Task 1 Project Report

* Introduction
  + The core objective is to carry out an investigative analysis observing the performance of various machine learning algorithms, concerning fake news detection.
  + To further investigate performance levels by changing certain hyperparameters and documenting the obtained results.
  + The dataset being used for the purpose of this research is the LIAR dataset, a political dataset made up of roughly between 11k to 12k articles vetted by Politifact.com. One of the objectives is to argue whether it is possible to identify a particular algorithm that performs better against political fake news.
* Completed work
  + As of this moment I have made a good draft for the literature review as it is not finalised and finished the related studies and dataset section in the methodology.
  + Additionally, started working on the prototype by loading, inspecting, and cleaning the dataset aswell as evaluating what classifiers are to be used (from the ones mentioned in the literature review).
  + The literature review has been structured as follows:
    - Introduction
      * Although fake news has been around for a long time throughout history, the introduction talks about the origination of the term ‘fake news’, when people labelled it as such. It also makes mention of a groundbreaking discovery regarding the 2016 US presidential election and the fake news being generated around it.
    - The Dangers of False Information
      * In this section I mention situations that happened or could possibly happen as a repercussion of fake news. I used a number of sources both academic and from news websites detailing certain tragedies which were a direct product of said repercussions. Additionally, I mentioned aswell the motive and what the people who create fake news stand to gain (Murayama et al. 2021).
    - Current Mitigation Techniques
      * This section talks about what the technological industry is doing to help mitigate the spread of fake news. It mentions what steps Google, FaceBook, Twitter and the European Commission are taking into closing and deleting fake accounts and bots, aswell as drafting policy initiatives (Corbu et al. 2020). In conjunction more organisations such as the FTC are also taking steps in combating misinformation and disinformation by shutting down a number of fake news websites (Levi 2017).
    - Academic Mitigation Techniques
      * This is where the bulk of the research towards my paper is found. I talk about what is the previous and current research being carried out, famous datasets being used such as the LIAR, what effective algorithms are being utilised such as classifiers and feature extractors, various data gathering techniques such as the Amazon Mechanical Turk (AMT) (P ́erez-Rosas et al. 2017) and showing the major results found in my research. This section helps setting up the context and tone of my research paper, by indicating what kind of classifiers, feature extractors and dataset will I be using.
    - Classifiers and feature extractors
      * In this section I talk about the various classifiers and feature extractors which were taken into consideration to be used in this paper, although the final product might not incorporate every single one that is mentioned. I give in depth explanations on how these algorithms work and function.
    - Chapter Conclusion
      * In the conclusion, I outline the best performing models as viewed in the literature as I make mention of what the methodology will entail.
  + The Methodology has been structured as follows
    - Introduction
      * The introduction will give a brief overview of how the methodology is structured and what kind of information is covered in this research paper.
    - Related studies
      * This section is an extension of the research discussed in the literature review. While the research in the literature review covered what the authors in those papers used for datasets and models and the major results achieved, this section of the methodology covers the application side of things. Sometimes referring directly to papers mentioned in the literature review and other times it makes mention of fresh papers found academically. This section also covers in the detail the results and discussions of the related studies found to further support the means of my research.
    - Dataset
      * The dataset section explains what datasets were taken into consideration and why. A total of three datasets were being considered, RealNews, FakeNewsNet and LIAR. Eventually selecting LIAR as I found it to be the most suited for this research due to its high article count and is one of the most favoured according to a number of academic papers. Having chosen the dataset I then proceeded to load and inspect it to figure out how it is sectioned depending on the article label, documenting my findings as I go along.
* Work to be done
  + Abstract
    - The abstract will be a summary of my research and it contents. The person reading my paper will get the idea of what my research is about and what it managed to achieve. It will contain brief information about the models and dataset used, aswell as the yielding results.
  + Introduction
    - The introduction will be the opening of my research, presenting the hypothesis and research questions. It will also give a detailed background of my research topic and the problems fake news is presenting, providing aswell an in-depth description of what my research hopes to achieve.
  + Literature Review
    - Since the literature review is still a draft, it is yet to be finished. In the Classifiers and Feature Extractors section an explanation for the Long Short-Term Memory (LSTM) and Decision Trees (DT) are yet to be added since they are being taken into consideration.
  + Methodology
    - Describing the nature of my research by mentioning Deductive vs Inductive and Quantitative vs Qualitative and then stating which category my paper falls under and why.
    - Explaining the prototype implementation process by discussing what classifiers and feature extractors are to be used.
    - Answer the hypothesis and research questions through the results achieved.
  + Prototype
    - Clean and preprocess the dataset
    - Construct the models with the chosen classifiers and feature extractors
    - Run multiple experiments of the constructed model and change hyperparameters along the way, documenting any results achieved per experiment.
  + Results and discussion
    - Compare and contrast all of the obtained results from the experiments. Outlining the best performing ones while criticizing the least performing models aswell as mentioning if possible, the reason as to why they might underperforming.
  + Conclusion
    - In the conclusion I will finally mention any future improvements that could be done and stating what limitations held back my research.
* Conclusion
  + Work on the dissertation is progressing steadily. Some minor problems are to be expected when work on the preprocessing of the dataset starts due to LIAR having unconventional labelling, therefore some extensive research on this matter will need to be carried out. Once the prototype is finalised the results will help resolve the hypothesis and research questions.

Task 2

My mind map wasn’t an easy one to make and this is partly due to the structure of my dissertation. One major limitation I found myself in was the creation of relationships. This could be because most of the research I found doesn’t overlap that much as it sprawls out covering multiple subjects within my research topic. Therefore, finding similarities between one paper and another proved to be more difficult. Another problem my mind map presents is the hierarchical format it is presented in. Although this is quite effective and makes the map quite readable, when there are a lot of nodes branching out it can get quite cluttered and congested.

Task 3

* Study 1
  + The purpose of the research carried out by Wang. 2017 presents an experimental study, quantitative in nature, with the purpose to present and test a new dataset called LIAR. The study’s aim is to investigate automatic fake news detection based on surface-level linguistic patters using a hybrid convolutional neural network, integrating metadata with text. Wang. 2017 makes use of several classifiers which are: SVM, LR, Bi-LSTM and CNN as independent variables and compares the yielded results of the validating and testing sets for dependent variables. Exploring the findings, Wang. 2017 points out how the combination of text with metadata significantly improved accuracy performance. The control variable in this research seems to be the dataset, as it never changes except during preprocessing (Wang. 2017).
* Study 2
  + This research carried out by Vijayaraghavan’s et al. 2020 has a purpose of being an experimental study to observe the effectiveness of various Natural Language Processing (NLP) models such as the TF-IDF, CountVectorizer (CV) and Word2Vec (W2V). The objective is to find out which model is able to preserve most of the contextual information while efficiently determining whether the news is real or fake. The independent variables in this paper were the classifiers, which are the Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Long Short-Term Memory (LSTMs), and Artificial Neural Network (ANNs). From the investigative analysis carried out by Vijayaraghavan et al. 2020 finds out the best performing model was LSTM with CV. The obtained results are the mean test scores of each classifier with each NLP model which are the dependent variables. The dataset is a fixed dataset where the only change is during preprocessing, therefore it makes it a control variable (Vijayaraghavan et al. 2020).
* Study 3
  + The purpose for Jehad & Yousif’s study was to carry out an experimental investigation to observe the performance of 2 algorithms (Random Forest (RF) and Decision Tree (DT)) on a dataset of 20, 761 samples aswell as a TF-IDF as a feature extractor with the aim to deduct which algorithm is the superior one within this research’s parameters. The independent variables were found to be the algorithms used to help construct the model (RF, DT and TF-IDF). The dependent variables are the accuracy results achieved from the experiment. From their results the authors stated that Decision Tree outperformed Random Forest, possibly due to the fact that Random Forest works better on smaller datasets. The control variable is the dataset due to the fact it was fixed and no records were altered expect in the preprocessing stage. (Jehad & Yousif, 2020)

Task 4

1. Since I will be using the LIAR dataset for my dissertation, I shall be choosing it for the purpose of this task. The LIAR dataset consists of 12,800 manually labelled short statements and consists of 6 labels rather than the more conventional method of true or false. The labels are as follows; True, Mostly True, Half True, Barely True, False, Pants of Fire. Taking this into consideration we can conclude that my dataset falls under the quantitative category since most of the work will be dealt with numerical and statistical data. To further evaluate the findings, a confusion matrix will be used to determine the True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) aswell as calculating the F1 Score, Recall and Accuracy of the models. This will help give a better understanding of where the models stand when it comes to fake news detection, by comparing the aforementioned findings with previous studies as a benchmark.