Data selection proposal

Data Set:

https://www.uvic.ca/engineering/ece/isot/datasets/fake-news/index.php

The selected data set is the largest verified fake news data set online. It clearly indicates the classification of true or false, combined with a relatively large dataset size, making it fitting for this specific project

Methodology:

1. Preprocessing:
    a. Note: Numerous instances of the data set are provided by the provider “Reuters” therefore this word needs to be removed from the dataset to avoid unnecessary confusion.
    b. 80% of the dataset will be used to train our machine learning model, 10% will be used for validation, 10% will be used for testing.
    c. Text Lemmatization to improve language consistency :
        i. “Feet” and “foot” are two different words but has the same meanings
        ii. Named Entity recognition to make sure names are recognised as it is

    d. Feature extraction: (Potential methods) :
        i. Bags of words:
        ii. Word2Vec
        iii. Glove
        iv. Sentiment detection using VADER
        v. Note the same method can be repeated on the headline for extra features, although if the input is an article only, we’ll have to implement text extraction to make up new headers, and this could cause inconsistencies.
2. Machine learning algorithm:
    a. Naive Bayes due to its simplicity and low computation requirement.
    b. Alternatively, use deep learning (due to the limit dataset size, this may not be possible) or Support vector machine
3. Matrix of evaluation:
    a. Confusion matrix. Because our end goal is a binary classification, a confusion matrix is quite convenient.
4. Conceptualisation: A webapp that takes input of String, and outputs the prediction as “Real news” or “Fake news”.