Tweet Sentiment Extraction

Jia Huang

Xi'an Shiyou University

Chinese Academy of Sciences

January 15, 2021

Overview

Project Overview

- Project Background And Purpose

Data Pre-Processing

- Date Processing
- Feature Item
- Features Item

Constructing Dataset Generator

- Constructing Dataset Generator

Modeling

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.

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

Constructing Dataset Generator

Constructing Dataset Generator

Build a dataset generator to analyze each sentence in a comment.

- Input_ids: The id number of each word in the dictionary .
- Attention_mask: Which words can be used in the sentence.
- Input_type_ids: Distinguish between the preceding sentence and the following sentence.

- Target_start: The beginning position in selected_text.
- Offsets: Mark the offsets of each word after the participle.
- Target_end: End position in selected_text.
- Tweet: The original sentence.
- Selected_text: Emotional sentences.
- Sentiment: The emotional polarity of this sentence.

Modeling

Build tokenizer to generate the dictionary

Because text needs to be processed before natural language processing can take place. So we need to turn text into a computer-friendly language. The Tokenizer class is used to count the words in the text and generate a document dictionary to support vector representations of the generated text based on the lexicon order.
The 2,323 copies are duplicates and we need to delete them. We will also evaluate the position of the data points using the coordinates.