A Replication of: Alex Warstadt and Samuel R. Bowman. 2020. Linguistic Analysis of Pretrained Sentence Encoders with Acceptability Judgments

Eli Heakins **Rahul Nair**

Dept. of Computer Science
Colgate University
13 Oak Drive
Hamilton, NY 13346

Abstract

This paper takes a look at the ability of pretrained Text-Classification models (i.e. GPT2 and BERT) to evaluate sentences as grammatical. This is done through the COLA annotated dataset, which looks at multiple grammatical structures in English. This is all done referring to a separate paper for its methods and data-set(Warstadt and Bowman, 2020). We found similar results to the original paper for our BERT models, however, it failed to capture similar results for GPT-2.

1 Introduction

The original paper (Warstadt and Bowman, 2020) attempted to evaluate TextClassification models (BERT and GPT-2) on their COLA-annotated data set. This was to see how these models performed overall at TextClassification over a wide dataset and how these models were able to do with TextClassification of different grammatical structures present in English. This paper looked to solidify COLA as a dataset useful in evaluating TextClassification models and to see how well these commonly used models classified text.

The original paper found these models achieved near human TextClassification, with BERT achieving better metrics than GPT-2. They also found that these models were able to achieve classification with simpler grammatical structures, but struggled with more complex ones.

We wished to replicate this entire pipeline, with the same data sets. We wished to see if the NLP-scholar pipeline produced similar data to the original paper.

Much like the original paper we found that BERT performed better than GPT-2 at evaluation of the COLA-annotated set, with similar results to the original paper. However, our GPT-2 model was ineffective at this task with low performance metrics. Also mirroring the results of the original paper we found both models performed far better at the simpler grammatical structures, and struggled with the more complex ones.

2 Background

Many previous evaluations of TextClassification language models have used different data sets. One of the authors of the paper we are replicating created a data set called the Corpus of Linguistic Acceptability (COLA) which covers a wide range of grammatical structures. These structures were later annotated to show which grammatical structure they were using.

The original paper trains both the BERT and GPT-2 models on this data set and then uses a separate section for testing these models. This allowed them to identify which grammatical structures these models either succeeded with classification or struggled with.

We replicated both the training on the same data set, using the two models. However, we differed by using the NLP-scholar pipeline for training instead of their specific training methods. We then analyzed and evaluated the same dataset, again using the NLP-scholar pipeline.

3 Methods

We used the NLP-Scholar pipeline to train these two models. We then used the pipeline to evaluate these trained models. Then we employed a Python script to compile data for analysis. Our Python code, data sets, and evaluation data are found in (Heakins and Nair, 2024).

3.1 Models

We used the same models that the original paper used, that being GPT-2 and BERT. We trained each of the models three times on the training dataset. Each of these six models were used for analysis.

3.2 Datasets

We used the same dataset used in the original paper, the COLA-annotated data set. Present was both a test and a training data set. The testing data set is over 1000 sentences, with 13 grammatical structures being annotated. These features are not exclusive for the sentences, with multiple being present in most of them. We had to make some changes to the data-sets for the use of it in the NLP-Scholar pipeline. However all of these changes were superficial, just resulting in the renaming of some of the fields, no changes to the actual data were made. We used the COLA annotated dataset to create both test and validation sub-datasets. The validation dataset represents 10% of the data.

3.3 Evaluation

We used the NLP-scholar pipeline to evaluate our data, using 3 different trained models for both BERT and GPT-2. We evaluated over the COLA-annotated dataset. This dataset has sentences considered grammatical and ungrammatical with labels attached to each sentence to look at specific grammatical structures. These labels were ignored in the evaluation but were used for the analysis.

3.4 Analysis

To evaluate the performance of the three trained GPT models and three trained BERT models on our evaluation set, we analyzed their ability to classify sentences as grammatically correct or incorrect. We compared the models' predictions to the true labels from the evaluation set and utilized several metrics from the scikit-learn library to assess their performance. First, we computed the Matthews Correlation Coefficient (MCC) for each model using scikit-learn's matthews_corrcoef function. The Matthews Correlation Coefficient (MCC) is a performance metric that measures the quality of binary classifications by considering true and false positives and negatives and combining them all in an equation. To calculate the actual number of true positives, true negatives, false positives, and false negatives for each model's predictions we applied sci-kit-learn's confusion_matrix function. In addition to the MCC, we employed scikit-learn's classification_report function to generate a detailed report for each model, including key metrics such as precision, recall, and F1 score. To analyze the models' performance on specific grammatical features, we used a supplementary dataset that annotated each sentence in our evaluation set with the grammatical points it employed. We scripted a process to isolate sentences that used particular grammar points from both the evaluation set and the models' result sets. For each grammatical feature, we again calculated their MCC using scikit-learn's matthews_corrcoef function. Finally, for both the GPT and BERT models, we averaged all the results from the three different trained versions of each model to obtain a more general assessment of their performance. We used pandas to read, filter, and manipulate the evaluation sets and the model results.

4 Results

Overall performance The BERT pre-trained model was evaluated far more effectively than the GPT-2 over the entire data set. We can see this from the MCC in Table 1. Comparing to the original paper we see how BERT model has reasonably similar performance (0.528 mean MCC in the original). However the GPT-2 model performed far worse in our replication (0.528 for GPT-2 in the original paper) (Warstadt and Bowman, 2020).

	Mean	Max
BERT	0.549	0.564
GPT-2	0.192	0.309

Table 1: Performance (MCC) on the CoLA test set, including mean and max of a given model

BERT model performance Due to the issues with the GPT-2 model performance we will look at the performance of the BERT model as compared to the original paper. Looking at Figure 1 we can see the performance of the model at specific grammatical structures. As per the original paper, we see the model does the best at the *Simple* sentences with a MCC above the overall mean. All other features were scored either at mean or below mean, showing the struggle with more complex sentences. Like the original paper, we also see the model does well at argumentative grammatical features (Argument Type and Arg. Altern) reiterating the original paper's thoughts that the model trains effectively on this feature.

Poor performance of GPT-2 model Given our much lower performance numbers for our GPT-2 model than the original paper, we can reason that there was some flaw with our training on the

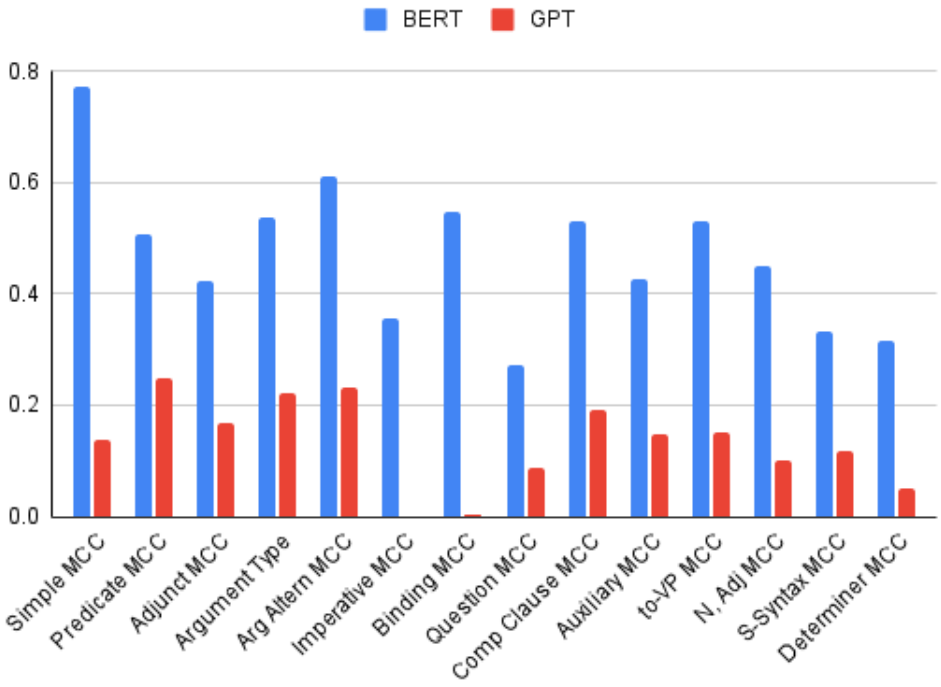

Figure 1: Performance (MCC) on our analysis set by major feature. These were calculated as per our description in section 3.4.

GPT-2 model. Looking closer at this with Table 2, we can see that both models, but more specifically GPT-2 exhibits extremely high positive recall and much lower negative recall. This suggests the model struggles to correctly identify negative cases, leading to a high false positive rate. This point is also echoed in Table 3. with a much lower Negative F1 for both models, showing an overall lack of ability to determine sentences to be ungrammatical. This leads us to believe that the model is over-determining the sentences to be grammatical. Given our training methods were different (using the NLP-scholar pipeline) we believe that this is more likely to be a flaw in the training of our model. Due to the issues with the GPT-2 performance, those results can mostly be ignored.

	+ Precision	- Precision	+ Recall	- Recall
BERT	0.820	0.804	0.940	0.545
GPT-2	0.710	0.679	0.986	0.113

Table 2: Mean Recall and Precision for both positive (+) and negative (-) on the CoLA test set of a given model

	Positive F1	Negative F1
BERT	0.875	0.650
GPT-2	0.825	0.188

Table 3: F1 for both positive and negative on the CoLA test set of a given model

5 Discussion

This paper looked to analyze the effectiveness of BERT and GPT-2 at evaluating sentences as grammatical over the COLA-annotated data set to verify findings in the reference paper.

Our GPT-2 model did not have strong results, being extremely dissimilar to what was found in the original paper. Thus we concluded there was some flaw in the training specific to that model. We found similar performance for our BERT models as compared to the original paper.

A major limitation in our replication of the original paper was that we did not use the same training methods. The paper spends extensive time explaining the methods they used for training. We simply used the NLP-scholar pipeline in its standard configuration, holding to none of the

methods of the original paper. This could explain the results we had for the GPT-2 model.

Throwing away the results from the GPT-2 models, we found that BERT trained and evaluated on this data set resulted in high performance metrics for evaluating grammaticality. This confirms the finding of the original paper, that BERT trained on this data set can determine the grammaticality of sentences at a near-human level, being stronger at simpler sentences.

References

Eli Heakins and Rahul Nair. 2024. Cosc 426 paper replication.

Alex Warstadt and Samuel R. Bowman. 2020. Linguistic analysis of pretrained sentence encoders with acceptability judgments.