

# Detecting Fake News

The world that we live in is filled with news. From Facebook and Instagram to different news agencies like BBC or Fox News we are always being fed with information. This information can be related to anything, from politics to Economy or health and entertainment, but how can we tell if the information that we are getting is real or fake? Fake news is a very big problem that world is facing currently. Facebook, Instagram, and other multimedia platforms have been under pressure to detect fake news and block the content of it. This is the main reason that I picked my project about fake news detection.

## DATA:

In order to start my project I took my data from Kaggle: <https://www.kaggle.com/c/fake-news/data>. The dataset contained about 40000 fake and real news articles divided equally.

## DATA PREPROCESSING:

The data consisted of 4 columns being the title, news article, date of publication, and if the news is real or fake. Looking at the data it became evident to me that all the real news articles started with name of the news agency and then continued with the news. My first step was to remove the agency name from the articles and put them in a new column called news agency.

My second step was to perform text cleaning and tokenization. This step was done by creating a function that lower cased all the words, removed stop words, removed punctuations, tokenized sentences, and transformed each word to its stem using the NLTK library.

## MODEL TRAINING:

After the data was cleaned, the dataset was divided into test and training sets using the Scikit-Learn library. Then an LSTM model was created with Keras including an embedding layer at the first layer and a dropout layer after the LSTM layer. The 20000 of the most used words were turned into sequences and the length of the sequence for the news articles was around 8500 words and for the articles title it was 29.

For the news article the training was done for 2 epochs and for the news title model the training was done for 10 epochs. The accuracy of prediction was 98 percent for news article model and for the title model it was 96 percent.

## MODEL PREDICTION:

By combining the two models of news article and news title the total performance of the models on test data was 99% with an F1 score of 99% which is amazing!

## **MODEL Deployment:**

The Streamlit library was used to create an API for news detection and after creation of the api it was deployed on AWS using docker.