Demonstrating the Utility of the Annotated CoLA Set via Linguistic Analysis of Pretrained Sentence Encoders with Acceptability Judgements

Jose Paredes-Larios and **Jacob Denekamp** and **Christian Xique**

Abstract

The use of data sets is central to the improvement of models used for NLP tasks. The annotated CoLA set addresses the void left by other data sets by providing sentences that span grammatical phenomena, while also labeling sentences with the phenomena they exhibit. Warstadt and Bowman demonstrated the utility of the data set by uncovering weaknesses in BERT and GPT's understanding of grammar. To validate their findings and see if successors to BERT and GPT could outperform them, roBERTa and GPT2-Med were evaluated across 4 major (Predicate, Arg Alt, N, Adj, and Aux) and 4 minor phenomena (Copula, PP Arg VP, Coord, and High Arity). We found that roBERTa and GPT2-Med possess a similar level of grammatical knowledge. Furthermore, both models struggled with phenomena involving long range dependencies, while also failing to best their predecessors. Our findings help justify the functionality of the annotated CoLA set, as it helped unveil several aspects of roBERTa and GPT2-Med's grasp of grammar.

1 Introduction

There is no denying that natural language processing (NLP) is deeply integrated into the lives of millions. NLP enables the functionality of search engines, chatbots, smartphones, social media, and numerous other things (Otter et al., 2021). Thus, it is key for the models involved in NLP, such as BERT and GPT, to possess a solid understanding of grammar. Edifying these models involves leveraging datasets to evaluate the grammatical knowledge of the models. In other words, the advancement of models is closely tied to the development of datasets.

Currently, the majority of datasets used to evaluate models are highly focused on a single grammatical feature, such as subject-verb agreement. These types of datasets excel at assessing a model's understanding of a specific grammatical feature but fail to test a model's general grammatical grasp (Warstadt and Bowman, 2019).

The Corpus of Linguistic Acceptability (CoLA) [1] begins to address the issue of highly focused data sets by consisting of thousands of sentences that span several grammatical domains. However, CoLA falls short in that it does not distinguish between grammatical features. As such, the data cannot be used to asses a model's comprehension of a distinct feature. Warstadt and Bowman tackle CoLA's greatest limitation by labeling sentences with expert annotations that signal if the sentence exhibits a major/minor grammatical phenomenon or not [2].

To prove the worth of the annotated CoLA data set, Warstadt and Bowman leveraged it to run a text classification experiment on BiL-STM, which served as the base-line, and top-of-the-line models, $\text{BERT}_{\text{large}}$ and GPT. Each model was trained on CoLA to produce twenty classifiers (e.g. from BERT, BERT1, BERT2, ..., and BERT20 were created). Each classifier was tasked with labeling groups of sentences from the annotated CoLA as grammatically acceptable or not. Lastly, the predicted labels were utilized to obtain the Matthews Correlation Coefficient (MCC). The approach taken by the authors is commendable, for they understood that they could not just release a unique data set and expect people to incorporate it into their experiments due to its distinctness alone. Running an experiment with the data set highlights how invaluable results can be extrapolated by using their data set.

Warstadt and Bowman found that the transformer models BERT and GPT possess an erudition for local phenomena but struggle with long-distance dependencies. These findings signal that

[1] CoLA can be downloaded here: https://nyu-mll.github.io/CoLA/

[2] The annotated CoLA set can be downloaded here: https://nyu-mll.github.io/CoLA/grammatical annotations

BERT and GPT are not perfect models; there are several aspects of grammar that they do not fully comprehend. Uncovering this weakness is pivotal because it allows the creators of the models to know what areas need to be edified. These results could not be obtained as easily without annotated CoLA; furthermore, this is only one application of it. Indeed, the data set facilitates exciting discoveries and helps enable progress.

Replicating experiments is essentially dogma in science. It increases the accuracy, reliability, and validity of previous findings. It also allows one to innovate by changing some of the methodology of the original experiment. Thus, we decided to replicate Warstadt and Bowman's experiment.

The original paper did not involve multiple experiments. Rather, it was one experiment that assessed 3 models across different grammatical features. We conducted the same experiment, but we down-scaled a couple of aspects. We are only evaluating 2 models, creating 5 classifiers for each model, and assessing the models across 4 major and 4 minor grammatical phenomena. We ensured that the phenomena we selected covered local and long range dependencies. Furthermore, we wanted to showcase the utility of annotated CoLA by gauging the grammatical knowledge of newer models not used in the original paper, roBERTa and GPT2-Med.

Thanks to the annotated CoLA set, we found that roBERTa and GPT2-Med possess a similar and solid understanding of grammar, with a mean MCC value across classifiers around 0.5. Furthermore, we observed that both models are more proficient at dealing with local phenomena than phenomena involving long range dependencies. Lastly, We did not find concrete evidence that roBERTa and GPT2-Med are better at text classification than their predecessors, BERT and GPT, for all four models possessed homogeneous mean MCC values across classifiers.

Our findings are not absolute and should be approached with a bit of caution because of the down scaled nature of our experiment. Nevertheless, we obtained results that resembled Warstadt and Bowman's findings. In other words, we justify the creation and existence of the annotated CoLA set. It allowed us to assess state-of-the-art models across numerous phenomena and identify points of weakness. Indeed, the annotated CoLA set is a powerful tools that enables the advancement of models.

2 Background

To replicate Warstadt and Bowman's work and enhance the clarity of subsequent parts of our paper, we will briefly discuss several key topics.

2.1 CoLA Data Sets

There are two different versions of CoLA we leveraged to conduct our experiment. CoLA is composed of 10,000 sentences, which have been expertly labeled as grammatically acceptable or not. The sentences were obtained from 23 theoretical linguistics publications; thus, they cover several aspects of grammar (Warstadt and Bowman, 2019). Furthermore, CoLA is split into a training set, in-domain dev set, and out-domain dev set. The training set is used to train models, while the dev set is used as validation during the training process.

Annotated CoLA is the one we are most interested in, for we are trying to justify its existence. It consists of every sentence in both of the CoLA dev sets, which totals to 1043 sentences. Each sentence was manually annotated by one of the authors of the original paper (Warstadt and Bowman, 2019). More over, annotated CoLA comes in two flavors, major and minor. Major labels the major grammatical phenomena exhibited by the sentence, while minor labels the minor grammatical phenomena. There are a total of 13 major and 59 minor grammatical phenomena.

2.2 Grammatical Phenomena

We evaluated both models on 4 major (predicate, argument alternation, noun adjectives, and auxiliary) and 4 minor (copula, prepositional phrase argument verb phrase, coordination, and high arity) grammatical phenomena. Let us quickly discuss what each phenomena entails.

2.2.1 Predicate

Predicates are simply the part of the sentence that informs us about what the subject is doing (Warstadt and Bowman, 2019). In "the panda is using strings to learn", "is using strings to learn" is the predicate, for it gives insight into what the panda is doing.

2.2.2 Argument Alternation

Argument alternation refers to the fact that the arguments of verbs can be altered while preserving the overall meaning of the sentence (Warstadt and Bowman, 2019). Consider: "the panda ate the sandwich" vs "the sandwich was eaten by the panda".

Both sentences convey the idea of a panda eating a sandwich, but the first sentence is in active voice, while the second one is in passive voice. Note that there are different types of argument alternations.

2.2.3 Noun and Adjectives

This major grammatical phenomenon is quite broad and encompasses all minor phenomena centered around nouns and adjectives, such as noun-noun compounds, relational nouns, complex noun prepositions, etc. (Warstadt and Bowman, 2019).

2.2.4 Auxiliary

Auxiliary sentences involves the presence of helper verbs, such as "be", "can", "have", etc., which support the main verb (Warstadt and Bowman, 2019). In "the panda can eat sandwiches", "can" signals that the panda has the ability to eat sandwiches.

2.2.5 Copula

Copula sentences implement copular verbs, such as "is", "was", "are", etc., to link a subject to a subject complement (Warstadt and Bowman, 2019). In "Maxwell is quite a doctor", "is" links Maxwell to the subject complement "quite a doctor".

2.2.6 Prepositional Phrase Argument Verb Phrase

Here a prepositional phrase (PP) serves as an argument to a verb phrase (VP) (Warstadt and Bowman, 2019). In "the panda studied in the cafe", "in the cafe" is the PP that is an argument to the VP "the panda studied".

2.2.7 Coordination

Coordination sentences involve a coordinating conjunction, such as "and", "but", "or", etc., linking two elements. In "the panda and the goose ran", "and" links two nouns. Note that the two elements that are linked must be syntactically equivalent (i.e. verb to verb and noun to noun) (Warstadt and Bowman, 2019).

2.2.8 High Arity

High arity sentences implement verbs that take at least 3 arguments (Warstadt and Bowman, 2019). In "the panda ate his sandwich in the park", "ate" takes the arguments "panda", "his sandwich", and "in the park".

2.2.9 Sentence Encoders and Classifiers

BERT and GPT are pretrained sentence encoders, which turn sentences into embeddings. We can think of it as a model that takes human language and translates it into something a machine can understand, an embedding. Embeddings are simply vectors that capture the linguistic features of a sentence (Le and Mikolov, 2014).

BERT is bidirectional, which means it considers the context to the left and right of words when generating its embeddings (Devlin et al., 2018). On the contrary, GPT is unidirectional; therefore, it only evaluates the context to the left of words when synthesizing its embeddings (Radford and Narasimhan, 2018).

We can use these sentence encoder to help produce a classifier by training it. During training, we provide the classifier with labeled sentences; the classifier learns to associate certain patterns in embeddings with specific labels. Thus, trained classifiers can apply what they learn and predict labels for a given sentence.

2.3 Matthews Correlation Coefficient

MCC values serve as a measure of how well a model can correctly predict a label. These values range from -1 to 1, where -1 signifies that that model predicted the wrong label for every sentence, while 1 signals that the model correctly predicted every label.

3 Methods

Replicating a portion Warstadt and Bowman's text classification experiment involved a couple of main steps. It should be noted that NLPScholar was utilized to help us conduct the experiment by training and evaluating models (Lhoest et al., 2021)(Prasad and Davis, 2024)(Wolf et al., 2020).

For both RoBERTa and GPT2-Med, five classifiers were created by training the models on CoLA. Next, evaluation sets for different grammatical phenomena were produced via annotated CoLA. The classifiers were then evaluated on the aforementioned evaluation sets to produce output files. Each output file contained a model's predictive label; as such, an MCC value was calculated for each output file. Lastly, for each set of output files produced for a given phenomenon by a model, an ensemble MCC value was computed. Detailed steps of our experimental protocol have been published [3].

[3]The pipeline we followed and code we utilized to generate results can be found here: https://github.com/Jose-Paredes-Larios/Leveraging-Annotated-CoLA-to-Assess-roBERTA-and-GPT2.git

3.1 Models

Warstadt and Bowman's experiment implemented BiLSTM, $\text{BERT}_{\text{large}}$, and GPT. However, because NLPScholar does not support long short term memory (LSTM) models, the BiLSTM model was not incorporated into our work. Furthermore, we decided to conduct our experiment with GPT2-Med, for it is generally regarded as an improvement to GPT. GPT2-Med contains significantly more parameters and was trained on a larger and more diverse data set than GPT (Radford et al., 2019). Thus, we wanted to observe if GPT2-Med could out do GPT's performance in the original paper. Similarly, we also opted for RoBERTa over $\text{BERT}_{\text{large}}$, for it outperform $\text{BERT}_{\text{large}}$ in several natural language understanding tasks (Liu et al., 2019). As such, we were curious if RoBERTa would yield higher MCC values than $\text{BERT}_{\text{large}}$.

3.2 Datasets

Our experiment could not be conducted without datasets to train and evaluate the models on. Because the main purpose of our experiment is to spotlight the utility provided by annotated CoLA, the set was used to help us generate our evaluations sets. CoLA was employed to train the models. Both of these data sets were also implemented in Warstadt and Bowman's experiment.

3.2.1 Training the Models

From CoLA, the train and in-domain dev sets were used as the training and validation data sets, respectively. The raw versions were implemented, for the models tokenize the data sets themselves. Because of how NLPScholar works, slight modifications had to be made to the original train and dev files; a header was added to each that designates columns two and four as "label" and "text", respectively. "text" contains the sentences of interest, while "label" indicates if the sentence is grammatical or not. A 1 signals that the sentence is grammatical, while a 0 signifies that it is ungrammatical. The altered sets were then used to train RoBERTa and GPT2-Med via a training config file that was ran with NLPScholar. For each model, 5 classifiers were created.

Note that Warstadt and Bowman trained 20 classifiers for each model. We settled on 5, as we are running a scaled down version of the original experiment.

3.2.2 Choosing Grammatical Phenomena

In the original paper, each model was evaluated across all 13 major and 59 minor grammatical phenomena. Because we are not replicating the entire experiment, we decided to evaluate the models across 4 major and 4 minor phenomena. Nevertheless, we did not select the phenomena at random. For both major and minor, we picked 2 phenomena that are typically local (predicate, argument alternation, copula, and PP argument VP) and 2 that involve long range dependencies (noun adjective, auxiliary, coordination, and high arity). The original paper made note of BERT's and GPT's struggles with long range dependencies relative to local phenomena. As such, we wanted our selected phenomena to cover both domains, so that we can see if we observe a similar pattern.

3.2.3 Evaluating the Models

The data sets we implemented to evaluate our models are all derived from annotated CoLA. To obtain an evaluation set, both the annotated CoLA major and minor files were altered. The "source" and "domain" columns were removed, for they were not relevant to our experiment. The "sentence", "acceptability", "[phenomenon]" columns were renamed to "text", "target", and "condition", respectively. Furthermore, a new column was added in the beginning called "text id", which simply enumerates the sentences. All of these changes were made so that the evaluation sets could be used in NLPScholar.

For a given phenomenon, its respective column was kept and the rest were deleted. Any sentence that did not contain a one in the phenomenon column was deleted, for it does not contain the phenomenon of interest.

Consider data set X with 10 sentences and data set Y with 100 sentences. Say a classifier only mislabels ten sentences for each data set. The classifier will perform much better with data set Y. However, is this because it possesses a better understanding of the type of sentences in Y or is it because X has so few sentences? In other words, labeling sentences correctly/incorrectly in smaller data sets has a larger consequence that can skew the data. To circumvent this concern, every major set was made to contain 250 sentences. None of the minor phenomenon contain 250 sentences, which is why each minor set only contains 150 sentences.

Once an evaluation set for a given phenomenon was created, it was leveraged to evaluate all ten

classifiers. Each classifier yielded an output file that contained its predictive labels.

3.3 Analyzing the Models

The output file created via the evaluation of a classifier contains the acceptability of a sentence given by an expert and the classifier itself. There are four combinations of acceptability that can arise (1 indicates acceptable and 0 signals unacceptable):

- Expert: 1 Model: 1 (True Positive (TP))
- Expert: 0 Model: 0 (True Negative (TN))
- Expert: 0 Model: 1 (False Positive (FP))
- Expert: 1 Model: 0 (False Negative (FN))

From these combinations we can compute an MCC value via the following formula:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

For a given phenomena and model, the MCC value produced by the classifies was averaged. Thus, eight averages were computed for a given model. The overall average MCC value of a model was calculated by taking the average of the eight averages computed.

The classifiers are not identical; thus, some will outperform others in certain phenomenon. One model's poor performance can bring down the overall performance of the model. To account for this, we can compute the ensemble MCC value of a model. Calculating ensemble is a matter of determining what label the majority of the classifiers predicted. For example, if four classifiers predicted that the sentence is acceptable and one did not, we consider the overall predictive label as acceptable. For every sentence in a given phenomena, the majority predictive labels are compared to the expert labels to compute an MCC value, which is our ensemble.

Obtaining MCC values in this experiment is pivotal because it allows us to assess the grammatical knowledge of models across phenomena. For example, we can observe what aspects of grammar roBERTa/GPT2-Med excel at and struggle with. Uncovering and documenting these weaknesses are the first steps to improving the models. Such discoveries help us prove the utility of the annotated CoLA set, a tool that can be implemented to assess a model's understanding of numerous parts of grammar.

	Mean (STD)	**Max**	**Ensemble**
roBERTa	0.514 (0.114)	**0.672**	0.530
BERT	**0.582 (0.032)**	0.622	**0.601**
GPT2-Med	0.457 (0.129)	0.605	0.475
GPT	0.528 (0.032)	0.575	0.567

Table 1: Each model's performance (MCC) on the annotated CoLA set, which includes the mean, max, and ensemble across the classifiers for a given model. Results for BERT and GPT were taken from Warstadt and Bowman

4 Results

4.1 Overall Performance

Table 1 provides an overview of each model's performance. For each grammatical phenomenon, the mean MCC across the classifiers of a model was computed. The mean of means give rise to the overall mean of the model. Max is simply the highest MCC average acquired across a model's classifiers. Lastly, The ensemble of a model is the average of ensemble values obtained for each phenomenon.

We will notice that roBERTa slightly outperformed GPT2-Med in every measure. However, the difference between the two models is not statistically significant (p=0.371). Thus, it would be more appropriate to claim that the models perform similar to one another. Both BERT and GPT surpassed roBERTa and GPT in every aspect, except for max, where roBERTa and GPT scored higher than their predecessors. Overall, BERT appears to be the most proficient in grammatical knowledge.

4.2 Performance across Phenomena

As highlighted by table 1, the successors to BERT and GPT performed worse, which is a bit strange at first glance. However, we must keep in mind that roBERTa and GPT2-Med were not evaluated across all major and minor phenomena. Therefore, there are a lot phenomena that the model may have excelled at, which would have increased its performance. Of course, the opposite is also true in that there are several phenomena that the model was not evaluated on that would have tainted its overall performance. With all of this in mind, we can obtain a better understanding of how the models compare to their predecessors by comparing their mean performance for a given phenomena.

As illustrated by figure 1, roBERTa managed to outperform every other model in almost every phenomena it was evaluated on. It's predecessor,

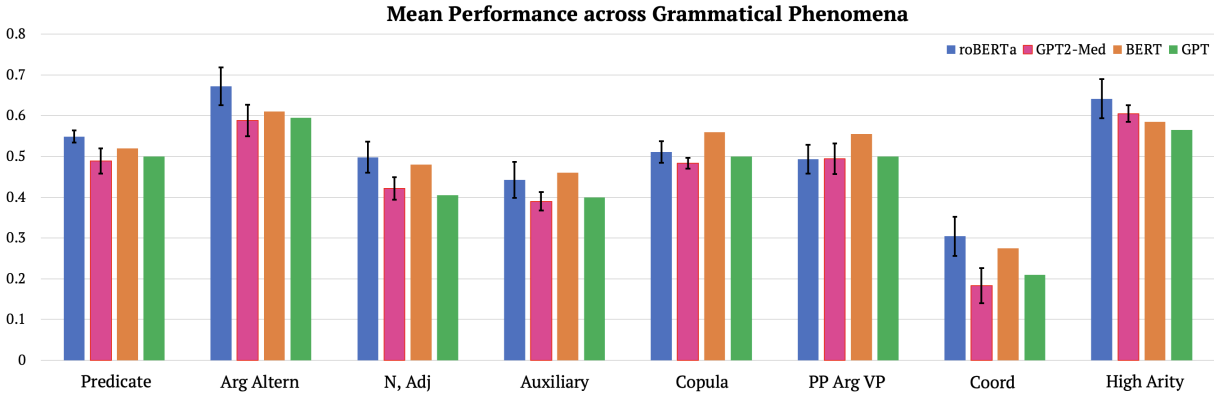

Figure 1: The mean MCC values across classifiers for a given grammatical phenomena. Error bars represent the standard deviation for a phenomena. Because we do not have direct access to the Warstadt and Bowman's data, error bars could not be displayed for BERT and GPT.

BERT, bested it in auxiliary, copula, and PP Arg VP. As for GPT2-Med, it consistently performed worse than GPT, as it only outperformed it in N, Adj and high arity. It appears that roBERTa is a slight improvement to BERT, while GPT2-Med fails to meet expectations.

It must be emphasized that we should approach the data in figure 1 with caution. The mean performance of BERT and GPT is across 20 classifiers, not 5 like roBERTa and GPT2-Med. Currently, we have a solid idea of roBERTa and GPT2-Med's knowledge of a certain phenomena, but it could be refined with more classifiers. Additionally, when we state that a model does better than another for a given phenomena, it is not by a drastic amount. We will notice that a model and its predecessor performed at very similar levels for each phenomena we assessed.

4.3 Local vs. Long Range Dependencies

Figure 1 also signals the idea that the models have a better understanding of local grammatical phenomena than those that involve long range dependencies. Every local phenomena (Predicate, Arg Alt, Copula, and PP Arg VP) yielded a higher mean MCC value for the models than long range phenomena (N, Adj, Aux, and Coord) with one glaring exception, high arity.

High arity was one of our long range phenomenon, yet it produces some of the highest MCC values for each model. High arity is often considered a long range phenomenon because it involves a verb and multiple arguments. However, these sentences frequently bundle the arguments near the verb. Consider this high arity sentence from the evaluation set: "The sailors rode the breeze clear of the rocks". Notice that the arguments for "rode" ("The sailors", "the breeze", "clear of the rocks") span the entire sentence, yet there are no words in between them and the verb. Perhaps we should consider high arity more of a pseudo long range phenomenon. Lastly, the other phenomena that yielded high MCC values across the models was Arg Altern. Note that high arity is a member of Arg Altern. Thus, it appears that all of the models have an erudition for argument structure.

5 Discussion

The NLP field is ever-expanding and is making many aspects of life much more convenient. At the forefront of the field are models that conduct NLP tasks. To improve the field, we must edify the models carrying out the work. One way to improve these models is by unveiling their weaknesses, which can be accomplished with data sets. Warstadt and Bowman acknowledge this idea and created the annotated CoLA set, a data set that allows models to be assesses across several grammatical features. To prove the worth of the annotated CoLA set and validate Warstadt and Bowman's findings, we evaluated state-of-the-art models, roBERTa and GPT2-Med, across 4 major (Predicate, Arg Alt, N, Adj, and Aux) and 4 minor (Copula, PP Arg VP, Coord, and High Arity) grammatical phenomena contained within the data set.

We found that roBERTa slightly outperforms GPT2-Med, BERT, and GPT based off of mean MCC and Ensemble values across classifiers. Additionally, BERT based models marginally bested GPT based models. Across the phenomena we evaluated roBERTa and GPT2-Med, every model

performed incredibly similar to one another. There was no model that dominated the others in grammatical knowledge. With all of this in mind, we believe it is appropriate to say that all models possess a very similar understanding of grammar, with a mean MCC across classifiers all around 0.5, which signals that the models predicted correct labels more often than not. Thus, we cannot state that roBERTa and GPT2-Med outperformed their predecessors.

Similar to Warstadt and Bowman, we found that the models performed worse when evaluated on long range phenomena (N, adj, Aux, and Coord). The mean MCC values across these phenomena were lower when compared to mean MCC values associated with local phenomena (Predicate, Arg Alt, Copula, and PP Arg VP), which reinforces the idea that transformers struggle with long range dependencies. One exception to this finding was high arity, which is a long range phenomena that every model performed exceptionally well on. However, we attributed this anomaly to the idea that high arity sentences all follow a similar structure of grouping a verb closely to its arguments.

As alluded to a couple of times already, the greatest limitation of our experiment is that it is significantly down scaled from the original. We only trained 5 classifiers for each model instead of 20. A single classifier performing incredible well or inadequately will have a greater impact on the mean MCC value of a model. Additionally, we only assessed the model across 8 grammatical phenomena. It is entirely possible that we bypassed certain phenomena that roBERTa and GPT2-Med excel at or perform poorly in, which would alter their mean MCC value.

Although we have demonstrated that the annotated CoLA set is a useful tool for assessing a model's grammatical comprehension across different domains, it is not without flaws. One of the biggest problems plaguing the data set is that not all features possess a similar number of sentences. For example, some minor features can have as little as 22 sentences while others contain around 200 sentences. As such, it is not a fair comparison when comparing a model's performance across such phenomena. Another limitation of the set is that its sentences are in standard American English. The set cannot be leveraged to analyze a model's understanding of dialects. Lastly, we also cannot ignore the fact that the sentences in the set lack context. Context is important to consider, for a sentence may be acceptable in one context but not in another.

There is no doubt that the annotated CoLA set is invaluable tool for those looking to evaluate models, despite its flaws. The set allowed us to demonstrate that roBERTa and GPT2-Med exhibit a similar and solid grasp of several grammatical features. It also helped us illustrate that roBERTa and GPT2-Med's comprehension of long range phenomena could be edified. Indeed, our distilled experiment justified the existence of the annotated CoLA set. It is exciting to think of the advancements of models the set will help come to fruition.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Quoc V. Le and Tomás Mikolov. 2014. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. 2021. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):604–624.

Grusha Prasad and Forrest Davis. 2024. Training an NLP scholar at a small liberal arts college: A backwards designed course proposal. In *Proceedings of the Sixth Workshop on Teaching NLP*, pages 105–118,

Bangkok, Thailand. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alex Warstadt and Samuel R. Bowman. 2019. Grammatical analysis of pretrained sentence encoders with acceptability judgments. *CoRR*, abs/1901.03438.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.