Using Machine Learning to Detect Insincerity in Online Forums

Jonathan Ho, Marisa Lenci, Dylan Sivori

W207 Applied Machine Learning

10 December 2024

Berkeley
SCHOOL OF INFORMATION

Agenda

1. Introduction: Background & Problem Motivation
2. Data: Preprocessing & EDA
3. Modeling: Baseline
4. Modeling: Improvement Over Baseline
5. Modeling: Final Model & Results
6. Conclusions & Implications

Introduction

Background & Problem Motivation

Introduction: Background & Problem motivation

Online content platforms and forums provide a place for everyone on the internet to connect, regardless of boundaries (geographic, social, economic, etc.)

These platforms must maintain a safe environment for everyone, which poses a challenge given the scale and prevalence of content.

Introduction: Background & Problem motivation

Quora's mission is to share and grow the world's knowledge.

Quora is an online content platform where people can ask and respond to questions, connecting with others in the platform who contribute unique insights and quality answers.

In maintaining their platform, they face a challenge in needing to **identify and manage insincere questions** – those founded upon false premises, or that intend to make a statement rather than look for helpful answers.

Data

Preprocessing & Exploratory Data Analysis

Data

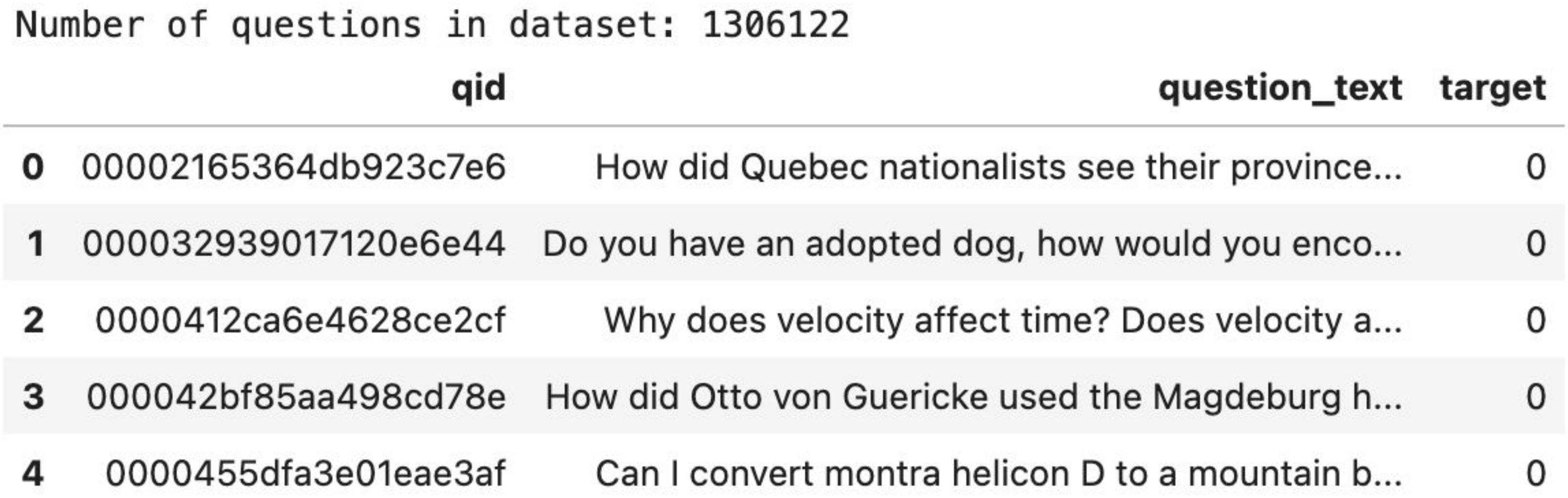
Kaggle dataset from Quora of Questions

- **qid** - unique question identifier
- **question_text** - Quora question text
- **target** - a question labeled "insincere" has a value of `1`, otherwise `0`

Defining Target Insincerity

- An insincere question is defined as a question intended to make a statement rather than look for helpful answers.
    - Has a non-neutral tone
    - Is disparaging or inflammatory
    - Isn't grounded in reality
    - Uses sexual content (incest, bestiality, pedophilia) for shock value, and not to seek genuine answers

Data: Preprocessing & EDA



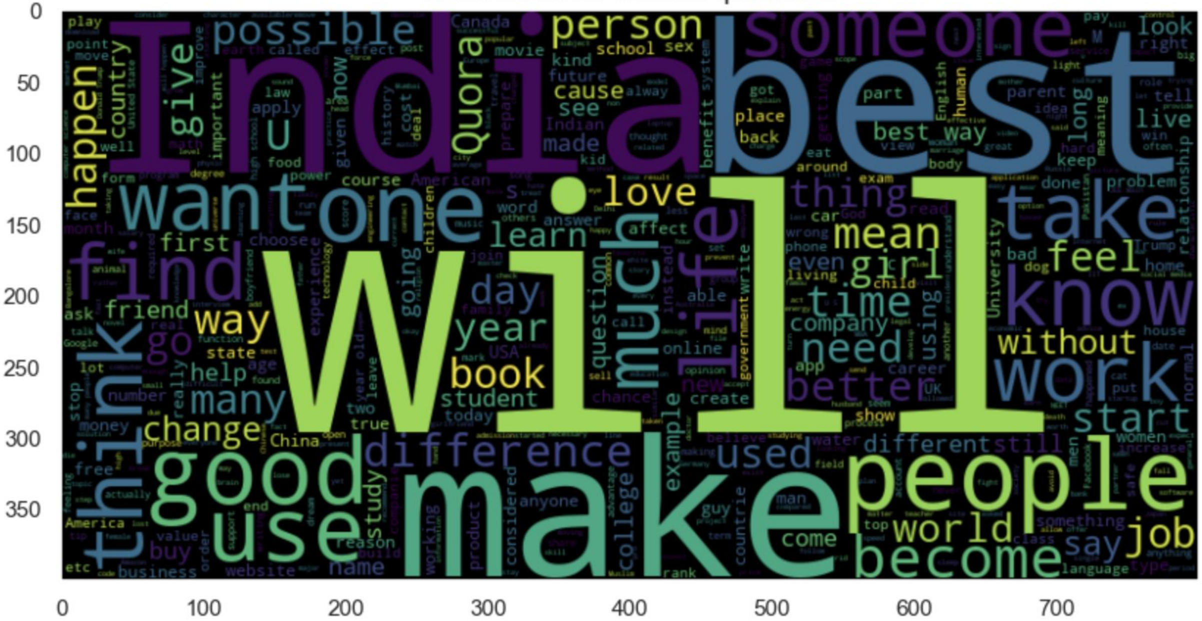
```
Number of questions in dataset: 1306122
```

	qid	question_text	target
0	00002165364db923c7e6	How did Quebec nationalists see their province...	0
1	000032939017120e6e44	Do you have an adopted dog, how would you enco...	0
2	0000412ca6e4628ce2cf	Why does velocity affect time? Does velocity a...	0
3	000042bf85aa498cd78e	How did Otto von Guericke used the Magdeburg h...	0
4	0000455dfa3e01eae3af	Can I convert montra helicon D to a mountain b...	0

Data: Preprocessing & EDA

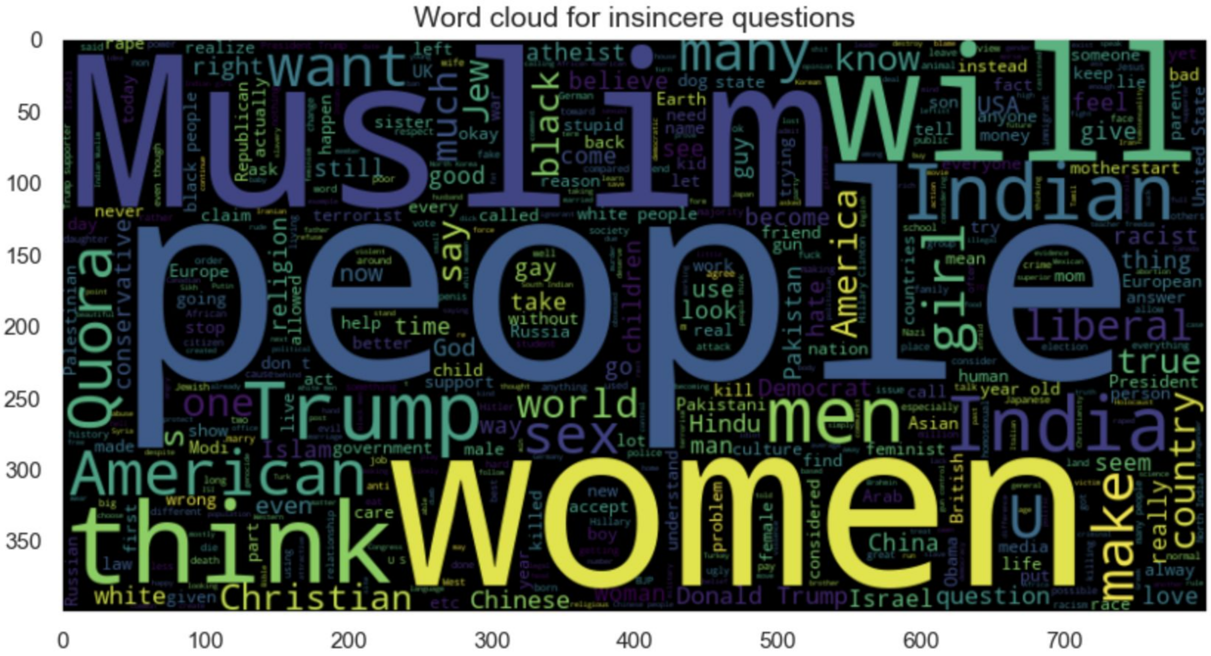
Understanding sincere questions

Word cloud for sincere questions



Data: Preprocessing & EDA

Understanding insincere questions



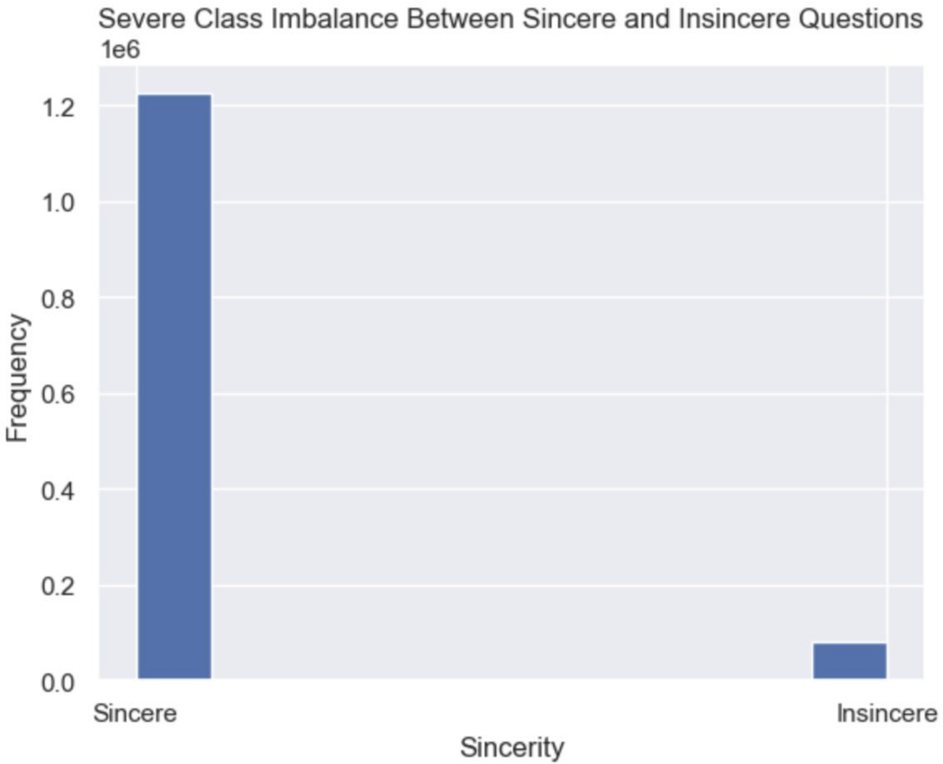
Data: Preprocessing & EDA

Findings

- Severe class imbalance favoring **sincere** class

Solution

- Rebalancing dataset for **1:3 ratio** of insincere:sincere questions



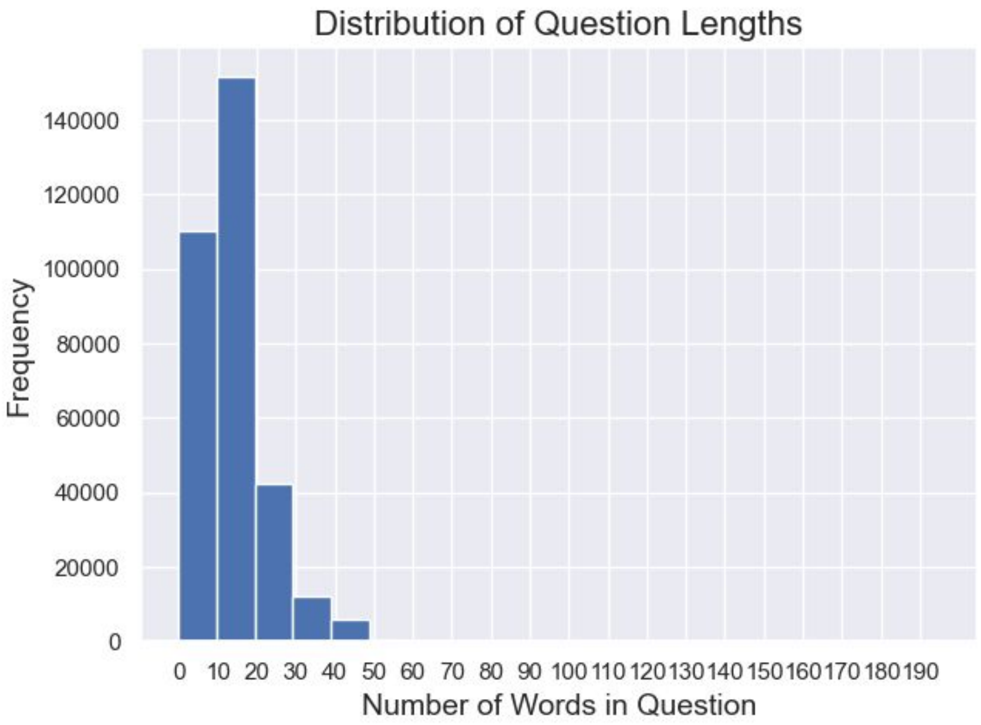
Data: Preprocessing & EDA

Findings

- Most questions between 10-20 words, right skew

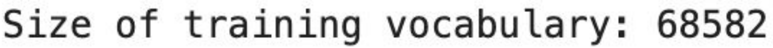
Solution

- Truncate questions to 20 words maximum – reduces dimensionality and potential for overfitting; reduces computational load

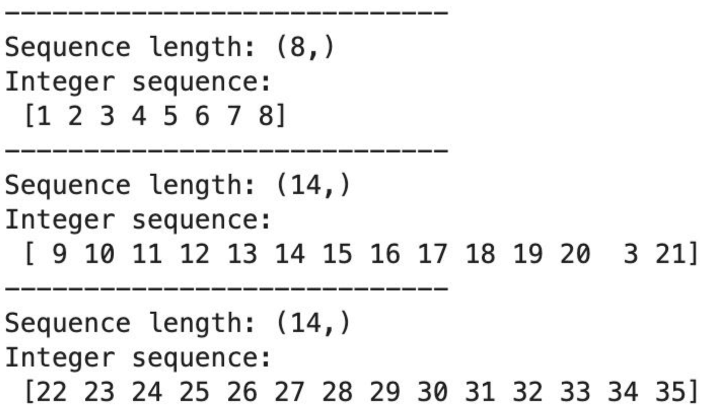


Data: Preprocessing & EDA

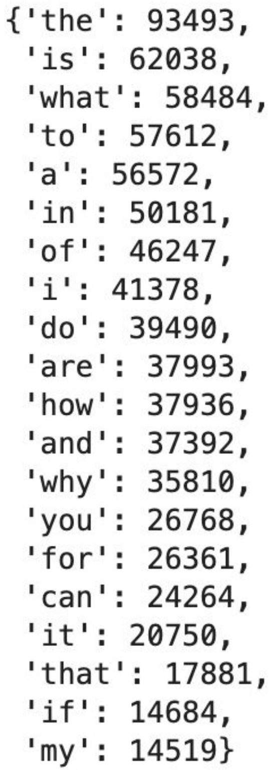
Embeddings



```
Size of training vocabulary: 68582
```



```
----------------------------
Sequence length: (8,)
Integer sequence:
 [1 2 3 4 5 6 7 8]
----------------------------
Sequence length: (14,)
Integer sequence:
 [ 9 10 11 12 13 14 15 16 17 18 19 20  3 21]
----------------------------
Sequence length: (14,)
Integer sequence:
 [22 23 24 25 26 27 28 29 30 31 32 33 34 35]
```



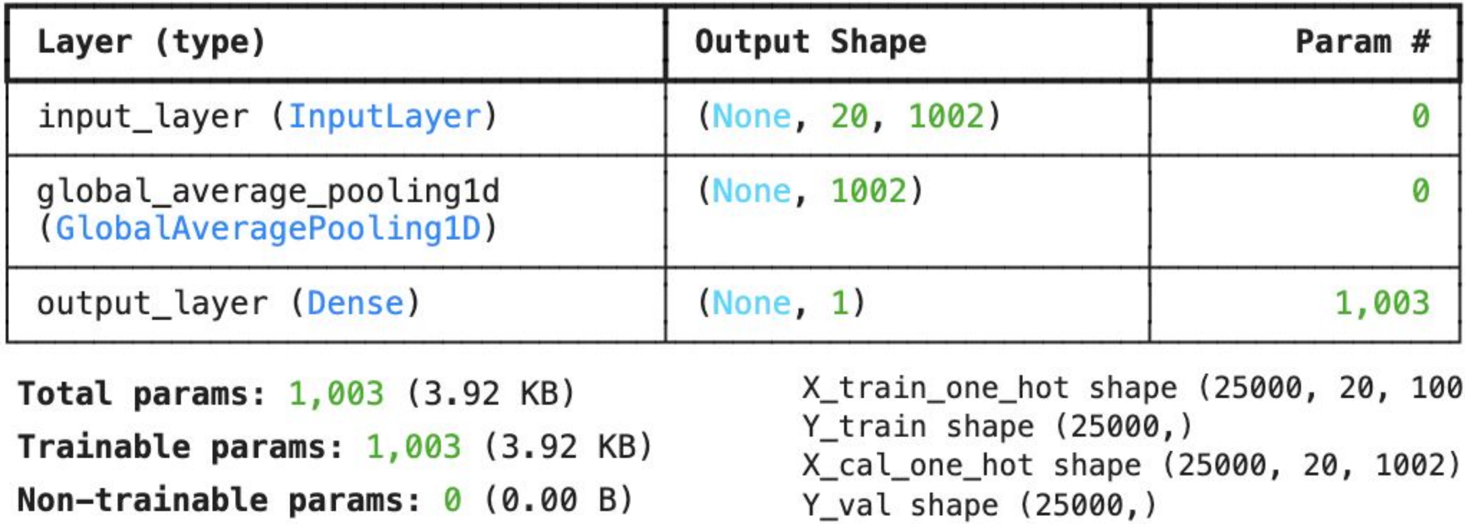
```
{'the': 93493,
 'is': 62038,
 'what': 58484,
 'to': 57612,
 'a': 56572,
 'in': 50181,
 'of': 46247,
 'i': 41378,
 'do': 39490,
 'are': 37993,
 'how': 37936,
 'and': 37392,
 'why': 35810,
 'you': 26768,
 'for': 26361,
 'can': 24264,
 'it': 20750,
 'that': 17881,
 'if': 14684,
 'my': 14519}
```

Modeling

Baseline

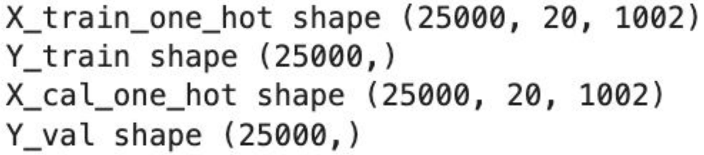
Modeling: Baseline

- Baseline model using one-hot encoded representation of question in a FFNN



Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 20, 1002)	0
global_average_pooling1d (GlobalAveragePooling1D)	(None, 1002)	0
output_layer (Dense)	(None, 1)	1,003

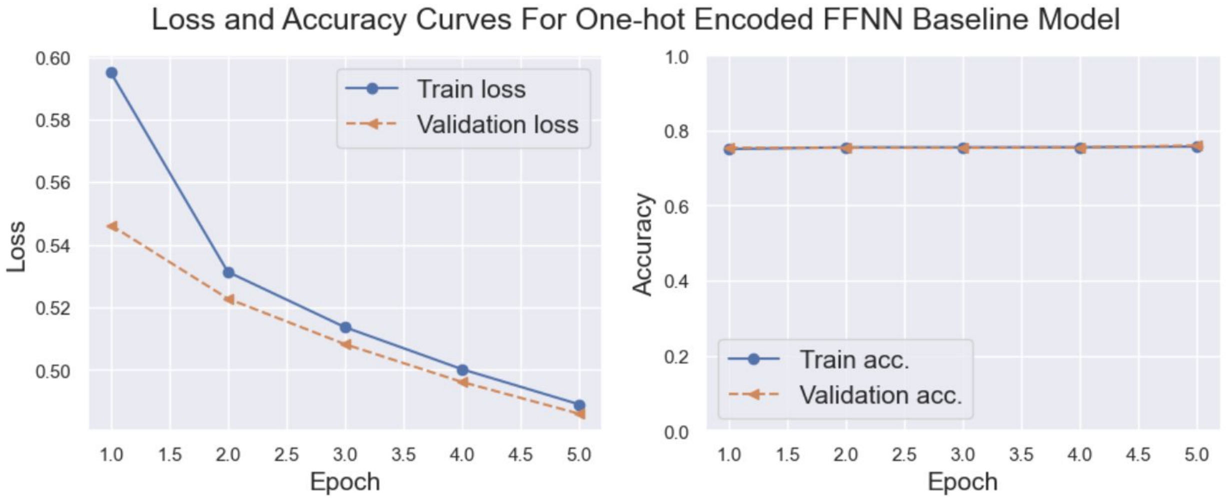
```
Total params: 1,003 (3.92 KB)
Trainable params: 1,003 (3.92 KB)
Non-trainable params: 0 (0.00 B)
```

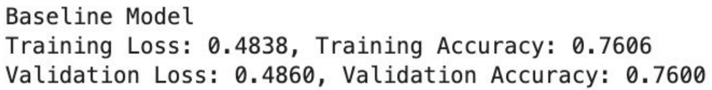


```
X_train_one_hot shape (25000, 20, 1002)
Y_train shape (25000,)
X_cal_one_hot shape (25000, 20, 1002)
Y_val shape (25000,)
```

Modeling: Baseline

- Meaningful distances between tokens not captured at this stage but still performing at a reasonable level





```
Baseline Model
Training Loss: 0.4838, Training Accuracy: 0.7606
Validation Loss: 0.4860, Validation Accuracy: 0.7600
```

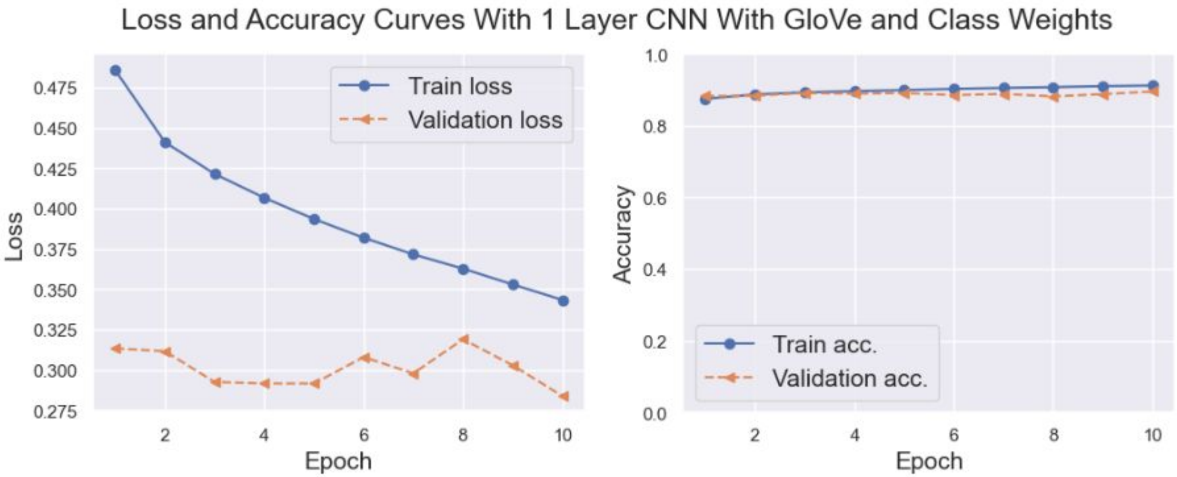
Modeling

Improvement Over Baseline

Modeling: Improvement over Baseline

Model 1: 1 convolutional layer with pre-trained GloVe embeddings and class weights

- Strong improvement over baseline, high accuracy and low loss for train & validation sets



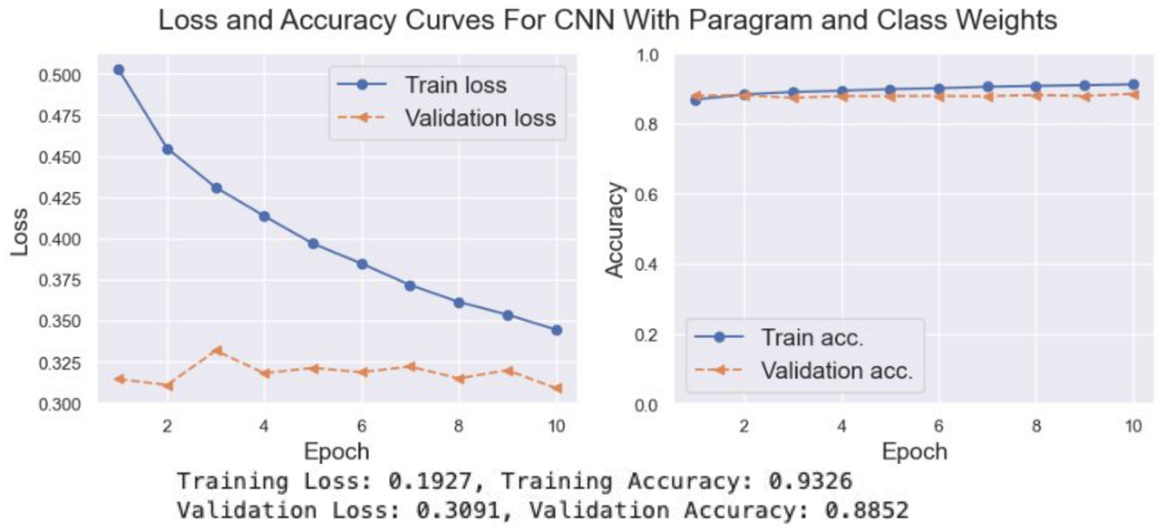
Loss and Accuracy Curves With 1 Layer CNN With GloVe and Class Weights

```
Training Loss: 0.2038, Training Accuracy: 0.9280
Validation Loss: 0.2837, Validation Accuracy: 0.8953
```

Modeling: Improvement over Baseline

Model 2: 1 convolutional layer with pre-trained Paragram embeddings and class weights

- No improvement over Model 1 – Slightly higher loss and lower accuracy on train and validation sets, GloVe embeddings appear to perform better so we select GloVe moving forward



Loss and Accuracy Curves For CNN With Paragram and Class Weights

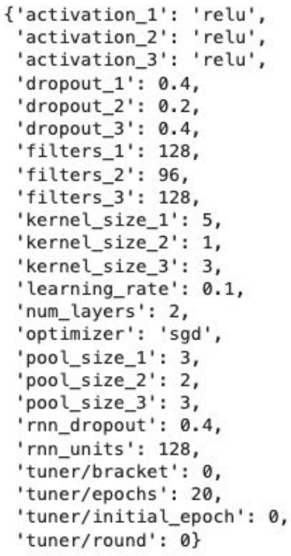
```
Training Loss: 0.1927, Training Accuracy: 0.9326
Validation Loss: 0.3091, Validation Accuracy: 0.8852
```

Modeling: Improvement over Baseline

Model 3: 1 recurrent layer, multiple convolutional layers, Hyperband finetuning

- Model features: 1 Recurrent layer, 1-3 Convolutional layers, Global Average pooling layer, Dense layer, Class weights

- Hyperparameter tuning with Hyperband:
  - Recurrent layer
    - Filters (32-128)
    - Dropout (0.2-0.4)
  - Convolutional layers
    - Number of layers (1-3)
    - Filters (32-128)
    - Kernel size (1-5)
    - Activation (relu, tanh)
    - Dropout (0.2-0.4)
    - Max pooling size (1-5)
  - Other
    - Learning rate (0.1, 0.01, 0.001)
    - Optimizer (SGD, Adam)



```
{'activation_1': 'relu',
 'activation_2': 'relu',
 'activation_3': 'relu',
 'dropout_1': 0.4,
 'dropout_2': 0.2,
 'dropout_3': 0.4,
 'filters_1': 128,
 'filters_2': 96,
 'filters_3': 128,
 'kernel_size_1': 5,
 'kernel_size_2': 1,
 'kernel_size_3': 3,
 'learning_rate': 0.1,
 'num_layers': 2,
 'optimizer': 'sgd',
 'pool_size_1': 3,
 'pool_size_2': 2,
 'pool_size_3': 3,
 'rnn_dropout': 0.4,
 'rnn_units': 128,
 'tuner/bracket': 0,
 'tuner/epochs': 20,
 'tuner/initial_epoch': 0,
 'tuner/round': 0}
```

- Results: No significant improvement over Model 1
  - High computation costs of running the Hyperband in Model 3

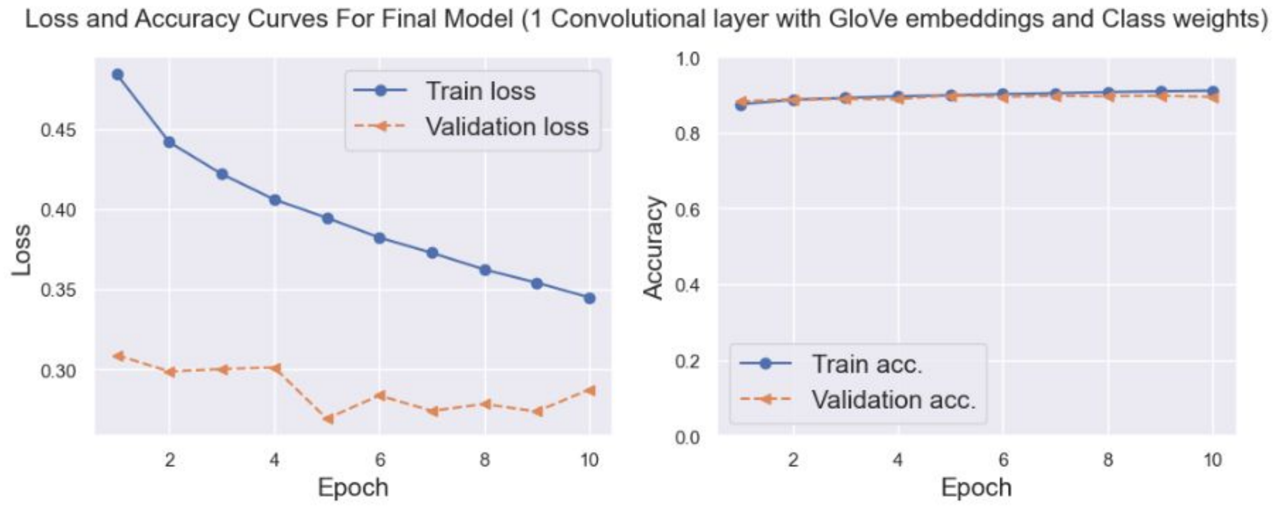
Modeling: Improvement over Baseline

Model 4: 1 convolutional layer, RandomSearch finetuning

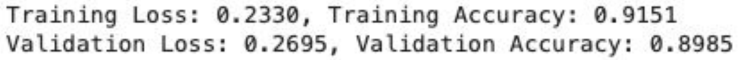
- Model features: 1 Convolutional layer, Dropout, Max Pooling and Global Average pooling layers, Dense layer, Class weights
- Hyperparameter tuning with RandomSearch
- Final model specifications:
  - Convolutional layer
    - Filters = 256
    - Kernel size= 2
  - Other
    - Learning rate = 0.0005
- Results: No significant improvement over Model 1
  - Higher loss for training and validation, and bigger difference between training and validation accuracy suggests Model 4 may be overfitting

Modeling: Improvement over Baseline

Model 5: 1 convolutional layer (128 filters); Dropout (0.5); MaxPooling (2); Global Average Pooling, Dense sigmoid layer; GloVe embeddings; early stopping; class weights



Loss and Accuracy Curves For Final Model (1 Convolutional layer with GloVe embeddings and Class weights)



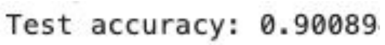
```
Training Loss: 0.2330, Training Accuracy: 0.9151
Validation Loss: 0.2695, Validation Accuracy: 0.8985
```

Modeling

Final Model

Modeling: Final Model Results

Results of Model 5: 1 convolutional layer with GloVe embeddings, early stopping



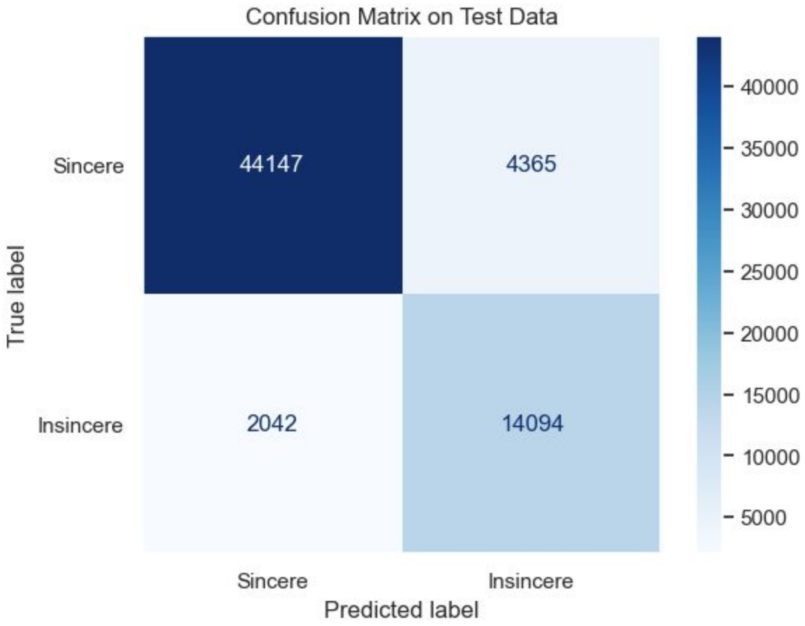
```
Test accuracy: 0.90089
```

Generalizability

- Test accuracy: 90%
- Model 5 shows strong generalizability

Precision: 76.4%

Recall: 87.3%



Conclusions & Implications

Ethical Considerations, Limitations, Further Research

Conclusions & Implications

Ethical Considerations

- Whenever there is online content moderation there will be discussion of censorship and free speech. Ensuring a safe platform also means that the censorship of insincere questions must be carefully considered.

- If our model were taken out of Quora's context, there could be a negative social impact. For example, if a news or political organization were to use it to suppress genuine and potentially critical comments on their material.

- What constitutes a "sincere" question can be somewhat subjective and context-dependent. For example, Quora defines *insincere questions* as those that are not "grounded in reality" which may differ greatly based on different contexts. Quora also considers *insincere questions* those that "use sexual content for shock value, and not to seek genuine answer", which may be difficult to determine.

Conclusions & Implications

Limitations and Further Research

- It is possible that there are biases present in the training data, leading to a misrepresentation of Quora questions which could limit the efficacy of this model applied in Quora's context. It is also possible that there is bias in the labeling, given the subjective nature of what is sincere.
- Understanding how the data was collected and testing our model on additional sets of Quora questions could help us understand and mitigate the presence of biases.
- Model use must also stay within online forum question domain. Different models must be trained for different moderation such as YouTube comments or live streaming chats
- Language and context can be complex, and can vary across countries and cultures; what is sincere in one culture may seem insincere in another.
- Understanding how Quora's questions vary across various subsets such as geographic location may be an important next step for further research, which could also alleviate some ethical considerations.

Conclusions & Implications

- **Conclusion:** Online content moderation can be achieved with a reasonably high degree of accuracy and is worth investing resources into. Misclassified insincere questions can always be handled with an appeal functionality that pushes the question for human review.

Group Contributions

Data Preprocessing & EDA

- Dylan Sivori, Jonathan Ho

Modeling: Baseline

- Dylan Sivori, Marisa Lenci, Jonathan Ho

Modeling: Improvement Over Baseline

- Jonathan Ho, Dylan Sivori, Marisa Lenci

Modeling: Final Model Results

- Jonathan Ho, Marisa Lenci

Conclusions & Implications

- Marisa Lenci, Dylan Sivori

Framing Narrative and Slides

- Marisa Lenci, Dylan Sivori, Jonathan Ho

Citations

Kaggle: Quora Insincere Questions Classification

- https://www.kaggle.com/competitions/quora-insincere-questions-classification/overview
  - Data
  - Definition & interpretation of data, class labels
  - Pre-trained embeddings (GloVe, Paragram)

Thanks!