# Capstone Project 2
# "Sentiment analysis of Tweets"

## Proposal

- **The problem I want to solve**

I would like to perform a sentiment analysis on Tweets containing *Trump* and *Ukraine*.

- **My potential client and why he cares about this problem**

The Trump–Ukraine scandal is an ongoing political scandal in the United States. It revolves around efforts by U.S. President Donald Trump to coerce Ukraine and other foreign countries into providing damaging narratives about 2020 Democratic Party presidential candidate Joe Biden and about Russian interference in the 2016 United States elections.

Different types of people could be interested to know the proportion of positive vs negative tweets on this topic: politics, governments, individual people for or against Trump.

- **The data I will be using and how I will acquire it**

I will be using the Twitter API to collect a Test set based on keywords. A function will return a list of tweets that contain our keywords selected. Each tweet's text will see itself attributed a label ('positive' or 'negative') to classify each tweet as positive or negative. The Training set will be downloaded because it has to be labelled into 'positive' or 'negative' on a big amount of tweets. The Training set is critical to the success of the model since our model will "learn" how to do create a sentiment analysis based on the Training set.

- **How I will solve this problem**

The steps that we will follow to perform the sentiment analysis are:

- acquiring a Test set
- acquiring and preparing a Training set
- pre-processing tweets in both Test set and Training set (remove punctuation, tokenize)

- build a vocabulary/list of words in our training data set
- match tweet content against our vocabulary
- build our word feature vector
- training the classifier
- testing the model

- **My deliverables**

A jupyter notebook containing the code, a slide deck , and a final report in PDF.