COMP SCI 7327 – Assignment 3

Project Overview:
With the rapidly changing landscape of digital data, humans are struggling to keep up with the information which also makes it extremely difficult for them to distinguish what is true and fake. Given unstructured textual data, identifying whether it provides valid or invalid information is a popular application area in Natural Language Processing (NLP). The need for such systems is further highlighted by the proliferation of fake news related to COVID-19. Due to the timely nature of this application area, in this project, you will develop deep learning models to detect whether a given textual data is a fake news or not. In other words, this will be a binary classification task. Use the following steps, during your project development.

Step 1: Download the dataset and perform Exploratory Data Analysis (EDA)
Use the following dataset for your project: https://www.cs.ucsb.edu/~william/data/liar_dataset.zip (link: https://sites.cs.ucsb.edu/~william/papers/acl2017.pdf)

EDA is a crucial step in any data science project, which enables us to perform some initial investigations on data so as to discover whether

- the dataset has any missing values
- to identify whether the dataset is having any duplicate values
- to spot anomalies
- to check whether it is necessary to normalise or scale any numerical variables
- to check whether there are categorical variables and encode them
- to check the distribution of the target variable and check whether we need to incorporate any data imbalacing techniques

Hints for Step 1:

- Decide which of the above points are relevant to this project and perform the EDA for the selected points.
- Use matplotlib and seaborn to draw the required plots.

Step 2: Preprocess the dataset

Text preprocessing is a step in the machine learning workflow to clean the text data and make it ready to feed to the models you develop. Why do you need preprocessing? Text data contains noise such as, punctuation, text in a different case, stop words, etc. that may not useful in the modelling stage. Mentioned below are some of the preprocessing techniques that you could use in this project.

- Lowercase
- Remove punctuation

- Remove stop words using NLTK
- Stemming

Step 3: Construct word embeddings and develop your LSTM deep learning models using Keras sequential API.

For word embeddings use one of the following options.

- Embedding Layer in Keras
- Word2Vec
- GloVE

You have the freedom to pick your own deep learning architecture.

Step 4: Parameter Tuning

In machine learning, hyperparameter tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. In this step, select one or two parameters and optimise them to further improve the performance of your LSTM model.

Step 5: Model Evaluation

Use standard evaluation metrics such as precision, recall, F-measure and AUC to evaluate the performance of your models.