Quora Insincere Questions Classification Project Milestone Report II

The goal of the project is to identify insincere questions based on question wordings. To achieve this the raw text data are transformed into feature vectors which can be used in a machine learning model. The following methodologies are implemented to transform text data into feature matrix: Term Frequency (TF), Term Frequency-Inverse Document Frequency (TF-IDF) and word embeddings. Different machine learning/ deep learning models are trained using different feature matrices. Predictions based on the test data are submitted to Kaggle and are evaluated using F1 Score.

Models:

TF Model

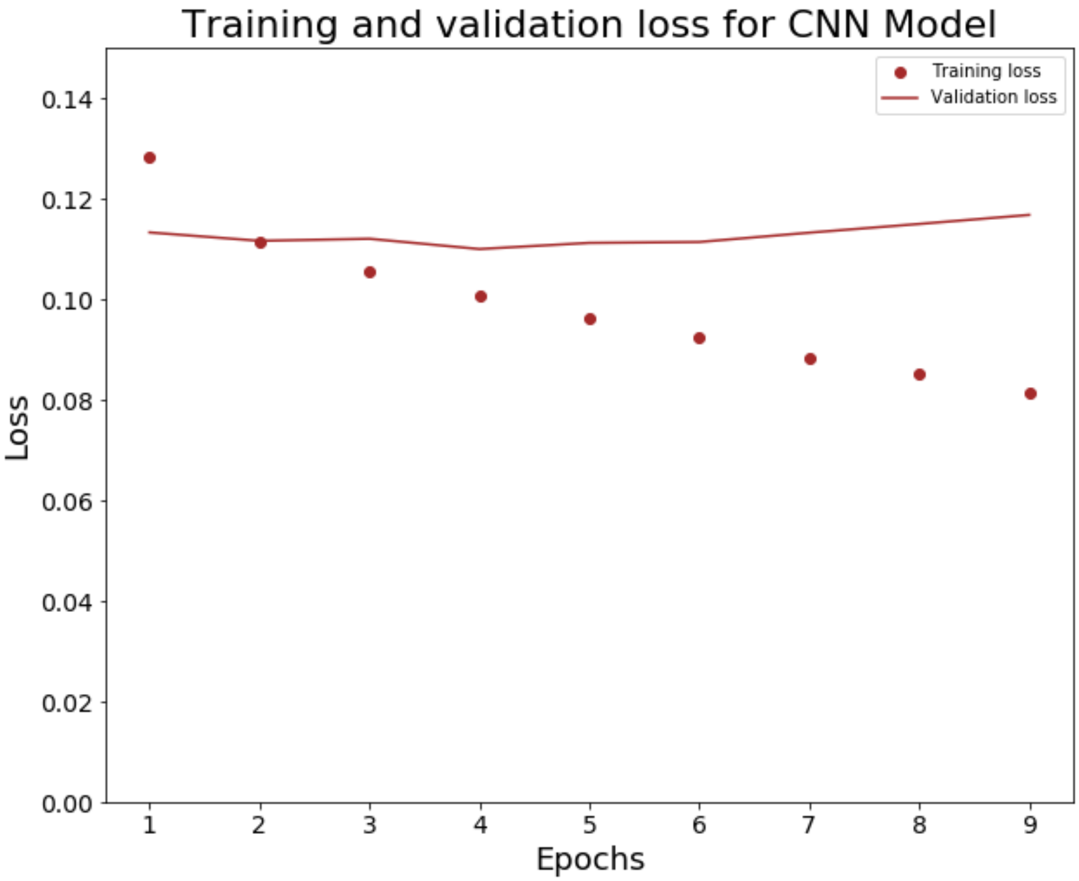
- A logistic regression model, a multinomial naïve bayes model and a random forest model are trained respectively using the TF matrix.
- Although the multinomial naïve bayes model gives the best out-of-sample F1 score (0.5360), the model is not really useful in identifying insincere questions as the F1 score is close to 0.5.

TF-IDF Model

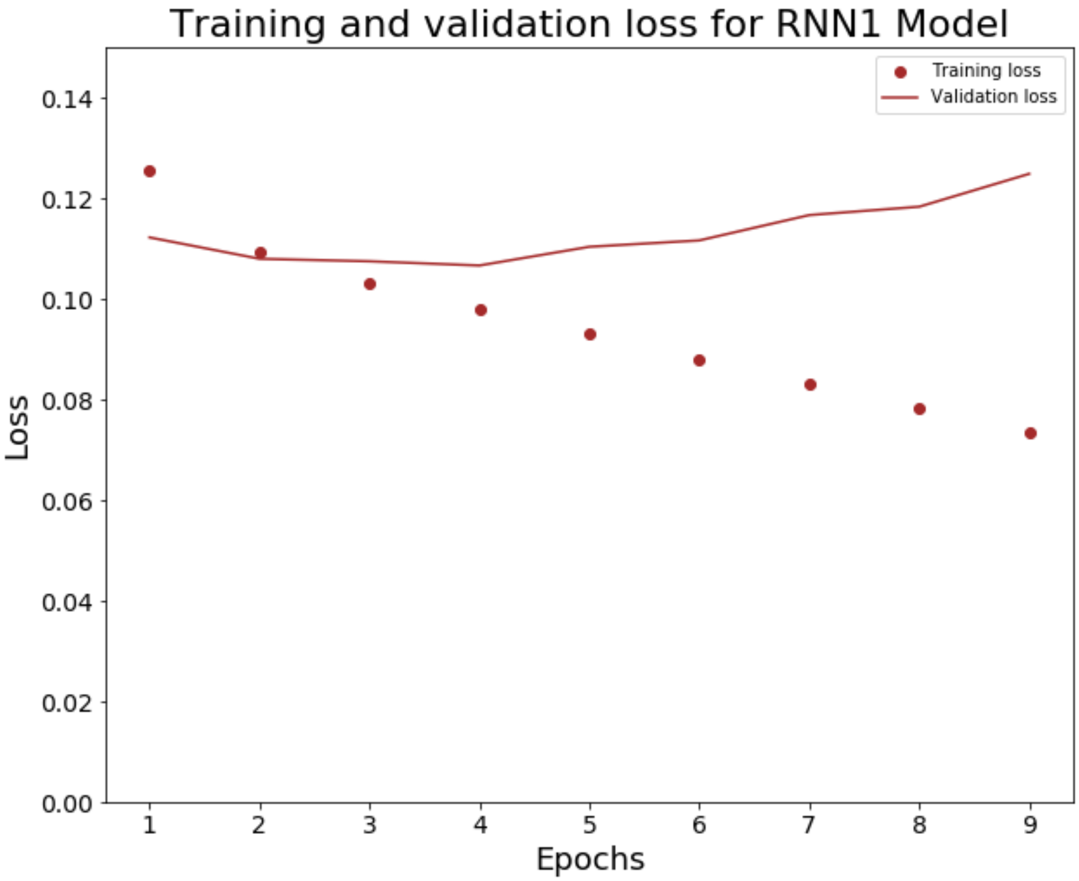
- A logistic regression model and a random forest model are built using the TF-IDF matrix.
- Yet, both models perform even worse than a random model (logistic regression: 0.4603) (random forest: 0.2476).

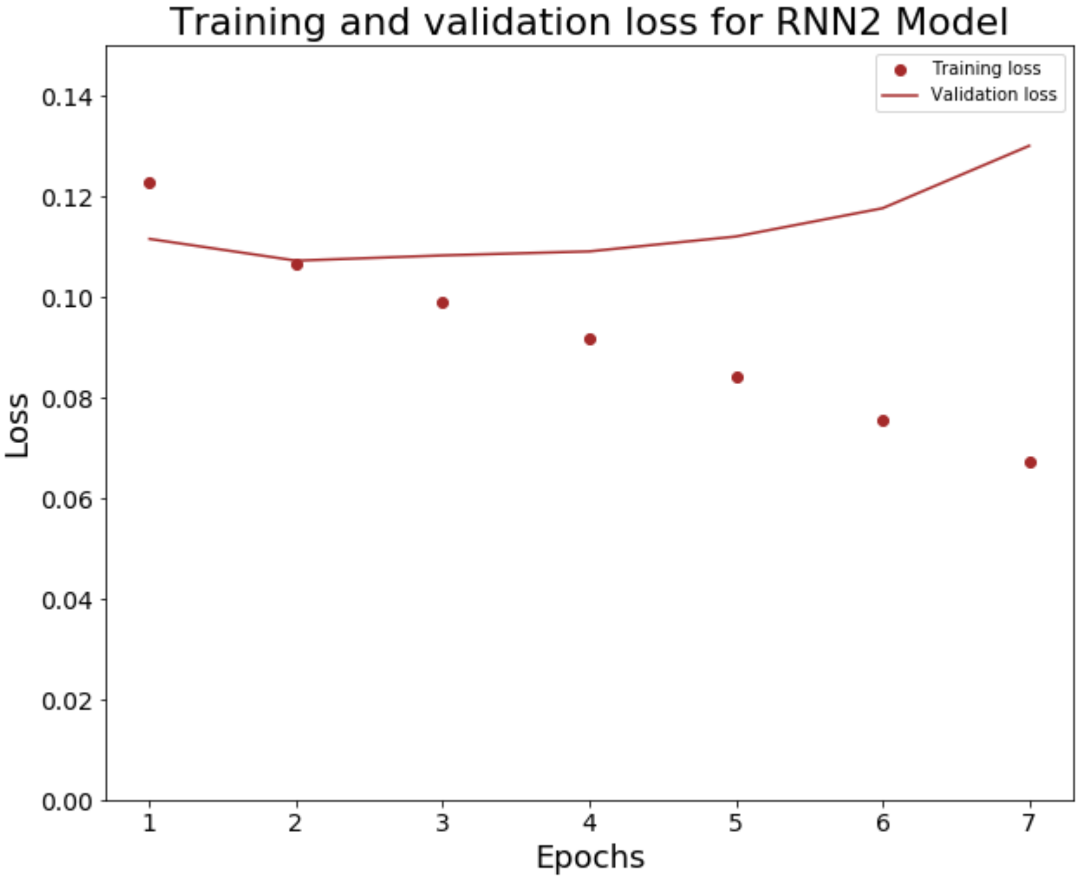
Word Embeddings

- Pre-trained Glove and Paragram word embeddings are utilized to transform text into dense vectors, and the transformed matrix is then fed into different Neural Networks. Early stopping is implemented to avoid overfitting.
- A Convoluted Neural Network with 1-Dimensional convolution layers, maxpooling layers and dropout is built. It gives an out-of-sample F1 score of 0.6642.



- Two Recurrent Neural Networks are constructed. One of the models (RNN1) consists of LSTM and dropout while the other one (RNN2) has bidirectional LSTM layers and dropout. RNN1 gives an F1 score of 0.6689 while RNN2 gives a slightly higher F1 score of 0.6697.





My best performing model is RNN2 Model (F1 score: 0.6697).