Tweet Analysis for COVID-19

FAKE NEWS DETECTION

ML Project | Group 4

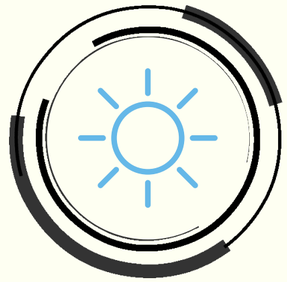
PREPROCESSING

Remove URLS

Remove Punctuations

Remove Stopwords

Remove Frequent Words



Remove Rare Words

BASELINE RESULTS

Passive Aggressive
0.9439

SVM
0.9431

Logistic Regression
0.9224

Naive Bayes
0.9192

Decision Tree
0.8503

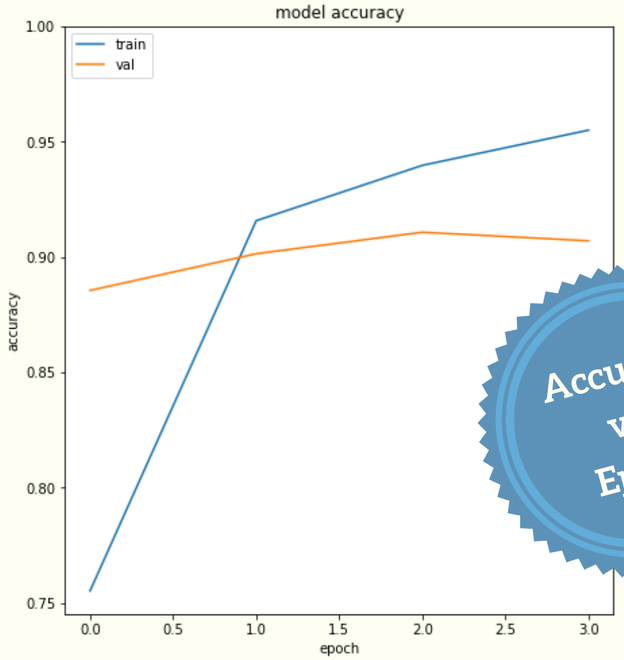
KNN
0.6926

#0.9276

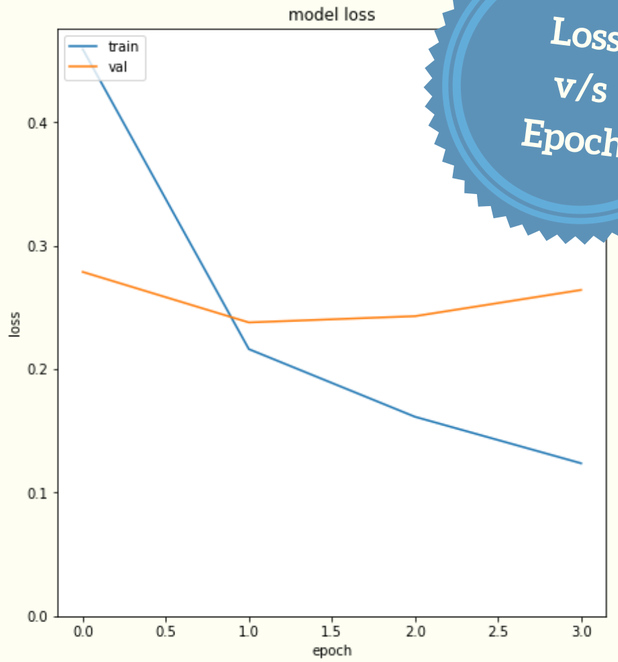
top-5 base-line

VOTING

A voting was done on the top-5 best baseline solutions to generate a new solution. These included Passive Aggressive, SVM, Logistic Regression, Naive Bayes, and Decision Tree. The resultant performed better than the bottom three but failed to perform better than PA and SVM.



Accuracy v/s Epoch



Loss v/s Epoch

trainable embeddings

RNN

An RNN with the following architecture was constructed that took tokenized, preprocessed words as inputs and returned the label for the text as REAL or FAKE.

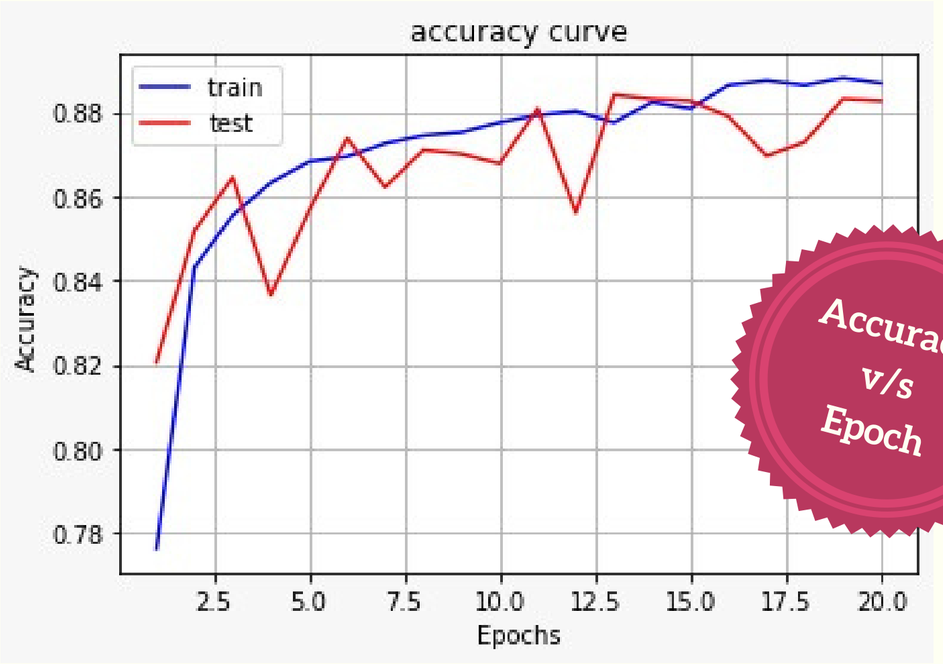
1. **Vocab-Length X 64 Embedding**
2. **64 o/p Bidirectional**
3. **32 o/p Bidirectional**
4. **64 o/p Dense (ReLU)**
5. **1 o/p Dense**

#0.9188

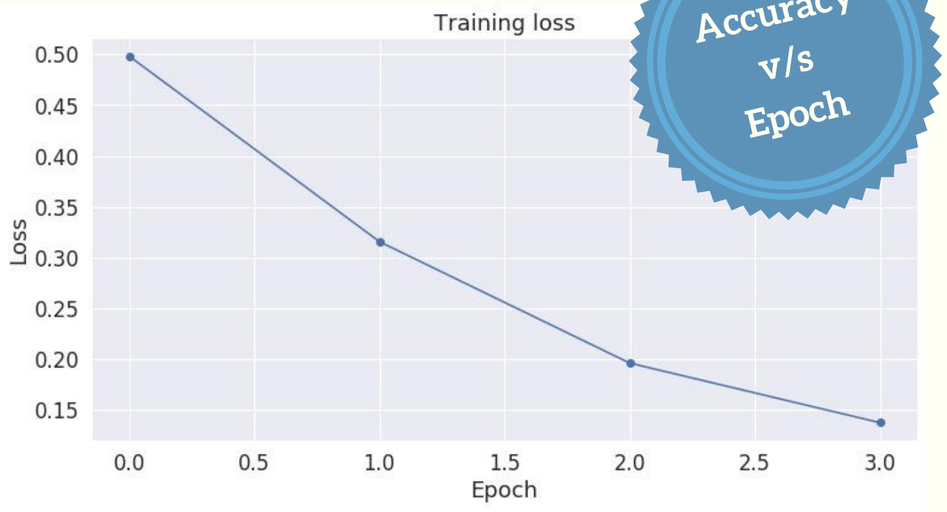
GloVe embeddings

RNN

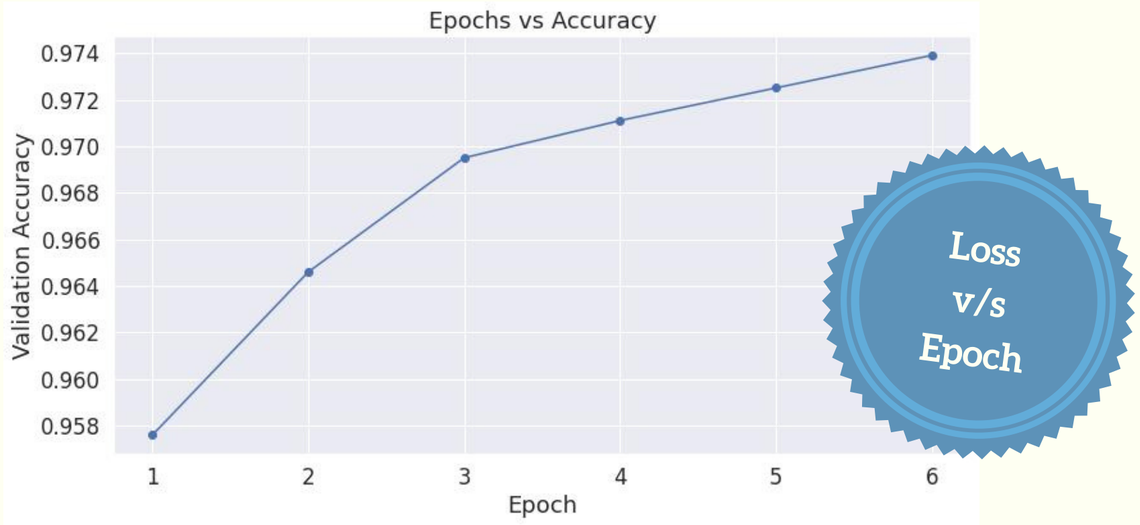
Another RNN architecture was trained using GloVe embeddings instead of trainable embeddings as the input to the LSTM layer. The architecture for the RNN was similar in terms of the layers only the dimensions of the embedding layer were modified to accomodate the GloVe embeddings with vector lengths 50



#0.8827



Accuracy v/s Epoch



Loss v/s Epoch

fine-tuning

BERT

A pre-trained model of BERT for English Text classification was used for Text-based classification post fine-tuning. A 2-layer classification architecture was added over the embeddings genearted by BERT in order to classify the text as REAL or FAKE news.

#0.9739

01

best PERFORMANCE

BERT was the best model with an **accuracy of 97.39%**.

02

individual COMPARISONS

- RNN with **trainable embeddings** performed **better than** RNN with **GloVe Embeddings**.
- RNNs were better than a few base-line models but not all. **Passive Aggressive**, **SVM**, **Logistic Regression** and **Naive Baye's** performed **better than RNN**.