Is this question sincere? Identifying insincere questions on Quora using BERT and variations

Alex Wang
Department of Computer Science
Stanford University
Stanford, CA 94305
jwang98@stanford.edu

Vince Ranganathan
Department of Mathematics
Stanford University
Stanford, CA 94305
akranga@stanford.edu

Abstract

The flagging of insincere questions on internet forums is an ongoing challenge for online communities that rely heavily on the faith and support of their users. We develop and train three variations of the BERT model on a dataset provided by Quora to predict whether posted questions are sincere or insincere. We compare our model's performance to the well-performing alternative architecture of LSTM + GRU and present an extensive discussion on the observed advantages and weaknesses of the BERT model in this context.

1 Introduction

We seek to develop models that take as input the text of a question in English, and outputs a 0 or 1 corresponding to whether the question should be approved as "sincere" or flagged as "insincere."

This is a highly relevant problem today, with online forums that act as platforms for people to resolve their curiosities—such as Reddit, Yahoo Answers, and StackExchange—struggling to handle posted questions that violate the guidelines of the forum. Though there often exist humans that act as moderators, usually individuals passionate about the discussion topic, the number of questions to monitor far outweighs the capacity of these moderators. The ability to automatically flag disingenuous, malicious, discriminatory questions—broadly, "insincere" questions—will help remove such negative content from these communities quickly and deter potential posters.

In this research, we use a dataset of questions and noisy labels provided by Quora. Insincere questions, according to Quora, are defined as ones that meet any of the following (non-comprehensive) criteria:

- *Has a non-neutral tone*, such as an exaggeration to emphasize a comment about a particular group of people
- *Is disparaging or inflammatory*, such as an attempt to seek confirmation of a stereotype or present a discriminatory remark, a comment based on an outlandish premise about a group of people, or the disparaging of a natural, not fixable, or immeasurable characteristic
- *Is not grounded in reality*, such as a reference to false information or absurd assumptions
- *Uses sexual content for shock value*, such as references to pedophilia, incest, or bestiality outside of an attempt to find genuine answers

This task is difficult due to the wide range of topics, moods, and phrasings that could be considered "insincere." Certain words, such as "stupid" or "dumb" or "ugly" may often lend themselves to discriminatory comments, but descriptors that are broad such as "worthless" (for example, in the question, "Why are 512 MB flash drives effectively worthless these days?"), or meaning-overloaded such as "bad" (such as in "How long does it take for milk to go bad?" or "Why is Michael Jackson's *Bad* so famous?") are much more difficult to extract information from. Additionally, nuances such as

the tone of the question, the use of jargon, slang, or abbreviations that are beyond the scope of the word embeddings, and the differences in phrasing between statements, exclamations, and questions contribute to the difficulty in identifying the intentions of a question asker.

The primary model used for question classification tasks is a model based on stacked bidirectional LSTMs and GRUs (which we henceforth refer to as LSTM + GRU) [1]. We approach this problem using BERT and variations of BERT, which are pretrained models with a final linear layer that is finetuned to the pertinent classification task. We choose BERT primarily because we anticipate that it will perform better on the dataset with noisy labels, since it has magnitudes of order fewer parameters to fit and hence is more likely to generalize its assessment.

2 Model

To takle this challenge, we implement three vartations of the BERT model: vanilla BERT, BERT + CNN, and BERT + Linear. We also implement an LSTM + GRU model as a baseline, which is a known well-performing model on this task.

2.1 LSTM + GRU (Baseline)

We reimplement a LSTM + GRU model has been shown to perform well for this task [1]. The architecture of the LSTM + GRU model is as follows:

1. Bidirectional LSTM with attention on input sequence
2. Bidirectional GRU with attention on LSTM output sequence
3. Separate average-pool and max-pool layers on the GRU output sequence
4. Linear layer with $W \in \mathbb{R}^{16\times1024}, b \in \mathbb{R}^{16}$
5. ReLU layer over linear layer outputs
6. Dropout layer with $p = 0.1$

The diagram in Figure 3 below illustrates this architecture:

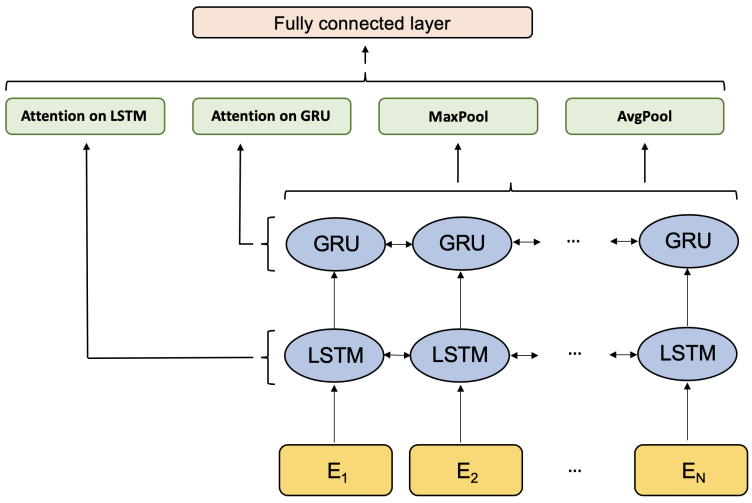
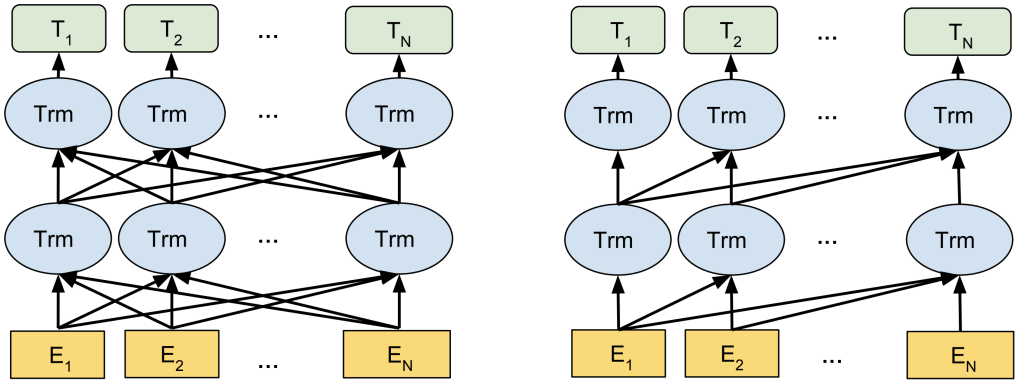


Figure 1: The architecture of the LSTM + GRU baseline model. The attention and pooling outputs are concatenated as an input to the fully connected layer.

2.2 BERT & Variations

We implement BERT models, which is a recent architecture developed by Devlin et al [3]. BERT has shown significant improvement over existing architectures on sequential classification tasks, obtaining state-of-the-art results on the GLUE benchmark, MultiNLI accuracy, and SQuAD v1.1 F1-score.

BERT is based on the OpenAI GPT [4] in the stacking of transformers, but uses bidirectional ones instead of just left-to-right. The figures below demonstrates this distinction:



(a) BERT (b) OpenAI GPT

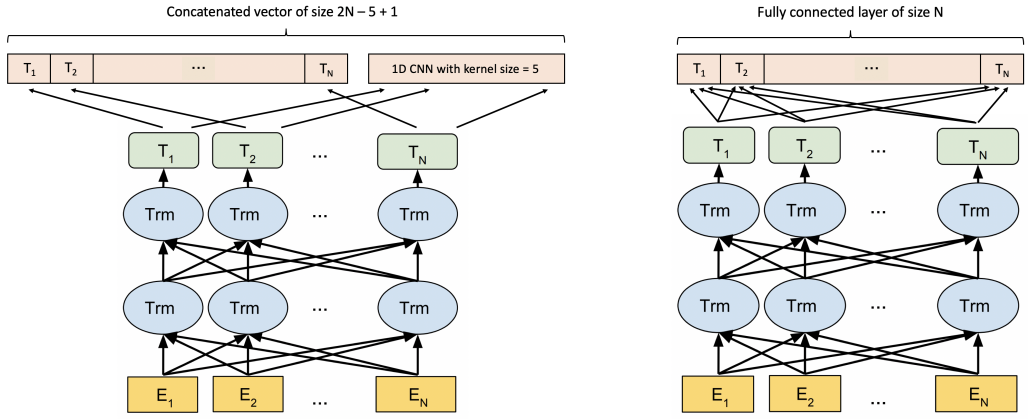
Figure 2: BERT uses bidirectional transformers, as opposed to GPT which uses only left-to-right. [3]

Our implementation on BERT is based on the implementation by Hugging Face AI at [5], which has available a pre-trained BERT model with a linear layer after the pooled output. We fine-tune this trained model by training it on the Quora dataset.

Our set of hyperparameters is as follows:

- **Train batch size:** 16
- **Learning rate:** 1e-4 with an Adam optimizer
- **Number of epochs:** 5 or until convergence (few are needed since BERT models are pre-trained.)

The architectures of the two BERT variations—BERT + CNN and BERT + Linear— are depicted in the Figure 2 below:



(a) BERT + CNN (b) BERT + Linear

Figure 3: The two variations on BERT that we test, with the modifications identified in pale red. All models are followed by a fully connected hidden layer that is not depicted in the diagrams.

3 Experiments

3.1 Data and preprocessing

The dataset that we use is provided by Quora through their "Insincere Questions Classification" challenge on Kaggle. We have split the dataset into 1.05m train and 262k dev, and the challenge provides a 376k test (in terms of number of question-label pairs), for a 62-16-22 train-dev-test split. The columns of the train dataset are simplified to: question ID, question text, and target value (0 or 1).

Insincere questions have a low prevalence of 6.187% across the train and dev sets. Additionally, Quora has suggested that there is an *unidentified amount of noise in the labels*, which places an upper bound on the theoretical performance capacity of the model.

Preprocessing of the question data is based on the public kernel written by Viel at [2]. There are five steps in the process:

1. Converting all alphabet characters to lowercase, e.g. replacing "Word" with "word"
2. Using a pre-defined contractions dictionary map to expend contractions, e.g. replacing "shouldn't" with "should not"
3. Using a pre-defined special-characters dictionary map to clean out non-English language characters and add spaces around punctuation and special characters, e.g. replacing "well, resumé" with "well , resume" or an exponent " ^{2} " with " 2 "
4. Autocorrect the 50 most common misspellings in the dataset by building a list of words that remain outside the embeddings, e.g. replacing "qoura" with "quora"
5. Replacing digits with a fixed token, e.g. converting "$ 350" to "$ ###"

We preprocess all the question text in the train and dev sets before beginning to train the model.

The Kaggle challenge additionally provides two sets of word embeddings that we use to train the LSTM + GRU baseline model:

1. **glove.840B.300d**, i.e the ones created during the GloVe research project at Stanford. There are four embedding sets: one based on Wikipedia, two based on Common Crawl, and one based on Twitter.
2. **paragram_300_sl999**, a.k.a. Paragram Embeddings, developed by the Cognitive Computation Group at the University of Pennsylvania.

3.2 Evaluation method

The evaluation metric for our models is the F_1 score (the harmonic average of precision and recall), which penalizes the model's approval of insincere questions and flagging of sincere ones.

3.3 Results

Model Type	Accuracy	Precision	Recall	F1 Score
LSTM + GRU	0.956	0.635	0.709	0.670
BERT	0.959	0.645	0.765	0.700
BERT + CNN	0.958	0.639	0.742	0.687
BERT + Linear	0.956	0.625	0.749	0.681

We found that BERT by itself performed the best in terms of every metric, with an accuracy of 95.9%, a precision of 0.645, a recall of 0.765, and a F1 score of 0.700. Additionally, each BERT model performed better than the baseline LSTM + GRU model in terms of accuracy and F1 score. The original BERT model had an improvement of 0.03 over the LSTM + GRU's F1 score.

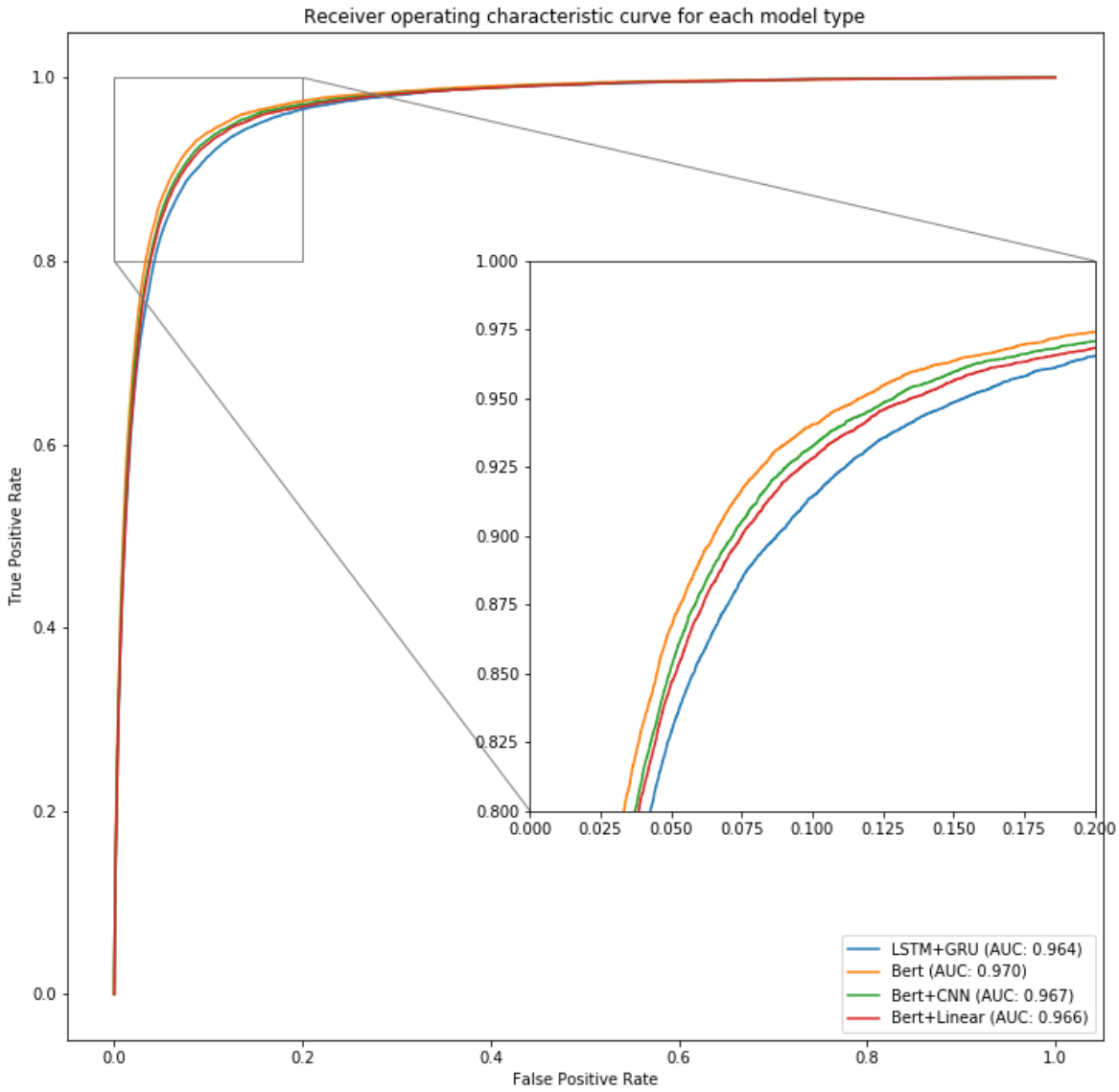


Figure 4: The ROC curve for each type of model.

In Figure 4, we can see that the ROC curves follow a similar pattern to the accuracies and F1 score, with Bert having the highest AUROC and LSTM+GRU the lowest AUROC. However, the difference between each model type is very minimal, with all of the AUROC values within 0.006 of each other. While this does show that the models perform very well for the task, **the high AUROC and accuracy could also be attributed to the large imbalance in the dataset**, with only about 6% of the data having positive labels. Thus, F1 score is a much better measurement of performance in this case, and it is clear that Bert by itself does much better than LSTM+GRU and also the other three models in this respect.

4 Error Analysis & Discussion

In this section we investigate, to the extent possible the trends in BERT's predictions and predictions from BERT vs. predictions from LSTM + GRU.

4.1 Analysis of BERT's predictions

We consider the questions where BERT's absolute error is greatest (i.e. $|p^{(i)}_{\text{BERT}} - \delta^{(i)}|$ is largest, where p_{BERT} is BERT's prediction in the form of a probability, δ is the truth value, and i is the question index), sorted in descending order. Our first observation is that of the 10,664 questions that BERT guessed incorrectly, the 735 with the highest errors (between 0.8392 and 0.9812) all occurred when *BERT guessed 0 where the target score was 1*. We consider a closer inspection of such questions; the top 10 are shown in the table below:

Question text (pre-processed)	**Target**	**BERT output**	**BERT error**
how do daughter cells split apart after mitosis ?	1	0	0.981168244
what is hypercet cholesterol formula ?	1	0	0.981115192
what is the function of charcoal mask blackhead remover mask ?	1	0	0.980969890
are transparent electrical circuit exists ?	1	0	0.980964214
does cold water make you catch colds ?	1	0	0.980286047
what if iam having bleeding after my regular periods on my ovulation day ?	1	0	0.978926698
if we use nitric acid + iron + water + copper contain ammonia solution ?	1	0	0.978781434
what are the requirements to study and work in canada for international students ?	1	0	0.978120390
what is the need of the seo tool wordai ?	1	0	0.977356493
how much more comfortable are plastic corrugated sheets to use than our sheets ?	1	0	0.977141572

Table 1: Questions that are incorrectly labeled 'insincere', which BERT guesses "incorrectly" with a large error (i.e. really guesses correctly with a small error).

Interestingly, *none* of these questions belong to the category of "insincere" questions, though the Target=1 label indicates that this is how they are marked in the Quora dataset. This means that **many of BERT's "errors" are actually correct guesses, marked incorrectly by noise in the dataset**.

There are, however, several questions where BERT predicts that the question is innocuous when it certainly is not. Consider the following examples:

Question text (pre-processed)	**Target**	**BERT output**	**BERT error**
how much wood could a wood chuck chuck if a wood chuck could chuck wood ?	1	0	0.973201755
why is my urine yellow instead of the usual opaque red ?	1	0	0.972459955
how many days are there in a week ?	1	0	0.971635840
how can you face your problem when your problem is your face ?	1	0	0.968160596
how do i make my computer over heat and burst ?	1	0	0.966216303
where do pencils go on vacations ?	1	0	0.964070443
what do you look like in a revealing dress ?	1	0	0.962398200
what can i do to actually get worse at games ?	1	0	0.958109543
i started living my life as a baby # years ago .,?	1	0	0.94519221
how can i find someone to do my essay ?	1	0	0.940543577

Table 2: Questions that are correctly labeled 'insincere', but which BERT guesses incorrectly with a large error.

These are cherry-picked examples (for the purposes of this paper, avoiding content that may be considered offensive) of questions where the Target=1 label is correct, and BERT guesses a very low probability (between 0.036 and 0.061). A common theme in these questions is that the vocabulary and question structures are regular—nine of the ten questions between with what/why/how—and vocabulary is suggestive of genuine thoughts—consider "problem", "urine", "essay", "vacations". We contrast this with the places where BERT correctly identifies insincere questions with the lowest errors (once again, with profanity filtered and outright obscene answers ommitted for the purposes of presentation in this paper):

Question text (pre-processed)	Target	BERT output	BERT error
quora stupid moderators „they collapse my superior answers „how can i destroy their inferiority ?	1	1	0.181789300
did luke search for meaning and redemption during his exile by castrating himself ?	1	1	0.181917100
is it just me „or is obama an utter moron ?	1	1	0.182803100
why is johnathan carl with abc news such a **** ?	1	1	0.183435499
how about all you quorians shut up for a minute and start asking goddamn questions that make sense ?	1	1	0.183615600
why is george kidd such a little ***** ?	1	1	0.183880570
exactly how big of an **** is donald trump ?	1	1	0.184041440
why is obama a stumble **** „relying on his communist goals to screw over the american people	1	1	0.184071840
is not hillary clinton is full of **** and above the law ?	1	1	0.184578540
why indian politician are this much big ****** ?	1	1	0.185157800

Table 3: Questions where BERT correctly identifies insincere questions, using the definition of error given above

BERT is clearly capable of identifying questions that include profanity or other strongly indicative words (e.g. "stupid", "moron", "shut up", "goddamn"). The differences between Table 2 and Table 3 are that the questions in the former have a much more inquisitive tone and lack specific vocabulary that BERT would learn to associate with insincere questions. In other words, **BERT uses vocabulary as a huge indicator** in deciding when to flag questions, and **sometimes the lack of malicious vocabulary causes BERT to miss flagging insincere questions**.

4.2 Predictions from BERT vs. LSTM + GRU

There are 6,495 questions where BERT and LSTM + GRU disagree in their output. Of these, LSTM + GRU is correct 2,901 times, while BERT is correct the other 3594 times. In Table 4 we consider some examples of where each is correct and incorrect for both Target=0 and Target=1.

For the questions 6–10 where LSTM + GRU "correctly" avoids flagging while BERT chooses to flag, upon closer inspection we see that some of the questions ("do indian chosen representatives...", "when did america...", "once you realized quora...") are actually insincere; in these situations, the LSTM + GRU model has clearly overfit to the training data more than BERT.

For the questions 11–15 where BERT "incorrectly" avoids flagging while LSTM + GRU chooses to flag, we observe that some of the questions ("how do tuition...", "why do not gun control...", "why is it necessary...") have a debatable target status, i.e. based on Quora's definition of insincere, they may or may not meet the criteria. As such, it is difficult to penalize the BERT prediction for these questions. One reasonable link that we observe from these results is that **BERT is able to generalize better than LSTM + GRU on the noisy dataset**.

5 Conclusion

After developing a baseline model and three variations of BERT, our experiments and analysis lead us to arrive at the following conclusions:

	Question text (pre-processed)	Target	LSTM + GRU output	BERT output
1	can women be charged for rape in the uk ?	0	1	0
2	why mustangs get all the hate ?	0	1	0
3	why do the wealthy have trophy wives or mistresses who only love their money ?	0	1	0
4	why are people blaming the rise in antisemitism in the us on trump,',s election ?	0	1	0
5	are most teachers narcissists ?	0	1	0
6	do indian chosen representatives do not know how to cast a valid vote and value of that ?	0	0	1
7	should the quora community coin a short term for,',read the comment before you answer this question,',?	0	0	1
8	when did america fall behind and how long before they can admit it to themselves ?	0	0	1
9	once you realized quora is a stage that attracts narcisssists ,,did things seem to make more sense to you ?	0	0	1
10	any women in quora had a love at first sight ?	0	0	1
11	how do tuition teachers able to get sex from their students ?	1	1	0
12	how did you deal with a feminazi ?	1	1	0
13	why do not you die of shame mr .,nitin kamath,(zerodha),?	1	1	0
14	why do not gun control advocates think of the non,-, urban communities ?	1	1	0
15	why is it necessary to assasinate a us president when they stay in office only a few years ?,cannot the opponents just wait them out ?	1	1	0
16	why there are so many social justice warriors on quora ?	1	0	1
17	should straight people be included in the lgbt +,community ?	1	0	1
18	is there any law in india which is incongruent with the attitudes of most indians ?	1	0	1
19	should all people in brazil talk to all people in italy ?	1	0	1
20	what would trump do if he was sent to the naughty corner in the oval office ?,how long would it take him to realise there are no corners ?	1	0	1

Table 4: 10 questions where Target=0 and 10 where Target=1, with LSTM + GRU and BERT offering dissenting predictions.

- BERT performs better than the existing LSTM + GRU architecture, increasing the F_1 score from 0.670 to 0.700, with a significant Recall increase from 0.709 to 0.765. This is likely because BERT is a pretrained model, and so transfering the learning to classification on the Quora dataset is a matter of retraining the final linear layer, which contains so few parameters that it is much more likely to avoid overfitting than LSTM + GRU, which has orders of magnitude more parameters to fit. In fact, we propose that *BERT may be better for working with noisily labeled datasets* due to its propensity to underfit and generalize.
- The variations of BERT do not perform as well as standalone BERT, likely because the original BERT model is so complex and internally connected that adding additional layers dilutes the model's refined outputs.
- The noise in the dataset's target labels is a significant limiting factor in assessing the models' performance. We observe that there are several situations where BERT outputs what any reasonable human familiar with the 'insincere' definition would, but where the target labels are faulty.
- BERT relies heavily on 'giveaway' words (e.g. profanity) in flagging questions, and so occasionally misses questions that use more advanced vocabulary and that are phrased as smoothly as sincere questions. This could potentially be mitigated with better, more extensive preprocessing methods.

References

[1] Lukyanenko, A., 2019, "Text modelling in PyTorch", available at https://www.kaggle.com/artgor/text-modelling-in-pytorch

[2] Viel, T., 2019, "Improve your Score with some Text Preprocessing", available at https://www.kaggle.com/theoviel/improve-your-score-with-some-text-preprocessing

[3] Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. "BERT: Pre-training of deep bidirectional transformers for language understanding". arXiv preprint arXiv:1810.04805.

[4] Radford, A., Wu, J., Child, R., Luan, D., Amodie, D., and Sutskever, I. 2018. "Language Models are Unsupervised Multitask Learners".

[5] Hugging Face AI, "PyTorch Pretrained BERT: The Big & Extending Repository of pretrained Transformers", 2019, GitHub repository, https://github.com/huggingface/pytorch-pretrained-BERT