# Fake and Real news classification

Group Project proposal

Meghana Gantla

Kohisha Aruganti

DATS 6312 - NLP for Data science

Instructor - Amir Jafari

# Project Proposal

## Problem Statement

Fake news has become a major problem in the digital era, especially with the increasing use of social media platforms. The spread of fake news can cause significant harm to individuals, organizations, and even nations. Therefore, it is important to develop an accurate and reliable fake news classification model that can distinguish between real and fake news.

## Dataset

We will use the "Fake and real news dataset" available on Kaggle. This dataset contains 44898 news articles, of which 21417 are labeled as real news and 23481 as fake news. Each article has a 'title', 'text', 'subject', and 'date'. We combined true.csv and fake.csv into a single csv file and added ‘target’ (0 & 1) column to differentiate.
*https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset*

## NLP Methods

We will use various NLP methods such as text preprocessing, tokenization, stopword removal, and lemmatization. We will also use feature extraction techniques such as bag-of-words and TF-IDF to represent the text data. We will experiment with various classical machine learning models and try out more advanced language models down the line.

## Packages

We will use the following Python packages for our project:

- pandas and numpy for data preprocessing and manipulation
- nltk for text preprocessing and feature extraction
- scikit-learn for machine learning models and evaluation metrics
- matplotlib and seaborn for data visualization

## NLP Tasks

Our main NLP task will be text classification, where we will classify news articles as real or fake. We will also perform various text preprocessing tasks such as removing stopwords, tokenization, and lemmatization to prepare the text data for classification.

**Performance Evaluation**

We will evaluate the performance of our model using various evaluation metrics such as accuracy, precision, recall, and F1-score. We will also use a confusion matrix to visualize the performance of our model.

**Project Schedule**

Week 1: Data collection, preprocessing, exploratory data analysis, and visualization

Week 2: Feature extraction and model training

Week 3: Model evaluation, tuning, and final model selection

Week 4: Testing, documentation, and final report