Machine Learning
Fake News Detection
Team Project

Team 21 :
Antoine ALEXANDRE 24170169
Valentin GIROLD 24170128

Problem definition

The spread of fake news is a critical issue in today's digital age, affecting public opinion and decision-making. For this reason, the problem addressed in our project is the **automatic detection of fake news**. The primary objective is to develop a Machine Learning model capable of classifying a given news article as either "fake" or "true" based on its textual content. With the widespread dissemination of information online, an automated solution to detect fake news could be valuable for platforms, researchers, and users to assess the credibility of news articles.

Dataset explanation

The **Fake News Detection**[1] dataset on Kaggle is structured into two separate .csv files, each representing a class of news articles: one for fake news (False.csv) and one for real news (True.csv). Extracts from each of these files are attached below :

True.csv (53.58 MB)

Detail | Compact | Column

title	text	subject	date
20826 unique values	21192 unique values	politicsNews 53% worldnews 47%	2016-01-13 2017-12-31
As U.S. budget fight looms, Republicans flip their fiscal script	WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, who voted...	politicsNews	December 31, 2017
U.S. military to accept transgender recruits on Monday: Pentagon	WASHINGTON (Reuters) - Transgender people will be allowed for the first time to enlist in the U.S. m...	politicsNews	December 29, 2017

Fake.csv (62.79 MB)

Detail | Compact | Column

title	text	subject	date
17903 unique values	[empty] 3% AP News The regu... 0% Other (22851) 97%	News 39% politics 29% Other (7590) 32%	2015-03-31 2018-02-19
Donald Trump Sends Out Embarrassing New Year's Eve Message; This is Disturbing	Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he had...	News	December 31, 2017
Drunk Bragging Trump Staffer Started Russian Collusion Investigation	House Intelligence Committee Chairman Devin Nunes is going to have a bad day. He s been under the as...	News	December 31, 2017

[1] Fake News Detection dataset on Kaggle : https://www.kaggle.com/datasets/jainpooja/fake-news-detection

Each .csv file includes similar columns, which help differentiate the articles based on their content, theme, and publication date. This structure allows for easy labeling and classification, as each file represents a single class:

- **title:** This column contains the headline or title of each article. Titles provide concise information and often encapsulate the main point or angle of the article, which can be helpful for classification models that rely on shorter text inputs.
- **text:** The main body of the article, typically including paragraphs of content. This field is the richest in terms of textual information and allows the model to analyze deeper language patterns and contextual clues to differentiate between fake and real news.
- **subject:** The topic or general category of the article, such as politics, world news, or technology. Subjects provide context, as fake and real news articles may target different subjects with varying styles and rhetoric.
- **date:** The publication date of the article. Although this may not directly impact the content, some studies suggest that fake news stories may peak around certain events or political cycles (e.g. presidential campaign), so this column might hold potential value if temporal patterns are relevant.

Expected results

The primary goal of this project is to develop a reliable model that can accurately classify, in a binary way, news articles as either "Fake" or "Real." We expect the model to perform well enough to be a practical tool in identifying misinformation, with a high level of accuracy and low rate of misclassification.

Expected Performance Metrics:

- Accuracy: The model should ideally achieve an accuracy of at least 90% to be practically useful.
- Recall and F1-score: High recall is crucial to minimize cases where fake news is mistakenly classified as real. An average F1-score of at least 0.85 is targeted.