Tweet Sentiment Extraction

Jia Huang

[1] Xi'an Shiyou University, China

Introduction

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? In this competition you will need to pick out the part of the tweet (word or phrase) that reflects the sentiment.

The Dataset The data set includes three files: the training set, the data set, and the sample set provided.

Data Format Each row contains text, a tweet, and a sentiment tag.In the training set, you extract a word or phrase from the tweet (selected_text) that contains the emotion provided.

Predicted Results You are trying to predict words or phrases from the tweets to illustrate the emotions offered.A word or phrase should include all characters in the range (that is, including commas, Spaces, etc.).

The Dataset

There are 27481 pieces of data in the training set, among which one piece of data is invalid and shall be deleted.

- *TextID - The unique ID of each text segment.*
- *Text - Tweets.*
- *Sentiment - General sentiment of a tweet.*
- *Selected_text -(only trains) text that supports the sentiment of the tweet.*

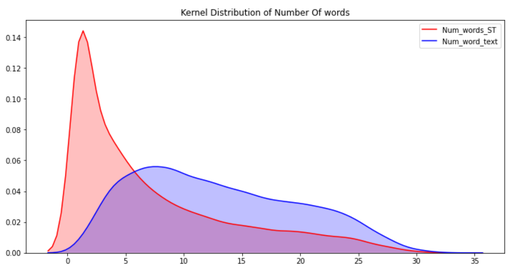
Exploratory Data Analysis

We can see that across the entire dataset, there are 111,117 tweets for neutral emotions, 8,582 tweets for positive emotions, and 7,781 tweets for negative emotions.

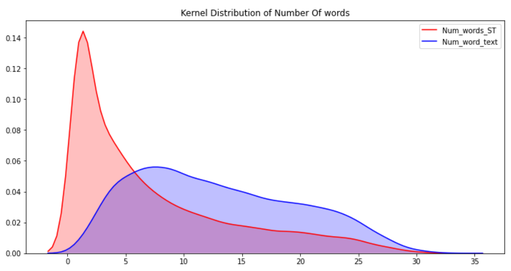
Nature	11117
positive	8582
negative	7781

Jaccard similarity between Text and selected_text. As can be seen from the above output results,such as I'd have responded, if I were goingThis kind of words has the highest similarity with his emotional polarity, with a similarity value of 1, followed by words like leave me alone, with a negative emotional polarity, with a similarity value of 0.6.

Nuclear distribution graph.

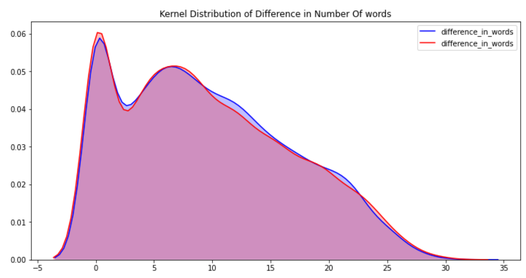


Flag.1



Flag.2

Exploratory Data Analysis



Flag.3

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.

It's also possible to use the word cloud to show how often words appear in different categories of Twitter.

The three word clouds show the most frequently used words in tweets for each of the three E-polarities. For example, we can clearly see that we usually use words like happy, good and so on to express our happy mood, and these words are also obvious in our output word cloud.

Missing figure Testing figcolor

Model Construction and Training.

Model for training in the large-scale corpus, it concluded that the word vector can be used in different field of NLP, but all of these training requires a lot of GPU, for the average person is unable to model parameters, used after fine-tuning in our NLP tasks.Here I used Huggingface's open source Transformers library, which provides a lot of directly called and built using PyTorch or TensorFlow.

- Divide the training set into 5 copies, and take 4 copies of training and 1 copy for verification.
- Three epochs are trained each time, and parameters of the epoch with the lowest Loss in verification set are taken.We get a new model at the end of each training session, so we end up with 5 models.
- When making predictions, the prediction results of these five models will be averaged.
- The predicted values of the five trained models were averaged.

Missing figure Testing figcolor

New Orleans Pelicans on FT%

Missing figure Testing figcolor

New Orleans Pelicans on FTA

Conclusion

Through this project, we further understand the related contents of big dat
ysis, and also have a preliminary understanding of natural lan
I believe this project experience will be of great help i

Acknowledgement

- International Cooperation Project (Y7Z0511101) of IIE, Chinese Academy of Sciences