For this assignment I focused on the review comments made by different people for their stay in Airbnb. The review comments made by a user can be used to get an understanding of the overall quality of the service provided by the Hotel/House/Rooms etc.

For simplicity I decided to label the user comments as either positive or negative. I could have picked other criteria for classifying the user comments on star basis (like the items listed in amazon). A **1-star rating** for bad user comments and **5-star** **rating** for a great user comment or could have labelled the comments in 3 different classed namely **positive, negative & neutral**.

**Text Classification**

1. **Text classification** also known as text tagging or text categorization is the process of assigning tags/labels to unstructured text. Using Natural Language Processing (NLP), text classifiers can automatically analyze text and then assign a set of pre-defined tags or categories based on its content.
2. **Motivation to use a Transformer architecture-based model -** The transformers are the latest and advanced models that give the state-of-the-art results for a wide range of tasks such as **text / sequence classification**, **named entity recognition (NER)**, **question answering**, **machine translation**, **text summarization**, **text generation**, etc.
3. **What is Zero-Shot Classification -** Zero-Shot learning is a specific area of machine learning where we want the model to classify data based on very few or even no training example. In **Zero-shot learning** the classes covered in the training data and the classes we wish to classify are completely different.

As we don’t have a training labels for the review comments made by the user in our dataset, the zero-shot classifier will be the best choice for predicting the labels of the user comments.

The **ZeroShotClassifier** model of the [Hugging Face Transformers](https://huggingface.co/transformers/v3.0.2/index.html) library  **classifies an input sequence from a list of candidate labels**. The model assumes by default that only one of the candidate labels is true and returns a list of scores for each label which add up to 1. Visit [this link](https://huggingface.co/models?pipeline_tag=zero-shot-classification) to learn more about the available models for **zero-shot-classification** task.

One thing I noticed while going through the user comments is that the comments are made in languages other than **English**. I saw user comments in **Russian** languages as well. So, we need a model that can handle multi-lingual text as well.

The **ZeroShotClassifier** model has been fine-tuned on [XNLI](https://cims.nyu.edu/~sbowman/xnli/) corpus which includes 15 languages: Arabic, Bulgarian, Chinese, English, French, German, Greek, Hindi, Russian, Spanish, Swahili, Thai, Turkish, Urdu, and Vietnamese. So, this model can be used to classify **multi-lingual** text as well.

As this assignment was time bound and I only have 8GB GPU on my local system, I decided to get labels for comments made in year 2021 only. Same approach can be taken to get labels for all the user comments in the csv but will eventually take lot of time to get the target labels/sentiments.

Once we have the labels/sentiments for each of the user comments made in year 2021, then it was very easy to do the analysis on

* General public sentiment of people staying in Airbnb for the year 2021
* People views on different neighborhood
* People views on different room-types
* People views on accommodations categorized under different price range.
* **General public sentiment of people staying in Airbnb for the year 2021**

Once we have the user comment sentiments, we can plot a histogram depicting the count of the positive and negative comments. (Graph in the notebook)

* **People views on different neighborhood**

Of all the 33 neighborhoods how are they compared to each other. A higher number of comments for a given neighborhood also indicates that people will generally prefer a locality.

But if we can label the people reviews for a given neighborhood as well, then we get an idea of what people are saying (positive/negative) about a given neighborhood/locality, i.e. if a locality is good or bad. For this I plotted a stacked bar graph for the count of positive and negative review for every neighborhood. (Graph in the notebook)

* **People views on different room type**

Of all the 4 room-types how are they compared to each other. A higher number of reviews for a given room type also indicates that people will generally give preference to a room-type.

But if we can label the people reviews for a given room-type as well, then we get an idea of what people are saying (positive/negative) about a given room-type. For this I plotted a stacked bar graph for the count of positive and negative review for every room-type. (Graph in the notebook)

* **People views on accommodations categorized under different price range**

Same analysis can be done for the accommodations categorized under different price range categories. I have decided to categorize accommodations into these price ranges (Graph in the notebook) –

* 1. Price less than 50
  2. Price >= 50 but less than 100
  3. Price >= 100 but less than 200
  4. Price >= 200 but less than 500
  5. Price >= 500 but less than 1000
  6. Price >= 1000