# Project Proposal

CMPE 255 - 02

## Group 1 Members:

Sachin Pothukuchi - 015276989
Adam Romayor - 015447731
Supreeth Chandrasekhar - 015919566
Shreyas Kulkarni - 015966054

## Link to Dataset:

Reddit jokes dataset:
https://www.kaggle.com/pavellexyr/one-million-reddit-jokes?select=one-million-reddit-jokes.csv
CNN-Dailymail dataset:
https://cs.nyu.edu/~kcho/DMQA/

## About the dataset:

Reddit jokes dataset:
The Reddit jokes dataset contains jokes from Reddit posts, the joke in the text format and other metadata about the post itself.

CNN-Dailymail dataset:
The CNN / DailyMail Dataset is an English-language dataset containing just over 300k unique news articles as written by journalists at CNN and the Daily Mail.

## Project Description:

The goal of this project is to create a model that can detect whether a line of text is a joke or not. We will use the Reddit Jokes Dataset and the CNN-Dailymail dataset to train our model. We will label all data in the Reddit dataset as a “joke” and the data in the CNN-Dailymail dataset as “no joke.” Once the data is correctly labelled, we will combine both datasets into one large one. Various preprocessing techniques can be applied to the dataset before the model is trained.

Once the model is trained, a user can type their own input, and the model will determine if their sentence is a joke or not.

## Project Methodologies/Technologies:

Technologies: Python3, Sklearn, Matplotlib, Seaborn
Methodologies: TF-IDF, Sliding window (N-grams or C-mers) tokenization, KNN Clustering, Classification, Dimensionality Reduction