# Introduction:

1.6 million tweets have been tagged for sentiment analysis using emoticons to classify polarity in the sentiment 140 dataset. Produced under remote supervision, it provides a scalable method for data labeling. A study report and thorough documentation are examples of resources.

# Data Collection:

**Data Set:** This data used for this assignment is publicly enviable at Kaggle with the name of Sentiment140 dataset with 1.6 million tweets.

# Methodology:

**Data Preprocessing:** Used RegEx for preprocessing to clean data and removed emojis, hashtags, newlines, and converted text to lowercase.

**Data Lebling:** Labeled the whole data set for model fitting.

**Class Distribution:** Checked the class distribution with labels and found the count of labels.

The label 4 has count 800000 and label 0 has 799999.

# Modelling:

**Bag-of-Words:** Used scikit-learn to construct Bag-of-Words (BOW).

**Performance:** Creating the metrics to check the performance of classification model build using sklear.

**AdaBoost Classifier:** AdaBoost Classifier is a technique used for Ensemble Method in Machine Learning. That has given the performance on train data as Accuracy = 0.66086, Recall = 0.8836, Precision = 0.611144, F1 = 0.722572. And given the performance on Test Data as Accuracy = 0.661221, Recall = 0.884107, Precision = 0.611808, F1 = 0.723174.

**Gradient Boosting:** Gradient Boosting is a technique of Machine Learning ensemble technique that combines the predictions of multiple weak learners. That has given the performance on train data as Accuracy = 0.668078, Recall = 0.870602, Precision = 0.619495, F1 = 0.723891. And given the performance on Test Data as Accuracy = 0.668183, Recall = 0.870442, Precision = 0.620047, F1 = 0.724212.

# Experimental Setup:

**Hardware and Software:** I used Google Collab to solve this problem.