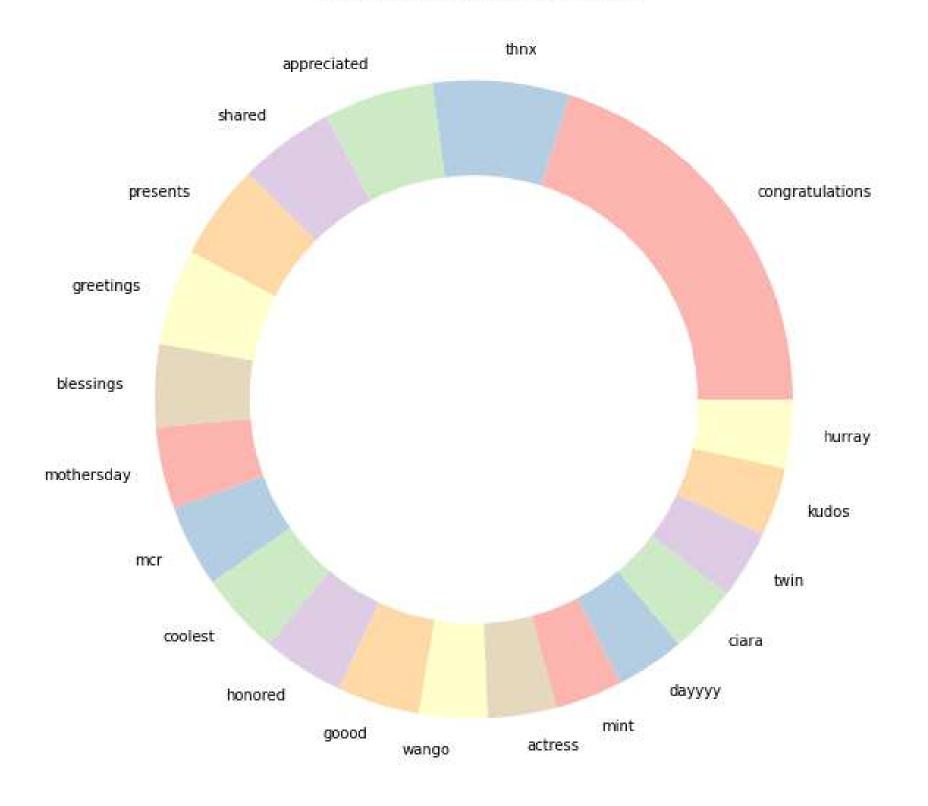
Data Visualization

Data Process

Build The Model

Conclusion

DoNut Plot Of Unique Positive Words







Problem Definition

Data Visualization

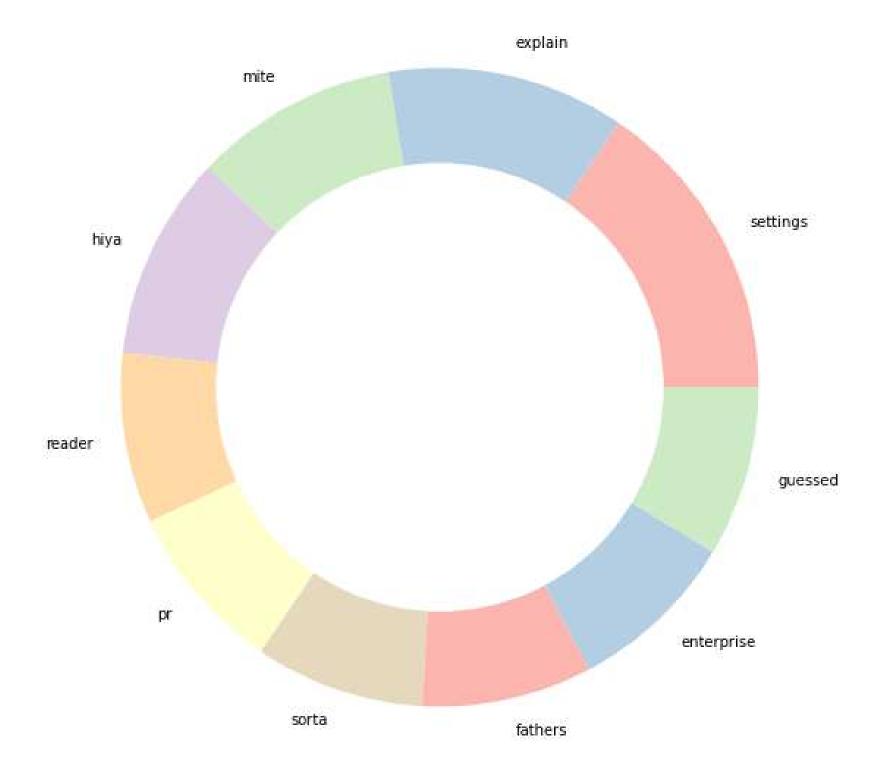
Data Visualization

Data Process

Build The Model

Conclusion

DoNut Plot Of Unique Neutral Words







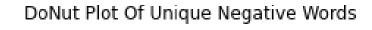
Problem Definition

Data Visualization

Data Visualization

Data Process

Build The Model









Problem Definition

Data Visualization

Data Process

Build The Model

Conclusion

Data Process





Data Process

Problem Definition

Data Visualization

Data Process

Build The Model

- Observe the given training set, and the extracted words are part of the original sentence.
- The data processing part will only delete the blank data in the given training set.





Problem Definition

Data Visualization

Data Process

Build The Model

Model:CRF+LSTM

Model:roBERTa

Conclusion

Build The Model





Model:CRF+LSTM

Problem Definition

Data Visualization

Data Process

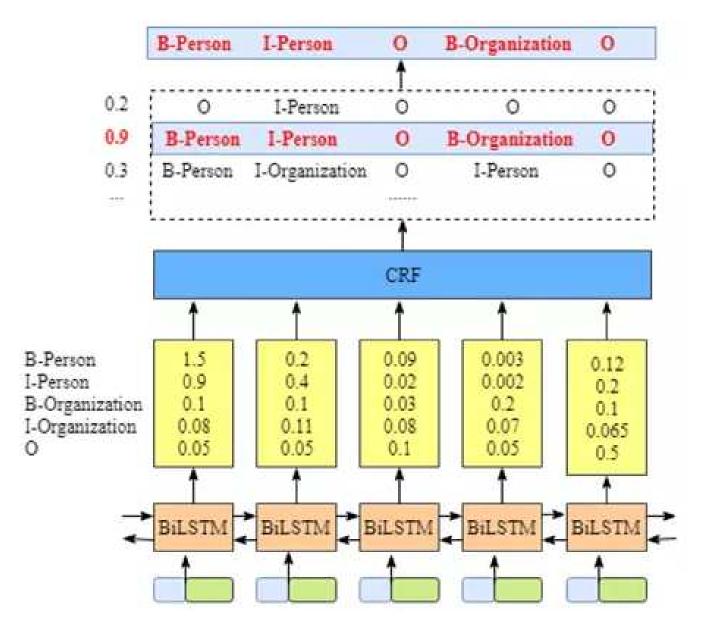
Build The Model

Model:CRF+LSTM

Model:roBERTa

Conclusion

■ First, the CRF+LSTM model is used.





Model:CRF+LSTM

Problem Definition

Data Visualization

Data Process

Build The Model

Model:CRF+LSTM

Model:roBERTa

Conclusion

- 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.

FLIP01 Final Assessment

- The optimizer is "SGD".
- The loss function is "categorical_crossentropy".
- \blacksquare Epochs = 10.





Model:CRF+LSTM

Problem Definition

Data Visualization

Data Process

Build The Model

Model:CRF+LSTM

Model:roBERTa

Conclusion

	textID	text	selected_text	sentiment
0	d93afa85cf	Car not happy, big big dent in boot! Hoping t	Car not happy, big big dent in boot! Hoping th	neutral
1	fab6b7d16c	im an avid fan of **** magazine and i love you	avid fan of	positive
2	2e7082d1c8	MAYDAY?!	MAYDAY?!	neutral
3	684081e4e7	RATT ROCKED NASHVILLE TONITEONE THING SUCKED	RATT ROCKED NASHVILLE TONITEONE THING SUCKED	neutral
4	c77717b103	I love to! But I'm only available from 5pm	I love to!	positive

■ Finally, the loss rate of the trained model is 0.5393.



Problem Definition

Data Visualization

Data Process

Build The Model

Model:CRF+LSTM

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.





Problem Definition

Data Visualization

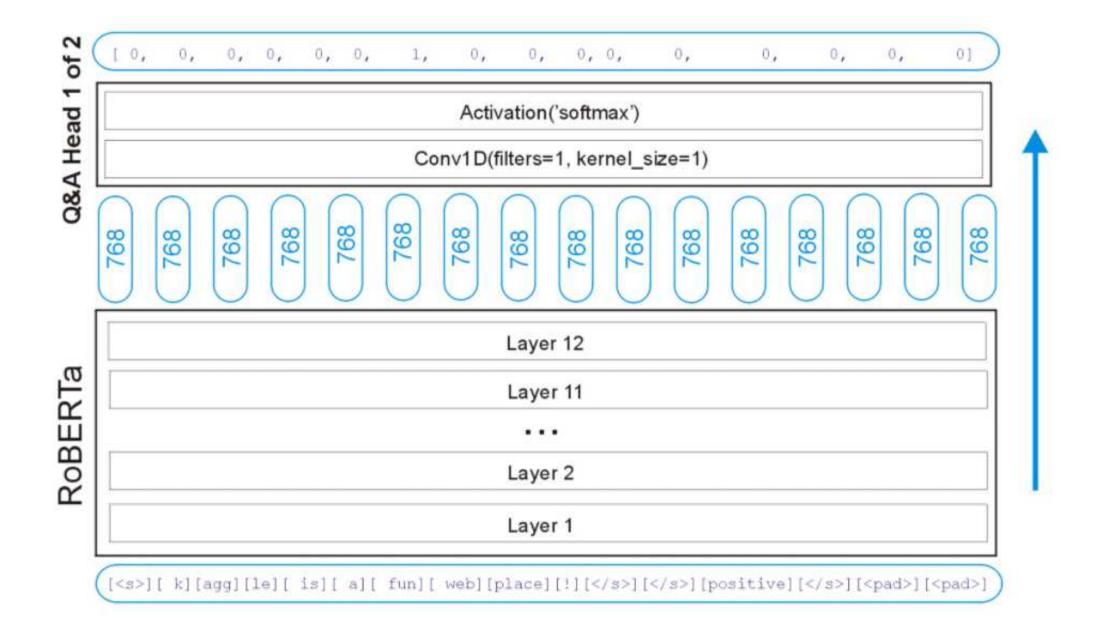
Data Process

Build The Model

Model:CRF+LSTM

Model:roBERTa

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





Problem Definition

Data Visualization

Data Process

Build The Model

Model:CRF+LSTM

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.





Problem Definition

Data Visualization

Data Process

Build The Model

Model:CRF+LSTM

Model:roBERTa

Conclusion

	textID	text	selected_text	sentiment
0	eae9c20c8d	#followfriday thank you so much. I`m so be	day thank you so mu	positive
1	404e86f215	(cont) when told him that I love beans on toa	I love beans on toast. SO CUTE!	positive
2	81a83e8d9a	why am i up so early	why am i up so early	negative
3	c0d5b45663	Joined you on facebook!	Joined you on facebook!	neutral
4	c3c1abb017	trying to find some friends and not having any	not having any luck	negative

■ Finally, the loss rate of the trained model is 0.7072782233322157.



Problem Definition

Data Visualization

Data Process

Build The Model

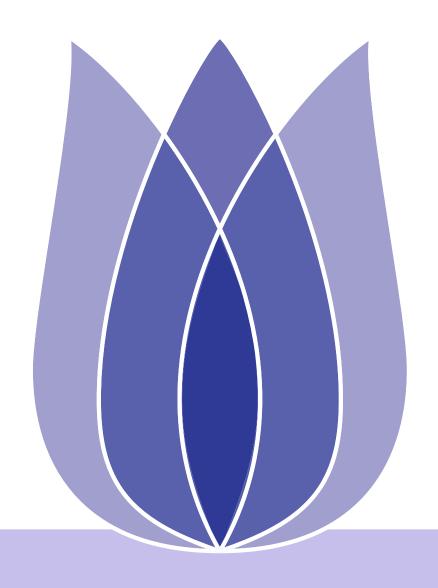
Conclusion

Contact Information





Contact Information



Made By Cong Ma QingDao Technological University