**TWEET SENTIMENT EXTRACTION**

What I have done so far:

* I started by exploring the tweet dataset of kaggle and various EDA notebooks. After analysing them following conclusions were drawn:
  1. Discussion on this link: <https://www.kaggle.com/c/tweet-sentiment-extraction/discussion/138520> pointed out that neutral tweets have a jaccard similarity of about 97 percent between text and selected\_text.
  2. URLs do not make much sense for positive and negative sentiments. They are more inclined towards the neutral side.
  3. Average length of words in the selected text is around 7. Also, selected text is always a continuous segment of words from the tweet.
  4. Also, for the best jaccard similarity, we need to extract the exact words from the tweet as selected text. Even a change of punctuation will lead to comparatively bad jaccard similarity.
* Next, I analysed solution notebooks present on kaggle to find out various solution approaches for the problem. The problem can be solved in the following approaches:
  1. Use an approach similar to question answering problem using Roberta model as indicated by the following links: <https://www.kaggle.com/cdeotte/tensorflow-roberta-0-705> <https://www.kaggle.com/jonathanbesomi/question-answering-starter-pack>
  2. Simulate the problem as NER problem. [https://www.kaggle.com/rohitsingh9990/ner-training-using-spacy-ensemble](https://www.kaggle.com/rohitsingh9990/ner-training-using-spacy-ensemble/comments) <https://www.kaggle.com/doomdiskday/full-tutorial-eda-to-dnns-all-you-need/notebook>
  3. Use model interpretation tools i.e. first train a model to predict the sentiment from the original text and then use tools like LIME to find words that indicate this sentiment.
* After analysing these approaches and the public scores that are achieved by other members on kaggle, we decided to go ahead with the BERT based models.

Understanding BERT:

* For BERT, I read the paper titled RoBERTa: A Robustly Optimized BERT Pretraining Approach <https://arxiv.org/pdf/1907.11692.pdf>
* And the article on the link: <https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270>.
* Next, I have to read the paper on<https://arxiv.org/abs/1706.03762> and use the knowledge gained so far to solve the problem.