# 7PAM2002-0901-2024 - Data Science Project

Student Name: Kamali Bakthavatchalam

Mail id: kb23aat@herts.ac.uk

Student id: 22081557

**FAKE NEWS DETECTION**

# Project Overview

**Project Title:**

Detection of Fake News: Analysing and Classifying News Articles

**Summary:**

The rapid spread of misinformation and fake news has significant implications for public perception and democracy. This project aims to utilize machine learning techniques to identify and classify news articles as real or fake, thereby contributing to more informed media consumption. The project leverages the Fake News Dataset from Kaggle, which contains labelled news articles for training and testing classification algorithms.

**Research Question:**

How effective are machine learning algorithms in distinguishing between real and fake news articles using textual features?

**Project Objectives:**

 **Data Preparation:** Clean and analyse the Fake News Dataset to make it ready for machine learning processes.

 **Algorithm Implementation:** Apply different machine learning algorithms, such as support vector machines (SVM) or random forests, to classify news as real or fake.

 **Performance Evaluation:** Measure the accuracy, precision, and recall of these models to determine their effectiveness.

 **Feature Analysis:** Identify which textual features (like words, phrases) contribute the most to distinguishing between fake and real news.

**Reference List:**

* Barzilai-Nahon,K. (2018). The Politics of Fake News: The Media’s Role in the Production of Fake News. Journal of Media Ethics, 33(1), 45-58[(https://www.tandfonline.com/toc/hmme20/current)](https://d.docs.live.net/72586ECF77183282/Documents/(https:/www.tandfonline.com/toc/hmme20/current))
* Shu, K., Sliva, A., Wang, S., Jiang, J., & Tang, J. (2017). Fake News Detection on social media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36. [(https://dl.acm.org/journal/sigkdd)](https://d.docs.live.net/72586ECF77183282/Documents/(https:/dl.acm.org/journal/sigkdd))
* Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151. Here are the hyperlinks for the references you asked for: [(https://www.science.org/journal/science)](https://d.docs.live.net/72586ECF77183282/Documents/(https:/www.science.org/journal/science))

# Project Timeline

|  |  |  |  |
| --- | --- | --- | --- |
| **Task** | **Description** | **Start Date** | **End Date** |
| Literature Review | Review existing research on fake news detection methodologies. | Sept 25, 2024 | Oct 1, 2024 |
| Data Pre-processing | Clean and pre-process the Fake News Dataset for analysis. | Oct 2,  2024 | Oct 8, 2024 |
| Feature Engineering | Extract relevant features from the dataset for model training. | Oct 9,  2024 | Oct 15, 2024 |
| Model Implementation | Implement machine learning algorithms (e.g., SVM, Random Forest). | Oct 16, 2024, | Oct 29, 2024 |
| Model Evaluation | Evaluate model performance using accuracy, precision, and recall. | Oct 30, 2024, | Nov 5, 2024 |
| Final Report and Presentation Prep | Compile results and prepare presentation materials. | Nov 6,  2024 | Nov19,2024 |
| Review and Revisions | Make necessary revisions to report and presentation based on feedback. | Nov20,  2024 | Nov 26, 2024 |
| Final Presentation Prep | Finalize presentation materials and practice delivery. | Nov27, 2024 | Dec 1, 2024 |

# Data Management Plan

**Overview of the Dataset:**

* **Source:** Kaggle Fake News Dataset.
* **Description:** Contains labelled news articles (fake vs. real) with textual content, including headlines and body text.
* **Size:** Approximately 20,000 records; expected size of the dataset is around 1 GB.

**Data Collection:**

* The dataset can be accessed at Kaggle Fake News Dataset – [Link](https://www.kaggle.com/code/therealsampat/fake-news-detection/input)

**Metadata:**

* **Format:** CSV files.
* **Size:** Approx. 1 GB.

**Document Control:**

* **GitHub Repository:**[link](https://github.com/kamalibakthavatchalam/pdm-plan)
* **Commit Frequency:** Weekly commits will be made to track code development.
* **File Naming Convention:** Descriptive names like fake-news-preprocessing.py, model-result.py, etc.

**ReadMe File:**

The ReadMe file will include:

* Overview of the project.
* Instructions for dataset setup and code execution.
* Description of the algorithms used and how to replicate the results.

**Security and Storage:**

* **Backup Frequency:** Data and code will be backed up weekly.

* **Storage Locations:** Data will be stored on GitHub and OneDrive to ensure security and accessibility.

**Ethical Requirements**

1. **GDPR Compliance:** The dataset does not contain personal data; thus, it does not fall under GDPR.
2. **UH Ethical Policies:** The project conforms to university ethical guidelines regarding research involving data.
3. **Permission for Data Use:** The dataset is publicly available on Kaggle, allowing for academic use.
4. **Data Collection Ethics:** The data was collected and labelled by credible sources, ensuring ethical standards were upheld during its compilation.