A Weakly Supervised Classifier and Dataset of White Supremacist Language

Michael Miller Yoder[1]   **Ahmad Diab**[2]   **David West Brown**[3]   **Kathleen M. Carley**[1]

[1]Software and Societal Systems Department, Carnegie Mellon University
[2]Department of Computer Science, University of Pittsburgh
[3]Department of English, Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
yoder@cs.cmu.edu ahd23@pitt.edu dwb2@andrew.cmu.edu carley@cs.cmu.edu

Abstract

We present a dataset and classifier for detecting the language of white supremacist extremism, a growing issue in online hate speech. Our weakly supervised classifier is trained on large datasets of text from explicitly white supremacist domains paired with neutral and anti-racist data from similar domains. We demonstrate that this approach improves generalization performance to new domains. Incorporating anti-racist texts as counterexamples to white supremacist language mitigates bias.

1 Introduction

The spread of white supremacist extremism online has motivated offline violence, including recent mass shootings in Christchurch, El Paso, Pittsburgh, and Buffalo. Though some research in natural language processing has focused on types of hate speech, such as anti-Black racism (Kwok and Wang, 2013) and misogyny (Fersini et al., 2018), little work has focused on detecting specific hateful ideologies. Practitioners have called for such systems, particularly for white supremacism (ADL, 2022; Yoder and Habib, 2022).

To detect white supremacist language, we build text classifiers trained on data from a large, diverse set of explicitly white supremacist online spaces, filtered to ideological topics.[1] In a weakly supervised set-up, we train discriminative classifiers to distinguish texts in white supremacist domains from texts in similar online spaces that are not known for white supremacism. These classifiers outperform prior work in white supremacist classification on three annotated datasets, and we find that the best-performing models use a combination of weakly and manually annotated data.

Hate speech classifiers often have difficulty generalizing beyond data they were trained on (Swamy

et al., 2019; Yoder et al., 2022). We evaluate our classifiers on unseen datasets annotated for white supremacism from a variety of domains and find strong generalization performance for models that incorporate weakly annotated data.

Hate speech classifiers often learn to associate any mention of marginalized identities with hate, regardless of context (Dixon et al., 2017). To address this potential issue with white supremacist classification, we incorporate anti-racist texts, which often mention marginalized identities in positive contexts, as counter-examples to white supremacist texts. Evaluating on a synthetic test set with mentions of marginalized identities in a variety of contexts (Röttger et al., 2021), we find that including anti-racist texts helps mitigate this bias.

2 The Language of White Supremacist Extremism

This work focuses on white supremacist extremism, social movements advocating for the superiority of white people and domination or separation from other races (Daniels, 2009). This fringe movement both exploits the bigotries widely held in societies with structural white supremacism and makes them explicit (Ferber, 2004; Berlet and Vysotsky, 2006; Pruden et al., 2022). Key beliefs of white supremacist extremism are that race and gender hierarchies are fixed, that white people's "natural" power is threatened, and that action is needed to protect the white race (Ferber and Kimmel, 2000; Brown, 2009; Perry and Scrivens, 2016; Ansah, 2021).

Many qualitative studies have examined the language of white supremacism (Thompson, 2001; Duffy, 2003; Perry and Scrivens, 2016; Bhat and Klein, 2020). Computational models have been developed to identify affect (Figea et al., 2016), hate speech (de Gibert et al., 2019), and violent intent (Simons and Skillicorn, 2020) within white supremacist forums.

---

[1]See https://osf.io/274z3/ to access public parts of this dataset and others used in this paper.

Two other studies have built models to detect white supremacist ideology in text. Alatawi et al. (2021) test Word2vec/BiLSTM models, pre-trained on a corpus of unlabeled white supremacist forum data, as well as BERT models. To estimate the prevalence of white supremacism on Twitter after the 2016 US election, Siegel et al. (2021) build a dictionary-based classifier and validate their findings with unlabeled alt-right Reddit data. In contrast, we use a large, domain-general white supremacist corpus with carefully selected negative training examples to build a weakly supervised discriminative classifier for white supremacism.

2.1 Hate speech and white supremacism

The relationship between hate speech and white supremacism has been theorized and annotated in different ways. Some have annotated the glorification of ideologies and groups such as Nazism and the Ku Klux Klan separately from hate speech (Siegel et al., 2021; Rieger et al., 2021), which is often defined as verbal attacks on groups based on their identity (Sanguinetti et al., 2018; Poletto et al., 2021; de Gibert et al., 2019). A user of Stormfront, a white supremacist forum, notes this distinction to evade moderation on other platforms: "Nationalist means defending the white race; racist means degrading non-white races. You should be fine posting about preserving the white race as long as you don't degrade other races."[2]

We aim to capture the expression of white supremacist ideology beyond just hate speech against marginalized identities (see Figure 1). In contrast, de Gibert et al. (2019) ask annotators to identify hate speech within a white supremacist forum. They note that some content that did not fit strict definitions of hate speech still exhibited white supremacist ideology. Examples of this from data used in the current paper include "diversity means chasing down whites" (white people being threatened) and "god will punish as he did w/ hitler" (action needed to protect white people).

3 Weakly Annotated Data

It is difficult for annotators to determine whether the short texts commonly used in NLP and computational social science, such as tweets, express white supremacism or other far-right ideologies. Alatawi et al. (2021) struggle to reach adequate
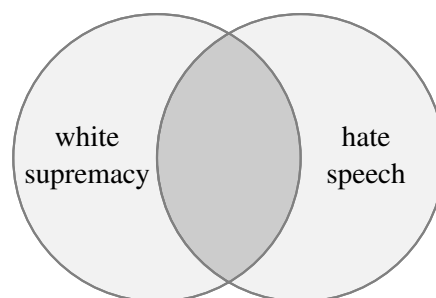

Figure 1: Conceptualization of the relationship between hate speech and white supremacism used in this paper. Much of white supremacist language includes text that would be considered hate speech, i.e. attacks against those with marginalized identities. However, we also aim to capture text that expresses white supremacist ideology without direct hate speech, such as the glorification of Nazism. Finally, some hate speech would not fit as expressing a white supremacist ideology, such as antisemitism within a Black Nationalist context.

inter-annotator agreement on white supremacism in tweets. Hartung et al. (2017) note that individual tweets are difficult to link to extreme right-wing ideologies and instead choose to annotate user tweet histories.

Instead of focusing on individual posts, we turn to *weak supervision*, approaches to quickly and cheaply label large amounts of training data based on rules, knowledge bases or other domain knowledge (Ratner et al., 2017). Weakly supervised learning has been used in NLP for tasks such as cyberbullying detection (Raisi and Huang, 2017), sentiment analysis (Kamila et al., 2022), dialogue systems (Hudeček et al., 2021) and others (Karamanolakis et al., 2021). For training the discriminative white supremacist classifier, we draw on three sources of text data with "natural" (weak) labels: white supremacist domains and organizations, neutral data with similar topics, and anti-racist blogs and organizations.

3.1 White supremacist data

We sample existing text datasets and data archives from white supremacist domains and organizations to build a dataset of texts that likely express white supremacist extremism. Table 1 details information on source datasets.

Sources include sites dedicated to white supremacism, such as Stormfront, Iron March, and the Daily Stormer. When possible, we filter out non-ideological content on these forums using existing topic structures, for example, excluding

<hr>

[2]Quotes in this paper are paraphrased for privacy (Williams et al., 2017)

Data source	Platform	# Posts	Excerpt from example post
Papasavva et al. (2020)	4chan	2,686,267	africans are inferior animals
Stormfront archive	Stormfront	751,980	help the white race
Jokubauskaitė and Peeters (2020)	4chan	578,650	we need to drop the nazism no , we need to do the opposite
Iron March archive	Iron March	179,468	disgusting looking fat ch*nk cuckold
Qian et al. (2018)	Twitter	84,695	keep illegal immigrants out
Patriot Front archive	Discord	39,577	interracial dating i find that appalling
Calderón et al. (2021)	Daily Stormer, Amer. Renaissance	26,099	black - on - white murders it never ends
Pruden et al. (2022)	books, manifestos	17,007	preventing the ongoing islamisation
ElSherief et al. (2021)	Twitter	3,480	desert barbarians will destroy the west

Table 1: Information on white supremacist corpus before filtering and sampling. ***Warning: offensive examples.***

"Computer Talk" and "Opposing Views" forums on Stormfront. We also include tweets from organizations that the Southern Poverty Law Center labels as white supremacist hate groups (Qian et al., 2018; ElSherief et al., 2021). In Papasavva et al.'s (2020) dataset from the 4chan /pol/ "politically incorrect" imageboard, we select posts from users choosing Nazi, Confederate, fascist, and white supremacist flags. We also include 4chan /pol/ posts in "general" threads with fascist and white supremacist topics (Jokubauskaitė and Peeters, 2020). From Pruden et al. (2022), we include white supremacist books and manifestos. We also include leaked chats from Patriot Front, a white supremacist group. Details on these datasets can be found in Appendix A.

With over 230 million words in 4.3 million posts across many domains, this is the largest collection of white supremacist text we are aware of. Contents are from 1968 through 2019, though 76% of posts are from 2017-2019 (see distributions of posts over time in Appendix A).

Outlier filtering and sampling This large dataset from white supremacist domains inevitably contains many posts that are off-topic and non-ideological. To build a weakly supervised classifier, we wish to further filter to highly ideological posts from a variety of domains.

We first remove posts with 10 or fewer words, as these are often non-ideological or require context to be understood (such as "reddit and twitter are cracking down today" or "poor alex, i feel bad").

We then select posts whose highest probability topic from an LDA model (Blei et al., 2003) are ones that are more likely to express white supremacist ideology. LDA with 30 topics separated themes well based on manual inspection.

One of the authors annotated 20 posts from each topic for expressing a tenet of white supremacism, described in Section 2. We selected 6 topics with the highest annotation score for white supremacy, as this gave the best performance on evaluation datasets. These topics related to antisemitism, anti-Black racism, and discussions of European politics and Nazism (details in Appendix B). To balance forum posts with other domains and approximate domain distributions in neutral and anti-racist datasets, we randomly sample 100,000 forum posts. This white supremacist corpus used in experiments contains 118,842 posts and 10.7 million words.

3.2 Neutral data

We also construct a corpus of "neutral" (not white supremacist) data that matches the topics and domains of the white supremacist corpus. To match forum posts, we sample r/politics and r/Europe subreddits. To match tweets, we query the Twitter API by sampling the word distribution in white supremacist tweets after removing derogatory language. For articles, we sample random US news from the News on the Web (NOW) Corpus[3], and use a random Discord dataset to match chat (Fan, 2021). For each of these domains, we sample the same number of posts per year as is present in the white supremacist corpus. If there is not significant time overlap, we sample enough posts to reach a similar word count. This corpus contains 159,019 posts and 8.6 million words.

3.3 Anti-racist data

Hate speech classifiers often overpredict mentions of marginalized identities as hate (Dixon et al.,

---

[3]https://www.corpusdata.org/now_corpus.asp

2017). Assuming our data is biased until proven innocent (Hutchinson et al., 2021), we design for this issue. We hypothesize that texts from anti-racist perspectives may help. Oxford Languages defines anti-racism as movements "opposing racism and promoting racial equality". Anti-racist communications often mention marginalized identities (as do white supremacist texts), but cast them in positive contexts, such as a tweet in our anti-racist dataset that reads, "stand up for #immigrants".

We construct a corpus of anti-racist texts to match the domain and year distribution of the white supremacist corpus. For forum data, we sample comments in subreddits known for anti-racism: r/racism, r/BlackLivesMatter, and r/StopAntiAsianRacism. We include tweets from anti-racist organizations listed by the University of North Carolina Diversity and Inclusion office[4]. To match articles, we scrape Medium blog posts tagged with "anti-racism", "white supremacy", "racism", and "BlackLivesMatter". As with other corpora, data from each of these sources was inspected for its perspective. This anti-racist corpus contains 87,807 posts and 5.6 million words.

4 Classification

Due to the success of BERT-based hate speech models (Mozafari et al., 2019; Samghabadi et al., 2020), we select the parameter-efficient Distil-BERT model (Sanh et al., 2019) to compare data configurations[5]. We use a learning rate of 2×10^{-5}, batch size of 16, and select the epoch with the highest ROC AUC on a 10% development set, up to 5 epochs. Training each model took approximately 8 hours on an NVIDIA RTX A6000 GPU.

We train models on binary white supremacist classification. All posts in the white supremacist corpus, after sampling and filtering, are labeled 'white supremacist'. Posts in neutral and anti-racist corpora are labeled 'not white supremacist'. We also test combining weakly labeled data with manually annotated data from existing datasets (see below) and our own annotation of white supremacist posts in LDA topics. Since there is relatively little manually annotated data, we duplicate it 5 times in these cases, to a size of 57,645 posts.

4.1 Evaluation

Evaluating weakly supervised classifiers on a held-out weakly supervised set may overestimate performance. Classifiers may learn the idiosyncrasies of domains known for white supremacy in contrast to neutral domains (4chan vs. Reddit, e.g.) instead of learning distinctive features of white supremacy. We thus evaluate classifiers on their ability to distinguish posts manually annotated for white supremacy within the same domains, in the following 3 datasets:

Alatawi et al. (2021): 1100 out of 1999 tweets (55.0%) annotated as white supremacist. Like our work, they conceptualize white supremacy as including hate speech against marginalized groups.

Rieger et al. (2021): 366 out of 5141 posts (7.1%) from 4chan, 8chan, and r/the_Donald annotated as white supremacist. This work uses a more restricted definition of white supremacy largely distinct from hate speech. We sample examples labeled as white supremacist or neither white supremacist nor hate speech. Examples only annotated as hate speech are excluded since they may or may not fit our broader conception of white supremacism.

Siegel et al. (2021): 171 out of 9743 tweets (1.8%) annotated as white supremacist. Since they use a more restrictive definition of white supremacy, we sample posts annotated as white supremacist or neither white supremacist nor hate speech.

The proportions of white supremacist posts in these annotated evaluation datasets vary widely, so we report ROC AUC instead of precision, recall, or F1-score, which assume similar class proportions between training and test data (Ma and He, 2013). Precision and recall curves are also available in Figure 5 in Appendix C.

Generalization evaluation To test the ability of classifiers to generalize, we perform a leave-one-out test among annotated datasets. During three runs for each model that uses manually annotated data, we train on two of the annotated datasets and test performance on the third. To test generalization to a completely unseen domain, we use a dataset of quotes from offline white supremacist propaganda, extracted from data collected by the Anti-Defamation League (ADL)[6]. 1655 out of 1798 quotes (92.0%) were annotated by two of the authors as exhibiting white supremacist ideology.

---

[4]https://diversity.unc.edu/anti-racism-resources/
[5]Code for experiments and dataset processing is available at https://github.com/michaelmilleryoder/white_supremacist_lang.

[6]https://www.adl.org/resources/tools-to-track-hate/heat-map

Baselines We evaluate our approaches against the best-performing model from Alatawi et al. (2021), BERT trained on their annotated Twitter dataset for 3 epochs with a learning rate of 2×10^{-5} and batch size of 16. We also compare against Siegel et al. (2021), who first match posts with a dictionary and then filter out false positives with a Naive Bayes classifier. Though Rieger et al. (2021) also present data annotated for white supremacy, they focus on analysis and do not propose a classifier.

HateCheck evaluation for lexical bias To evaluate bias against mentions of marginalized identities, we use the synthetic HateCheck dataset (Röttger et al., 2021). We filter to marginalized racial, ethnic, gender and sexual identities, since white supremacy is a white male perspective interlinked with misogyny and homophobia (Ferber, 2004; Brindle, 2016). We select sentences that include these identity terms in non-hateful contexts: neutral and positive uses; homonyms and reclaimed slurs; and counterspeech of quoted, referenced, and negated hate speech. This sample totals 762 sentences.

5 Results

Table 2 presents performance of single runs on randomly sampled 30% test sets from Alatawi et al. (2021), Rieger et al. (2021), and Siegel et al. (2021). Classifiers trained with both weakly annotated data and a combination of all manually annotated data average the best performance across evaluation datasets. On the Alatawi et al. (2021) dataset, their own classifier performs the best. All models have lower scores on this challenging dataset, which human annotators also struggled to agree on (0.11 Cohen's κ). In generalization performance (Table 3), we find that using weakly annotated data outperforms using only manually annotated data in almost all cases, and that combining weakly and manually annotated data enables classifiers to generalize most effectively.

5.1 Anti-racist corpus

Training with both neutral and anti-racist negative examples improves accuracy on the HateCheck dataset to 69.2 from 60.5 when using a similar number of only neutral negative examples. This supports our hypothesis that incorporating anti-racist texts can mitigate bias against marginalized identity mentions. Adding anti-racist texts slightly decreases performance on the other 4 evaluation datasets, to 82.8 from 84.3 mean ROC AUC.

Model	A	R	S	Mean
S	60.3	61.8	61.3	61.2
A	**74.0**	81.2	89.7	81.6
Annotated	65.3	86.1	92.9	81.4
Weak	71.6	87.8	90.3	83.2
Weak + Ann	70.9	**90.3**	**96.8**	**86.0**

Table 2: ROC AUC scores of models (rows) on test splits of evaluation datasets (columns). A = Alatawi et al. (2021), R = Rieger et al. (2021), S = Siegel et al. (2021).

Model	A	R	S	ADL
S	56.3	61.9	-	57.2
A	-	81.9	83.9	89.1
Annotated	55.2	82.0	84.7	68.5
Weak	**71.0**	87.8	87.3	85.1
Weak + Ann	70.0	**89.8**	**88.9**	**89.2**

Table 3: Generalization performance (ROC AUC). For Annotated and Weak + Annotated models, the first 3 columns report scores on evaluation datasets when trained on data from the other two datasets. The final column reports on the unseen ADL dataset. Scores on datasets used by baseline models for training are not reported since this table focuses on generalization.

6 Conclusion

Ideologies such as white supremacy are difficult to annotate and detect from short texts. We use weakly supervised data from domains known for white supremacist ideology to develop classifiers that outperform and generalize more effectively than prior work. Incorporating texts from an anti-racist perspective mitigates lexical bias.

To apply a white supremacist language classifier to varied domains, our results show the benefit of using such weakly supervised data, especially in combination with a small amount of annotated data. Other methods for combining these data could be explored in future work, such as approaches that use reinforcement learning to select unlabeled data for training (Ye et al., 2020; Pujari et al., 2022). Incorporating social science insights and looking for specific tenets of white supremacist extremism could also lead to better classification. This classifier could be applied to measure the prevalence or spread of white supremacist ideology through online social networks.

Limitations

The presented classifier and dataset are only from English-speaking sources, a major disadvantage in detecting white supremacist content globally. The dataset also is predominantly sourced from data between 2015-2019 and reflects white supremacist extremist responses to current events from that period, including the Black Lives Matter movement. This limits its effectiveness in detecting white supremacist content from other time periods.

Though including anti-racist data helps mitigate bias tested by our sample of the HateCheck dataset, an accuracy of 69.2% shows room for improvement. There is still a risk of overclassifying posts with marginalized identity mentions as white supremacist.

Ethics Statement

There are significant ethical issues to consider in developing text classifiers for ideologies. Since this research has clear social implications, we wish to be explicit about values and author positionality beyond a sense of "objectivity" in selecting research questions (Schlesinger et al., 2017; D'Ignazio and Klein, 2020; Waseem et al., 2021). The authors come from European- and American-dominated university contexts and consider working against racism and white supremacy a priority. Most identify as white and some identify as people of color. This research proceeded with values of racial justice and places those values at the center of assessing knowledge claims (Collins, 1990; Daniels, 2009). Our choice of focusing on white supremacy among other ideologies stems from those values. White supremacist extremism, as well as structural white supremacism, is responsible for substantial harms against those with marginalized identities. This research responds to a need from practitioners for more nuanced classifiers than for broad categories of hate speech or abusive language. We thus choose to pursue this research, though caution that developing classifiers for other ideologies should be done with careful consideration and a clear statement of motivating values.

There are significant risks which we consider, and attempt to mitigate, in such a dataset and classifier. First, there is the risk of misuse of a large corpus of white supremacist data, as has been seen in building and releasing a hate speech "troll bot" from 4chan data[7]. For this reason we build a discriminative, not generative, classifier, and only plan on releasing our dataset through a vetting process instead of publicly.

There are also privacy risks in how such a classifier could be used. Our classifier only identifies language that is likely similar to white supremacist content. The intended use of this classifier is to measure the prevalence of such an ideology on particular platforms or within networks for research purposes, not to label individuals as holding or not holding white supremacist ideologies. Using the classifier for this purpose poses significant risks of misclassification and could increase harmful surveillance tactics. We strongly discourage such a use. Our hope is that our proposed classifier and dataset can increase knowledge about the nature and extent of white supremacist extremist movement online and can inform structural interventions, such as platform policies, not interventions against individuals.

Hate speech classifiers, developed by researchers with similar equity-based values, have been found to contain biases against marginalized groups (Sap et al., 2019; Davidson et al., 2019). We measure and mitigate this bias from the start by incorporating anti-racist data, though caution that this risk still exists.

References

Anti-Defamation League: ADL. 2022. Deplatform Tucker Carlson and the "Great Replacement" Theory.

---

[7]https://www.vice.com/en/article/7k8zwx/ai-trained-on-4chan-becomes-hate-speech-machine

Hind S. Alatawi, Areej M. Alhothali, and Kawthar M. Moria. 2021. Detecting White Supremacist Hate Speech Using Domain Specific Word Embedding with Deep Learning and BERT. *IEEE Access*, 9:106363–106374.

Tawia Ansah. 2021. Violent words: strategies and legal impacts of white supremacist language. *Virginia Journal of Social Policy & the Law*, 28(3):305–340.

Chip Berlet and Stanislav Vysotsky. 2006. Overview of U.S. White Supremacist Groups. *Journal of Political and Military Sociology*, 34(1):11–48.

Prashanth Bhat and Ofra Klein. 2020. Covert Hate Speech: White Nationalists and Dog Whistle Communication on Twitter. In Gwen Bouvier and Judith E. Rosenbaum, editors, *Twitter, the Public Sphere, and the Chaos of Online Deliberation*, pages 151–172. Springer International Publishing, Cham.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the Natural Language Toolkit*. O'Reilly Media, Inc.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Andrew Brindle. 2016. *The language of hate: A corpus linguistic analysis of white supremacist language*. Routledge.

Christopher Brown. 2009. WWW.HATE.COM: White supremacist discourse on the internet and the construction of whiteness ideology. *Howard Journal of Communications*, 20(2):189–208.

Fernando H. Calderón, Namrita Balani, Jherez Taylor, Melvyn Peignon, Yen-Hao Huang, and Yi-Shin Chen. 2021. Linguistic Patterns for Code Word Resilient Hate Speech Identification. *Sensors*, 21(23):7859.

Patricia Hill Collins. 1990. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. Routledge.

Jessie Daniels. 2009. *Cyber racism: White supremacy online and the new attack on civil rights*. Rowman & Littlefield Publishers.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35. Association for Computational Linguistics.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2019. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the Second Workshop on Abusive Language Online (ALW2)*, pages 11–20.

Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism*. Strong Ideas. MIT Press.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2017. Measuring and Mitigating Unintended Bias in Text Classification. In *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*.

Margaret E. Duffy. 2003. Web of hate: A fantasy theme analysis of the rhetorical vision of hate groups online. *Journal of Communication Inquiry*, 27(3):291–312.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jess Fan. 2021. Discord dataset. https://www.kaggle.com/jef1056/discord-data. V5.

Abby L. Ferber, editor. 2004. *Home-grown hate: Gender and organized racism*. Psychology Press.

Abby L. Ferber and Michael Kimmel. 2000. Reading right: the Western tradition in white supremacist discourse. *Sociological Focus*, 33(2):193–213.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, Turin, Italy.

Leo Figea, Lisa Kaati, and Ryan Scrivens. 2016. Measuring online affects in a white supremacy forum. In *IEEE International Conference on Intelligence and Security Informatics: Cybersecurity and Big Data, ISI 2016*, pages 85–90. Institute of Electrical and Electronics Engineers Inc.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Matthias Hartung, Roman Klinger, Franziska Schmidtke, and Lars Vogel. 2017. Identifying Right-Wing Extremism in German Twitter Profiles: a Classification Approach. In *Proceedings of the 22nd International Conference on Applications of Natural Language Processing to Information Systems (NLDB 2017)*. Springer International Publishing.

Vojtěch Hudeček, Ondřej Dušek, and Zhou Yu. 2021. Discovering Dialogue Slots with Weak Supervision. 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 2430–2442, Online. Association for Computational Linguistics.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 560–575, New York, NY, USA. Association for Computing Machinery.

Emilija Jokubauskaitė and Stijn Peeters. 2020. Generally Curious: Thematically Distinct Datasets of General Threads on 4chan/pol/. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 863–867.

Sabyasachi Kamila, Walid Magdy, Sourav Dutta, and MingXue Wang. 2022. AX-MABSA: A Framework for Extremely Weakly Supervised Multi-label Aspect Based Sentiment Analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6136–6147, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Giannis Karamanolakis, Subhabrata Mukherjee, Guoqing Zheng, and Ahmed Hassan Awadallah. 2021. Self-training with weak supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 845–863, Online. Association for Computational Linguistics.

Irene Kwok and Yuzhou Wang. 2013. Locate the Hate: Detecting Tweets against Blacks. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 1621–1622.

Yunqian Ma and Haibo He. 2013. *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2019. A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. In *International Conference on Complex Networks and Their Applications.*, pages 928–940.

Antonis Papasavva, Savvas Zannettou, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2020. Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 885–894.

Barbara Perry and Ryan Scrivens. 2016. White pride worldwide: Constructing global identities online. In *The Globalization of Hate: Internationalizing Hate Crime?* Oxford University Press.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. In *Language Resources and Evaluation*, volume 55, pages 477–523. Springer Science and Business Media.

Meredith L. Pruden, Ayse D. Lokmanoglu, Anne Peterscheck, and Yannick Veilleux-Lepage. 2022. Birds of a Feather: A Comparative Analysis of White Supremacist and Violent Male Supremacist Discourses. In *Right-Wing Extremism in Canada and the United States*, Palgrave Hate Studies, pages 215–254. Palgrave Macmillan.

Rajkumar Pujari, Erik Oveson, Priyanka Kulkarni, and Elnaz Nouri. 2022. Reinforcement guided multitask learning framework for low-resource stereotype detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6703–6712, Dublin, Ireland. Association for Computational Linguistics.

Jing Qian, Mai Elsherief, Elizabeth Belding, and William Yang Wang. 2018. Hierarchical CVAE for Fine-Grained Hate Speech Classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3550–3559.

Elaheh Raisi and Bert Huang. 2017. Cyberbullying detection with weakly supervised machine learning. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017*, pages 409–416. Association for Computing Machinery, Inc.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid Training Data Creation with Weak Supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, pages 269–282.

Diana Rieger, Anna Sophie Kümpel, Maximilian Wich, Toni Kiening, and Georg Groh. 2021. Assessing the Extent and Types of Hate Speech in Fringe Communities: A Case Study of Alt-Right Communities on 8chan, 4chan, and Reddit. *Social Media and Society*, 7(4).

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B 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*, pages 41–58.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas Pykl, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 11–16.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*, pages 2798–2895.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1668–1678.

Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. 2017. Intersectional HCI: Engaging Identity through Gender, Race, and Class. In *CHI '17: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5412–5427.

Alexandra A. Siegel, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 2021. Trumping Hate on Twitter? Online Hate Speech in the 2016 U.S. Election Campaign and its Aftermath. *Quarterly Journal of Political Science*, 16:71–104.

B. Simons and D. B. Skillicorn. 2020. A Bootstrapped Model to Detect Abuse and Intent in White Supremacist Corpora. In *Proceedings - 2020 IEEE International Conference on Intelligence and Security Informatics, ISI 2020*. Institute of Electrical and Electronics Engineers Inc.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying Generalisability Across Abusive Language Detection Datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, pages 940–950, Hong Kong, China. Association for Computational Linguistics.

Kevin C. Thompson. 2001. Watching the Stormfront: White Nationalists and the Building of Community in Cyberspace. *Social Analysis: The International Journal of Anthropology*, 45(1):32–52.

Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied Machine Learning: On the Illusion of Objectivity in NLP. pages 1–8. ArXiv: 2101.11974.

Matthew L. Williams, Pete Burnap, and Luke Sloan. 2017. Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. *Sociology*, 51(6):1149–1168.

Zhiquan Ye, Yuxia Geng, Jiaoyan Chen, Jingmin Chen, Xiaoxiao Xu, SuHang Zheng, Feng Wang, Jun Zhang, and Huajun Chen. 2020. Zero-shot Text Classification via Reinforced Self-training. In *Proceedings*

of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3014–3024, Online. Association for Computational Linguistics.

Michael Miller Yoder and Hana Habib. 2022. Research Needs for Countering Extremist Hate. Technical report, Collaboratory Against Hate.

Michael Miller Yoder, Lynnette Ng, David West Brown, and Kathleen Carley. 2022. How Hate Speech Varies by Target Identity: A Computational Analysis. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 27–39, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A White supremacist corpus details

We sample 9 datasets and data dumps to construct our white supremacist corpus (see Section 3.1). Here we provide details on how each of these data sources was processed and sampled, as well as other details of the corpus.

Papasavva et al. (2020): 4chan /pol/ allows users to select "troll" flags to use instead of the default country flag detected from their IP address. We filter this dataset[8] to posts from users that chose to post with Nazi, White Supremacist, Confederate, or Fascist troll flags. From a qualitative check, samples of posts from users with these flags often expressed white supremacist ideology. We remove posts with duplicate texts, as well as posts that are also found in the 4chan /pol/ dump from Jokubauskaitė and Peeters (2020). Our sample of this dataset contains posts from 2017 through 2019.

Stormfront data archive: Stormfront, a popular white supremacist forum, is no longer active. We sample from an Internet Archive dump of its content taken in 2017[9]. We extract forum text from the HTML files and exclude threads that are not in English and are non-ideological. Specifically, we exclude the following threads: Nederland & Vlaanderen, Srbija, Español y Portugués, Italia, Croatia, South Africa, en Français, Russia, Baltic / Scandinavia, Hungary, Opposing Views Forum, Computer Talk. Our sample of this dataset contains posts from 2001 through 2017.

---

[8]Available at https://zenodo.org/record/360681 0#.Y8lkkBXMKF6, accessed 19 January 2023. This dataset is under a Creative Commons Attribution 4.0 International license.

[9]Available at https://archive.org/details/stormf ront.org_201708, accessed 11 January 2023

Jokubauskaitė and Peeters (2020): We select posts in this dataset of "general" 4chan /pol/ threads[10] that we find to be related to white supremacy and fascism: kraut/pol/, afd, national socialism, fascism, dixie, kraut/pol/, ethnostate, white, chimpout, feminist apocalypse, (((kraut-gate))). This dataset contains posts from 2001 through 2017.

Iron March data archive: Data from Iron March, a now defunct neo-Nazi and white supremacist message board, was obtained through an Internet Archive data dump[11] referenced in Simons and Skillicorn (2020). This dataset contains posts from 2011 through 2017.

Qian et al. (2018): We rehydrate tweet IDs from this dataset, graciously provided by the authors, by the ideology of the tweet author according to the Southern Poverty Law Center. After qualitatively checking sample tweets from each ideology to see how closely they match tenets of white supremacism, we select tweets from the following ideologies: neo-Confederate, neo-Nazi, Ku Klux Klan, racist skinhead, anti-immigration, white nationalist, anti-Semitism, hate music, holocaust identity, Christian Identity. 44.9% of tweets were able to be rehydrated from the original set in September 2022. Our rehydrated tweets ran from 2009 through 2017.

Patriot Front data archive: We select Discord chat posts from servers operated by the white supremacist group, Patriot Front. These chats were leaked by Unicorn Riot[12]. After manual inspection for which threads are most ideological, we select the 'general' channels from 3 servers: Vanguard America-Patriot Front (2017), Front and Center (2018), MI Goy Scouts Official (2018).

Since chat data may contain names, we remove the top 300 US first names from a 1990 list[13].

---

[10]Available at https://zenodo.org/record/360329 2#.Y8lmTxXMKF5, accessed 19 January 2023. This dataset is under a Creative Commons Attribution 4.0 International license.

[11]Available through links at https://www.bellingcat.c om/resources/how-tos/2019/11/06/massive-white-s upremacist-message-board-leak-how-to-access-and -interpret-the-data/, accessed 11 January 2023

[12]https://unicornriot.ninja/2022/patriot-front -fascist-leak-exposes-nationwide-racist-campaig ns/, accessed 11 January 2023

[13]https://namecensus.com/first-names/, accessed 11 January 2023

Calderón et al. (2021): We include articles from two white supremacist news websites, the Daily Stormer and American Renaissance, graciously provided by Calderón et al. (2021). This data contains posts from 2005 through 2017.

Pruden et al. (2022): We include white supremacist books and manifestos collected and provided by Pruden et al. (2022). These are: Enoch Powell's "Rivers of Blood" speech (1968), Jean Raspail's *Camp of the Saints* (1973, English translation), William Pierce's *The Turner Diaries* (1978), David Lane's "White Genocide" manifesto (2012), Anders Breivik manifesto (2011), Renaud Camus' *The Great Replacement* (2012, English translation). These books and manifestos are split into paragraphs (split at newlines) for experiments.

ElSherief et al. (2021): From this dataset of implicit hate speech tweets[14], we select two portions: 1) tweets labeled for "white grievance" by annotators, and 2) when rehydrated, tweets by users identified as holding selected white supremacist ideologies by Qian et al. (2018) (these papers draw on similar datasets). When we rehydrated these tweets in August 2022, we were only able to access 36.8%. Rehydrated tweets spanned from 2009 through 2017.

We lowercase and tokenize all data sources with spaCy 3.1.1 for forum posts and articles, and NLTK's TweetTokenizer (Bird et al., 2009) for tweets and chat data.

Figure 3 shows the time spans of data from different sources in the full corpus, and Figure 4 shows the distribution of posts over time in the dataset. These figures exclude historical data from Pruden et al. (2022) for readability.

B Outlier topic removal

This appendix describes details of removing non-ideological content from our white supremacist corpus. We run LDA over the full white supremacist corpus and decide on 30 topics after manually inspecting topics for coherence. We also tried BERTopic (Grootendorst, 2022), but LDA gave a less skewed distribution of documents per topic.

After a brief initial annotation period, one of the authors annotated 20 instances per topic as white supremacist (coded 1), neutral/undecided (0), or not white supremacist (-1). The criteria was the
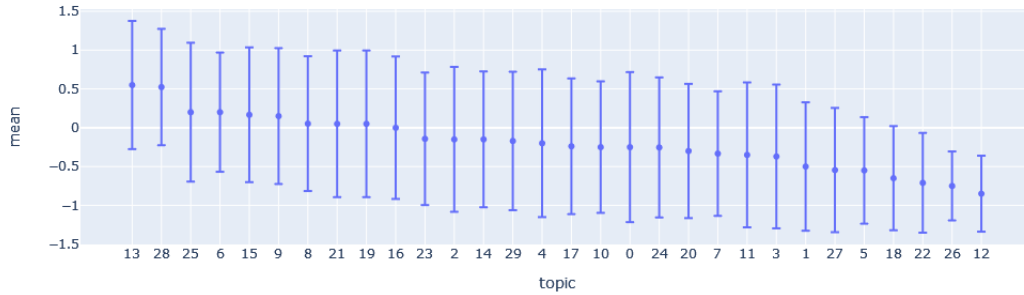
---

[14]Available at https://github.com/SALT-NLP/implici t-hate, accessed 19 January 2023

Figure 2: Mean white supremacism annotations by LDA topic in the white supremacist corpus.

Topic	Top words	Mean ann.
13	jews jewish jew israel kike anti holocaust kikes zionist goyim	0.55
28	white people whites race non black blacks racist hate want	0.52
25	eu russia russian europe france french european turks country sweden	0.20
6	national state people government power nation political socialism society right	0.20
15	war hitler germany german did germans nazi world army nazis	0.17
9	black crime gun kill blacks killed africa rape guns people	0.15

Table 4: LDA topics selected for the white supremacist corpus used in experiments. These are the 6 topics with the highest mean annotation values for white supremacy. **_Warning: offensive and hateful terms._**
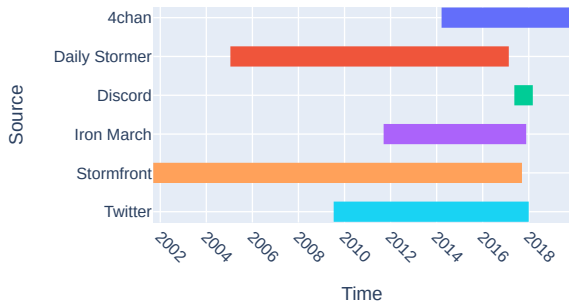

Figure 3: Time spans of data included in full white supremacist corpus, separated by source. Historical data from Pruden et al. (2022) is excluded.
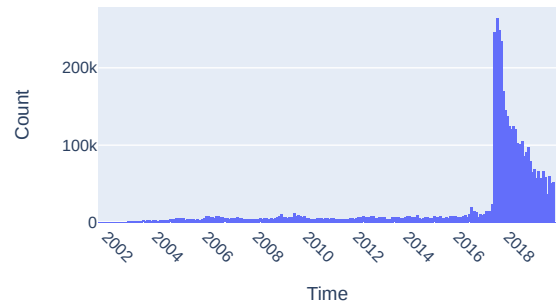

Figure 4: Histogram of post counts in full white supremacist corpus over time, binned monthly. Historical data from Pruden et al. (2022) is excluded.

presence of at least one tenet of white supremacism, described in Section 2. Mean distribution of these annotations over topics are presented in Figure 2.

As can be seen, most topics have mean scores less than 0, i.e., that they contain more posts annotated as neutral or not white supremacist than white supremacist. This matches results from Rieger et al. (2021), who find 24% of posts in a sample from fringe far-right platforms to be hate speech, high compared to other online spaces but certainly not the majority of posts. This motivates outlier removal, and we found that removing outlier topics provided an advantage in classification on the eval-

uation datasets. Assigning posts to the highest-likelihood topic, we find that filtering to posts within the 6 topics with the highest mean annotations for white supremacy provides the best performance. As seen in Figure 2, beyond 6 topics the mean drops to close to a 0 (neutral) rating. These topics related to antisemitism, anti-Black racism, and discussions of European politics and Nazism. Top words for these 6 topics are listed in Table 4.

C   Evaluation datasets

This appendix describes the details of sampling and processing datasets manually annotated for white

supremacy used to evaluate classifiers.

We also present precision and recall curves for our best-performing Weak + Annotated model on evaluation datasets in Figure 5 for decision thresholds every 0.01 between [0, 1]. Class probabilities were calculated from a softmax over the output class logits. There is particular room for improvement on precision for Rieger et al. (2021) and Siegel et al. (2021) datasets.

Alatawi et al. (2021): From the full annotated dataset of tweets from Alatawi et al. (2021)[15], we choose the combined annotator labels for white supremacy as the label of white supremacy or not.

Rieger et al. (2021): This dataset, provided by the authors, contains posts on fringe platforms (4chan /pol/, 8chan /pol/, and r/the_Donald) annotated for many aspects of hate speech, including white supremacist ideology. We sample examples labeled for 'white supremacy/white ethnostate' or 'National Socialist' ideology as examples of white supremacy. For negative examples, we sample posts that are not labeled as white supremacist or as hate speech for negative examples, since their definition of white supremacy is more restrictive Specifically, we sample posts not labeled for 'white supremacy/white ethnostate', 'National Socialist', 'general insult', 'personal insult' or 'violence'. Direct requests for this dataset to the authors.

Siegel et al. (2021): We use training data from Siegel et al. (2021), provided by the authors. From lists of tweets annotated for white nationalism and hate speech, we select those marked as positive for white nationalism and as negative examples, those annotated as neither white nationalism nor hate speech. Requests for this dataset should be directed to the authors.

Alatawi et al. 2021 test set
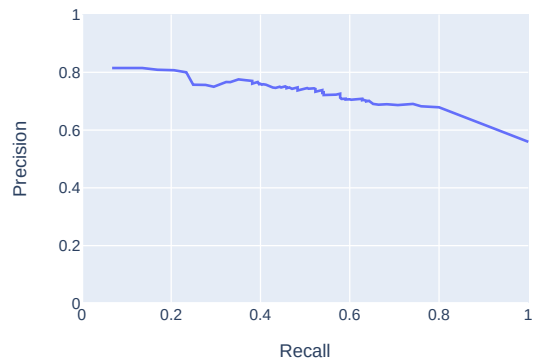
Rieger et al. 2021 test set
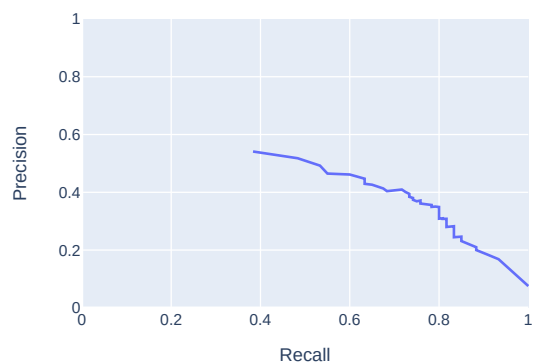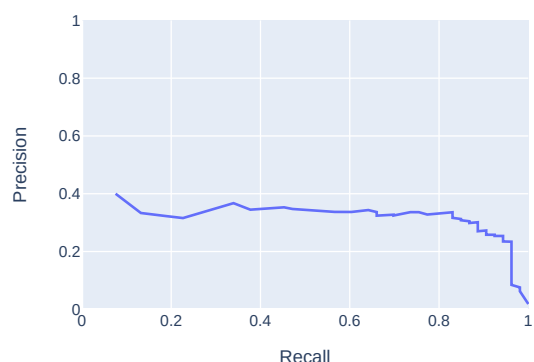

Siegel et al. 2021 test set

Figure 5: Precision and recall curves on test splits of evaluation datasets for the best-performing Weak + Annotated model.

---

[15]Accessed from https://github.com/Hind-Saleh-Alatawi/WhiteSupremacistDataset on 11 January 2023.

A For every submission:

☑ A1. Did you describe the limitations of your work?
Unnumbered "Limitations" section at the end.

☑ A2. Did you discuss any potential risks of your work?
Unnumbered "Ethics Statement" section at the end.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and section 1, Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3 "Weakly Annotated Data" Section 4 "Classification" (a model)

☑ B1. Did you cite the creators of artifacts you used?
Sections 3.1 and 4.1, more details in Appendices A and C

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 3.1 Appendix A

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Ethics Statement

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Appendix A (the Patriot Front Discord chat dataset)

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3.1 Table 1

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.