**Proposal**

**Members:** Marcus, Ivana, Gianna

**Project title:** Fake News Detection using Machine Learning

**Motivation**

* Fake news has become an issue in today’s climate, especially with the prevalence of social media and easily shareable content. Fake news can often spread quickly and can be taken as fact when it is bias and/or opiniated. It is also sometimes difficult/time consuming to determine what is fake news or not. This is why we have decided explore what makes fake news, fake news and how to distinguish it from real news.

**Project Description**

* The goal of this project is to create a model using machine learning that can determine whether an article is fake news or not. To interact with this model, a web page will be created that will allow a user to enter a article title/paragraph/URL and have the model determine whether that specific article is fake news or real news. An analysis will also be conducted to see what the characteristics of a fake news article are and compare to the characteristic of real news article and help make fake news more identifiable.

**Research question to answer (specific)**

* What do you plan to show with your data analysis?
  + The characteristics of fake news and compare them to the characteristic of real news
  + How to spot fake news
* What research questions do you plan to answer? That is, what are your hypotheses?
  + Do fake news article titles have contain more words?
  + Do fake news articles contain more emotive words?
  + Do fake news articles contain more unprofessional words?
  + Can you determine what is fake news by looking at the source/author of an article?
* What do you expect to see as a result of your analysis?
  + Expect **higher** word counts in article titles in fake news articles
  + Expect **higher** counts of emotive words in fake news articles
  + Expect **higher** counts of unprofessional words in fake news articles

**­**

**Datasets required**

* What data sets will you access?
  + **Article dataset 1:** https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset
  + **Article dataset 2:** https://drive.google.com/file/d/1er9NJTLUA3qnRuyhfzuN0XUsoIC4a-\_q/view
* How will you access the data? (e.g. via an API or downloading the data)
  + Download CSV files from Kaggle
* What format will the data be in?
  + CSV files
* What problems might arise given your choices above?
  + Data may not have the columns we want, so we may need to create them ourselves using the data we have
  + May not be able to find the source or author of an article
  + Some rows may be repeated when we combine datasets
* Is your data sufficiently large?
  + Yes, one data set has 6000 rows
  + The other has over 30000
* Is your data messy or relatively clean?
  + Relatively clean