**Project proposal**

# Question and needs

Nowadays, social media produces a high quantity of textual content that grows exponentially. As they become the main communication channel, accessed by all segments of society. For instance, Twitter records more than 500 million tweets per day. Therefore, such platforms are the floor to express opinions, share feelings, build and develop relationships either by posts, comments, likes or direct messages. Therefore, we conclude the importance of social media from its capability of granting access to a huge amount of data. Therefore, the following motivations and needs can be concluded:

* + Can we extract valuable patterns from this data to make decisions?
  + How can we prepare this data to feed it into various machine learning models?

Speaking of needs, a use case is to analyze US Airline Service, where Airline companies are based on traditional feedback forms which are time consuming. Thus, twitter data acts as a valuable source to extract customer feedback tweets to build a sentiment classifier.

# Data set

The data set we plan to use is a twitter data set available at: [https://www.kaggle.com/](https://www.kaggle.com/crowdflower/twitter-airline-sentiment) [crowd flower/twitter-airline-sentiment](https://www.kaggle.com/crowdflower/twitter-airline-sentiment) . It is released by Crowd Flower and has a total of 14640 tweets. Tweets of six major US Airlines are evolved: US Airways, Southwest, United, Virgin America, Delta. The tweets are labeled as negative, positive and neutral sentiment. Thus, the ability to apply our supervised learning models.

# Methodology

Our proposed model will be based on comparative study of 3 ML models: Random Forest, logistic regression and Adaptive boosting. The predicted target is the customer sentiment, and the input features are text features after the necessary processing. The data will be fragmented into train and validation sets, then a supervised model will be trained for each approach. The test data will be used to test each model against multiple metrics including: Precision, Recall and overall accuracy. Adapting multiple metrics will give a clear vision about the prediction accuracy of each of the labels (sentiments).

# Tools

Python libraries will be used in all the steps: data preparation, train and test. The main libraries are Pandas and nltk to load and prepare the data, sckit-learn to build and train the 3 models and also for accuracy measurements. On the other hand, Jupyter notebook is the framework that we will adapt for implementation for its simplicity and ability to write documentation along with python code. It is a strong tool in representing the whole project with explanations in one single report.

# MVP goal

* + **Data loading**: It will be my first time where I will load the data.
  + **Data preparation:** Clear the data from any missing values, Extract numerical features from text tweets, Split the data to train and test sets.
  + **Data visualisation**: Plot some statistics about the data like the number of tweets for each label.
  + **Training:** Define each of the tree models and train it on training data.
  + **Hyper-parameters tuning:** Re-train each of the models for a combination of hyper-parameters to get the optimal set for each one.
  + **Testing:** Test each model using different classification metrics like Recall, precision and F1-score.
  + **Comparison**: Based on metrics values, we will compare the 3 models to choose the best one.
  + **Discussion:** Results interpretation.
  + **Conclusion:** Set the final results in brief and open challenges for future improvements.