### **CS 6375.003 Machine Learning**

### **Final Project**

**--- Project Topic ---**

Twitter Sentiment Analysis with LSTM - Importance of the Neutral Classification

**--- Team Members ---**

* Rishi Dandu - rkd160030
* Melissa Fulfer - mdf150230
* Yuncheng Gao - yxg151230
* Usuma Thet - uxt130030

Late days being used: 1 day

Link to dataset (sentiment140): <https://www.kaggle.com/kazanova/sentiment140>

**--- Instructions on how to run code ---**

Preprocessing:

splitdata.py

example: python splitdata.py training\_size sequence\_length dataset.csv

Model Training:

train.py

To train the model, use command line to run "train.py"

The hyperparameters are manually set inside of the program.

This program outputs a trained model using LSTM network with the 1.6 million

tweets dataset to predict the sentiment of tweets.

Model Evaluation:

evaluate.py

example: python evaluate.py testSet.csv x\_test.csv model.h5

Use the program "evaluate.py" to evaluate the performance of the model.

Running instructions:

First, run splitdata.py, specify the training size, sequence length and the path to

the full dataset.

This program will output a couple of files required to train and do model eval:

testSet.csv - the test set for our model

trainSet.csv - the train set for our model

x\_test.csv - the parameters for our test set

Second, run train.py with the training dataset. It will train a model using

hyperparameters specified inside of the train.py file.

This program outputs a fully trained model ready for evaluation.

Third, run the evaluate.py with the testset, parameters for the test set and also

the trained model to evaluate the model being tested.

This program will output the final evaluation metrics for the model as a

confusion matrix.

To efficiently evaluate our models, you can download the pre-trained models

from the google drive link provided below:

https://drive.google.com/drive/folders/16SFLQ0mKEd\_nCTdNlr4Z144le5EfW61w?usp=sharing

**--- Log file of experiments and parameters ---**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SEQUENCE\_LEN | EPOCHS | BATCH\_SIZE | THRESHOLD 1 | THRESHOLD 2 | Results |
| 300 | 6 | 1024 | 0.4 | 0.7 | Precision: 0.79  Recall: 0.79  F1-store: 0.79  Accuracy: 0.7894 |
| 400 | 4 | 1024 | 0.3 | 0.7 | Precision: 0.78  Recall: 0.78  F1-store: 0.78  Accuracy: 0.7845 |
| 300 | 8 | 512 | 0.3 | 0.7 | Precision: 0.78  Recall: 0.79  F1-store: 0.79  Accuracy: 0.7887 |
| 250 | 5 | 512 | 0.3 | 0.7 | Precision: 0.79  Recall: 0.79  F1-store: 0.79  Accuracy: 0.7889 |
| 250 | 8 | 256 | 0.4 | 0.7 | Precision: 0.78  Recall: 0.78  F1-store: 0.77  Accuracy: 0.7712 |