**Data Stream - Homework 2**

**Group Members**

* Victoria
* Janka Gordos
* Princess Mavuru
* Nicola Sagay
* S R (Rose) ?

**TASK**

1. Your group needs to decide what kind of project you are going to work on and lock in your decision.

2. You need to submit a free style paper that describes your project on a high level.

Please cover the following questions:

1. What kind of data research and analysis are you going to take on?

2. What industry or areas does it cover?

3. What questions are you planning to answer?

4. What data sources are you planning to use?

5. Describe the team approach to the project work: how are you planning to distribute the workload, how are you managing your code, how are you planning to work on your project?

**SOLUTION**

**Project ‘Fake News detection’**

**1.** Check where and what useful data is available (open to the public).

Check datasets on Kaggle for the 'Fake News detection' project. Download dataset using Kaggle API.

Check other sources and ways to extract the data.

Check if the data is labelled and how can we merge different datasets (on which features) to extend the dataset and lower variability for new data samples.

Explore and clean data using Pandas. As we are dealing with text data rather than numbers, we will clean the text of any punctuation and common English words (eg.: 'a', 'the', 'he', 'she'…) and cut the number of features to some reasonable amount of words based on their frequency, probably using Sklearn inbuild methods.

Explore the data visually using matplotlib or other suitable libraries.

Decide what classification algorithm to use to predict if the news is fake.

How could we validate and measure the predictive outcomes.

**2.** Fake News recognition is one of the most challenging cybersecurity problems of our time. We have never been able to access information so easily before. (You can read the latest news on online news outlets, blogs, vlogs or articles shared on social media platforms.) On one hand, the social media platforms such as Facebook, Twitter, Youtube or Instagram create a space where the users can discuss and share their ideas about basically anything: from politics and education to climate change concerns. On the other hand, these sites were built to run by algorithms: recommendation and search engines, autocompletes or trending topics that are prioritised and pushed to the users. However, these algorithms cannot distinguish between fake or real news, which content is propaganda or which content is fact-checked. Their only job is to share new content related to the users, which they do remarkably well. Unfortunately, these algorithms can be used for creating biased opinions, polarisation, spreading absurdities, radicalising and manipulating the thoughts of the users. The widespread use of fake news can cause serious social and political damage. The United Nation reported about the role of Facebook that played a role of fueling a genocide in Myanmar as follows: “The role of social media is significant. Facebook has been a useful instrument for those seeking to spread hate, in a context where for most users Facebook is the Internet… The extent to which Facebook posts and messages have led to real-world discrimination and violence must be independently and thoroughly examined.’ Or more recently, the spread of fake news related to COVID-19 has put global health at risk, of which the WHO released a warning explaining the effects of misinformation related to the pandemic. Not surprisingly, more and more research focuses on the necessity to identify the impact of fake news, to come up with solutions, and to reduce the spread of fake news. Our aim is to make an analysis on the detection of fake news and to explore different machine learning models to find the most effective one.

**3.** Find the patterns or words used in fake news to distinguish them from real articles.

Our aim is to make an analysis on the detection of fake news and to explore different machine learning models to find the most effective one. We aim to create a model that can differentiate between fake or real facts within an article based on its words, phrases or titles by applying supervised machine learning on a pre-labeled dataset. The dataset we use were manually classified using fact-checking sites.

**4.** ISOT Fake News Dataset -found on Kaggle.

The dataset we use has been manually classified using fact-checking sites into two categories - fake and real news. It contains different types of articles but mainly focuses on political news.

<https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset/metadata>

**5.** The plan is to use the Agile Development methodology:

Create a Backlog as a living document that contains tasks that have to be done to get the final product, priority to choose tasks for each sprint, estimated deadline, and result.

Weekly Sprint document that contains tasks, assigned team member(s) to the task, reviewed by whom, deadline, result, and schedule for meetings in this sprint.

We have suggested roles for individual group members to choose from. They will have the overarching responsibility of their role, though we will ultimately work together to create a cohesive piece and to complete the tasks for the role.

Work will be carried out both individually and as a group through Zoom calls and ‘Code with Me Sessions’ in between sprints.

GitHub repository for project code flow tracking. Include: ‘data’ folder for all datasets (row\_data\_1,..., merged\_df, cleaned\_df, preprocessed\_df, …), ‘src’ folder for .ipynb files, ‘pickled’ folder to store pickled objects such as trained model or other, README.md for project description, PPT final presentation.