Task 1: Fake News Detection

Sharlene Moetjie – Data Science Intern

Objective:

- The objective of the fake news detection task is to develop a model that is capable of accurately distinguishing between fake and real news articles. This task involves preprocessing textual data, training machine learning models and evaluating their performance to ensure they can effectively identify false information.

1.Project Selection

- For this task, the aim is to develop a fake news detection model that can accurately classify textual data into predefined label categories (FAKE or REAL). In this task, I aim to classify news articles as either FAKE or REAL based on textual content.

2.Data Selection

- For this detection model, I utilized two datasets of fake and real news articles, from Kaggle, that are majorly focused on political and world news topics. Each dataset contains more than 12 600 articles from varying sources.

3.Preprocessing

- Cleaning the data texts by removing HTML tags,non-alphabetic characters, URLs, numbers and punctuation then converting them to lowercase.

4. Model Development

- I used a systematic process approach that included: data preprocessing, feature extraction, model training and evaluation. The implementations are as follows:
  - **Data Collection:** Imported the datasets, combined them into a single Data Frame and shuffled them to ensure randomness.
  - **Data Preprocessing:** Defined a preprocessing function to clean the text data by converting the texts to lowercase, removing non-alphabetic characters, URLs, punctuation, HTML tags, digits, etc.
  - **Feature Extraction:** Used the TfidfVectorizer to convert and vectorize the text data into numerical features that can be used by the machine learning model.
  - **Train_Test_Split:** To split the combined dataset into training and testing sets in order to evaluate the model's performance on unseen data.
  - **Model Training:** Train the classifier models on the vectorized data
  - **Model Evaluation:** Evaluate the model's performance using the testing data.
  - **Prediction Function:** Define a function to preprocess new article text inputs, vectorize it and make predictions using the trained model.

5.Training and Evaluation

- The training and evaluation phase includes splitting the data, training the model and evaluating its performance using appropriate metrics such as accuracy, precision, recall and F1-score.

Results Presentation

CLASSIFICATION REPORT:

	precision	recall	f1-score	support
FAKE	0.99	0.98	0.98	4678
REAL	0.98	0.98	0.98	4302
accuracy			0.98	8980
macro avg	0.98	0.98	0.98	8980
weighted avg	0.98	0.98	0.98	8980

Summary

- The results indicate that the model performs well, with significant accuracy in distinguishing between FAKE and REAL news.

References

- Lutz Hamel. Fake News Detection in Python. [Video]. YouTube. https://www.youtube.com/watch?v=ZE2DANLfBIs&t=746s
- Simplilearn. Fake News Detection Using Machine Learning | Machine Learning Projects In Python | Simplilearn. [Video]. YouTube. https://www.youtube.com/watch?v=U6ieiJAhXQ4