Name: Abdulalrahman Husham A Razzaq Alabbas Student ID:1905983

1. Introduction

The objective of the project is the prediction of the disease that a patient might have, given a short textual representation of the symptoms. This is a multi-class classification problem, with 30 classes, each of which is a disease. This project analyses the use of four machine learning and deep learning algorithms on the same healthcare dataset. The four algorithms that were developed for the project are TF-IDF with Logistic Regression, Feed-Forward Neural Network (FFNN), the Recurrent Neural Network (RNN), and the Long Short-Term Memory Network (LSTM). Each of these models used the same dataset.

2. Dataset Description

The data used for the model is the “Healthcare Symptoms-Disease Classification Dataset” obtained from Kaggle, comprising 25,000 samples of patients. Each sample contains data on the patient’s age, gender, a brief textual representation of their symptoms, the number of symptoms, and the respective disease identified. The target variable comprises 30 types of diseases like “Anxiety,” “Diabetes,” “COVID-19,” “Arthritis,” and “Bronchitis,” among others. This data is somewhat balanced, with each disease occurring approximately 800-900 times. The symptom texts provided are short and come with overlapping words for differing diseases, adding complexity to the classification problem.

3. Data Preprocessing

The same preprocessing steps were used for all the models. Missing values were removed. The Gender variable was changed to a numerical representation using label encoding, and the disease variables were changed to integers. The numerical variables (Age, Gender, Symptom_Count) were rescaled using StandardScaler.

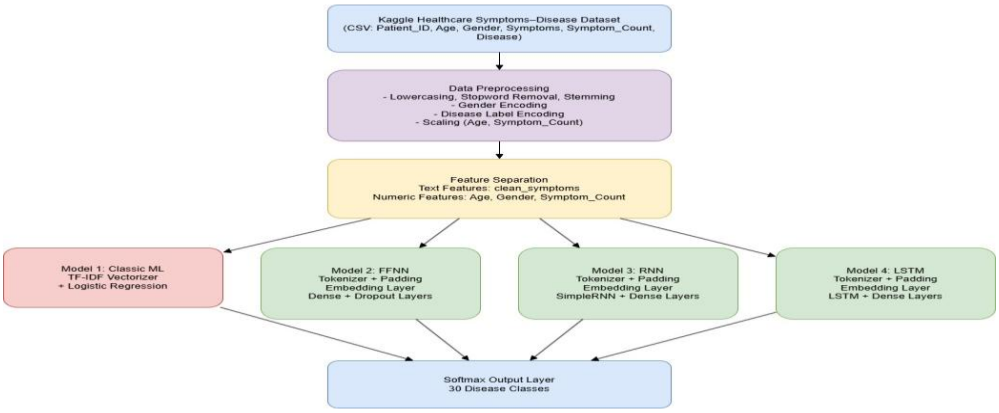
For the TF-IDF model, the symptom text was pre-processed by removing stop words, lemmatized, and vectorized using the TF-IDF vectorization technique. For the neural network approaches, the symptom text was pre-processed by tokenization, conversion of the tokens into integers, and padding the data into fixed-length sequences of 50. The train and test data were split using a fixed random seed (SEED = 42).

4. Model Architectures

In the first model, the TF-IDF vectorization technique is used, and the model is a multinomial Logistic Regression model that uses class weights. The other model is the Feed-Forward Neural Network using the embedding layer, the global average pooling layer, the dense layers with 128 and 64 units, and the dropout regularization technique. The numerical variables are concatenated with the textual variables.

In the third model, the pooling layer is exchanged for a SimpleRNN layer with 64 hidden units. In the fourth model, the RNN layer is exchanged for an LSTM layer. Everything else remains the same. Each of the neural networks was trained with the Adam optimizer and the categorical cross-entropy loss function. Early stopping was used for the RNN and LSTM networks. This helps the networks stop training when the validation loss ceases to improve.

5. System Architecture



The system integrates a complete machine learning cycle ranging from data loading, preprocessing, feature engineering, model building, evaluation, and eventually model saving. The different models, tokenizers, scalers, and label encoders have been saved so that the predictions can be repeated during the live session on the GitHub platform.

6. Experimental Results

Each of the four models was assessed for test accuracy, classification reports, and confusion matrices. The TF-IDF and Logistic Regression model had a test accuracy of about 3.0%, while the Feed-Forward Neural Network model attained about 3.6% accuracy. However, the highest accuracy of about 3.7% was attained by the RNN model, with the LSTM model attaining about 3.5% accuracy.

As the number of disease classes is 30, the number of random predictions would stand at around 3.3% of the total. The confusion matrices also reveal that the predicted outputs were dominated by a handful of disease classes. The fact is that very few correct predictions, that is, none for a large number of classes, were made for the symptom data.

7. Discussion and Conclusion

In all the four models, the accuracy level was very close to the random guess. This poor performance of the model is predominantly because of the overlap of symptoms among various diseases, the generic nature of symptom reports, and the high number of classes for the output. It is also easy to identify common symptoms of various diseases, including dizziness, vomiting, fatigue, and pain, among patients, because of which it becomes very tough for the model to identify specific patterns. It appears that the data has the properties of partially synthetic data or data that is not strongly correlated, hence not optimal for machine learning, deep learning, or both.

Although the RNN performed the best among the models, the improvement was slight. This project also illustrates that the complexity of the model does not always ensure success, provided the data does not have sufficient cues for the model. Any progress in the model will thus require improvements in the quality of the clinical data used, the symptom reports, and the medical embeddings.