Data Collection Report

Prepared by: Alexandra Leung Dat Wan (2410909), Bhoovana Dinoo (2410920), Kentish Thummanah (2411831), Luvish Kumar Lokye (2410762), Nahla Dinmahamed (2414467)

1. **How did you obtain your data, if it's a Data-driven problem?**

Our fake news detection AI model is a data-driven problem. We sourced our data from a dataset published by author **Mahdi Mashayekhi** on Kaggle. Below is our D-SOAKED problem formulation:

→ D – Decision: refers to the decision whether the data is real or fake
→ S – States: we will have several states namely, initial state, tokenization state, vector state, model internal state, decision state, and finally contextual state.
→ O – Observations: the observation will be our original data.
→ A – Action: our action will be what we do to move from one state to the next.
→ K – Knowledge: refers to the prior knowledge, what the machine learns.
→ E – Environment: refers to the whole dataset from Kaggle.
→ D – Desirability: refers to the cost.

2. **How large is your data?**

Our data consists of 20,000 volumes of unique data that were registered in the datasets. It will be stored in a tabular format, and the board of configuration will consist of 7 columns for each sample of data.

3. **In what format are you storing your data or states? Describe the abstract data type, not just the file format.**

The datasets will be stored in csv format (excel spreadsheet); 7 columns with 20,000 rows. Each row will represent an instance, and each column will represent a feature/attribute. The 7 attributes are title, text, date, source, author, category and label.

4. **Did you need to process the original data to get it into an easier, more compressed format (e.g., convert from one format to another one)?**

Data in the real world is messy. We need to clean the original data, that is, remove duplicates, handle missing values and fix inconsistencies. The original data will be tokenized then converted into a vector. From this, the weight of the data will be calculated, using this as a reference to decide if the news is real or fake.

5. **How would you simulate similar data?**

To stimulate similar data, we will test the ai with new real data (articles found on twitter/facebook).