Tweet Sentiment Analysis

SENTIMENT ANALYSIS OF TWEETS ON A THEME

Sreedhar V | Deep Learning for Natural Language Processing | 18/4/2021

Coding Assignment

PROBLEM

Process the training data consisting of various recent themes centered around current events. There are a series of sentiments given in the training data. Predict the sentiment of the test data.

OVERVIEW OF THE CODE

DATA FIELDS

- UserName - an anonymous id
- ScreenName - the id of screen
- TweetAt - the location of tweet
- Tweet - the tweet
- Sentiment - the sentiment

This notebooks aims at presenting the Sentiment Analysis on Covid-19 tweets. The emphasis has been on the text (tweets) and its in-depth analysis for pre-processing. The modeling will be done in a separate notebook. This notebook is prepared on the Covid-19 tweets which are tagged manually from Highly Negative to Highly Positive - i.e. five classes.

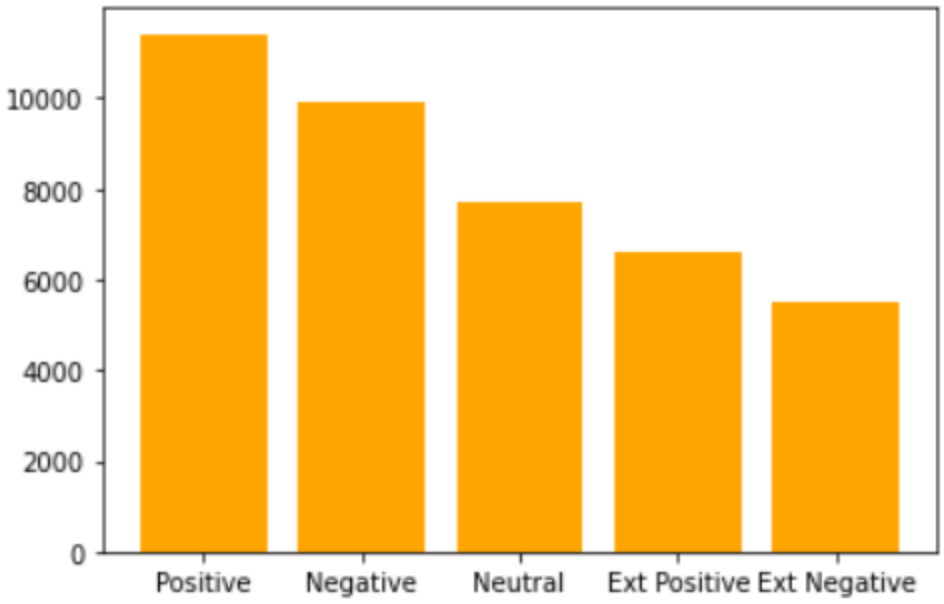
Data Files

1. train.csv - For training the models, we provide a labelled dataset of

41157 tweets.

2. There is 1 test file test.csv - The test data file contains same contents as train except the sentiment field.

EDA on Sentiment:

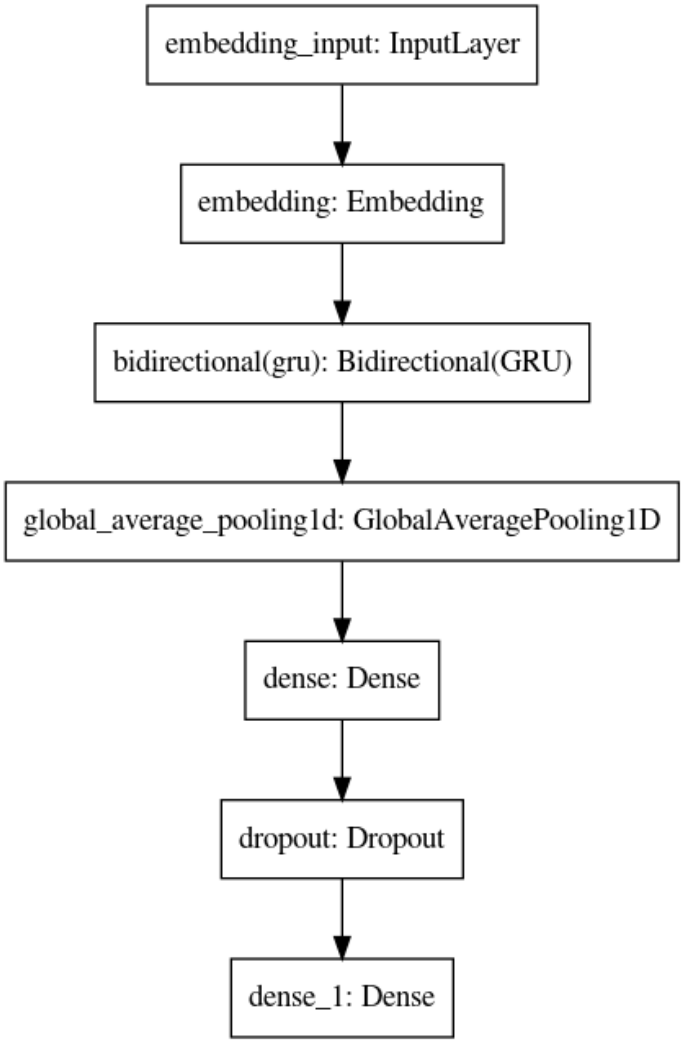


Data Pre-processing:

The test data is contains many unwanted text pattern such as URLs, HTML links, Numbers , Punctuations , Mentions, hashtags and stopwords We will remove them from our data and vectorize out train data into arrays for feeding into the model using Tokenizer in scikit-learn. We can also use Count vectorizer to vectorize out data.

MODEL:

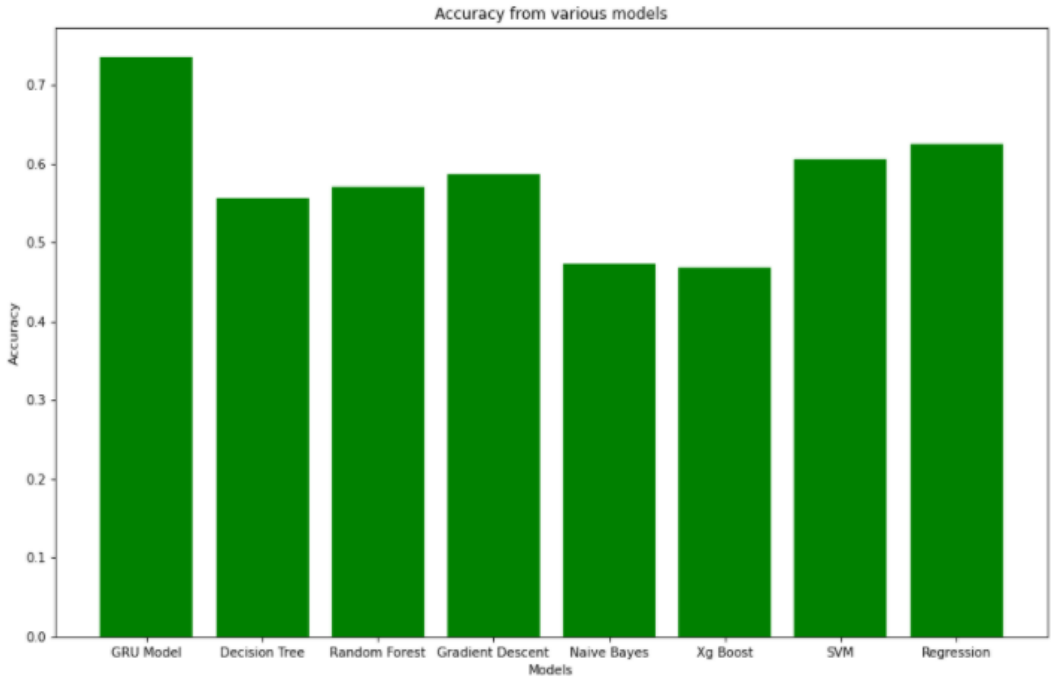
- GRU Model architecture.



OTHER MODELS:

- Decision Tree
- RandomForest Classifier
- SGDClassifier
- Naïve Bayes
- Xgboost
- SVC
- Logistic Regression

I have used basic model with the default hyper-parameters for all these algorithms.



From this graph , GRU model has an edge over all the other models since it has 75% accuracy on Validation set. We will use this model for testing out test data and generate the output.

RESULTS:

Since it has 5 class target output, our validation accuracy is pretty low. We can get better accuracy if it is a binary or three class output.

```
Accuracy on training data is:- 95.36218643188477 %
Loss 17.02379584312439
Accuracy on validation data is:- 73.57871532440186 %
Loss 110.077965259552
```