A Novel Methodology for Developing Automatic Harassment Classifiers for Twitter

Ishaan Arora
Columbia University
ia2419@columbia.edu

Julia Guo
Columbia University
jzg2110@columbia.edu

Susan E. McGregor
Columbia University
sem2196@columbia.edu

Sarah Ita Levitan
Hunter College (CUNY)
sarah.levitan@hunter.cuny.edu

Julia Hirschberg
Columbia University
julia@cs.columbia.edu

Abstract

Most efforts at identifying abusive speech online rely on public corpora that have been scraped from websites using keyword-based queries or released by site or platform owners for research purposes. These are typically labeled by crowd-sourced annotators – not the targets of the abuse themselves. While this method of data collection supports fast development of machine learning classifiers, the models built on them often fail in the context of real-world harassment and abuse, which contain nuances less easily identified by non-targets. Here, we present a mixed-methods approach to create classifiers for abuse and harassment which leverages direct engagement with the target group in order to achieve high quality and ecological validity of data sets and labels, and to generate deeper insights into the key tactics of bad actors. We use women journalists' experience on Twitter as an initial community of focus. We identify several structural mechanisms of abuse that we believe will generalize to other target communities.

1 Introduction

Harassment is a significant problem in online spaces. In 2017, one in four Americans reported experiencing online harassment, with more than 60% describing it as a "major problem" (Duggan, 2017).

For journalists, a social media presence is essentially a professional requirement, as it is both a mechanism for locating sources and for promoting stories (Ferrier and Garud-Patkar, 2018); as of 2018, more Americans (roughly 20%) get their news from social media than from printed newspapers (Shearer, 2018). At the same time, journalists receive an inordinate volume of hateful and harassing messages via social media. In a recent survey conducted by the Committee to Protect Journalists (CPJ), 90% of American journalists described online harassment as the biggest threat facing journalists today, with women and minority journalists being disproportionately targeted online (Westcott and Foley, 2019).

This harassment can have devastating effects. In 2016, 10% of women journalists said that they had considered leaving the profession out of fear (Nilsson and Örnebring, 2016), while others avoided certain coverage areas in an effort to mitigate the risk of harassment. Still others may choose not to enter the field at all.

At a time when there is a major need to retain skilled journalists and diversify newsrooms (Scire, 2020), our goal is to develop a research methodology to address this critical threat facing journalists, and ultimately, our free press.

Our contributions in this paper include:

a) Identifying gaps in current anti-harassment tools provided by Twitter;

b) Identifying key strategies used by harassers to circumvent these tools and reach their targets; and

c) Development of a direct-engagement research process and data collection platform to curate datasets with high ecological validity, which will ultimately be used to train better machine learning classifiers for harassment detection.

2 Motivation and Approach

Currently, there are limited options available for journalists to deal with harassing messages on Twitter. Twitter has three primary mechanisms through which a user can control their interactions on the platform: muting, blocking, and the recently introduced "conversations" controls, all of which have a slightly different impact on the content a user can access. For example, muting and blocking can both prevent content from certain users from appearing in some user A's timeline (Twitter, c) (Twitter, d).

However, muted users can still follow and interact with A, while blocked users are no longer able to see A's tweets, and if they visit A's profile, they will see they have been blocked (Twitter, b). The new "conversations" feature, meanwhile, allows user A to specify whether everyone, everyone they follow, or only specific users can reply to a specific tweet (Twitter, a).

While these tools offer impressive granularity, many journalists have both large followings and a professional mandate to interact with their audiences on social media. This makes many of the available controls impractical or ineffective. Moreover, two of the three tools Twitter offers are only effective retrospectively, meaning the targeted user must still read blocked users' offensive tweets before they can choose to mute or block them. Not only does this require journalists to experience harm in order to achieve any potential remediation, if they are targeted by a large number of accounts, the manual effort becomes time-prohibitive.

Shared blocklists have been touted as a means for addressing some of these issues (Geiger, 2016). However, for journalists this can result in blocking users who may be sharing legitimate critiques of their work (Jhaver et al., 2018). As a whole, journalists as a community have expressed desire for more effective user engagement management tools (Saridou et al., 2019).

Furthermore, while many social media platforms do already have automated mechanisms for filtering harassment and hate speech, these are largely based on keyword matching, requiring manual creation with no guarantee of accuracy. Due to the large scale of problematic content on social media worldwide, manual efforts by moderators and filters have also been insufficient (Gerrard, 2018).

The goal of this work is, therefore, to contribute a robust, generalizable mixed-methods approach to constructing harassment training datasets with strong ecological validity, in order to support the development of truly effective classifiers for proactively identifying real-world abusive, harassing, and demeaning speech towards specific communities on Twitter.

Working with journalists, we are collecting a large-scale corpus of personally-harassing messages they have received on Twitter, and have developed an easily-employed annotation method to label messages by degree of observed harassment. Using this data, we then build machine learning classifiers to distinguish between hateful, abusive and neutral tweets. Ultimately, we plan to integrate our trained models into a tool to help journalists navigate and avoid having to see these unwanted, harassing messages.

3 Related Work

Prior work on automatic detection of hateful and abusive speech toward journalists is limited. In (Charitidis et al., 2020), researchers used a manually-validated seed set of journalism-related Twitter accounts to generate a list of target accounts across five languages. Using the Twitter API to conduct keyword-based searches, they then manually annotated hate vs. non-hate tweets. This yielded highly imbalanced corpora, with more "hate" than "non-hate" tweets for each language. Deep learning models trained on each language corpus achieved best macro-F1 scores over .80 for English, French and Greek but somewhat lower for Spanish and German.

Other work has addressed the more general problem of automatic identification of hate speech and abusive language online. In (Waseem, 2016), researchers found that crowd-sourced annotations performed poorly. This indicates the importance of expert annotators, which (Blackwell et al., 2017) situates specifically in terms of classifying harassment.

In (Warner and Hirschberg, 2012), researchers using data from Yahoo and the American Jewish Congress found that anti-Semitic hate speech differed linguistically from speech that targeted other religious or ethnic groups, highlighting the need for a community-specific approach to studying hate speech. (Salem et al., 2016) used content from self-identified hate communities, instead of keywords from hand-coded speech or manually coded hate speech terms, as training data for their work on hate speech detection with some success. In (Nobata et al., 2016), researchers studied abusive language in online user comments on news and finance forums using linguistic, syntactic, and distributed semantic features as well as lexicon-based features. Their dataset has been used to benchmark performance in hate speech detection, as has (Waseem and Hovy, 2016). In (Kshirsagar et al., 2018), researchers developed deep learning models for hate speech detection on Twitter, using transformed word embeddings to classify hate speech on three public datasets.

Researchers in journalism have also used more qualitative methods to study abusive and hateful speech towards journalists. For example, UT Austin's School of Journalism published results from in-depth interviews with 75 female journalists describing how rampant online sexual harassment disrupts their ability to do their jobs (Chen et al., 2018). The Committee to Protect Journalists reported similar findings in 2019 (Westcott and Foley, 2019).

Finally, we note that developers have created tools (e.g. Twitter Block Chain (Wren, 2019) and the recently discontinued Block Together (Hoffman-Andrews, 2020) and the forthcoming Block Party app (Chou, 2020)) specifically designed to address the manual nature of Twitter's muting and blocking functions. While these efforts appear to address an important limitation of Twitter's current systems, they remain a reactive, rather than proactive, approach.

Our proposed methodology for training data collection and annotation incorporates and improves on these approaches as follows: (1) We conduct background interviews with our target community of women journalists in order to identify common heuristics used to carry out harassment on Twitter, in order to develop a more nuanced and balanced dataset for annotation; (2) Annotations are performed by the targets of harassment, guaranteeing a unique level of ecological validity; (3) Our approach takes an empowering rather than exploitative approach to the detection process, promoting harm reduction by allowing harassment targets to participate constructively in the creation of classifiers that can better support their needs.

4 Methodology

We employ a mixed-methods approach that integrates qualitative and quantitative data collection and analysis. We begin by directly engaging with our target group of women journalists who have experienced online harassment. We recruit participants by circulating calls to participation in key networks of women journalists, followed by semi-structured pilot interviews with select participants, in which we question them about patterns of harassment that they have experienced or observed, and about potential tools or interventions that would improve their experience on social media. Despite our convenience sample, two key themes emerged across several pilot interviews, providing valuable insights about the mechanisms of harassment on Twitter, which we describe in Section 5.

Results of these interviews are then integrated into our quantitative data collection pipeline. Using patterns of harassing language and behaviors on Twitter described by interview participants, we develop computational methods to automatically identify those patterns and then use these methods to sample potentially hateful messages from participants' Twitter archives for them to annotate. We describe this data selection process in Section 6.1. Through the process of direct engagement with our target community, we are able to curate a high quality dataset of labeled tweets to support the development of more robust harassment classifiers.

5 Pilot Interviews

To generate a well-balanced training set of tweets, we conducted pilot interviews with several women journalists who have faced significant harassment on Twitter. Through these interviews we learned about specific forms of the "sub-tweeting" and "snitch-tweeting" heuristics that are used to target these and other women journalists with abusive and harassing messages.

The primary form of "sub-tweeting" described to us consists of perpetrators capturing screenshots that contain the target's Twitter profile or username. They then tweet these out with implicit or explicit calls for their followers to tweet at the same target. This behavior constitutes "sub-tweeting" because the absence of the target's username in the text of the original tweet means that target will not be notified of the instigating tweet, and will therefore be caught off-guard by an influx of often abusive tweets, sometimes numbering in the thousands over a period of less than a day. (See (Tufekci, 2014) for more details and examples of "sub-tweeting.") We note that none of Twitter's currently available tools can mitigate this attack; even if the perpetrator has already been blocked by the target, they can simply log out of Twitter and view the target's profile in a web browser in order to obtain the required media.

While the effect of sub-tweeting is to mask the identity of the perpetrator, "snitch-tweeting" is a means of drawing the target into a sub-tweeted thread about themselves to expose them to abuse. Because sub-tweeting intentionally circumvents Twitter's notification systems, targets of abuse will typically be unaware of such sub-tweeting, unless, as described above, it is used to direct traffic to their account. "Snitch-tweeting" consists of adding a target's handle to a thread about them, thus triggering a notification. The goal is for the target then to review the notification and thus to view the abusive thread that precedes the snitch-tweet. Taken together, these results helped us inform our design for the tweet selection portion of our data pre-processing, as described below.

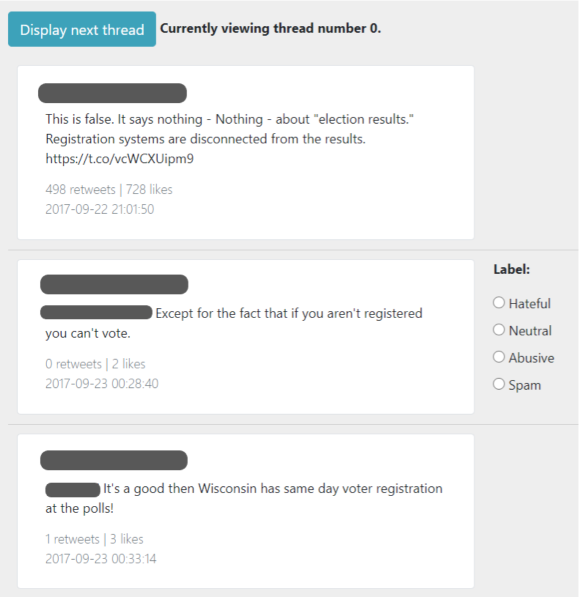
6 Platform Design

In order to curate a high-quality training dataset from participating journalists' tweets, we designed and implemented a two-part, web-based platform to facilitate the data collection and annotation processes. This web platform was designed to balance the proportion of abusive vs. non-abusive tweets that are presented for annotation, without relying on keywords, which are often too coarse-grained to serve as a reliable indicator of abusive content. Instead, we develop heuristics using insights from our pilot interviews as well as private data from the participant's account to include a more nuanced and representative range of potentially abusive tweets for annotation.

The platform is also designed to maximize the efficiency and accuracy of the annotation process, in order to generate a large volume of high-quality training data for deep learning models. We achieve this via batched contextual annotation: participants annotate tweets within the context of the original conversation or tweet thread, rather than annotating them in isolation, simulating how they would have viewed the conversation initially on Twitter. In addition to the annotation tool described above, we have also built a tool for secure data upload, as described below.

6.1 Platform Structure

The process of using our web annotation tool is split into 2 stages, each of which can be accessed via secure, password-protected URLs. First, the study participant securely logs in to the upload platform using a uniquely generated username and password. We ask participants to upload three distinct files, which can be extracted from their Twitter data archive: (1) tweet.js, which contains all of their tweets; (2) muted.js, which contains the list of accounts they have muted, and (3) blocked.js, which contains the list of accounts they have blocked.



Currently viewing thread number 0.

[redacted]

This is false. It says nothing - Nothing - about "election results." Registration systems are disconnected from the results. https://t.co/vcWCXUipm9

498 retweets | 728 likes
2017-09-22 21:01:50

[redacted]

[redacted] Except for the fact that if you aren't registered you can't vote.

0 retweets | 2 likes
2017-09-23 00:28:40

[redacted]

[redacted] It's a good then Wisconsin has same day voter registration at the polls!

1 retweets | 3 likes
2017-09-23 00:33:14

Figure 1: Annotation platform user interface.

Because participants' Twitter archives may contain anywhere from hundreds to tens of thousands of tweets, asking them to label all tweet threads is impractical. Moreover, our goal is to build a training corpus that is approximately equally split between hateful/abusive examples and neutral examples — a very different distribution than we expect to see across the entire corpus, making random sampling inefficient for these purposes.

In order to capture more varied and nuanced examples of problematic data than are likely to be generated by common techniques like keyword filtering, we use multiple heuristics inspired by the participant's muted and blocked lists and the insights gained from our pilot studies to curate a manageable sample of tweets for annotation. Applying these heuristics involves a combination of manual and scripted processing, resulting a gap of several hours to one day between data upload and the availability of data for annotation by each participant. A list of balanced tweet threads fetched from both of these heuristics described below is used to populate the annotation interface.

Our first heuristic using muted and blocked lists uses a Python script to identify all tweets in the tweet.js file that contain any username present in either the muted.js or blocked.js files. Because the presence of a username in these lists reflects an intentional choice on the part of the participant to have these accounts' tweets hidden or blocked from their timeline, we believe the proportion of harmful tweets involving these usernames is likely to be higher than what is present in the corpus as a whole. We then use the thread-retrieval algorithm described in Section 6.1.1 to construct the thread for each relevant tweet.

Our second heuristic searches sub-tweets (described in 5) targeting the study participant, using the query *"[real name] -from:[username] -@[username]"* where "username" is the participant's Twitter handle, and "real name" is the participant's real name. This method allows us to find and capture Tweets in which the study participant was "sub-tweeted" over the most recent 30 days (using Twitter's non-premium Search API). Each of these tweets is then passed through the procedure in Algorithm 1 to once again obtain the corresponding tweet threads.

We find that this methodology retrieves a few interesting threads, but has several shortcomings. First, many of these tweets are positive, and praise the journalist for their work, which makes sense as their name is directly mentioned. Second, and relatedly, we are unable to find sub-tweets where the journalist's name is not mentioned, i.e. the post merely consists of a screenshot of their tweet. These tweets are presumably more negative, as they avoid easy attention from the target. In order to find these sub-tweets, we would have to implement computer vision methods to search for their name in images across Twitter, though it could be difficult to know where to look for these screenshots in the first place. We will investigate this further in future work.

We have also attempted to build a third heuristic using the study participant's Twitter archive to capture scenarios where they had been "snitch-Tweeted" into one of these sub-tweet threads, i.e. find a thread of the structure [image, ..., mention of their username, their response], but we did not find any such threads. We plan to revisit this with future annotators.

To balance the potentially negative threads identified through these heuristics, we also select a random sample of tweets made to non-blocked, non-muted users, and retrieve their corresponding threads. We also exclude from this non-negative sample tweet threads constructed by the participant through self-replies.

6.1.1 Annotation Platform

After data upload and preprocessing, the annotation platform is deployed and sent to the study participant. Participants annotate each tweet sent to them within a retrieved tweet thread. This provides better context to the participant while annotating, addressing a key limitation of many existing datasets, where tweets are presented without context.

Algorithm 1 presents pseudocode for computing a tweet thread from a given tweet. To see the full codebase which joins this algorithm with the aforementioned heuristics into a complete data processing pipeline, please refer to the GitHub repository linked below.[1]

Algorithm 1 Fetch thread from tweet

```
procedure FETCH_THREAD(id, api)
    thread = []
    users = set()
    while id $\neq$ None do
        tweet = api.get_status(id)
        thread.add(tweet)
        users.add(tweet.user_name)
        id = tweet.in_reply_to_id
    if len(thread) $>$ 2 then
        thread.reverse()
    if len(users) $==$ 1 then
        handle("no conversation")
    return thread
```

Label Choices Study participants are currently presented with the following labels: hateful, abusive, neutral, or spam.

- **Hateful speech** is defined as language used to express hatred towards a targeted individual or group, or which is intended to be derogatory, to humiliate, or to insult members of the group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender.
- **Abusive language** is defined as any strongly impolite, rude or hurtful language using profanity, that debases someone or something, or shows intense negative emotion.
- **Spam** includes posts consisting of related or unrelated advertising / marketing, selling products of adult nature, linking to malicious websites, phishing attempts and other kinds

[1] The code for all of our tweet filtering heuristics and thread retrieval methods can be accessed at the following GitHub repository: `https://github.com/ishaan007/woah_emnlp_2020`

of unwanted information, usually executed repeatedly.

- **Neutral** is all tweets that do not fall into any of the prior categories.

We drew these labels from (Founta et al., 2018)'s work, which created a hate speech dataset of 80,000 tweets labeled by crowdsourced annotators, using several iterations of labels (including "offensive", "aggressive", etc.), narrowing them down to these terms. We plan to further iteratively add and remove labels based on insights from interviews and annotation sessions (see 8).

7 Modeling

While we are recruiting more journalists as study participants into our data collection pipeline, we have in parallel been building models of both feature engineering and neural network-based approaches, and testing them on historical hate speech datasets. We plan to take the insights we acquire from these experiments and apply them to classifiers built on our own data once we have accumulated a sufficient amount. We also plan to check the cross-performance between models trained on our own and historical corpora as quality assurance.

The data which we have accumulated so far gives us a good idea of which historical corpora are most similar to our own. We explored several corpora, including (Waseem and Hovy, 2016) and (Founta et al., 2018), but focused on Task 5 of SemEval 2019, "Multilingual detection of hate speech against immigrants and women in Twitter (HatEval)" in English (Basile et al., 2019), as it is most recent and they are all of similar genre.

Both Task 5 subtasks used the same dataset (cicl2018/HateEvalTeam, 2019) but with different labels. Subtask A was a binary classification task to assign a label of "hate" or "non-hate" to each tweet. Subtask B was a multi-class classification task to assign two additional label pairs to each tweet in addition to "hate" or "non-hate": "individual" or "group" and "aggressive" or "non-aggressive". The split across train and development datasets was 9000 to 1000 tweets; these have been open-sourced by the organizing team. The true labels for the test set have not, however, so we evaluate only on the development set.

We replicated the winning approach (Indurthi et al., 2019) for sub-task A in English, which used SMOTE to over-sample the "hate" class as a preprocessing step, followed by the use of Universal Sentence Encoder (Cer et al., 2018) to generate a vector representation of the tweet, and SVM (RBF kernel) to classify the tweet. We also implemented a transformer-based approach for this subtask, based on (MacAvaney et al., 2019), which uses pre-trained BERT for sequence classification, fine-tuned for 10 epochs. This approach in fact outperforms the aforementioned winning approach.

For sub-task B, the multi-classification task, we replicated the winning approach (Bauwelinck et al., 2019) by training three separate classifiers to classify three label pairs individually; these classifiers used a linear SVM on handcrafted syntactic, lexical and bag-of-words features. The optimal hyperparameters were found using grid search. Our experiments with these corpora have given us insights about best practices for training effective models of hate speech, which we plan to apply to our new corpus as we collect more data from participating women journalists. We have additionally been exploring experiments on our collected data with various novel *model architectures* as opposed to *data corpora*, which are elaborated upon in 10.

8 Results and Discussion

Although testing of our platform is still in the pilot phase, early users have shared positive feedback regarding its usability, and have also been able to perform the annotation task with good efficiency, on the order of ~300 tweets per hour. Given the size of previously-collected datasets in this space, our methodology is efficient enough to generate sufficient training data in less than 40 hours, making it both a cost-effective and robust approach. Given the high fidelity of our labels and the near-perfect ecological validity of the training data, we believe that classifiers trained on data collected using our methods will significantly outperform existing classifiers on hateful and abusive speech in the wild.

From early feedback, we have also identified additional labels that participants found relevant, such as "campaign" or "brigade", used to indicate a lexically generic Tweet that is still part of a harassment campaign, as in 2019's "Learn to code" campaign (Molloy, 2019). In addition, our pilot interviews suggest that including a fill-in "other" label may be useful for generating more nuanced classifiers, especially as there has historically been a lack of annotator agreement on what constitutes hateful speech, which tends to vary in severity and lexical nature depending on the situation (Waseem et al., 2017).

9 Limitations

Currently, our approach is limited by its dependence on a feature allowing Twitter users to download an archive of their data; this feature was suspended for roughly two months of the research period in response to the social-engineered hacking of more than 100 accounts (Conger and Popper, 2020). Moreover, some blocked or muted users identified in pre-processing may have been suspended by Twitter, making it impossible to include their potentially harassing messages in our corpus. Finally, while our platform yielded a useful annotation rate, we note that there are inherent limitations to developing classifiers using strictly hand-labeled data.

10 Directions for Future Work

Given the interruption in data collection, we propose to augment our data-access pipeline by building a sufficiently-permissioned Twitter app to download the required data directly from participants' accounts. This would not only provide similarly high-quality data with less burden on participants, it would also provide an ongoing source of test data with which we could refine and improve our classifiers in much closer to real-time.

By leveraging the methods presented in (Wulczyn et al., 2017), moreover, we also believe we could augment and improve the classifiers built from our hand-labeled data using a combination of machine learning and crowdsourcing. We are in general investigating ways to overcome the inherent shortcomings of manual expert annotation, while retaining its significant benefits; for example, augmenting our data annotation tool with active learning annotation (Vlachos, 2006), so that participants only need to annotate the most unclear instances of hateful/harassing/neutral speech.

In regards to model-building, we are exploring ways we can take advantage of the contextual thread annotation scheme present in our annotation platform. Specifically, we have investigated methods using LSTMs (Huang et al., 2016), and are presently investigating graph attention networks (Veličković et al., 2017); these architectures and others like them could allow us to take advantage of the rich metadata and parent tweet text embeddings present in tweet threads, and have the potential to achieve significantly boosted classification performance compared to that of models built on text embeddings of the potentially harassing tweet alone (Mishra et al., 2019).

For the purpose of building the eventual tool to aid journalists in the field, we could alternatively address the relatively small size of our manually-labelled datasets for training deep learning classifiers, by augmenting them against the large, popular corpora already in existence. We could investigate whether this addition would boost performance compared to classifiers trained only on those large, crowd-sourced corpora, as a measure of effectiveness of our methodology.

Finally, we note that while certain semantic features of the classifiers developed using our methodology will differ depending on the community of focus, we hypothesize that by studying several communities with this level of detail and quality, we will eventually be able to identify generalizable features of harassment activities.

11 Conclusion

This work has focused on outlining a novel and generalizable methodology for generating better training datasets for the detection of abusive and harassing speech on Twitter, using women journalists as a test community. By directly engaging the targets of harassment in our research, we have not only created an efficient annotation platform using insights about the structural mechanisms of harassment, but we have offered these victims a constructive way to engage with what are otherwise totally negative experiences. We look forward to continuing to work with women journalists to build data-driven tools against abuse and harassment that allow them to maintain their personal needs while working to uphold our free press.

References

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

Nina Bauwelinck, Gilles Jacobs, Véronique Hoste, and

Els Lefever. 2019. LT3 at SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter (hatEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 436–440, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proc. ACM Hum.-Comput. Interact.*, 1(24):19pp.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Polychronis Charitidis, Stavros Doropoulos, Stavros Vologiannidis, Ioannis Papastergiou, and Sophia Karakeva. 2020. Towards countering hate speech against journalists on social media. *Online Social Networks and Media*, 17:100071.

Gina Masullo Chen, Paromita Pain, Victoria Y Chen, Madlin Mekelburg, Nina Springer, and Franziska Troger. 2018. 'you really have to have a thick skin': A cross-cultural perspective on how online harassment influences female journalists. *Journalism*, page 1464884918768500.

Tracy Chou. 2020. Block party. https://www.blockpartyapp.com/.

cicl2018/HateEvalTeam. 2019. *HateEval 2019 Task 5 Data Files*. https://github.com/cicl2018/HateEvalTeam/tree/master/Data%20Files/Data%20Files.

Kate Conger and Nathanial Popper. 2020. Florida teenager is charged as 'mastermind' of twitter hack. *The New York Times*. https://www.nytimes.com/2020/07/31/technology/twitter-hack-arrest.html.

Maeve Duggan. 2017. Online harassment 2017. https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/.

Michelle Ferrier and Nisha Garud-Patkar. 2018. Trollbusters: Fighting online harassment of women journalists. In Jacqueline Ryan Vickery and Tracy Everbach, editors, *Mediating Misogyny: Gender, Technology, and Harassment*, pages 311–332. Springer International Publishing.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *arXiv preprint arXiv:1802.00393*.

R. Stuart Geiger. 2016. Bot-based collective blocklists in twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society*, 19(6):787–803.

Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, 20(12):4492–4511.

Jacob Hoffman-Andrews. 2020. Block together. https://twitter.com/blocktogether?lang=en.

Minlie Huang, Yujie Cao, and Chao Dong. 2016. Modeling rich contexts for sentiment classification with lstm. *arXiv preprint arXiv:1605.01478*.

Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. 2019. FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 70–74, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Trans. Comput.-Hum. Interact.*, 25(2):33pp.

Rohan Kshirsagar, Tyrus Cukuvac, Kathleen McKeown, and Susan McGregor. 2018. Predictive embeddings for hate speech detection on twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 26–32.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PLOS ONE*, 14:1–16.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Abusive language detection with graph convolutional networks. *arXiv preprint arXiv:1904.04073*.

Parker Molloy. 2019. How a myth about journalists telling miners to "learn to code" helped people justify harassment. *Media Matters*. https://www.mediamatters.org/erick-erickson/how-myth-about-journalists-telling-miners-learn-code-helped-people-justifyharassment.

Monica Löfgren Nilsson and Henrik Örnebring. 2016. Journalism under threat. *Journalism Practice*, 10(7):880–890.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Haji Mohammad Salem, Kelly P Dillon, Susan Benesch, and Derek Ruths. 2016. A web of hate: Tackling hateful speech in online social spaces. In *First Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS 2016) at the International Conference on Language Resources and Evaluation (LREC2016)*.

Theodora Saridou, Kosmas Panagiotidis, and Andreas Veglis. 2019. Towards a semantic-oriented model of participatory journalism management: Perceptions of user-generated content. *Redefining Communication: Social Media and the Age of Innovation*, page 27.

Sarah Scire. 2020. A window into one newsroom's diversity opens, but an industry-wide door shuts (for now). *NiemanLab*. https://www.niemanlab.org/2020/05/a-window-into-one-newsrooms-diversity-opens-but-an-industry-wide-door-shuts-for-now.

Elisa Shearer. 2018. Social media outpaces print newspapers in the u.s. as a news source. https://www.pewresearch.org/fact-tank/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/.

Zeynep Tufekci. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *arXiv preprint arXiv:1403.7400*.

Twitter. a. About conversations on twitter. https://help.twitter.com/en/using-twitter/twitter-conversations.

Twitter. b. How to block accounts on twitter. https://help.twitter.com/en/using-twitter/blocking-and-unblocking-accounts.

Twitter. c. How to mute accounts on twitter. https://help.twitter.com/en/using-twitter/twitter-mute.

Twitter. d. How to use advanced muting options. https://help.twitter.com/en/using-twitter/advanced-twitter-mute-options.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Andreas Vlachos. 2006. Active annotation. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)*.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada.

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

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *CoRR*, abs/1705.09899.

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

Lucy Westcott and James W Foley. 2019. Why newsrooms need a solution to end online harassment of reporters. https://cpj.org/2019/09/newsrooms-solution-online-harassment-canada-usa/.

Cecilia Wren. 2019. Twitter block chain. https://chrome.google.com/webstore/detail/twitter-block-chain/dkkfampndkdnjffkleokegfnibnnjfah?hl=en/.

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