# Fake News Detection with Word2Vec and BERT

**Description**:

In this assignment we will implement a machine learning model to decide whether a given news article is trustworthy or not trustworthy. There are two main approaches to do so: One is the powerful Word2Vec algorithm which transforms words into semantic word vectors. The word vectors are able catch a word's meaning which simple bag-of-words (bow) strategies do not. A problem of Word2Vec, however, is that it does not look at a word's context in a sentence. BERT might thus perform better as it, too, can generate word vectors.

Both approaches are based on neural networks (i.e., deep learning). Whereas Word2Vec needs an additional machine learning algorithm like SVMs, BERT can be directly trained via a procedure called "finetuning".

**Data:**

There are two datasets:

1. train.csv
2. test.csv

We will use the train dataset to explore the data and develop our machine learning algorithm and then validating it with the test.csv at the end. We will not touch that dataset to evaluate our progress.

**Assignment:**

1. Learn about neural networks, starting with the perceptron
2. Learn about deep neural networks
3. Learn about Transformers and Self-Attention (http://jalammar.github.io/illustrated-transformer/)
4. Explore the data and get a basic understanding
5. Write a simple tokenizer or use a prebuild one
6. Download and install the English models
    a. Word2Vec (https://code.google.com/archive/p/word2vec/)
    b. BERT (https://www.youtube.com/watch?v=Hnvb9b7a_Ps, https://colab.research.google.com/drive/1pTuQhug6Dhl9XalKB0zUGf4FIdYFlpcX)
7. Extract meaningful sequence vectors from the news data for word2vec (https://arxiv.org/abs/1405.4053)
8. Train a machine learning model on the vector given by step 4
9. Finetune an BERT-Model (see step 3) and evaluate it on a validation dataset