# Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection

Harini Suresh
hsuresh@mit.edu
Data + Feminism Lab, MIT, USA

Rajiv Movva
rmovva@mit.edu
Data + Feminism Lab, MIT, USA

Amelia Lee Dogan
dogan@mit.edu
Data + Feminism Lab, MIT, USA

Rahul Bhargava
r.bhargava@northeastern.edu
School of Journalism, Northeastern University, USA

Isadora Cruxên
i.cruxen@qmul.ac.uk
School of Business and Management, Queen Mary University of London, UK

Ángeles Martinez Cuba
angelesm@mit.edu
Data + Feminism Lab, MIT, USA

Giulia Taurino
g.taurino@northeastern.edu
Khoury College of Computer Science, Northeastern University, USA

Wonyoung So
wso@mit.edu
Data + Feminism Lab, MIT, USA

Catherine D'Ignazio
digazio@mit.edu
Data + Feminism Lab, MIT, USA

## ABSTRACT

Data ethics and fairness have emerged as important areas of research in recent years. However, much work in this area focuses on retroactively auditing and "mitigating bias" in existing, potentially flawed systems, without interrogating the deeper structural inequalities underlying them. There are not yet examples of how to apply feminist and participatory methodologies *from the start*, to conceptualize and design machine learning-based tools that center and aim to challenge power inequalities. Our work targets this more prospective goal. Guided by the framework of data feminism, we co-design datasets and machine learning models to support the efforts of activists who collect and monitor data about feminicide — gender-based killings of women and girls. We describe how intersectional feminist goals and participatory processes shaped each stage of our approach, from problem conceptualization to data collection to model evaluation. We highlight several methodological contributions, including 1) an iterative data collection and annotation process that targets model weaknesses and interrogates framing concepts (such as who is included/excluded in "feminicide"), 2) models that explicitly focus on intersectional identities rather than statistical majorities, and 3) a multi-step evaluation process — with quantitative, qualitative and participatory steps — focused on context-specific relevance. We also distill insights and tensions that arise from bridging intersectional feminist goals with ML. These include reflections on how ML may challenge power, embrace pluralism, rethink binaries and consider context, as well as the inherent limitations of any technology-based solution to address durable structural inequalities.


*FAccT '22, June 21–24, 2022, Seoul, Republic of Korea*

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533132

**ACM Reference Format:**
Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxên, Ángeles Martinez Cuba, Giulia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3531146.3533132

## 1 INTRODUCTION

Data ethics and fairness have become active areas of research in recent years. Work in this vein often focuses on "mitigating bias" in harmful systems, building "fair" or "transparent" algorithms, or performing retroactive audits. While these developments are important, they typically locate the source of injustice in individual people or specific technical systems, and solutions that emerge often take the form of "technological Band-Aids" [29, p. 60].

Alternate framings of data and algorithms are emerging which are rooted in considerations of power and justice [6, 10, 24, 25, 29, 61]. These trace the root cause of "biased" systems not to individual programmers or design decisions, but rather, to the deeper structural inequalities in which data-driven systems are embedded. Inspired by this framing, and with a feminist lens, we ask the question posed by D'Ignazio and Klein in *Data Feminism*: "why should we settle for retroactive audits of potentially flawed systems if we can design with a goal of co-liberation from the start?" [29, p. 63]

However, to our knowledge, there are not yet examples of how to apply feminist and participatory methodologies throughout the entire machine learning (ML) life cycle, to conceptualize and design ML tools that center and aim to challenge power inequalities. We target this more prospective goal through a concrete case study that asks how digital, data-driven tools can support the efforts of activists who collect and monitor data on the topic of feminicide (or femicide) — broadly understood as gender-related killings of women and girls. Many of these activist organizations use news articles to find and record instances of feminicide in their region. This process typically relies on using search queries that return a large fraction

of irrelevant results, adding to the arduous labor of monitoring this kind of violence [28]. Through co-design sessions with groups, we conceptualized, built, and deployed an ML-based system to deliver more relevant media results to activists on a regular basis, thus helping to facilitate their monitoring work.

In our initial pilot of the system, we found that it worked well for groups that broadly monitored all feminicides in a given region. However, the results were consistently not relevant for groups who focused on specific, racialized forms of feminicide (e.g., Black women in the US killed by police). In this paper, we focus on our subsequent efforts to perform iterative data collection, modeling and evaluation steps to build out context-specific models for two organizations that monitor gender-related killing as it intersects with white supremacy, state violence and colonialism. Throughout this process, we take an explicitly feminist approach, both in our overaching process—which we strive to make iterative, reflexive, contextual, and participatory—as well as the technology we build. In particular, we draw on four principles of data feminism that are most salient to our work: *challenge power, embrace pluralism, consider context*, and *rethink binaries and hierarchies*. We describe how these goals shaped each stage of our approach, from problem conceptualization to data collection to model evaluation.

We highlight several methodological contributions, including 1) a data collection process in which we iteratively identify and incorporate context-specific examples that target model weaknesses, 2) annotation and modeling methods that incorporate sociohistorical context and explicitly focus on intersectional identities, and 3) a three-stage evaluation process — with quantitative, qualitative and participatory steps — focused on real-world, context-specific usefulness.

At the same time, we acknowledge the ongoing tensions we grappled with, and areas where our tools currently fall short. In doing so, our contribution is two-fold: we describe our approach to this specific case study in detail, but also aim to provide inspiration for how to mobilize intersectional feminist values in technology more generally. We conclude with the idea that intersectional feminist and participatory ML is possible, but that the creators of such systems should consider themselves in the humble and bounded role of supporting and sustaining activist efforts to shift power.

## 2 BACKGROUND AND RELATED WORK

### 2.1 Power, Oppression, and Intersectional Feminism

Our work builds on intersectional feminist thought, and stems from the acknowledgement that power is not equally distributed in the world. By *power*, we refer to configurations of structural privilege, in which certain groups experience unearned advantages — i.e., because they control the dominant institutions of law, education and culture [17, 29, p. 24]. *Systems of oppression* arise because of the unequal distribution of power, and involve the systematic mistreatment of certain groups of people by others. There are many dominant and marginalized identities in society, and forms of oppression manifest differently across them. For example, gender oppression takes the form of cissexism, heterosexism and patriarchy, while racial oppression manifests in racism and white supremacy. Other forms of oppression include ableism, colonialism and classism.

These various dimensions of disempowerment converge in complex and unique ways for groups and individuals [21]. The concept of *intersectionality* provides a contrast to single-axis or additive views of how discrimination manifests. Intersectional feminism is grounded in a long history of Black feminist thought stretching back to at least the mid-1800s [60] and into the present day [18], and conceptualized in particular by Kimberlé Crenshaw [20] and the Combahee River Collective [16]. Intersectionality comprises an analytical framework based on the premise that different systems of power are interdependent, "mutually constructing one another" and producing complex social inequalities that fundamentally shape both individual and group experiences [18, p. 16].

An intersectional view situates all systems of oppression as interlocking axes that construct what Patricia Hill Collins names the *matrix of domination* [17]. The matrix of domination describes four interrelated domains that operate at different scales of granularity, from interpersonal to institutional, shaping society and human actions. These include the structural domain, where oppression is organized and codified in law and policy; the disciplinary domain, where it is enforced through bureaucracy and hierarchy; the hegemonic domain, in which oppressive ideas are circulated through culture and media; and the interpersonal domain, which captures the everyday lived experiences of individuals. Different systems of oppression manifest across each domain to different degrees and generate dynamics of subordination and vulnerability in varying ways. Ultimately, they converge to create very real consequences, such as the invisibility of violence against women of color [21].

Together, intersectionality and the matrix of domination show that there are not clean cut distinctions between victims and oppressors; rather, each individual "derives varying amounts of penalty and privilege from the multiple systems of oppression which frame everyone's lives" [17]. This understanding motivates the goal of *co-liberation*: the idea that dismantling interlocking systems of oppression is necessary for everyone's collective freedom.

### 2.2 Participatory and Feminist ML

While intersectionality and the matrix of domination are *conceptual models* for how inequality is structured and reinforced, expanding participation is often suggested as a *method* for rectifying imbalances of power. Sloane et al. [72] survey three main forms that "participation" has taken in ML. Most frequently, much of ML involves *participation as work*, where people (often unknowingly) contribute examples [35] and annotations [63] to large-scale datasets, e.g., through systems that monitor activity [66] or scrape web data [26]. This work is typically unacknowledged and poorly compensated (if at all), and it does not meaningfully integrate user perspectives [7, 40].

More recently, there has been increasing awareness of the importance of more meaningful community and end user participation during the development of machine learning systems [52]. This can take the form of *participation as consultation* [72], where stakeholders are consulted at specific points throughout the development process for need-finding or feedback [9, 13, 56]. While potentially promising, this setup is inherently top-down, "designing *for*" groups rather than committing to their ongoing inclusion [36, 57, 77]. In

contrast, *participation as justice* involves more long-term relationships with participants "based on mutual benefit, reciprocity, equity and justice" [72]. These types of approaches focus on "designing *with*" communities to ensure outcomes are valuable to diverse and minoritized stakeholders [45]. Participation as justice — i.e., centering the voices of marginalized groups throughout the whole ML life cycle — is critical if we want to design for the goal of co-liberation.

However, while such participatory approaches have been more widely explored in peripheral domains (e.g., participatory action research [5], design justice [19], disability justice [41], environmental justice [3]), they are fairly rare in practice during ML development. Some examples from recent years focus on community-oriented educational materials: for example, the Algorithmic Equity Toolkit is a set of tools for recognizing and understanding algorithmic systems co-designed with community stakeholders [48]; similarly, "A People's Guide to AI" [62] aims to create accessible educational materials about AI tools and their consequences. Others target specific points during the development process: e.g., the Contextual Analysis of Social Media (CASM) approach is a method for data labeling in which community expertise shapes a team-based annotation process [64]. In our work, we use intersectional feminist thought as a guide for understanding how to incorporate meaningful participation throughout the entire ML life cycle, from problem conceptualization to system evaluation and deployment.

Other work has considered the implications of feminist epistemologies to areas of ML. For example, Hancox-Li and Kumar [42] apply the frameworks of situated knowledge and standpoint theory [44] to understand the values implicit in feature importance methods. Barabas et al. [4] explore how the concept of "studying up" [59] could be used to reorient ML research questions to better confront power. Gray and Witt [39] outline a mix of technical, social and cultural interventions that constitute a feminist data ethics of care approach to ML. And Buolamwini and Gebru [11] bring an intersectional lens to model evaluation, considering the performance of facial analysis systems not only across skin tones or across genders, but also across specific intersections of them. Our own work specifically builds on the framework laid out in *Data Feminism*, which introduces seven principles for thinking about and using data that are "informed by direct experience, by a commitment to action, and by intersectional feminist thought" [29].

### 2.3 Feminicide, Counterdata Collection and Media Analysis

The term femicide [68] is broadly understood to mean the gender-related killings of women and girls [34]. Latin American feminists have built on this work and introduced the term *feminicidio (feminicide)*, as a way to capture the systemic nature of this violence and the role of the state in enabling it through either omission, negligence or complicity [54]. Activists in Latin America have also played a pivotal role in bringing worldwide attention to the issue of feminicide through powerful demonstrations and movement-building, and eighteen countries have instituted legislation criminalizing feminicide [14]. However, policies to ensure adequate information collection have not followed, and official government data on gender violence and feminicide remain incomplete, difficult to access, infrequently updated, contested, and underreported due to stigma, victim-blaming, or matters of legal interpretation [8, 37, 50, 70, 76]. Incomplete or inaccurate records of feminicide can be understood as *missing data*, resulting from power imbalances — across all four domains of the matrix of domination — in the collection environment. Missing data masks the systemic nature of this violence, allowing it to go unpunished. In this sense, the lack of information can be interpreted as an active form of "women disempowerment" [53], in the constructed process of gendered delegitimization that results from heteroimposed "patriarchal pacts" [2] and normative violence [12]. "In such a framing, women are set up to be forgettable. Ignorable. Dispensable — from culture, from history, and from data. And so, women become invisible" [22, p. 21]. This is particularly true for groups at the intersection of patriarchy and other forces of domination, like settler colonialism, and the movement around Missing and Murdered Indigenous Women, Girls and Two Spirit people (MMIWG2) was founded to challenge the invisibility of this violence.

When the state and its institutions fail to collect important data, civil society organizations increasingly step in to fill these gaps, performing counterdata collection as a way to regain empowerment, legitimization, and visibility [23, 29, 58]. Counterdata science practices mount an explicit challenge to the data practices (collection, analysis, deployment, visualization, ethics, values) of mainstream, well-resourced "counting institutions" such as governments and corporations [27]. As Alice Driver notes in the case of Mexico, "the most accurate records of feminicide are still kept by individuals, researchers, and journalists, rather than by the police or a state or federal institution" [30, p. 7].

In the absence of reliable "official" data sources, anti-feminicide activists in the field build significant data acquisition pipelines to support creating databases of incidents from media reports. From a data perspective, this work is similar to media analysis projects undertaken in computational social science where informatic tasks include event detection, content extraction, classification, and entity extraction. Platforms such as Media Cloud [69] and GDelt [55] aggregate and collate news stories from the open web to support media research projects. Various projects support extracting and annotating content from news corpora to identify entities, dates, and more [33, 46]. Researchers have combined those systems, and built others, to study the emergence of protest movements, creating automated classification systems to analyze their representation in the news [43]. Others have built systems to automatically detect and extract victims of police killings in the US [49], and analyze shifts in narrative frames employed by the media to discuss them [81].

## 3 CASE STUDY: DATA AGAINST FEMINICIDE

The Data Against Feminicide project — co-organized by the Data + Feminism Lab at MIT, Feminicidio Uruguay, and the Latin American Initiative for Open Data (ILDA) — is an initiative intended to support feminicide data activists through knowledge-sharing, technology development and community building. Since 2020, the project has conducted semi-structured interviews with 31 monitoring organizations based mainly in the Americas, with a focus on Latin America. Through these interviews, it became clear that most organizations use media articles to identify cases of feminicide in their regions

[28]. However, a consistent issue is that many of the articles retrieved via search queries are not relevant. In practice, activists spend much of their time reading articles that are not instances of feminicide (but that might describe other violent or traumatizing events) in order to find the minority that are feminicide, which is both emotionally taxing and time consuming.

Through co-design sessions with groups, we conceptualized, built, and deployed an ML-based system to deliver more relevant media results to activists on a regular basis (referred to from here as Email Alerts). The system uses news content from Media Cloud, an open source platform for analysis of online news [69]. A particular organization using the Email Alerts system can customize a search query and set of place-based media sources to best suit their project needs. Media Cloud then retrieves matching articles from its continually updated database of global news stories, which are run through a machine learning model we developed that predicts the probability that the article will be relevant to the organization (i.e., describes an instance of feminicide). Articles above a particular probability threshold (which defaults to 0.75) are sorted by the probability of feminicide and delivered in a daily email digest (matching activists' existing workflows) and can also be viewed in an online dashboard.

Our focus in this paper is on the development and evaluation of the ML models used to filter articles. For our initial prototype, we collected and annotated two datasets of 399 and 424 articles, respectively: the first in English, in collaboration with Women Count USA (a US-based activist group), and the second in Spanish, in collaboration with a Feminicidio Uruguay (a Uruguay-based activist group). This data was used to train two language-specific logistic regression models to predict the probability of feminicide from the text of an article. The English and Spanish models achieved 84.8% and 81.6% accuracy in 5-fold cross-validation, respectively. Further details about data collection, annotation, and model performance for this initial iteration can be found in D'Ignazio et al. [32].

Our results with this initial version of the model served as a proof-of-concept that such a system could reduce the burden of labor for activists in this space, and in Spring 2021, we ran a two-month pilot with seven groups to gauge if and how it would help in practice. The pilot was run simultaneously in Spanish and English, with four groups based in the United States, one group in Uruguay, and two groups in Argentina. Among the four English groups, we received dramatically different feedback about the model's performance. Women Count USA, which monitors all US feminicides, reported that the results were overall very relevant and useful. Another organization, Black Femicide US, monitors femicides of Black women and reported mixed but still useful results, with around 4 out of every 10 articles the system sent being relevant. Both have continued using the system in their work. However, the system did not source relevant results for two organizations in particular, both of which monitor specific, intersectional types of feminicide: 1) Sovereign Bodies Institute (SBI), a group who tracks missing and murdered Indigenous women, girls, and two spirit people (MMIWG2) and 2) the African American Policy Forum (AAPF), who monitor police violence against Black women as part of the #SayHerName campaign. Feedback from these groups consistently showed a lack of relevant articles being returned by the system, despite modifying the search queries to add relevant terms. The groups' frustration could be heard in comments in focus groups and weekly surveys—for example, an activist from SBI wrote, "The majority of articles are not relevant to our focus, which means I'm actually spending more time than usual trawling through potential additions because I'm reviewing so many more news articles than usual." As we reached the conclusion of the pilot, it became apparent that groups dealing with general feminicides (i.e., all women killed in a specific region) were much more satisfied with the tools than groups that monitored more intersectional forms of such violence.

Our participatory development and evaluation processes made it clear that the model needed to be adjusted to better serve projects with more targeted monitoring needs. A commitment to intersectionality means that systems should work not only for the mainstream, majority use case, but also for those on the margins; and it means acknowledging that this might require dedicating additional time and resources to these specific intersectional use cases. With agreement from the two groups, we went through further iterative data collection, modeling and evaluation steps to deploy new models for their specific monitoring needs. The rest of the paper focuses on motivating and describing this development process, the resulting models, and the underlying theoretical prompts this project creates for those working to create participatory approaches to machine learning informed by intersectional feminism.

## 4 IMAGINING AN INTERSECTIONAL FEMINIST APPROACH TO ML

Throughout this project, we strive to take an explicitly feminist approach, both in the technology we built and our overarching process. To do so, we build on the intersectional feminist principles proposed by *Data Feminism* [29]. In this section, we focus on four principles that are most salient to our work: *challenge power, embrace pluralism, consider context, and rethink binaries and hierarchies*. We describe how they shaped our research questions, our approach, and the resultant tools we built. At the same time, we acknowledge the ongoing tensions we grappled with, and areas where our tools currently fall short. Our goal is both to illustrate our approach to this specific case study, as well as provide inspiration for how to mobilize intersectional feminist values in technology more generally.

### 4.1 Challenge Power

Focusing on the problem of gender-related violence might seem like it is inherently challenging power. However, there are many possible directions within this space that could be pursued — not all of which challenge power to the same extent. A deeper examination of how unequal power manifests in each domain of the matrix of domination guides what we choose to build and who we work with. For example, in the structural domain, we understand that data about feminicide is missing in large part because the state and its institutions neglect to collect and report adequate information about it. Activists who collect counterdata challenge and hold these institutions accountable, reclaiming power in the process. With an intersectional lens, we can further understand how unequal power manifests differently for people or groups fighting multiple, intersecting systems of oppression. In the disciplinary domain, for example, victim blaming and a failure to investigate are particularly

prominent when the victim is Black or when the case involves police violence [31]. In the hegemonic domain, biased media narratives misgender or ignore trans people, disregard Indigenous identity, or stigmatize and blame sex workers for violence inflicted on them [73, 79].

Throughout our work, then, rather than support governmental organizations or large international NGOs, we collaborate with and build tools in service of civil society activists. Moreover, to truly challenge power involves not only getting our tools to work for the broadest, majority group, but also for those on the margins who are face multiple intersecting oppressions. Here, we choose to work with a range of organizations monitoring feminicides, including those with specific, intersectional focuses (e.g., Black women in the US killed by the police). And in practice, throughout the development process, the problems we focus on center around getting our system to work equally well for groups monitoring intersectional violence, rather than only improving upon a singular version that primarily benefits the general, majority cases.

## 4.2 Embrace Pluralism

To embrace pluralism means insisting that "the most complete knowledge comes from synthesizing multiple perspectives, with priority given to local, Indigenous, and experiential ways of knowing" [29, p. 125]. Embracing pluralism explicitly calls for the use of participatory methods throughout the ML process, from project inception through deployment, in support of "locally informed, ground-truthed insights that derive from many perspectives."

This principle fundamentally shapes our research process, which is iterative and participatory. The project is co-led by three people with diverse backgrounds and positionalities, including a counterdata activist who helped surface the value of activists' situated knowledge from the start. The ideas we chose to develop were brainstormed in participatory co-design sessions with two activist groups working in different contexts, and piloted with seven organizations across four countries. A *pluralistic process* helps enable *conceptual pluralism* — wherein the framing of the problem is cognisant of and affirms the value of multiple perspectives rather than enforcing a single ground truth. For example, while developing our datasets and models, we collected and annotated data in collaboration with each group, with the understanding that a case that was relevant to one group might not be relevant to another, and vice versa. And importantly, our goal is to support the work that activists are already doing — not to replace or override it.

## 4.3 Consider Context

To consider context means acknowledging that "data are not neutral or objective" [29, p. 149]. Rather, data — and missing data — are the product of unequal social relations, and this context is critical for accurate and ethical analysis.

In our project, the importance of context became clear after our first phase of model development, in which we used the same language-specific feminicide-detection model for each of the pilot organizations. While some groups (in particular, those that broadly monitored all types of feminicide) found they were receiving relevant results, others (those that focused on specific, intersectional cases, such as MMIWG2) struggled to find any relevant results at all. If we think about the data (which in our case, are news articles) as being produced within the context of intersecting systems of oppression, we can understand that the way that different forms of violence are reported about (as well as if they are reported about at all) can be vastly different. For example, cases of MMIWG2 are often under-reported; and when they are reported, the articles often omit Indigenous identity or other key information [38]. Cases involving police violence are often written with biased, victim-blaming narratives [31]. In addition, we found that many US-based Indigenous news sources are published as PDFs, a model of distribution that wasn't readily ingestible by Media Cloud, which was built to support the dominant form of RSS-based syndication and distribution. As we found, by not considering this context, a one-size-fits-all model fails on these types of cases.

In subsequent rounds of development and evaluation, we have built out different models for groups that work in different contexts and face different challenges. In part, this involved iterative, context-specific data collection/annotation in collaboration with each group. Our evaluation is also multi-step and contextual, acknowledging that a model that works for one group might not work for another, and that a model that performs well on a test sample might perform differently once deployed in a real-world context.

## 4.4 Rethink Binaries and Hierarchies

Rethinking binaries and hierarchies means "challenging the gender binary, along with other systems of counting and classification that perpetuate oppression" [29, p. 18]. As mentioned previously, we intentionally work not only with groups that view their work as recording "feminicide," but also with those that monitor other types of gendered and racialized violence. We annotate the datasets for each group in accordance with what they consider relevant, rather than an arbitrary binary gender label.

However, our ability to challenge binaries and categorization schemes is an ongoing limitation of this work. While many activist organizations do include murders of trans people in their definition of feminicide, for example, these cases tend to be recorded much less frequently. The reasons for this are complex, but include a lack of legal protections in the structural domain, misgendering and victim-blaming in the disciplinary domain, and media bias in the hegemonic domain. LGBTQ+ activists have overcome some of these barriers by relying on community members and ally networks to identify trans violence. News article-based ML classifiers would not be as helpful for this work, and may even perpetuate erasure by learning victims' incorrect genders.

Even beyond gender, choosing categorization schemes to use in each dataset brought up unresolved tensions. For example, AAPF focuses on recording cases of Black women in the US killed in police violence. We annotated the corresponding dataset with "police violence" and "feminicide" labels, but did not include a "race" annotation due to the difficulty of inferring it from most news articles. As a result, ML classifiers trained on this data are not able to capture the specific intersectional cases of interest, and we aren't able to internally evaluate the model's performance on Black women specifically.

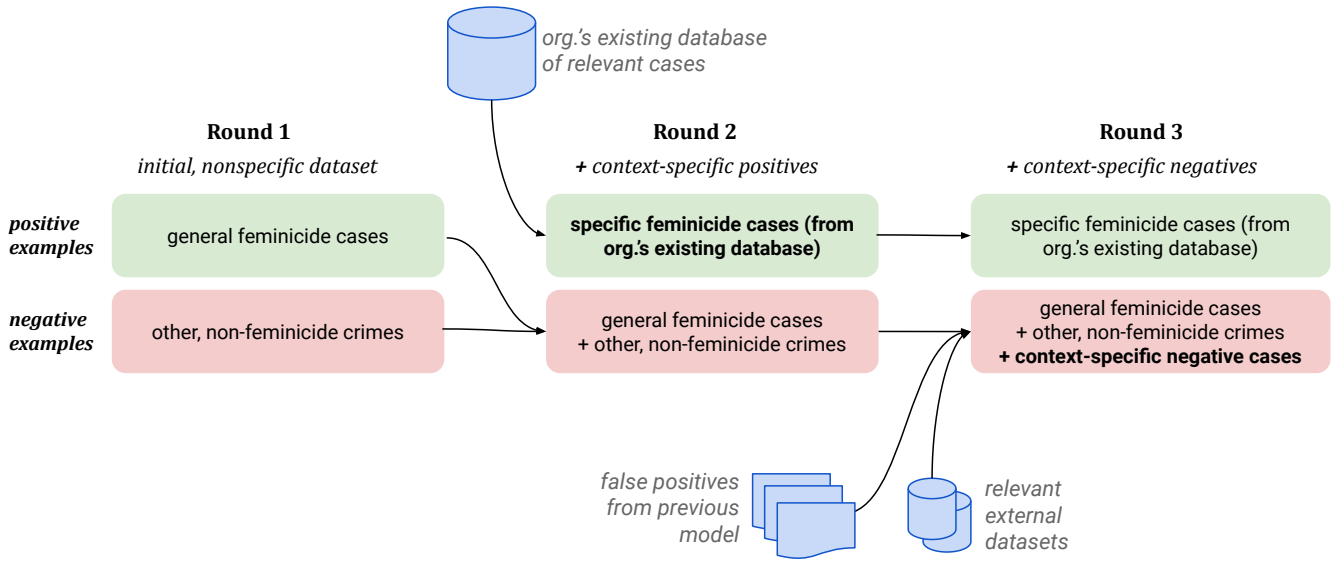


**Figure 1: Our data collection process involves iteratively collecting context-specific positives (e.g., by sourcing ground-truth articles from organizations' existing databases) and context-specific negatives (e.g., by identifying and collecting types of negative examples close to the decision boundary, for which the model is underspecified).**

## 5 DEVELOPING CONTEXT-SPECIFIC FEMINICIDE DETECTION MODELS

In this section, we describe the methods we used to develop and evaluate context-specific models for SBI and AAPF. This consists of four main steps: 1) data collection, where we iteratively collect context-specific examples to target model weaknesses, 2) data annotation, where we re-annotate data with multiple relevant attributes, 3) modeling, where we explore how to build and combine models that center specific intersectional identities, and 4) evaluation, where we define appropriate metrics of success and how to assess them.

### 5.1 Data Collection

For SBI and AAPF, we collected additional data and trained models in an iterative process to target model weaknesses. The data we use are news articles, and a single example is a string containing the article full text.

***Round 1.*** The initial pilot model was trained with general feminicide cases as positive examples and non-feminicides (usually other crimes) as negative examples. While this model was useful for broadly monitoring feminicide, it was not context-specific, and returned mostly unrelated cases for groups with a specific intersectional focus.

***Round 2.*** In the subsequent round of context-specific data collection, we asked groups to send us articles they had already collected in their existing databases, which we used as positive examples. In line with *embracing pluralism*, this entailed a shift from a predefined notion of "feminicide" to a framing of predicting "relevant cases" for a particular group. We used a sample of cases from our initial dataset (both general feminicides and non-feminicides) as negative examples. We then trained models using this data, and deployed them to the Email Alerts system, to monitor their performance in a real-world context (i.e., on the thousands of unseen articles pulled in daily from Media Cloud).

While the returned articles were vaguely relevant to the context (e.g., related to police violence), we still found that they were not finding the specific cases of interest. With AAPF, for example, we found that the list of returned articles was often dominated by police violence against Black men or cases where the police were investigating other violence, both of which are much more common in the media than Black women killed by the police. This shortcoming may be due to an underspecified decision boundary: while the positive examples we curated from the groups reflected specific and intersectional cases of interest, the negative examples were much more general, and did not include many cases close to the decision boundary. In this failure mode, the model may learn that positive cases involve police violence, but nothing drives it to learn the more specific intersectional identities of interest.

For SBI, we noticed a similar pattern — many of the highly-scored articles were generally about MMIWG2 developments (e.g., task forces being formed), rather than cases of a particular victim. Without context-specific negatives (e.g., MMIWG2-related articles that *don't* describe a particular case of feminicide), the model may have learned that any MMIWG2 terms indicate a positive article.

***Round 3.*** In the next round of data collection, we focused on collecting more tailored *negative* examples, using both select external datasets as well as a sample of the previously unseen false positives returned by the previous model. For AAPF, we focused on collecting articles about police violence against Black men and cases where the police were investigating non-police-related violence. We sourced

| | | *Feminicide* | |
|---|---|---|---|
| | | **Y** | **N** |
| *Police Violence* | **Y** | 249 | 168 |
| | **N** | 189 | 136 |

**Table 1: Data breakdown for AAPF.**

| | | *Feminicide or missing victim* | |
|---|---|---|---|
| | | **Y** | **N** |
| *MMIWG2-related or Indigenous news* | **Y** | 217 | 110 |
| | **N** | 201 | 151 |

**Table 2: Data breakdown for SBI.**

articles from external datasets such as the Washington Post's *Fatal Force* dataset [67], Guardian's *The Counted* dataset [74], and the *Whose Deaths Matter?* dataset [81], as well as a sample of the new false positives returned by the prior model. For SBI, we collected context-specific negative examples of Indigenous non-feminicides (e.g. missing or murdered Indigenous men, or articles generally related to the MMIWG2 movement but not referencing a specific case) by sampling new false positives.

The data collection process for both organizations highlights the importance of targeted data collection to find negative examples: while activists provided lists of articles that can be used as positives, our work relies on iterative error analysis and data collection to develop a classifier with high specificity.

## 5.2 Data Annotation

Even given more context-specific data, the model may still be limited by the default binary classification setup, in which examples either have a positive or negative label. In the datasets we compiled for AAPF and SBI, a positive label indicates the intersection of multiple identities and/or attributes. For AAPF, for instance, a positive example is one where there was police violence *and* the victim was a Black woman. For SBI, a positive example is one where the victim is Indigenous *and* a woman, girl, or two spirit person. With just a single, positive label, we leave it to the model to learn the complex decision boundary delineating examples with these intersecting attributes.

Moreover, treating all negative examples as equally irrelevant misses the ways in which they may actually be similar to the positive examples. For example, for AAPF, a case where police shot a Black man and a case where a white woman was killed by an intimate partner are both given the same negative label. However, the former shares the institutional violence and racism present in the positive cases, while the latter shares the root causes of sexism and patriarchy.

Annotating articles with multiple attributes allows us to build models that explicitly take into account the ways in which different systems of oppression manifest and interplay. For AAPF, we re-annotated each example with two labels: "police violence" and "feminicide". For SBI, we annotated articles with both "MMIWG2-related/Indigenous news" and "feminicide or missing victim". The breakdown of examples (after additional data collection and re-annotation) in the resulting datasets is shown in Tables 1 and 2. In both cases, the *intersection* of the two new annotations represents a "positive" case. This framing allows us to incorporate this prior knowledge of the domain into the model architecture rather than expecting the model to automatically learn the specific intersection of interest. We expand upon these modeling options in the following section.

## 5.3 Model Development

Given a context-specific set of examples and multiple annotations, we explored a few different modeling approaches to identify the intersectional subgroups of interest. In both datasets, there are two labels: one for whether the article is related to the class of violence we want to monitor (police violence / missing and murdered Indigenous people), and one for whether the victim is a woman or girl (or two spirit person, in the case of SBI). We want to identify cases where both labels are true, i.e. cases of Black police feminicides for AAPF and cases of missing and murdered Indigenous women/girls/two spirit people for SBI.

***Featurization.*** We tested two strategies to embed article full-texts: term frequency-inverse document frequency (TF-IDF) and TensorFlow (TF) Universal Sentence Encoder [15]. In TF-IDF, each article is converted to a word count vector, with values being normalized based on their frequency across the training dataset. Rare (< 5th percentile frequency) and common (> 95th percentile) words are excluded, along with stop words. Meanwhile, TF's sentence encoder is a Transformer-based model [75] trained on unsupervised text from Wikipedia and news corpuses, and supervised data from the Stanford NLI (SNLI) corpus. The model outputs general-purpose embeddings with a variety of higher-order text features, and has been shown to perform well on tasks such as identifying semantically similar sentences.

We acknowledge limitations of both these approaches. Since TF-IDF only relies on word counts, it has trouble disambiguating the victim, perpetrator, and other people mentioned in an article. For example, TF-IDF would yield a similar featurization for an article describing a female victim or an article describing a female perpetrator, since both of those articles would typically have high counts of she/her pronouns. Though the sentence encoder can theoretically distinguish these cases by accounting for interactions between words, it may have other flaws: e.g. (1) it is trained on a very general text distribution, so it may ignore or poorly handle the language specific to our intersectional article set; (2) it condenses each article into a 512-dim embedding vector, which may lose relevant semantics and context.

***Classifiers.*** As alluded to above, we went with the simplest models that would effectively discriminate our training set. We used logistic regression (LR) due to its quick training iteration times with our rapidly evolving datasets and limited computational resources. More complex decision trees or neural networks did not improve performance and tended to overfit, while a naive Bayes classifier did not perform as well as LR.

We tested three different modeling approaches to predict the intersectional labels which combined two annotations. These approaches are summarized below:

(1) Joint: In this approach, we train a single LR on the AND of the two labels. Positives are articles with both labels TRUE, while all others are negatives. Because this approach treats all victims not in the specific category of interest equally as negatives, Joint is equivalent to a single-label baseline.
(2) Hybrid: We train two LRs independently on all articles, one for the feminicide label and one for the police violence / Indigenous-related label. These two predictors' outputs are then combined into a single intersectional prediction by using the product of their probabilities (multiplying worked better than addition or other weighting schemes).
(3) Contextual Hybrid: This model is similar to Hybrid in that we multiply the predictions from two LRs, one for each label. However, the feminicide predictor is made contextual, in that we only train it on the articles where the auxiliary label is TRUE. For example, the contextual model for police feminicides works by training one LR on all articles to identify police violence, and combining it with another LR trained only on the police violence articles to distinguish police feminicides from non-feminicide police violence.

We trained classifiers in Python with the `scikit-learn` package [65]. The LR models use L2-regularization, with the strength optimized via 5-fold cross-validation.

### 5.4 Model evaluation

Our evaluation methodology consists of three main stages: 1) 5-fold cross-validation on our current datasets; 2) a monitoring phase where team members assess model performance on unseen, possibly out-of-distribution data in the real-world deployment context, and 3) an extended, participatory evaluation with the partner organization. The aim of this multi-level approach is to ensure that the models actually serve activists' needs in deployment, but also to build a degree of confidence in the models before requesting partner feedback, to avoid unintentionally over-burdening them.

In **Stage 1**, we compute internal validation performance by averaging across 5 training-validation splits of the dataset. In **Stage 2**, we deploy models to the Email Alerts system, using the same queries and media source configurations as each partner organization to reflect the actual deployment context. We then internally monitor the results returned by the system each day for approximately two weeks.

Our quantitative evaluation metrics for Stages 1 and 2 include the Area Under Precision-Recall and Receiver Operating Characteristic curves (AUPRC, AUROC), and Precision@K where K = 50 and 100. Precision@K measures how many of the articles ranked in the top K are indeed positives, where the ideal ratio is 1. This metric is especially relevant to the downstream use case, where we would want to surface as many relevant cases as possible for an activist that only has bandwidth to look at a limited set of articles.

We proceed to **Stage 3** once the results of Stages 1 and 2 indicate an improved set of models. In this stage — which is currently ongoing — activists incorporate the new models into their regular workflows. Each week, we check in to gauge their feedback, both qualitatively (through semi-structured discussion) and quantitatively (through a small set of survey questions focused on result relevance).

## 6 RESULTS

Here, we describe both quantitative and qualitative observations for Stages 1 and 2 of our evaluation. Stage 3 is currently ongoing, and is a longer-term evaluation in which each organization integrates the system into their workflow over a two-month period. While this is a necessary and important part of a feminist and participatory evaluation process, the analysis and results from this stage comprise a separate and significant contribution outside of the scope of this paper.

We also note that "success" can manifest in different ways beyond typical metrics — for example, in the extent to which trust is built with partner organizations, power and resources are shared, community is built, or future work is imagined [29]. In our project, beyond these three evaluative stages, we were humbled by both AAPF and SBI's willingness to extend their partnership with us, despite the initial tool not meeting their needs. Both participated as panel speakers in a community event our team organized for feminicide activists and spoke to the value of the collaboration [1].

### 6.1 Stage 1

After our additional data collection and annotation steps, we performed an internal quantitative evaluation of the different model architectures for both SBI and AAPF with 5-fold cross-validation.

We found that for the classifiers predicting the auxiliary label (e.g., police violence / Indigenous-related), TF-IDF featurization seemed to work better than the Universal Sentence Encoder embeddings. For example, for AAPF, the police violence predictor was able to identify police violence from words alone: many of these articles describe the officers involved and police-specific types of force, while they do not include words about domestic violence or other types of civilian assaults. Similarly, for SBI, cases explicitly framed as MMIWG2 typically include the exact phrase "missing and murdered," which can be identified effectively via word counts. Thus, a word count-based featurization (TF-IDF) was sufficient.

The reverse was true for the feminicide predictors, where the Universal Sentence Encoder was more effective than TF-IDF. We speculate that this is due to the relative complexity of this task, which involves disambiguating entities in the text, and the difficulty of using only word counts to so. The sentence embeddings may contain some additional information, e.g., to distinguish the perpetrator's gender from the victim's. There were fewer false positives of female perpetrators killing male victims when we used embeddings, which supports this hypothesis.

Therefore, for the Hybrid and Contextual Hybrid models for both organizations, we used TF-IDF featurization for the police violence and Indigenous-related predictors, and sentence embeddings for the feminicide predictors. Joint only uses one featurization, so we compared TF-IDF and embeddings and found that the latter worked better for both organizations. The Stage 1 results for AAPF and SBI are shown in Tables 3 and 4 .

For both organizations, Contextual Hybrid achieves the highest aggregate metrics (AUROC, AUPRC), as well as the highest Precision@100, which is a particularly relevant metric for an activist's use case (minimizing false positives in a finite number of top-scoring articles). Importantly, both hybrid models performed better than Joint, which is the baseline in which we train one model

| | *AUPRC* | *AUROC* | *Precision@50* | *Precision@100* |
|---|---|---|---|---|
| Joint | 0.881 | 0.913 | **1** | 0.97 |
| Hybrid | 0.902 | 0.931 | 0.98 | 0.96 |
| Contextual Hybrid | **0.921** | **0.944** | 0.98 | **0.99** |

**Table 3: AAPF model comparison.**

| | *AUPRC* | *AUROC* | *Precision@50* | *Precision@100* |
|---|---|---|---|---|
| Joint | 0.894 | 0.941 | **0.98** | 0.96 |
| Hybrid | 0.913 | 0.954 | 0.96 | 0.95 |
| Contextual Hybrid | **0.92** | **0.959** | **0.98** | **0.97** |

**Table 4: SBI model comparison.**

on a single label representing the intersectional cases of interest. The improved performance over Joint highlights that incorporating multiple annotations yields better intersectional performance.

Further, Contextual Hybrid's success over base Hybrid underscores the importance of *considering context*. Because the feminicide predictor in the former is trained only on articles related to the context (e.g., only police violence articles), we can interpret it as predicting the probability of feminicide *conditioned* on context. It both reaffirms and is able to utilize our prior knowledge that these intersectional cases manifest and are written about in unique ways.

## 6.2 Stage 2

For the next phase of evaluation, we launched projects on the Email Alerts system with the Contextual Hybrid models trained for both AAPF and SBI. For both projects, there were fewer articles returned overall when compared with the original model before targeted data collection (for which there were significant false positives).

***AAPF.*** While the articles returned were overall much more relevant than the previous, general model, many were still false positives, despite few false positives on our internal dataset. Specifically, of the 37 cases returned between 12/07/21 and 12/24/21, 12 described cases of Black women killed by the police. This finding reflects an inherent difficulty in logging feminicides, especially for an intersectional subgroup: though many Black women are unfortunately killed by police, these victims are a small fraction of the overall amount of violence that involves Black victims, women, and/or police. Additionally, Black women victims are systemically undercovered by most media outlets [78]. Compared to our relatively balanced training dataset, real-world news media displays extreme skew towards the "negative" article class, and hence it is nearly impossible to match internal metrics in deployment settings. While straightforward, this was an humbling realization to frame our expectations during practical evaluation.

Still, we tried to improve our model by learning from this stage. Many of the false positives made logical sense: for example, there were feminicides committed by civilians where the article described police arriving at the scene, or cases of female police officers killing Black male victims (such as female officer Kim Potter killing Daunte Wright, which had an ongoing trial during our observation period). We collected 30 such false positives and added them to our dataset, and retrained the hybrid models. Interestingly, their internal performance decreased, with Contextual Hybrid's AUPRC dropping by 6%, implying that these challenge examples were difficult for the model to adapt to. Despite lower metrics, there was some adaptation: when actually deployed, we observed fewer new false positives.

***SBI.*** The returned articles from 12/01/21 to 12/15/21 were overall much more relevant than the previous, general model, with 17 out of 20 related to MMIWG2, and 10 out of 20 describing specific cases of MMIWG2. The false positives that appeared included a few cases where Indigenous men or boys were killed, but female relatives were also involved or quoted in the article. Other false positives included articles where the MMIWG2 movement was referenced outside of the context of a specific feminicide case. As we iteratively improve our models through collecting additional context-specific positives and negatives, we imagine adding a representative set of such false positives to our dataset to target those model weaknesses, mirroring our process for AAPF.

Overall, for both organizations, this stage yielded important insights: (1) During internal validation, a worse-performing model on a more challenging dataset may yield more relevant articles in practice. This connects back to *considering context* in model evaluation as in model development, since a model that appears to perform well in a lab setting may perform worse once deployed in its real-world context, and conversely, a model that performs more poorly in a lab setting may actually yield strong results when deployed in situ. (2) Our models have the most trouble disambiguating the roles of different entities in an article, which suggests that we might want to explore other featurization options (such as including hand-crafted linguistic features in addition to a learned sentence embedding [80]) in future work.

# 7 DISCUSSION

In this paper, we describe the process of building an ML-based system with feminist and participatory methods from the start.

While prior approaches to addressing harmful ML systems have focused on post-hoc technical fixes or considering user input at discrete points in the development process, we instead pose the more forward-facing question of how to conceptualize and design ML systems in support of co-liberation. We consider a concrete case study in which we co-designed datasets and machine learning models to support the efforts of activists who collect and monitor data about feminicide. In particular, we focus on our efforts to create ML models to detect instances of feminicide from news media that work well not only for the majority cases, but also for groups that monitor specific, intersectional forms of gender-related violence.

Guided by the framework of data feminism, we demonstrate how intersectional feminist goals shaped each stage of our process, from problem conceptualization all the way to evaluation and deployment. For example, we highlight how the goal of *challenging power* led us to work with a diverse group of counterdata activists; how the goal of *considering context* motivated our iterative data collection process and CONTEXTUAL HYBRID model architecture; or how the goal of *embracing pluralism* influenced our multi-stage evaluation focused on practical relevance for each group. We find that the resulting models return more relevant results for two intersectional monitoring organizations than a general model that does not take context into account.

Through this process, we distill some practical lessons learned from approaching ML with an intersectional feminist lens. A commitment to intersectionality means that systems should work not only for the mainstream, majority use case, but also for those on the margins. In contrast to dominant values of speed and efficiency, this requires dedicating (possibly a lot of) additional time and resources towards these more specific use cases. Project plans can be developed from the start to anticipate this additional time, reduce the sense of urgency which often leads to overlooking marginal identities, and appreciate the opportunity to build trust and community with impacted groups [47]. In addition to taking longer, a truly participatory process necessarily must be iterative and ongoing. While we discuss the results of a particular set of models here, we also understand that they are imperfect, that the needs of our partner organizations may evolve, and that the data we deal with may also shift substantially as reporting around feminicide changes. We have already begun additional rounds of data collection, development and evaluation, and imagine this as an ongoing process — not in pursuit of a "perfect" model after which we declare the project finished, but towards a sustained, trusted collaboration in which we can continually support activist efforts.

Looking forward, we consider the interplay between generalizability and context-specificity. As part of the overarching Data Against Feminicide project, we are onboarding 20 more monitoring organizations into the Email Alerts system. We imagine that there are likely to be other groups for whom our existing set of models will not work well — in particular, those who focus on subcategories of gender-related violence that sit at the intersection of many forces of domination (e.g., trans violence, feminicides of sex workers, or Indigenous land defenders). The number of relevant cases reported in the news media for these subcategories is also likely to be relatively small compared to the (unfortunately) larger numbers of feminicide. While these forms of violence are important to understand on their own, and while we will likely need to build separate, context-specific models for each focus, many aspects of our approach are generalizable. For example, the iterative process of collecting context-specific positives (beginning with cases the groups have already identified) and context-specific negatives (based on iteratively identifying areas for which the model is underspecified) is generalizable to many contexts in which the positive cases of interest comprise a highly-specific type of example.

We also acknowledge the tensions brought up by our current approach. For example, while we annotated AAPF's data with "police violence" and "feminicide", we did not include a race annotation even though their focus is specifically on Black women. While technologies exist to infer race (e.g., from names or photos), they are often ethically fraught. As a result of excluding this information from the dataset, the model cannot learn the specific intersectional identity of interest. In the case of SBI, due to the difficulty of inferring Indigenous identity, we primarily used articles that were explicitly framed as an MMIWG2 case — which means our classifier may miss out on relevant articles that fall outside of this framing. More generally, we see an unavoidable tension in this work between classifying people with rigid categorization systems (e.g., race, gender) and the inherent fluidity of these categories [51, 71]. Staying with this tension is part of data feminism's commitment to *rethinking binaries and hierarchies*. It is also one of the reasons that our system locates the ultimate decision-making power to determine whether an article is relevant in the human activists themselves.

Finally, for scholars who may struggle to understand the focus on intersectional groups as the path towards universal liberation, we wish to end with this quote from the Combahee River Collective. They write: "If Black women were free, it would mean that everyone else would have to be free, since our freedom would necessitate the destruction of all the systems of oppression" [16]. In other words, intersectional analysis and action are of utmost importance to secure co-liberation — to free us all from the matrix of domination that harms us all.

## ACKNOWLEDGMENTS

We wish to acknowledge the Sovereign Bodies Institute, the African American Policy Forum, and all of the people and organizations interviewed in the Data Against Feminicide project for their time, labor, care and dedication to righting the balance of justice.

## REFERENCES

[1] Data Against Feminicide 2021. 2021. https://datoscontrafeminicidio.net/en/2021-edition/.
[2] Celia Amorós. 1990. Violencia contra las mujeres y pactos patriarcales. (1990).
[3] Isabelle Anguelovski. 2014. *Neighborhood as Refuge Community Reconstruction, Place Remaking, and Environmental Justice in the City.* https://doi.org/10.7551/mitpress/9780262026925.001.0001
[4] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 167–176.
[5] Fran Baum, Colin MacDougall, and Danielle Smith. 2006. Participatory action research. *Journal of Epidemiology & Community Health* 60, 10 (2006), 854–857. https://doi.org/10.1136/jech.2004.028662 arXiv:https://jech.bmj.com/content/60/10/854.full.pdf
[6] Ruha Benjamin. 2019. Race after technology: Abolitionist tools for the new jim code. *Social Forces* (2019).
[7] Janine Berg, Marianne Furrer, Ellie Harmon, Uma Rani, and M Six Silberman. 2018. Digital labour platforms and the future of work. *Towards decent work in*

*the online world. Genf: International Labour Organization ILO* (2018).
[8] Artigo 19 Brasil. 2018. Dados Sobre Feminicidio No Brasil. https://artigo19.org/wp-content/blogs.dir/24/files/2018/03/Dados-Sobre-Feminic%C3%ADdio-no-Brasil-.pdf
[9] Tone Bratteteig and Guri Verne. 2018. Does AI Make PD Obsolete? Exploring Challenges from Artificial Intelligence to Participatory Design. In *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial - Volume 2* (Hasselt and Genk, Belgium) *(PDC '18)*. Association for Computing Machinery, New York, NY, USA, Article 8, 5 pages. https://doi.org/10.1145/3210604.3210646
[10] Joy Buloamwini. 2022. *Facing the Coded Gaze with Evocative Audits and Algorithmic Audits.* PhD dissertation. Massachusetts Institute of Technology.
[11] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.
[12] Judith Butler. 2004. *Undoing gender.* Psychology Press.
[13] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (nov 2019), 24 pages. https://doi.org/10.1145/3359206
[14] CEPAL. 2018. Al menos 2.795 mujeres fueron víctimas de feminicidio en 23 países de América Latina y el Caribe en 2017. https://www.cepal.org/es/comunicados/cepal-al-menos-2795-mujeres-fueron-victimas-feminicidio-23-paises-america-latina-caribe
[15] 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 for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 169–174.
[16] Combahee River Collective. 1977. *A Black Feminist Statement.*
[17] P.H. Collins. 2000. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment.* Routledge.
[18] Patricia Hill Collins. 2019. *Intersectionality as critical social theory.* Duke University Press.
[19] Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need.* The MIT Press.
[20] Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.* (1989), 139.
[21] Kimberle Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review* 43, 6 (1991), 1241–1299. http://www.jstor.org/stable/1229039
[22] Caroline Criado-Perez. 2019. *Invisible Women: Data Bias in a World Designed for Men.* Abrams Press.
[23] Morgan Currie, Britt S Paris, Irene Pasquetto, and Jennifer Pierre. 2016. The conundrum of police officer-involved homicides: Counter-data in Los Angeles County. *Big Data & Society* 3, 2 (2016), 2053951716663566. https://doi.org/10.1177/2053951716663566 arXiv:https://doi.org/10.1177/2053951716663566
[24] Lina Dencik, Arne Hintz, and Jonathan Cable. 2016. Towards data justice? The ambiguity of anti-surveillance resistance in political activism. *Big Data & Society* 3, 2 (2016).
[25] Lina Dencik, Arne Hintz, Joanna Redden, and Emiliano Treré. 2019. Exploring data justice: Conceptions, applications and directions. *Information, Communication & Society* 22, 7 (2019), 873–881.
[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848
[27] Catherine D'Ignazio. 2023. *Counting Feminicide: Data Feminism in Action.* MIT Press.
[28] Catherine D'Ignazio, Isadora Cruxên, Ángeles Martinez, Mariel García-Montes, Helena Suárez Val, Silvana Fumega, Harini Suresh, and Wonyoung So. 2022. Feminicide & Counterdata Collection: Activist Efforts To Monitor And Challenge Gender-based Violence. *In submission* (2022).
[29] Catherine D'ignazio and Lauren F Klein. 2020. *Data feminism.* MIT press.
[30] Alice Driver. 2015. *More or less dead: Feminicide, haunting, and the ethics of representation in Mexico.* University of Arizona Press.
[31] Kristin Nicole Dukes and Sarah E Gaither. 2017. Black racial stereotypes and victim blaming: Implications for media coverage and criminal proceedings in cases of police violence against racial and ethnic minorities. *Journal of Social Issues* 73, 4 (2017), 789–807.
[32] Catherine D'Ignazio, Helena Suárez Val, Silvana Fumega, Harini Suresh, and Isadora Cruxên. 2020. Feminicide & machine learning: detecting gender-based violence to strengthen civil sector activism. *Mechanism Design for Social Good (MD4SG '20)* (2020).
[33] Jenny Rose Finkel, Trond Grenager, and Christopher D Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. 363–370.
[34] Rosa-Linda Fregoso and Cynthia Bejarano. 2010. Introduction: A cartography of feminicide in the Americas. In *Terrorizing Women*. Duke University Press, 1–42.
[35] Evgeniy Gabrilovich and Shaul Markovitch. 2009. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research* 34 (2009), 443–498.
[36] Mariel García-Montes. 2020. View source: making secure communication tools more USABLE. *XRDS: Crossroads, The ACM Magazine for Students* 26, 4 (2020), 16–19.
[37] Claudia Garcia-Moreno, Alessandra Guedes, Wendy Knerr, R Jewkes, S Bott, and S Ramsay. 2012. Understanding and addressing violence against women. *World Health Organization, Issue brief No. WHO/RHR/12.37)(S. Ramsay, Ed.)* (2012).
[38] Kristen Gilchrist. 2010. "Newsworthy" Victims? *Feminist Media Studies* 10, 4 (2010), 373–390. https://doi.org/10.1080/14680777.2010.514110 arXiv:https://doi.org/10.1080/14680777.2010.514110
[39] Joanne Gray and Alice Witt. 2021. A feminist data ethics of care for machine learning: The what, why, who and how. *First Monday* (2021).
[40] Mary L Gray and Siddharth Suri. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass.* Eamon Dolan Books.
[41] Aimi Hamraie and Kelly Fritsch. 2019. Crip technoscience manifesto. *Catalyst: Feminism, Theory, Technoscience* 5, 1 (2019), 1–33.
[42] Leif Hancox-Li and I Elizabeth Kumar. 2021. Epistemic values in feature importance methods: Lessons from feminist epistemology. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 817–826.
[43] Alex Hanna. 2017. MPEDS: Automating the generation of protest event data. (2017).
[44] Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14, 3 (1988), 575–599. http://www.jstor.org/stable/3178066
[45] Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.
[46] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017).
[47] Kenneth Jones and Tema Okun. 2001. White supremacy culture. *Dismantling Racism: A Workbook for Social Change* (2001).
[48] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and PM Krafft. 2020. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 45–55.
[49] Katherine A Keith, Abram Handler, Michael Pinkham, Cara Magliozzi, Joshua McDuffie, and Brendan O'Connor. 2017. Identifying civilians killed by police with distantly supervised entity-event extraction. *arXiv preprint arXiv:1707.07086* (2017).
[50] Tamil Kendall. 2020. *A Synthesis of Evidence on the Collection and Use of Administrative Data on Violence against Women: Background Paper for the Development of Global Guidance.* UN Women.
[51] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–22.
[52] Bogdan Kulynych, David Madras, Smitha Milli, Inioluwa Deborah Raji, Angela Zhou, and Richard Zemel. 2020. Participatory approaches to machine learning. In *International Conference on Machine Learning Workshop*.
[53] Marcela Lagarde y de los Ríos. 1996. *Género y feminismo: desarrollo humano y democracia.* Siglo XXI Editores México.
[54] Marcela Lagarde y de los Ríos. 2010. Preface: Feminist Keys for Understanding Feminicide: Theoretical, Political, and Legal Construction. In *Terrorizing Women: Feminicide in the Americas*. Duke University Press. https://doi.org/10.1215/9780822392644
[55] Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, Vol. 2. Citeseer, 1–49.
[56] Donald Martin Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. 2020. Participatory problem formulation for fairer machine learning through community based system dynamics. *International Conference on Learning Representations (ICLR)* (2020).
[57] Laurenellen McCann. 2015. Building Technology With, Not For Communities: An Engagement Guide for Civic Tech. https://medium.com/organizer-sandbox/building-technology-with-not-for-communities-an-engagement-guide-for-civic-tech-b8880982e65a
[58] Amanda Meng and Carl DiSalvo. 2018. Grassroots resource mobilization through counter-data action. *Big Data & Society* 5, 2 (2018), 2053951718796862. https://doi.org/10.1177/2053951718796862 arXiv:https://doi.org/10.1177/2053951718796862
[59] Laura Nader. 1972. Up the anthropologist: Perspectives gained from studying up. (1972).
[60] Jennifer C Nash. 2018. *Black feminism reimagined.* Duke University Press.

[61] Safiya Umoja Noble. 2018. *Algorithms of oppression.* New York University Press.

[62] Mimi Onuoha and Diana Nucera. 2018. A People's Guide to AI. (2018).

[63] James O'Malley. 2018. Captcha if you can: How you've been training AI for years without realizing it. Tech Radar. *URL https://bit. ly/37H1esA* (2018).

[64] Desmond U Patton, William R Frey, Kyle A McGregor, Fei-Tzin Lee, Kathleen McKeown, and Emanuel Moss. 2020. Contextual analysis of social media: The promise and challenge of eliciting context in social media posts with natural language processing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 337–342.

[65] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[66] Claudia Perlich, Brian Dalessandro, Troy Raeder, Ori Stitelman, and Foster Provost. 2014. Machine learning for targeted display advertising: Transfer learning in action. *Machine learning* 95, 1 (2014), 103–127.

[67] Washington Post. 2016. Fatal force [database]. https://www.washingtonpost.com/graphics/investigations/police-shootings-database/

[68] Jill Radford and Diana EH Russell. 1992. *Femicide: The politics of woman killing.* Twayne Publishers.

[69] Hal Roberts, Rahul Bhargava, Linas Valiukas, Dennis Jen, Momin M Malik, Cindy Sherman Bishop, Emily B Ndulue, Aashka Dave, Justin Clark, Bruce Etling, et al. 2021. Media Cloud: Massive Open Source Collection of Global News on the Open Web. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 1034–1045.

[70] Camilo Bernal Sarmiento, Miguel Lorente Acosta, Françoise Roth, and Margarita Zambrano. 2014. Latin American model protocol for the investigation of gender-related killings of women (femicide/feminicide). *New York: United Nations High Commissioner for Human Rights (OHCHR) and UN Women* (2014).

[71] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–35.

[72] M Sloane, E Moss, O Awomolo, and L Forlano. 2020. Participation is not a Design Fix for Machine Learning. *Participatory Approaches to Machine Learning* (2020).

[73] Susan Strega, Caitlin Janzen, Jeannie Morgan, Leslie Brown, Robina Thomas, and Jeannine Carriere. 2014. Never innocent victims: Street sex workers in Canadian print media. *Violence against women* 20, 1 (2014), 6–25.

[74] Jon Swaine, Oliver Laughland, Jamiles Lartey, and Ciara McCarthy. 2016. The Counted: people killed by police in the US. (2016).

[75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[76] Sandra Walklate, Kate Fitz-Gibbon, Jude McCulloch, and JaneMaree Maher. 2019. *Towards a global femicide index: Counting the costs.* Routledge.

[77] David Werner. 1998. *Nothing About Us Without Us: Developing Innovative Technologies For, By, and With Disabled Persons.* Healthwrights.

[78] Sherri Williams. 2016. # SayHerName: Using digital activism to document violence against black women. *Feminist media studies* 16, 5 (2016), 922–925.

[79] Frank Wood, April Carrillo, and Elizabeth Monk-Turner. 2019. Visibly unknown: Media depiction of murdered transgender women of color. *Race and Justice* (2019), 2153368719886343.

[80] Minghao Wu, Fei Liu, and Trevor Cohn. 2018. Evaluating the Utility of Hand-crafted Features in Sequence Labelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2850–2856.

[81] Ethan Zuckerman, J Nathan Matias, Rahul Bhargava, Fernando Bermejo, and Allan Ko. 2019. Whose Death Matters? A Quantitative Analysis of Media Attention to Deaths of Black Americans in Police Confrontations, 2013–2016. *International Journal of Communication* 13 (2019), 27.