Is this question sincere? Identifying insincere questions on Quora using BERT and variations
Alex Wang (jwang98), Vince Ranganathan (akranga)

Background & Problem
- Flagging “insincere” questions is an ongoing challenge for online forums
- Insincere = non-neutral tone, disparaging, false assumptions, ...
- Difficult—words like “bad”, “worthless” can be used genuinely!
- **Flag input question as sincere or not**
- Dataset provided by Quora + Kaggle
- Compare LSTM+GRU baseline to BERT, BERT+CNN, and BERT+Linear

Pre-processing
1. everything.toLower()
2. Expand the contractions using a map
3. Remove Special chars using map
4. Autocorrect top 50 misspellings (“qoura”)
5. Replace digit5 with token='#'

Models
1. LSTM + GRU baseline
2. BERT (excluding pink)
3. BERT + Linear
4. BERT + CNN

Conclusion & Discussion
- **BERT performs better than LSTM+GRU**
- Likely because fewer parameters => less propensity to overfit on noisy dataset
- BERT variations perform worse; possibly redundant parameters complicating the model
- Noise in the dataset is a huge problem! Many BERT predictions falsely marked “incorrect”

Results & Outputs

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM + GRU</td>
<td>0.956</td>
<td>0.635</td>
<td>0.709</td>
<td>0.670</td>
</tr>
<tr>
<td>BERT</td>
<td>0.959</td>
<td>0.645</td>
<td>0.765</td>
<td>0.700</td>
</tr>
<tr>
<td>BERT + CNN</td>
<td>0.958</td>
<td>0.639</td>
<td>0.742</td>
<td>0.687</td>
</tr>
<tr>
<td>BERT + Linear</td>
<td>0.956</td>
<td>0.625</td>
<td>0.749</td>
<td>0.681</td>
</tr>
</tbody>
</table>

ROC Curve