# Capstone Project 2 : MileStone Report

## Project Title :

"Fake News Classifier"

Project Mentor : Vikas

### *What is the Problem to be solved?*

In Daily routine we will read a number of articles about trends , politics , technology,food etc., but we just read and will come to our opinion and we will not think whether that article is Fake or Real

### *Who is this project beneficial for?*

This project is used for all of us who have a habit of reading news articles. So this project will help us to identify a given article is Fake or Real

### So *What is the data used for the project? And where I received the data from?*

So the data consists of all News records about different topics in the world and I have extracted the data from Kaggle website. This is a multivariate data consisting of Id, title, author, text,label, etc.

## *Techniques that I would use to solve the problem:*

1) Data Wrangling - I will observe the data for all the missing values, incorrect entries and based on the observation I will clean the data for further use and apply the Modern NLP techniques to transform text data into vector form

2) Exploratory Data Analysis - I would plot the relation between different parameters in the dataset and based on that I can find the trends for storytelling.

3) Machine Learning Model / Deep Learning - After all the initial work on the dataset it would be time to build a Machine Learning Model to find the given article is Fake or Real i may use the LSTM or a simple Naive Bayes.

## *What is the methods I will use to deliver the project?*

I will be using the following tools to work on the project :-

- Jupyter Notebook for the major part of the project. - I will be using Python version 3.x for writing the code. - I will use Google doc for submission on the in process project reports. - Powerpoint for Data Storytelling presentation.

## *References:*

- Springboard Curriculum material - DataCamp courses - Google, Stackoverflow, Github, etc. - And any other Open Resources.

# Data Wrangling Report:

## Data Source:

This dataset is downloaded from the Kaggle website which holds many competitions for the Data science learners with real-world problem statements and data sets.

## Initial Data Observations:

The data is split into two datasets: train and test. The train dataset has a shape of (18285,5) and the test dataset is of the shape (2671,4) with label column dropped. The features of the train dataset are [‘id’ ,’title’ , ‘author’, ‘text’ ,’label’']. The column types of the dataset are as follows:

**Column Name Type**

id int64

title object

author object

text object

label int64

## Data Wrangling process:

## Missing Value Treatment:

I can see multiple missing values for title , author , text since they are text fields so I dropper all the missing values because we can’t impute like for numeric type with mean , median etc.,

# Outlier Treatment:

Here we can't say for the text fields whether they are outliers or not . Outliers can only identified for numeric columns so I Ignored Outlier treatment .

Now the dataset is mostly clean and ready for further process of Exploratory Data Analysis, Data Visualization and Machine Learning modelling.

**Data Processing :**

Now all the text data should be handled . I am considering the text or title field for identifying given text is Fake or Real because it will not dependent on the type of Id or author based on the content we can say it is Fake or Real .

Here for that I defined a Corpus with all the text field data then I did the following steps :

- Remove all the stopwords because they are no use for our prediction and they will decrease our accuracy of model if we are not removed
- Bring the root words for each word by using Stemming or lemmatizing
- Now convert all the words into lower case and split the sentence and append to the list
- Now Each sentence is list of words with no stop words and with root words now we have to do count vectorizer or tf-idf vector to convert words into vectors become model can't take text fields .Count vectorizer is Bag-of-model technique and Tf-idf model will assign weights to words based on their importance if the word is repeated in all the sentences then weight will be less and if word is unique to that sentence then it will have more weight