Fake news is a major concern in our society right now. It has gone hand-in-hand with the rise of the data-driven era – not a coincidence when you consider the sheer volume of data we are generating every second!

Fake news for various commercial and political purposes has been appearing in large numbers and is widespread in the online world in recent years as a result of the booming development of online social networks. Online social network users can easily become infected by online fake news by using deceptive words, which has already had a huge impact on offline society.

Examples of Fake News:

* Muzzafarnagar riots of 2013: fake video fuelled communal passions
* UNESCO has declared ‘Jana Gana Mana’ best national anthem in the world (WhatsApp)
* President Kovind makes Twitter debut; gains 3 million followers in one hour (Republic, Zee news, TOI etc.)
* GPS tracking nanochip in 2000 Rupee notes (Nov 2016)
* Ministry of Home Affairs (MHA) annual report used a picture of the Spain-Morocco border to show Indian border floodlighting Missing JNU student Najeeb Ahmed has joined the ISIS

Using Natural Language Processing, we will be predicting whether a news headline is genuine or fake.

**Objective:**

1. Our objective is to classify the news from the dataset to fake or true news.
2. Extensive EDA of news
3. Selecting and building a powerful model for news classification

**About the data set:**

There are two datasets one for fake news and one for true news. In true news, there is 21417 news, while in fake news, there are 23481 data points. Both datasets have a label column in which 1 for fake news and 0 for true news. We are combined both datasets using pandas built-in function.

In this Dataset there are no missing values otherwise we have to remove that information or we have to impute some value. In the final dataset, we got a balanced dataset as both categories (fake news and real news) have the approximate same no. of examples.

**Libraries Used:**

* Numpy
* Pandas
* Gensim
* NLTK
* Matplotlib
* Seaborn
* Sklearn
* Tensorflow

**Algorithms Used:**

* Natural Language Processing or NLP is a field of Artificial Intelligence that gives machines the ability to read, understand and derive meaning from human languages.

There are two main phases to natural language processing: data preprocessing and algorithm development.

Data preprocessing involves preparing and "cleaning" text data for machines to be able to analyze it, preprocessing puts data in a workable form and highlights features in the text that an algorithm can work with.

Once the data has been preprocessed, an algorithm is developed to process it. There are many different natural language processing algorithms, but two main types are commonly used are Rules-based systems and Machine learning-based systems.

* **LSTM:** Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. This is a behaviour required in complex problem domains like machine translation, speech recognition, and more.

**Insights from Exploratory Data Analysis:**

* Political News and World News hold the most domination counts in the data set that we have considered
* The maximum number of titles range between 7-8 words. Would be a bit tricky to judge if the news is true or fake only with these few words. So we hope we shall not be getting a great amount of accuracy using the title alone
* News articles are within a range of 100-200 words. Anything beyond this point is not very catchy, and even the media tends to avoid it. But yes when the number of words is large, it will be much easier for the model to predict.

**Model Evaluation:**

* Model accuracy is around 99 percent using the LSTM algorithm
* Model precision comes to be around 0.98

**Conclusion:**

* We have been successfully able to reach the target for classifying a News- and are able to identify if this is a True or Fake
* Its not much feasible to predict the news just from its title. Titles may be misleading.
* A much better Idea was obtained when we considered the context for the News. The accuracy percentages shot up
* Considering both the title and the context, we have been able to reach the highest percentages of accuracy