Language Models (GPT-2 and BERT) vs Humans on Detecting Spam Messages

Changhyun Lee cl4017@nyu.edu
Jennifer Rodriguez Trujillo jr5951@nyu.edu
Yeong Koh yk2678@nyu.edu
Yoobin Cheong yc5206@nyu.edu
New York University - Center for Data Science

Abstract

In this paper, we use randomly collected SMS messages that have been classified as either "spam" or "ham" and create human evaluations using 0-1 classifications based on our subjective opinion of our data. We then take these human evaluations and compare them to GPT-2 and BERT models' evaluations. By inspecting the accuracy scores and correlations between our human evaluations and the models' ratings, we determine how well each model resembles human evaluation. Also, we take a deeper dive into how the two language models get trained on the lexical and grammatical structure of texts. Furthermore, we explore connections between the models and the human language acquisition process. After fine-tuning the models with the training data, we explore possible causes of GPT-2 outperforming both BERT and human predictions based on their final performance. GPT-2 has been pre-trained with informal language: through Reddit pages and dictionary data. This allows the programmer flexibility to optimize the model to learn informal language patterns and detect spam text messages. On the other hand, BERT uses Wikipedia pages, where sentence structure and grammar are less colloquial or less similar to text message format.

Introduction

The Generative Pre-trained Transformer 2 (GPT-2) is an open-source artificial intelligence program created by OpenAI. One of the advantages of the GPT-series models includes their ability to be trained on larger corpora than previous NLP models. Their use of transformer architecture enables massive parallelization. In particular, this program consists of architecture implementations of deep neural networks. Its implementation allows the model to focus on segments of input text and attempts to predict an appropriate response.

GPT-2 uses a token that has undergone Byte Pair Encoding as an input unit. The token dictionary is then created using the divided words into char units, and it repeats the same process of combining the most frequent pairs. Through decoder blocks, the final output receives the final self-attention value which is used to find the probability of each word appearing as the next word. Then, the highest probability becomes the input of the next step, resembling a recurrent neural network (RNN). Like other natural language processing models, GPT-2 takes an embedding matrix as input with an additional step of Positional Encoding to add order information to each word. It is similar to passing the order of words in a sentence in human languages. The passed order information can

adopt the transformer architecture, the biggest advantage of RNN, that can generate high-quality fluent texts (Adelani et al., 2019).

It is a subjective opinion as to what a "good" language model should be capable of, but we can infer that it is ideal for a language model to have cognitive aspects that reflect abilities to be influenced by cognitive human characteristics. More specifically, a language model should not only encompass the ability to pick up on the semantic and syntactic queue of human language, but it should also have the ability to reflect the context and grammatical structure presented. In other words, languages require contextual information, not just simple combinations of words. In GPT-2, we can achieve a higher level of understanding of compositional meaning including lexical components and grammatical meaning through the Attention mechanism.

Although GPT-2 is powerful, it also contains limitations that are yet to be accounted for. Some of these limitations include repetitive text, misunderstanding technical language and highly complex topics that are out of the scope of a human's daily language use, and a misunderstanding of contextual phrases. These limitations may approbate the machine's ability to misclassify.

Cognitively, humans go through a lifetime of not only "street-language" training but also go through formal training of language through academia and years of contextual training on highly technical terminology. The GPT-2 model, on the other hand, does not learn through the same years of learning, but it still demonstrates abilities to learn from human language. More specifically, this model is trained with the objective of making a prediction based on previously given information.

Even though there exist weaknesses within the model, these weaknesses are also similar to the cognitive aspect of the way humans learn. There exist individual differences in semantic comprehension driven by individuals' differences in their ability to make sense of narratives. In connection to these thoughts, we can think back to children's cognitive capabilities of learning new content through self-assessment. Not only do they ponder back at their previous knowledge and experiences, but they also ponder on the present and the past simultaneously to make predictions.

Contrary to the model's weaknesses, it also has strengths worth noting. The model's learning gaps allow for flexibility to customize the model to one's use. In addition, this model has the ability to outperform other models on tasks related to question answering and common sense reasoning. We can conclude that it contains strengths in context classification.

On the other hand, we have the Bidirectional Encoder Representation from Transformers (BERT), an unsupervised language model, and an open-source learning framework for the natural learning process. This model is unique from other existing models in that during training, it looks to the left, right, or a combination of the two. Furthermore, BERT trains on the dataset and the labels while making predictions based on left and right observations. BERT has the ability to learn contextual relations between input tokens and through the use of "Transformer", a type of encoder that allows the processing of words to take place bidirectionally.

In the general scope of BERT, the size of the data used within the model matters and its performance is noticeable in small-scale tasks. The more training data there is, the more training steps will lead to higher accuracy. BERT's

bidirectional ability might result in slower convergence, but its ability to outperform has been generally notable.

One difference between the way BERT and GPT-2 are set up is the manner in which the datasets are handled. More specifically, what makes BERT's model powerful is its ability to pre-train, allowing the model to learn more. Another key aspect is its ability to be fine-tuned, adding an additional layer to NLP tasks. Furthermore, along with being able to fine-tune the model comes the ability to frequently update the model, allowing the programmer to achieve an optimal fine-tuned model.

Although it contains a handful of key qualities in comparison to GPT-2, it also encompasses weaknesses. Although it is a small drawback, the model is computationally costly in training its structure. Additionally, the model is designed to be used as an input into another system to be fine-tuned, which can result in unforeseen road bumps.

Cognitively, the model has aspects that are similar to human cognition. For instance, BERT has been examined for its link between common sense and question and answer data sets. Because of the bidirectional aspects of BERT, we can see the connection between children's cognitive development and the model itself, where we can observe that children learn through their environment. Not only do they learn things from past experiences, but also from their surrounding environments. Their learning takes place bi-directionally, where the people around them, the experiences that take place, and their cognitive bias permits them to learn bidirectionally. Likewise, the model not only learns through prior experiences but also through posterior experiences analyzed from observations.

In this study, we primed GPT-2 and BERT with "spam" versus "ham" SMS text messages that were collected from the University of Singapore's student population. GPT-2 uses transformer decoder blocks. Within the decoder blocks, we have one main mechanism occurring, where the blocks only have self-attention with respect to the decoder. In other words, GPT-2 only encompasses blocks that are masked in a manner that permits Self-Attention toward previous tokens. On the other hand, we have BERT, where we take advantage of its encoder in order to compare predictions produced by both of these models along with our human-evaluation results.

Data Processing

The data we used is a set of SMS tagged messages collected by the Department of Computer Science at the National University of Singapore for a research study. It contains one set of SMS messages in English with a sample size of 5,574 messages, tagged accordingly as *spam* or *ham*. We then took the original data that included messages with Latin-1 encoding, leading to inferences of plausible confusion when classifying which texts are spam. Dropping rows of data that contained unusual encodings resulted in a total of 4,951 messages, 4,473 of which were *ham* and 478 were spam. After pre-processing and doing a stratified 80-20 split into training and test datasets, we were left with 991 texts to perform human evaluations using 0-1 classifications (0: ham, 1: spam) based on our subjective opinion and compared our human accuracy and correlation with the performance of BERT and GPT-2.

Modeling

For the GPT-2 model, we used Hugging Face's transformer library, implementing the GPT-2 Tokenizer and AdamW optimizer with their setting derived from the pre-trained GPT-2. Previously split, 80% of the data was trained in a batch size of 32 over 4 epochs. The model was also set to truncate over 60 text sequences for training. The remaining 20% of the data was tested to calculate the model's loss and accuracy. GPT-2 model loss result shows no signs of overfitting (Figure 1).

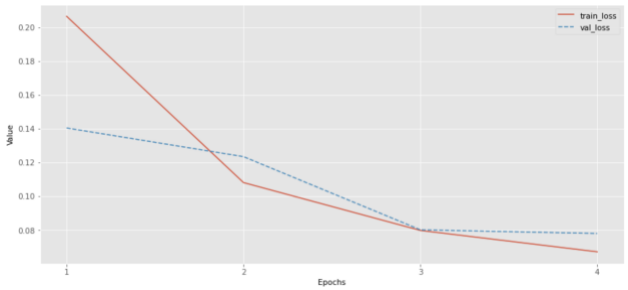

Figure 1. Train and validation loss for GPT-2

For the BERT model, we utilized the Tensorflow library and implemented a BERT-en-uncased variant with Adam optimizer and binary cross-entropy loss. We took a different approach here, training of model by using a down-sampled dataset of 475 spam and 475 non-spam messages. The original data contained a higher count for non-spam messages than spam messages such that if the model was to classify all messages as non-spam, it would still achieve high accuracy. To avoid this mistake, we input an equal amount of training sets for both spam and non-spam messages. Although there may be the disadvantage of possibly decreasing the overall accuracy, this downsampling method makes the BERT model impartial and highly relevant. This model was fitted in 32 batches over 2 epochs, and the original test set of 991 rows was used to predict in order to interpret the prediction data alongside the human evaluation and GPT-2 results.

Results

Results of Human Evaluation

Each group member classified the set of messages into spam or ham. The data concluded accuracy rates of 94.3%, 97.9%, 95.7%, and 87.8%. The mean accuracy of all members combined is 93.9% (Figure 2).

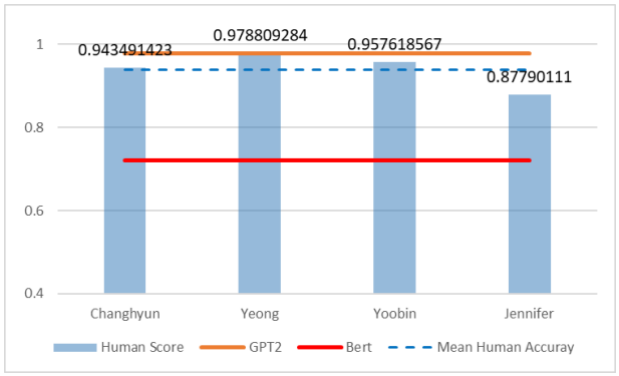

Figure 2. Comparisons among classification accuracies of humans, GPT-2 and BERT

To increase the confidence level of our evaluations, we added a new measure of human evaluation by classifying the message as spam only when at least three of us labeled it as spam. From the four different sets of human evaluations completed by our group members, we combined the results based on how the majority of members (at least 3 out of 4) classified the data. The evaluations agreed by the majority now represent human evaluations and they get compared with the model predictions.

Results on GPT-2

After completing the fine-tuning process with our training data, the GPT-2 outperformed the human predictions. The GPT-2 accuracy was roughly 3.8% higher than the accuracy of human evaluation.

After close observation of the misclassifications in human evaluation, we came up with a few possible causes for misclassification by humans as the following: 1) humans tend to classify subscription-based advertisements as non-spam, 2) humans apply their past experiences when they make predictions, 3) humans are more prone to make careless mistakes than machine models.

One of the examples for the first case in our data is "*Burger King - Wanna play footy at a top stadium? Get 2 Burger King before 1st Sept and go Large or Super with Coca-Cola and walk out a winner" (spam).* It appeared to be that most of these messages were categorized as non-spam by individuals. From a human's perspective, a plausible conclusion for this misclassification is their tone of being a voluntary subscription email. In this case, humans' prior assumptions or experiences can be falsely applied when it comes to classifying a spam message, and there are cases where these additional subjective opinions could negatively affect and lead to misclassification.

For the second case, an example from our data set is *"Today is ACCEPT DAY..U Accept me as? Brother Sister Lover Dear1 Best1 Clos1 Lvblefrnd Jstfrnd Cutefrnd Lifpartnr Belovd Swtheart Bstfrnd No rply means enemy".* Based on prior experiences, well-wishes messages were often fraudulent and came from unknown users so the human evaluation misclassified them as spam even though these messages do not take the form of advertisements or inappropriate words. On the other hand, the majority of human evaluations misclassified the following example, *"2/2 146tf150p",* as non-spam, assuming it could be sent unintended. Here, humans allow room for mistakes which may lead to higher misclassification. In both cases, the language model could classify these messages correctly since these messages do not likely stimulate any spam-related indicator. Some of these misclassified messages also contained words such as "sir", where these types of messages often led to explicit words or content. Messages like these, which contain undirected salutations, raise suspicions for humans which might not raise enough suspicion for GPT-2 to classify as spam.

In the last category of human misclassification, humans are more prone to make careless mistakes than models. We could observe that there are some messages that were relatively easy to classify as spam or ham but the majority of us got it wrong. One example can be *"Someonone you know is trying to contact you via our dating service! To find out who it could be call from your mobile or landline 09064015307 BOX334SK38ch"* and we could clearly say that it is spam but human evaluation misclassified it as non-spam.

Although the GPT-2's prediction resulted in relatively high accuracy (98% accuracy which is almost 4% higher than the human evaluations), we could observe that the model still produces misclassifications at the rate of roughly around 2% of the sample. More specifically, we observed three main categories of misclassification: 1) Punctuation errors, 2) Limitation or inability to carry on patterns of street languages, and 3) Weak spam detection on uncommon structures or formats.

In human reality, punctuation errors or typos are very common so humans have the ability to distinguish between the messages containing these punctuation errors and spam messages. Punctuation errors above include missing a question mark after a questioning sentence or vice versa, or a random period mark appearing in the middle of sentences. However, GPT-2 misclassified these messages as spam.

In the second case of misclassifications, the language model may not capture tones, slang, or street languages as in *"2 much"* or *"2mrw"* and misclassify them as spam. This limitation has great potential to be improved as the model's learning process continues on the newly created languages or trendy abbreviations.

The last possible cause of the model's misclassification is the weakness of structures that are not common or not trained in the learning process. If a certain structure of the spam messages has not been trained in pre-trained or fine-tuning, GPT-2 fails to capture them. As in the model's misclassification example, *"ringtoneking 84484",* might appear like an address to the model and perhaps lead to being unable to distinguish the difference.

Results on BERT

Most of the major misclassification patterns in GPT-2 were also observed in BERT after inspecting its evaluations. However, compared to the GPT-2 model, BERT returned a lower accuracy score in classifying spam messages. BERT seemed to apply a lower threshold when classifying spam messages; whenever text messages have punctuation errors or out-of-dictionary abbreviations, BERT classified them as spam messages. Some examples that BERT misclassified as spam include *"2mro i am not coming to gym machan. Goodnight.", "Hmm ok, i'll stay for like an hour cos my eye is really sore!",* and *"Some friends want me to drive em someplace, probably take a while".* These messages are relatively easy to classify from a human perspective, but they seem to not follow proper grammar and it causes BERT to detect them as spam. What humans would view as a minor grammatical flaw or commonly used abbreviations of words resulted in BERT classifying a message as spam. Considering that BERT was pre-trained using Wikipedia pages - a website that follows formal structural language - and most text messages are written in spoken languages, this result is not surprising. In a similar sense, BERT classified all messages containing text emoji faces such as *"S:)8 min to go for lunch:)" ":-(that's not v romantic!"* and *"Update your face book status frequently :)"* as spam whereas both GPT-2 and humans correctly classified them as non-spam.

Additionally, humans are more likely to classify texts as spam when a sign of urgency is implied in the messages. The following are some examples of messages that humans misclassified as spam while BERT was able to correctly classify them as ham. Some examples include the following: *"Good Morning plz call me sir", "Sir, I am waiting for your mail.", "Hello which the site to download songs its urgent pls".*

One possible explanation for this misclassification by humans is that humans tend to put emphasis on feelings behind text messages that could raise suspicion of fraudulence. This tendency leads them to incorrectly classify texts as spam while BERT correctly classifies them as non-spam.

Due to BERT's leniency in classifying text messages as spam, BERT achieves a relatively high spam detection ratio (95%) but it is only

because it plays safe and classifies even very mildly suspicious messages as spam. Because of this, BERT on the other hand performs poorly (69% accuracy) on correctly classifying non-spam messages.

Discussion

We have informally collected our human data amongst ourselves by randomly shuffling and grabbing 991 samples from the dataset in order to make predictions based on BERT's, GPT-2, and human's ability to correctly predict whether SMS text messages are spam or not spam. Our first assumptions relied on the information provided via the web on how the two models run. More specifically, articles proposed that BERT's bidirectional abilities made it a stronger model, but on the contrary, our data has made other discoveries.

In existing experiments, we have seen that GPT-2 tends to be used for answers/questions, text, summaries of passages, and also to predict the probability of a sentence. On the other hand, BERT is a pre-trained model used for sentence classification. Our results demonstrated that GPT-2 made more accurate predictions for this particular dataset than BERT. One notable factor as to why GPT-2 may have performed better is due to the nature of our data. Although it is text-oriented, it does encompass slang and symbols not used in a common text. Additionally, BERT is pre-trained and has patterned knowledge that is not applicable to our data. For instance, our data contains symbols, unrecognizable text, and discernible messages.

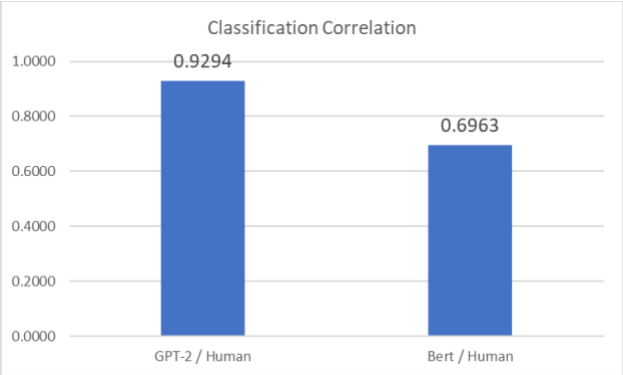

Figure 2. Correlations between classifications by humans vs. classifications GPT-2 and BERT

Aside from accuracy, our results show that human classification in this particular dataset is significantly more correlated with GPT-2 than with BERT (Figure 2). The fact that GPT-2 has been pre-trained with casually written languages whereas BERT has been pre-trained with sources like Wikipedia provides a possible explanation for this difference in correlation between humans vs. BERT and GPT-2.

From the results, some things that stood out the most were GPT-2's ability to outperform BERT. In literature, BERT is commonly defined as using encoders with powerful abilities to make predictions. Its bidirectional abilities brought expectations of BERT outperforming GPT-2, but on the contrary GPT-2's performance proved otherwise.

Our difference in expectations could be explained by cognitive differences and abilities amongst the models. We concluded there were three types of differences in the classification of text messages: within human evaluations, the model's evaluation, and both the model and machine's evaluation. Our cognitive abilities and inferences explain why these differences exist. As mentioned previously, seeing GPT-2's outperformance is logical in that the model was trained on Reddit text. In terms of our data, the

model better represents the informal text our data contains. We could also observe from the experiment that the way each model gets trained may surpass the limitation of the technical structure of the model if it is fine-tuned with enough sophisticated data.

All in all, GPT-2 and BERT both maintain their own strengths and weaknesses. As a result of our research, we were able to see how well these two models perform on SMS text messages. We have observed that from the two models, after fine-tuning, GPT-2 better represents accuracy and human cognition in detecting spam messages. More specifically, GPT-2 encompasses an ability to better represent cognitive abilities regarding informal language. We also know that the model learns through self-assessment, allowing the model to learn from the past and present, a method of learning humans learn to take that begins in childhood. Although it is able to pick up on informal language, we can notably state that the model lacks to pick up on more formal language, creating an inability to pick up on queues of spam emails that carry a more formal tone. Even if the model failed in this sense, it is plausible in that most of the emails that were misclassified by the model were also misclassified by humans. This brings us to the conclusion that there are some messages that still fall in this gray area of unpredictability. In this sense, both the model and humans lack this cognitive ability.

References

Almeida, T.A., GÃ³mez Hidalgo, J.M., Yamakami, A. Contributions to the Study of SMS Spam Filtering: New Collection and Results. Proceedings of the 2011 ACM Symposium on Document Engineering (DOCENG'11), Mountain View, CA, USA, 2011.

Budzianowski, P., & Vulić, I. (2019, August 4). *Hello, it's GPT-2 -- how can I help you? towards the use of pretrained language models for task-oriented dialogue systems*. arXiv.org. Retrieved May 10, 2022, from https://arxiv.org/abs/1907.05774

David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen, arXiv: 1907.09177v2, Dec 2019.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May 24). Bert: Pre-training of deep bidirectional Transformers for language understanding. arXiv.org. Retrieved May 10, 2022, from https://arxiv.org/abs/1810.04805

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (n.d.). *Language models are unsupervised multitask learners - openai*. OpenAI. Retrieved May 11, 2022, from https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

UCI Center for Machine Learning and Intelligence Systems (2012, June 22). *SMS Spam Collection Data Set (classification)*. UCI. Retrieved May 8, 2022, from https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection