Addressing Content Selection Bias in Creating Datasets for Hate Speech Detection

Md Mustafizur Rahman
School of Information
The University of Texas at Austin
nahid@utexas.edu

Dinesh Balakrishnan
Department of Computer Science
The University of Texas at Austin
dinesh.k.balakrishnan@utexas.edu

Dhiraj Murthy
School of Journalism and Media
The University of Texas at Austin
Dhiraj.Murthy@austin.utexas.edu

Mucahid Kutlu
Department of Computer Engineering
TOBB Economy and Tech. University
m.kutlu@etu.edu.tr

Matthew Lease
School of Information
The University of Texas at Austin
ml@utexas.edu

Abstract

A key challenge in building a dataset for hate speech detection is that hate speech is relatively rare, meaning that random sampling of tweets to annotate is highly inefficient in finding hate speech. To address this, prior work often only considers tweets matching known "hate words", but restricting the dataset to a pre-defined vocabulary only partially captures the real-world phenomenon we seek to model. Our key insight is that the rarity of hate speech is akin to rarity of relevance in information retrieval (IR). This connection suggests that well-established methodologies for creating IR test collections can be usefully applied to build more inclusive datasets for hate speech. Applying this idea, we have created a new hate speech dataset for Twitter that provides broader coverage of hate, showing a drop in accuracy of existing detection models when tested on these broader forms of hate. This short paper highlights our NeurIPS 2021 Datasets and Benchmarks Track paper [21].

1 Introduction

Online hate speech constitutes a vast and growing problem in social media[13, 17, 15, 9, 25, 5]. For example, Halevy et al. [13] note that the wide variety of content violations and problem scale on Facebook defies manual detection, including the rate of spread and harm such content may cause in the world. Automated detection methods can be used to block content, select and prioritize content for human review, and/or restrict circulation until human review occurs. This need for automated detection has naturally given rise to the creation of labeled datasets for hatespeech (e.g., [20]).

In general, benchmark datasets play a crucial role in machine learning, translating real-world phenomena into a surrogate research environments within which we formulate computational tasks and perform modeling. Training data defines the totality of what models have the opportunity to learn, while testing data provides the means by which we measure empirical success and field progress. Benchmark datasets thus serve to catalyze research and define the world within which our models operate. However, research to improve models is often prioritized over research to improve the data environments in which models operate, even though mismatch between dataset and real-world can lead to significant system failures in practical deployments [23, 18].

35th Conference on Neural Information Processing Systems (NeurIPS 2021): Workshop on Data-Centric AI.

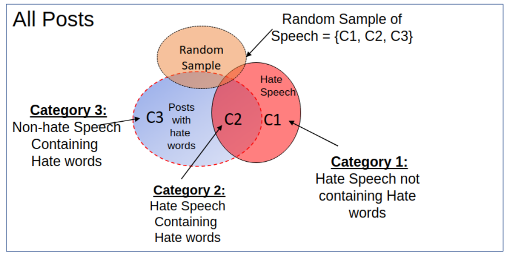

Figure 1: Hate speech coverage based on which social media posts are annotated. Given some list of "hate words" to filter posts, some matching posts are indeed hateful (region C2, *true positives*) while other matching posts are benign (region C3, *false positives*). Region C1 indicates *false negatives*: hate speech missed by the filter and mistakenly excluded. Random sampling correctly overlaps C1+C2 but is highly inefficient in coverage.

Fortunately, many valuable datasets for detecting hate speech already exist [7, 20, 30, 28, 5, 10, 6, 11]. However, each dataset can be seen to embody an underlying design tradeoff (often implicit) in how to balance cost vs. coverage of the phenomenon of hate speech, in all of the many forms of expression in which it manifests. At one extreme, random sampling ensures representative coverage but is highly inefficient (e.g., less than 3% of Twitter posts are hateful [10]). At the other extreme, one can annotate only those tweets matching a pre-defined vocabulary of "hate words" [14, 19] whose presence is strongly correlated with hateful utterances [30, 29, 5, 11, 12]. By restricting a dataset to only those tweets matching a pre-defined vocabulary, a higher percentage of hateful content can be found. However, this sacrifices representative coverage for cost-savings, yielding a biased dataset whose distribution diverges from the real world we seek to model and to apply these models to in practice [16]. If we only look for expressions of hate matching known word lists, we will completely miss in our dataset inclusion of any expressions of hate beyond this prescribed vocabulary. This is akin to traditional AI models that relied entirely on hand-crafted, deterministic rules for classification and failed to generalize beyond their narrow rule sets, providing only partial representation for the real-world phenomenon of interest. We illustrate this in the Venn diagram in **Figure 1**. only annotating posts matching known "hate words" [5, 30, 12, 11] covers only regions C2-C3. The prevalence of hate in such datasets is also limited by the word lists used [26]. In short, "Making effective detection systems for abusive content relies on having the right training datasets" [27].

Our key insight is that the rarity of hate speech are akin to that of relevance in information retrieval (IR) [24]. This suggests that established methods for creating IR *test collections* might also be similarly applied to create better hate speech datasets. To intelligently and efficiently select which content to annotate for hate speech, we applied two IR techniques: *pooling* [24] and *active learning* (AL) [4, 22]. In both cases, we begin with a very large random sample of social media posts to search (i.e., the *document collection*). With pooling, we use existing hate speech datasets and models to train a diverse ensemble of predictive models. We then prioritize posts for annotation by their predicted probability of being hateful, restricting annotation to the resulting *pool* of highly-ranked posts. For nearly 30 years, NIST TREC has applied such pooling techniques with a diverse set of ranking models in order to optimize the coverage vs. cost tradeoff in building test collections for IR, yielding benchmark datasets for fair and robust evaluation of IR systems. Active learning, on the other hand, requires only an initial set of *seed* posts from which a classifier is progressively trained to intelligently select which posts should be labeled next by human annotators. The tight human-AI feedback loop provides greater efficiency and does not require (nor is biased by) existing hate speech datasets. We find that AL can effectively find around 80% of the hateful content at 50% of the cost of pooling. To be clear, such findings are known in the field of IR for building IR test collections [3]; our translational contribution is showing the utility of these techniques for building hate speech datasets.

With regard to broad coverage of hate speech in our dataset[1], pooling yields 14.60% *relative coverage*, far better than the best prior work [10]'s combination of random sampling with keyword-based filtering (10.40%). Regarding efficiency in selecting which tweets to annotate, the *prevalence* of 14.12% of annotated tweets we find to be hateful exceeds a number of prior datasets [5, 9, 12] while crucially also providing the aforementioned greater coverage. Just as precision and recall are balanced in classification, we wish to balance prevalence (efficiency) and coverage (fidelity) in creating a benchmark datasets that is faithful to the phenomenon while remaining affordable to create. We benchmark several recent hate speech detection models [8, 1, 2] and find that the performance of these models drops drastically when tested on these broader forms of hate in our dataset.

[1] github.com/mdmustafizurrahman/An-Information-Retrieval-Approach-to-Building-Datasets-for-Hate-Speech-Detection

Acknowledgements. We thank the many talented Amazon Mechanical Turk workers who contributed to our study and made it possible. This research was supported in part by Wipro, the Knight Foundation, the Micron Foundation, and Good Systems (`https://goodsystems.utexas.edu`), a UT Austin Grand Challenge to develop responsible AI technologies. Our opinions are our own.

References

[1] Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *European Conference on Information Retrieval*. Springer, 141–153.

[2] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 759–760.

[3] Ben Carterette, James Allan, and Ramesh Sitaraman. 2006. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 268–275.

[4] Gordon V. Cormack and Maura R. Grossman. 2015. Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review. *arXiv:1504.06868 [cs]* (April 2015). arXiv: 1504.06868.

[5] 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*, Vol. 11.

[6] 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)*. Association for Computational Linguistics, Brussels, Belgium, 11–20. `https://doi.org/10.18653/v1/W18-5102`

[7] Leon Derczynski. 2021. *Hate speech data.* `https://hatespeechdata.com/`.

[8] 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* (2018).

[9] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 1–30.

[10] Antigoni 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 *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[11] Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, et al. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*. 229–233.

[12] Lara Grimminger and Roman Klinger. 2021. Hate Towards the Political Opponent: A Twitter Corpus Study of the 2020 US Elections on the Basis of Offensive Speech and Stance Detection. *arXiv preprint arXiv:2103.01664* (2021).

[13] Alon Halevy, Cristian Canton Ferrer, Hao Ma, Umut Ozertem, Patrick Pantel, Marzieh Saeidi, Fabrizio Silvestri, and Ves Stoyanov. 2020. Preserving integrity in online social networks. *arXiv preprint arXiv:2009.10311* (2020).

[14] Hatebase. [n.d.]. The world's largest structured repository of regionalized, multilingual hate speech. `https://hatebase.org/`.

[15] David Jurgens, Eshwar Chandrasekharan, and Libby Hemphill. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. *arXiv preprint arXiv:1906.01738* (2019).

[16] Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2018. The Gab Hate Corpus: A collection of 27k posts annotated for hate speech. (2018).

[17] 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 (2019), e0221152.

[18] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems: Datasets and Benchmarks Track*.

[19] NoSwearing. [n.d.]. List of Swear Words, Bad Words, & Curse Words - Starting With A. `https://www.noswearing.com/`.

[20] 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 (2021), 477–523.

[21] Md Mustafizur Rahman, Dinesh Balakrishnan, Dhiraj Murthy, Mucahid Kutlu, and Matthew Lease. 2021. An Information Retrieval Approach to Building Datasets for Hate Speech Detection. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS): Datasets and Benchmarks Track*.

[22] Md Mustafizur Rahman, Mucahid Kutlu, Tamer Elsayed, and Matthew Lease. 2020. Efficient Test Collection Construction via Active Learning. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval* (Virtual Event, Norway) *(ICTIR '20)*. Association for Computing Machinery, New York, NY, USA, 177–184. `https://doi.org/10.1145/3409256.3409837`

[23] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[24] Mark Sanderson. 2010. Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval* 4, 4 (2010), 247–375.

[25] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International workshop on natural language processing for social media*. 1–10.

[26] Sara Sood, Judd Antin, and Elizabeth Churchill. 2012. Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1481–1490.

[27] Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS one* 15, 12 (2020), e0243300.

[28] 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*. 138–142.

[29] Zeerak Waseem. 2016. *Automatic hate speech detection*. Ph.D. Dissertation. Master's thesis, University of Copenhagen.

[30] 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*. Association for Computational Linguistics, San Diego, California, 88–93. `https://doi.org/10.18653/v1/N16-2013`