Detection of Question Sincerity on Online Forums

Kate Bosshart & Siyao Peng, Georgetown University

ABSTRACT

Research was performed to determine whether sincerity of questions posted on online question and answer (Q&A) forums could be systematically and accurately identified. Utilizing a training set of labeled questions sourced from Quora, the data was augmented and cleaned. From there, three neural networks were developed and trained, and then tested on a held-out portion of the dataset. Ultimately, it was determined that an Attention Model was able to successfully flag insincere questions in the training data.

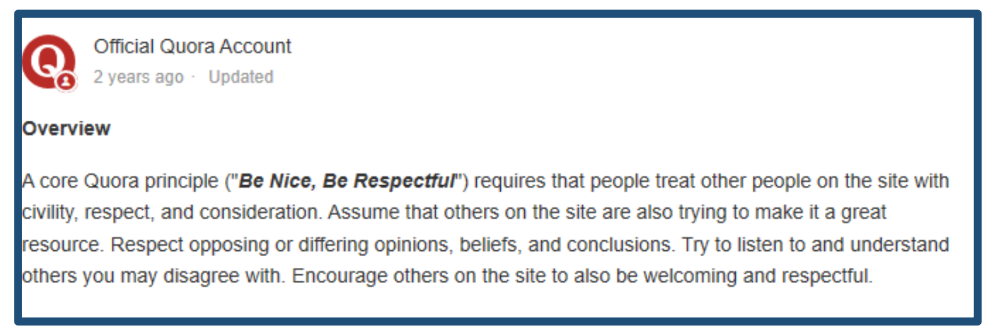
INTRODUCTION

OVERVIEW & PROBLEM STATEMENT

Our project seeks to **develop an approach to correctly identify the sincerity of questions posted on online Q&A forums,** utilizing those posted on the website Quora as a way of handling the problem of toxic internet content. We leverage multiple techniques to interpret the content of the questions and detect toxic and divisive questions, including using embeddings for text preprocessing, and text augmentation.

GOAL Our goal is to **create a scalable solution to the problem of insincere questions online**.

DEFINING SINCERITY



Official Quora Account
2 years ago · Updated

Overview

A core Quora principle ("***Be Nice, Be Respectful***") requires that people treat other people on the site with civility, respect, and consideration. Assume that others on the site are also trying to make it a great resource. Respect opposing or differing opinions, beliefs, and conclusions. Try to listen to and understand others you may disagree with. Encourage others on the site to also be welcoming and respectful.

Quora defines an insincere question as one founded on false premises with out intention of seeking helpful answers, potentially signified by:

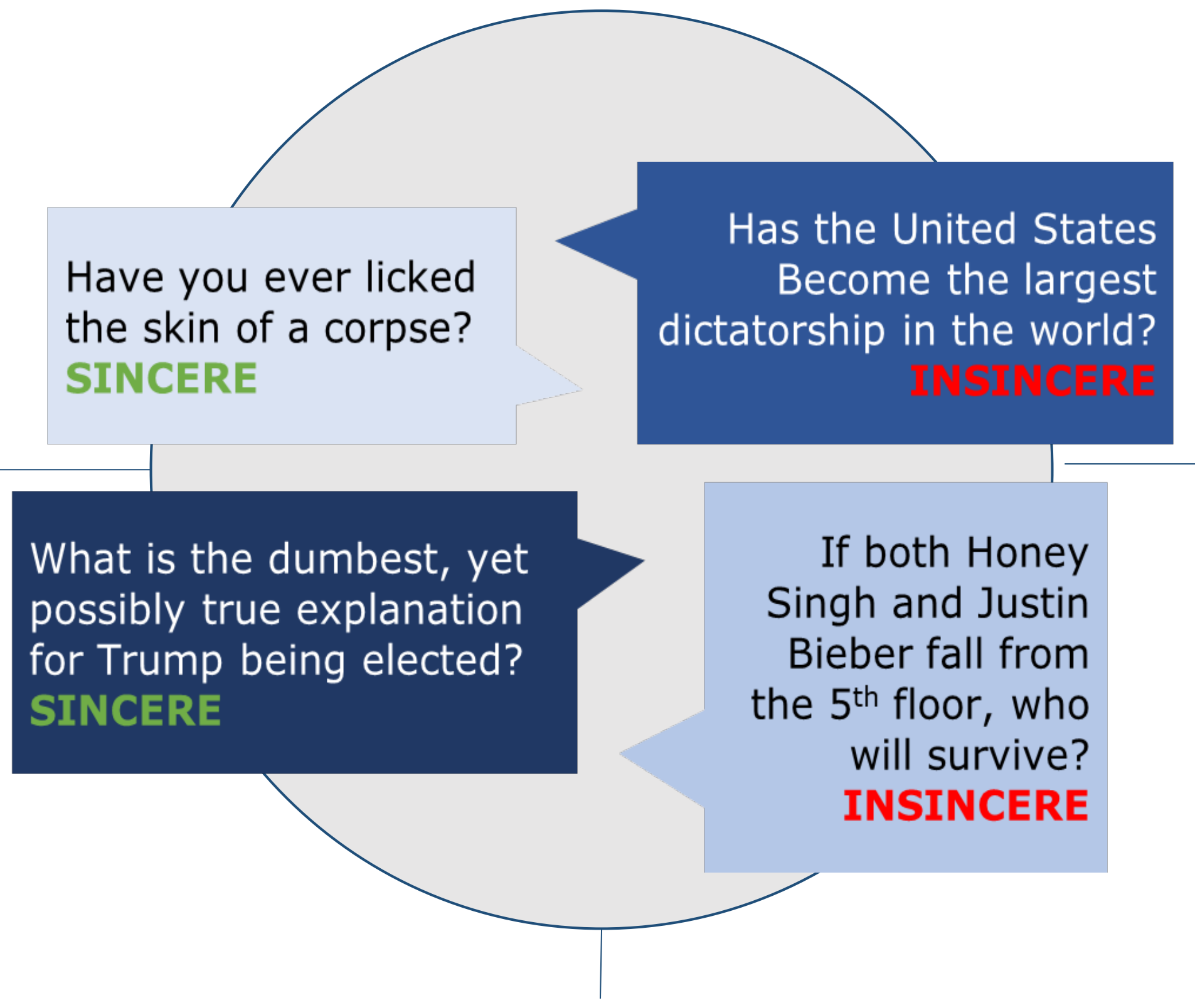
- **Non-Neutral Tone** – heavily exaggerated and/or rhetorical
- **Disparaging** – discriminatory content or premises
- **Not Grounded in Reality** – based on false information, or absurd assumption(s)
- **Sexual Content** – incest, bestiality, pedophilia and/or other sexual content included for shock-value rather than necessity

DESCRIPTION OF DATA

DATA EXPLORATION Training data pulled from Kaggle contained 1,306,122 rows – each containing a Question ID, Question Text, and indicator of whether the question was as sincere (0) or insincere (1).

INVESTIGATION OF DATA LABELS

Some labels **appear to favor falsely classifying as sincere**, rather than incorrectly classifying as insincere. This could affect model accuracy.



Have you ever licked the skin of a corpse? **SINCERE**

Has the United States Become the largest dictatorship in the world? **INSINCERE**

What is the dumbest, yet possibly true explanation for Trump being elected? **SINCERE**

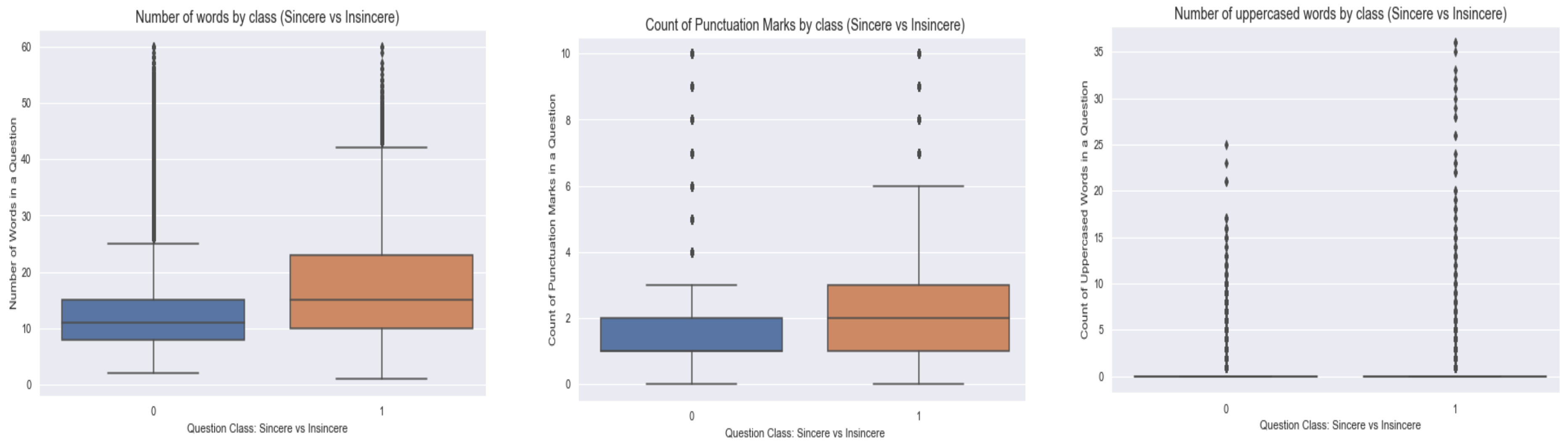
If both Honey Singh and Justin Bieber fall from the 5th floor, who will survive? **INSINCERE**

DATA AUGMENTATION

Only **6% of the training questions were insincere.** Augmentation techniques were used to create more insincere questions, and outliers were removed from the sincere question population to **create more balanced training data**.

INVESTIGATION OF CLASS DIFFERENCES

Initial exploration sought to determine if simple characteristics could easily separate sincere from insincere questions, including: **Length, Punctuation**, and **Capitalization**

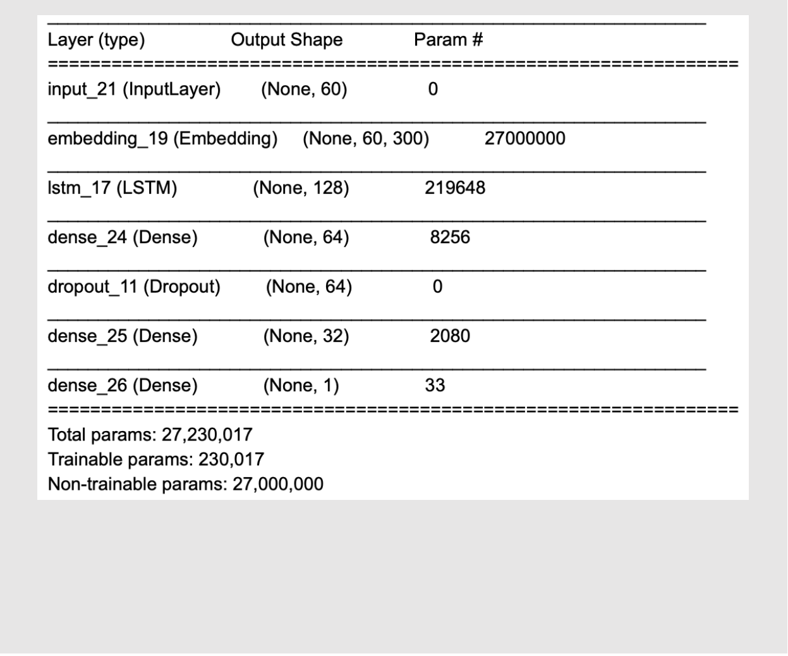


The boxplots suggest **more complex models may be required** to separate the classes.

DESCRIPTION OF MODELS

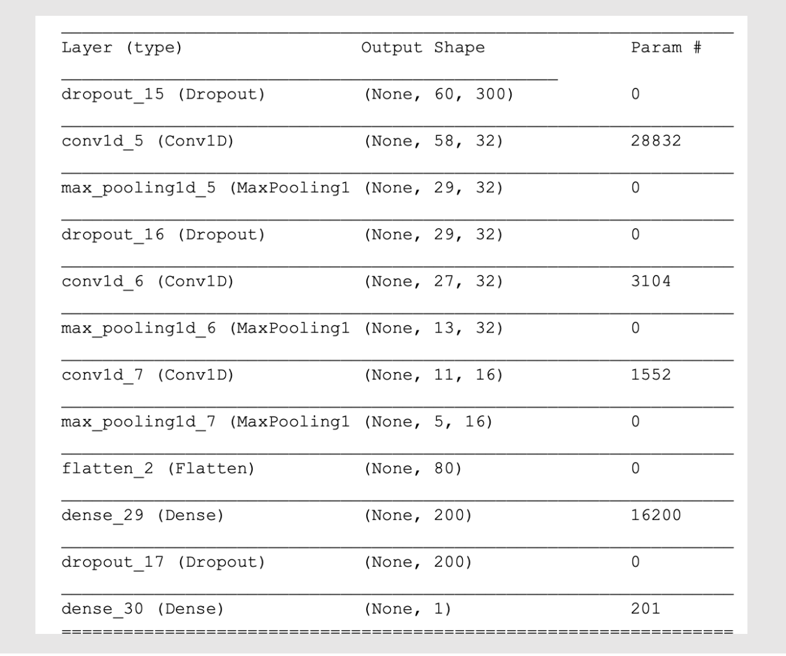
METHODOLOGY Three models were developed to assist in the identification of insincere questions within the cleaned and labelled Quora questions dataset:

LSTM Model



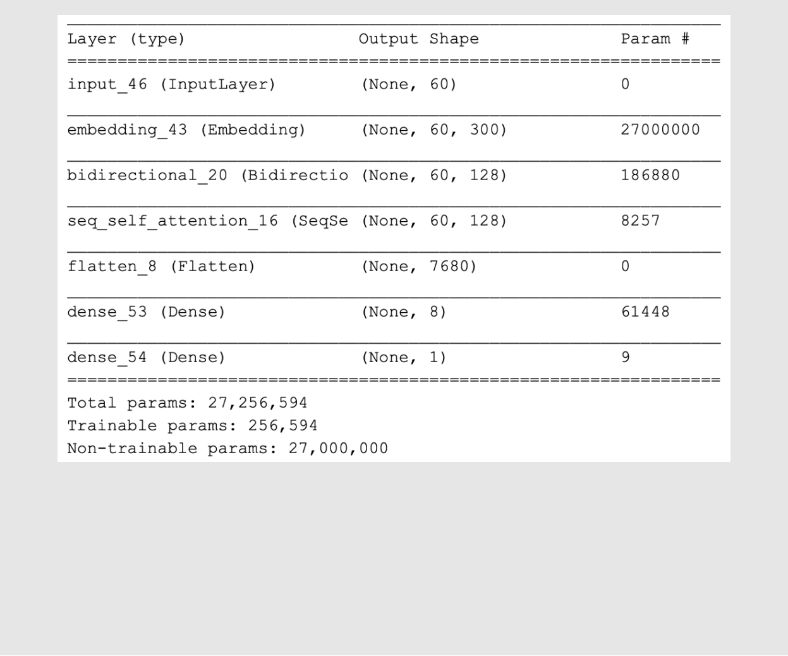
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_21 (InputLayer)        (None, 60)                0
_________________________________________________________________
embedding_19 (Embedding)     (None, 60, 300)           27000000
_________________________________________________________________
lstm_17 (LSTM)               (None, 128)               219648
_________________________________________________________________
dense_24 (Dense)             (None, 64)                8256
_________________________________________________________________
dropout_11 (Dropout)         (None, 64)                0
_________________________________________________________________
dense_25 (Dense)             (None, 32)                2080
_________________________________________________________________
dense_26 (Dense)             (None, 1)                 33
=================================================================
Total params: 27,230,017
Trainable params: 230,017
Non-trainable params: 27,000,000
```

Word-Level 1D CNN Model



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
_________________________________________________
dropout_15 (Dropout)         (None, 60, 300)           0
_________________________________________________________________
conv1d_5 (Conv1D)            (None, 58, 32)            28832
_________________________________________________________________
max_pooling1d_5 (MaxPooling1 (None, 29, 32)            0
_________________________________________________________________
dropout_16 (Dropout)         (None, 29, 32)            0
_________________________________________________________________
conv1d_6 (Conv1D)            (None, 27, 32)            3104
_________________________________________________________________
max_pooling1d_6 (MaxPooling1 (None, 13, 32)            0
_________________________________________________________________
conv1d_7 (Conv1D)            (None, 11, 16)            1552
_________________________________________________________________
max_pooling1d_7 (MaxPooling1 (None, 5, 16)             0
_________________________________________________________________
flatten_2 (Flatten)          (None, 80)                0
_________________________________________________________________
dense_29 (Dense)             (None, 200)               16200
_________________________________________________________________
dropout_17 (Dropout)         (None, 200)               0
_________________________________________________________________
dense_30 (Dense)             (None, 1)                 201
=================================================================
```

Attention Model



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_46 (InputLayer)        (None, 60)                0
_________________________________________________________________
embedding_43 (Embedding)     (None, 60, 300)           27000000
_________________________________________________________________
bidirectional_20 (Bidirectio (None, 60, 128)           186880
_________________________________________________________________
seq_self_attention_16 (SeqSe (None, 60, 128)           8257
_________________________________________________________________
flatten_8 (Flatten)          (None, 7680)              0
_________________________________________________________________
dense_53 (Dense)             (None, 8)                 61448
_________________________________________________________________
dense_54 (Dense)             (None, 1)                 9
=================================================================
Total params: 27,256,594
Trainable params: 256,594
Non-trainable params: 27,000,000
```

The LSTM and 1D CNN Models were **trained for 7 epochs** and the Attention Model **trained for 3 epochs** on the cleaned and augmented training dataset.

ANALYSIS OF RESULTS

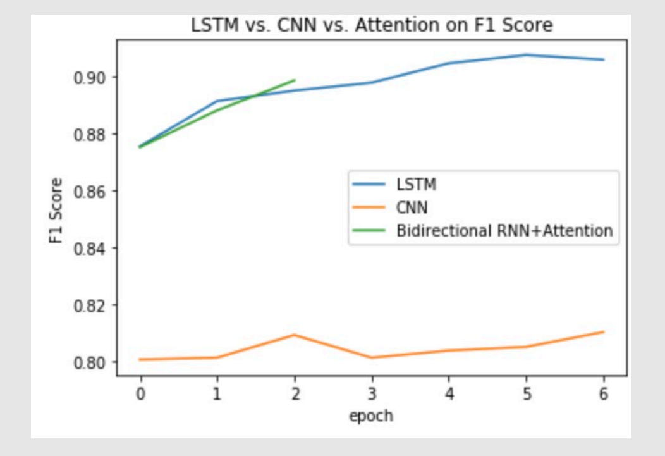
CONCLUSIONS

Each of the models was tested using a hold-out portion of the cleaned and augmented labelled data, and then assessed for performance based on a comparison of **F1 Score, Sensitivity,** and **Precision**:

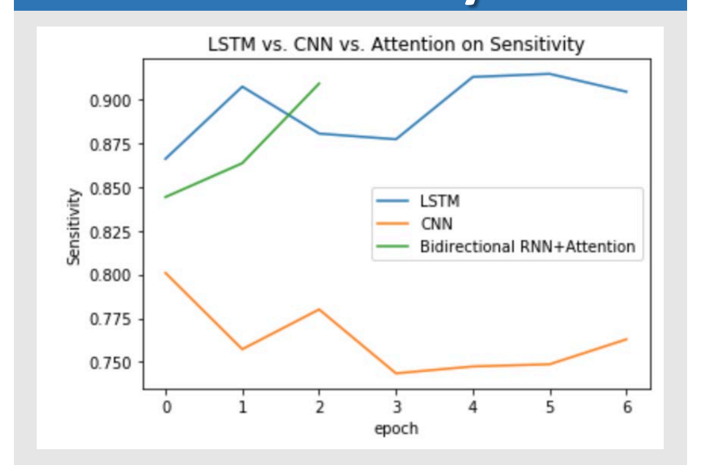
- **1D CNN** is **not suitable** for complex text classification.
- **LSTM Model** has the **highest precision**.
- **Attention Model** performs the best on classification. The Sensitivity and F1 Score plots suggest it correctly captures the most insincere questions and suggests the Attention Model has **potential for increased performance if trained longer** with additional epochs. The precision plot suggests the Attention Model tends to classify more sincere questions as insincere – however, our intuition did the same when exploring the data labels by hand.

The Attention Model is the most successful detector of question sincerity.

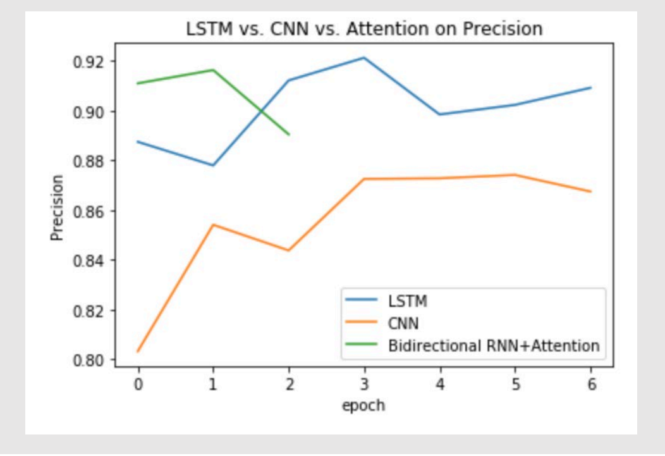
F1 Score



Sensitivity



Precision



NEXT STEPS Follow-on research could be performed to further enhance model performance including:

- **Hand-Label Data –** Initial exploration suggested data labels potentially incorrectly flagged questions as sincere. Hand-labelling could be performed to improve data accuracy, and potentially results.
- **Train Models Longer –** Due to resource limitations, the LSTM and CNN Models were trained for 7 epochs, and the Attention model was trained for 3. These models can be trained for longer and re-evaluated for performance.
- **Scalability –** Expand research to additional online Q&A forums to investigate model performance and scalability beyond Quora

Georgetown University

Georgetown University | ANLY590 Final Project | Fall 2019