**COMSATS University   
Park Road, Chak Shahzad, Islamabad Pakistan**

**Project Proposal**

**(Scope Document)**

**for**

**Fake News Detector**

***By***

**Waleed Butt CU/SP18-BCS-170/ISB**

**Shariq Ahmed CU/SP18-BCS-151/ISB**

**Umar Khalid CU/SP18-BCS-164/ISB**

***Course Instructor*Mr. Usman Yaseen**

Class/Section: BCS-6B

Submission Date: 10/10/2020

*Bachelor of Science in Computer Sc­ience (2018-2022)*

# Problem statement

Whenever there are elections in the country, or a famous personality being accused of a crime, lots of fake news start circulating. This has been happening through out the history of mankind but, due to social media and technology this has reached its peak. If we want to counter this problem we are going to need technology to do so.

So how will you detect the fake news? The answer is Python. By practicing this advanced python project of detecting fake news, you will easily make a difference between real and fake news.

# Related work

Kaggle in 2016, released a dataset of fake news from 13000 articles during elections cycle. Later they grabbed 5279 real news articles through web scrapping. This dataset has been widely used to train models that would detect the fake news and real news.

Several projects were done to perform this task. Sci-kit Learn’s GridSearch (Solutions, 2019) is a one of the famous functionality to execute this task with 91-92% of success rate. Also data-flair.training (data-flair, 2020) has a blog that goes through step by step teaching the way of creating such software.

# Methodology

## What is Fake News

Before we get into the ways of detecting fake news, we need to know what the fake news are. Fake news encapsulates pieces of news that may be hoaxes and is generally spread through social media and other online media. This is often done to further or impose certain ideas and is often achieved with political agendas. Such news items may contain false and/or exaggerated claims, and may end up being viralized by algorithms, and users may end up in a filter bubble.

## TfidfVectorizer

TF (Term Frequency): The number of times a word appears in a document is its Term Frequency. A higher value means a term appears more often than others, and so, the document is a good match when the term is part of the search terms.

IDF (Inverse Document Frequency): Words that occur many times a document, but also occur many times in many others, may be irrelevant. IDF is a measure of how significant a term is in the entire corpus.

The TfidfVectorizer converts a collection of raw documents into a matrix of TF-IDF features.

## PassiveAggressiveClassifier

Passive Aggressive algorithms are online learning algorithms. Such an algorithm remains passive for a correct classification outcome, and turns aggressive in the event of a miscalculation, updating and adjusting. Unlike most other algorithms, it does not converge. Its purpose is to make updates that correct the loss, causing very little change in the norm of the weight vector.

## Steps

Using sklearn, we build a TfidfVectorizer on our dataset. Then, we initialize a PassiveAggressive Classifier and fit the model. In the end, the accuracy score and the confusion matrix tell us how well our model fares.

# References

data-flair. (2020). *Advanced Python Project - Detecting Fake News*. Retrieved from data-flair.training: https://data-flair.training/blogs/advanced-python-project-detecting-fake-news

Solutions, P. T. (2019, November 22). *Fake News Detection Using Python*. Retrieved from pantechsolutions.com: https://www.pantechsolutions.net/fake-news-detection-using-machine-learning