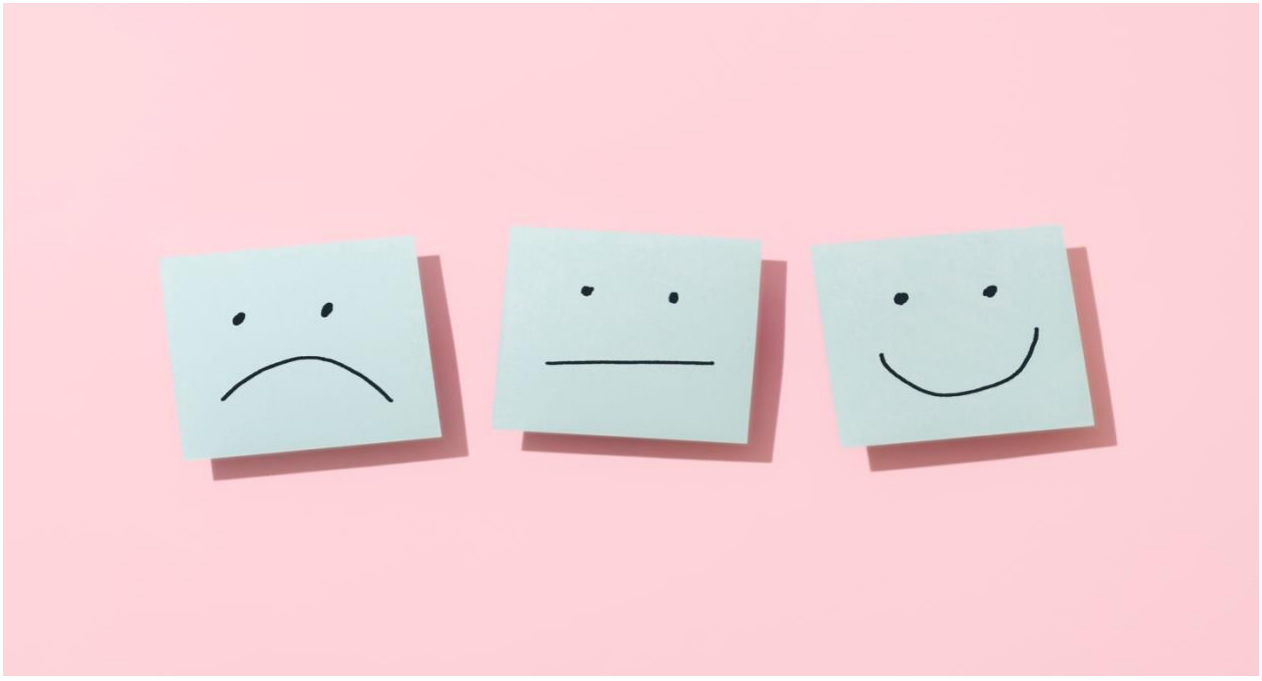
Twitter Sentiment Analysis: Classification Vs Deep Learning Approach



Project Proposal by: Himani Kaushik

- **Question/need:**
  - What is the question behind your analysis or model and what practical impact will your work have?
    - The purpose of the model is to analyze tweets and predict whether it expresses positive or negative sentiment. The model aims to implement and evaluate Naive Bayes model for classification and LSTM model for deep learning, and then select the better performing model for the client.
  - Who is your client and how will that client benefits from exploring this question or building this model/system?
    - A local client wants to implement a sentiment analysis model utilizing the tweets obtained about the product. This will help them detect angry customers or negative emotions before they escalate.

- **Data Description:**
  - What dataset(s) do you plan to use, and how will you obtain the data?
    - The data will be obtained from the kaggle website - https://www.kaggle.com/datasets/kazanova/sentiment140 .
    - The dataset has 1.6 million tweets extracted using the twitter api. The tweets can be used to detect sentiment, as they have been annotated.
    - Since the dataset is too huge to work on a regular machine, I plan to trim the dataset to $1/4^{th}$ of its original size. The data balance will be maintained while trimming the data.
  - What is an individual sample/unit of analysis in this project?
    - The individual sample includes six fields:
      - target: Polarity of the tweet (*0* = negative, *2* = neutral, *4* = positive)
      - ids: Id of the tweet
      - date: Date of the tweet
      - flag: Query. If there is no query, then this value is NO_QUERY.
      - user: User that tweeted
      - text: Text of the tweet

- **Tools:**
  - How do you intend to meet the tools requirement of the project?
    - Following Python tools would be used:
      - Pandas and Numpy: For cleaning data and preprocessing.
      - NLTK, scikit-learn, keras: For processing data, Naive Bayes classification, GloVe embedding, LSTM
      - Matplotlib and Seaborn: For visualizing the data.

- **MVP Goal:**
  - What would a minimum viable product (MVP) look like for this project?
    - MVP for the project would include few data visualizations after the data is cleaned, tokenized and lemmatized. It will also include the initial runs of Naive Bayes model and word embeddings.