TWEET SENTIMENT EXTRACTION

JIA HUANG

Abstract. 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 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.

Contents

Date: 2020-10-15.
1991 *Mathematics Subject Classification.* Artificial Intelligence.
Key words and phrases. Emotional extract.

1. 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.

Help build your skills in this important area with this broad dataset of tweets. Work on your technique to grab a top spot in this competition. What words in tweets support a positive, negative, or neutral sentiment? How can you help make that determination using machine learning tools?

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.

2. Data Exploratory Analysis

Now let's take a look at the training data provided by this competition and do some exploratory data analysis to get a detailed understanding of the training set.

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

Flag.1

(1) Text and selected_text of one record in training set are null, so delete this data.
(2) Check the distribution of the sentiment category in the training set data.

Out[21]:

	sentiment	text
1	neutral	11117
2	positive	8582
0	negative	7781

Flag.2

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.

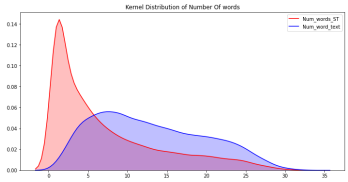
(3)Look at the Jaccard similarity between Text and selected_text.

	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

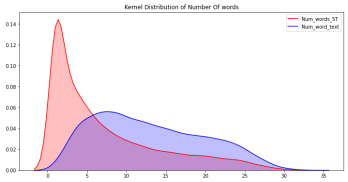
Flag.3

As can be seen from the above output results,such as I'd have responded, if I were going This 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.

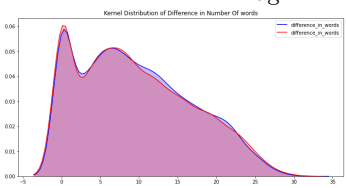
(4)The kernel distribution graph of word length.



Flag.4



Flag.5



Flag.6

The above three pictures are text and selected_text, respectively. The kernel distribution picture of Jaccard similarity for Sentiment classified as positive or negative word length difference.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.

(5) The word cloud shows how often words appear in different categories of Twitter.

Flag.8

Flag.9

Flag.10

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.

3. Model Construction and Training

These models 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 afford the cost of theraining from

the start, so we can use has trained model parameters, used after fine-tuning in our NLP tasks.Here I used Huggingface's open source Transformers library, which provides a lot of pre-trained models with a very good and easy-to-use interface that can be directly called and built using PyTorch or TensorFlow.

- Construct tokenizer, and transform the text of training set and test set into token.
  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.
- 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.

	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	sooooooo 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...

Flag.3

4. Conclusions

Through this project, we further understand the related contents of big data analysis, and also have a preliminary understanding of natural language processing. I believe this project experience will be of great help in the following study.

List of Todos

(A. 1) School of Computer Science,, Xi'an Shiyou University, Shaanxi 710065, China
Email address, A. 1: xxx@tulip.academy