THERMOSTAT: A Large Collection of NLP Model Explanations and Analysis Tools

Nils Feldhus   **Robert Schwarzenberg**   **Sebastian Möller**

German Research Center for Artificial Intelligence (DFKI)

Speech and Language Technology Lab, Berlin, Germany

`{firstname.lastname}@dfki.de`

Abstract

In the language domain, as in other domains, neural explainability takes an ever more important role, with feature attribution methods on the forefront. Many such methods require considerable computational resources and expert knowledge about implementation details and parameter choices. To facilitate research, we present THERMOSTAT which consists of a large collection of model explanations and accompanying analysis tools. THERMOSTAT allows easy access to over 200k explanations for the decisions of prominent state-of-the-art models spanning across different NLP tasks, generated with multiple explainers. The dataset took over 10k GPU hours ($>$ one year) to compile; compute time that the community now saves. The accompanying software tools allow to analyse explanations instance-wise but also accumulatively on corpus level. Users can investigate and compare models, datasets and explainers without the need to orchestrate implementation details. THERMOSTAT is fully open source, democratizes explainability research in the language domain, circumvents redundant computations and increases comparability and replicability.

1 Introduction

Deep neural networks are state-of-the-art in natural language processing (NLP) but due to their complexity they are commonly perceived as opaque (Karpathy et al., 2015; Li et al., 2017). For this reason, explainability has seen heightened attention in recent years (Belinkov and Glass, 2019; Wallace et al., 2020; Danilevsky et al., 2020).

A prominent class of explainability methods, referred to as *feature attribution methods* (in the following used interchangeably with *explainers*), attributes the output of a complex model to its input features. Feature attribution methods arguably have become a cornerstone of explainability research in NLP: For example, Arras et al. (2016, 2017); Atanasova et al. (2020); Chen et al. (2021); Neely

et al. (2021) analyze different model architectures with feature attributions.

There is now also a large body of work comparing explainers in the language domain. Explainers are compared with count-based metrics (Poerner et al., 2018; De Cao et al., 2020; Tsang et al., 2020; Nguyen and Martínez, 2020; Bodria et al., 2021; Ding and Koehn, 2021; Yin et al., 2021; Hase et al., 2021; Kokhlikyan et al., 2021; Zafar et al., 2021; Sinha et al., 2021) and against human judgement (Nguyen, 2018; Lertvittayakumjorn and Toni, 2019; Hase and Bansal, 2020; Prasad et al., 2020). Feature attribution scores have also been incorporated into model training (Ross et al., 2017; Liu and Avci, 2019; Erion et al., 2021; Pruthi et al., 2020).

The feature attribution maps produced and used in the above cited works arguably are the most crucial component of the studies. Unfortunately, none of the above cited papers explicitly links to the generated attribution maps. Easy access to a wide variety of such feature attribution maps, across models, datasets and explainers, however, holds a lot of potential. A central hub

1. would increase the comparability and replicability of explainability research,
2. would mitigate the computational burden,
3. would mitigate the implementational burden since in-depth expert knowledge of the explainers and models is required.

Put differently, a central data hub containing a wide variety of feature attribution maps and offering easy access to them would (1) democratize explainability research to a certain degree, and (2) contribute to green NLP (Strubell et al., 2019) and green XAI (Schwarzenberg et al., 2021) by circumventing redundant computations.

For this reason, we compiled THERMOSTAT,[1] an easily accessible data hub that contains a large

---

[1] The term Thermostat is inspired by the Greek word for "warm" ("thermos"), hinting at heatmaps being a central application of our contribution. "stat" can mean both (1) "immediately" referring to the immediate access of attribution maps, and (2) "statistics" hinting at cumulative statistics applications.

```
1  import thermostat
2  data = thermostat.load("imdb-bert-lig")
3  example = data[0]
```

> 'attributions' = {list: 512} [-0.18760254979913606, -0.0315956249833107, 0.04854373633861542, 0.00658783596009016, 0.0

'idx' = {int} 0

> 'input_ids' = {list: 512} [101, 2092, 1010, 1045, 7166, 2000, 3422, 3152, 2005, 2028, 1997, 2093, 4436, 1012, 6854, 1010, 2045,

'label' = {int} 1

> 'predictions' = {list: 2} [-3.4371631145477295, 4.042327404022217]

```
4  example.render()
```

[CLS] amazing movie . some of the script writing could have been better (some cliched language) . joyce ' s " the dead " is alluded to throughout the movie . beautiful scenery and great acting . very poetic . highly recommend . [SEP]

Figure 1: Code examples. Top: Loading a dataset and extracting a single instance. Bottom: Visualizing the instance as a heatmap on token level.

quantity of feature attribution maps, from numerous explainers, for multiple models, trained on different tasks. Alongside the dataset, we publish a compatible library with analysis and convenience functions. In this paper, we introduce the data hub, showcase the ease of access and discuss use cases.

2 THERMOSTAT

THERMOSTAT is intended to be a continuous, collaborative project. As new models, datasets and explainers are published, we invite the community to extend THERMOSTAT. In what follows, we describe the current state which is published under https://github.com/DFKI-NLP/thermostat.
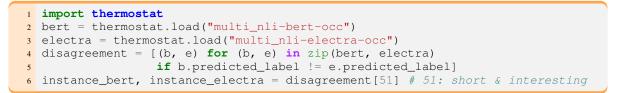
2.1 Demonstration

First, we demonstrate the ease of access. Downloading a dataset requires just two lines of code, as illustrated in the snippet in Fig. 1, in which we download the attribution maps as returned by the (Layer) Integrated Gradients explainer (Sundararajan et al., 2017) for BERT classifications (Devlin et al., 2019) on the IMDb test set (Maas et al., 2011). In addition to the list of feature attributions, the input IDs, the true label of every instance given by the underlying dataset, and the logits of the model's predictions are shipped. Switching the explainer, model or dataset only requires to change the configuration identifier string ("imdb-bert-lig" in Fig. 1). All configuration identifiers in THERMOSTAT consist of three coordinates: dataset, model, and explainer. A visualization tool which returns

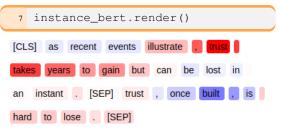heatmaps like the one shown in Fig. 1 is also contained in the accompanying library.

The object that holds the data after download inherits from the Dataset class of the datasets library (Lhoest et al., 2021). This is convenient, because data is cached and versioned automatically in the background and processed efficiently in parallel. Furthermore, datasets provides many convenience functions which can be applied directly, such as filtering and sorting.

Let us see how this helps us to efficiently compare models and explainers in THERMOSTAT. Let us first compare models. In what follows we consider BERT (Devlin et al., 2019) and ELECTRA (Clark et al., 2020), both trained on the MultiNLI (Williams et al., 2018) dataset. We are particularly interested in instances that the two models disagree on. Downloading the explanations and filtering for disagreements is again only a matter of a few lines, as demonstrated in Fig. 2a.
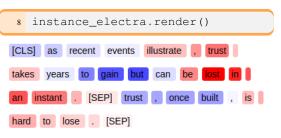
We derive explanations for the output neuron with the maximum activation. In Fig. 2b, we observe that the Occlusion (Zeiler and Fergus, 2014) explainer does not attribute much importance to the phrase "can be lost in an instant". This is plausible since the heatmap explains a misclassification: the maximum output activation stands for entailment, but the correct label is contradiction and the phrase certainly is a signal for contradiction. In contrast, in the case of ELECTRA (Fig. 2c) which correctly classified the instance the signal phrase receives much higher importance scores.

```
1  import thermostat
2  bert = thermostat.load("multi_nli-bert-occ")
3  electra = thermostat.load("multi_nli-electra-occ")
4  disagreement = [(b, e) for (b, e) in zip(bert, electra)
5                  if b.predicted_label != e.predicted_label]
6  instance_bert, instance_electra = disagreement[51]  # 51: short & interesting
```

(a) Code example that loads two THERMOSTAT datasets. We create a list of instances (disagreement) where the two models (BERT and ELECTRA) do not agree with each other regarding the predicted labels. We then select a demonstrative instance from it.

(b) Heatmap visualization of the selected instance from Fig. 2a. BERT predicted "entailment" for this example, while the true label is "contradiction".

(c) Heatmap visualization of the selected instance from Fig. 2a. ELECTRA correctly predicted "contradiction" for this example.

Figure 2: Code examples for comparing models instance-wise.

After we have now demonstrated how to compare models, let us compare explainers across datasets, as done in previous works. Here, we partly replicate the experiments of Neely et al. (2021). The authors compute the rank correlation in terms of Kendall's τ between feature attribution maps. If two explainers mostly agree on the importance rank of input features, the value should be high; low otherwise. The authors find that the Integrated Gradients explainer and the LIME (Ribeiro et al., 2016) explainer have a higher τ value (agree more) for a MultiNLI model (.1794) than when used to rank features for an IMDb-trained classifier (.1050).

Fig. 3 demonstrates how THERMOSTAT allows to conduct such experiments concisely. The output of the experiment in Fig. 3 reproduces the findings of Neely et al. (2021) to a reasonable degree, i.e. the τ value for MultiNLI (.1033) is larger than the τ value for IMDb (.0257).[2]

2.2 Maintenance

However, explainers such as LIME involve several hyperparameters (number of samples, sampling method, similarity kernel, ...) and thus results can deviate for other choices. THERMOSTAT datasets are versioned and for each version a configuration file is checked in that contains the hyperparameter choices.

If new best practices or bugs emerge, e.g. in terms of hyperparameters, an updated dataset can be uploaded in a new version. This increases comparability and replicability.

There is also a seamless integration with Hugging Face's datasets as mentioned above. This is why explanation datasets that are published through datasets can be used in THERMOSTAT directly. When contributing new explanation datasets, users simply add the metadata about the three coordinates and make sure that the fields listed in Fig. 1 are contained. As soon as the dataset is published on the community hub, it can be downloaded and parsed by THERMOSTAT. More details are provided in the repository.[3]

2.3 Current State

After discussing use and maintenance, we will now present the current state of THERMOSTAT. Please recall that THERMOSTAT is intended to be a continuous, collaborative project.

With the publication of this paper, we contribute the explanations listed in Tab. 1. In total, the compute time for generating the explanations already amounts to more than 10,000 GPU hours; computational resources that the community does not have

---

[2]Neely et al. (2021) compare DistilBERT (Sanh et al., 2020) based models, we compare BERT-based models. They further constrain their evaluation to 500 instances while we are calculating the values for the entire datasets.

[3]Users contributing to THERMOSTAT should be aware that the THERMOSTAT project follows the Code of Ethics of ACL and ACM.

```
1  import thermostat
2  from scipy.stats import kendalltau
3  lime, ig = thermostat.load("imdb-bert-lime"), thermostat.load("imdb-bert-lig")
4  lime, ig = lime.attributions.flatten(), ig.attributions.flatten()
5  print(kendalltau(lime, ig))
6  # KendalltauResult(correlation=0.025657302000906455, pvalue=0.0)
```

Figure 3: Code example for investigating the rank correlation between LIME and (Layer) Integrated Gradients explanations on IMDb + BERT. The analogous calculation of Kendall's τ for MultiNLI is left out for brevity. We simply state the value in the last line.

to invest repeatedly now.[4] Please note that the table is organized around the three coordinates: datasets, models, and explainers, the choices of which we discuss in the following.

Datasets Currently, four datasets are included in THERMOSTAT, namely IMDb (sentiment analysis, Maas et al., 2011), MultiNLI (natural language inference, Williams et al., 2018), XNLI (natural language inference, Conneau et al., 2018) and AG News (topic classification, Zhang et al., 2015). We chose these datasets, because arguably, they are prominently used in NLP research.

We hypothesize that instances that the model did not encounter at training time are more informative than known inputs. This is why we concentrated our computational resources on the respective test splits. In total, these amount to almost 50,000 instances already.

Models The second coordinate in THERMOSTAT is the model. Currently, five model architectures are included, namely ALBERT (Lan et al., 2020), BERT (Devlin et al., 2019), ELECTRA (Clark et al., 2020), RoBERTa (Liu et al., 2019), and XL-Net (Yang et al., 2019). We chose community-trained fine-tuned classifiers as they are aptly available through the `transformers` library (Wolf et al., 2020), several of which are provided by TextAttack (Morris et al., 2020). The repository that we publish can be used to quickly include new models, if they are provided through the `transformers` library.

Explainers Provided through the Captum library (Kokhlikyan et al., 2020), there are five prominent feature attribution methods included in THERMOSTAT. (Layer) Gradient x Activation (Shrikumar et al., 2017) is an efficient method without hyperparameters. Integrated Gradients (Sundararajan

et al., 2017), LIME (Ribeiro et al., 2016), Occlusion (Zeiler and Fergus, 2014) and Shapley Value Sampling (Castro et al., 2009) can be considered computationally challenging and involve hyperparameters. The choice of parameters in THERMOSTAT follows best practices and, as mentioned above, is well-documented and can be updated and extended.

3 Related Work

To the best of our knowledge, the closest work to our contribution is the Language Interpretability Tool (LIT) by Tenney et al. (2020) which offers a graphical interface for exploring saliency maps, counterfactual generation and the visualization of attention and embeddings. We also draw connections to Hoover et al. (2020) and Lal et al. (2021) who developed interfaces for analyzing the attention mechanism and embeddings of Transformer architectures. These again are visualization and analysis tools and as such complementary to the collection of explanations that is our primary contribution. Thus, an interface that bridges THERMOSTAT and LIT, for instance, is an interesting future direction.

There exist libraries, such as Captum (Kokhlikyan et al., 2020) or transformers-interpret (Pierse, 2021), that facilitate the generation of neural explanations. These libraries do not, however, free the user of the computational burden, nor are they easily accessible to non-technical researchers.

As noted in Section 2, the `datasets` library (Lhoest et al., 2021) functions as the backbone of THERMOSTAT. The novelties our work brings with it are (1) attributions from a variety of models that took over 10,000 GPU hours to compute in total, and (2) the support for explainability-specific visualizations and statistics like heatmap visualization and rank correlation between multiple explainers.

Finally, tangentially related to our work are

---

[4] To produce the feature attribution maps, we used up to 24 NVIDIA GPUs in parallel, namely GTX 1080Ti, RTX 2080Ti, RTX 3090, Quadro RTX 6000 and RTX A6000.

Dataset	Split (Subset)	# instances	# classes	Explainers				
IMDb	Test	25000	2	**LGxA**	**LIG**	**LIME**	**Occ**	**SVS**
textattack/albert-base-v2-imdb (ALBERT)				✓	✓	✓	✓	✓
textattack/bert-base-uncased-imdb (BERT)				✓	✓	✓	✓	✓
monologg/electra-small-finetuned-imdb (ELECTRA)				✓	✓	✓	✓	✓
textattack/roberta-base-imdb (RoBERTa)				✓	✓	✓	✓	✓
textattack/xlnet-base-cased-imdb (XLNet)				✓	✓	✓	✓	✓
MultiNLI	Validation Matched	9815	3	**LGxA**	**LIG**	**LIME**	**Occ**	**SVS**
prajjwal1/albert-base-v2-mnli (ALBERT)				✓	✓	✓	✓	
textattack/bert-base-uncased-MNLI (BERT)				✓	✓	✓	✓	✓
howey/electra-base-mnli (ELECTRA)				✓	✓	✓	✓	✓
textattack/roberta-base-MNLI (RoBERTa)				✓	✓	✓	✓	✓
textattack/xlnet-base-cased-MNLI (XLNet)				✓	✓	✓	✓	✓
XNLI	Test (en)	5010	3	**LGxA**	**LIG**	**LIME**	**Occ**	**SVS**
prajjwal1/albert-base-v2-mnli (ALBERT)				✓	✓	✓	✓	
textattack/bert-base-uncased-MNLI (BERT)				✓	✓	✓	✓	✓
howey/electra-base-mnli (ELECTRA)				✓	✓	✓	✓	✓
textattack/roberta-base-MNLI (RoBERTa)				✓	✓	✓	✓	
textattack/xlnet-base-cased-MNLI (XLNet)				✓	✓	✓	✓	✓
AG News	Test	7600	4	**LGxA**	**LIG**	**LIME**	**Occ**	**SVS**
textattack/albert-base-v2-ag-news (ALBERT)				✓	✓	✓	✓	✓
textattack/bert-base-uncased-ag-news (BERT)				✓	✓	✓	✓	✓
textattack/roberta-base-ag-news (RoBERTa)				✓	✓	✓	✓	✓

Table 1: Overview of feature attribution maps in THERMOSTAT. Dark green cells (86 out of 90) denote available configurations. Gray cells denote configurations that are work-in-progress.

datasets that supply explanations on top of texts and labels, usually collected from human annotators. e-SNLI (Camburu et al., 2018) probably is the most famous example in this line of work. The reader is referred to Wiegreffe and Marasović (2021) for a concise survey. In contrast to our work, the above mentioned papers present human ground truths instead of machine-generated explanations.

4 Conclusion

We present THERMOSTAT, an easily accessible data hub containing a large collection of NLP model explanations from prominent and mostly expensive explainers. We demonstrate the ease of access, extensibility and maintainability. New datasets can be added easily. Furthermore, we showcase an accompanying library and outline use cases. Users can compare models and explainers across a variety of datasets.

THERMOSTAT democratizes explainability re-search to a certain degree as it mitigates the computational (environmentally and financially) and implementational burden. Machine-generated explanations become accessible to non-technical researchers. Furthermore, comparability and replicability are increased.

It becomes apparent when consulting the literature in Section 1 that interpretation beyond classification (e.g. machine translation) is still an open problem (Wallace et al., 2020). Hence, we focus on these four text classification problems that are well-trodden paths.

References

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Explaining predictions of non-linear classifiers in NLP. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 1–7, Berlin, Germany. Association for Computational Linguistics.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168, Copenhagen, Denmark. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. In *Transactions of the Association for Computational Linguistics*, volume 7, pages 49–72.

Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. 2021. Benchmarking and survey of explanation methods for black box models. *arXiv preprint arXiv:2102.13076*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the shapley value based on sampling. *Comput. Oper. Res.*, 36(5):1726–1730.

Hanjie Chen, Song Feng, Jatin Ganhotra, Hui Wan, Chulaka Gunasekara, Sachindra Joshi, and Yangfeng Ji. 2021. Explaining neural network predictions on sentence pairs via learning word-group masks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3917–3930, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. 2020. How do decisions emerge across layers in neural models? interpretation with differentiable masking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3243–3255, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shuoyang Ding and Philipp Koehn. 2021. Evaluating saliency methods for neural language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5034–5052, Online. Association for Computational Linguistics.

Gabriel Erion, Joseph Janizek, Pascal Sturmfels, Scott Lundberg, and Su-In Lee. 2021. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, pages 1–12.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online. Association for Computational Linguistics.

Peter Hase, Harry Xie, and Mohit Bansal. 2021. Search methods for sufficient, socially-aligned feature importance explanations with in-distribution counterfactuals. *arXiv preprint arXiv:2106.00786*.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics:*

System Demonstrations, pages 187–196, Online. Association for Computational Linguistics.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Narine Kokhlikyan, Vivek Miglani, Bilal Alsallakh, Miguel Martin, and Orion Reblitz-Richardson. 2021. Investigating sanity checks for saliency maps with image and text classification. *arXiv preprint arXiv:2106.07475*.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Vasudev Lal, Arden Ma, Estelle Aflalo, Phillip Howard, Ana Simoes, Daniel Korat, Oren Pereg, Gadi Singer, and Moshe Wasserblat. 2021. InterpreT: An interactive visualization tool for interpreting transformers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 135–142, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-grounded evaluations of explanation methods for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5195–5205, Hong Kong, China. Association for Computational Linguistics.

Quentin Lhoest, Patrick von Platen, Thomas Wolf, Albert Villanova del Moral, Yacine Jernite, Suraj Patil, Mariama Drame, Julien Chaumond, Julien Plu, Lewis Tunstall, Joe Davison, Simon Brandeis, Teven Le Scao, Victor Sanh, Kevin Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, and François Lagunas. 2021. huggingface/datasets: 1.8.0.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Frederick Liu and Besim Avci. 2019. Incorporating priors with feature attribution on text classification.

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6274–6283, Florence, Italy. 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. *arXiv preprint arXiv:1907.11692*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Michael Neely, Stefan F. Schouten, Maurits J. R. Bleeker, and Ana Lucic. 2021. Order in the court: Explainable ai methods prone to disagreement. *arXiv preprint arXiv:2105.03287*.

Anphi Nguyen and María Rodríguez Martínez. 2020. On quantitative aspects of model interpretability. *arXiv preprint arXiv:2007.07584*.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

Charles Pierse. 2021. transformers-interpret v0.5.1. GitHub. Accessed: 2021-06-29.

Nina Poerner, Hinrich Schütze, and Benjamin Roth. 2018. Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 340–350, Melbourne, Australia. Association for Computational Linguistics.

Grusha Prasad, Yixin Nie, Mohit Bansal, Robin Jia, Douwe Kiela, and Adina Williams. 2020. To what extent do human explanations of model behavior align with actual model behavior? *arXiv preprint arXiv:2012.13354*.

Danish Pruthi, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C. Lipton, Graham Neubig, and William W. Cohen. 2020. Evaluating explanations: How much do explanations from the teacher aid students? *arXiv preprint arXiv:2012.00893*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2662–2670.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Robert Schwarzenberg, Nils Feldhus, and Sebastian Möller. 2021. Efficient explanations from empirical explainers. *arXiv preprint arXiv:2103.15429*.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.

Sanchit Sinha, Hanjie Chen, Arshdeep Sekhon, Yangfeng Ji, and Yanjun Qi. 2021. Perturbing inputs for fragile interpretations in deep natural language processing. *arXiv preprint arXiv:2108.04990*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, Online. Association for Computational Linguistics.

Michael Tsang, Sirisha Rambhatla, and Yan Liu. 2020. How does this interaction affect me? interpretable attribution for feature interactions. In *Advances in Neural Information Processing Systems*, volume 33, pages 6147–6159. Curran Associates, Inc.

Eric Wallace, Matt Gardner, and Sameer Singh. 2020. Interpreting predictions of NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 20–23, Online. Association for Computational Linguistics.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable nlp. *arXiv preprint arXiv:2102.12060*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang. 2021. On the faithfulness measurements for model interpretations. *arXiv preprint arXiv:2104.08782*.

Muhammad Bilal Zafar, Michele Donini, Dylan Slack, Cedric Archambeau, Sanjiv Das, and Krishnaram Kenthapadi. 2021. On the lack of robust interpretability of neural text classifiers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3730–3740, Online. Association for Computational Linguistics.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833, Cham. Springer International Publishing.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.