ECE 7995: ST AI for NLP
Assignment 2
Due date: October 15, 2023

The goal of this assignment is to implement a deep neural network (DNN) model for fake news detection. We will also implement a named entity recognition (NER) DNN model to extract features from the dataset and feed them to the fake news classifier.
In this assignment, students have to use the same tools we use in the lectures, such as nltk, sklearn, keras, etc. (Note: for the implementation of DL models, you can also use tensorflow or pytorch)

For each dataset (fake news and NER) separately, load the data using pandas DataFrame. Check the data balance and report it. (**Note:** you need to remove unlabeled samples (if any).)
Then, clean the data, build the vocabulary, convert tokens to indices, and split the data into (train, validation, and test) sets. (**Note:** if the date is split into train and test, use 10% of the train set for validation)

Follow these steps to build the classification model:

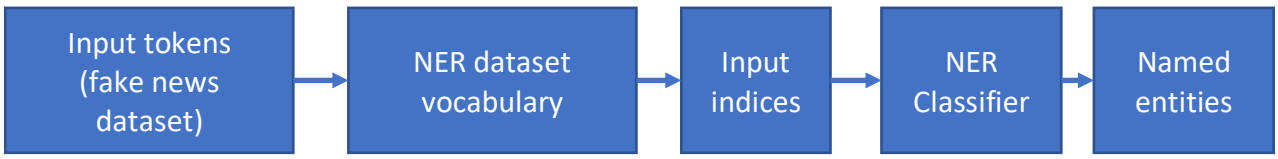
1. Implement the baseline fake news detection model which is a DNN for binary classification. Trian the model on the train set. Then, evaluate the model on the test set, and report the evaluation metrics, such as accuracy, confusion matrix, f-score, etc.

2. Compare the performance of the model in step-1 when a) embeddings are randomly initialized, b) embeddings are initialized using a pre-trained embeddings and fine-tuned during the DNN model training c) embeddings are initialized using a pre-trained embeddings and frozen during the DNN model training. Select the best model to use in step 4. (**Note:** select a pretrained embeddings model from the genism api. Make sure you consider the OOV)

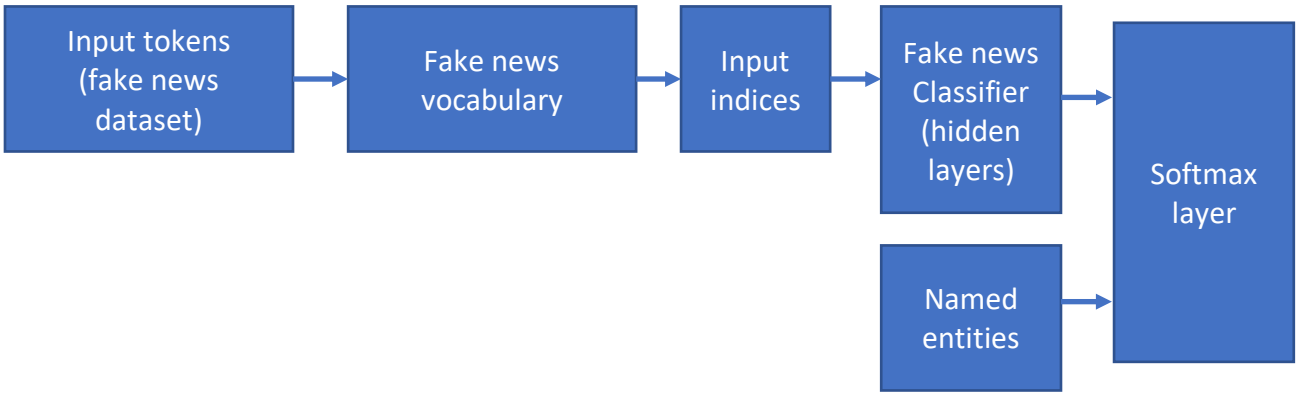
3. Evaluate the embeddings of the three models in step 2 on the word analogies in this file (analogy-test.txt). Report the accuracy of each word embeddings and elaborate on the results. (Note: you can access the embedding vectors from the trained DNN model. The code for analogy computing is available on canvas)

4. Implement a NER model to extract features for the fake news classification model.

    a. Train a DNN classification model on the NER dataset
    b. Use the trained model in (step 4-a) to extract features from the fake news dataset. (Hint: you can run this model offline on the fake news dataset and extract these features to use them when you train the fake news classifier in (step 4-c))



c. Implement a DNN classifier for fake news detection. The model has two inputs 1) the input documents (as in step 1), and the extracted features from step 4-b. Trian the model on the train set. Then, evaluate the model on the test set, and report the evaluation metrics, such as accuracy, confusion matrix, f-score, etc.



Note: submit one "Jupyter Notebook" that includes the whole code, explanation, and comments.