An Interactive Exploratory Tool for the Task of Hate Speech Detection

Angelina McMillan-Major[1,3] and **Amandalynne Paullada**[2] and **Yacine Jernite**[3]
Department of Linguistics, University of Washington, Seattle, USA[1]
Department of Biomedical Informatics & Medical Education, University of Washington, Seattle, USA[2]
Hugging Face[3]
aymm@uw.edu, paullada@uw.edu, yacine@huggingface.co

Abstract

With the growth of Automatic Content Moderation (ACM) on widely used social media platforms, transparency into the design of moderation technology and policy is necessary for online communities to advocate for themselves when harms occur. In this work, we describe a suite of interactive modules to support the exploration of various aspects of this technology, and particularly of those components that rely on English models and datasets for hate speech detection, a subtask within ACM. We intend for this demo to support the various stakeholders of ACM in investigating the definitions and decisions that underpin current technologies such that those with technical knowledge and those with contextual knowledge may both better understand existing systems.

1 Introduction

The field of natural language processing (NLP) is organized into *tasks*, definitions of which minimally include the combination of a modeling paradigm and benchmark datasets (Vu et al. (2020); Reuver et al. (2021); Schlangen (2021); see also BIG-bench[1]). This organization, however, is not necessarily apparent to those outside of NLP research. Making these established tasks outwardly visible is one step towards the recent push for accessible documentation of NLP (Bender and Friedman, 2018; Holland et al., 2018; Mitchell et al., 2019; Arnold et al., 2019; McMillan-Major et al., 2021; Gebru et al., 2021) and promoting the importance of careful data treatment (Paullada et al., 2021; Sambasivan et al., 2021b).

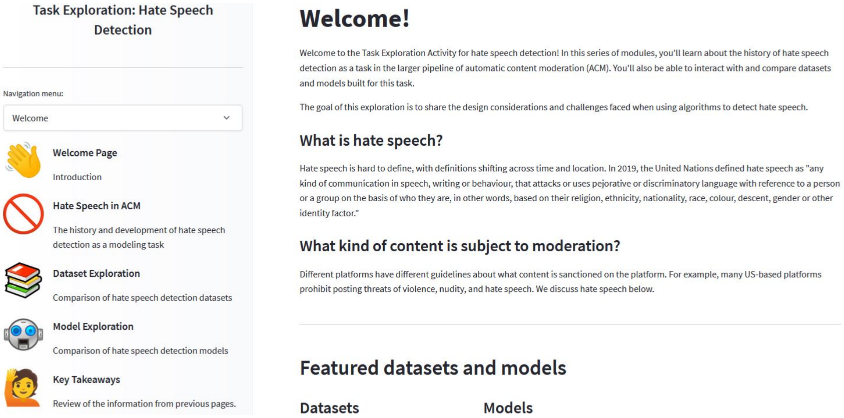
One task that has attracted sustained interest in NLP is the problem of content moderation. While many manual and hybrid paradigms for content moderation exist (Pershan, 2020), several major platforms have invested heavily in automated methods that they see as necessary to support scaling up moderation to address their colossal content loads (Gillespie, 2020). Automatic Content Moderation (ACM) includes strategies that range from keyword- or regular expression-based approaches, to hash-based content recognition, to data-driven machine learning models. These approaches employ different families of algorithms, resulting in various downstream effects and necessitating documentation and algorithmic accountability processes that address the needs of a variety of stakeholders.

Synchronizing research around consistent modeling paradigms and benchmark datasets is an ongoing problem for ACM (Fortuna et al., 2020; Madukwe et al., 2020), with experts calling for more grounding in related areas in the social sciences, communication studies and psychology (Vidgen and Derczynski, 2020; Kiritchenko et al., 2021). Without this grounding and without consideration for the contexts into which ACM is integrated, the technology intended to prevent harms ends up magnifying them, especially for vulnerable communities (Dias Oliva et al., 2021).

The present paper proposes an interactive tool aimed at allowing a diverse audience to explore examples of NLP data and models used in data-driven ACM, focusing on the subtask of hate speech detection. Our tool outlines various aspects of the social and technical considerations for ACM, provides an overview of the data and modeling landscape for hate speech detection, and enables comparison of different resources and approaches to the task. Our goal is to understand the role of multidisciplinary education and documentation in promoting algorithmic transparency and contestability (Vaccaro et al., 2019). We provide a brief overview of ACM as well as the interactions between its many stakeholders (§2) and describe related work in dataset and model exploration (§3). We then present our demo (§4), highlighting its constituent sections and describing our rationale for each. We conclude with a summary of limitations and future work (§5).

[1] https://github.com/google/BIG-bench/

Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pages 11 - 20
July 15, 2022 ©2022 Association for Computational Linguistics



Task Exploration: Hate Speech Detection
Navigation menu:
Welcome
Welcome Page
Introduction
Hate Speech in ACM
The history and development of hate speech detection as a modeling task
Dataset Exploration
Comparison of hate speech detection datasets
Model Exploration
Comparison of hate speech detection models
Key Takeaways
Review of the information from previous pages.

Welcome!

Welcome to the Task Exploration Activity for hate speech detection! In this series of modules, you'll learn about the history of hate speech detection as a task in the larger pipeline of automatic content moderation (ACM). You'll also be able to interact with and compare datasets and models built for this task.

The goal of this exploration is to share the design considerations and challenges faced when using algorithms to detect hate speech.

What is hate speech?

Hate speech is hard to define, with definitions shifting across time and location. In 2019, the United Nations defined hate speech as "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor."

What kind of content is subject to moderation?

Different platforms have different guidelines about what content is sanctioned on the platform. For example, many US-based platforms prohibit posting threats of violence, nudity, and hate speech. We discuss hate speech below.

Featured datasets and models

Datasets **Models**

Figure 1: Introduction page to the demo

2 Background: Content Moderation

Content moderation is the process by which online platforms manage which kinds of content, in the form of images, video, or text, that users are allowed to share. Policies for content moderation, which vary across platforms, are often guided by a combination of legal, commercial, and social pressures. Broadly, these policies tend to prohibit explicit sexual content, graphic depictions of violence, hate speech[2], and harassment or trolling between platform users (Gillespie, 2018). Platforms take a variety of actions to moderate content, including removal of the offending content, reducing the visibility of the content, adding a flag or warning, and/or suspending accounts that violate content guidelines. Moderation decisions can, however, lead to undesired reactions. For example, removing conspiracy theory content tends to reinforce conspiracy theory claims, and ousting hateful groups from larger platforms can result in these groups flocking to smaller platforms with fewer resources for moderation (Pershan, 2020).

Conflicts in moderation decisions often arise due to the size and diversity of a platform's community members and a divergence in priorities between community members and platform managers. A report from the Brennan Center for Justice found that 'double standards' pervade in content moderation actions, and that inconsistently applied content policies overwhelmingly silence marginalized voices (Díaz and Hecht-Felella). For example, Facebook erroneously labeled hashtags referencing Al-Aqsa, a mosque in a predominately Palestinian neighborhood of Jerusalem, as pertaining to a terrorist organization, and was also found to censor derogatory speech against white people more frequently than slurs against Black, Jewish, and transgender people (Eidelman et al., 2021). To address the often stark gap between model performance on intrinsic metrics and performance in real-world, user-facing scenarios for toxic content classifiers, Gordon et al. (2021) propose an evaluation paradigm that takes into account inter-annotator disagreements on training data.

Even when moderation rules are applied consistently, they may result in over-moderating communities that use terms that are deemed 'explicit' outside the community but are acceptable to the community members themselves, as often happens for LGBTQ communities online (Dias Oliva et al., 2021). These kinds of harms show that content moderation algorithms must be developed with transparency, care for the context in which the algorithms will be integrated, and mechanisms for the community to contest moderation decisions. One approach to consulting diverse perspectives on 'toxic' content relies on *jury learning*, as in a model proposed by Gordon et al. (2022).

In addition to calling for more inclusion by various stakeholders in decision-making processes for each platform, Pershan (2020) advocates for the development of regional policies that consider the moderation styles of smaller platforms as well as larger ones. Regional policies are especially important as the large platforms, primarily located in the US, are ported outside the US with moderation policies that are ill-equipped to support local communities appropriately, for example in India where hate speech may also occur on the basis of caste (Sambasivan et al., 2021a).

Approaches to content moderation commonly involve a hybrid strategy that uses reports from users

[2] We define *hate speech* in §4.

and algorithmic systems to identify content that may violate platform guidelines, and then relies on human review to determine a course of action (i.e., retain, obscure, or remove the content). This process exposes human moderators to high volumes of violent and hateful content (Roberts, 2014), motivating a push for enhanced automatic methods to alleviate the burden on human moderators. Automated content moderation can rely on analyses of the content itself using NLP or computer vision (CV), features of user dynamics, and hashing to match instances of pre-identified forbidden content. Within the realm of text-based ACM, approaches vary from wordlist-based approaches to data-driven models. When platforms opt not to build their own systems, Perspective API[3] is commonly used to flag various kinds of content for moderation.

Forbidding hate speech on online platforms is seen as a way to prevent the proliferation of hateful discourse from leading to hate-driven violence offline[4]. Common datasets used for training and evaluating hate speech detectors can be found at `https://hatespeechdata.com/`. We refer readers to Kiritchenko et al. (2021) for a comprehensive overview of definitions, resources, and ethical challenges incurred in the development of hate speech detection technologies.

3 Related Work: Interactive Dataset and Model Exploration

A variety of methods and tools that enable dataset users to explore and familiarize themselves with the contents of the datasets have been proposed. For example, Know Your Data[5], provided by Google's PAIR research group, aims to provide users with views of datasets that surface errors or issues with particular instances, systematic gaps in representation, or problematic content that requires human judgment to assess. This tool thus far has focused on image datasets. The Dataset Cartography method, proposed by Swayamdipta et al. (2020), uses model training dynamics to create maps of dataset instances organized by difficulty or ambiguity, which can surface problematic instances. Recently, Xiao et al. (2022) released a tool for comparing datasets aimed at enabling dataset users to understand potential sources of bias in the data. While much previous work has focused on exploratory tools for dataset *users*, our tool is meant to cater to an audience who will not necessarily be training machine learning models, but constitute a variety of impacted or interested stakeholders.

Wright et al. (2021) tackle the problem of interrogating a toxicity detection model using a tool they call RECAST. They fine-tune a BERT-based Transformer model on the Jigsaw Kaggle dataset of toxic comments from Wikipedia and provide an online text-editing application that visually highlights words that the models detects as toxic, suggesting alternate phrases that may be less toxic using both word embeddings and language modeling predictions. They evaluate the tool using a text-editing task, presenting user study participants with comments drawn from both the Kaggle dataset and Twitter threads, and show that the users in their study are learning about the model behavior by editing toxic comments to be less toxic according to the model prediction scores.

4 Demo Development and Structure

We aim to make the exploration tool as accessible and useful as possible to the many stakeholders involved in ACM. Particularly in light of the closed nature of many contemporary content moderation pipelines that impact people who use social media, our demo familiarizes these stakeholders with the general framework of how such systems might work behind the scenes. In order to conceptualize the breadth of uses that ACM stakeholders may have for such an exploratory tool, we considered the stakeholders and their goals detailed in Pershan (2020) using the framework developed by Suresh et al. (2021). Rather than identifying stakeholders based on their roles, they propose mapping stakeholders based on the type of knowledge they hold and the context of that knowledge, such as technical, domain, and contextual knowledge.

In mapping out our envisioned stakeholders, we tried to consider how they might use the tool towards their goals. Policymakers, journalists and impacted communities may use the demo to understand where and how things go wrong in hate speech detection in order to advocate for changes to platform policies. Domain experts may use the tool to understand where their work is used in a pipeline, such as in label definitions, and envision potential locations in the pipeline where additional domain information could be useful. Students and current developers may use the tool to reflect upon

[3] `https://perspectiveapi.com/`

[4] Discord Off-Platform Behavior Update

[5] `https://knowyourdata.withgoogle.com/`

their own design decisions in light of the historical and sociotechnical framing we provide for ACM and consider new possibilities for research development. Finally, we imagine that our demo may generally provide common ground for these and other stakeholders in order to facilitate more productive discussions on how to develop ACM technologies.

Additionally, in order to more fully understand the perspectives of stakeholders outside of the academic context, we discussed our demo and the state of the field of hate speech detection with several experts in the field, particularly those with experience deploying models in the industry context and working with non-technical stakeholders. Following these discussions, we built the interactive, openly available demo using Streamlit[6], the first page of which is shown in Fig. 1. We provide screenshots of the other modules in Appendix A.

S1. Welcome and Introduction

The introduction to the demo is intended to provide common ground for the various stakeholders with key terms and the kinds of data that are subject to moderation. The key terms include *hate speech* and *content moderation*, for which we provide the following definitions to help build a shared understanding given the broad audience we identified:

Hate speech Any kind of communication in speech, writing, or behaviour that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor (United Nations, 2019).

Content moderation A collection of interventions used by online platforms to partially obscure, or remove entirely from user-facing view, content that is objectionable based on the company's values or community guidelines.

Additionally, we provide a list of the datasets and models that we feature in the tool along with links to further documentation for each resource.

S2. Context of ACM

To contextualize automatic hate speech detection tools, we describe of the kinds of content that moderation is intended to target and how automatic methods are used to support manual approaches to content moderation, as discussed in §2 and §3. We also illustrate the ongoing challenges in hate speech detection with links to platforms' content guidelines and press releases in addition to critical works in response to content moderation.

S3. Hate Speech Dataset Exploration

Meaningfully exploring datasets composed of up to hundreds of thousands of instances constitutes a signifcant difficulty. To address this challenge, we rely on hierarchical clustering to group similar examples at different levels of granularity, using SentenceBERT (Reimers and Gurevych, 2019) embeddings of the example text to evaluate closeness. For each cluster (including the top-level one corresponding to the full dataset), the text of a selection of examplars for that cluster may be viewed along with their labels, as well as the distribution of labels within the entire cluster. This allows users of our system to zoom in on specific regions, and gain insights into what sorts of examples are represented in a dataset and how different topics are labeled. Comparison across datasets also illustrates the different assumptions that are made at the time of dataset creation even within the same established task. For this demo, we pre-selected datasets constructed for hate speech detection in English. These include the FRENK Dataset of Socially Unacceptable Discourse in English (Ljubešić et al., 2019), the Measuring Hate Speech dataset (Kennedy et al., 2020), and the Twitter Sentiment Analysis dataset (Sharma, 2019).

S4. Hate Speech Model Exploration

In the model exploration section, we provide two ways of probing models. The first allows viewers to submit one or more test inputs to a single model. The results are then shown such that viewers may select a label and order the output scores for the inputs based on that label. The second module compares the same input sentence with two different models. The module then returns the label and score given by each model as well as a confidence graphs for each model. The confidence graphs show the model's accuracy, errors, and scores over hateful and non-hateful instances as well as out-of-domain content from the Hateful Memes and Open Subtitles datasets (Kiela et al., 2020; Tiedemann, 2016). These graphs provide context for an individual instance of model behavior, showing whether the model is likely to be more or less confident when labeling an instance, regardless of the model's overall accuracy. Again,

[6] https://streamlit.io/

for the demo we pre-selected models including a RoBERTa model trained on the FRENK dataset (Ljubešić et al., 2019), a RoBERTa model trained on the TweetEval benchmark (Barbieri et al., 2020), and a DeHateBERT model trained on Twitter and StormFront data (Aluru et al., 2021).

To support viewers in probing the models, the demo also supplies test case instances from Röttger et al. (2021)'s HateCheck tests. The tests use templates to generate examples of text that are likely to mislabeled by hate speech detection models. The templates are organized into classes based on the linguistic properties of the sentence, such as whether it contains negation or counter speech, and have placeholders for social groups that are frequently targeted by hate speech. The counter speech template class, for example, contains sentences that directly reference or quote hate speech, as in "Statements like '[IDENTITY] are scum' are deeply hurtful." (examples from Röttger et al. (2021), p.45). These kinds of sentences are often labeled as hateful by automatic systems due to the direct reference, but the sentence as a whole provides context that shows that the intent is not hate speech. With 29 different groups of examples over 10 different classes in addition to counter speech, the HateCheck tests encourage users to think of the variety of forms that hate speech can take and how they may be misclassified by automatic systems.

While the single model exploration module is similar in function to the RECAST tool (Wright et al., 2021), we note significant differences in the imagined use cases of ours and the RECAST tool. Wright et al. emphasize RECAST's use in real time as a comment-editing tool. Our tool on the other hand is not intended for integrated use, but rather as a self-directed learning tool. While stakeholders could compare several edits of the same comment using our tool, stakeholders are not limited to this method of exploration. We instead encourage stakeholders to consider comparisons, between inputs and between models, as a way to surface expected and unexpected model behavior.

S5. Demo Feedback Questionnaire

To end the demo, we ask the user for feedback on their role and experience with the modules. The questions focus on what the user learned from the modules about the sociotechnical aspects of ACM and the resources for hate speech detection. In particular, we are interested in seeing how the modules were more or less informative for different stakeholder groups. See Appendix B for the specific questions asked.

5 Limitations and Future Work

While our tool is aimed at promoting a shared vocabulary and common ground between (1) those who build and design hate speech detection datasets and models, (2) those who are on the receiving end of moderation decisions on social media platforms, and (3) researchers and journalists who are interested in understanding some of the mechanics of automated content moderation, the tool is not designed to be a platform for facilitating connection and engagement *between* these groups. However, the tool can serve as a foundation for such discussions and could be integrated into a larger system designed for engagement.

We plan to update the demo based on feedback from the questionnaire. Once the demo has been finalized, user studies aimed at gathering perspectives from a broader set of stakeholders, including those we did not consider in our initial design process such as content moderation workers, would help to outline how different stakeholders actually use the tool and evaluate the effectiveness of the tool with respect to the participants' use cases and contexts. Following these studies, future versions of the tool could expand to consider more issues within content moderation beyond hate speech detection or be designed to provide context for other kinds of NLP tasks. While this current demo is focused on English resources, future versions could also include resources and contexts for other languages as well as more complex configurations of datasets and models beyond binary labeling schemas.

We began this work with the intention to help provide clarity into the organization of the field of NLP into various tasks. While this demo has focused on the task of ACM, we would expect that similar demos could be developed to contextualize other well-known tasks in NLP such as machine translation, information retrieval, and automatic speech recognition.

Acknowledgements

Thank you to Zeerak Talat, Dia Kayyali, Bertie Vidgen, and the Perspective API Team for their insightful comments during the development of the demo, and to the anonymous reviewers for their very thoughtful and helpful suggestions for improving this publication.

A.P. is supported by the National Institutes of Health, National Library of Medicine (NLM) Biomedical and Health Informatics Training Program at the University of Washington (Grant Nr. T15LM007442). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2021. A deep dive into multilingual hate speech classification. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*, pages 423–439, Cham. Springer International Publishing.

Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, Darrell Reimer, John Richards, Jason Tsay, and Kush R. Varshney. 2019. Factsheets: Increasing trust in ai services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6:1–6:13.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & Culture*, 25(2):700–732.

Ángel Díaz and Laura Hecht-Felella. Double standards in social media content moderation. *Brennan Center for Justice at New York University School of Law*.

Vera Eidelman, Adeline Lee, and Fikayo Walter-Johnson. 2021. Time and again, social media giants get content moderation wrong: Silencing speech about al-aqsa mosque is just the latest example. *American Civil Liberties Union*.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Tarleton Gillespie. 2020. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards.

Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. The frenk datasets of socially unacceptable discourse in slovene and english. In *Text, Speech, and Dialogue*, pages 103–114, Cham. Springer International Publishing.

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161, Online. Association for Computational Linguistics.

Angelina McMillan-Major, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. 2021. Reusable templates and guides for documenting datasets and models for natural language processing and generation: A case study of the HuggingFace and GEM data and model cards. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 121–135, Online. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Claire Pershan. 2020. Moderating our (dis)content: Renewing the regulatory approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, page 113, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Myrthe Reuver, Antske Fokkens, and Suzan Verberne. 2021. No NLP task should be an island: Multi-disciplinarity for diversity in news recommender systems. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 45–55, Online. Association for Computational Linguistics.

Sarah T Roberts. 2014. *Behind the screen: The hidden digital labor of commercial content moderation*. Ph.D. thesis.

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.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021a. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 315–328, New York, NY, USA. Association for Computing Machinery.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021b. *"Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI*. Association for Computing Machinery, New York, NY, USA.

David Schlangen. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics.

Roshan Sharma. 2019. Twitter sentiment analysis.

Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. *Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and Their Needs*. Association for Computing Machinery, New York, NY, USA.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3518–3522, Portorož, Slovenia. European Language Resources Association (ELRA).

United Nations. 2019. United nations strategy and plan of action on hate speech.

Kristen Vaccaro, Karrie Karahalios, Deirdre K Mulligan, Daniel Kluttz, and Tad Hirsch. 2019. Contestability in algorithmic systems. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, pages 523–527.

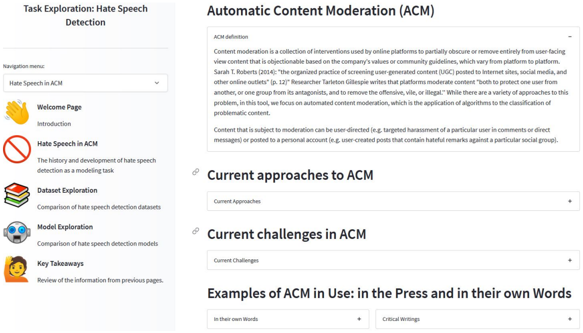
Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS one*, 15(12):e0243300.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926, Online. Association for Computational Linguistics.

Austin P Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Duen Horng Chau, and Diyi Yang. 2021. Recast: Enabling user recourse and interpretability of toxicity detection models with interactive visualization. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–26.

Yang Xiao, Jinlan Fu, Weizhe Yuan, Vijay Viswanathan, Zhoumianze Liu, Yixin Liu, Graham Neubig, and Pengfei Liu. 2022. Datalab: A platform for data analysis and intervention. *arXiv preprint arXiv:2202.12875*.

A Demo Screenshots



Task Exploration: Hate Speech Detection

Navigation menu:

Hate Speech in ACM

Welcome Page
Introduction
Hate Speech in ACM
The history and development of hate speech detection as a modeling task
Dataset Exploration
Comparison of hate speech detection datasets
Model Exploration
Comparison of hate speech detection models
Key Takeaways
Review of the information from previous pages.

Automatic Content Moderation (ACM)

ACM definition

Content moderation is a collection of interventions used by online platforms to partially obscure or remove entirely from user-facing view content that is objectionable based on the company's values or community guidelines, which vary from platform to platform. Sarah T. Roberts (2014): "the organized practice of screening user-generated content (UGC) posted to Internet sites, social media, and other online outlets" (p. 12)" Researcher Tarleton Gillespie writes that platforms moderate content "both to protect one user from another, or one group from its antagonists, and to remove the offensive, vile, or illegal.'' While there are a variety of approaches to this problem, in this tool, we focus on automated content moderation, which is the application of algorithms to the classification of problematic content.

Content that is subject to moderation can be user-directed (e.g. targeted harassment of a particular user in comments or direct messages) or posted to a personal account (e.g. user-created posts that contain hateful remarks against a particular social group).

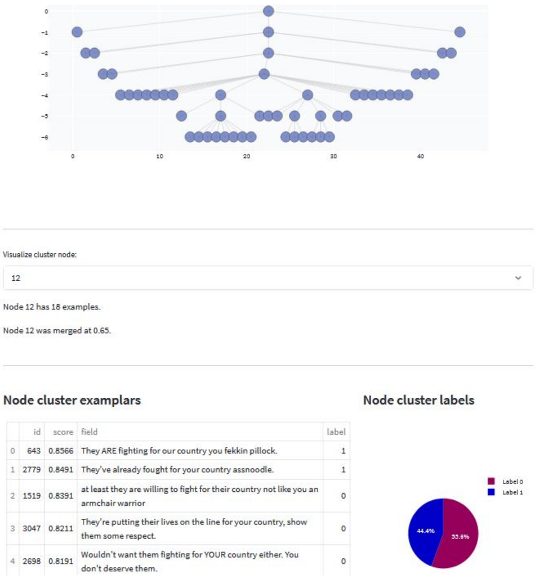
Current approaches to ACM

Current challenges in ACM

Examples of ACM in Use: in the Press and in their own Words

Figure 2: Context of ACM Module

Figure 2 shows the Context of Automatic Content Moderation module (Section 4). By introducing the demo users to some of the relevant context outlined in Section 2 and to selected writings both by content platforms and independent writers on their approach to (automatic) content moderation, we aim to help them better understand the information presented in the following sections.



Visualize cluster node:

12

Node 12 has 18 examples.

Node 12 was merged at 0.65.

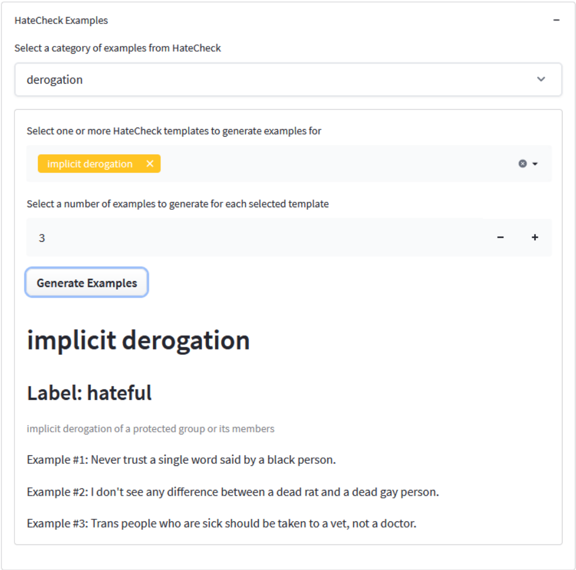
Node cluster examplars

	id	score	field	label
0	643	0.8566	They ARE fighting for our country you fekkin pillock.	1
1	2779	0.8491	They've already fought for your country assnoodle.	1
2	1519	0.8391	at least they are willing to fight for their country not like you an armchair warrior	0
3	3047	0.8211	They're putting their lives on the line for your country, show them some respect.	0
4	2698	0.8191	Wouldn't want them fighting for YOUR country either. You don't deserve them.	0

Node cluster labels

Figure 3: Dataset Exploration Module

Figure 3 provide a screenshot of the Dataset Exploration Section (4). The top half presents a graphical representation of the dataset hierarchical clustering, summary information about a cluster is provided in a tooltip when the user hovers over the corresponding node. The user can then select a specific cluster for which they want to see more information, and the app shows a selected numbers of exemplars (examples that are closest to the cluster centroid) along with the distribution of labels in the cluster.


HateCheck Examples

Select a category of examples from HateCheck

derogation

Select one or more HateCheck templates to generate examples for

implicit derogation ✕

Select a number of examples to generate for each selected template

3

Generate Examples

implicit derogation

Label: hateful

implicit derogation of a protected group or its members

Example #1: Never trust a single word said by a black person.

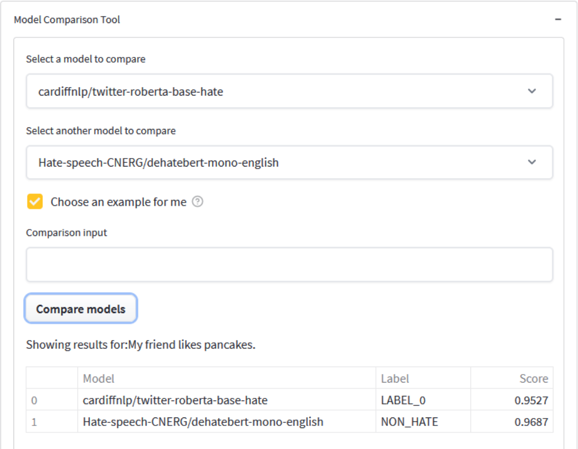
Example #2: I don't see any difference between a dead rat and a dead gay person.

Example #3: Trans people who are sick should be taken to a vet, not a doctor.

Figure 4: Examples using the HateCheck templates

Figures 4, 5, and 6 correspond to the Model Exploration Section (4).

The first module in this Section (Figure 4) allows the user to generate text examples from Röttger et al. (2021)'s HateCheck tests. These tests are designed to examine the models' behaviors on cases that are expected to be difficult for Automatic Content Moderation system and allow users to explore their likely failure cases.


Model Comparison Tool

Select a model to compare

cardiffnlp/twitter-roberta-base-hate

Select another model to compare

Hate-speech-CNERG/dehatebert-mono-english

☑ Choose an example for me

Comparison input

Compare models

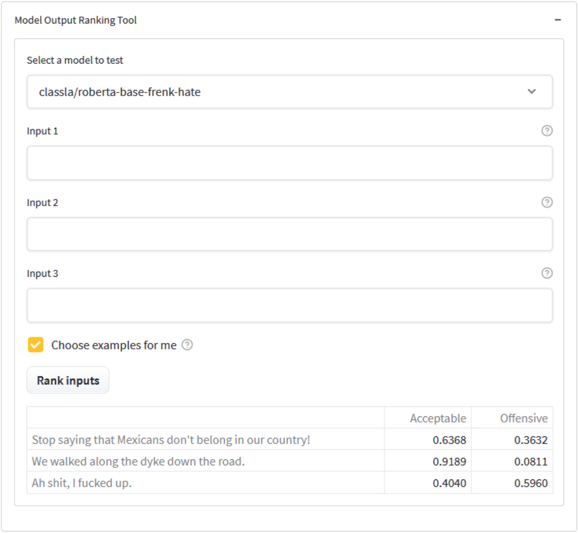
Showing results for:My friend likes pancakes.

	Model	Label	Score
0	cardiffnlp/twitter-roberta-base-hate	LABEL_0	0.9527
1	Hate-speech-CNERG/dehatebert-mono-english	NON_HATE	0.9687

Figure 5: The model comparison section of the model exploration module

Figure 5 presents the model comparison module. Models trained on different datasets might behave differently on similar examples. Being able to test them side by side should allow users to assess their fitness for specific use cases.

Figure 6 presents the example ranking module. Whereas the model comparison module helps users compare model behaviors on similar examples, this one allows them to view a given models' predictions side by side for a set of selected examples, to allow them to explore for example the effect of small variations in the text or the behavior of the model on different categories of tests featured in the HateCheck module.


Model Output Ranking Tool

Select a model to test

classla/roberta-base-frenk-hate

Input 1

Input 2

Input 3

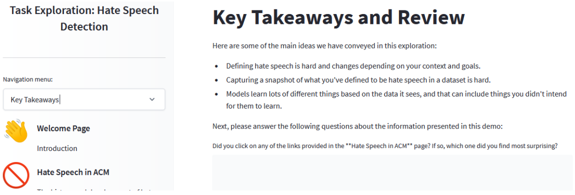
☑ Choose examples for me

Rank inputs

	Acceptable	Offensive
Stop saying that Mexicans don't belong in our country!	0.6368	0.3632
We walked along the dyke down the road.	0.9189	0.0811
Ah shit, I fucked up.	0.4040	0.5960

Figure 6: The model ranking section of the model exploration module

B Feedback Questions


Task Exploration: Hate Speech Detection

Navigation menu:

Key Takeaways

Welcome Page

Introduction

Hate Speech in ACM

Key Takeaways and Review

Here are some of the main ideas we have conveyed in this exploration:

- Defining hate speech is hard and changes depending on your context and goals.
- Capturing a snapshot of what you've defined to be hate speech in a dataset is hard.
- Models learn lots of different things based on the data it sees, and that can include things you didn't intend for them to learn.

Next, please answer the following questions about the information presented in this demo:

Did you click on any of the links provided in the **Hate Speech in ACM** page? If so, which one did you find most surprising?

Figure 7: The key takeaways and feedback module

Figure 7 presents the concluding Section (4), which summarizes some key points presented in the demo and asks users to answer a feedback questionnaire, which includes questions such as:

- How would you describe your role?
- Why are you interested in content moderation?
- Which modules did you use the most?
- Which module did you find most informative?
- Which application were you most interested in learning more about?

- What surprised you most about the datasets?
- Which models are you most concerned about as a user?
- Do you have any comments or suggestions?