

1 - Live Testing Final Model Folder

Contains a notebook which uses the FINAL LSTM Model to do real time predictions by only typing a sentence which the user wants to classify.

2 - LSTM Folder

Contains a notebook with all the code necessary for building the final LSTM model including some preprocessing and other experimental campaigns like the two LSTM model, the hybrid LSTM and RNN model, etc. See the writeup for more information. The dataset required to run the notebook can be downloaded from the following links:

<https://drive.google.com/open?id=1M9KWPjMQNwr753cstlhCBSsTYPHR5U0F> – message_length.csv

3 - Main Notebook Folder

Manual_label_edits.pdf: includes information about the label reduction process.

boosting_bagging.ipynb: contains the code related to boosting, bagging, and random forest model.

Preprocessing_KNN_DT.ipynb: contains all the code necessary for preprocessing the dataset including the KNN clustering that was performed for label reduction. It also includes the Decision Tree Model

The dataset is beyond the bCourses upload limit therefore the following links will provide access to the data. ALL of them must be downloaded to the main folder in order for the MAIN_NOTEBOOK to work. The notebook contains code for reading the data directly of the google drive folder since our code was written in Google Colab. This means that in order to run the MAIN_NOTEBOOK you must upload it to a google drive folder along with the following data files:

https://drive.google.com/open?id=11QfWSGZX1N28pdKoAbm_hUpU_FdKKwZL – data.filtered.csv.zip

<https://drive.google.com/open?id=1zIO3lbkrKE1dt2Z6Guv5AWCUxAdjilM7> – email.csv.zip

4 - Clustering Label Visualization Folder

Contains a notebook that creates visualization of the message embedding using PCN and TSNE for dimensionality reduction.

5 - Doc2Vec Folder

Contains a notebook of the incomplete Doc2Vec model. See the writeup for more information. Dataset required to run the notebook:

https://drive.google.com/open?id=1luAmJxsFVudZakhlbYoPvbZ_3xNx9GL7 – medium_data.csv