# Capstone III: Project Ideas

## 1. Predicting Fraudulent News Articles Using NLP

- The aim of this project is to detect fake contents using Natural Language Processing.
- The data for this project include:
    - **id**: unique id for a news article
    - **title**: the title of a news article
    - **author**: author of the news article
    - **text**: the text of the article; could be incomplete.
    - **label**: a label that marks the article as potentially unreliable.
        - 1: unreliable
        - 0: reliable

    that can be downloaded from:

    1) Source 1
    2) Source 2

Note: This might be a challenging project, but it helps with improving my NLP skills.

## 2. Credit Card Fraud Detection

- The aim of this project is to classify credit card transactions into normal and abnormal.
- The data can be downloaded from:
    1) Source 1

Note: This project is not as challenging as the previous one but might be more helpful for getting a job at a bank.

## Problem Statement

By examining historical news articles labeled as true or fake, can we create superior software that outperforms competitors in the market and attracts social media giants like Facebook and Twitter?

### 1. Context

A fact-checking organization is developing a fake news detection software to be sold to companies like Twitter and Facebook. For that, an analysis is to be conducted to answer the following questions:

I. Is it possible to differentiate between fake news and true news based on information such as the title, author, and content of the news?
II. How accurate the model will be?
III. What are the possible ways to improve the fake news detection model?

The answers to these questions will help the company to outperform its competitors and sell its software to a larger client base.

### 2. Criteria for Success

Ability to detect fake new with at least 70% accuracy.

### 3. Scope of Solution Space

I. Different preprocessing techniques will be used to gain first insights into the text features.
II. Different models will be developed and compared to find the one with the best level of accuracy.
III. The possibility of using a similar approach to conduct other tasks will be assessed.

### 4. Constraints within solution space

- Limited number of training samples.
- The generalization of the proposed models.

### 5. Stakeholders to provide key insight

- CFO
- CEO

### 6. Key data sources

The data for this project include news articles in a tabular format with the following features:

a. **id**: unique id for a news article
b. **title**: the title of a news article
c. **author**: author of the news article
d. **text**: the text of the article; could be incomplete.

e. **label**: a label that marks the article as potentially unreliable.