Supplementary Material:
An Information Retrieval Approach to Building 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 University of Economics and Technology
m.kutlu@etu.edu.tr

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

Abstract

This appendix provides supplementary material for our 2021 NeurIPS article [63].

A Related Work

A considerable amount of research [77, 76, 17, 29, 31, 25, 22, 18, 37] has been conducted to construct datasets for hate speech. These approaches can be mainly categorized into three groups: i) keyword-based search [77, 76, 17, 29, 31]; ii) random sampling [18, 37]; and iii) random sampling with the keyword-based search [25, 22].

Keyword-based search. Because less than 3% of tweets are hateful [25], prior studies for constructing hate speech are largely based on the keyword-based search. In particular, a set of manually curated "hate words" are defined and documents containing any of these keywords are selected for annotation. While defining this list of keywords, prior work often considers hate words that are typically used to spread hatred towards various targeted groups. For example, Waseem and Hovy [77] identify 17 different keywords as hate words which covers hate under racism and sexism categories. Waseem and Hovy find 130K tweets containing those hate words and annotate 16,914 of them as racist, sexist, or neither.

Golbeck et al. [29] also define their own set of 10 hate words which are used to cover racism, Islamophobia, homophobia, anti-semitism, and sexism related hate speech. Similarly, Warner and Hirschberg [74] also construct a hate speech dataset containing 9,000 human-labeled documents from Yahoo News! and *American Jewish Congress* where they consider hate words targeting Judaism and Israel. Apart from using keywords targeted towards specific groups, prior work also utilizes general hate words. For instance, Davidson et al. [17] utilize keywords from the HateBase [33], a crowd-sourced list of hate words, whereas Founta et al. [25] use keywords from both the HateBase and an offensive words dictionary [56].

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

Prior work [42, 35] on hate speech dataset construction also focused on profiles of social media users who are known to generate hateful content. For example, Kwok and Wang [42] search for keywords from Twitter users who claim themselves as racist or are deemed racist based on the news sources they follow.

Random Sampling. Kennedy et al. [37] argue that the datasets constructed by the keyword-based search are biased towards those keywords and therefore are not representative of the real-world. Thus the authors randomly sample 28,000 Gab (`gab.com`) posts for annotation. Similarly, de Gibert et al. [18] collect documents for annotation from a White Supremacist forum (`https://www.stormfront.org`) by selecting documents uniformly at random.

Random Sampling with Keyword Search. Prior work has also combined keyword-based search and random sampling to select posts for annotation. For example, Founta et al. [25] develop a hate speech dataset that contains 91,951 annotated tweets categorized into four categories: abusive, hateful, spam, and normal. While many annotated tweets are randomly sampled from the Twitter API, they also select some tweets for annotation based on the keyword-based search to increase prevalence.

Similarly, Wulczyn et al. [82] develop a dataset of personal attacks from Wikipedia comments that contains 37,000 randomly sampled comments and 78,000 comments from users who are blocked. The authors mention that since the prevalence of personal attacks on those 37,000 randomly sampled comments is only 0.9%, they increase the prevalence of personal attacks comments by searching over the blocked users' comments.

We do not use the keyword-based search method in our work and do not apply only random sampling. Instead, we adapt the well-known pooling method [67] for constructing IR test collections to select documents to be annotated. To our knowledge, this is the first work using pooling method to construct a hate-speech dataset.

B Hate Speech Challenges

There are various challenges associated with developing a dataset for hate speech, and in this section, we will discuss those challenges. However, we should note that most of these challenges are widely debated issues in machine learning research including dataset bias [72], annotator bias [28], documenting datasets [27], task decomposition [36], selection of annotators [4] and others. If adequate steps are not taken to mitigate various issues associated with the challenges, datasets will reflect various forms of biases. Consequently, researchers and practitioners who deprioritize the biases in the dataset would run the risk of inflicting greater harm to human society by deploying automated systems trained on these biased datasets.

B.1 Definition of Hate Speech

Even experts disagree on what constitutes hate speech [25, 77, 17, 68]. It is a complex phenomenon typically associated with relationships between groups and depends on the nuances of languages. Since there is no legal definition of hate speech, various international organizations, social media platforms, and research articles [17, 77] define hate speech differently. There are two notable similarities between these definitions: 1) hate speech incites violence or is intended to be derogatory, and 2) hate speech is directed towards certain targeted groups.

However, these definitions are not comprehensive enough to cover the real-world representation of hate speech. For example, MacAvaney et al. [46] point out these definitions cover whether someone is attacked or humiliated in hate speech. However, praising a particular group (e.g., KKK, Nazi) may also be considered hate speech and this is not covered by the existing definitions.

B.2 Annotation Schema

Hate speech is a relatively complex phenomenon because the difference between other related concepts (e.g., cyberbullying [14], abusive language [55], discrimination [71], etc.) and hate speech is not obvious [24]. As a result, different hate speech datasets have different annotation schema for hate speech and other related concepts [17, 25].

The binary annotation schema is the basic schema that labels a post as either hate speech or normal speech. However, prior studies mostly annotate hate speech using non-binary schema. Since offensive language is prevalent in social media and does not necessarily always represent hate speech, Davidson et al. [17] annotate their dataset using three categories: i) hate speech, ii) offensive language, and iii) normal speech. Mathew et al. [48] also follow these three categories in annotation. Apart from hate speech, offensive language, and normal speech categories, Founta et al. [25] annotate their dataset into four (4) other categories, namely: i) abusive language, ii) aggressive behavior, iii) cyberbullying, and iv) spam.

Non-binary schemes based on the intensity of hate speech are also utilized in prior studies. For example, Del Vigna et al. [19] implement strong hate, weak hate, and no hate, Kumar et al. [39] categorize posts into overtly aggressive, covertly aggressive, not aggressive categories. Poletto et al. [61] compare binary annotation scheme against rating scale and best-worst ranking scale for hate speech annotation and find that rating scale is comparatively better than the other two schemes.

B.3 Annotation Guidelines

Often, it is very challenging for the annotators to decide whether a particular post or document is hateful or not [65, 68]. Thus a carefully designed annotation guideline is crucial to have a better quality hate speech dataset. However, prior work also significantly differs from each other in terms of designing annotation guidelines. Most of the time, researchers only specify the definition of categories (e.g., hate or offensive) [17, 25] but do not provide any additional clarification about how to interpret each of those categories.

Furthermore, since annotators are not provided with any contextual information regarding the social media post, different authors provide different types of guidelines to their annotators to resolve this absence of context. For example, Davidson et al. [17] instructed the annotators not only to consider the presented tweets but also to think about the context in which tweets might appear before making the judgment. However, such practices risk making the task of annotating hate speech more subjective.

B.4 Selection of Annotators

Previous studies regarding hate speech also vary in terms of hiring annotators. Given the nuances of language and the degree of difficulty of annotating hate speech, expert annotators can play an important role in achieving a higher inter-annotator agreement [26]. Prior work hired experts from different backgrounds including feminists and anti-racism activists [77], content moderators [57], PhD students in Linguistics [40], experts in Natural Language Processing [49]. In addition, since expert annotators typically have domain knowledge, it is expected that expert annotators tends to agree more with other experts in annotating hate speech. For example, Waseem [75] find that crowd-workers have a lower inter-annotator agreement score than experts.

However, hiring expert annotators is expensive, and there is also a scalability issue. Thus for a large-scale annotation task, prior work typically employs crowd-workers [17, 25]. To make sure that crowd-workers have the necessary expertise to perform the annotation task, prior work sometimes restrict annotation tasks to workers meeting certain qualifications (e.g., Amazon Mechanical Turk). However, crowd-workers lacking proper training are more prone to do the "keyword-spotting" while labeling hate speech. As a result, crowd-workers may be more likely to label a post as hate speech than experts [75].

B.5 Annotators' Bias

Prior work also investigates how annotators' demographics (e.g., gender, race, first language) affect the perception of the annotators to hate speech. Gold and Zesch [30] find that female annotators who are typically part of the targeted group in hate speech are more likely to annotate a possible gender-related post as sexist than their male counterparts (i.e., *gender bias*). A similar type of observation (i.e., *racial bias*) is also made by Kwok and Wang [42] while working on racist hate speech. Additionally, Sap et al. [69] report that annotators who are unfamiliar with the African American English (AAE) dialect are more likely to label documents containing AAE as racist, although those same documents may be considered non-racist by native AAE speakers.

The *racial bias* problem is even more severe when we consider the crowd-workers; for example, in Amazon Mechanical Turk, non-AAE speakers are overrepresented [34]. Additionally, the annotators' political ideology can also unintentionally manifest in yielding a politically biased dataset [79]. Furthermore, it has been found that both expert and crowd-workers are prone to similar types of bias while annotating hate speech [16]. Given this counter-intuitive observation, prior work [16] argues to develop a better training process for the annotators to mitigate the annotators' bias.

B.6 Measurement of Annotator Agreement

Previous studies diverge significantly in reporting on the quality of the annotations, especially the inter-annotator agreement score [24]. Typically, Cohen's κ, Fleiss κ, Krippendorf's α, or a plain observed agreement percentage are reported in prior work. On the other hand, there are also many studies in hate speech that do not report any inter-annotator agreement score [24].

Since many factors are involved in annotation (e.g., annotation scheme, annotation guidelines, annotators' background), the reported agreement scores among prior studies vary widely. For example, Bohra et al. [11] report a Cohen κ score of 0.982, whereas Del Vigna et al. [19] report a Fleiss κ score of 0.19. Furthermore, different studies argue for different thresholds for an acceptable inter-annotator agreement score [5, 23]. Typically more complex annotation schemes [19, 68] produce a lower inter-annotator agreement score than a simple, binary annotation scheme [11, 17].

We are not familiar with any prior work on hate speech annotation using self-consistency checks [84], which we believe complements traditional use of annotator agreement measures. Conceptually, for objective tasks with a single true answer, we expect reasonable annotator agreement, while on more subjective tasks [53] (e.g., favorite ice cream flavor) we do not expect annotators to agree. While an annotator can be expected to be self-consistent for either task type, self-consistency seems particularly valuable for subjective tasks when annotators are expected to disagree with one another. Hate speech annotation lies in the spectrum between objective vs. subjective tasks. Some objectivity is necessary to yield consistent data for training detection models, but low annotator agreement remains common. This is why we believe self-consistency measures can complement traditional practice.

B.7 Absence of a Benchmark Hate Speech Dataset

Although hate speech is a widely discussed topic and there are many publicly available hate speech datasets, there is no commonly accepted benchmark dataset for hate speech detection [70, 60, 47]. This is largely due to the fact that in hate speech, data degradation is a known issue [78, 13]. This is because researchers primarily collect hate speech from social media and release only the IDs of the social media posts in the hate speech domain. For example, Watanabe et al. [78] and Chaudhry and Lease [13] report that a number of tweets released initially by [77] are not available anymore. Furthermore, standard benchmark datasets (e.g., SQUAD [64], MSMARCO [54]) provide a standard train-test-validation split, whereas most of the hate speech datasets lack in providing this train-test-validation split [47].

B.8 Less Generalizability of Automated Hate Speech Detection Models

Although the generalization capability of a model can be largely attributed to the complexity of the model itself, Gröndahl et al. [32] argue that for hate speech, the nature and composition of the datasets are more important than the model itself. This is because researchers differ from each other regarding various related issues of annotating hate speech, including definition, categories, annotation guidelines, types of annotators, aggregation of annotations. Consequently, different hate speech datasets have different natures and compositions. As a result, automated hate speech detection systems trained on one hate speech dataset exhibit poor generalization performance on another hate speech dataset. For example, Arango et al. [3] show that the state-of-the-art hate speech detection models [8, 1] provide very poor cross-data generalization performance when trained on the dataset created by Waseem and Hovy [77] but tested on the HateEval dataset [9].

B.9 Summary of Challenges

In conclusion, all these issues discussed in this section should provide the practitioners a general overview about why they should be vigilant in performing their due diligence while deploying

automated hate speech detection systems using any constructed hate speech dataset, including our own. While solving all these issues is beyond the scope of this work, here, we particularly focus on the data sampling process (i.e., which post to select for annotation) so that the final hate speech dataset has a better coverage of hate speech from all categories (C1 and C2 of Figure 1) with a limited budget for annotation. Prior work regarding the data sampling process of hate speech is discussed in the next section.

C Tweet Corpus Collection

We construct a collection of documents (e.g., tweets) by collecting a random sample of tweets from the Twitter Public API, which usually provides 1% random sample of the entire Twitter stream in a given time range. In our case, we have collected tweets from May-2017 to Jun-2017.

Next, we apply regular expressions to get rid of tweets containing retweets, URLs, or short videos. Tweets are also anonymized by removing the @username tag. However, we do not remove any emojis from tweets as those might be useful for annotating hate speech. We filter out any tweets as non-English unless two separate automated language detection tools, Python Langdetect[1] and Python Langid[2], both classify the Tweet as English. Finally, after removing duplicate tweets, approximately 13.6 million English tweets remain as our tweet corpus collection.

D Document Annotation Process

D.1 Annotation Guidelines

Our designed annotation guidelines consist of a clear definition of what constitutes as hate speech and examples covering various cases of hate speech. Following Davidson et al. [17], we also instruct the annotators not to annotate any post as hateful if the derogatory language used in the post does not have any target associated with a protected group. Furthermore, annotators are explicitly instructed that they should not label any pornographic content as hateful.

D.2 Annotation Interface

Instructions ask annotators to follow these steps in order:

1. Highlight any words or phrases in the post **INCITING VIOLENCE**.
2. Highlight any **DEROGATORY LANGUAGE** in the post on the basis of group identity.
3. If the post **IMPLICITLY** incites violence or denigrates an individual or group on the basis of group identity, select that option. **[INCITING VIOLENCE / DEROGATORY LANGUAGE]**
4. If the target is **EXPLICIT**, highlight the INTENDED TARGET in the post. If the target is implicit, **name the target**.
5. Identify the type of group targeted (explicit or implicit). **[BODY / GENDER / IDEOLOGY / RACE / RELIGION / SEXUAL ORIENTATION / OTHER]**
6. Based on your answers to the above steps, do you believe the post is hateful? **[YES / NO]**
7. We welcome any additional explanation of your labeling decisions you would like to provide. **[TEXTBOX INPUT]**

This annotation scheme requests the annotators to identify both the violating content (Steps 1-2) and the demographic group targeted (Steps 3-5). When the annotators reach Step 6, they have already completed several sub-tasks. They then decide whether they believe the post is hateful or not.

Targeted group identification. Once annotators identify the target of hate (implicit or explicit), they select the group identity of the target. There are seven (7) categories of targeted groups in our interface, as listed in the interface (Step 5).

[1] https://pypi.org/project/langdetect/
[2] https://github.com/saffsd/langid.py

Categorization of highlighted terms. While Mathew et al. [48] ask the annotators to only highlight terms that are related to hate speech, in our interface, annotators have to do both highlighting and categorization of terms that are related to the actions and the targets of hate speech. For example, for this post "Good morning Kanye. Shut the fuck up", the annotators have to highlight "Kanye" as the target and they also have to highlight terms "Shut" and "fuck" and categorize those terms as derogatory terms. In addition to the potential downstream value of the collected rationales, it is known from prior work [51, 41] that requiring annotator rationales improves label quality, even if the rationales are ignored.

D.3 Collecting Annotations

We hire annotators from Amazon Mechanical Turk. To ensure label quality, only annotators with at least 5,000 approved HITs and a 95% HIT approval rate are allowed. We pay $0.16 USD per tweet. Since three annotators annotate each tweet, including the platform fees, we have paid $4,930.17 USD in total to annotate 9,667 tweets. We apply majority voting to compute the final label.

Guidelines indicate that if a tweet is hateful, annotators must identify both the targets and the actions related to hate speech; otherwise, their work will be rejected. Furthermore, typically only a few words or phrases are related to targets and actions of hate speech; highlighting all words will yield rejection. One might consider this rule as too prohibitive, because highlighting all words might be necessary in some cases. However, in our pilot study we observed that annotations in which entire text is highlighted corresponded to low quality work in almost all cases.

To facilitate quality checks, we collect the annotations in iterative small batches. The quality check typically includes randomly sampling some annotated tweets and checking the annotations. Finally, if we reject any HIT, we notify the worker why we have done so and re-assign the task to others.

E State-of-the-art Models

I. LSTM. The LSTM model implemented by Badjatiya et al. [8] achieves 93% F_1 on the hate speech dataset created by Waseem and Hovy [77] (though it is unclear whether they report macro or micro F_1). The deep learning architecture starts with an embedding layer with dimension size of 200. Then it is followed by a Long Short-Term Memory (LSTM) network. Their final layer is a fully connected layer with a soft-max activation function to produce probabilities across three classes, namely sexist, racist, and non-hateful, at the output layer. To train the model, they use the categorical cross-entropy as a loss function and the Adam optimizer. In our case, we modify the output layer with two nodes and use Sigmoid as an activation function as we have a binary classification task. Finally, for the loss function, we use the binary cross-entropy loss function. The model is trained for ten epochs following Badjatiya et al. [8].

II. BiLSTM. Agrawal and Awekar [1] design a BiLSTM architecture that achieves $\approx 94\%$ score in terms of both micro and macro averaged F_1 on the hate speech dataset constructed by Waseem and Hovy [77]. Their architecture consists of the following layers sequentially: 1) Embedding layer, 2) BiLSTM Layer, 3) Fully Connected layer, and 4) output layer with three nodes. They also use the softmax activation function for the final layer and the categorical cross-entropy as loss function with the Adam optimizer. For this BiLSTM model, we also perform the same modification as we do for the LSTM model. In addition, we train the model for 30 epochs.

Note. For both LSTM and BiLSTM models, we use the corrected versions of these models reported by Arango et al. [3].

III. BERT. We also utilize Bidirectional Encoder Representations from Transformers (BERT) [20] which achieves 67.4% macro averaged F_1 on the hate speech dataset created by Mathew et al. [48]. With pooling, we use the BERT-base-uncased model with 12 layers, 768 hidden dimensions, 12 attention heads, and 110M parameters. For fine-tuning BERT, we apply a fully connected layer with the output corresponding to the CLS token. The BERT model is fine-tuned for five epochs.

Document Pre-processing. We pre-process tweets using `tweet-preprocessor`[3]. Then we tokenize, and normalize those pre-processed tweets. For the TF-IDF representation, we further stem those tweets using Porter Stemmer [81].

[3] https://pypi.org/project/tweet-preprocessor/

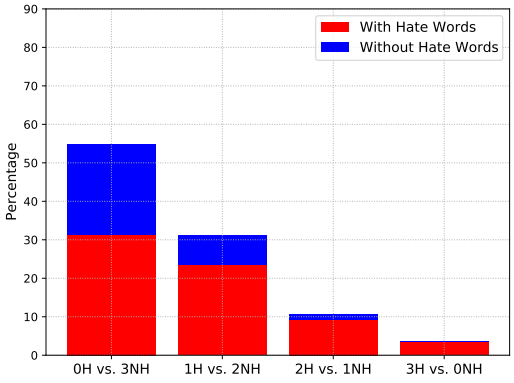

Figure 1: Distribution of annotators' agreement

Document Representation. For LSTM and BiLSTM models, documents are represented using a word embedding where the embedding layer is initialized using the Twitter pre-trained GloVe embedding [58] which is pre-trained on 2 billion tweets. For the Logistic Regression and Naive Bayes models, we generate the TF-IDF representation [66] of documents using bigram, unigram, and trigram features following the work of Davidson et al. [17].

F Additional Dataset Properties

Effect of the presence of hate words on annotation. We also analyze how the presence of hate words affects the decision-making process of the annotators. To achieve that, we plot the frequency distribution of the number of annotators who agree regarding the label of tweets in **Figure 1**. For example, the 1H vs. 2NH entry on the x-axis of Figure 1 represents how many times one annotator labels a tweet as hateful, but two annotators annotate that tweet as non-hateful. The frequency distribution is divided into two sets where one set does not contain hate words, and another set contains hate words. For example, 3.36% of the time, all three annotators label a tweet as hateful when there are hate words in tweets. In contrast, only 0.19% of the time, three annotators label a tweet as hateful when there is no hate word in that tweet (3H vs. 0NH). This observation is also true for the other entries on the x-axis. Annotators agree more given known hate words in tweets.

Targeted group label. The targeted group label of a hateful tweet is computed using majority voting. However, we find that there are 86 hateful tweets where all three annotators assign different targeted groups. In those cases, the targeted group label is assigned to "UNDECIDED". A closer inspection on this "UNDECIDED" category reveals that for 83 tweets out of 86 UNDECIDED tweets, one annotator out of three provides "NONE" as a targeted group, contrary to our guidelines. Furthermore, for 35 tweets, one annotator selects "GENDER", but another annotator mostly picks "IDEOLOGY" or "RACE". For example, one annotator selects "GENDER" but the other annotator picks "IDEOLOGY"

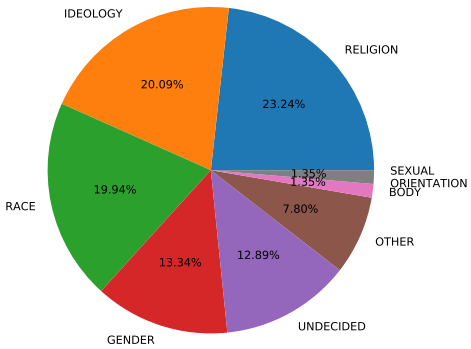

Figure 2: Hate speech distribution over targeted groups.

for tweet "F**k off you old socialist millionaire Clinton b***h". Here we find "UNDECIDED" category because the tweets can be interpreted as hateful to two different target groups.

Figure 2 presents the percentage of hate speech under each targeted group. We see that the largest three targeted groups are religion, ideology, and race, accounting for more than 62% of hate speech in the dataset. Those top three groups have almost equal share ($\approx$ 20%). To understand further each of these targeted groups, we analyze the top 10 most frequent terms appearing under each of these targeted groups and find that some of these targeted groups cover a broad range of hate speech which is one of our goals. For example, the "Race" group consists of hate speech on the basis of ethnicity, race, colour, or descent. Similarly, the "Religion" group covers hate towards Muslims and Christians. Furthermore, under the "Ideology" group, targeted groups include individual or groups having various ideologies such as liberals, republicans, and feminists. The same observation holds if we consider the "sexual orientation" group. However, if we consider the "Gender" group, we find that this group mainly covers hate towards only women. Finally, 66% and 20% of the hateful tweets are explicit and implicit, respectively, whereas the remaining 14% falls under the "UNDECIDED" category of annotator disagreement.

Inter-annotator agreement for rationales. Recall that our annotators were asked to distinguish three cases of implicit vs. explicit labeling decisions: types of hate (derogatory language or inciting violence) and demographic group targeted. In explicit cases (only), annotators provided a rationale by highlighting a portion of the tweet supporting their labeling decision. This leads to three complications in how to measure inter-annotator agreement for rationales. Firstly, there are three different categories of rationales. For simplicity, we ignore this and simply report one statistic over all categories combined. Secondly, since rationales are only provided in explicit cases, the number of rationales per labeling decision depends on how many annotators identified explicit evidence for their labeling decision. Since inter-annotator agreement measures typically assume the same number of annotations per item, we make another simplifying assumption that all annotators provided rationales, but that for implicit cases, no tokens were part of the rationale. Thirdly, our interface has annotators highlight rationales at the character level, which we then map to binary token level labels as follows: 1 if all characters in the token are highlighted, and 0 otherwise. After this, we can then calculate annotator agreement as a binary labeling task over all tokens (over all tweets). This yields raw agreement of 95% (most tokens are not part of rationales), Fleiss $\kappa = 0.07$, and Gwet's $AC_1 = 0.95$.

G Wellness Risks for Hate Speech Annotators and Moderators

To the best of our knowledge, none of the previous work has analyzed the effect of frequent exposure to hate speech on the well-being of human annotators. However, prior evidence suggests that exposure to online abuse has serious consequences on the mental health of the workers [83]. The studies conducted by Boeckmann and Liew [10] and Leets [43] to understand how people experience hate speech have found that low self-esteem, symptoms of trauma exposure, etc., are associated with the constant exposure to hate speech.

In addition, according to the premier diagnostic manual for psychological disorders, DSM-5 [6], a person can suffer from post-traumatic stress disorder (PTSD) via "repeated exposure" to indirect traumatic material, which in our case is online hate speech. Prior work [45, 50] regarding the psychological effect of indirect trauma also recognizes PTSD as "secondary traumatic stress", "compassion fatigue", and "vicarious traumatization". Several studies have been conducted to understand the consequences of constant exposure to trauma, e.g., Kleim and Westphal [38] perform research on the first responders, Perez et al. [59] investigate the police officers, Wagaman et al. [73] study the social workers, etc. Although there is no prior study regarding the consequences due to the constant exposure to hate speech, recently, some employees have sued Microsoft. In their lawsuit, they claimed that due to repeated exposure to traumatic contents (e.g., child pornography) as a part of their work, they are being diagnosed with PTSD [44].

From the above discussion, it is evident that there are some serious consequences for the constant exposure to indirect trauma, and the same is true for annotating hate speech. As a result, we can see that there are some efforts from organizations to improve the work environment of employees. To be more specific, more than 12 technology companies (e.g., Adobe, Apple, Dropbox, Facebook, GoDaddy, Google, Kik, Microsoft, Oath, PayPal, Snapchat, Twitter) have implemented some guidelines

developed by The Technology Coalition [15] to "support of those employees who have exposure to online child pornography in the course of their work".

Some of the key mitigating steps proposed in the Employee Resilience Guidebook [15] are: i) limiting the amount of time an employee can spend on moderating child pornography contents, and ii) acquiring informed consent from employees so that they have a clear understanding of the role as a moderator. The latter strategy has also been emphasized by the University Institutional Review Boards (IRBs).

Owing to the fact that we work with crowd-workers via a crowd-platform where we do not have any direct control over their work environment, directly implementing the above-mentioned strategies is beyond our control. Consequently, to reduce the risks for the crowd-workers associated with annotating hate speech, we have posted a disclaimer as shown below at the very beginning of the annotation task.

> Our research seeks to reduce the spread of hate speech on social media by training computer programs to automatically detect hate speech. To accomplish this, we ask human annotators to read tweets and label hate speech. We understand that this labeling task requires content that can be disturbing to read. If you prefer to return this task rather than work on it, we understand. In general, if you ever experience mental or emotional distress, please know that help is available online. Helplines include `https://suicidepreventionlifeline.org` in the USA and `http://suicide.org/international-suicide-hotlines.html` internationally. For additional reading on this subject, please consult our research article, "The Psychological Well-Being of Content Moderators"(`https://www.ischool.utexas.edu/~ml/papers/steiger-chi21.pdf`).

In the spirit of *informed consent*, this disclaimer helps the crowd-workers to make an informed decision about whether or not to accept the task. It also suggests where to seek help regarding any mental or emotional distress.

H Discussion and Limitations

The primary research goal of this work is to develop a hate speech dataset that covers a broader range of hate speech while maintaining a comparable prevalence of hate. While developing this dataset, we have made various operational decisions regarding various issues discussed in Section B. In this section, we discuss the practical implication of those operational decisions in the constructed dataset.

I. While defining hate speech, we emphasize that the presence of both targets and actions is necessary to consider a post as hate speech. This definition is consistent with the prior definition of hate speech used in the datasets created by Davidson et al. [17], Founta et al. [25], de Gibert et al. [18] and Nobata et al. [55]. We select this hate speech definition because it covers a wide range of hate with a generalized set of targets. However, this also creates room for different interpretations among the annotators, which is reflected in the inter-annotator agreement score of our dataset. On the other hand, Waseem and Hovy [77] provide an eleven (11) steps approach, including the presence of specific hashtags to consider a post as either sexist or racist. Note that, our annotation interface also has multiple steps to determine whether a post is hateful or not. Another practical limitation of the definition used in this dataset is that it does not cover those hate speech related to praising certain groups (e.g., praising Nazi).

II. Our dataset has been annotated using a binary scheme considering only hate speech and normal speech. However, non-binary schemes are very prevalent in the hate speech domain [17, 25] because it helps us to understand hate and other related concepts (e.g., offensiveness, aggressiveness) using the same annotation effort. One practical limitation of the binary scheme used in this work is that annotators might label an offensive post as hate speech because they have no other categories to specify. For example, pornographic-related posts are typically offensive, and there is a clear instruction regarding this in our guidelines, and yet many annotators label these offensive posts as hate speech in our annotated dataset.

III. Unlike prior work [77, 17, 25, 18] where a simple annotation interface has been employed, by adapting the suggestion of Sanguinetti et al. [68], we have designed a hierarchical, structured annotation interface to annotate hate speech. The rationale behind this hierarchical interface is to perform the task decomposition, which can help the annotators navigate their decision-making to label a post as hate or normal speech. However, following prior work of Sanguinetti et al. [68], we have also noticed that the use of this structured interface does not necessarily improve the inter-annotator agreement score. Further investigation regarding the annotation interface is needed to understand how different annotation interfaces affect the quality of the annotated data.

IV. Following prior work [25, 17, 48], we have used crowd-workers to annotate hate speech. As mentioned earlier, since training the crowd-workers is practically challenging, we select crowd-workers with specific qualifications (e.g., a minimum HIT approval rate). However, since the crowd-workers are more prone to annotate a post as hate speech based on the keyword-spotting [75], the quality of the annotated data might be affected. To compensate for this issue, we have designed a thorough annotation guideline with various examples considering different boundary cases of hate speech. Furthermore, by noticing the fact that many previous studies do not disclose their guidelines [24], we have made our annotation guidelines publicly available with our dataset.

V. We have assumed that annotators can complete the annotation task effectively without any contextual information irrespective of their demographics, expertise, ideologies, etc. Note that this assumption holds for both pooling and active learning methods. However, prior work by Al Kuwatly et al. [2] has shown that if the demographic factors (e.g., first language, gender, etc.,) are not properly handled, potential *annotation bias* might arise in the dataset. For example, it has been found that native English speakers are better at detecting toxic comments [2] than non-native English speakers when the annotation task is in English.

VI. Our Twitter-specific dataset is not necessarily representative of how hate is expressed on other social media platforms and forums. For example, tweets have a fixed maximum length (i.e., 280 characters), so our dataset does not cover any hateful expressions longer than this limit.

Other related biases that we should be concerned about regarding the dataset constructed in this work are: i) temporal bias [52], ii) user bias [3] and iii) pooling bias [12]. Although we have collected tweets from May-2017 to Jun-2017 using a uniform random sample, familiar topics discussed during that time frame would be over-represented in the constructed corpus and thus introduce the temporal bias in the dataset. Furthermore, Arango et al. [3] mention that 65% of hate speech annotated in the dataset created by Waseem and Hovy [77] are generated by only two (2) Twitter users. Since we have taken a random sample of tweets from the Twitter API and another random sample from the pooled tweets, user bias should be less prevalent in the constructed hate speech dataset.

VII. The pooling process introduces two additional biases : i) pool depth bias and ii) system bias and here, we discuss those biases in the context of hate speech. When the pool depth is very shallow, many posts remain unjudged. If several of those unjudged documents are hate speech, that introduces a pool depth bias. Typically, employing a wide pool depth helps reducing this bias. On the other hand, system bias is introduced when the number of machine learning models in pooling is very few and those models are not diverse. This phenomenon reduces the prevalence of hate in the annotated dataset drastically. Generally, system bias can be addressed by increasing the number and diversity of machine learning models [12].

Although handling annotation-related biases (Section B) and other biases discussed above is not the key contribution of this work, readers should be aware of those biases (e.g., racial bias, political bias, gender bias, etc.) while designing an automated hate speech detection system using our hate speech dataset. Specifically, the presence of these biases might adversely impact the quality of the constructed hate speech dataset. Moreover, when machine learning models are trained on biased datasets, those models typically learn and exasperate those biases. For example, these biased automated hate speech detection systems might flag posts written using American English dialect (AAE) as hate speech or impair political debates on social media platforms because the systems are politically biased.

Apart from the above-discussed issues of the constructed hate speech dataset, we have made some key assumptions related to pooling and active learning methods that are crucial to achieving our research goal. Here, we discuss those assumptions and their corresponding limitations.

Assumption I. Recall that hate speech is relatively rare in social media (only 3% of social media posts are hateful [24]), and annotating everything is not feasible. Thus to maximize the prevalence of

hate speech for a given budget, during the pooling, documents (e.g., tweets) that exceed a certain threshold in terms of their likely hatefulness are only considered in the pooled document set. Note that due to this assumption, the process for constructing the pooled document set is a *non-random sampling* process, which is prone to *sampling bias* because of its nature. In other words, documents having a likely hatefulness score less than the provided threshold are not present in the final dataset, and some of those discarded documents might be hateful. Note that sampling bias is also an issue for active learning [62]. In addition, since machine learning models are employed in both pooling and active learning, the selection of posts for annotation is also affected by *model bias* [52]. This could be most pronounced with our active learning approch because only a single model is used to select tweets, which may reduce the diversity in selection vs. the pooling approach across models.

Typically prior work on constructing hate speech datasets is mostly based on searching hate words [77, 9, 25] and/or finding potential hateful social media users [17]. Because of their nature, they are also heavily criticized for having a strong sampling bias. For example, only two users are responsible for generating 70% of sexist tweets, and only one user generates 99% of racist tweets [80] in the dataset constructed by Waseem and Hovy [77]. Unlike prior work, we do not rely on keyword-based searches or finding hateful users. In addition to that, to mitigate the sampling bias, we have two random sampling steps at two different stages of the pipeline: 1) corpus construction phase and 2) final sampling of documents for annotation. However, potential users of the dataset constructed in this work should be aware of this potential selection bias. They might adopt some de-biasing strategies discussed in prior work [21, 7] while designing their automated hate speech detection systems trained on our hate speech dataset.

Assumption II. Another key assumption made in the pooling-based approach is that there exist prior hate speech datasets on which prediction models can be trained to kickstart the pooling technique. However, this assumption does not always hold, especially for the less-studied languages (e.g., Amharic, Armenian, etc.). Additionally, since the pooling technique relies on prior hate speech datasets, any known limitations of those datasets will influence the document selection process of the pooling technique. For example, the dataset constructed by Waseem and Hovy [77] covers sexist posts from the sports domain, and the dataset by Grimminger and Klinger [31] covers political hate speech covering the 2020 US Election topic. This type of *topical bias* for the sake of identifying hate speech in the existing hate speech datasets can also be propagated through the pooling technique, and the constructed hate speech dataset can have the same type of topical bias.

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] Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 184–190.

[3] Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*. 45–54.

[4] Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* 36, 1 (2015), 15–24.

[5] Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, 4 (2008), 555–596.

[6] American Psychiatric Association et al. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

[7] Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*. 49–59.

[8] 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.

[9] 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*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 54–63. `https://doi.org/10.18653/v1/S19-2007`

[10] Robert J Boeckmann and Jeffrey Liew. 2002. Hate speech: Asian American students' justice judgments and psychological responses. *Journal of Social Issues* 58, 2 (2002), 363–381.

[11] Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*. Association for Computational Linguistics, New Orleans, Louisiana, USA, 36–41. `https://doi.org/10.18653/v1/W18-1105`

[12] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Voorhees. 2006. Bias and the limits of pooling. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*. ACM Press, Seattle, Washington, USA, 619. `https://doi.org/10.1145/1148170.1148284`

[13] Prateek Chaudhry and Matthew Lease. 2020. *You Are What You Tweet: Profiling Users by Past Tweets to Improve Hate Speech Detection*. Technical Report. University of Texas at Austin. `http://arxiv.org/abs/arXiv:2012.09090` arXiv:2012.09090.

[14] Ying Chen. 2011. Detecting offensive language in social medias for protection of adolescent online safety. (2011).

[15] The Technology Coalition. 2013. Employee Resilience Guidebook for Handling Child Sex Abuse Images. (2013). `https://www.thorn.org/wp-content/uploads/2015/02/EmployeeResilienceGuidebookFinal7-13-1.pdf`

[16] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516* (2019).

[17] 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.

[18] 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`

[19] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*. 86–95.

[20] 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).

[21] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 67–73.

[22] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*. 29–30.

[23] Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. *Computational linguistics* 30, 1 (2004), 95–101.

[24] 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.

[25] 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.

[26] Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. *arXiv preprint arXiv:1710.07395* (2017).

[27] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[28] Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. 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)*. Association for Computational Linguistics, Hong Kong, China, 1161–1166. `https://doi.org/10.18653/v1/D19-1107`

[29] 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.

[30] Michael Wojatzki Tobias Horsmann Darina Gold and Torsten Zesch. 2018. Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments. (2018).

[31] 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).

[32] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*. 2–12.

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

[34] Paul Hitlin. 2016. Research in the crowdsourcing age: A case study. (2016).

[35] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909* (2015).

[36] Huan Jiang and Shigeo Matsubara. 2014. Efficient Task Decomposition in Crowdsourcing. In *PRIMA 2014: Principles and Practice of Multi-Agent Systems*, Hoa Khanh Dam, Jeremy Pitt, Yang Xu, Guido Governatori, and Takayuki Ito (Eds.). Springer International Publishing, Cham, 65–73.

[37] 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).

[38] Birgit Kleim and Maren Westphal. 2011. Mental health in first responders: A review and recommendation for prevention and intervention strategies. *Traumatology* 17, 4 (2011), 17–24.

[39] Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1–11. `https://www.aclweb.org/anthology/W18-4401`

[40] Ritesh Kumar, Aishwarya N Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of hindi-english code-mixed data. *arXiv preprint arXiv:1803.09402* (2018).

[41] Mucahid Kutlu, Tyler McDonnell, Yassmine Barkallah, Tamer Elsayed, and Matthew Lease. 2018. What Can Rationales behind Relevance Judgments Tell Us About Assessor Disagreement?. In *Proceedings of the 41st international ACM SIGIR conference on Research and development in Information Retrieval*. 805–814.

[42] Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 27.

[43] Laura Leets. 2002. Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speech. *Journal of social issues* 58, 2 (2002), 341–361.

[44] Sam Levin. 2017. Moderators who had to view child abuse content sue Microsoft, claiming PTSD. (2017). `https://www.theguardian.com/technology/2017/jan/11/microsoft-employees-child-abuse-lawsuit-ptsd`

[45] Marné Ludick and Charles R Figley. 2017. Toward a mechanism for secondary trauma induction and reduction: Reimagining a theory of secondary traumatic stress. *Traumatology* 23, 1 (2017), 112.

[46] 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.

[47] 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*. 150–161.

[48] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *arXiv preprint arXiv:2012.10289* (2020).

[49] Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018. Did you offend me? classification of offensive tweets in hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 138–148.

[50] Casey L May and Blair E Wisco. 2016. Defining trauma: How level of exposure and proximity affect risk for posttraumatic stress disorder. *Psychological trauma: theory, research, practice, and policy* 8, 2 (2016), 233.

[51] Tyler McDonnell, Mucahid Kutlu, Tamer Elsayed, and Matthew Lease. 2017. The many benefits of annotator rationales for relevance judgments. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 4909–4913.

[52] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[53] An Thanh Nguyen, Matthew Halpern, Byron C. Wallace, and Matthew Lease. 2016. Probabilistic Modeling for Crowdsourcing Partially-Subjective Ratings. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*. 149–158.

[54] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

[55] 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*. 145–153.

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

[57] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep learning for user comment moderation. *arXiv preprint arXiv:1705.09993* (2017).

[58] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. `http://nlp.stanford.edu/data/glove.twitter.27B.zip`. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[59] Lisa M Perez, Jeremy Jones, David R Englert, and Daniel Sachau. 2010. Secondary traumatic stress and burnout among law enforcement investigators exposed to disturbing media images. *Journal of Police and Criminal Psychology* 25, 2 (2010), 113–124.

[60] 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.

[61] Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate speech annotation: Analysis of an italian twitter corpus. In *4th Italian Conference on Computational Linguistics, CLiC-it 2017*, Vol. 2006. CEUR-WS, 1–6.

[62] Ameya Prabhu, Charles Dognin, and Maneesh Singh. 2019. Sampling bias in deep active classification: An empirical study. *arXiv preprint arXiv:1909.09389* (2019).

[63] 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 Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS): Datasets and Benchmarks Track*.

[64] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

[65] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118* (2017).

[66] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24, 5 (1988), 513–523.

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

[68] Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[69] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 1668–1678.

[70] 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*. Association for Computational Linguistics, Valencia, Spain, 1–10. `https://doi.org/10.18653/v1/W17-1101`

[71] Neil Thompson. 2016. *Anti-discriminatory practice: Equality, diversity and social justice*. Macmillan International Higher Education.

[72] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. 2017. A deeper look at dataset bias. In *Domain adaptation in computer vision applications*. Springer, 37–55.

[73] M Alex Wagaman, Jennifer M Geiger, Clara Shockley, and Elizabeth A Segal. 2015. The role of empathy in burnout, compassion satisfaction, and secondary traumatic stress among social workers. *Social work* 60, 3 (2015), 201–209.

[74] 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*. 19–26.

[75] 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.

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

[77] 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`

[78] Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. *IEEE Access* 6 (2018), 13825–13835. `https://doi.org/10.1109/ACCESS.2018.2806394`

[79] Maximilian Wich, Jan Bauer, and Georg Groh. 2020. Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 54–64.

[80] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)*. 602–608.

[81] Peter Willett. 2006. The Porter stemming algorithm: then and now. *Program* (2006).

[82] 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*. 1391–1399.

[83] Michele L Ybarra, Kimberly J Mitchell, Janis Wolak, and David Finkelhor. 2006. Examining characteristics and associated distress related to Internet harassment: findings from the Second Youth Internet Safety Survey. *Pediatrics* 118, 4 (2006), e1169–e1177.

[84] Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. 2014. Multidimensional Relevance Modeling via Psychometrics and Crowdsourcing. In *Proceedings of the 37th international ACM SIGIR conference on Research and Development in Information Retrieval*. 435–444.