Contextualizing Hate Speech Classifiers with Post-hoc Explanation

Brendan Kennedy[*] and **Xisen Jin**[*] and **Aida Mostafazadeh Davani**

Morteza Dehghani and **Xiang Ren**
University of Southern California
{btkenned,xisenjin,mostafaz,mdehghan,xiangren}@usc.edu

Abstract

Hate speech classifiers trained on imbalanced datasets struggle to determine if group identifiers like "gay" or "black" are used in offensive or prejudiced ways. Such biases manifest in false positives when these identifiers are present, due to models' inability to learn the contexts which constitute a hateful usage of identifiers. We extract post-hoc explanations from fine-tuned BERT classifiers to detect bias towards identity terms. Then, we propose a novel regularization technique based on these explanations that encourages models to learn from the context of group identifiers in addition to the identifiers themselves. Our approach improved over baselines in limiting false positives on out-of-domain data while maintaining or improving in-domain performance.[†]

1 Introduction

Hate speech detection is part of the ongoing effort to limit the harm done by oppressive and abusive language (Waldron, 2012; Gelber and McNamara, 2016; Gagliardone et al., 2015; Mohan et al., 2017). Performance has improved with access to more data and more sophisticated algorithms (e.g., Mondal et al., 2017; Silva et al., 2016; Del Vigna12 et al., 2017; Basile et al., 2019), but the relative sparsity of hate speech requires sampling using keywords (e.g., Olteanu et al., 2018) or sampling from environments with unusually high rates of hate speech (e.g., de Gibert et al., 2018; Hoover et al., 2019). Modern text classifiers thus struggle to learn a model of hate speech that generalizes to real-world applications (Wiegand et al., 2019).

A specific problem found in neural hate speech classifiers is their over-sensitivity to group identifiers like "Muslim", "gay", and "black", which are only hate speech when combined with the right

---
[*] Authors contributed equally
[†] Code is available here

"[F]or many <u>Africans</u>, the most threatening kind of ethnic hatred is <u>black</u> against <u>black</u>." - *New York Times*

"There is a great discrepancy between <u>whites</u> and <u>blacks</u> in SA. It is … [because] <u>blacks</u> will always be the most backward race in the world." Anonymous user, *Gab.com*

Figure 1: Two documents which are classified as hate speech by a fine-tuned BERT classifier. Group identifiers are underlined.

context (Dixon et al., 2018). In Figure 1 we see two documents containing the word "black" that a fine-tuned BERT model predicted to be hate speech, while only the second occurs in a hateful context.

Neural text classifiers achieve state-of-the-art performance in hate speech detection, but are uninterpretable and can break when presented with unexpected inputs (Niven and Kao, 2019). It is thus difficult to contextualize a model's treatment of identifier words. Our approach to this problem is to use the Sampling and Occlusion (SOC) explanation algorithm, which estimates model-agnostic, post-hoc feature importance (Jin et al., 2020). We apply this approach to the Gab Hate Corpus (Kennedy et al., 2020), a new corpus labeled for "hate-based rhetoric", and an annotated corpus from the Stormfront white supremacist online forum (de Gibert et al., 2018).

Based on the explanations generated via SOC, which showed models were biased towards group identifiers, we then propose a novel regularization-based approach in order to increase model sensitivity to the context surrounding group identifiers. We apply regularization during training to the explanation-based importance of group identifiers, coercing models to consider the context surrounding them.

We find that regularization reduces the attention given to group identifiers and heightens the importance of the more generalizable features of hate speech, such as dehumanizing and insulting language. In experiments on an out-of-domain test set

of news articles containing group identifiers, which are heuristically assumed to be non-hate speech, we find that regularization greatly reduces the false positive rate, while in-domain, out-of-sample classification performance is either maintained or improved.

2 Related Work

Our work is conceptually influenced by Warner and Hirschberg (2012), who formulated hate speech detection as disambiguating the use of offensive words from abusive versus non-abusive contexts. More recent approaches applied to a wide typology of hate speech (Waseem et al., 2017), build supervised models trained on annotated (e.g., Waseem and Hovy, 2016; de Gibert et al., 2018) or heuristically-labeled (Wulczyn et al., 2017; Olteanu et al., 2018) data. These models suffer from the highly skewed distributions of language in these datasets (Wiegand et al., 2019).

Research on bias in classification models also influences this work. Dixon et al. (2018) measured and mitigated bias in toxicity classifiers towards social groups, avoiding undesirable predictions of toxicity towards innocuous sentences containing tokens like "*gay*". Similarly, annotators' biases towards certain social groups were found to be magnified during classifier training Mostafazadeh Davani et al. (2020). Specifically within the domain of hate speech and abusive language, Park et al. (2018) and Sap et al. (2019) have defined and studied gender- and racial-bias, emphasizing issues of undetected dialect variation and imbalanced training data, respectively. Techniques for bias reduction in these settings include data augmentation by training on less biased data, term swapping during training (i.e., swapping gender words), and using debiased word embeddings (Bolukbasi et al., 2016).

Complementing these works, we directly manipulate models' modeling of the context surrounding identifier terms by regularizing explanations of these terms. Specifically, we use post-hoc explanation algorithms to interpret and modulate fine-tuned language models like BERT (Devlin et al., 2018), which achieve state of the art performance on many hate speech detection tasks (MacAvaney et al., 2019; Mandl et al., 2019). We focus on post-hoc explanation approaches, which interpret model predictions without elucidating the mechanisms by which the model works (Guidotti et al., 2019). These explanations reveal either word-level (Ribeiro et al., 2016; Sundararajan et al., 2017) or phrase-level importance (Murdoch et al., 2018; Singh et al., 2019) of inputs to predictions.

3 Data

We selected two public corpora for our experiments which highlight the rhetorical aspects of hate speech, versus merely the usage of slurs and explicitly offensive language (see Davidson et al., 2017). The "Gab Hate Corpus" (GHC; Kennedy et al., 2020) is a large, random sample ($N = 27,655$) from the Pushshift.io data dump of the Gab network [*], which we have annotated according to a typology of "hate-based rhetoric", a construct motivated by hate speech criminal codes outside the U.S. and social science research on prejudice and dehumanization. Gab is a social network with a high rate of hate speech (Zannettou et al., 2018; Lima et al., 2018) and populated by the "Alt-right" (Anthony, 2016; Benson, 2016). Similarly with respect to domain and definitions, de Gibert et al. (2018) sampled and annotated posts from the "Stormfront" web domain (Meddaugh and Kay, 2009) and annotated at the sentence level according to a similar annotation guide as used in the GHC.

Train and test splits were randomly generated for Stormfront sentences (80/20) with "hate" taken as a positive binary label, and a test set was compiled from the GHC by drawing a random stratified sample with respect to the "target population" tag (possible values including race/ethnicity target, gender, religious, etc.). A single "hate" label was created by taking the union of two main labels, "human degradation" and "calls for violence". Training data for the GHC (GHC$_{train}$) included 24,353 posts with 2,027 labeled as hate, and test data for the GHC (GHC$_{test}$) included 1,586 posts with 372 labeled as hate. Stormfront splits resulted in 7,896 (1,059 hate) training sentences, 979 (122) validation, and 1,998 (246) test.

4 Analyzing Group Identifier Bias

To establish and define our problem more quantitatively, we analyze hate speech models' bias towards group identifiers and how this leads to false positive errors during prediction. We analyze the top features of a linear model and use post-hoc explanations applied to a fine-tuned BERT model in order to measure models' bias towards these terms. We then establish the effect of these tendencies on
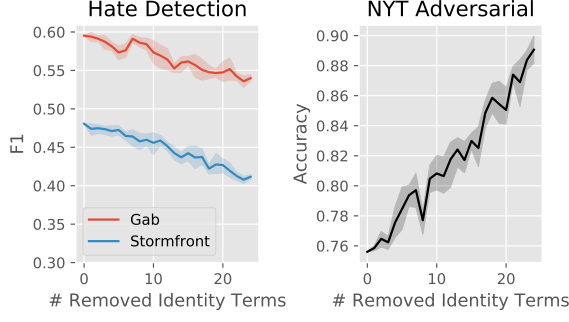
---

[*] https://files.pushshift.io/gab/

Figure 2: BoW F1 scores (trained on GHC_{train} and evaluated on GHC_{test}) as a function of how many group identifiers are removed (left). Accuracy of same models on NYT dataset with no hate speech (right).

model predictions using an adversarial-like dataset of New York Times articles.

4.1 Classification Models

We apply our analyses on two text classifiers, logistic regression with bag of words features and a fine-tuned BERT model (Devlin et al., 2018). The BERT model appends a special CLS token at the beginning of the input sentence and feeds the sentence into stacked layers of Transformer (Vaswani et al., 2017) encoders. The representation of the CLS token at the final layer is fed into a linear layer to perform 2-way classification (hate or non-hate). Model configuration and training details can be found in the Section A.3.

4.2 Model Interpretation

We first determine a model's sensitivity towards group identifiers by examining the models themselves. Linear classifiers can be examined in terms of their most highly-weighted features. We apply a post-hoc explanation algorithm for this task of extracting similar information from the fine-tuned methods discussed above.

Group identifiers in linear models From the top features in a bag-of-words logistic regression of hate speech on GHC_{train}, we collected a set of twenty-five identity words (not restricted to social group terms, but terms identifying a group in general), including "homosexual", "muslim", and "black", which are used in our later analyses. The full list is in Supplementals (A.1).

Explanation-based measures State-of-the-art fine-tuned BERT models are able to model complicated word and phrase compositions: for example, some words are only offensive when they are com-

posed with specific ethnic groups. To capture this, we apply a state-of-the-art Sampling and Occlusion (SOC) algorithm which is capable of generating hierarchical explanations for a prediction.

To generate hierarchical explanations, SOC starts by assigning importance score for phrases in a way that eliminates compositional effect between the phrase and its context $\mathbf{x}_\delta$ around it within a window. Given a phrase $\mathbf{p}$ appearing in a sentence $\mathbf{x}$, SOC assigns an importance score $\phi(\mathbf{p})$ to show how the phrase $\mathbf{p}$ contribute so that the sentence is classified as a hate speech. The algorithm computes the difference of the unnormalized prediction score $s(\mathbf{x})$ between "hate" and "non-hate" in the 2-way classifier. Then the algorithm evaluates average change of $s(\mathbf{x})$ when the phrase is masked with padding tokens (noted as $\mathbf{x} \backslash \mathbf{p}$) for different inputs, in which the N-word contexts around the phrase $\mathbf{p}$ are sampled from a pretrained language model, while other words remain the same as the given $\mathbf{x}$. Formally, the importance score $\phi(\mathbf{p})$ is measured as,

$$\phi(\mathbf{p}) = \mathbb{E}_{\mathbf{x}_\delta}[s(\mathbf{x}) - s(\mathbf{x} \backslash \mathbf{p})] \qquad (1)$$

In the meantime, SOC algorithm perform agglomerative clustering over explanations to generate a hierarchical layout.

Averaged Word-level SOC Explanation Using SOC explanations output on GHC_{test}, we compute average word importance and present the top 20 in Table 2.

4.3 Bias in Prediction

Hate speech models can be over-attentive to group identifiers, as we have seen by inspecting them through feature analysis and a post-hoc explanation approach. The effect of this during prediction is that models over-associate these terms with hate speech and choose to neglect the context around the identifier, resulting in false positives. To provide an external measure of models' over-sensitivity to group identifiers, we construct an adversarial test set of New York Times (NYT) articles that are filtered to contain a balanced, random sample of the twenty-five group identifiers (Section A.1). This gives us $12,500$ documents which are devoid of hate speech as defined by our typologies, excepting quotation.

It is key for models to not *ignore* identifiers, but to match them with the right context. Figure 2 shows the effect of ignoring identifiers: random
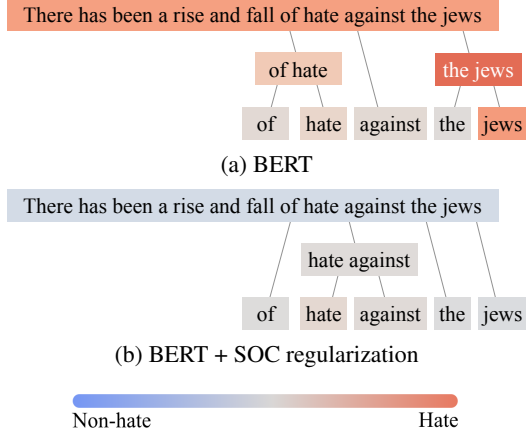
Figure 3: Hierarchical explanations on a test instance from GHC$_{test}$ before and after explanation regularization, where false positive predictions are corrected.

subsets of words ranging in size from 0 to 25 are removed, with each subset sample size repeated 5 times. Decreased rates of false positives on the NYT set are accompanied by poor performance in hate speech detection.

5 Contextualizing Hate Speech Models

We have shown hate speech models to be over-sensitive to group identifiers and unable to learn from the context surrounding these words during training. To address this problem in state-of-the-art models, we propose that models can be regularized to give no explained importance to identifier terms. We explain our approach as well as a naive baseline based on removing these terms.

Word Removal Baseline. The simplest approach is to remove group identifiers altogether. We remove words from the term list found in Section A.1 from both training and testing sentences.

Explanation Regularization. Given that SOC explanations are fully differentiable, during training, we regularize SOC explanations on the group identifiers to be close to 0 in addition to the classification objective $\mathcal{L}'$. The combined learning objective is written as follows.

$$\mathcal{L} = \mathcal{L}' + \alpha \sum_{w \in \mathbf{x} \cap S} [\phi(w)]^2, \qquad (2)$$

where S notes for the set of group names and $\mathbf{x}$ notes for the input word sequence. α is a hyperparameter for the strength of the regularization.

In addition to SOC, we also experiment with regularizing input occlusion (OC) explanations, defined as the prediction change when a word or

phrase is masked out, which bypass the sampling step in SOC.

6 Regularization Experiments

6.1 Experiment Details

Balancing performance on hate speech detection and the NYT test set is our quantitative measure of how well a model has learned the contexts in which group identifiers are used for hate speech. We apply our regularization approach to this task, and compare with a word removal strategy for the fine-tuned BERT model. We repeat the process for both the GHC and Stormfront, evaluating test set hate speech classification in-domain and accuracy on the NYT test set. For the GHC, we used the full list of 25 terms; for Stormfront, we used the 10 terms which were also found in the top predictive features in linear classifiers for the Stormfront data. Congruently, for Stormfront we filtered the NYT corpus to only contain these 10 terms ($N = 5,000$).

6.2 Results

Performance is reported in Table 1. For the GHC, we see an improvement for in-domain hate speech classification, as well as an improvement in false positive reduction on the NYT corpus. For Stormfront, we see the same improvements for in-domain F$_1$) and NYT. For the GHC, the most marked difference between BERT+WR and BERT+SOC is increased recall, suggesting that baseline removal largely mitigates bias towards identifiers at the cost of more false negatives.

As discussed in section 4.2, SOC eliminates the compositional effects of a given word or phrase. As a result, regularizing SOC explanations does not prohibit the model from utilizing contextual information related to group identifiers. This can possibly explain the improved performance in hate speech detection relative to word removal.

Word Importance in Regularized Models We determined that regularization improves a models focus on non-identifier context in prediction. In table 2 we show the changes in word importance as measured by SOC. Identity terms' importance decreases, and we also see a significant increase in importance of terms related to hate speech ("poisoned", "blamed", etc.) suggesting that models have learned from the identifier terms' context.

Visualizing Effects of Regularization We can further see the effect of regularization by considering Figure 3, where hierarchically clustered expla-

Training set	GHC				Stormfront			
Method / Metrics	Precision	Recall	F1	NYT Acc.	Precision	Recall	F1	NYT Acc.
BoW	62.80	56.72	59.60	75.61	36.95	58.13	45.18	66.78
BERT	69.87 ± 1.7	66.83 ± 7.0	67.91 ± 3.1	77.79 ± 4.8	**57.76 ± 3.9**	54.43 ± 8.1	55.44 ± 2.9	92.29 ± 4.1
BoW + WR	54.65	52.15	53.37	89.72	36.24	55.69	43.91	81.34
BERT + WR	67.61 ± 2.8	60.08 ± 6.6	63.44 ± 3.1	89.78 ± 3.8	53.16 ± 4.3	**57.03 ± 5.7**	54.60 ± 1.7	92.47 ± 3.4
BERT + OC (α=0.1)	60.56 ± 1.8	**69.72 ± 3.6**	64.14 ± 3.2	89.43 ± 4.3	57.47 ± 3.7	51.10 ± 4.4	53.82 ± 1.3	95.39 ± 2.3
BERT + SOC (α=0.1)	**70.17 ± 2.5**	69.03 ± 3.0	**69.52 ± 1.3**	83.16 ± 5.0	57.29 ± 3.4	54.27 ± 3.3	**55.55 ± 1.1**	93.93 ± 3.6
BERT + SOC (α=1.0)	64.29 ± 3.1	69.41 ± 3.8	66.67 ± 2.5	**90.06 ± 2.6**	56.05 ± 3.9	54.35 ± 3.4	54.97 ± 1.1	**95.40 ± 2.0**

Table 1: Precision, recall, F_1 (%) on GHC$_{test}$ and Stormfront (Stf.) test set and accuracy (%) on NYT evaluation set. We report mean and standard deviation of the performance across 10 runs for BERT, BERT + WR (word removal), BERT + OC, and BERT + SOC.

BERT	Δ Rank	Reg.	Δ Rank
ni**er	+0	ni**er	+0
ni**ers	-7	fag	+35
kike	-90	traitor	+38
mosques	-260	faggot	+5
ni**a	-269	bastard	+814
jews	-773	blamed	+294
kikes	-190	alive	+1013
nihon	-515	prostitute	+56
faggot	+5	ni**ers	-7
nip	-314	undermine	+442
islam	-882	punished	+491
homosexuality	-1368	infection	+2556
nuke	-129	accusing	+2408
niro	-734	jaggot	+8
muhammad	-635	poisoned	+357
faggots	-128	shitskin	+62
nitrous	-597	ought	+229
mexican	-51	rotting	+358
negro	-346	stayed	+5606
muslim	-1855	destroys	+1448

Table 2: Top 20 words by mean SOC weight before (BERT) and after (Reg.) regularization for GHC. Changes in the rank of importance as a result of regularization are also shown. Curated set of group identifiers are highlighted.

nations from SOC are visualized before and after regularization, correcting a false positive.

7 Conclusion & Future Work

Regularizing SOC explanations of group identifiers tunes hate speech classifiers to be more context-sensitive and less reliant on high-frequency words in imbalanced training sets. Complementing prior work in bias detection and removal in the context of hate speech and in other settings, our method is directly integrated into Transformer-based models and does not rely on data augmentation. As such, it is an encouraging technique towards directing models' internal representation of target phenomena via lexical anchors.

Future work includes direct extension and validation of this technique with other language models such as GPT-2 (Radford et al., 2019); experimenting with other hate speech or offensive language datasets; and experimenting with these and other sets of identity terms. Also motivated by the present work is the more general pursuit of integrating structure into neural models like BERT.

Regularized hate speech classifiers increases sensitivity to the compositionality of hate speech, but the phenomena remain highly complex rhetorically and difficult to learn through supervision. For example, this post from the GHC requires background information and reasoning across sentences in order to classify as offensive or prejudiced: "Donald Trump received much criticism for referring to Haiti, El Salvador and Africa as 'shitholes'. He was simply speaking the truth." The examples we presented (see Appendix 4 and 5) show that regularization leads to models that are context-sensitive to a degree, but not to the extent of reasoning over sentences like those above. We hope that the present work can motivate more attempts to inject more structure into hate speech classification.

Explanation algorithms offer a window into complex predictive models, and regularization as performed in this work can improve models' internal representations of target phenomena. In this work, we effectively applied this technique to hate speech classifiers biased towards group identifiers; future work can determine the effectiveness and further potential for this technique in other tasks and contexts.

References

Andrew Anthony. 2016. Inside the hate-filled echo chamber of racism and conspiracy theories. *The guardian*, 18.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.

Thor Benson. 2016. Inside the twitter for racists: Gab the site where milo yiannopoulos goes to troll now.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international AAAI conference on web and social media*.

Fabio Del Vigna12, Andrea Cimino23, Felice DellOrletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73. ACM.

Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online hate speech*. Unesco Publishing.

Katharine Gelber and Luke McNamara. 2016. Evidencing the harms of hate speech. *Social Identities*, 22(3):324–341.

Ona de Gibert, Naiara Perez, Aitor García Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93.

Joseph Hoover, Mohammad Atari, Aida Mostafazadeh Davani, Brendan Kennedy, Gwenyth Portillo-Wightman, Leigh Yeh, Drew Kogon, and Morteza Dehghani. 2019. Bound in hatred: The role of group-based morality in acts of hate. *PsyArxiv Preprint 10.31234/osf.io/359me*.

Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2020. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *International Conference on Learning Representations*.

Brendan Kennedy, Mohammad Atari, Aida M Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs Jr., Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Gabriel Cardenas, Alyzeh Hussain, Austin Lara, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, Yong Zhang, Beth Meyerowitz, and Morteza Dehghani. 2020. The gab hate corpus: A collection of 27k posts annotated for hate speech.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Lucas Lima, Julio CS Reis, Philipe Melo, Fabricio Murai, Leandro Araujo, Pantelis Vikatos, and Fabricio Benevenuto. 2018. Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 515–522. IEEE.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8).

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, pages 14–17.

Priscilla Marie Meddaugh and Jack Kay. 2009. Hate speech or "reasonable racism?" the other in stormfront. *Journal of Mass Media Ethics*, 24(4):251–268.

Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The impact of toxic language on the health of reddit communities. In *Canadian Conference on Artificial Intelligence*, pages 51–56. Springer.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 85–94. ACM.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, Shreya Havaldar, and Morteza Dehghani. 2020. Hatred is in the eye of the annotator: Hate speech classifiers learn human-like social stereotypes (in press). In *31st Annual Conference of the Cognitive Science Society (CogSci)*.

W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from LSTMs. In *International Conference on Learning Representations*.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664.

Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush R Varshney. 2018. The effect of extremist violence on hateful speech online. In *Twelfth International AAAI Conference on Web and Social Media*.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Open AI Blog*.

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*, pages 1135–1144. ACM.

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.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Tenth International AAAI Conference on Web and Social Media*.

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

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*, pages 3319–3328. JMLR. org.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Jeremy Waldron. 2012. *The harm in hate speech*. Harvard University Press.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. 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 602–608.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399. International World Wide Web Conferences Steering Committee.

Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, pages 1007–1014. International World Wide Web Conferences Steering Committee.

A Appendices

A.1 Full List of Curated Group Identifiers

muslim jew jews white islam blacks muslims women whites gay black democat islamic allah jewish lesbian transgender race brown woman mexican religion homosexual homosexuality africans

Table 3: 25 group identifiers selected from top weighted words in the TF-IDF BOW linear classifier on the GHC.

jew jews mexican blacks jewish brown black muslim homosexual islam

Table 4: 10 group identifiers selected for the Stormfront dataset.

A.2 Visualizations of Effect of Regularization
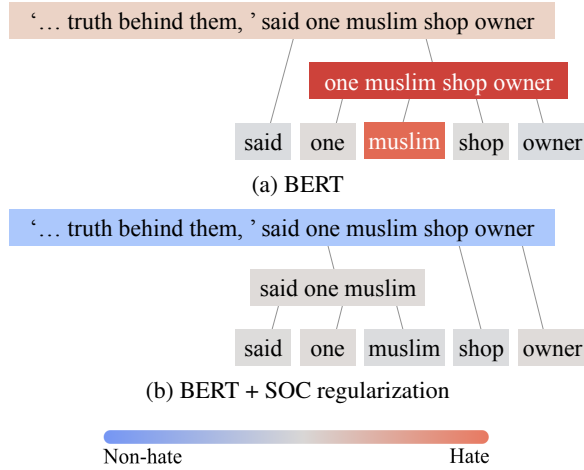

(a) BERT

(b) BERT + SOC regularization

Figure 4: Hierarchical explanations on a test instance from the NYT dataset where false positive predictions are corrected.

A.3 Implementation Details

Training Details. We fine-tune over the BERT-base model using the public code[†], where the batch size is set to 32 and the learning rate of the Adam (Kingma and Ba, 2015) optimizer is set to 2×10^{-5}. The validation is performed every 200 iterations and the learning rate is halved when the validation F1 decreases. The training stops when the learning rate is halved for 5 times. To handle the data imbalance issue, we reweight the training loss so that positive examples are weighted 10

[†] https://github.com/huggingface/transformers
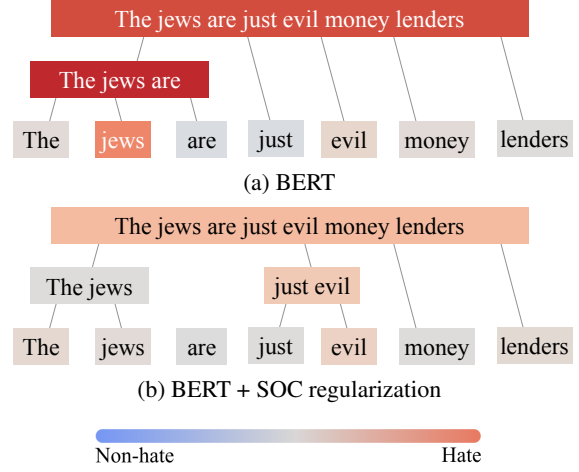

(a) BERT

(b) BERT + SOC regularization

Figure 5: Hierarchical explanations on a test instance from the Gab dataset where both models make correct positive predictions. However, the explanations reveal that only the regularized model is making correct predictions for correct reasons.

times as negative examples on the Gab dataset and 8 times on the Stormfront dataset.

Explanation Algorithm Details. For the SOC algorithm, we set the number of samples and the size of the context window as 20 and 20 respectively for explanation analysis, and set two parameters as 5 and 5 respectively for explanation regularization.

A.4 Cross-Domain Performance

In addition to evaluating each model within-domain (i.e., training on GHC_{train} and evaluating on GHC_{test}) we evaluated each model across domains. The results of these experiments, conducted in the same way as before, are presented in Table 5.

Method / Dataset	Gab → Stf. F1	Stf. → Gab F1
BoW	32.39	46.71
BERT	42.84 ± 1.2	53.80 ± 5.5
BoW + WR	27.45	44.81
BERT + WR	39.10 ± 1.3	55.31 ± 4.0
BERT + OC (α=0.1)	40.60 ± 1.6	56.90 ± 1.8
BERT + SOC (α=0.1)	41.88 ± 1.0	55.75 ± 2.1
BERT + SOC (α=1.0)	39.20 ± 2.7	56.82 ± 3.9

Table 5: Cross domain F1 on Gab, Stormfront (Stf.) datasets. We report mean and standard deviation of the performance within 10 runs for BERT, BERT + WR (word removal), BERT + OC, and BERT + SOC.