SMAI PROJECT

ASPECT BASED SENTIMENT ANALYSIS

TEAM MEMBERS:

REVANTH TALLURI: 2020102068

SRIRAM PRASAD: 2020102033

ROHIT REDDY: 2020102035

JOSH UJWAL: 2020102018

AIM

In this project we will do aspect-based sentiment analysis by modelling the interdependencies of sentences in a review with a hierarchical bidirectional LSTM.

DATASET

We obtained the dataset online which is based on the reviews of a restaurant. The main attributes present in the dataset are:

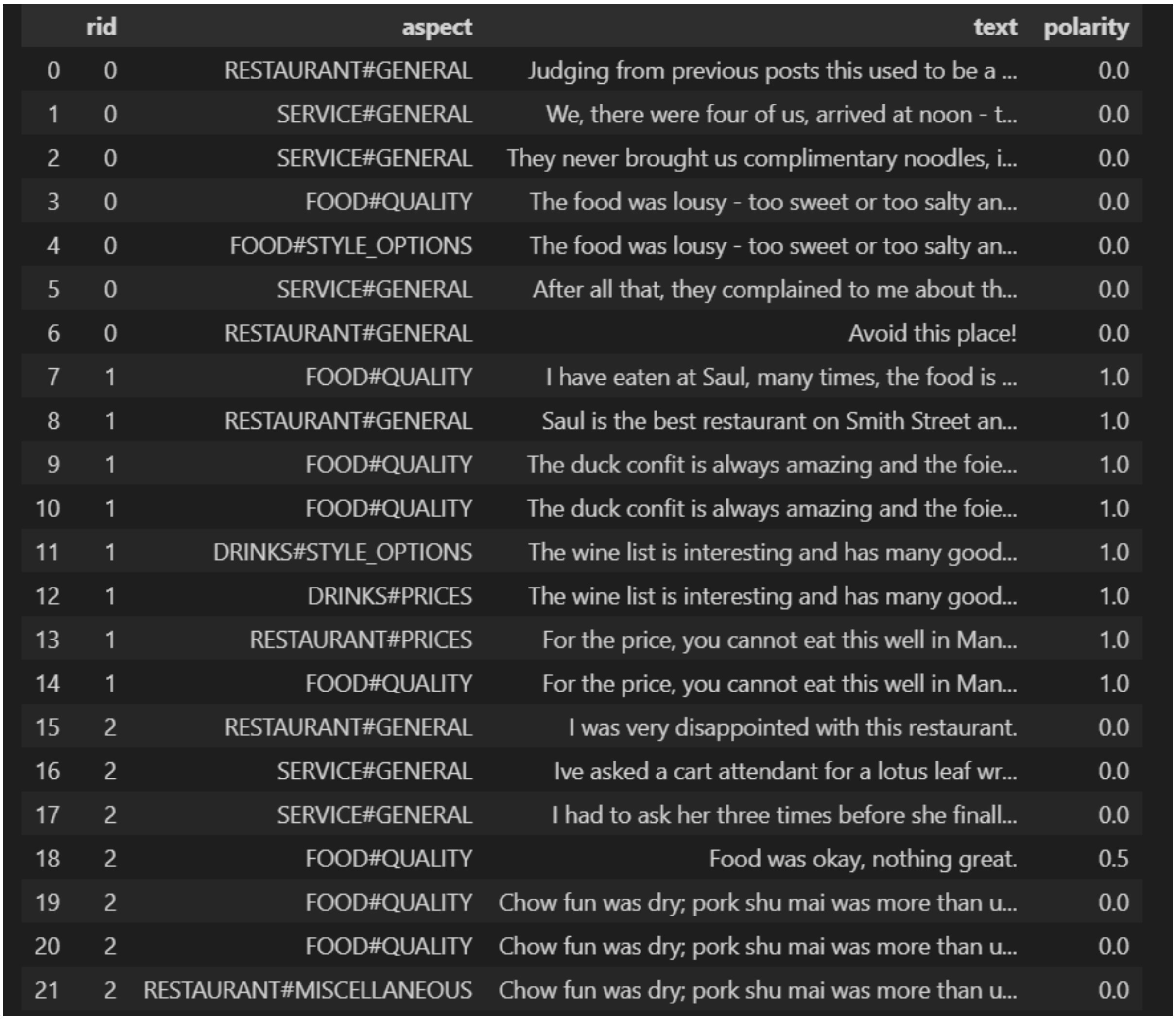
1. Review: There are many reviews, and each review has many sentences.
2. Aspect: It is defined for each sentence. It contains an entity and attribute. For e.g.: SERVICE#GENERAL, FOOD#QUALITY etc.
3. Text: The text which is present in the sentence.
4. Polarity: It is defined for each sentence.

   1 ----- positive
   0 ----- negative

APPROACH

Data Extraction

This is the firs step. Every sentence has some attributes. So, we extracted the required parameters from the dataset. After extracting the required parameters from the dataset, stored the data in a systematic form in a pandas dataframe. It is shown below as follows:



	rid	aspect	text	polarity
0	0	RESTAURANT#GENERAL	Judging from previous posts this used to be a ...	0.0
1	0	SERVICE#GENERAL	We, there were four of us, arrived at noon - t...	0.0
2	0	SERVICE#GENERAL	They never brought us complimentary noodles, i...	0.0
3	0	FOOD#QUALITY	The food was lousy - too sweet or too salty an...	0.0
4	0	FOOD#STYLE_OPTIONS	The food was lousy - too sweet or too salty an...	0.0
5	0	SERVICE#GENERAL	After all that, they complained to me about th...	0.0
6	0	RESTAURANT#GENERAL	Avoid this place!	0.0
7	1	FOOD#QUALITY	I have eaten at Saul, many times, the food is ...	1.0
8	1	RESTAURANT#GENERAL	Saul is the best restaurant on Smith Street an...	1.0
9	1	FOOD#QUALITY	The duck confit is always amazing and the foie...	1.0
10	1	FOOD#QUALITY	The duck confit is always amazing and the foie...	1.0
11	1	DRINKS#STYLE_OPTIONS	The wine list is interesting and has many good...	1.0
12	1	DRINKS#PRICES	The wine list is interesting and has many good...	1.0
13	1	RESTAURANT#PRICES	For the price, you cannot eat this well in Man...	1.0
14	1	FOOD#QUALITY	For the price, you cannot eat this well in Man...	1.0
15	2	RESTAURANT#GENERAL	I was very disappointed with this restaurant.	0.0
16	2	SERVICE#GENERAL	Ive asked a cart attendant for a lotus leaf wr...	0.0
17	2	SERVICE#GENERAL	I had to ask her three times before she finall...	0.0
18	2	FOOD#QUALITY	Food was okay, nothing great.	0.5
19	2	FOOD#QUALITY	Chow fun was dry; pork shu mai was more than u...	0.0
20	2	FOOD#QUALITY	Chow fun was dry; pork shu mai was more than u...	0.0
21	2	RESTAURANT#MISCELLANEOUS	Chow fun was dry; pork shu mai was more than u...	0.0

Each review has many sentences.

The columns represent:

Rid --- > Review Id

Aspect --- > It represents the aspect

Text --- > Text in the sentence

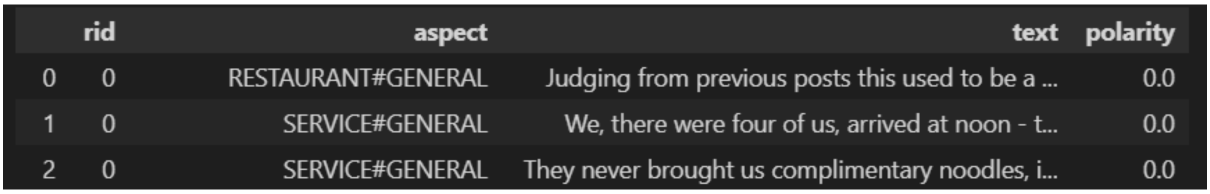
Polarity --- > Positive/Negative (1 – positive, 0 – negative)

Data Cleaning

In this step we remove the unnecessary and unwanted information from the data. We remove HTML, emotes, links and punctuations from the dataset. We also remove the stop words (like “a”, “the” etc) from the dataset to improve the efficiency.

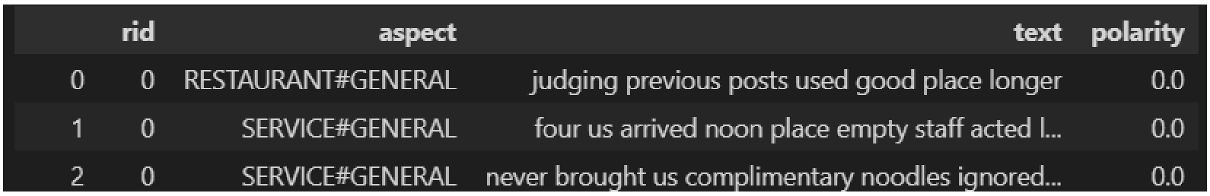
Below it is shown how the data looked before and after data cleaning:

Before



	rid	aspect	text	polarity
0	0	RESTAURANT#GENERAL	Judging from previous posts this used to be a ...	0.0
1	0	SERVICE#GENERAL	We, there were four of us, arrived at noon - t...	0.0
2	0	SERVICE#GENERAL	They never brought us complimentary noodles, i...	0.0

After



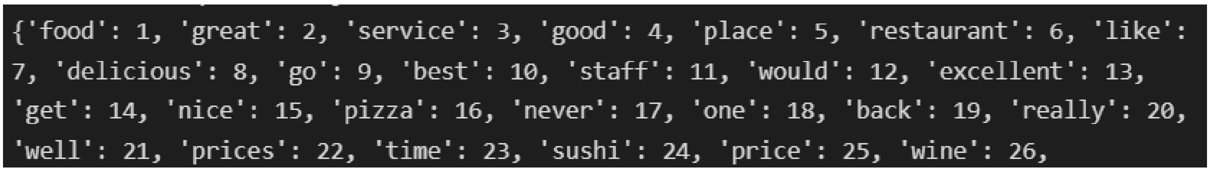
	rid	aspect	text	polarity
0	0	RESTAURANT#GENERAL	judging previous posts used good place longer	0.0
1	0	SERVICE#GENERAL	four us arrived noon place empty staff acted l...	0.0
2	0	SERVICE#GENERAL	never brought us complimentary noodles ignored...	0.0

In this first text we can see that the before data cleaning words like “this”, “from”, “to” are present but after data cleaning they have been removed. This has been done for all sentences.

Data Pre-processing

Now comes data pre-processing. Here we will convert each sentence into vectors of sentence size and represent each word using word index.

The word index is shown below:



```
{'food': 1, 'great': 2, 'service': 3, 'good': 4, 'place': 5, 'restaurant': 6, 'like':
7, 'delicious': 8, 'go': 9, 'best': 10, 'staff': 11, 'would': 12, 'excellent': 13,
'get': 14, 'nice': 15, 'pizza': 16, 'never': 17, 'one': 18, 'back': 19, 'really': 20,
'well': 21, 'prices': 22, 'time': 23, 'sushi': 24, 'price': 25, 'wine': 26,
```

Now we will do padding to ensure all the vectors are of the same length. Also, we will prepare the required deliverables for the embedding matrix.

Embeddings

We used Glove 27B, 100-dimensional pre-trained word embeddings.

Now we used the Glove model to create word embeddings for every word in the corpus