Reducing Target Group Bias in Hate Speech Detectors

Anonymous ACL submission

Abstract

The ubiquity of offensive and hateful content on online fora necessitates the need for automatic solutions that detect such content competently across target groups. In this paper we show that text classification models trained on large publicly available datasets despite having a high overall performance, may significantly under-perform on several protected groups. On the Vidgen et al. (2020) dataset, we find the accuracy to be 37% lower on an under annotated Black Women target group and 12% lower on Immigrants, where hate speech involves a distinct style. To address this, we propose to perform token-level hate sense disambiguation, and utilize tokens' hate sense representations for detection, modeling more general signals. On two publicly available datasets, we observe that the variance in model accuracy across target groups drops by at least 30%, improving the average target group performance by 4% and worst case performance by 13%.

1 Introduction

The diverse nature of hate speech against distinct target groups makes its automatic detection very challenging. In this paper, we study the impact of training machine learning models on two public hate speech datasets, where the content is organically driven by forum users, making the subsequent corpora unbalanced. While datasets should reflect content produced in the real world, we find models trained on such unbalanced datasets to perform with varying competence across target groups – demographic segmentations, often being poorer for protected groups. For example a BERT model (Devlin et al., 2019) trained and evaluated on the dataset in Vidgen et al. (2020), has a high variance in detection accuracy across different target groups, significantly under performing on attacks against Gay Men and Black Women (see Figure 1).

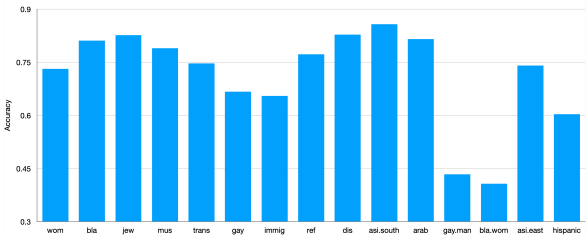

Figure 1: The performance of a state-of-the-art model on hate speech detection across different target groups on Vidgen et al. (2020).

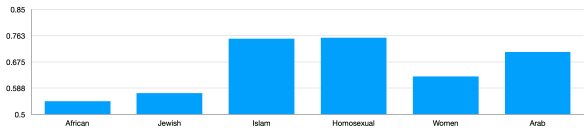

Figure 2: The performance of a state-of-the-art model on hate speech detection across different target groups on Mathew et al. (2020).

Our analysis of this bias – high variance in detection accuracy across target groups, shows that data distribution in these unbalanced datasets is a critical factor. Hate speech detection on a target group is more challenging with fewer corresponding training data points. Additionally, stylistic differences in hateful and offensive text against different minorities also plays a role in poor performance, as discussed in Section 2.

We propose to address this using a token level hate sense disambiguation based approach. Tokens like *kill* and *gay* can be hateful when targeting a particular group and used in malicious context. To distinguish the hateful application from benign, we implement a token-level model which predicts the hate sense (distribution over class labels) at every time-step while predicting the overall hate speech class. Subsequently, the classifier considers hate sense augmented token representations, allowing a more general detection solution. Experimentally, we show that our approach leads to a more balanced performance with a 30% drop in variance

Target Group	Training Data	Word Overlap	Eval Accuracy
Women	1652	0.65	0.73
Black	1580	0.79	0.81
Jew	891	0.70	0.83
Muslim	779	0.66	0.79
Transgender	640	0.64	0.75
Gay	580	0.71	0.67
Immigrants	545	0.58	0.66
Refugee	376	0.57	0.77
Disable	374	0.58	0.83
South Asian	274	0.51	0.86
Arab	262	0.53	0.82
Gay Men	217	0.43	0.43
Black Women	144	0.45	0.41
East Asian	144	0.47	0.74
Hispanic	57	0.15	0.60

Table 1: Performance of a BERT model on different target groups in Vidgen et al. (2020). Statistics on specific number of training data points and fraction of word overlap are also provided.

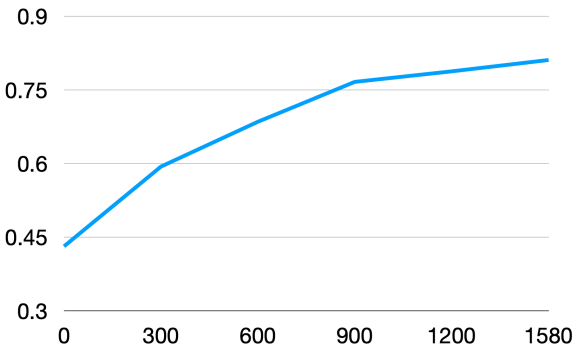

Figure 3: Model performance on the "Black" target group with an increasing number of "Black" training data points.

across target groups and has an at least 4% greater average-across-target groups performance than a BERT-based baseline.

In summary, the contribution of this paper include:
(1) We extensively highlight a crucial problem in NLP models having an unbalanced hate speech detection capabilities across different target groups.
(2) We propose a zero-shot token level hate sense disambiguation technique with no target group features to address this.
(3) Our technique leads to an absolute 3% gain in average detection accuracy over target group accuracy with a significant drop in group-wise variance.

2 Motivation and Analysis

In this section, we study the performance of a BERT model trained on Vidgen et al. (2020).

Biased Performance The BERT hate speech detection model has a biased performance as seen in Table 1. For instance, model accuracy on the Gay Men target group is 43% which is almost half of 85% on South Asian's. We hypothesize that these results are due to two factors: (1) Training data available for each target group; (2) Stylistic differences in hate text used across target groups.

Training Data We investigate the impact of training data available for every target group and the corresponding model performance. In particular, we look at the model performance on the Black target group with an increasing number of corresponding training data available. Figure 3 shows that performance on the Black target group improves with an increase in training data.

On the complete dataset we also see that target groups with more training data such as Black, Jew and Muslim target groups (Table 1) have a higher test performance than Gay Men, Black Women and Hispanics target groups which have fewer training data points. However, the size of training data is not the only deciding factor for performance. For instance, the performance on South Asian and Arab target groups is much higher than performance against Immigrant and Women target groups, the latter with far more training data. Overall, training data is an important but not exclusive factor in hate speech detection performance across target groups.

Stylistic Differences Hateful text varies according to the intended target group, hence making such datasets a mixture of unique sub-domains. Such stylistic differences have the potential to cause a variance in performance across target groups. Table 1 reports the token overlap for the most frequent tokens used against different target groups with most frequent tokens used in the rest of the data. A higher word overlap for Black, Jew, Women and Muslim target groups corresponds to a high test accuracy, while a lower word overlap for Immigrant, Hispanics and Black Women target groups corresponds to a lower test performance. Performance on Arabs and South Asians target groups with low word overlap is higher than the performance against Gay target group which has a higher word overlap. Overall, the stylistic differences does not explain all the bias but is a strong factor.

3 Towards Unbiased Modeling

In this section, we propose a token-level model which performs sense disambiguation enroute to the overall hate speech prediction. Specifically, we develop our model to detect hate speech related

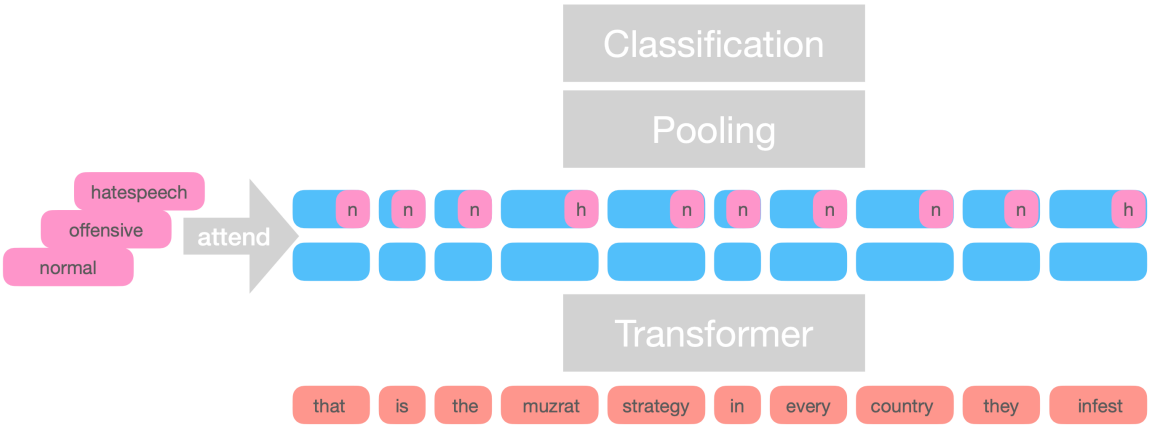

Figure 4: Figure illustrating the flow of our model.

senses for all tokens using their contextual information – conceptually similar to (Murdoch et al., 2018; Kennedy et al., 2020) but without utilizing any additional annotations. Apart from augmenting the tokens' hidden representations with their sense representations, our model is regularized to force the consensus of the token level hate sense predictions to agree with the target hate sense. This enables our model to rely on general signals for hate speech detection compared to vanilla models.

Architecture Figure 4 shows our model architecture. We consider a Transformer based text encoder E to represent our inputs. For a potential hateful text input $x = \{x_1, x_2, ..., x_n\}$, our model produces representations for every token $E(x) = [E(x_1), E(x_2), ..., E(x_n)]$. The named hate speech classes $C = \{c_1, c_2, ..., c_k\}$ are also represented by using E on their corresponding class names to get $\{E(c_1), E(c_2), ..., E(c_k)\}$ through encoding and subsequent pooling.

Every hidden state $E(x_i)$ in $E(x)$ is attended to the class representations $\{E(c_1), E(c_2), ..., E(c_k)\}$. The hate sense s_i for the hidden state $E(x_i)$ is categorized as token x_i's sense, where :

$$s_i = \arg\max_j \frac{\exp(cos(E(x_i), E(c_j)))}{\sum_{l=1}^{k} \exp(cos(E(x_i), E(c_l)))} \quad (1)$$

The final prediction, $f(x) = C(E(x))$ with C a Multi-layer Perceptron and Pooling classifier and E the encoder utilize this sense prediction s and attended hidden representations. Specifically, $f(x) = C([E(x_1) + E(c_{s_1}), E(x_2) + E(c_{s_2}), ..., E(x_n) + E(c_{s_n})])$ where a max-pooling and multi-layer perceptron classifier C is applied to the attended representations.

Optimization In addition to minimizing the final loss $L(f(x), y)$, we enforce constraints on the token level sense predictions having their consensus match the final hate speech label (M selects the max occurring hateful sense):

$$L(M(s_1, s_2, ..., s_n), y) \quad (2)$$

We enforce the number of unique hateful senses to be minimized (U selects all unique hateful senses):

$$||U(s_1, s_2, ..., s_n)||_{L_1} \quad (3)$$

We hypothesize that our sense prediction approach, implemented through these constraints, models sentence semantics better for robust detection.

4 Experiments

We report performance across target groups on two public datasets Learning from the Worst *LearningWorst* (Vidgen et al., 2020) and *HateXplain* (Mathew et al., 2020). Both these datasets have annotations on the target groups.[1] Tables 2 and 3 list the target groups in the respective datasets.[2]

We consider a BERT document level text-classification model (Devlin et al., 2019) for hate speech detection. We develop our token-level sense disambiguation model on top of this model. The models are implemented using the Huggingface library (Wolf et al., 2019).

Results Tables 2 and 3 report the results of the baseline BERT method and our debiasing approach

[1]To the best of our knowledge the performance per target group has not been previously reported.

[2]We consider all target groups with at least 25 data points in the test set.

Target Group	Baseline Performance	Method Performance
Women	**0.73**	0.71
Black	0.81	**0.83**
Jew	0.83	**0.85**
Muslim	0.79	**0.82**
Transgender	0.75	**0.78**
Gay	0.67	**0.74**
Immigrants	0.66	**0.69**
Refugee	0.77	0.77
Disable	**0.83**	0.80
South Asian	**0.86**	0.82
Arab	0.82	**0.85**
Gay Men	0.43	**0.57**
Black Women	0.41	**0.59**
East Asian	0.74	**0.79**
Hispanic	**0.60**	0.56
Test Performance	**0.78**	0.77
Average (across targets)	0.71	**0.74**
Performance Variance	0.14	**0.10**

Table 2: Comparison of baseline BERT and token-level classification model on *LearningWorst*.

Target Group	Baseline Performance	Method Performance
African	0.54	**0.75**
Jewish	0.57	**0.79**
Islam	**0.75**	0.71
Homosexual	**0.76**	0.73
Women	**0.63**	0.61
Arab	0.71	**0.74**
Test Performance	**0.77**	0.76
Average (across targets)	0.66	**0.72**
Performance Variance	0.09	**0.06**

Table 3: Comparison of baseline BERT and token-level classification model on *HateXplain*.

on the *LearningWorst* and *HateXplain* datasets respectively. Table 2 reports how our method is able to reduce the variance in accuracy across different target groups while improving the average accuracy across all target groups. The performance on several poor performing target groups like Gay Men, Black Women and Immigrants is significantly improved by our method. The lowest accuracy is now on the Hispanics target group at 56% which is significantly higher than the original lowest of 41% on the Black Women target group. This balanced performance comes at a slight cost of 1% drop in the overall test accuracy. Similarly, Table 3 reports how our method is able to reduce the variance in accuracy across different target groups while improving the average performance across all target groups. The accuracy on poor performing African and Jewish target groups is significantly improved by our method. The lowest performance is now on Women target group at 60% which is higher than the previous lowest on African target group at 54%. The balanced performance comes at a slight cost of 1% drop in the overall test accuracy.

Our method is effective in reducing the bias by performing better in scenarios with fewer training data points and greater stylistic differences.

5 Related Work

Bias in Hate Speech Detection The growth of hate and abuse online has inspired the collection of several datasets to study the phenomenon (Waseem and Hovy, 2016; Waseem, 2016; Davidson et al., 2017; Founta et al., 2018; Mandl et al., 2019, 2020; Kumar et al., 2018; Zampieri et al., 2019; Mathew et al., 2020). While these datasets form numerous benchmarks to compare machine learning solutions, several issues have been identified with hate speech training datasets – lack of linguistic variety and annotations (Vidgen et al., 2019a; Poletto et al., 2021; Röttger et al., 2021). In particular, sampling data by searching for keywords can lead to trained models that are biased towards keywords (Vidgen et al., 2019b; Wiegand et al., 2019). While Davani et al. (2020); Toutanova et al. (2021) highlight target group bias through counterfactuals, in our work we are identifying and reducing general bias in performance across target groups without using any group specific annotations.

Few Shot Sense Detection Word Sense Detection (Miller et al., 1993) is a long standing task of identifying the meaning of a word in a specific text. Recent methods (Huang et al., 2019; Blevins and Zettlemoyer, 2020; Bevilacqua and Navigli, 2020) have outperformed human performance on sense detection (Navigli, 2009). In scenarios where certain senses are rare, the performance of typical models is not optimal and a BERT based description of the senses helps alleviate the low resource problem (Blevins et al., 2021). In this work, we focus on identifying hateful senses as annotated in the training datasets, using their BERT representations. Despite having no sense annotations, we use the class names to assign token level hate senses.

6 Discussion

This paper demonstrates that models trained on hate speech datasets may have biased performance across different target groups. Our analysis shows that additional training data related to a target group is beneficial, highlighting the need for a more balanced collection of hateful text. We suggest a sense-based solution to address this issue, leading to a better average performance across different target groups.

References

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864.

Terra Blevins, Mandar Joshi, and Luke Zettlemoyer. 2021. Fews: Large-scale, low-shot word sense disambiguation with the dictionary. *arXiv preprint arXiv:2102.07983*.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss-informed biencoders. *arXiv preprint arXiv:2005.02590*.

Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren, and Morteza Dehghani. 2020. Fair hate speech detection through evaluation of social group counterfactuals. *arXiv preprint arXiv:2010.12779*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

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

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. In *Twelfth International AAAI Conference on Web and Social Media*.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. Glossbert: Bert for word sense disambiguation with gloss knowledge. *arXiv preprint arXiv:1908.07245*.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Forum for Information Retrieval Evaluation*, pages 29–32.

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.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

W. James Murdoch, Peter J. Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from lstms. *ArXiv*, abs/1801.05453.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou. 2021. Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019a.

Challenges and frontiers in abusive content detection. In *Proceedings of the third workshop on abusive language online*, pages 80–93.

Bertie Vidgen, Helen Margetts, and Alex Harris. 2019b. How much online abuse is there. *Alan Turing Institute*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2020. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761*.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.