BERT Sentiment Analysis- YouTube Comments & Restaurant Reviews

CS 688
Term Project
-Harshil Khara

Problem Statement

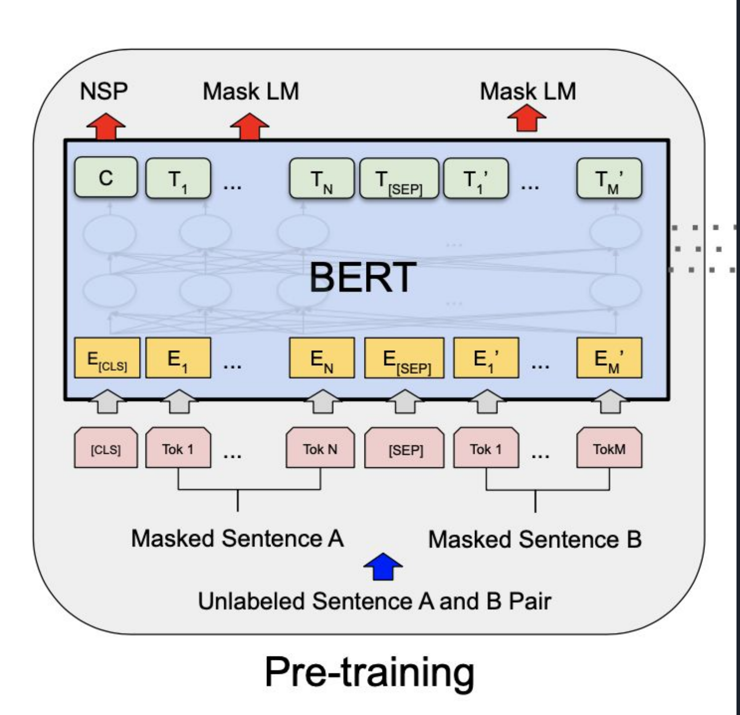
Using BERT NLP model apply it on the texts (comments & reviews) that are scrapped from YouTube and Restaurant reviews site and do a sentiment analysis based on it.

BERT (BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS)

BERT was pretrained on two tasks: **language modelling** (15% of tokens were masked and BERT was trained to predict them from context) and **next sentence prediction** (BERT was trained to predict if a chosen next sentence was probable or not given the first sentence). As a result of the training process, BERT learns contextual embeddings for words.

After pretraining, which is computationally expensive, BERT can be fine tuned with less resources on smaller datasets to optimize its performance on specific tasks.

- MLM enables/enforces bidirectional learning from text by masking (hiding) a word in a sentence and forcing BERT to bidirectionally use the words on either side of the covered word to predict the masked word. This had never been done before!

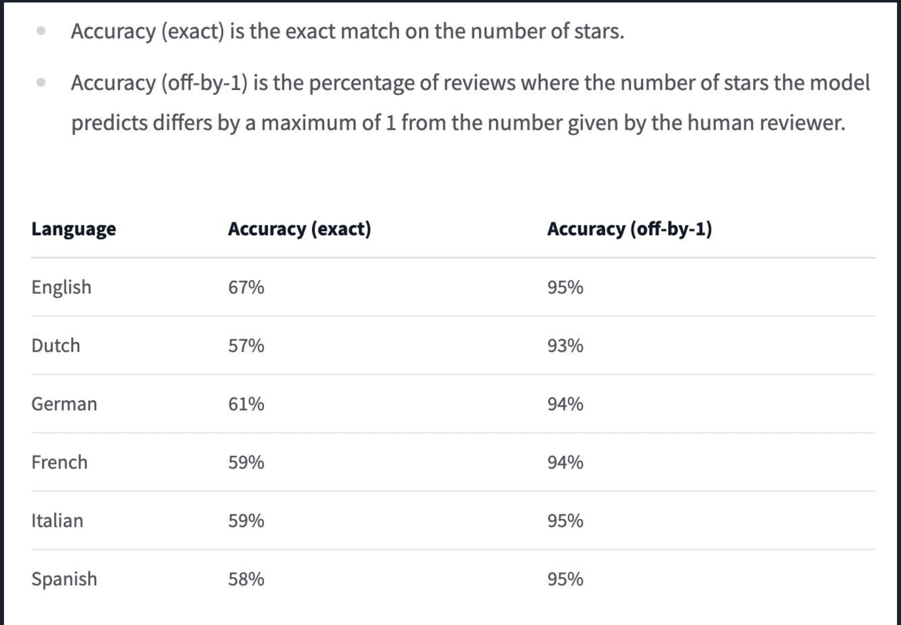


Pre-training

Picking the right BERT Model for our Sentiment Analysis

- This a bert-base-multilingual-uncased model finetuned for sentiment analysis on product reviews in six languages: English, Dutch, German, French, Spanish and Italian. It predicts the sentiment of the review as a number of stars (between 1 and 5).

- This model is intended for direct use as a sentiment analysis model for product reviews in any of the six languages above, or for further fine tuning on related sentiment analysis tasks.



- Accuracy (exact) is the exact match on the number of stars.
- Accuracy (off-by-1) is the percentage of reviews where the number of stars the model predicts differs by a maximum of 1 from the number given by the human reviewer.

Language	**Accuracy (exact)**	**Accuracy (off-by-1)**
English	67%	95%
Dutch	57%	93%
German	61%	94%
French	59%	94%
Italian	59%	95%
Spanish	58%	95%

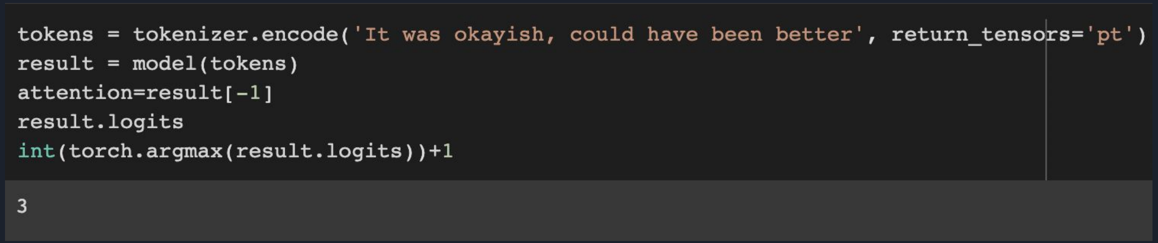
Link to this bert-base-multilingual sentiment analysis model-

https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

Passing a token and checking the sentiment score

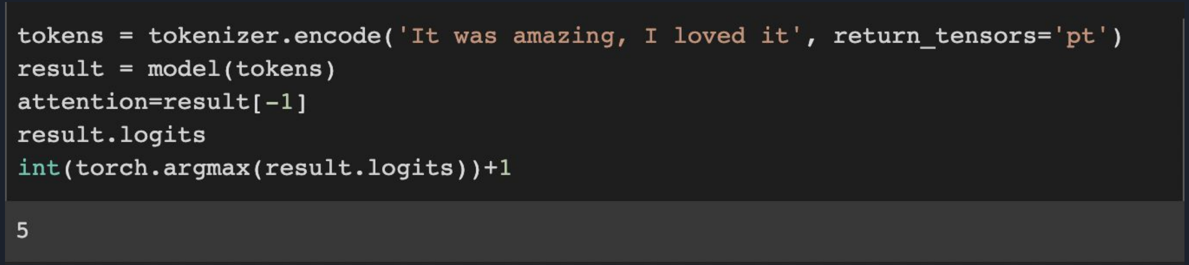
We pass some random tokens and check if the BERT model is correctly scoring it or not.

Note- Scoring is on a scale of 1 to 5.



```
tokens = tokenizer.encode('It was okayish, could have been better', return_tensors='pt')
result = model(tokens)
attention=result[-1]
result.logits
int(torch.argmax(result.logits))+1
```

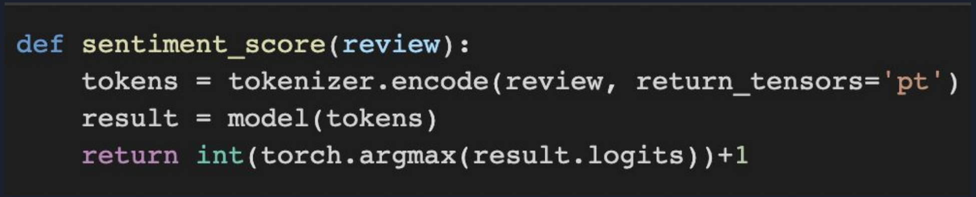
```
3
```



```
tokens = tokenizer.encode('It was amazing, I loved it', return_tensors='pt')
result = model(tokens)
attention=result[-1]
result.logits
int(torch.argmax(result.logits))+1
```

```
5
```

Our BERT Sentiment Score function



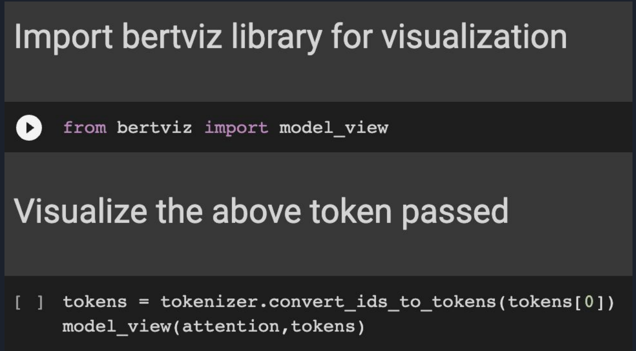
```
def sentiment_score(review):
    tokens = tokenizer.encode(review, return_tensors='pt')
    result = model(tokens)
    return int(torch.argmax(result.logits))+1
```

Visualizing our model

Using the bertviz we visualize the first token we passed in our model.

- CLS- Start of the sequence

- SEP- Next Sentence

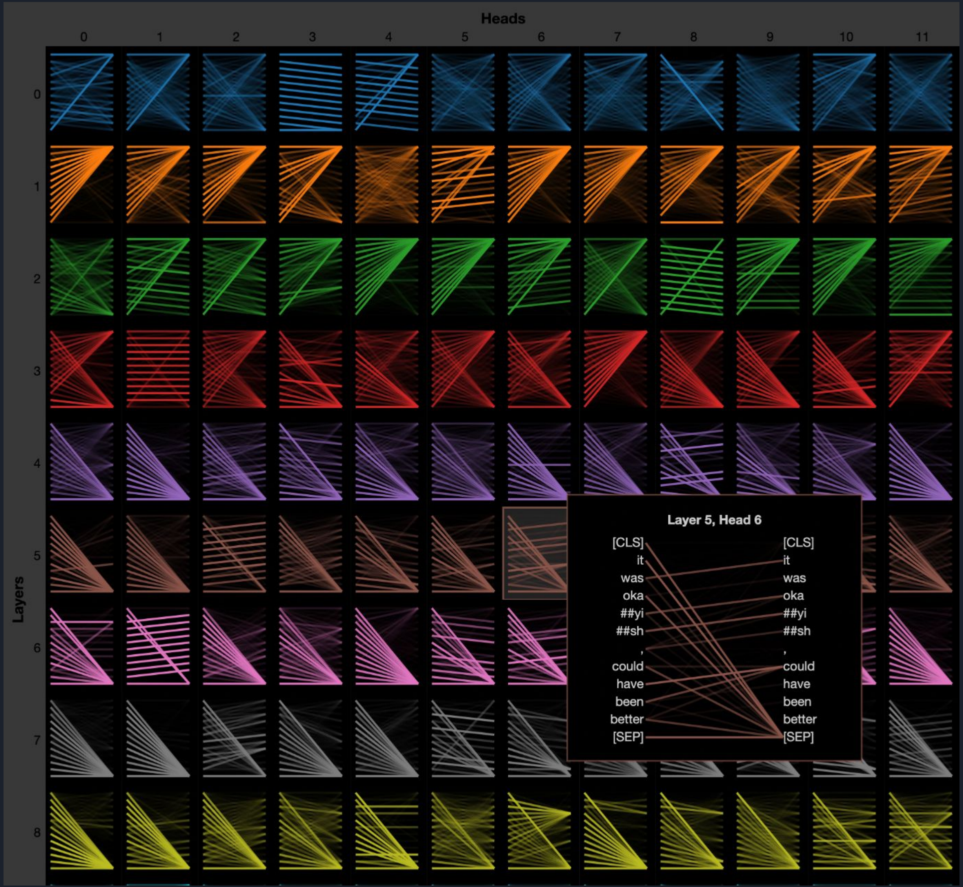


Import bertviz library for visualization

```
from bertviz import model_view
```

Visualize the above token passed

```
[ ]  tokens = tokenizer.convert_ids_to_tokens(tokens[0])
     model_view(attention,tokens)
```

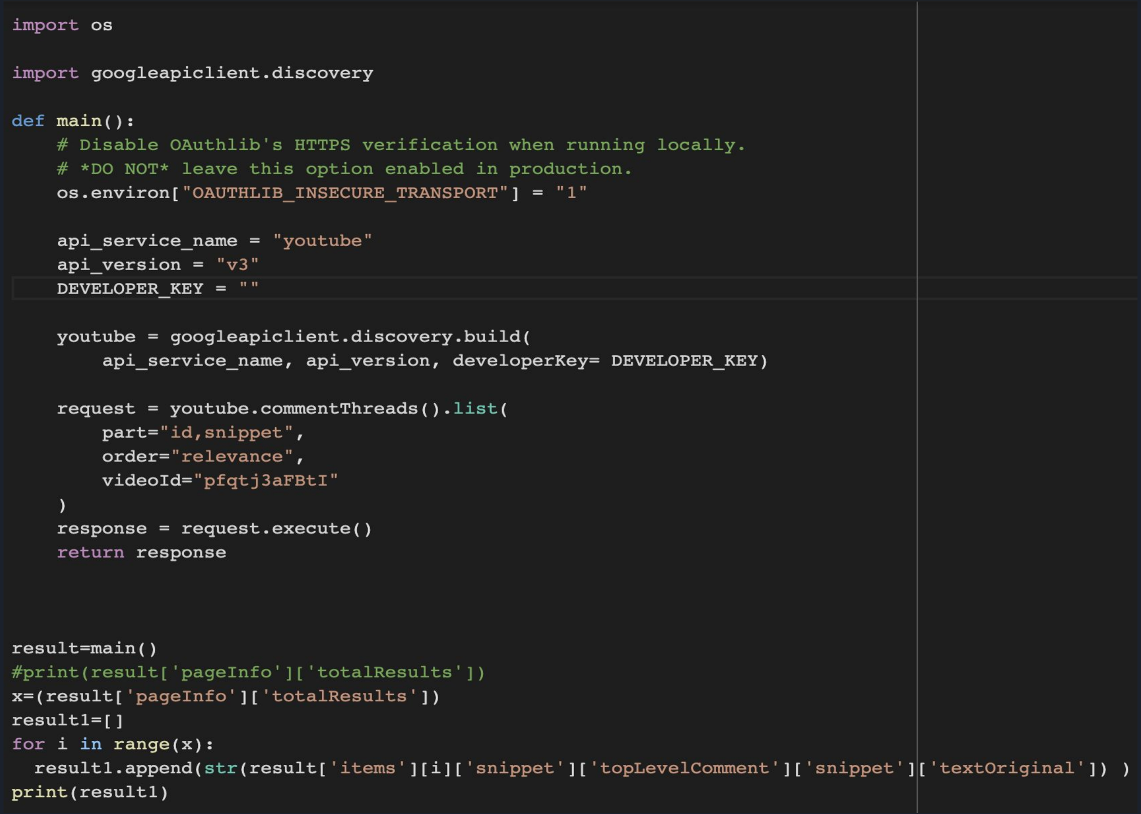


Scrapping the YouTube Comments

Now that YouTube has removed the dislike counter we will scrape the comments and have a sentiment analysis of comments on that video.

I have used YouTube API to scrape the comments (according to relevance)

Note- Each person will have their unique developer key

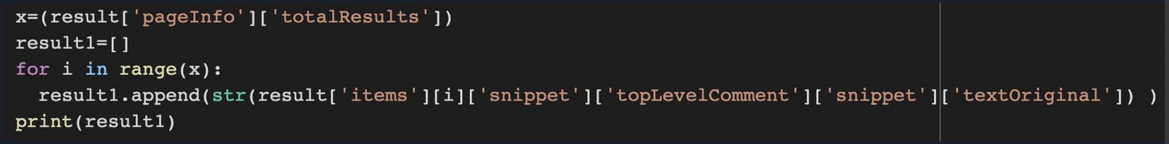


```python
import os

import googleapiclient.discovery

def main():
    # Disable OAuthlib's HTTPS verification when running locally.
    # *DO NOT* leave this option enabled in production.
    os.environ["OAUTHLIB_INSECURE_TRANSPORT"] = "1"

    api_service_name = "youtube"
    api_version = "v3"
    DEVELOPER_KEY = ""

    youtube = googleapiclient.discovery.build(
        api_service_name, api_version, developerKey= DEVELOPER_KEY)

    request = youtube.commentThreads().list(
        part="id,snippet",
        order="relevance",
        videoId="pfqtj3aFBtI"
    )
    response = request.execute()
    return response



result=main()
#print(result['pageInfo']['totalResults'])
x=(result['pageInfo']['totalResults'])
result1=[]
for i in range(x):
  result1.append(str(result['items'][i]['snippet']['topLevelComment']['snippet']['textOriginal']) )
print(result1)
```

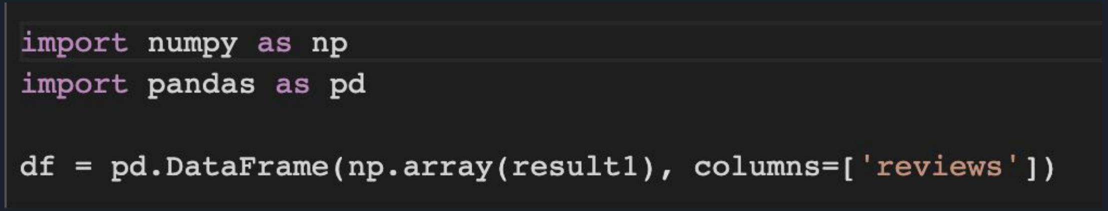
Output & Preprocessing of the YouTube Comments Scrapping

The output isn't directly in this dataframe form but comes out in JSON. You then have to extract the comments accordingly and pass it on to the dataframe.

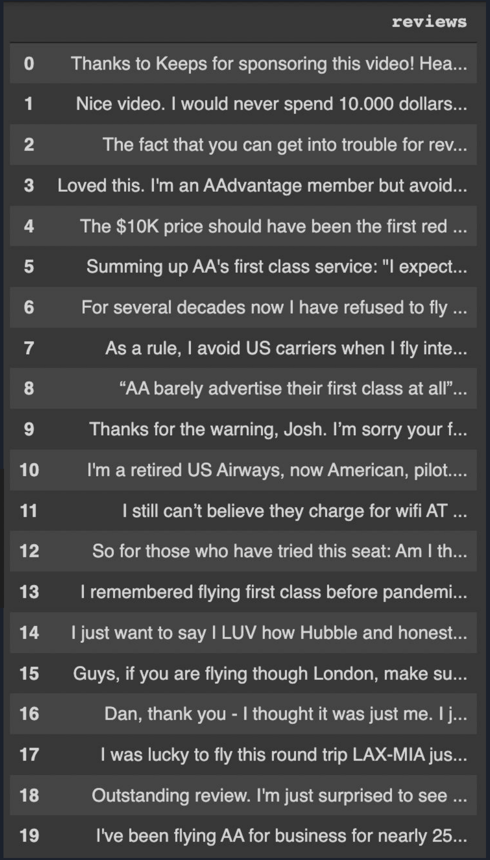
The snippet that fetches the comment we are scrapping-



```
x=(result['pageInfo']['totalResults'])
result1=[]
for i in range(x):
  result1.append(str(result['items'][i]['snippet']['topLevelComment']['snippet']['textOriginal']) )
print(result1)
```



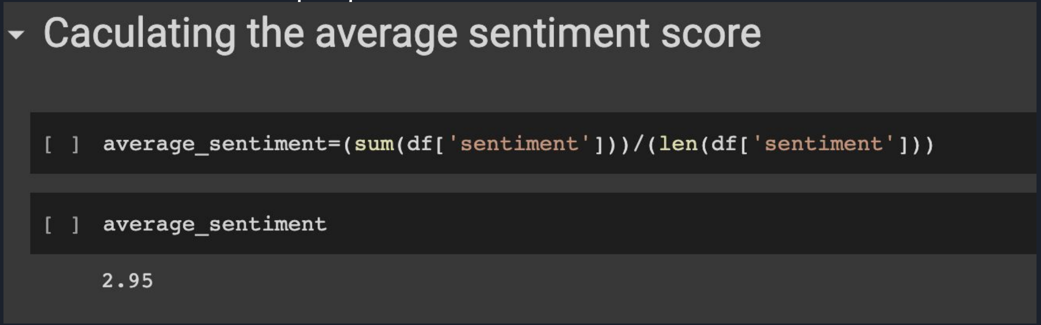
```
import numpy as np
import pandas as pd

df = pd.DataFrame(np.array(result1), columns=['reviews'])
```



	reviews
0	Thanks to Keeps for sponsoring this video! Hea...
1	Nice video. I would never spend 10.000 dollars...
2	The fact that you can get into trouble for rev...
3	Loved this. I'm an AAdvantage member but avoid...
4	The $10K price should have been the first red ...
5	Summing up AA's first class service: "I expect...
6	For several decades now I have refused to fly ...
7	As a rule, I avoid US carriers when I fly inte...
8	“AA barely advertise their first class at all”...
9	Thanks for the warning, Josh. I'm sorry your f...
10	I'm a retired US Airways, now American, pilot....
11	I still can't believe they charge for wifi AT ...
12	So for those who have tried this seat: Am I th...
13	I remembered flying first class before pandemi...
14	I just want to say I LUV how Hubble and honest...
15	Guys, if you are flying though London, make su...
16	Dan, thank you - I thought it was just me. I j...
17	I was lucky to fly this round trip LAX-MIA jus...
18	Outstanding review. I'm just surprised to see ...
19	I've been flying AA for business for nearly 25...

BERT Sentiment Analysis on the comments

After applying the BERT Text Sentiment Analysis on the above scrapped comments we get this



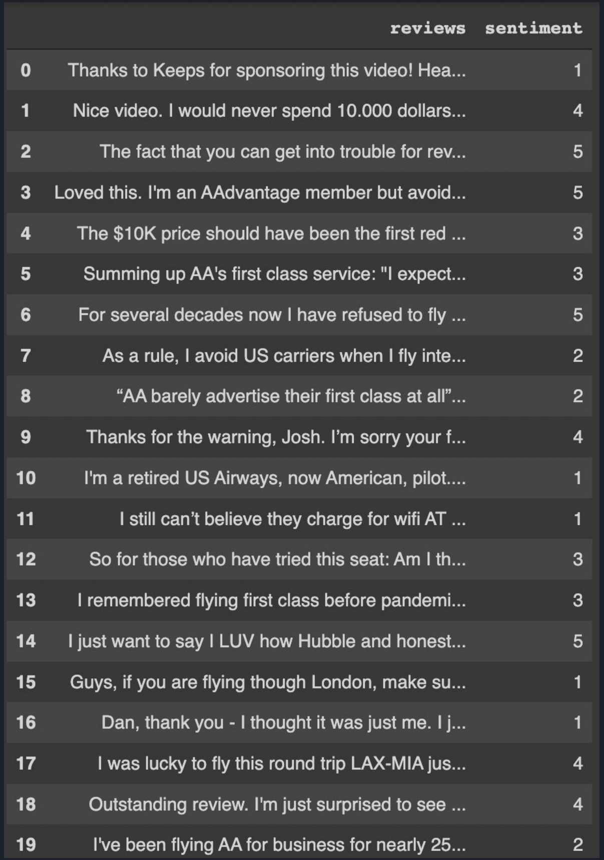
Caculating the average sentiment score

```
[ ]  average_sentiment=(sum(df['sentiment']))/(len(df['sentiment']))
```

```
[ ]  average_sentiment
```

```
2.95
```

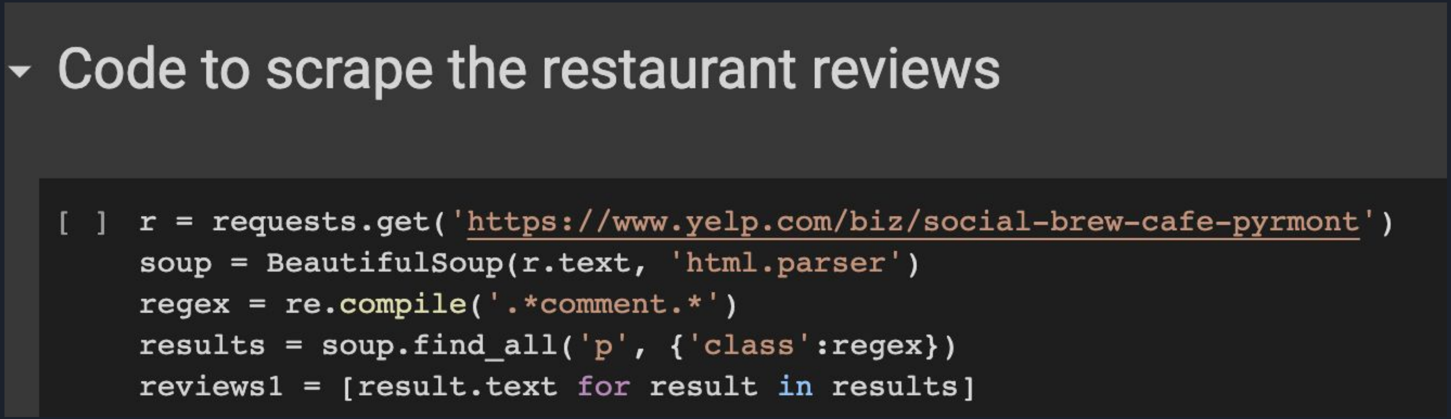
The average sentiment score comes out to be 2.95



	reviews	sentiment
0	Thanks to Keeps for sponsoring this video! Hea...	1
1	Nice video. I would never spend 10.000 dollars...	4
2	The fact that you can get into trouble for rev...	5
3	Loved this. I'm an AAdvantage member but avoid...	5
4	The $10K price should have been the first red ...	3
5	Summing up AA's first class service: "I expect...	3
6	For several decades now I have refused to fly ...	5
7	As a rule, I avoid US carriers when I fly inte...	2
8	“AA barely advertise their first class at all”...	2
9	Thanks for the warning, Josh. I’m sorry your f...	4
10	I'm a retired US Airways, now American, pilot....	1
11	I still can’t believe they charge for wifi AT ...	1
12	So for those who have tried this seat: Am I th...	3
13	I remembered flying first class before pandemi...	3
14	I just want to say I LUV how Hubble and honest...	5
15	Guys, if you are flying though London, make su...	1
16	Dan, thank you - I thought it was just me. I j...	1
17	I was lucky to fly this round trip LAX-MIA jus...	4
18	Outstanding review. I'm just surprised to see ...	4
19	I've been flying AA for business for nearly 25...	2

Scrapping the Restaurant Reviews

I have used beautifulsoup4 library to scrap the restaurant review.

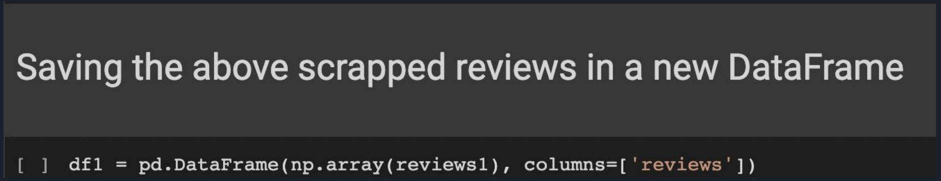


Code to scrape the restaurant reviews

```python
r = requests.get('https://www.yelp.com/biz/social-brew-cafe-pyrmont')
soup = BeautifulSoup(r.text, 'html.parser')
regex = re.compile('.*comment.*')
results = soup.find_all('p', {'class':regex})
reviews1 = [result.text for result in results]
```

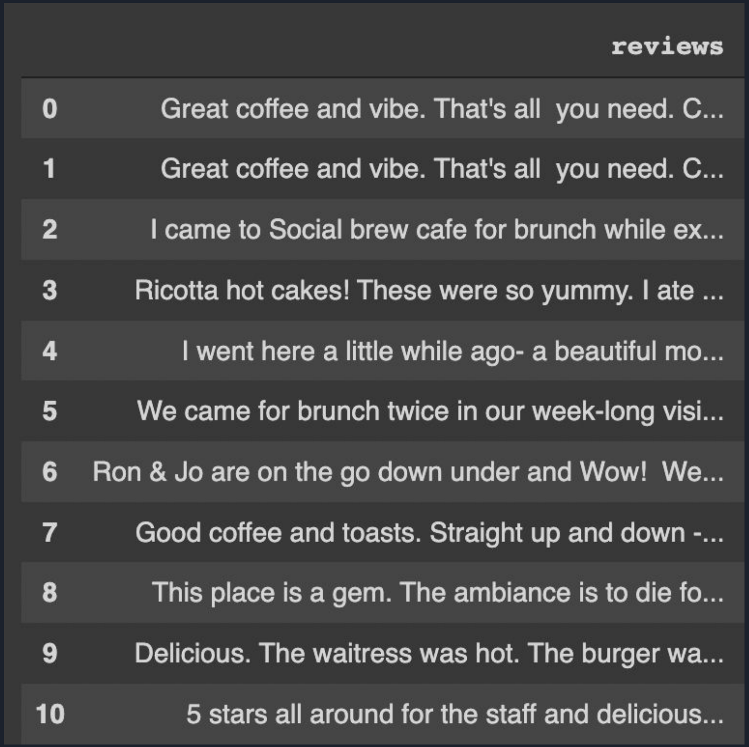
Output & Preprocessing of the reviews

The output isn’t directly in this dataframe form but comes out in text form. You then have to extract the reviews accordingly and pass it on to the dataframe.



Saving the above scrapped reviews in a new DataFrame

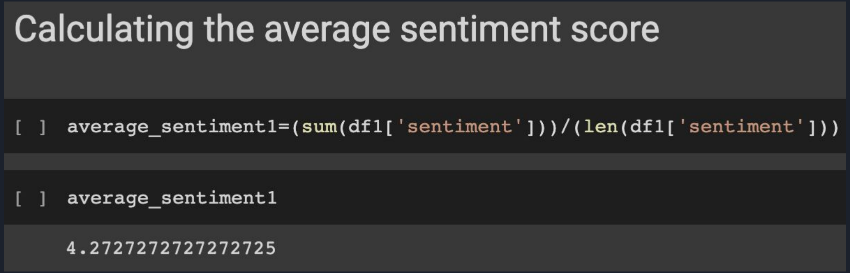
```
[ ]  df1 = pd.DataFrame(np.array(reviews1), columns=['reviews'])
```



	reviews
0	Great coffee and vibe. That's all you need. C...
1	Great coffee and vibe. That's all you need. C...
2	I came to Social brew cafe for brunch while ex...
3	Ricotta hot cakes! These were so yummy. I ate ...
4	I went here a little while ago- a beautiful mo...
5	We came for brunch twice in our week-long visi...
6	Ron & Jo are on the go down under and Wow! We...
7	Good coffee and toasts. Straight up and down -...
8	This place is a gem. The ambiance is to die fo...
9	Delicious. The waitress was hot. The burger wa...
10	5 stars all around for the staff and delicious...

BERT Sentiment Analysis on the reviews

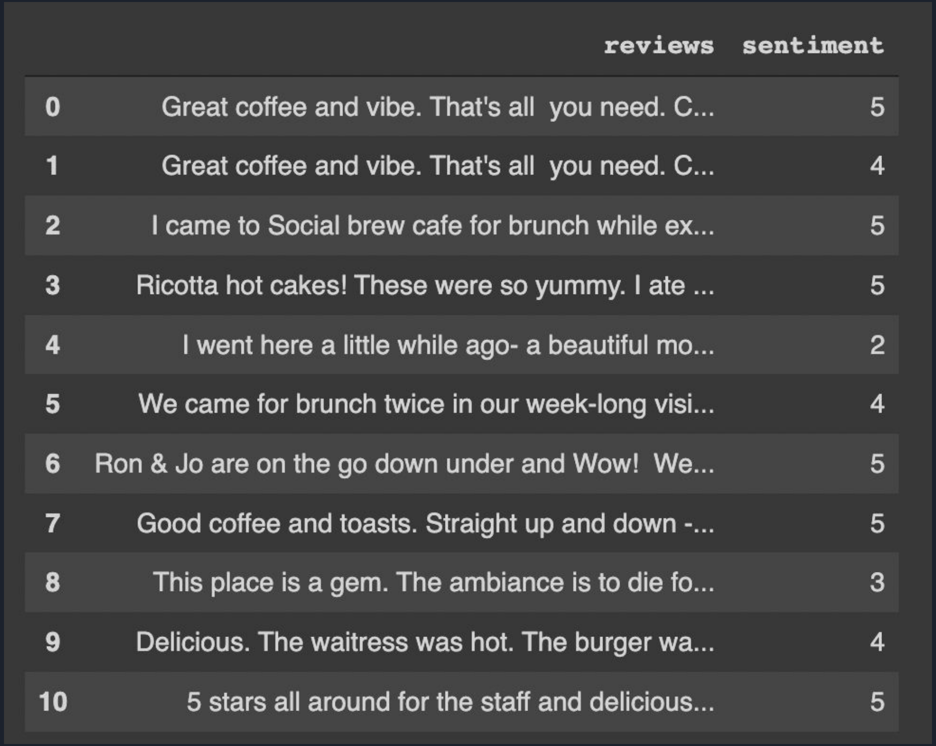
After applying the BERT Text Sentiment Analysis on the above scrapped reviews we get this output.



Calculating the average sentiment score

```
[ ]  average_sentiment1=(sum(df1['sentiment']))/(len(df1['sentiment']))
```

```
[ ]  average_sentiment1

     4.2727272727272725
```



	reviews	sentiment
0	Great coffee and vibe. That's all you need. C...	5
1	Great coffee and vibe. That's all you need. C...	4
2	I came to Social brew cafe for brunch while ex...	5
3	Ricotta hot cakes! These were so yummy. I ate ...	5
4	I went here a little while ago- a beautiful mo...	2
5	We came for brunch twice in our week-long visi...	4
6	Ron & Jo are on the go down under and Wow! We...	5
7	Good coffee and toasts. Straight up and down -...	5
8	This place is a gem. The ambiance is to die fo...	3
9	Delicious. The waitress was hot. The burger wa...	4
10	5 stars all around for the staff and delicious...	5

The average sentiment score comes out to be 4.27

Conclusion

- Thus, we have successfully applied BERT NLP model for sentiment analysis on the YouTube comments and restaurant reviews by using the tools of web analytics & mining.

- By using BERT NLP model for sentiment analysis we got a deeper insight on the scrapped comments and reviews which we can use for our decision making of that particular texts.

Thank you