Machine Learning Engineer Nanodegree

Capstone Project: Quora Insincere Questions Classifier

Karim Fateem

January 26th, 2019

I. Definition

Project Overview

A major challenge faced by social media and news websites is maintaining constructive non-toxic dialogue between users. The notion of what is considered constructive or non-toxic itself is a subjective determination and often varies across cultures. Social media companies spend significant resources on developing and enforcing community standards and are increasingly turning to machine learning to assist in applying those community standards.

One such company is Quora, a platform that empowers people to learn from one another. On Quora, people can ask questions and connect with others who contribute unique insights and quality answers. Quora has increasingly relied on ML in various parts of its products. Quora recently launched a Kaggle competition and dataset to classify insincere questions.

A flurry of developments in recent years have enabled the use of neural networks for sophisticated sentiment analysis and machine understanding. This project will explore using neural networks for classifying the insincere Quora questions data set.

Problem Statement

This project aims to classify whether a question asked on Quora is sincere or insincere. An insincere question is defined as a question intended to make a statement rather than to solicit helpful answers. Characteristics used for determining if a question is insincere are discussed in the “Datasets and Inputs” section below. The dataset includes labeled

data which allows for supervised training. Model accuracy and F1-score will be measured against the validation dataset.

The tasks involved are as follows:

1. Download and process the [Kaggle Quora Insincere Questions dataset](#).
2. Sample the data and explore patterns to assist with design decisions.
3. Preprocess the dataset by performing data transformations
4. Train a NN (Neural Network) to classify questions as sincere or insincere.
    a. analyze training performance
    b. tune model (architecture, normalization, regularization)
    c. tune training parameters (learning rate, class weights)
5. Analyze NN prediction performance. Update then repeat steps 3 - 5, until results are satisfactory.

Metrics

Given:

- *TP* = True Positives, *TN* = True Negatives, *FP* = False Positives, and *FN* = False Negatives.
- Precision = $\frac{TP}{TP + FP}$
- Recall (TPR) = $\frac{TP}{TP + FN}$
- Fall-out (FPR) = $\frac{FP}{FP + TN}$

The benchmark and solution models will be measured using the following metrics:

- F_β-score = $(1 + \beta^2) \times \frac{Precision \times Recall}{(\beta^2 \times Precision) + Recall}$

  A common value for β is 1, which produces an F1-score:

  F1-score = $2 \times \frac{Precision \times Recall}{Precision + Recall}$

  The F1-score measures the harmonic mean of precision and recall. An ideal score is 1.0, denoting perfect recall and precision.
  This is a primary metric useful for evaluating imbalanced datasets.The β parameter allows us to control the tradeoff of importance between precision and recall. $\beta<1$ focuses more on precision while $\beta>1$ focuses more on recall. A custom F_β could be useful metric customized for evaluating model performance

against production criteria. Assuming that questions flagged as insincere will go through human review, a custom β could be selected that balances human workload vs quality of unreviewed questions on the platform.

- ROC and AUC

  Receiver Operating Characteristic (ROC) is a curve plot of TPR against FPR. Area Under the curve for Receiver Operating Characteristic (AUROC), is a single metric that measures the area under a ROC curve. ROC will be used to classification threshold needs to be computed and used for predictions. It is useful to visually compare the optimal threshold against the When determining the optimal classification threshold it is useful to look at ROC.
  An ideal AUROC score is 1.0 and denotes a perfect true positive recall with no false positives.

II. Analysis

Data Exploration

Dataset description

Quora provides training data that includes the question that was asked, and whether the question was identified as insincere (target = 1). The ground-truth labels contain some amount of noise, since there is subjectivity when labelling a question as sincere or insincere. Some characteristics used to determine insincerity are listed below.

- Has a non-neutral tone
  - Has an exaggerated tone to underscore a point about a group of people
  - Is rhetorical and meant to imply a statement about a group of people
- Is disparaging or inflammatory
  - Suggests a discriminatory idea against a protected class of people, or seeks confirmation of a stereotype
  - Makes disparaging attacks/insults against a specific person or group of people
  - Based on an outlandish premise about a group of people
  - Disparages against a characteristic that is not fixable and not measurable
- Is not grounded in reality
  - Based on false information, or contains absurd assumptions

- Uses sexual content (incest, bestiality, pedophilia) for shock value, and not to seek genuine answers

The Quora Kaggle challenge also permits the use of the following word embeddings:

- GoogleNews-vectors-negative300 - https://code.google.com/archive/p/word2vec/
- glove.840B.300d - https://nlp.stanford.edu/projects/glove/
- paragram_300_sl999 - https://cogcomp.org/page/resource_view/106
- wiki-news-300d-1M - https://fasttext.cc/docs/en/english-vectors.html

The dataset can be downloaded from the following Kaggle competition link: https://www.kaggle.com/c/10737/download-all

Data samples

The following tables contain *sincere* and *insincere* question samples.

Sincere questions

qid	question_text	target
d2fcceacb58a80a3897e	What do people in a car in the winter turn the heat up to such a high temperature that passengers feel sick?	0
d3b05c76122e8e8becdb	I recently told someone I had feelings for them, but they revealed they were in a relationship and we just stopped talking. Is it wrong to shut people out if there's not much left to say between the two of you?	0
f0b768c1c4c26cf3072e	Can I respect a friend's autonomy while preventing their suicide?	0
97bf61ca5922ff1c4e5d	How is timex blink?	0
e32dadef8798cce2e813	Can we learn art from YouTube?	0
6a30e0225606f6f2d6e3	How can you prevent getting bumps on your stretch marks?	0
5f32692f26079c2ba9b	What are the strengths and weaknesses of TOGAF in the	0

4	implementation of Cloud computing?	
30482b2d8d498f0d728f	How do I suspend virtual machine from remote machine?	0
dc3202dbfb562a494c14	Will BTS add a new member next year?	0
a46369a25a809b41d40a	How can I protect my brain against Alzheimer?	0

Insincere questions

qid	question_text	target
46e49fc5bc2d93fc6409	Why don't the BJP and RSS supporters understand that they are being brainwashed and quite tactfully their focus is being shifted from real issues like rape and DEVELOPMENT to religion and casteism?	1
a42a66d06d159c865a9c	Indian christen say every Hindu god is a Satan while Western are believing hindusm, how Indian Christen can teach such hatefull things?	1
1bf82de1e3eaa8051166	If JFK wasn't a president would he have been a envelope adhesive taste tester?	1
41506c8988c1434a3863	Why does Quora allow a topic named “Palestine” to exist when there’s no such a thing as “Palestine”?	1
37be354a7f34190e5ac4	Does Harry Potter know a spell to castrate his enemies?	1
6e3724642ed7385b3ad2	Should we be using children to push policy changes?	1
f263e125f431dc1f6b1d	If a victim that got raped really enjoyed it and would love to do it again, is it bad?	1
a8139d34c569389bb8b0	Vast majority of Nepali people bitterly hate India despite the fact that Hinduism is the common religion among them. Why?	1
98b105aa54e129e59813	If Muslims think ISIS is doing wrong, then why don’t they oppose it openly? As they have supported Gaza against Israel, now they should speak against ISIS.	1
efa7300fdd3e1a3d9883	Can I get black tar delivered discreetly to me?	1

Dataset Analysis

Target Label Balance

The following table summarizes the ***target*** label ratios in the training data.

Target class	Count	Ratio
Sincere	1225312	0.94
Insincere	80810	0.06

The data is highly imbalanced, with *sincere* questions representing 94% of targets, and *insincere* questions representing 6%. Given this observation, the following normalization techniques may yield improvements during training:

- oversampling (e.g. using SMOTE)
- Undersampling
- using class weights to bias training towards insincere questions

Question Text Statistics

The table below summarizes various statistics computed for the ***question_text*** feature taken from training data:

character length	**min**	**max**	**mean**	**std**
Test	11	588	70.46	38.73
Training: all	1	1017	70.68	38.78
Training: sincere	5	752	66.87	36.74
Training: insincere	1	1017	98.06	55.19
uppercase to total characters ratio	**min**	**max**	**mean**	**std**
Test	0	0.78	0.05	0.04
Training: all	0	1	0.05	0.04
Training: sincere	0	0.9	0.05	0.04
Training: insincere	0	1	0.04	0.04
word count	**min**	**max**	**mean**	**std**
Test	2	87	12.75	7.01
Training: all	1	134	12.8	7.05
Training: sincere	2	134	12.51	6.75
Training: insincere	1	64	17.28	9.57
uppercase to total words ratio	**min**	**max**	**mean**	**std**
Test	0	0.94	0.03	0.06

Training: all	0	1	0.03	0.06
Training: sincere	0	1	0.04	0.06
Training: insincere	0	1	0.02	0.05
question mark count	**min**	**max**	**mean**	**std**
Test	0	5	1.06	0.26
Training: all	0	10	1.06	0.26
Training: sincere	0	10	1.05	0.25
Training: insincere	0	8	1.13	0.39
ends with question mark	**min**	**max**	**mean**	**std**
Test	0	1	0.98	0.15
Training: all	0	1	0.98	0.15
Training: sincere	0	1	0.98	0.14
Training: insincere	0	1	0.94	0.23
exclamation mark count	**min**	**max**	**mean**	**std**
Test	0	3	0.0012	0.0381
Training: all	0	5	0.0017	0.0468
Training: sincere	0	5	0.0014	0.0408
Training: insincere	0	5	0.0075	0.1001
inappropriate word count	**min**	**max**	**mean**	**std**
Test	0	2	0.0011	0.0357
Training: all	0	4	0.0012	0.0362
Training: sincere	0	2	0.0005	0.0224
Training: insincere	0	4	0.012	0.1161

Below is a summary of observations:

- The training and test data share similar metrics and appear to be from the same distribution.
- Sincere questions appear to be more concise with fewer total characters than inscienre questions. The ***target*** variable depends on the *character length* and *word count* variables. The *character length* and *word count* variables are interchangeable.

- Use of capitalized letters to signal shouting is often observed on the internet. Surprisingly, that convention does not seem to apply to the Quora data set. In fact, *sincere* questions have slightly more uppercase characters/words on average. The ***target*** variable slightly depend on the frequency of uppercase characters/words. Sampling has also shown that *sincere* questions contain more uppercase acronyms than *insincere* questions.
- Questions without question marks, or containing 3 or more question marks may indicate an *insincere* question. The ***target*** variable depends on the on the *question mark count* variable.
- Text that ends with a question mark tends to be more *sincere*. The ***target*** variable depends on the *ends with question mark* variable.
- Questions with more exclamation marks tends to be more *insincere*. The ***target*** variable depends on the *exclamation count* variable.
- Questions with more inappropriate words tend to be more insincere. The ***target*** variable depends on the *inappropriate word count* variable.

The above observations suggest that a number of engineered features can be useful inputs to the model. There is also an opportunity to generate sophisticated features using Natural Language Processing (NLP) techniques such as Part-of-Speech (POS) Taggers to analyze sentence structure. This could be useful for assigning context-based weights to words and scoring sentence grammar.

Word Embedding Coverage

The following table summarizes coverage of the data set vocabulary by each of the supplied word embeddings.

The *% of unique words* column is a measurement of unique words in the vocabulary with corresponding word vectors in each embedding.

The *% of question text* column is a measurement of the total words with corresponding word vectors in each embedding.

	Coverage (Training + Test)	
Embedding	% of unique	% of question

	words	text
glove.840B.300d	32.91%	88.16%
paragram_300_sl999	19.42%	72.21%
wiki-news-300d-1M	29.77%	87.66%
GoogleNews-vectors-negative300	24.05%	78.75%

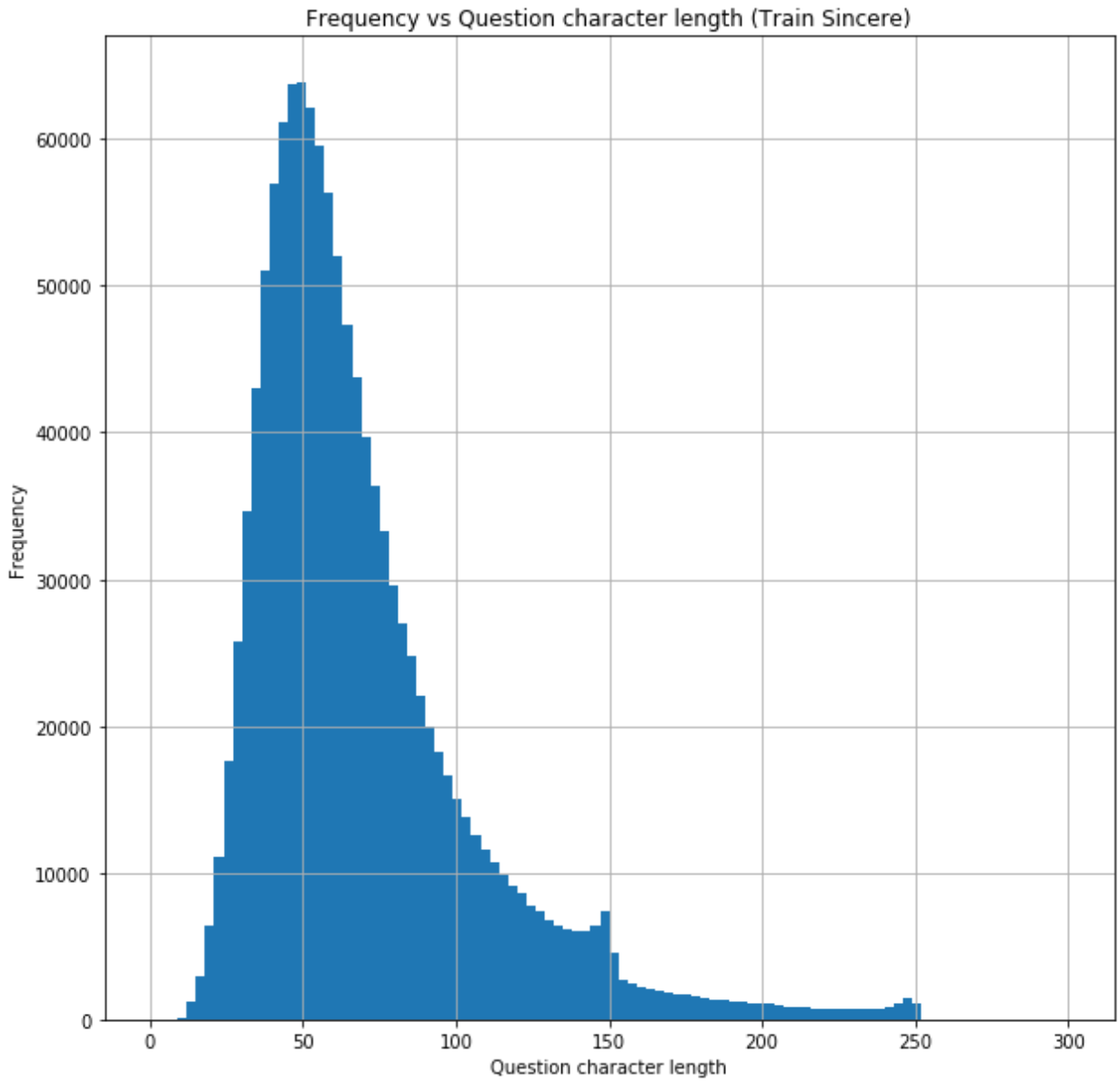
Embeddings greatly accelerate training and simplify the NN architecture. Sampling showed that contractions, American vs British spelling variations, punctuations and casing all contribute to low unique word coverage. There is an opportunity to increase training performance by improving unique word coverage data preprocessing.

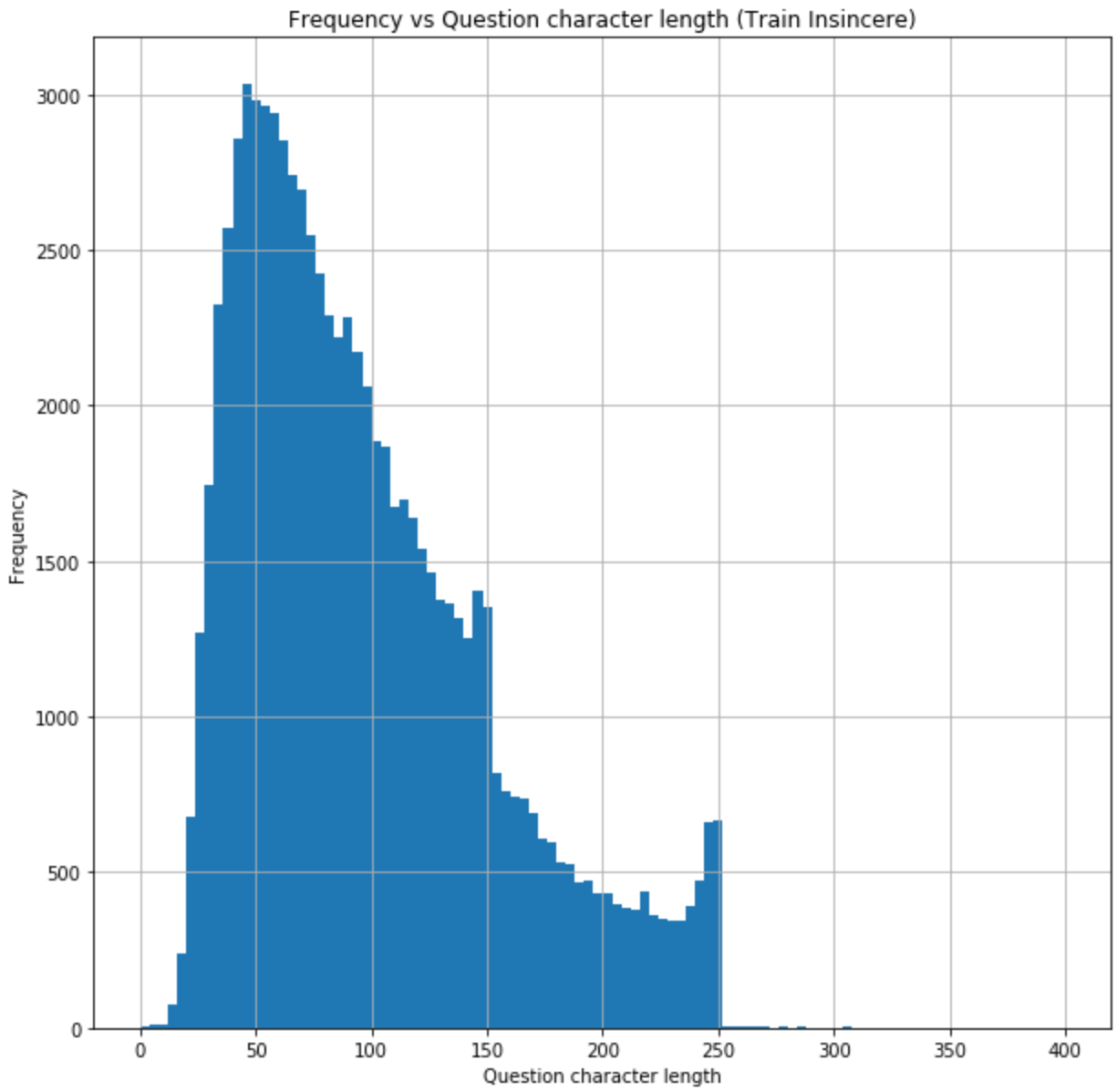
Exploratory Visualization

Below are a number of histograms to support observations in the “Data Analysis: Question Text Statistics” section.

Question character length

The following pair of histograms illustrate the frequency of the *sincere* and *insincere* questions by character length.

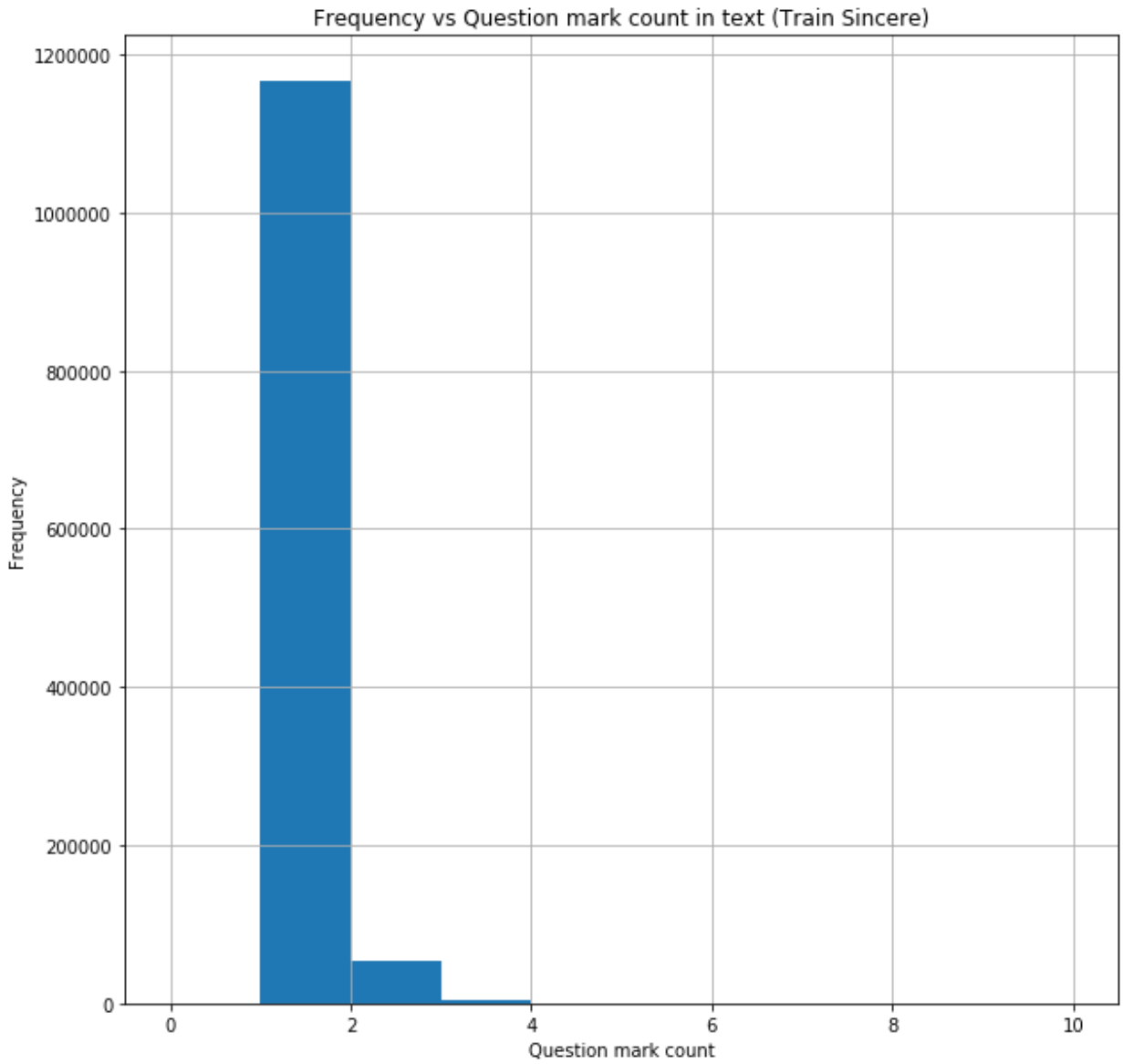

Frequency vs Question character length (Train Sincere)
Frequency
Question character length
0
10000
20000
30000
40000
50000
60000
0
50
100
150
200
250
300

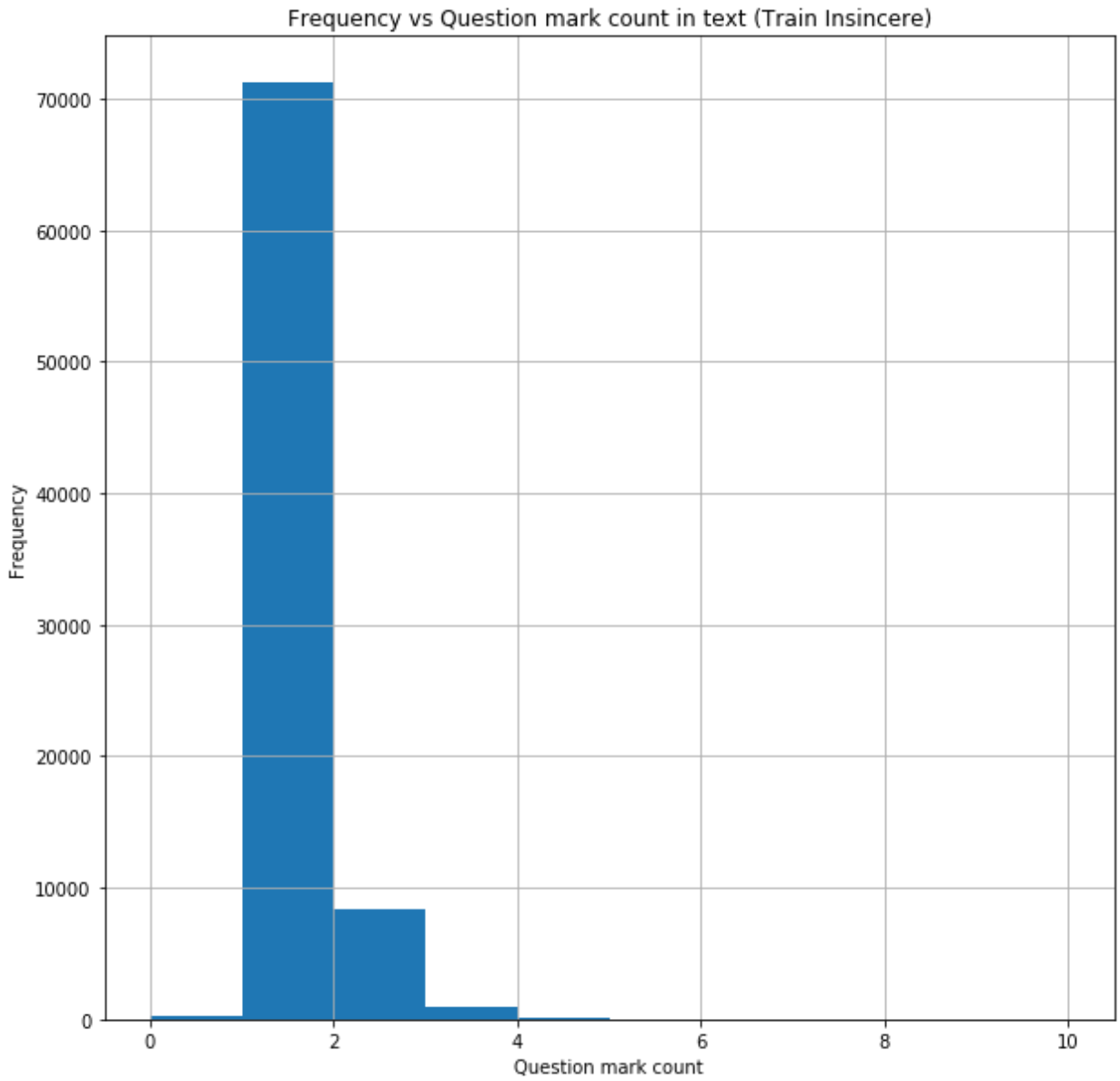


The histograms show the ratio of long text to short text is higher in *insincere* questions when adjusted by class population size.

Question mark count

The following pair of histograms illustrate the frequency of the *sincere* and *insincere* questions by question mark count.


Frequency vs Question mark count in text (Train Sincere)
Frequency
Question mark count
0
200000
400000
600000
800000
1000000
1200000
0
2
4
6
8
10



The histograms show that *insincere* questions have a higher occurrence of question marks when adjusted for class population size.

Algorithms and Techniques

The problem is tackled using a binary classifier using Recurrent Neural Networks (RNN) as follows:

- The network architecture is comprised of multiple Gradient Recurrent Units (GRU) with pre-trained embedding layers as inputs.
- All 4 pre-trained word embeddings are used.

- GRU outputs are concatenated in a merge layer.
- The classifier outputs a single non-discrete probability.
- An optimal threshold is used to convert the output probability to a binary prediction.
- Regularization techniques such as dropout and L2 kernel regularizers are used.
- Batch normalization is used to speed up training.
- Data preprocessing is used to:
    - increase embedding coverage to reduce misclassifications due to Out of Vocabulary (OOV)
    - synthesize useful input features that could improve network performance
- Experimentation with the following to improve network performance:
    - the number of network layers to improve performance
    - learning rate parameters
    - training parameters such as epochs and batch sizes

Benchmark

This Kaggle kernel contains a 2D CNN model with a pretrained glove.840B.300d word embedding layer. The public test F1-score is 0.671 and a local validation F1-score of 0.6744 at a threshold of 0.3. For reference, the highest F1-score on the Kaggle leaderboard is 7.200. This model was selected for its simplicity, relatively high F1-score and quick training time.

III. Methodology

Data Preprocessing

Data exploration determined that transformations could be used to improve embedding coverage. To improve coverage the ***question_text*** feature was transformed to ***treated_question*** as follows:

- **Converted questions to lowercase:** many embeddings did not have vectors for uppercase word.
- **Expanded contractions:** many embeddings did not have vectors for common of contractions. This was addressed by expanding common contractions using a mapping found on wikipedia, e.g. “can’t” to “cannot”.

- **Remapped punctuations and mathematical symbols:** certain punctuations and symbols did not have representation in embeddings but there were appropriate replacements. E.g. “π” to “pi”.
- **Corrected common misspellings:** sampled top misspelled words and replaced the text with the correct spelling.

The table below summarizes coverage improvements due to these transformations:

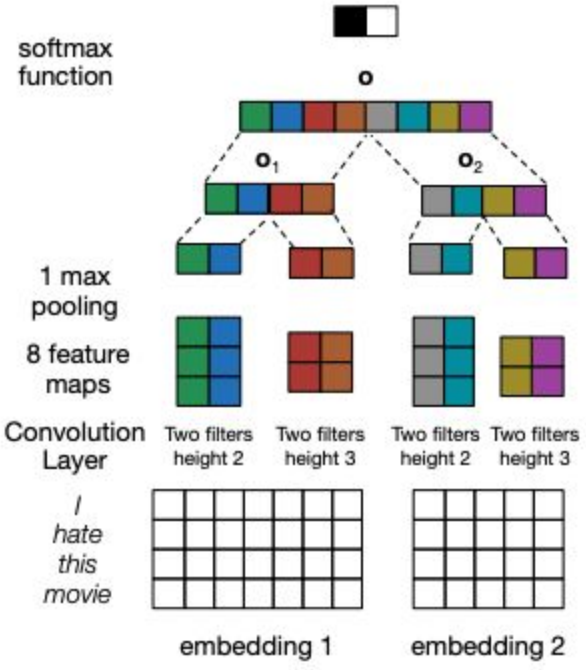
	Coverage (Before)		Coverage (After)	
Embedding	% of unique words	% of question text	% of unique words	% of question text
glove.840B.300d	32.91%	88.16%	63.10%	99.39%
paragram_300_sl999	19.42%	72.21%	74.06%	99.64%
wiki-news-300d-1M	29.77%	87.66%	48.05%	98.81%
GoogleNews-vectors-negative300	24.05%	78.75%	36.68%	78.84%

Implementation

Architecture

The implementation took a hybrid approach:

- Built on the technique of outlined in the MGNC-CNN paper which used multiple embeddings and merged outputs from each in a penultimate layer.



* Image source: MGNC-CNN: A Simple Approach to Exploiting Multiple Word Embeddings for Sentence Classification

- Replaced the CNN layer with a GRU based on historic strong performance of GRUs outlined in the this twitter analysis paper and HAN paper.

Complications

- The accuracy metric was not a good predictor of NN performance due to a highly imbalanced data set. F1-scores and confusion matrices were used to evaluate performance instead.
- Keras 2.0 removed f1-score metrics from the training stage summary. Custom callbacks needed to be implemented to emit f1-scores during training.
- Scikit learn implementations of F1-score, recall and precision functions are incompatible with Keras Tensorflow callbacks. Custom implementations were developed and used.
- Embeddings have a large memory footprint. High memory compute instances were provisioned to overcome this in development.
- Using a Telsa K80 GPU, the selected NN architecture required 20 minutes of wall time per training epoch. Upgraded to Tesla V100 GPUs during more intensive training cycles.

- The Kaggle Quora kernel competition uses a 14GB memory instance with a K80 GPU with a 2 hour limit on training and prediction time. To overcome this, optimizations were made to better utilize memory consumption. Large objects were flushed to disk when not used.

Refinement

Extensive iteration and experimentation was conducted to improve NN performance. Variables included:

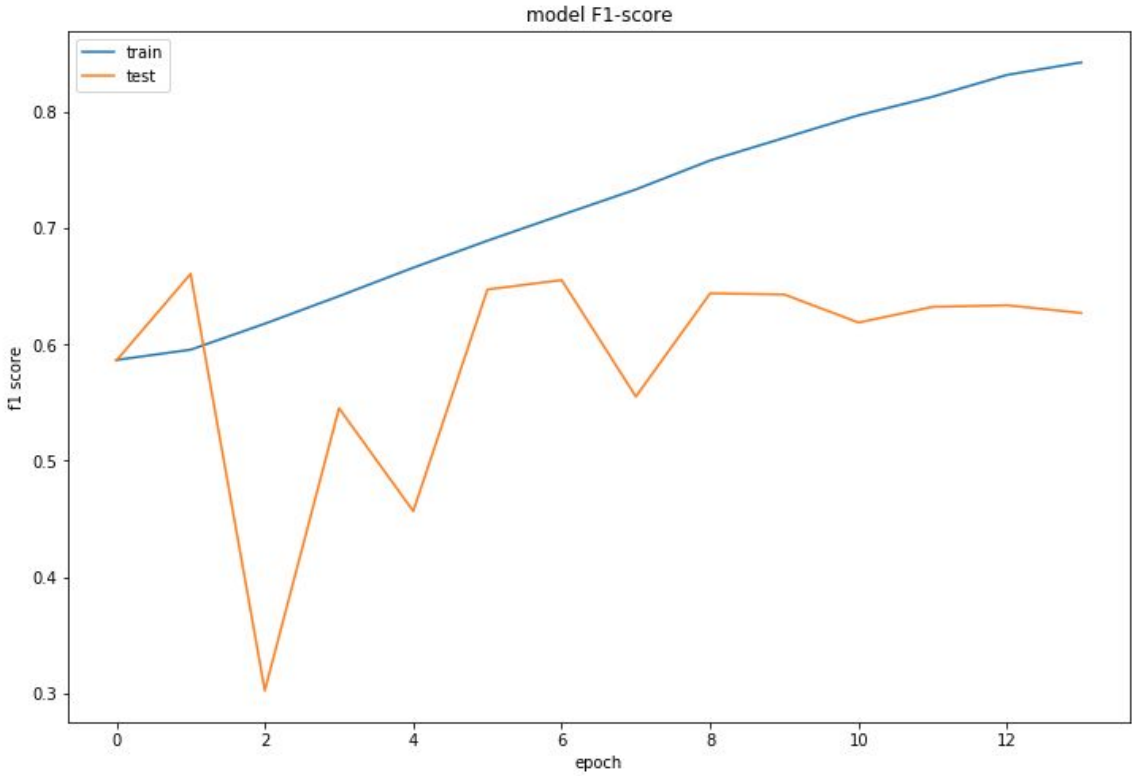
- the # of pre-trained embeddings used (ranging from none to 4).
- whether or not embeddings should be retrained.
- whether or not to apply data transformation techniques to improve vocabulary coverage.
- the maximum word dictionary size and sentence length to use for generating an embedding mapping.
- whether or not to apply class weights to boost learning on the minority data class.
- architecture layers.
- batch normalization.
- dropout and L2 regularization thresholds.
- the number of epochs and batches to use for training.
- the optimizer learning rate, early stopping and patience.

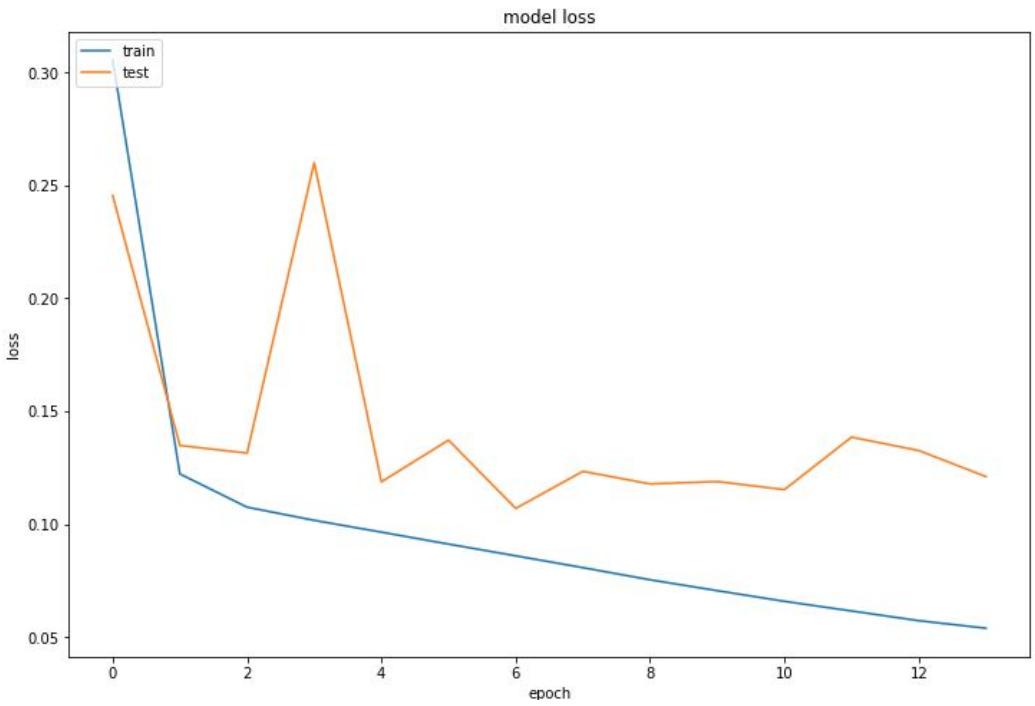
Below is a table that summarizes the different variables and F1-score outcomes.

Iteration	Arch	# pretrained emb	emb trainable	Data cleaning	MAX WORDS	MAX SEQ LENGTH	class weights	epoch	batch	optimal f1-score	optimal threshold	F1-score at 0.35	observation
benchmark	2D CNN	1	TRUE	FALSE	40000	70	FALSE	2	256	0.6744	0.30	0.6699	
capstoneV1	1GRU	0	TRUE	FALSE	50000	100	FALSE	2	256	0.6374	0.35	0.6374	
capstoneV2	1GRU	1	FALSE	FALSE	50000	100	FALSE	2	256	0.6417	0.30	0.6374	Improvement using Glove embedding
capstoneV3	4GRU	4	FALSE	TRUE	50000	100	TRUE	3	256	0.6729	0.23	0.6505	Imrpovement using class wieght + 4 pre-training embeddings + data cleaning
capstoneV4	4GRU	4	FALSE	TRUE	40000	70	TRUE	3	256	0.6773	0.38	0.6760	F1-improvement over previous models.
capstoneV5	1GRU	1	TRUE	TRUE	50000	100	TRUE	2	256	0.6736	0.18	0.6303	
capstoneV6	4GRU	4	TRUE	TRUE	50000	100	FALSE	2	256	0.6751	0.18	0.6322	Improvement over v3
capstoneV7	4GRU	4	TRUE	TRUE	50000	100	TRUE	2	256	0.6704	0.39	0.6676	
capstoneV8	4GRU	4	TRUE	TRUE	40000	70	TRUE	2	256	0.6738	0.29	0.6700	
capstoneV9	4GRU	4	TRUE	TRUE	40000	70	FALSE	2	256	0.6728	0.25	0.6572	

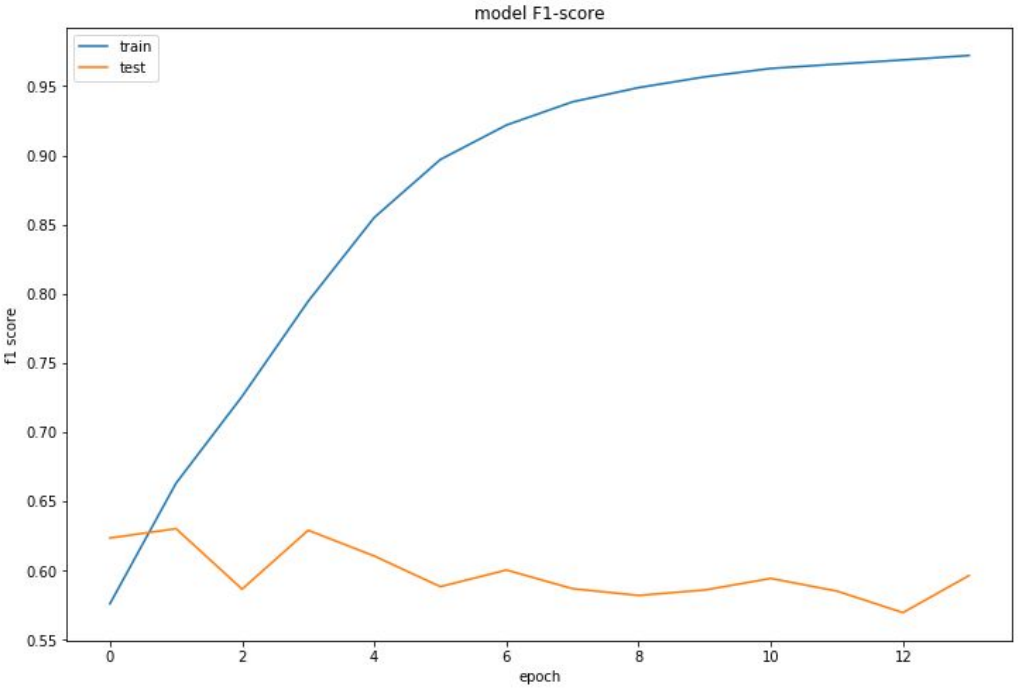
An initial solution was found but underperformed against the benchmark. By iteratively tuning the above variable a model was found, ***capstoneV4***, that outperformed the benchmark on validation data.

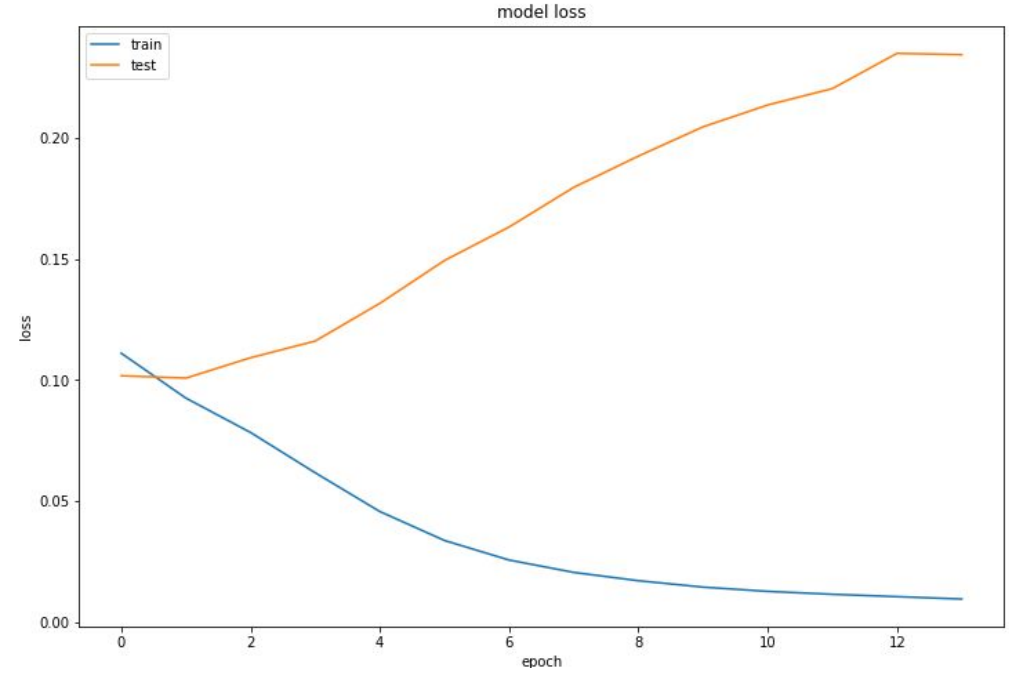
An Adam optimizer was used during training. The network stopped showing training improvement after 1 epoch and prediction f1-scores fluctuated noticeably when using the default learning rate.



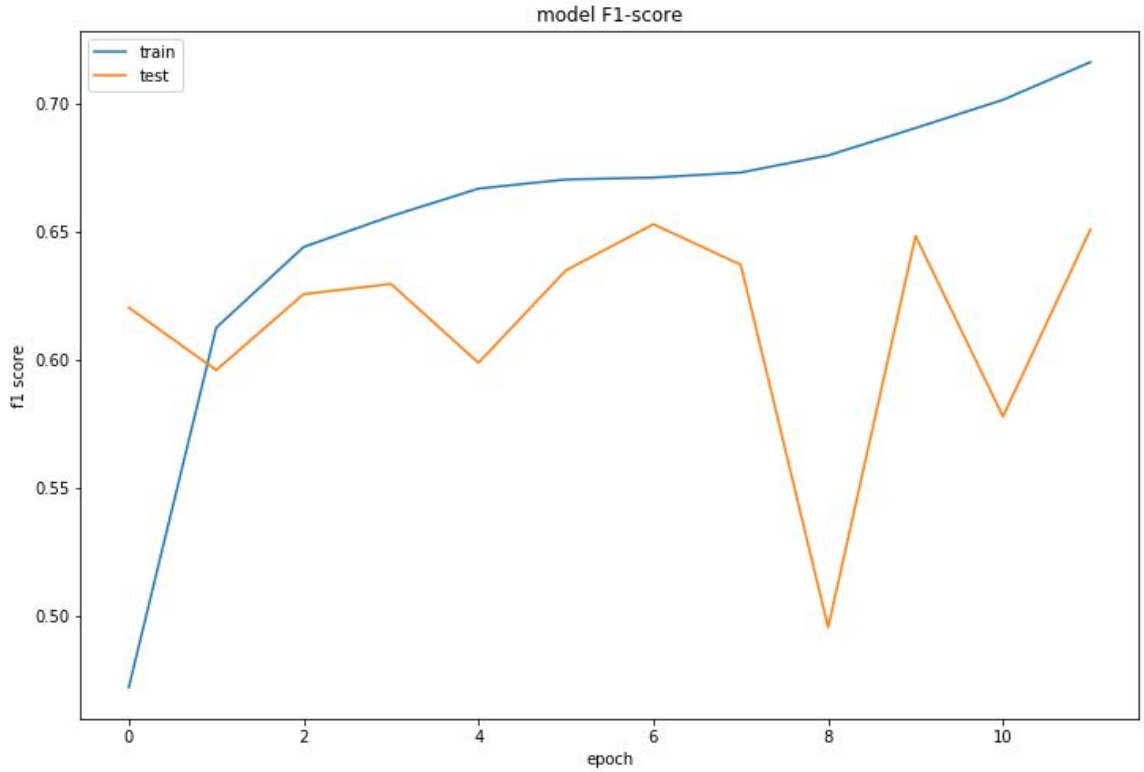


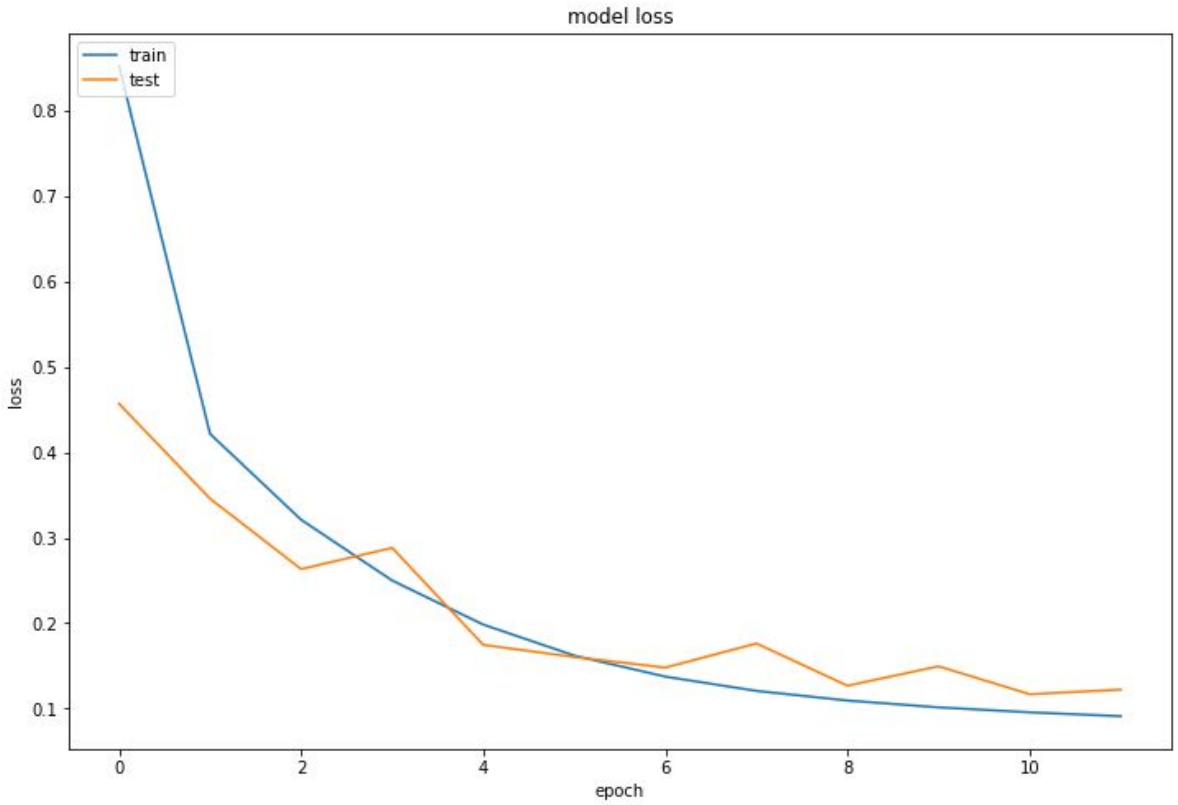
The benchmark model showed even more divergent results using the same optimizer, learning rate, epochs and batch size:


model F1-score
train
test
f1 score
epoch
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
0
2
4
6
8
10
12


model loss
train
test
loss
epoch
0.00
0.05
0.10
0.15
0.20
0
2
4
6
8
10
12

By lowering the learning rate from 0.001 to 0.0001 and using early stopping after 5 epochs of patience fluctuations between train and validation predictions improved but final prediction F1-score did not change significantly.


model F1-score
train
test
f1 score
epoch
0.70
0.65
0.60
0.55
0.50
0
2
4
6
8
10


model loss
train
test
loss
epoch
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
2
4
6
8
10

IV. Results

Model Evaluation and Validation

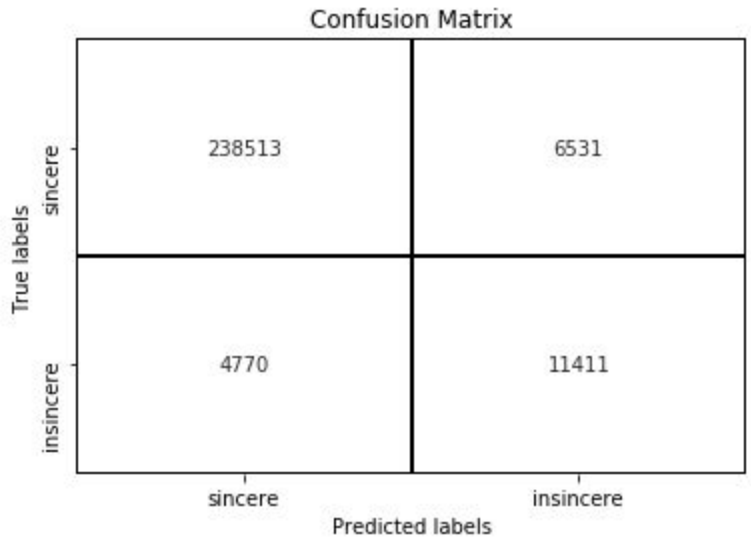
The final model and hyperparameters with strongest F1-score was selected. The base model contained a GRU with input from an embedding layer. More GRUs and embeddings were iteratively added and the results merge in a penultimate layer. Dropout layers, regularilzers and batch normalization were introduced and tuned to improve the network's ability to generalize.

The model was trained with a 20% hold out for validation. The model exhibited F1-score and loss fluctuations on the validation set rather than smoothly converging. The benchmark model performed similarly.

A classification report for the model shows the following recall and precision values for sincere and insincere questions:

	precision	recall	f1-score	support
sincere	0.98	0.97	0.98	245044
insincere	0.64	0.72	0.68	16181
avg/total	0.96	0.96	0.96	261225

A confusion matrix for validation predictions is below:



Overall the model did well classifying sincere and insincere questions with some Type I and II errors. The Quora dataset is know to contain noise. Samples of Type I and Type II misclassifications below demonstrate this noise:

```
[257]: false_positive_extreme.sample(10)
```

[257]:

	target	prediction	pred_val	question_text_rebuilt
41469	0	1	0.927667	why do some liberals ignore israelites ' sufferings and struggles for a free , democratic and independent israel ? how can they support those cruel islamists ?
224583	0	1	0.741720	do koreans use water to clean their bottom ? do they have bidet in toilet as muslim countries do or clean with toilet paper only ?
189608	0	1	0.750678	are the evangelicals in the usa racist , fascist or what ? and what is their relationship with gun lobby group ?
23054	0	1	0.974254	in this enlightened nation people believe sex offenders should be castrated , do you ? how about other criminals , should they be mutilated ad well ?
103058	0	1	0.782203	how the hell do most asian women get those big , strong , and beautiful muscular legs without exercising ?
130397	0	1	0.758577	did prophet muhammad made any woman swallow his semen ?
244774	0	1	0.773128	why do half of americans believe trump colluded with russia and half believe he was framed by the department of justice ?
149802	0	1	0.934981	why do black people in general face more racism than white people based on skin color ?
234506	0	1	0.767865	why should young south africans stay in the country ? considering the countries recent junk status .
102520	0	1	0.808862	why do not chinese officials suffer from incompetence despite the fact that china has a " repressive " system ?

```
false_negative_extreme.sample(10)
```

	target	prediction	pred_val	question_text_rebuilt
258563	1	0	0.142188	do you listen to the lies of god ?
93293	1	0	0.056738	what are some good things done by isis ?
177430	1	0	0.233607	i live like i am homeless , yet i have 2 million dollars sitting in the bank and a condo paid for . i basically just wander around all day doing nothing . is not life great ?
184247	1	0	0.237611	how does being in the same building that someone died in make your gun intelligence immediately ?
226969	1	0	0.050864	is it possible to sue my parents for me ?
69961	1	0	0.142670	why do people try to save suicidal persons they dont even know ?
26676	1	0	0.094318	why is my urine yellow instead of the usual opaque red ?
183571	1	0	0.015124	who wants the best dissertation writing service in usa ?
126414	1	0	0.140942	what does your butt look like ?
127575	1	0	0.041702	can you get a speeding ticket for traveling faster than the speed of light ? what cop is going to catch you ?

Justification

The model slightly outperformed the benchmark by 0.43% on validation data, with respective F1-scores of 0.6773 and 0.6744, and respective thresholds of 0.38 and 0.3. At a fixed threshold 0.35 the model had a F1-score of 0.6760 compared to the benchmark F1-score of 0.6699.

The model was submitted in the Kaggle Quora Insincere Question competition and slightly underperformed by 0.99% with a F1-score of 0.664 when compared to the benchmark F1-score of 0.671.

Overall, the model F1-score is comparable to the benchmark over a large number of runs, but the model is more complex and slower to train.

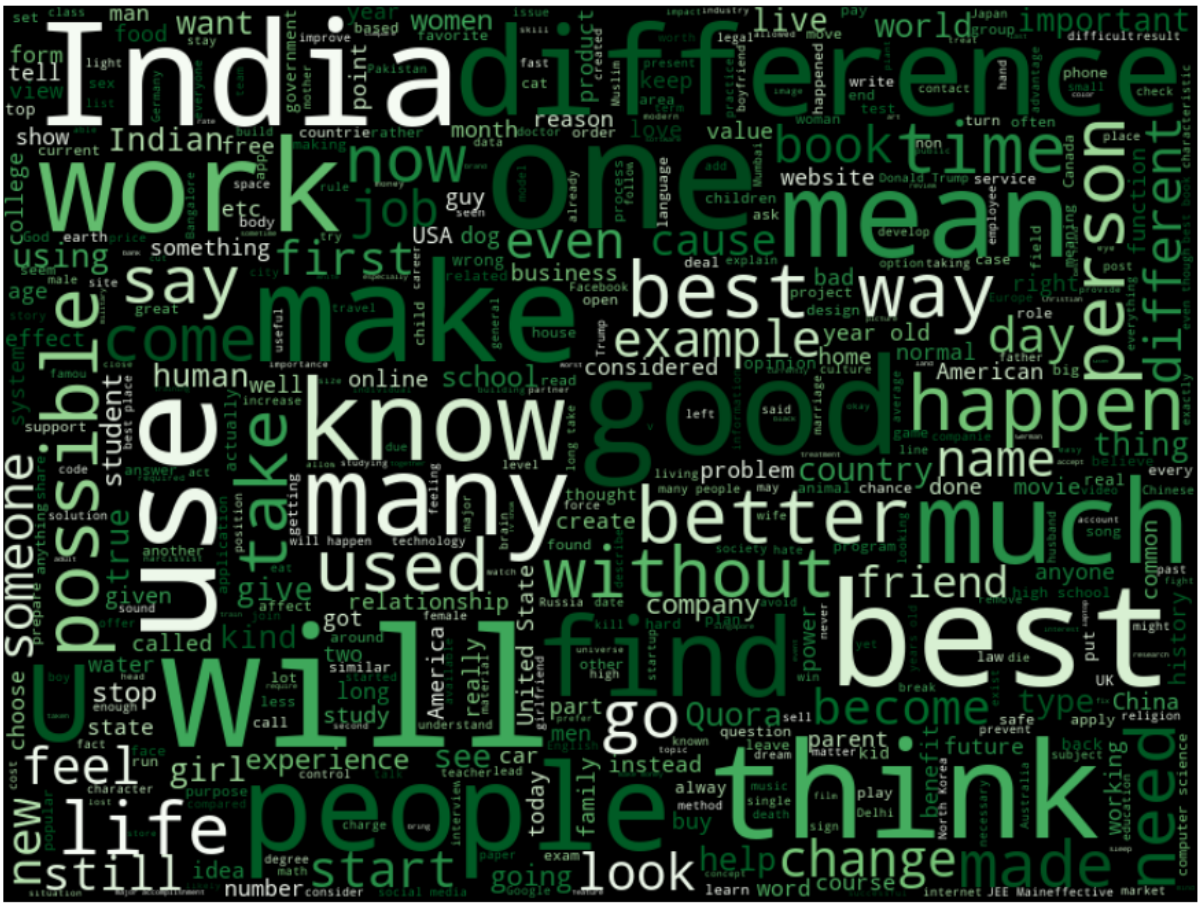
The model could be trusted to operate in an assistive capacity as an aid to human review. Its classifications of sincere questions but it will incorrectly let some inscinere questions through. This is likely acceptable given the already low occurrence of the minority class. Questions classified as insincere should go through additional human review since some will have been misclassified.

V. Conclusion

Free-Form Visualization

Below are word cloud visualization of *sincere* and *insincere* ***question_text*** feature.

Sincere question word cloud



Insincere question word cloud

These word clouds provide a frequency-adjusted representation of the 500 most used words in *sincere* and *insincere* questions. While there are some common words, some distinct trends are also apparent. These qualitative visuals demonstrate an opportunity to classify based on word frequency. Given there are distinct word clusters in each class, it was possible to use word embeddings to create associations and effectively generalize predictions to previously unseen questions.

Reflection

This project explored training a neural network for sentiment analysis using multiple word embeddings. Doing so entailed:

- setting up an environment in Google Cloud Platform to facilitate rapid and iterative experimentation using GPUs
- transferring in the Quora Insincere Question dataset
- analyzing the data set for:
    - insights to inform the solution design
    - anomalies that could pose challenges
- designing and iterating on a RNN architecture using GRUs in Keras Tensorflow
- making trade-offs between input dimensionality and network simplicity
- experimenting with regularization techniques such as using dropouts, regularizers and batch normalization
- experimenting with normalization techniques to deal with data imbalance such as SMOTE oversampling and using class weights
- tuning hyperparameters to improve training
- experiencing first-hand challenges with training a neural network on real-world dataset

This was my first attempt at sentiment analysis using machine learning. I was excited to get hands on experience with neural networks given recent breakthroughs with word embeddings. The results were very promising and I'm excited to continue refining my model.

Improvement

There are number of improvements that can likely produce improvements:

- using a momentum aware optimizer to deal with oscillations during learn.
- using feature transformation to synthesize new model inputs. Simple features were identified in the data analysis section and more complex features could be generated using NLP libraries to analyze POS.
- learning and using attention and capsules in conjunction with GRUs to build a HAN.
- revisiting the 2D CNN benchmark and iterating on that model as an alternate design.
- exploring how to use distributed GPUs to further speed up training.

The project has helped me develop an understanding of how to use ML for real-world complex tasks and I am excited to continue learning and applying ML to everyday challenges.

Works Cited

Quora Insincere Question Challenge | Kaggle, www.kaggle.com/c/quora-insincere-questions-classification.

Improve Your Score with Some Text Preprocessing | Kaggle, www.kaggle.com/theoviel/improve-your-score-with-some-text-preprocessing.

10153181162182282. “Illustrated Guide to LSTM's and GRU's: A Step by Step Explanation.” *Towards Data Science*, Towards Data Science, 24 Sept. 2018, towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21.

Hu, et al. “Listening to Chaotic Whispers: A Deep Learning Framework for News-Oriented Stock Trend Prediction.” *[Astro-Ph/0005112] A Determination of the Hubble Constant from Cepheid Distances and a Model of the Local Peculiar Velocity Field*, American Physical Society, 16 Mar. 2018, arxiv.org/abs/1712.02136.

Nabi, Javaid, and Javaid Nabi. “Machine Learning - Word Embedding & Sentiment Classification Using Keras.” *Towards Data Science*, Towards Data Science, 4 Oct. 2018, towardsdatascience.com/machine-learning-word-embedding-sentiment-classification-using-keras-b83c28087456.

“Parlio.” *Wikipedia*, Wikimedia Foundation, 16 Dec. 2018, en.wikipedia.org/wiki/Parlio.

“Quora - A Place to Share Knowledge and Better Understand the World.” - *Quora*, www.quora.com/.

Sebastian Ruder. “Word Embeddings in 2017: Trends and Future Directions.” *Sebastian Ruder*, Sebastian Ruder, 24 Oct. 2018, ruder.io/word-embeddings-2017/index.html#oovhandling.

Synced. “A Brief Overview of Attention Mechanism – SyncedReview – Medium.” *Medium.com*,

Medium, 25 Sept. 2017, medium.com/syncedreview/a-brief-overview-of-attention-mechanism-13c578ba9129.

Tien, et al. “Sentence Modeling via Multiple Word Embeddings and Multi-Level Comparison for Semantic Textual Similarity.” *[Astro-Ph/0005112] A Determination of the Hubble Constant from Cepheid Distances and a Model of the Local Peculiar Velocity Field*, American Physical Society, 21 May 2018, arxiv.org/abs/1805.07882.

Yang, Zichao, et al. “Hierarchical Attention Networks for Document Classification.” *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, doi:10.18653/v1/n16-1174.

Zhang, Ye, et al. “MGNC-CNN: A Simple Approach to Exploiting Multiple Word Embeddings for Sentence Classification.” *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, doi:10.18653/v1/n16-1178.

Zhao, Chuanjun, et al. “Deep Transfer Learning for Social Media Cross-Domain Sentiment Classification.” *Communications in Computer and Information Science Social Media Processing*, 2017, pp. 232–243., doi:10.1007/978-981-10-6805-8_19.