## Fine Tuning and pertaining with BERT Data

This section describes from where we got the data, and it discusses the process of discovering and preparing the data.

### Getting the Data

The type of data needed to train our model is sentences or paragraphs. Specifically the input is a set of labels and sentences/paragraphs. The target is to predict the sentimental tendency of the input text. The data used to train and test the module was generated manually from online reviews. It is spread across two csv files one for training and one for testing. Table 2 shows the data used for the model.

Table 2: Data used for the model

|  |  |  |
| --- | --- | --- |
| Model | Data used for training | Data used for testing |
| Our model | Gathered from online reviews | Gathered from online reviews |

### Legal Obligation and Authorization

The data is open source feel free to download it ###.

### Exploring the Data

In this chapter, we will go in depth to explore the dataset. Please note that we will only explore the training data, the test data will not be explored to prevent over fitting. The data is arranging in a data format similar to the one seen in *Figure 22*.

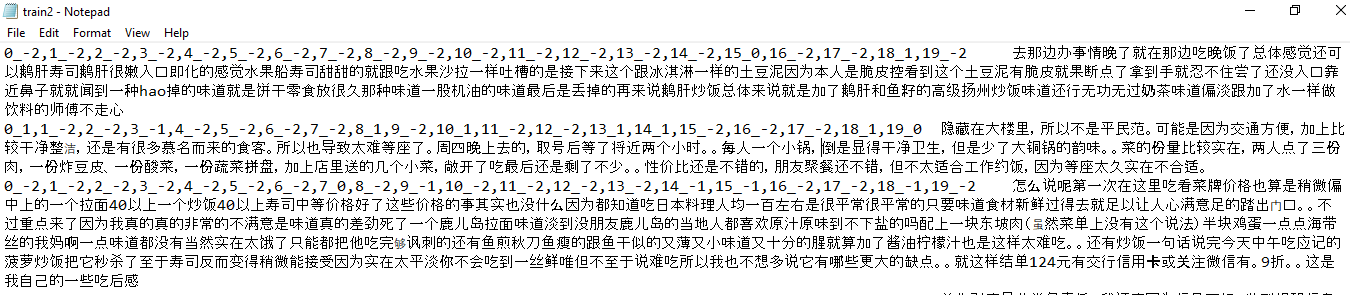
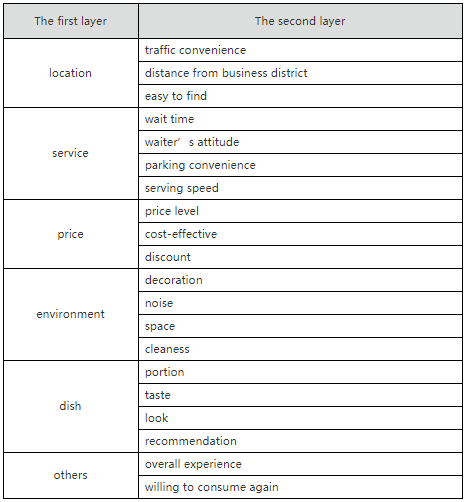


Figure 22: sample data

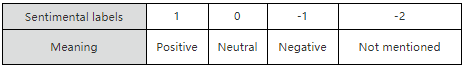
The dataset is built using a two-layer labeling system:  the first layer is the coarse-grained evaluation object, such as “service” and “location”; the second layer is the fine-grained emotion object, such as “waiter’s attitude” and “wait time” in “service” category. The specific description is shown in the Table 3.

Table 3: Description of how the data was built



There are four sentimental types for every fine-grained element: Positive, Neutral, Negative and Not mentioned, which are labelled as 1, 0, -1 and-2. The meaning of these four labels are listed below.

Table 4: sentimental labels

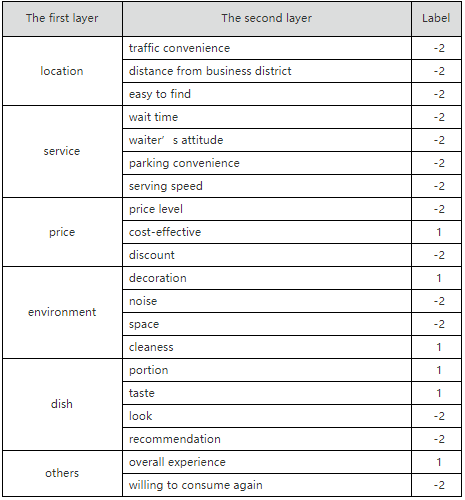


For example:

“味道不错的面馆，性价比也相当之高，分量很足～女生吃小份，胃口小的，可能吃不完呢。环境在面馆来说算是好的，至少看上去堂子很亮，也比较干净，一般苍蝇馆子还是比不上这个卫生状况的。中午饭点的时候，人很多，人行道上也是要坐满的，隔壁的冒菜馆子，据说是一家，有时候也会开放出来坐吃面的人。“

The previous paragraph has the following sentimental labels.

Table 5: example



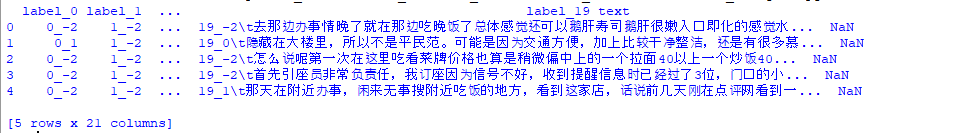
Now let’s take a look at the top five rows of the dataset see *Figure 23*. As seen in the figure each row represents one sentence or sentences as a paragraph. There are 20 labels/targets (in the figure label\_0 , label\_1 and label\_19 are shown): label\_0, label\_1,label\_2, label\_3,label\_4, label\_5,label\_6,label\_7,label\_8,label\_9,label\_10,label\_11,label\_12,label\_13,label\_14,label\_15,label\_16,label\_17,label\_18,label\_19. 

Figure 23: Top five rows in the dataset

Looking at *Figure 24* you can get a quick description of the data. Particularly the total number of rows, and each labels type and number of non-null values.

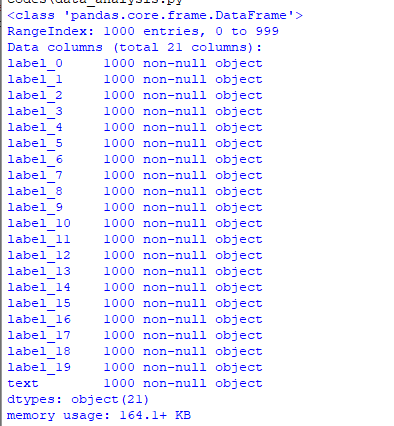


Figure 24: Data info

There are 1000 sentences or paragraphs in the dataset, no row is missing any value. All the attribute types are objects and since the data was loaded from a csv file then it is a text attribute.

There is no numerical attributes.

The data in the csv file only contains two columns as follows:

1. The first column has the target , one or multi label
2. The second column has the input strings

Please note that both the training data and the test data are treated the same way.