P3: Evaluating the efficacy of explanation methods in uncovering spurious correlations

ANIRUDH NAKRA and HENOK GEBEYEHU

1 INTRODUCTION

Modern day machine learning tools often exploit non-interpretable features from the train dataset. In addition to achieving state-of-the-art accuracy on benchmarks, there is also an underlying need to understand which concepts are being learned by the designed model. A natural consequence to designing without such robustness in mind is the fact that a model might learn from unwanted micro-signals that might be unexplainable through a human lens. In the context of Natural Language based tasks, spurious correlations often come from unexpected sources. Adversaries that flip decisions can be generated at a character level without changing the semantic content. [Ebrahimi et al. 2018a],[Ebrahimi et al. 2018b] Spurious correlations can also arise due to biased datasets. Models trained on such corpora are prone to learning unwanted correlations even though the model itself might be robust in theory.[Zhang et al. 2018], [Bolukbasi et al. 2016]. A great practical example of a spurious correlation arisen through biases in datasets is given by [Wang and Culotta 2020a] who find that words like *Spielberg* and *New York Subway* have high co-occurences with the sentiment of movie reviews.

We adopt the generic framework suggested in [Wang et al. 2021] for identifying and mitigating spurious correlations.

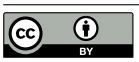
(1) Important Token Identification: Using the underlying basic causal mapping between the input tokens and the output class in a text classification problem, we evaluate different explainability techniques (such as LIME, IG, etc.) and regimes (such as local vs global explanations, post-hoc vs learned explanations) to identify how effective they are in recovering spurious correlations.
(2) Identifying spurious correlations: We adopt the approach outlined by [Wang et al. 2021] and use a cross-dataset consistency based approach with the goal being to identify globally meaningful tokens (out of domain importance) and locally meaningful tokens (in domain importance).

This is of particular importance to NLP tasks where we can have a lot of unintended bias built into the dataset due to broad web crawling based collection procedures. We are explaining the model's underlying working by uncovering data that might be noise such that researchers and broadly humans think these data samples are not useful for the task its being used for. In this study, we gear the explanations towards Machine Learning researchers. We provide all details at [1]

2 BACKGROUND AND RELATED WORK

Several approaches to tackling such problems have been explored in the literature with solutions such as dataset augmentation [Wu et al. 2022] and adding adversarial examples to training corpus [Jia and Liang 2017] but these efforts are focused on modifying the dataset. These techniques can be great when taken into consideration for generating new datasets, but are infeasible for already existing datasets considering the vast quantity of well-established datasests that are currently

[1]Code, appendix and analysis details available at the following link



available. Other works that explored spurious correlations for NLP focused on training both a biased model and a debiased model that could confirm if particular input tokens were spurious in nature [He et al. 2019] [Clark et al. 2019], but the limitation of these approaches are that they required spurious tokens to be pre-defined and could not be used to identify new spurious correlations. Currently, the best method for automatically identifying spurious correlations in NLP was proposed by [Wang et al. 2021]. They first identify which tokens in a sentence are considered important for a machine learning model to accomplish a certain task using attention scores. Then they perform a cross-dataset analysis to categorize each important token as either "genuine" (tokens necessary for the model to make an accurate prediction) or "spurious" (tokens that are useful in the training data, but are unreliable in general). The intuition behind the cross-dataset analysis is that if an input token has inconsistent attention scores across both datasets, then it is likely spurious, but if it has consistent attention scores in both datasets, then it is likely to be genuine.

The current limitation of this approach is that the identification of important tokens is solely done using transformer attention scores. It is still unclear whether attention can be considered explanation as summarized by [Bibal et al. 2022], and one of the major points from [Jain and Wallace 2019] was that "there are poor correlations between attention weights and gradient-based or leave-one-out methods for explanation." This implies that different input saliency techniques will select different input tokens as explanation. So by only utilizing a single input saliency method in the first stage of their framework, it is possible that some spurious correlations remain unidentified. This study will thus be of great importance to ML researchers since they can identify which features of their dataset are causing their models to be less generalizable through a comparison of different techniques. If spurious correlations are identified and their effects are mitigated, then the model will also be robust to domain shifts which has the potential to significantly improve perceived trust in algorithms for decision-making.

Dataset	Data Source	Train-Val-Test Split
Stanford Sentiment Treebank - 2 [Socher et al. 2013]	IMDb	96% : 1% : 3%
Yelp Polarity Dataset [Zhang et al. 2015]	Yelp	94% : N/A : 6%
Movie Rationales Dataset [Zaidan et al. 2007]	IMDb	80% : 10% : 10%

Table 1. Datasets used and their characteristics

3 APPROACH

In our work, we adapt the framework developed by [Wang et al. 2021] to identify spurious correlations. We modify the important token extraction process of their framework by replacing attention scores with various input saliency techniques. The goal is to identify whether using different saliency techniques in the important token extraction stage can uncover new spurious correlations. There is also an underlying concern that the cross-dataset analysis adopted by [Wang et al. 2021] may identify domain-specific tokens as spurious. We expand on the cross-dataset analysis performed by [Wang et al. 2021] by doing analysis on datasets from the same domain like SST-2 and Movie Rationales and by doing analysis of datasets from different domains like SST-2 and Yelp. Using this in-domain, out-of-domain setup, we try to see if context-specific tokens can be properly identified as genuine tokens. In a nutshell, we formalize our objectives by posing the following research questions:

RQ1: Are current feature importance and attribution methods generally effective in identifying spurious tokens?

RQ1b: Furthermore, how does the effectiveness of discovering spurious features vary across different paradigms such as gradient vs non gradient based methods, post-hoc vs learned methods and most importantly local vs global methods.

RQ2: Can cross dataset consistency reveal different spurious correlations depending on dataset context?

RQ3: Does human annotated evidence agree with uncovered genuine tokens?

Our methodology for finding candidate spurious tokens is broadly categorized into three steps. The first step involves calculating for the different saliency methods for all samples in our dataset. We explain the datasets of our choice, the XAI techniques we use and how they are used to create token level saliency scores in our experimental setup section. Once we have the scores for all of these techniques across all samples, in the second step we create a dictionary that forms the backbone for all of our analysis. This dictionary contains information about every unique token encountered, the respective frequency of that token and the values of the saliency scores of that token across all samples in our dataset. This dictionary is then used to normalize the saliency scores for the tokens based on how frequently they appear across the dataset and the saliency scores themselves to get meaningful scores that can be compared. In the third and last step, we perform our cross-dataset analysis (for in and out of domain setups) in which we compare scores for each unique token across two different datasets, and if the difference in their normalized saliency scores is large enough, we mark them as candidate spurious tokens. Attention weights have been shown to have poor correlation with other explanation techniques and different explanation techniques have been found to be intrinsically inconsistent. Our approach does not make the assumption that attention scores will identify all important tokens to a model's prediction but instead we motivate an ensemble based approach where we can learn from tokens and patterns found across different techniques rather than relying on Attention itself.

4 EXPERIMENTAL SETUP

Based on the taxonomy given by[Doshi-Velez and Kim 2017], we adopt the use of a functionally grounded approach by doing a purely computational study involving no humans in the loop. We believe this to be an adequate first step to guide further work in the area. In particular, we evaluate a proxy task which has been shown to be human grounded through the user studies done in [Wang et al. 2021]. For the purpose of the study, we assume that the observations of [Wang et al. 2021] have significant merit and have some small ϵ error/misalignment with respect to the human grounded evaluation done in the paper. So, any further evaluation would then consider an offset of ϵ and the evaluation will be done purely w.r.t the references in [Wang et al. 2021] like a surrogate setup.

Our prototype application focuses on sentiment analysis which is a text classification problem. For solving the problem, we consider using the exact same setup as suggested by [Wang et al. 2021]. Particularly we use the SST-2 dataset [Socher et al. 2013] involving various movie reviews, the Yelp-polarity dataset [Zhang et al. 2015] consisting of food reviews for OOD analysis and the Movie Rationales dataset [Zaidan et al. 2007] consisting of movie reviews for in domain analysis. For the purposes of identifying spurious correlations, we use a compressed BERT model called DistilBERT which has been fine-tuned on SST-2 and provided at [HF Canonical Model Maintainers 2022].

To collect important scores using: LIG, LxA, and LIME, we utilized the Captum library. We define a (callable) forward function which returns softmaxed model outputs that are needed for library function calls. Whereas LxA and LIG only require evaluating the gradients of the model, LIME requires fitting an additional local & interpretable model. The interpretable model that we used was a linear model trained with L1 regularization from sci-kit learn library. We considered a 1000 sample neighborhood for fitting the local classifier after analyzing faithfulness compute tradeoffs.

To extract the attention scores, we take the transformer block right before the classification head as the self-attention weights in this layer will likely contain the most information regarding the sentiment of each input token. Since the transformer utilizes multi-head attention, we take the average of the attention matrices from the 12 attention heads. This gives us a matrix that contains

the self-attention weights for the input sequence. We then take the average of each column to get token level saliency. Our attention scores differ slightly from [Wang et al. 2021]'s since they only take the row in self-attention that corresponds to the CLS token.

We train our Logistic Regression model using the sci-kit learn library. The input to the model is a one-hot encoding consisting of the presence/absence of a token. Every word in the dataset's vocabulary is included except for words that either occur less than 5 times in the entire dataset or are present in more than 80% of the sentences to remove nonsensical weights or stop words respectively.

In order to calculate alignment with respect to the reference tokens in [Wang et al. 2021], we adopt two metrics each of which gives us some information about the level of overlap between tokens that we find through our methods and tokens that form the reference for our research questions. The metrics are as follows:

(1) Intersection-Over-Union (IoU): IoU is defined as

$$IoU = \mathbb{1}(A \cap B)/\mathbb{1}(Unique(A \cup B))$$

(2) Recovery Percentage (RP): RP is defined as

$$RP = 100 * \mathbb{1}(A \cap B)/\mathbb{1}(B)$$

where $\mathbb{1}$ represents the count function, $Unique$ represents the number of unique tokens in the set, A represents the set uncovered by the technique being evaluated and B represents our reference tokens. It is trivial that the union in IoU can explode with an increase in size of B. This can lead to difficulty in analysis. To this extent, we also use a second metric called Recovery Percentage (RP).

For our experimental setup, we design three different computational goals for each research question. Firstly to answer RQ1, we choose to use an Averaged Attention score (Learned, Local technique) (different from [Wang et al. 2021]'s setup), a Logistic Regression based score (Learned, Global technique) [Wang and Culotta 2020b], LIME (Post-hoc, Local technique) [Ribeiro et al. 2016], Layer Integrated Gradient (Post-hoc, Local technique) [Sundararajan et al. 2017] and Layer Gradient x Activation (Post-hoc, Local technique). These techniques span different categories which motivate the analysis of RQ1b where we average alignments across different methods that fall under these categories.

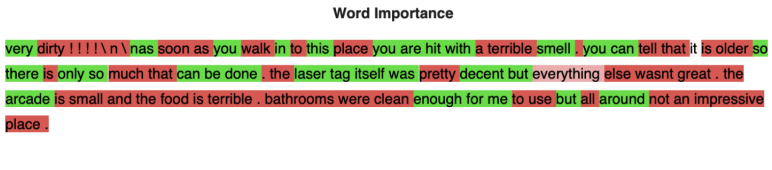


Fig. 1. Sample highlight on Yelp using LR, red means negative sentiment, green means positive

RQ2 is motivated by the need to identify domain agnostic and domain specific tokens that might be spurious in nature. To solve this problem, we solve for the alignment score between two setups. The first setup involves recovering spurious correlations by cross-dataset analysis of the SST-2 and Yelp dataset for recovering globally meaningful tokens. The second setup involves recovering spurious correlations by cross-dataset analysis of the SST-2 and MR dataset for recovering locally meaningful tokens. Analyzing their quantitative alignment using IoU and RP lets us know how much the setups agree on average across different techniques while doing a qualitative analysis on the tokens recovered lets us know variations in tokens we are recovering at a linguistic level.

Solving RQ3 is simpler than the previous RQs. The MR dataset already contains human annotated evidence for tokens that users have found relevant for the sentiment analysis task. Solving for

agreement between human annotated evidence and genuine tokens involves computing alignment between them using the chosen metrics. We formulate this as an inverse problem in which we calculate the alignment between the evidence and the spurious tokens and expect this alignment to be small since spurious tokens are meant to not be meaningful to a human annotator for the particular task.

5 FINDINGS

All the methods for extracting the candidate spurious tokens are significantly discriminative. Out of the approximately 31,000 unique words across the corpora, the different saliency methods are able to extract around 2000-4000 tokens. This is an order of magnitude smaller and reaffirms that such methods are indeed well suited for looking at spurious relationships of tokens with labels. We verify this by looking at the frequency of the spurious tokens along with the label skew. (listed in the appendix)

Problem Type	Metric	Attention	LR	LIME	LIG	LxA
RQ1 : Alignment with [Wang et al. 2021]	**IoU** (↑)	0.056	0.084	0.018	0.063	0.057
	RP (↑)	60.4%	81.06%	25.91%	76.74%	70.76%
RQ2 : Alignment of **SST+Yelp** and **SST+MR**	**IoU** (↑)	0.090	0.023	0.085	0.033	0.040
	RP (↑)	96.19%	70.37%	94.80%	58.12%	70.37%
RQ3 : Alignment with [Zaidan et al. 2007]	**IoU** (↓)	0.00062	0.00029	0.00063	0.00033	0.00119
	RP (↓)	1.87%	2.02%	1.58%	0.954%	3.31%

Table 2. Alignment scores for RQ1, RQ2 and RQ3

Problem Type	Metric	Local	Global	Gradient	Non-Gradient	Post-hoc	Learned
RQ1b: Alignment across paradigms	**IoU**	0.049	0.084	0.060	0.053	0.046	0.071
	RP	58.47%	81.06%	73.75%	55.81%	57.80%	70.76%

Table 3. Alignment scores for RQ1b

Looking at the five different saliency techniques in **Table 2**, we find that Attention. LR, LxA and LIG exhibit significant levels of alignment with the reference tokens with them recovering **60.4%, 81.06%, 76.74% and 70.76%** of the tokens enumerated in [Wang et al. 2021]. This is impressive since the reference tokes we consider are merely samples provided by the authors and thereby we get worst-case scores. We also notice that such trends remain consistent across both metrics with LR being better than all the other techniques, LxA, LIG and Attention being worse than LR and LIME being the worst of all which is expected due to its model agnostic nature. Interestingly, we also note that LxA and LIG outperform Attention based saliency despite them being employed post-hoc. We believe that this finding agrees with other studies [Jain and Wallace 2019] that find low alignment for attention with gradient based methods. In addition, we also note that Logistic Regression based saliency being the best performing score is unsurprising due to the global-learned nature of the method.

Overall, we find that when comparing across the six different paradigms of Local vs Global methods, Gradient based vs Non Gradient based methods & Post-hoc vs Learned methods, we reaffirm our intuition. We observe that LR (our Global technique) outperforms the rest which are Local **(IoU:0.084 vs 0.049, RP:81.06% vs 58.47%)** motivating future XAI methods to come up with global explanations. Similarly, we also observe that Learned techniques which are a result of the training process of the model itself like Attention or surrogate models like LR outperform Post-hoc techniques like LIG,LxA and LIME **(IoU: 0.071 vs 0.046 , RP: 70.76% vs 57.80%)**. We note that Gradient based saliency methods outperform non-Gradient methods on average (outperform Attention and LIME but not LR) **(IoU: 0.060 vs 0.053, RP: 73.75% 55.81%)**. For RQ1, we hence find that the general trend in efficacy is **LR>LIG>LxA>Attention>LIME** which follows from

theory. For RQ1b, we observe that the general trend in effectiveness across paradigms is that Learned methods outperform Post-hoc techniques, Gradient based saliency is generally better than Non-gradient approaches and that our Global method outperforms all other local methods.

Dataset	Technique	Count	Examples of Spurious Tokens
SST-2 + Yelp	**Attention**	3,078	*river, asian, american, college, office, recycled, genre, romance, stock, chicken*
	LR	2,835	*poet, exercise, martial, oscar, schumacher, korean, palestinian, rodriguez*
	LIME	4,050	*nature, karen, jews, orthodox, heroes, european, lewis, father, davies, seven*
	LIG	3,563	*parental, director, cinema, eastwood, pulp, mystery, college, sequel, auschwitz*
	LxA	3,627	*parental, conspiracy, directors, showcase, storyline, dialogue, rock & roll, noir*
SST-2 + MR	**Attention**	3,256	*movie, lee, television, century, production, animation, hugh, european, jim, adam*
	LR	2,751	*production, industry, theatrical, category, john, director, michael, performance*
	LIME	3,854	*william, action, lee, starring, music, movies, production, cartoon, industry, direct*
	LIG	2,690	*flash, animation, cartoon, hugh grant, hannibal, performance, michael, classic*
	LxA	2,482	*movie, films, production, drama, hannibal, satire, cinematic, screenplay, race*

Table 4. Cross dataset analysis details : Datasets and Techniques used, Count of spurious tokens uncovered and examples of spurious tokens uncovered

For evaluating RQ2, we calculate alignment scores between the spurious tokens recovered from SST-2 through out-of-domain dataset analysis and in-domain dataset analysis. A high alignment score means that the technique recovers similar tokens when the analysis is done in and out of domain. This would imply that the techniques might recover the same kind of spurious tokens regardless of dataset contexts which can be helpful when evaluating dataset-agnostic biases but might miss crucial variations in biases that we observe while changing contexts. On the other hand, low alignment will mean that the kind of spurious tokens we observe have too much variability to the extent that we uncover completely new kinds of spurious relations. Ideally, we would like to have a balance between such extreme situations. Experimentally, we notice that Attention **(IoU: 0.090 , RP:96.19%)** and LIME **(IoU:0.085 , RP:94.30%)** exhibit a high alignment between the spurious tokens uncovered in and out of domain. We postulate that this is because the DistilBERT model is not changed when computing scores in and out of the domain which means that the attention scores largely remain the same while the model agnostic nature of LIME might explain the other part. These techniques can thereby be insightful for broader patterns of biases but might not uncover new variations that might be helpful for domain experts. On the other hand, we see that techniques like LIG, LxA and LR have RP's between **50-70%** which demonstrates that they are possibly uncovering newer tokens that are spurious given differing contexts which is very interesting. Qualitatively, we find that spurious tokens that are unique to the SST-MR setup across LIG,LxA and LR are words like *" Jennifer", "Richard", "Barry", "Diane"* that refer to famous celebrities and T.V shows while spurious tokens across these techniques which are unique to the SST-Yelp setup are words like *"Washington", "mergers", "Diplomacy"* which are relatively domain agnostic. On the other hand, tokens uncovered by Attention and LIME are similar across both setups with *"heat", "animated" and "music"* being some sample tokens in the intersection. More examples of uncovered tokens can be found in the appendix. For RQ2, we thereby propose that it might be better to adopt techniques like LR, LIG and LxA that are versatile enough to uncover new spurious correlations given changing context. Note that these are also the set of techniques that perform the best for RQ1.

For RQ3, we find that all the investigated techniques have very low IoU and RP alignment scores with human evidence for the task **(IoU: 0.0002-0.0011 , RP: 0.9-3.3%)**. This is great news since this means that the set of tokens that are being identified as spurious by our chosen techniques are not considered valuable by the human annotators who form our reference. This further strengthens

our overall approach to uncovering biases in the chosen task. Based on the abovementioned experimental results, we claim that all three of our RQs are sufficiently justified.

6 LIMITATIONS

Even though we identify several key trends that can help ML researchers to study spurious correlations more effectively, there are still key limitations that remain in our work. Firstly, the automatic identification of spurious correlations for NLP tasks is still an emerging field of research, so currently, there does not exist many techniques for completing this task leading us to adopt the approach proposed by [Wang et al. 2021] but this does not guarantee the exact identification of spurious tokens, rather just candidates for them. We approach this problem both quantitatively and qualitatively but guarantees on properties of the candidate spurious tokens do not exist. We also do not implement the 3rd stage proposed by [Wang et al. 2021] on knowledge-aware perturbation which does synonym based analysis to refine candidate spurious tokens. Additionally, we also consider three standard datasets and limitations exist intrinsically due to the data we are studying. Our recovered relations might not be broadly representative of all spurious tokens that can be identified but we also do not claim to do so. In this sense, our refined context for the sake of the study might be an inherent limitation when broadening the scope of the work. We also note that issues are present in human annotations for the Movie Rationales dataset. Particularly, we note that such annotations are very unselective and large portions of the texts are highlighted by human users. Even though we uncover extremely low alignment between candidate spurious tokens and human annotations, this is still an essential limitation of our dataset choice. Furthermore, as inspired by Dr Ana Marasovic, the particular sentiment analysis task that is being analysed might not contain high amounts of risk making our evaluation restricted in application and encouraging more complicated tasks.

7 CONCLUSION

In this work, we thoroughly compare different XAI techniques and their roles in identifying possible spurious tokens. We find that spurious tokens identified using Logistic Regression had the highest alignment with the reference provided in [Wang et al. 2021], closely followed by gradient based techniques such as LxA and LIG which are in turn better than Attention and LIME based saliency. Furthermore, we intuitively observe that Learned techniques are better than Post-hoc analysis, Gradient based techniques outperform those not utilising gradient information and that Global explanations are preferable to local ones for our task. We also find that different XAI techniques uncover different variations in candidate spurious tokens. Tokens revealed by Attention and LIME vary less across contexts making them more robust to domain shift while techniques like LIG, LxA and LR uncover varying spurious tokens that allow the study of context specific data biases. In addition, we observe that human-annotations have extremely low alignment with uncovered spurious tokens validating our hypothesis that the tokens we find are indeed spurious in nature.

Although our evaluation is very thorough, it is not exhaustive. Other techniques like Universal Adversarial Triggers have shown promise in uncovering biases in NLI tasks which can be explored. Our study has been developed as a functionally grounded evaluation but can be expanded into a bigger application grounded study. Our comparison motivates standardised ways of evaluating dataset biases which motivates the need for new datasets specifically focused on such problems. An exciting direction of future research can be to analyse varying levels of granularity when considering domain shifts by trying to uncover unqiue biases across datasets that provide extremely different contexts. Future work can also try to incorporate directions we forsook while focusing on our pilot study such as the incorporation of knowledge aware perturbations and better proxies to user studies if not a user study itself.

REFERENCES

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is Attention Explanation? An Introduction to the Debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 3889–3900. https://doi.org/10.18653/v1/2022.acl-long.269

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. https://doi.org/10.48550/ARXIV.1607.06520

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. 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)*. Association for Computational Linguistics, Hong Kong, China, 4069–4082. https://doi.org/10.18653/v1/D19-1418

Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [stat.ML]

J. Ebrahimi, Daniel Lowd, and Dejing Dou. 2018a. On Adversarial Examples for Character-Level Neural Machine Translation. In *International Conference on Computational Linguistics*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018b. HotFlip: White-Box Adversarial Examples for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 31–36. https://doi.org/10.18653/v1/P18-2006

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn Dataset Bias in Natural Language Inference by Fitting the Residual. arXiv:1908.10763 [cs.CL]

HF Canonical Model Maintainers. 2022. distilbert-base-uncased-finetuned-sst-2-english (Revision bfdd146). https://doi.org/10.57967/hf/0181

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. arXiv:1902.10186 [cs.CL]

Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2021–2031. https://doi.org/10.18653/v1/D17-1215

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, San Francisco, CA, USA, August 13-17, 2016*. 1135–1144.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 1631–1642. https://aclanthology.org/D13-1170

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. arXiv:1703.01365 [cs.LG]

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2021. Identifying and Mitigating Spurious Correlations for Improving Robustness in NLP Models. https://doi.org/10.48550/ARXIV.2110.07736

Zhao Wang and Aron Culotta. 2020a. Identifying Spurious Correlations for Robust Text Classification. (2020). https://doi.org/10.48550/ARXIV.2010.02458

Zhao Wang and Aron Culotta. 2020b. Identifying Spurious Correlations for Robust Text Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3431–3440. https://doi.org/10.18653/v1/2020.findings-emnlp.308

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating Data to Mitigate Spurious Correlations in Natural Language Inference Datasets. https://doi.org/10.48550/ARXIV.2203.12942

Omar Zaidan, Jason Eisner, and Christine D. Piatko. 2007. Using "Annotator Rationales" to Improve Machine Learning for Text Categorization. In *North American Chapter of the Association for Computational Linguistics*.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. https://doi.org/10.48550/ARXIV.1801.07593

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf