Description

Implemented here is a LSTM (Long Short-Term Memory Network), which is a variation of a RNN (Recurrent Neural Network). The data is processed (stemming, stopwords removal case folding, and lemmatization) and processed into an Embedding layer represented as a one-hot vector. Once divided according to the cross validation it is passed through an embedding layer for dense representation. And then to dropout and fully-connected layers, there is than a parametrized number of hidden layers, reaching a sigmoid function which allows the output label to be determined (by rounding this output),

Hyperparameter Setting

nn.Dropout(0.2) : Dropout allows regularising of the inputs so that there is a decreased effect of overfitting, which would decrease the difference of train/test accuracy, and the classification for unseen (test data) is expected to be higher.

Number of hidden layers: The number of hidden layers should reflect the complexity of the problem, programmatically sensible values between 2-10 were tried. 5 Provided the best accuracy.

Embedding Dimensions: Another hyperparameter tuned b y systematically varying the dimensions from 32 - 200. Tradeoffs of time were considered.

Number of Epochs: The epochs are the number of times the dataset is passed through the network/update process, Systematically I tried values between 2-10 and found 6 a good accuracy/computational time tradeoff

Batch Size: The Number of Training examples used per epoch before updating the weights of the network, because of the noisy data after processing I was limited to a maximum of 16 batch size (Must be a multiple of 8), therefore this was used.

Performance

In order to create a realistic run-time trade-off I selected parameters that would be completed with < 30minutes.

- Range of averages for 80/20 Split from 60-80%
- The accuracy from 5CV fold was often less as using a smaller representation of the data would decrease the ability of the classifier to generalize.
    - Mean = 0.6333333333333333
- Accuracy was increased for processed (Stem, Lem) data.