Tweet Sentiment Extraction

Jia Huang

Xi'an Shiyou University

Chinese Academy of Sciences

January 16, 2021

Overview

Project Overview

Project Background And Purpose

Defn

- Background

  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. Capturing sentiment in language is important in these times where decisions and reactions are created and updated in seconds. But, which words actually lead to the sentiment description?

- Purpose

  In this competition we've extracted support phrases from Figure Eight's Data for Everyone platform. The dataset is titled Sentiment Analysis: Emotion in Text tweets with existing sentiment labels, used here under creative commons attribution 4.0. international licence. Your objective in this competition is to construct a model that can do the same - look at the labeled sentiment for a given tweet and figure out what word or phrase best supports it.

Data Pre-Processing

Date Processing

Two data sets were given during the match: train.csv and test.csv.Train.csv data were used to construct the model and to predict test.csv data. Now let's take a look at the training data provided by this competition, do some exploratory data analysis work, and learn more about the content of this training set.
The data sets train.csv and test.csv are respectively 27408 and 3534 pieces of data.The data is described in terms of four attributes.

	textID	text	selected_text	sentiment
0	cb774db0d1	I`d have responded, if I were going	I`d have responded, if I were going	neutral
1	549e992a42	Sooo SAD I will miss you here in San Diego!!!	Sooo SAD	negative
2	088c60f138	my boss is bullying me...	bullying me	negative
3	9642c003ef	what interview! leave me alone	leave me alone	negative
4	358bd9e861	Sons of ****, why couldn`t they put them on t...	Sons of ****,	negative

Figure 1

Feature Item

As you can see from the table above, the four feature items in the training set are textID, Text, and selected_text sentiment.

- textID
  - Write the ID of the comment.
- Text
  - The specific content of the comment.
- Selected_text
  - Are the keywords that we have chosen to judge the emotional state of the comment.
- Sentiment
  - The emotional polarity of the sentence.

Features Item

Out[21]:

	sentiment	text
1	neutral	11117
2	positive	8582
0	negative	7781

There are three types of sentence emotional polarity.Neurtal indicates that the sentence is emotionally neutral, positive means the comment is positive, and negative means the comment is negative.

Constructing Dataset Generator

Similarity

Look at the Jaccard similarity between Text and selected_text.

	textID	text	selected_text	sentiment	jaccard_score
0	cb774db0d1	I`d have responded, if I were going	I`d have responded, if I were going	neutral	1.000000
1	549e992a42	Sooo SAD I will miss you here in San Diego!!!	Sooo SAD	negative	0.200000
2	088c60f138	my boss is bullying me...	bullying me	negative	0.166667
3	9642c003ef	what interview! leave me alone	leave me alone	negative	0.600000
4	358bd9e861	Sons of ****, why couldn`t they put them on t...	Sons of ****,	negative	0.214286

The kernel distribution graph of word length

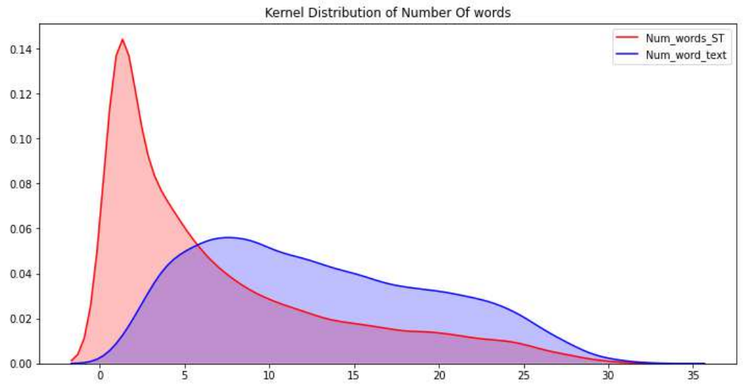

Figure 2

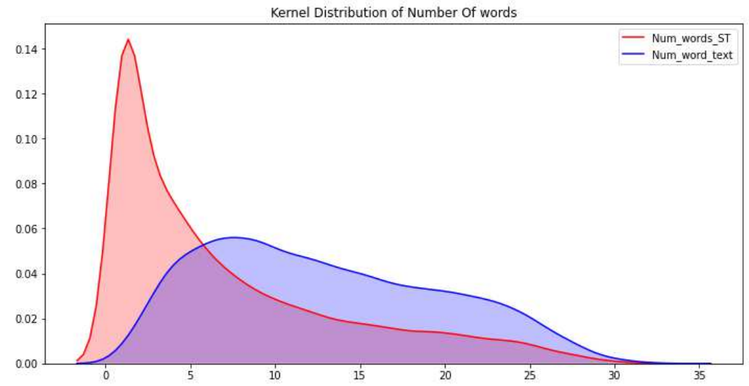

Figure 3

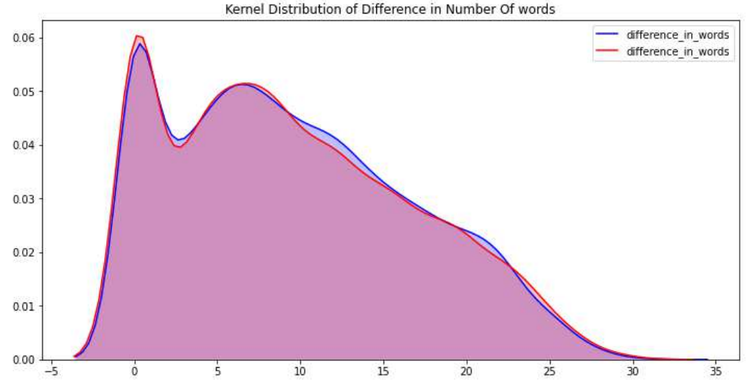

Figure 4

The kernel distribution graph of word length

As can be seen from the above figure, the Jaccard similarity of positive or negative text and selected_text has two sharp kurtosis around 1.0 or 0.1.The word length difference also has two kurtosis, where the difference of 0 is a sharp kurtosis.That means that a large percentage of positive or negative text is the same as selected_text.

The word cloud

Figure 5

Figure 6

Figure 7

Model Construction and Training

- Using Facebook's Roberta model, we first need to build Tokenizer to convert the training text into Roberta's token.
  - Build Tokenizer to convert the text of the training set and test set to Token.

To build the Roberta model,first load the pre-trained model,the input of the model is the above converted text token,the output is the vecor of BatchxMAX_LENx768, add two Q&A heads,one of which is responsible for predicting the beginning position of the answer and the other is responsible for predicting the end position of the answer.By making a full connection layer of 768x1 to the output vector to get the Head, the output becomes the vector of BatchxMAX_LENx1,then rehaspe is the vector of BatchxMAX_LEN,and then Softmax.

	textID	text	sentiment	selected_text
0	f87dea47db	Last session of the day http://twitpic.com/67ezh	neutral	last session of the day http://twitpic.com/67ezh
1	96d74cb729	Shanghai is also really exciting (precisely -...	positive	exciting
2	eee518ae67	Recession hit Veronique Branquinho, she has to...	negative	such a shame!
3	01082688c6	happy bday!	positive	happy bday!
4	33987a8ee5	http://twitpic.com/4w75p - I like it!!	positive	i like it!!
5	726e501993	that`s great!! weee!! visitors!	positive	that`s great!!
6	261932614e	I THINK EVERYONE HATES ME ON HERE lol	negative	hates
7	afa11da83f	sooooo wish i could, but im in school and my...	negative	blocked
8	e64208b4ef	and within a short time of the last clue all ...	neutral	and within a short time of the last clue all o...