LEXPLAIN: Improving Model Explanations via Lexicon Supervision

Orevaoghene Ahia[♢] **Hila Gonen**[♢] **Vidhisha Balachandran**[♠]
Yulia Tsvetkov[♢] **Noah A. Smith**[♢♡]
[♢]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♠]Language Technologies Institute, Carnegie Mellon University
[♡]Allen Institute for Artificial Intelligence
oahia@cs.washington.edu

Abstract

Model explanations that shed light on the model's predictions are becoming a desired additional output of NLP models, alongside their predictions. Challenges in creating these explanations include making them trustworthy and faithful to the model's predictions. In this work, we propose a novel framework for guiding model explanations by supervising them explicitly. To this end, our method, LEXPLAIN, uses task-related lexicons to directly supervise model explanations. This approach consistently improves the plausibility of model's explanations without sacrificing performance on the task, as we demonstrate on sentiment analysis and toxicity detection. Our analyses show that our method also demotes spurious correlations (i.e., with respect to African American English dialect) on toxicity detection, improving fairness.

1 Introduction

Extensive recent work has sought to advance NLP models so that they offer explanations for their predictions (Rajani et al., 2019; Lundberg and Lee, 2017; Camburu et al., 2018). Here we focus on methods that extract features from the input text to explain a classifier's prediction, known variously as "feature attribution" or "rationales" (Lundberg and Lee, 2017; Li et al., 2016).

Beyond high accuracy on unseen data, classifiers that offer explanations are expected to provide explanations that are faithful to the workings of the model and also intuitive to human users, goals that might be contradicting. We begin with an approach designed for faithfulness (SELFEXPLAIN, §2 and Rajagopal et al., 2021a) and introduce supervision that guides its explanations toward lexical clues already established to be associated with the classification task. Ancillary goals are to *improve* model accuracy through the construction of explanations, and to *remove reliance on spurious features* that can bias a classifier's output in unwanted ways.

Our method, LEXPLAIN (§3), encourages the model to be "confused" in the absence of words from a task-specific lexicon, i.e., to assign a uniform probability distribution across labels, and promotes model explanations that contain task-specific lexemes. We apply LEXPLAIN to sentiment analysis and toxicity detection tasks, and our controlled experiments (§5, §6) comparing LEXPLAIN to SELFEXPLAIN (which does not use supervision for explanations) show that:

(a) LEXPLAIN does not show an accuracy drop relative to the baseline. (b) LEXPLAIN not only promotes lexicon entries as explanations, but also generalizes to additional terms that are related to them but excluded from the lexicon. (c) LEXPLAIN's explanations are usually more sufficient than the baseline's explanations (i.e., the model makes the same prediction on the explanation as on the full input). (d) In toxicity detection, spurious correlations between the toxicity label and African American English (Sap et al., 2019) are reduced in the predictions of LEXPLAIN, relative to the baseline. We view this result as a positive side effect of guiding the model to use task-relevant lexemes. (e) Most importantly, LEXPLAIN's explanations are preferred by human judges 3–4× more often than the baseline's explanations.

We believe these results are encouraging, as they suggest that type-level (lexicon) supervision is a viable alternative to methods that require costly annotation of explanations (Zaidan and Eisner, 2008; Huang et al., 2021).[1]

2 Background: SELFEXPLAIN

Our goal is to improve model explanations in supervised text classification tasks. By supervising explanations, we incorporate inductive biases into models, making them robust to spurious artifacts. Our base model is SELFEXPLAIN (Rajagopal et al.,

[1]Code available at https://github.com/orevaahia/supex

*Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 207–216

2021a), a framework that explains a text classifier's predictions with phrase attribution. We describe SELFEXPLAIN (omitting the global interpretable layer, as we focus on local explanations) and in Section 3 present our proposed method, LEXPLAIN.

Starting with a neural classifier, let $\mathbf{u}_s$ be the masked LM's (Yang et al., 2019) final layer representation of the "[CLS]" token for one instance. $\mathbf{u}_s$ is passed through ReLU, affine, and softmax layers to yield a probability distribution over outputs; the loss is the negative log probability, summed over training instances i:

$$\boldsymbol{\ell} = \text{softmax}(\text{affine}(\text{ReLU}(\mathbf{u}_s))) \quad (1)$$

$$L_{task} = -\sum_i \log \boldsymbol{\ell}[y_i^*] \quad (2)$$

y_i^* is the correct label for instance i. Parameters of the affine layer are suppressed here for simplicity.

A set of phrases is extracted from the data with a phrase-structure parser (Kitaev and Klein, 2018). Let $\mathbf{u}_j$ be the average of the MLM representations of tokens in phrase j. The output distribution *without* phrase j is modeled by transforming the difference (Shrikumar et al., 2017; Montavon et al., 2017) between $\mathbf{u}_s$ and $\mathbf{u}_j$.

$$\mathbf{s}_j = \text{softmax}(\text{affine}(\text{ReLU}(\mathbf{u}_s) - \text{ReLU}(\mathbf{u}_j))) \quad (3)$$

Vector $\mathbf{s}_j$ is a probability distribution over labels, with phrase j absent: the closer $\mathbf{s}_j$ is to $\boldsymbol{\ell}$ (Eq. 1), the less important phrase j is. A secondary log loss L_{LIL} is formed from the probability assigned to the correct label without phrase j, taking a learned weighted sum over all of instance i's phrases, and interpolating with the original log loss (Eq. 2) with a hyperparameter α_1 to weight the secondary loss:

$$\text{loss} = L_{task} + \alpha_1 L_{LIL} \quad (4)$$

The relevance of each phrase j can be defined as the change in probability of the correct label when j is included vs. excluded:

$$r_j = [\boldsymbol{\ell}]_{y_i^*} - [\mathbf{s}_j]_{y_i^*} \quad (5)$$

where higher r_j signify more relevant phrases to the prediction, and as such serve as better explanations.

3 Supervising Explanations

On inspecting explanations retrieved from SELFEXPLAIN, in many cases they do not align intuitively with the predictions. Table 1 illustrates the problem: the explanation of SELFEXPLAIN sentence (1) is the phrase ***on this planet*** which is not a good explanation for the predicted toxic label, unlike ***the biggest idiot***, which can better explain the model's prediction, having the toxic word ***idiot***.

Our modeling innovation is to supervise the explanations encoded in the LIL, rather than letting them emerge from the secondary loss function (L_{LIL} in Equation 4). We incorporate a task lexicon as a source of supervision during training via a third loss component to encourage the model to prefer phrases that contain words in our lexicon as explanations. Table 1 lists examples in the datasets, showing the advantage of our method with more intuitive explanations that better reflect the predicted label.

Our proposed method, named LEXPLAIN, assumes that good explanations within the input are crucial for predictions, thus we encourage the model to be "confused" in the *absence* of lexicon entries, which we expect to be good explanations.

Formally, we minimize the KL divergence between the predicted label distribution $\mathbf{s}_j$, which stands for the distribution in the absence of phrase j (as described in Section 2) and the uniform distribution $\mathbf{s}_{unif}$, for every phrase j:

$$L_{\text{LEXPLAIN}} = D_{KL}(\mathbf{s}_j, \mathbf{s}_{unif}) \quad (6)$$

This objective is used for only lexicon phrases. LEXPLAIN interpolates the third loss, weighted by hyperparameter α_2, with the other two:

$$\text{loss} = L_{task} + \alpha_1 L_{LIL} + \alpha_2 L_{\text{LEXPLAIN}} \quad (7)$$

4 Experimental Setup

Datasets We experiment on three datasets and evaluate explanations based on alignment with model predictions and plausibility with humans. We focus on sentiment analysis and toxicity detection, as judging explanations is easy, intuitive and high-quality lexicons are available. Toxicity detection also allows us to analyze the efficacy of our method in demoting spurious racial correlations, as detailed in §6.

For sentiment analysis, we use the SST-2 dataset (Socher et al., 2013), where the task is to predict the sentiment of movie reviews. For toxicity detection we use **DWMW17** (Davidson et al., 2017) and **FDCL18** (Founta et al., 2018); both Twitter datasets annotated for toxicity and dialect: African

Input	SELFEXPLAIN	LEXPLAIN
she is the biggest idiot on the planet.	on this planet	the biggest idiot
Haha , says the little bitch who let someone take his phone . a real man would n't have let that happen . a little bitch would.	someone take his phone	a little bitch would
All you hoes wanna be like me so bad.	bad	you hoe s
I 'm so ugly & april fools bitch you thought.	you thought	so ugly
He draw (for " big bad love ") is a solid performance by arliss howard.	big bad love	a solid performance
A lackluster , unessential sequel to the classic disney adaptation of j.m. barrie 's peter pan	the classic dis ney adaptation	the classic disney adaptation

Table 1: Explanations from SELFEXPLAIN and LEXPLAIN for DWMW17, FDCL18 and SST2 (2 examples each). Predicted labels are toxic for DWMW17 and FDCL18. First and second SST2 examples are positive and negative, respectively. Explanations of LEXPLAIN align better with the model prediction and contain more task-related terms.

American English (AAE) and White American English. The AAE annotations are obtained from a demographically aligned ensemble model that learns a posterior distribution of topics corresponding to African American tweets (Blodgett et al., 2016). Our task lexicons and full experiment details are described in appendix section A.

Training We use SELFEXPLAIN as our baseline. When training both the baseline and LEXPLAIN, we keep the same hyperparameters and weights from the pretraining of the XLNet encoder and finetune the model for 5 epochs. In LEXPLAIN we do not use the GIL, since initial experiments showed no difference between adding and removing the GIL.

For LEXPLAIN, we perform hyperparameter tuning for $\alpha_1 \in \{0.01, 0.05, 0.1\}$ and $\alpha_2 \in \{0.8, 1.5.2.0\}$ on the development set. We report results on the best configuration on the test sets.

We extract phrases from sentences, by parsing each sentence with a constituency parser (Kitaev and Klein, 2018) and extracting all non-terminals with a token length of up to 5 words in the parse tree.

5 Evaluating Explanations

The goal of LEXPLAIN is to train models to produce plausible explanations that align with their predictions. We start with an intrinsic evaluation, verifying that LEXPLAIN indeed promotes *lexicon entries* as explanations. We then analyze the sufficiency of the explanations and conduct human evaluation to show that explanations from LEXPLAIN are more *plausible* and preferred by humans.

Intrinsic evaluation: are lexicon entries ranked higher as explanations of the model? The LIL outputs explanations as a rank of all input phrases. Following lexicon supervision, we expect to see that phrases ranked higher contain more lexicon entries, indicating that supervision was effective. To quantify this, we compute in Table 2 the mean reciprocal rank (MRR) of the lexicon entries within the ranked phrases of LEXPLAIN vs. the baseline.

Across all datasets, LEXPLAIN ranks lexicon entries higher than the baseline on average, showing the effectiveness of our supervision in providing explanations included in the task lexicon. We note that high-rank phrases should be the focus, thus in Appendix 2 we plot the raw counts of lexicon entries that appear in each rank, across sentences in each dataset. Clearly, LEXPLAIN puts more lexicon entries higher in the rank, this is especially noticeable in the highest ranked explanations (rank 1).

Dataset	Model	MRR(Full lexicon)	MRR(50% lexicon)
FDCL18	Baseline	0.29	0.31
	LEXPLAIN	**0.33**	**0.35**
DWMW17	Baseline	0.32	0.20
	LEXPLAIN	**0.35**	**0.24**
SST2	Baseline	0.23	0.18
	LEXPLAIN	**0.25**	**0.22**

Table 2: Mean reciprocal rank (MRR) of lexicon phrases across the full ranking of explanations on the test set.

Do explanations sufficiently reflect model predictions? *Sufficiency* measures how indicative explanations alone are of the model's predicted label (Jacovi et al., 2018; Yu et al., 2019). Sufficient explanations are expected to reflect the prediction of the predicted label on their own. To measure that, we use the FRESH pipeline (Jain et al., 2020): we train a BERT-based classifier to perform the task with only the explanations as input, and with the originally predicted labels as output. Higher accuracy on this task indicates that the explanations are more reflective of the model predictions. We train the sufficiency models with the top ranking explanations of each sentence as input.

Following Jain et al. (2020), we measure this with a BERT classifier trained with top ranked

phrases as input and predicted label as output. Higher accuracy indicates more sufficient explanations. Table 3 shows that LEXPLAIN explanations have higher predictive performance and are more sufficient on average compared to the baseline.

Dataset	Model	Top 1	Top 2
SST-2	Baseline	64.99	68.90
	LEXPLAIN	**68.00**	**70.00**
FDCL18	Baseline	82.25	87.37
	LEXPLAIN	**83.79**	**87.79**
DWMW17	Baseline	**88.16**	89.00
	LEXPLAIN	85.12	**91.10**

Table 3: Test set accuracy of sufficiency models trained on the top-1 and top-2 explanation as input.

Do humans prefer LEXPLAIN explanations? To evaluate how plausible our model's explanations are (Singh et al., 2019; Jin et al., 2020) we ask annotators to select their preferred explanations, comparing explanations from both the baseline and LEXPLAIN. We provide 3 annotators with 50 samples from the test set of each of our three datasets (9 annotators in total). All annotators are computer science graduate students and were already familiar with the tasks. Annotators were given a pair of explanations about the same input (one from the baseline, one from LEXPLAIN), in random order, and asked to select the one they prefer. They could also judge "both unsatisfactory" or "both satisfactory." The exact phrasing of the instructions can be found in Section B in the Appendix.

We analyse the human evaluations and take the max-vote preference of all three annotations per task. In Figure 1, we present the results of the human judgments. The differences between LEXPLAIN and the baseline are striking with LEXPLAIN being preferred about 3-4× more often than the baseline.

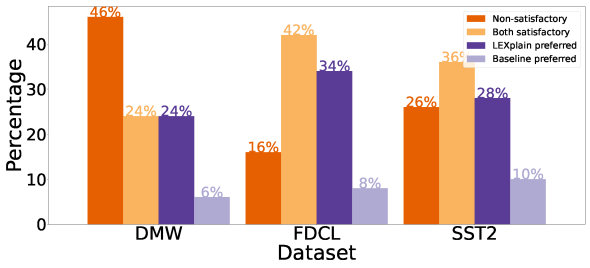

Figure 1: Results of human evaluation of explanation preference. LEXPLAIN is preferred by annotators 3-4× more often than the baseline.

6 Downstream Performance Analysis

To test our hypothesis that supervising explanations not only leads to plausible explanations, but robust models overfitting less to spurious confounds, we evaluate downstream classification performance.

Sentiment Analysis We obtain an accuracy of 93.92% and 93.35% for LEXPLAIN and SELF-EXPLAIN respectively. This slight improvement shows that the added supervision for explanations maintains the utility of the model.

Toxicity Detection We report the results on toxicity detection in Table 4. The accuracy results of LEXPLAIN are competitive with the baseline, also showing that additionally supervising explanations does not hurt the results of the classification task.

Dataset	Dialect	Model	Accuracy	FPR	FNR
FDCL18	All dialects	Baseline	93.94	**3.94**	10.05
		LEXPLAIN	94.10	4.24	**9.03**
	AAE	Baseline	93.60	**12.43**	4.21
		LEXPLAIN	93.60	13.87	**3.36**
DWMW17	All dialects	Baseline	96.06	10.98	**2.48**
		LEXPLAIN	96.30	**5.99**	3.24
	AAE	Baseline	98.00	21.69	**1.10**
		LEXPLAIN	97.95	**12.05**	1.59

Table 4: Toxicity accuracy, FPR, FNR on the test sets.

Demoting Spurious Correlations with Race Neural classifiers have been shown to rely on spurious artifacts in the training data (Kumar et al., 2019; Gururangan et al., 2018; McCoy et al., 2019), sometimes causing unfair predictions, when they relate to attributes like gender or race (Sap et al., 2019; Xia et al., 2020). We ask if guiding models to influential input phrases using lexicon reduces reliance on these artifacts and promote fairness.

Our toxicity data have dialect labels: African American English (AAE) and White American English. We inspect if our model demotes racial correlations. When a model relies on correlations harmfully, we expect higher false negatives rate (FNR), as more non-toxic instances are falsely labelled toxic because of reliance on dialectal features. In Table 4 we report the FPR (false positive rate) and FNR on DWMW17 and FDCL18. We get a much lower FPR on the full DWMW17, and more significant reduction on AAE samples. On the FDCL18 data, we see a slightly higher FPR than the baseline.

Lexicon Generalization We inspect the generalization abilities of LEXPLAIN: does it generalize and promote task related terms in explanations but

not present in the lexicon? We randomly select 50% of lexicon words and use them only to supervise while training. We compute MRR with respect to the other half not used for supervision on the same test set. If the phrases are ranked higher on average, even without being seen during training, it indicates that LEXPLAIN generalizes over lexicon phrases.

Table 2 shows the MRR of lexicon entries (not used as supervision). We show that our method generalizes consistently across all tasks: even lexicon entries absent during supervision are ranked higher with LEXPLAIN when compared to the baseline.

7 Related Work

Different works have approached interpreting models trained for various downstream tasks using post hoc (Simonyan et al., 2014; Jin et al., 2020; Smilkov et al., 2017) and intrinsic (Rajagopal et al., 2021b; Alvarez Melis and Jaakkola, 2018) methods. In this work we focus on intrinsic methods that highlight rationales (Denil et al., 2014; Rajani et al., 2019; Luo et al., 2021) – where parts of the input influential for prediction are extracted.

Some works leveraged interpretability methods to improve model performance (Han and Tsvetkov, 2021; Hase and Bansal, 2022). Wei et al. (2022) teach models to do commonsense tasks by providing step-by-step instructions. For classification tasks, Madaan et al. (2021) use free-form explanation generation and Hayati et al. (2022); Zaidan and Eisner (2008); Huang et al. (2021) use human rationales as model feedback. These methods require expensive annotation to elicit good explanations. We instead aim to supervise rationales using task lexicons, and show it yields improved explanations.

8 Conclusion

We propose LEXPLAIN, a method to improve model explanations by directly supervising them using task lexicons as the source of supervision. We show that our method is indeed able to promote dictionary entries as explanations, resulting in explanations that align well with the model's predicted label without sacrificing accuracy, and that the explanations are more plausible according to human evaluation. We also show that LEXPLAIN is able to generalize well to features that are not present in the supervising lexicon. Finally, we show that by promoting task related lexicon entries, we are able to demote spurious correlations with AAE annotations on toxicity datasets.

Limitations and Future Work

One limitation of LEXPLAIN stems from the reliance on task lexicons. First, a reliable task lexicon is required in order to adequately supervise explanations, and this might be non-trivial to create for an arbitrary task. We do show, however, that LEXPLAIN is able to generalize beyond lexicon entries, which suggests that even partial lexicon for the task at hand can provide a significant improvement in explanations. Second, the chosen lexicon might include certain biases itself, that might in turn be incorporated in the model and its explanations.

Another limitation, shared with the majority of existing interpretability methods, is that the faithfulness of interpretations is not guaranteed. In other words, there is no theoretical guarantee that the retrieved explanations reflect the actual mechanisms of the model in making predictions. We partially mitigate this by choosing SELFEXPLAIN as our base model. It is more faithful by design: it is trained to enforce the alignment between model outputs in the task classification and the LIL.

Finally, LEXPLAIN requires fine-tuning the model for the task and incorporating the LIL on top of a pretrained language model, and we established its success only with one model (XLNet). Future work should explore adaptations of other language models, and extensions to language generation, to facilitate model interpretability in new settings.

Ethics Statement

Our work aims at developing interpretable models that do not overfit to artifacts in the training data. However, there is no guarantee that we fully mitigate model reliance on all spurious correlations. Further, by incorporating new lexicons that might contain annotation biases (Sap et al., 2022), there is an additional risk to incorporate and amplify social biases. We mitigate these risks through manual analyses and fairness evaluations presented in §6.

We conduct fairness evaluations on the commonly used toxicity datasets (Davidson et al., 2017; Founta et al., 2018) annotated for AAE (Blodgett et al., 2016). These AAE annotations for the toxicity datasets are a useful but imperfect proxy for information about race. For example, these datasets are not annotated by in-group members and annotators had insufficient social context (Sap et al., 2019). Future work should focus on a more careful dataset curation that would enable a more reliable fairness evaluation.

Acknowledgments

This research is supported in part by by the National Science Foundation (NSF) under grants IIS2203097 and IIS2125201. This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *NeurIPS*.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Misha Denil, Alban Demiraj, and Nando de Freitas. 2014. Extraction of salient sentences from labelled documents. *ArXiv*, abs/1412.6815.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *ArXiv*, abs/1802.00393.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Xiaochuang Han and Yulia Tsvetkov. 2021. Influence tuning: Demoting spurious correlations via instance attribution and instance-driven updates. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4398–4409.

Peter Hase and Mohit Bansal. 2022. When can models learn from explanations? a formal framework for understanding the roles of explanation data. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 29–39, Dublin, Ireland. Association for Computational Linguistics.

Shirley Anugrah Hayati, Kyumin Park, Dheeraj Rajagopal, Lyle Ungar, and Dongyeop Kang. 2022. Stylex: Explaining styles with lexicon-based human perception. *arXiv preprint arXiv:2210.07469*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*.

Quzhe Huang, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. 2021. Exploring distantly-labeled rationales in neural network models. *ArXiv*, abs/2106.01809.

Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*.

Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. 2018. Understanding convolutional neural networks for text classification. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–65, Brussels, Belgium. Association for Computational Linguistics.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron Wallace. 2020. Learning to faithfully rationalize by construction. pages 4459–4473.

Xisen Jin, Junyi Du, Zhongyu Wei, X. Xue, and Xiang Ren. 2020. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. *ArXiv*, abs/1911.06194.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Sachin Kumar, Shuly Wintner, Noah A. Smith, and Yulia Tsvetkov. 2019. Topics to avoid: Demoting latent confounds in 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 4153–4163, Hong Kong, China. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *ArXiv*, abs/1612.08220.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Siwen Luo, Hamish Ivison, Soyeon Caren Han, and Josiah Poon. 2021. Local interpretations for explainable natural language processing: A survey. *ArXiv*, abs/2103.11072.

Aman Madaan, Niket Tandon, Dheeraj Rajagopal, Yiming Yang, Peter Clark, Keisuke Sakaguchi, and Eduard H. Hovy. 2021. Improving neural model performance through natural language feedback on their explanations. *ArXiv*, abs/2104.08765.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222.

Dheeraj Rajagopal, Vidhisha Balachandran, Eduard H. Hovy, and Yulia Tsvetkov. 2021a. Selfexplain: A self-explaining architecture for neural text classifiers. *CoRR*, abs/2103.12279.

Dheeraj Rajagopal, Vidhisha Balachandran, Eduard H Hovy, and Yulia Tsvetkov. 2021b. SELFEXPLAIN: A self-explaining architecture for neural text classifiers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 836–850, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *ICML*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034.

Chandan Singh, W. James Murdoch, and Bin Yu. 2019. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations*.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825.

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*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. 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 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Neural Information Processing Systems*.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi S Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. *arXiv preprint arXiv:1910.13294*.

Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *EMNLP*.

A Experimental Details

Training We use SELFEXPLAIN as our baseline. When training both the baseline and LEXPLAIN, we keep the same hyperparameters and weights from the pretraining of the XLNet encoder and finetune the model for 5 epochs. In LEXPLAIN we do not use the GIL, since initial experiments showed no difference between adding and removing the GIL.

For LEXPLAIN, we perform hyperparameter tuning for $\alpha_1 \in \{0.01, 0.05, 0.1\}$ and $\alpha_2 \in \{0.8, 1.5.2.0\}$ on the development set. We report results on the best configuration on the test sets.

Toxicity Dataset DWMW17 is a Twitter dataset with 25,000 tweets that have been annotated for hate speech, offensive, or none alongside dialect labels: African American English (AAE) and White American English. We merge the hatespeech and offensive examples and regard all of them as toxic. FDCL18 is also a Twitter dataset with 100,000 tweets annotated for hate, abuse, spam, and none. We select all instances, except for the ones labeled as spam. Again, we merge the hate and abuse examples and regard all of them as toxic. For all datasets we use the provided splits to train/dev./test.[2]

Task Lexicons Our sentiment lexicon of 2,470 words is derived by combining two existing lexicons: Hutto and Gilbert (2014) and Hu and Liu (2004). For toxicity detection, we use the lexicon from Wiegand et al. (2018), from which we extract 350 toxic words that appear in our datasets. We were only able to obtain a toxic lexicon. Our attempts to create a lexicon of non-toxic words by extracting the most salient words present in the non-toxic instances did not yield improved explanations. We opt to only supervise toxic instances in the training data.

B Human Evaluation

We ask annotators to select preferred explanations between the baseline and LEXPLAIN. They are presented with the model input, the original label and the predicted label and also All annotators are familiar with the tasks and are computer science graduate students.

Instructions given to human evaluators The task here is sentiment analysis. The labels are 0 for negative instances and 1 for positive instances. Please enter **X** or **Y** in the last column for the algorithm that provides the best explanation for the predicted label. If the explanations are the same for both algorithms, please enter **XY**. If the explanations for both algorithms are not satisfactory, please enter **NXY**. If explanations are not same, but both are satisfactory, please enter **SXY**.

[2]Train/dev./test: FDCL18: 54120/10145/11825, DWMW17: 17849/3001/3501, SST2: 66976/872/1821.

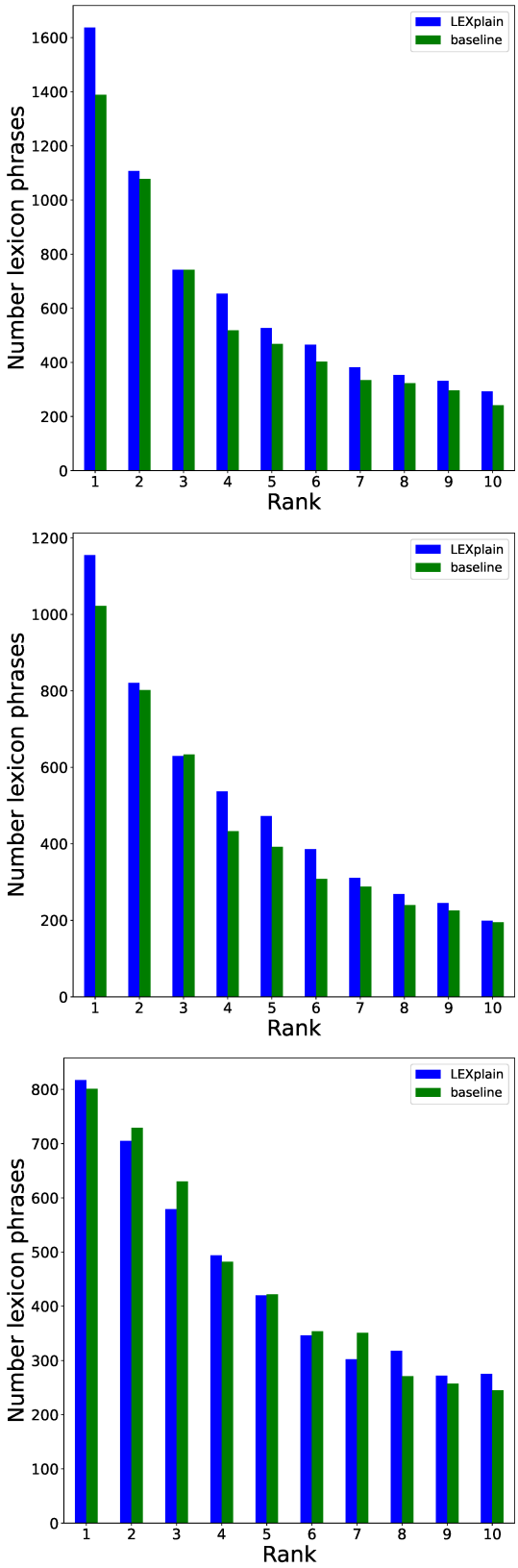

Figure 2: Number of lexicon entries in each rank across all sentences in each test set in the order of [FDCL18, DWMW17 and SST2].