Quora Insincere Questions Classification Project

Introduction

Research Objective

Quora is a knowledge sharing platform for asking questions and connecting with others who contribute unique insights and quality answers. Unfortunately, insincere questions are posted from time to time and these are the questions that are based upon false premises, intend to make a statement rather than look for helpful answers or even try to insult against a specific group of people. Such questions are against Quora's core values and might pose a threat to the Quora community.

The goal of this project is to identify insincere questions based on the question wordings so that Quora can develop scalable methods to detect toxic and misleading content.

Data

The datasets for this project are available on Kaggle. They are:

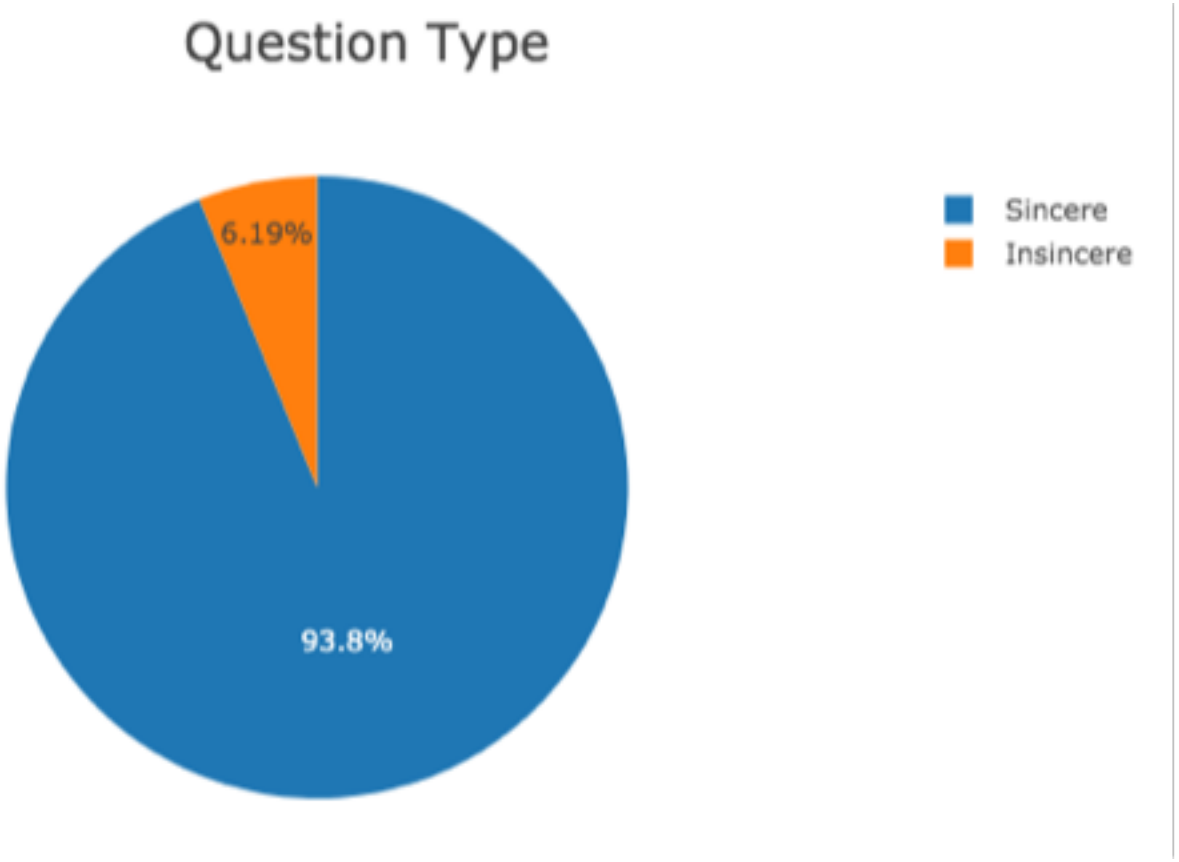
Training & Testing data:

- For each question the unique question identifier and question text are provided. For training data we also have the label that indicates whether the question is insincere or not.

Embeddings:

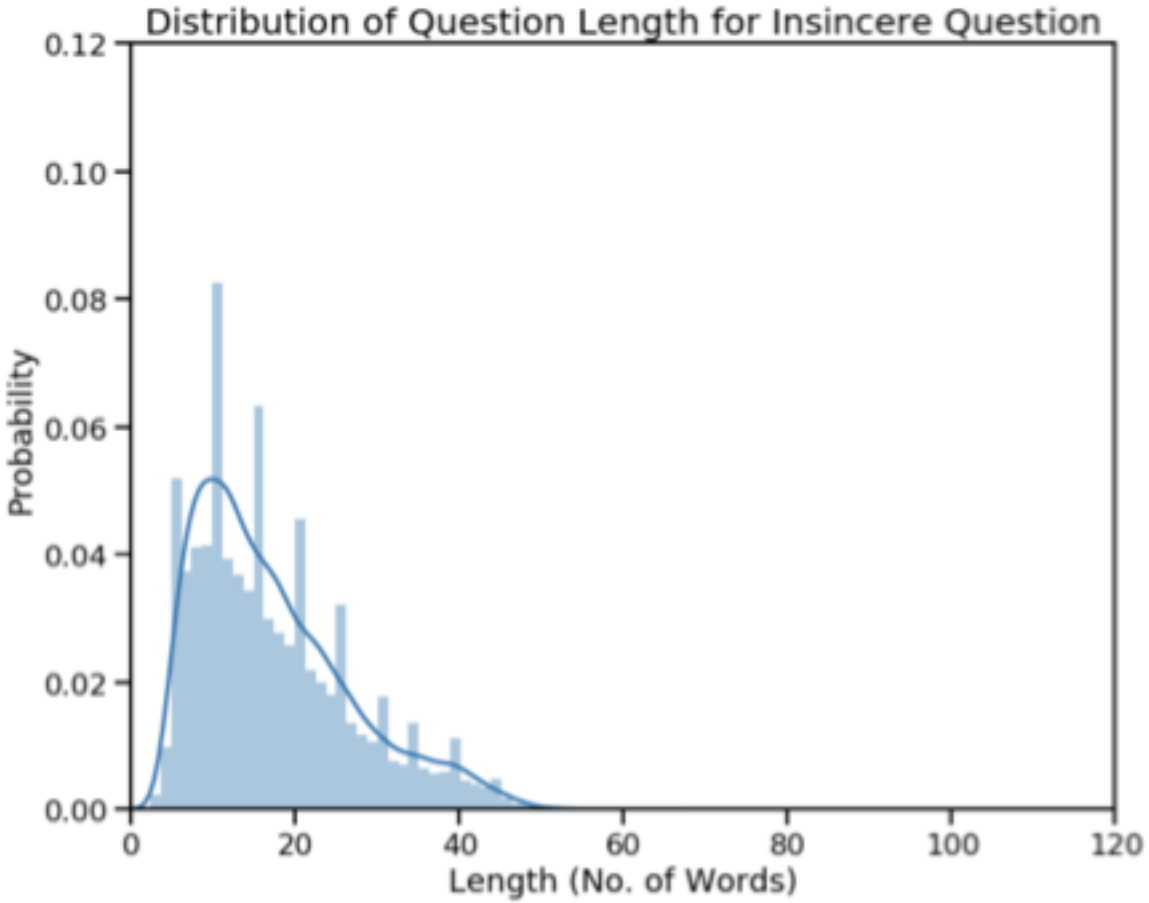
- The following embeddings are provided for text processing:
  - GoogleNews-vectors-negative300
  - Glove.840B.300d
  - Paragram_300_sl999
  - Wiki-news-300d-1M

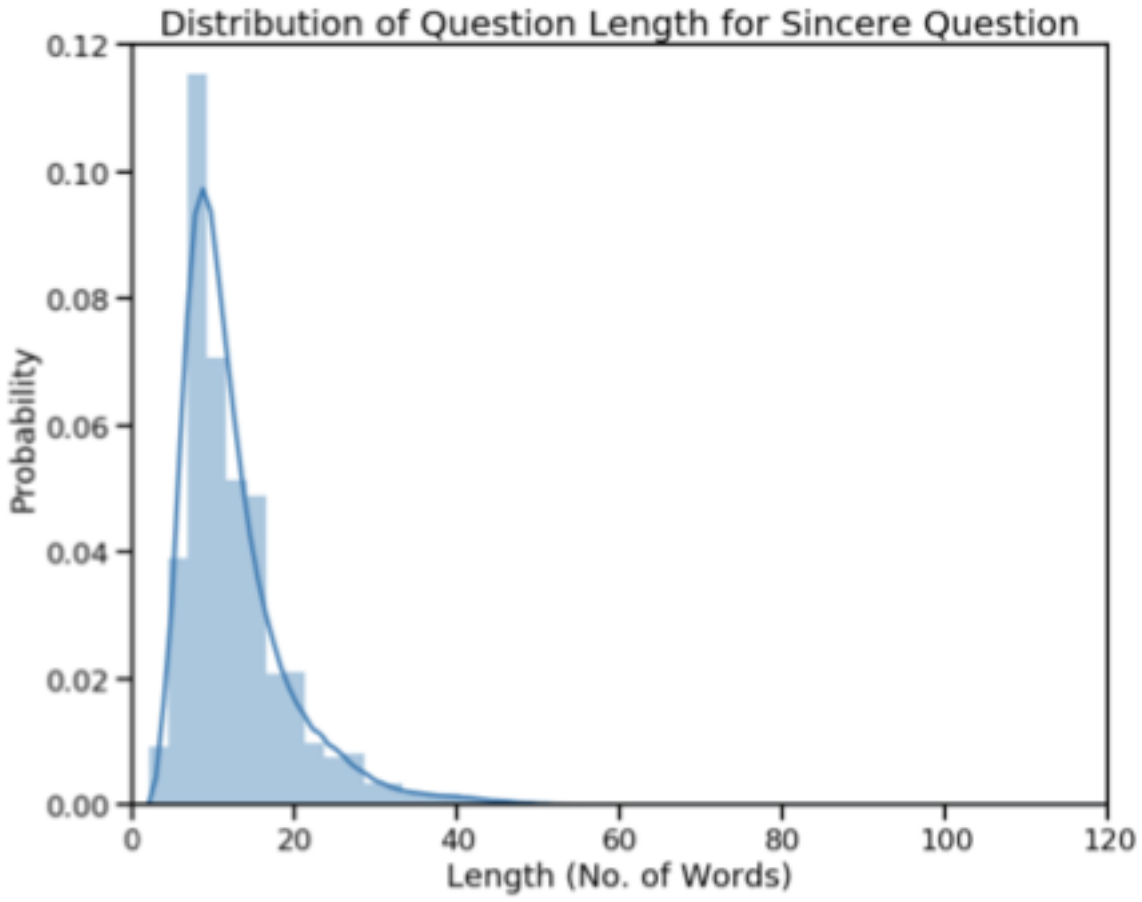
Exploratory Analysis



- We have an imbalance dataset with only 6% of questions labeled as ‘insincere’ and 94% as ‘sincere’.

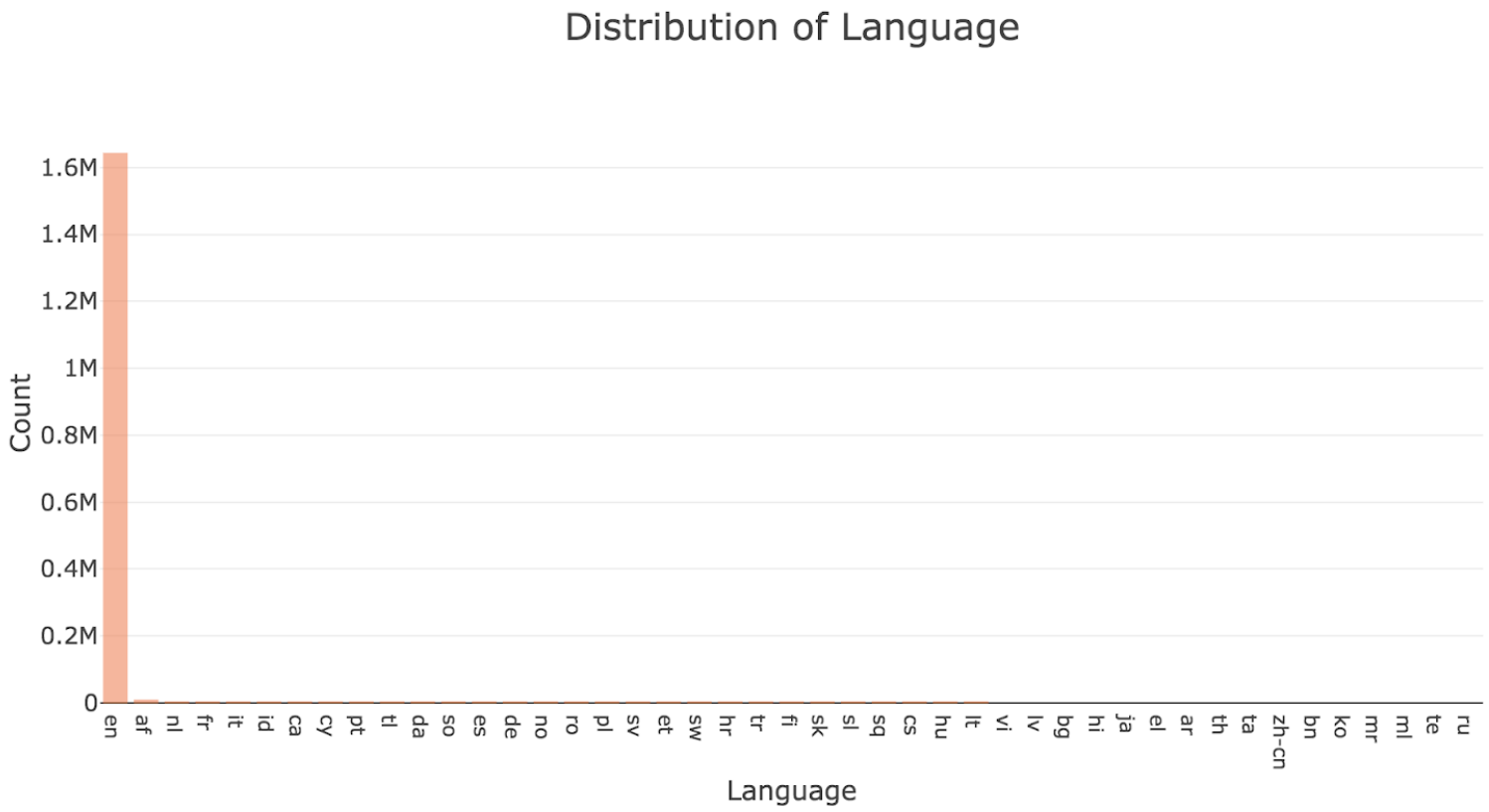
Exploratory Analysis





- The distributions of question length between insincere and sincere questions are quite different; insincere questions have a lower and broader peak.

Exploratory Analysis



- Most questions are in English. Since we only have very few questions in other languages, it is safe to ignore them.

Exploratory Analysis

Word-cloud for Insincere Questions

Word-cloud for Sincere Questions

- The word-cloud for insincere questions is mainly made up of words related to nationality/ race and gender. On the other hand, the word-cloud for insincere questions is more diverse and consists of different topics like knowledge, career, relationships, etc.

Modeling

- Classification
  - Outcome variable: Whether a question is insincere or not
  - Features: Features generated from 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.

TF Model

- A logistic regression model, a multinomial naïve bayes model and a random forest model are trained
- Best Model: multinomial naïve bayes model (F1 score: 0.5360) but a model with such a low F1 score is not useful in identifying insincere questions

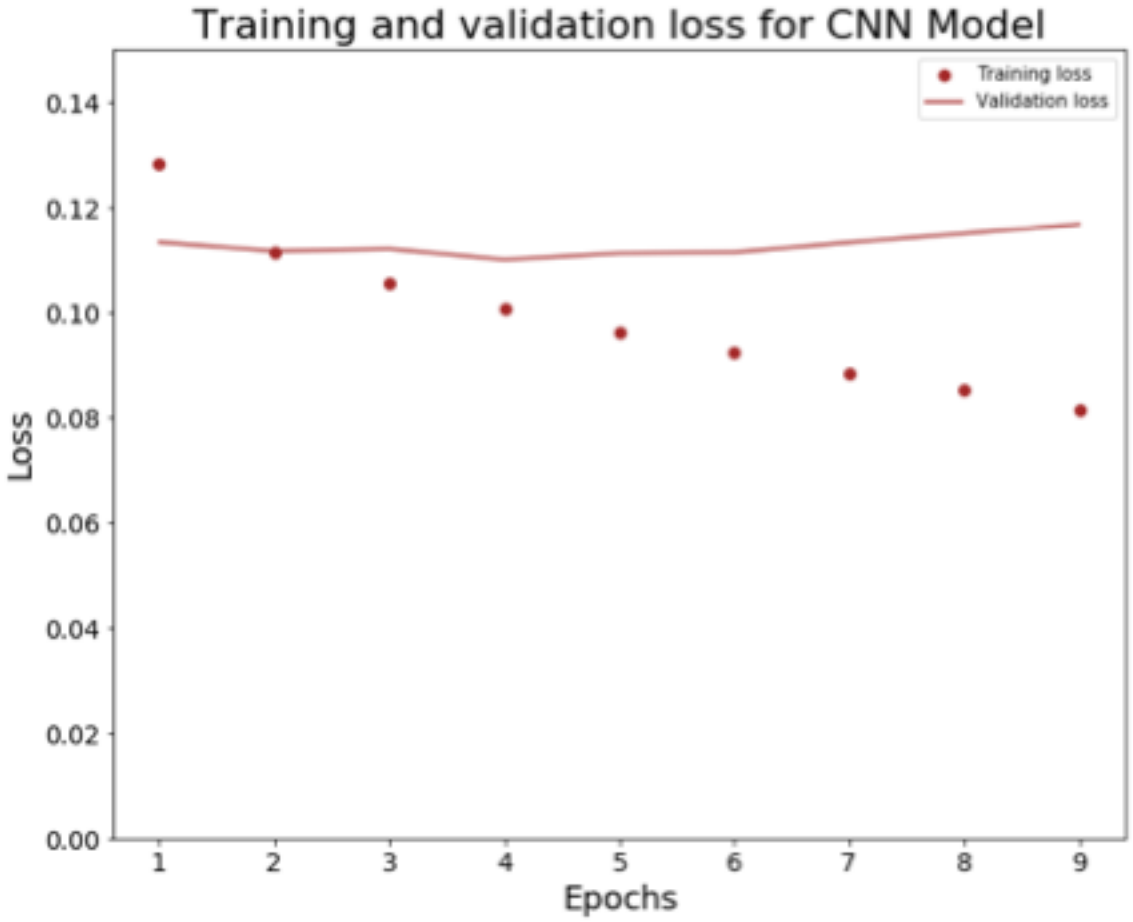
TF-IDF Model

- A logistic regression model and a random forest model are built
- Both models perform even worse than a random model (logistic regression: 0.4603, random forest: 0.2476)

Word Embeddings

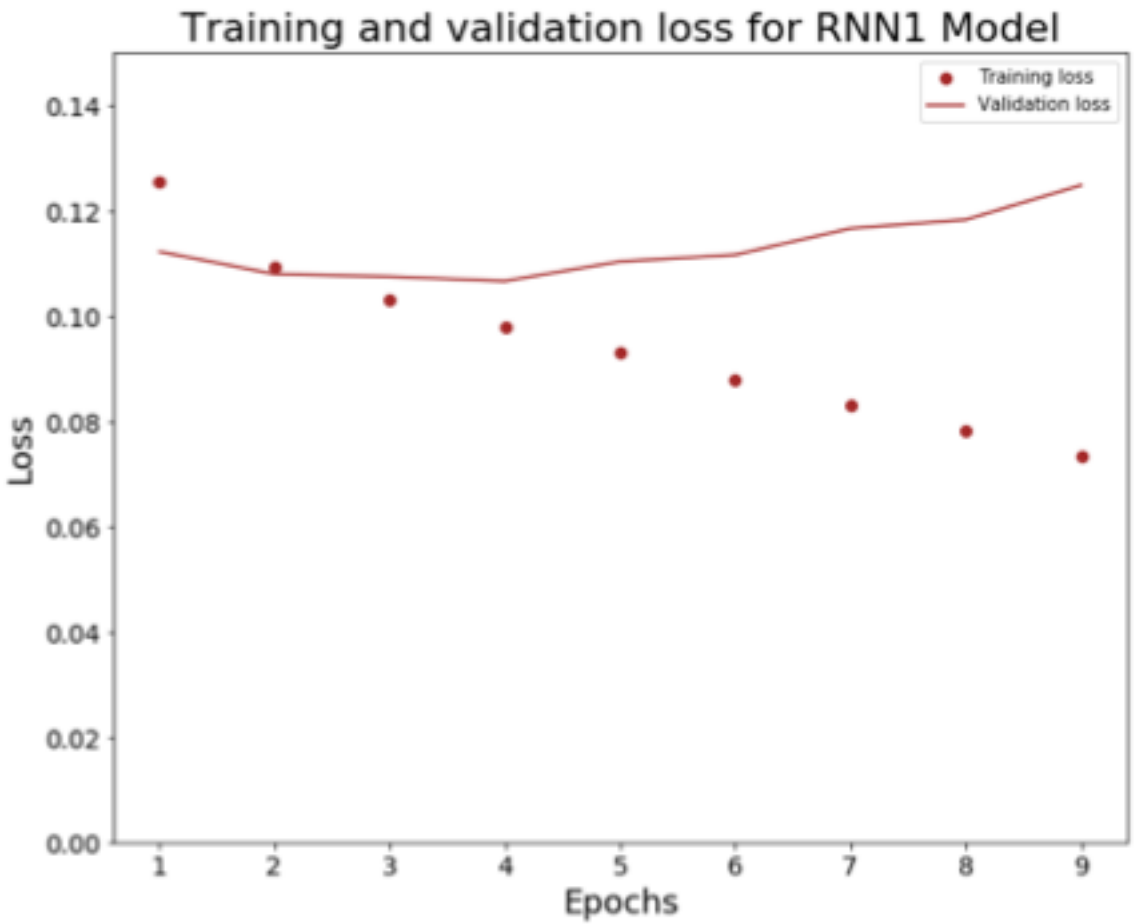
- Pre-trained Glove and Paragram word embeddings are utilized to transform text into dense vectors, and the transformed matrix is then fed into different Neural Networks.
- Early stopping is implemented to avoid overfitting.
- Different F1 scores are calculated based on different classification thresholds. The threshold which gives the highest validation F1 score will be chosen.

Convoluted Neural Network



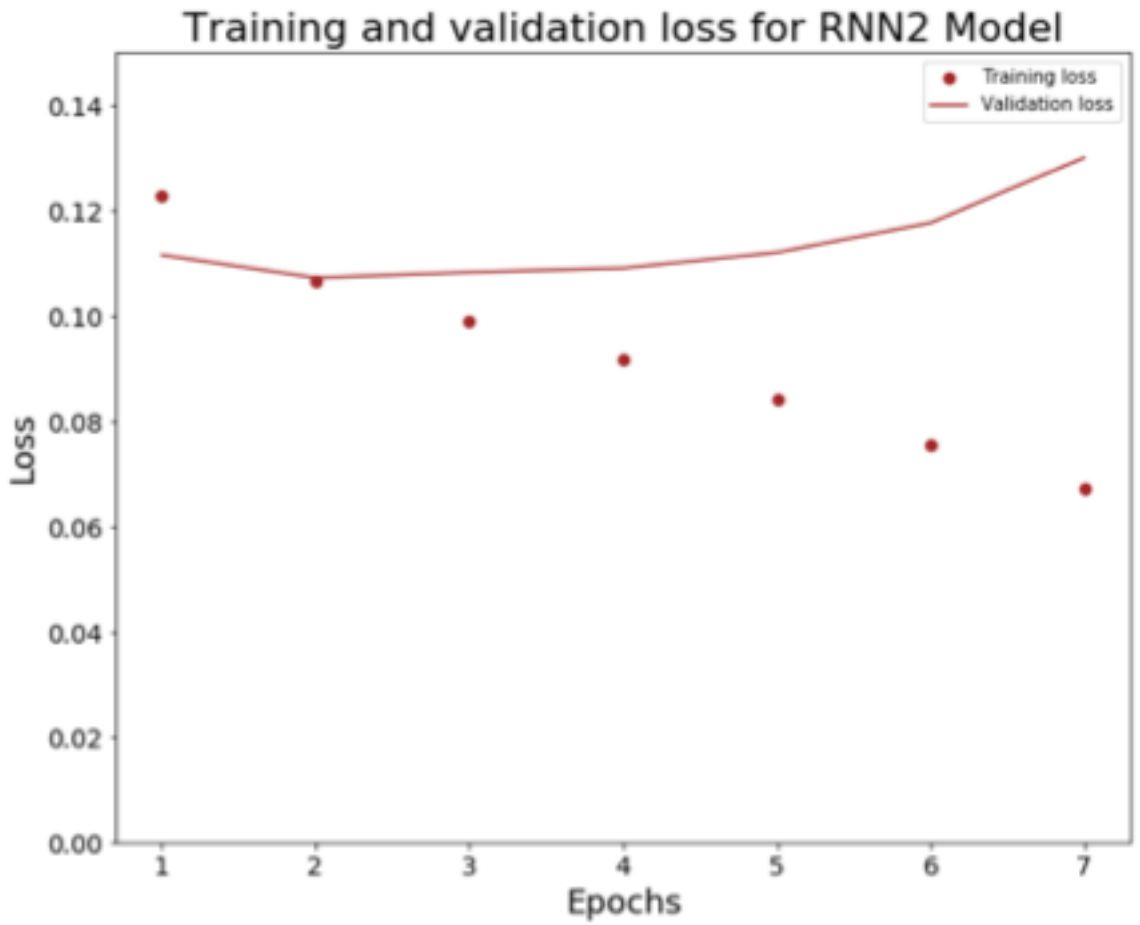
- With 1-Dimensional convolution layers, maxpooling layers and dropout
- Best Classification Threshold = 0.4
- Best F1 Score = 0.6642

Recurrent Neural Network 1



- With LSTM and dropout
- Best Classification Threshold = 0.4
- Best F1 Score = 0.6689

Recurrent Neural Network 2



- With bidirectional LSTM and dropout
- Best Classification Threshold = 0.4
- Best F1 Score = 0.6697

Final Model

Model	F1 Score
TF Naïve Bayes	0.5360
TF-IDF Logistic Regression	0.4603
Embedding CNN	0.6642
Embedding RNN1	0.6689
Embedding RNN2	0.6697

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

Future Research Recommendation

- Set up GPU :
  - Try to set up GPU for faster training of my Neural Networks if given more time and resources

- Try more RNN models:
  - Try implementing more RNN models with different layers such as the attention layer and the Spatial Dropout

Thank you!