Resources for Automated Identification of Online Gender-Based Violence

Citation for published version:
Abercrombie, G, Jiang, A, Gerrard-Abbott, P, Konstas, I & Rieser, V 2023, Resources for Automated Identification of Online Gender-Based Violence: A Systematic Review. in Y-L Chung, A Mostafazadeh Davani, D Nozza, P Rottger & Z Talat (eds), *7th Workshop on Online Abuse and Harms (WOAH).* Association for Computational Linguistics, pp. 170-186, 7th Workshop on Online Abuse and Harms 2023, Toronto, Ontario, Canada, 13/07/23. https://doi.org/10.18653/v1/2023.woah-1.17

Digital Object Identifier (DOI):
10.18653/v1/2023.woah-1.17

Link:
Link to publication record in Heriot-Watt Research Portal

Document Version:
Publisher's PDF, also known as Version of record

Published In:
7th Workshop on Online Abuse and Harms (WOAH)

Resources for Automated Identification of Online Gender-Based Violence: A Systematic Review

Gavin Abercrombie[1] and **Aiqi Jiang**[1,3] and **Poppy Gerrard-Abbott**[4,5]
and **Ioannis Konstas**[1,2] and **Verena Rieser**[1*]
[1]The Interaction Lab, Heriot-Watt University [2]Alana AI
[3]Computational Linguistics Lab, Queen Mary University of London
[4]School of Social and Political Science, University of Edinburgh [5]EmilyTest
{g.abercrombie, a.jiang, i.konstas, v.t.rieser}@hw.ac.uk
pgerrard@ed.ac.uk

Abstract

Online Gender-Based Violence (GBV), such as misogynistic abuse, is an increasingly prevalent problem that technological approaches have struggled to address. Through the lens of the GBV framework, which is rooted in social science and policy, we systematically review 63 available resources for automated identification of such language. We find the datasets are limited in a number of important ways, such as their lack of theoretical grounding and stakeholder input, static nature, and focus on certain media platforms. Based on this review, we recommend development of future resources rooted in sociological expertise and centering stakeholder voices, namely GBV experts and people with lived experience of GBV.

1 Introduction

We are in the midst of an 'epidemic of online abuse', which disproportionately affects women and minoritised groups and has worsened during and after the COVID-19 pandemic: 46% of women and marginalised gender identities such as transgender users experience gender-based online abuse, with non-binary people and Black and minority ethnic women at 50% (Glitch UK and EVAW, 2020).

In recent years, technology companies and computer science researchers have made efforts to automate the identification of hate speech and other toxic or abusive language, and have released datasets and resources for training machine classification systems (see e.g. Poletto et al., 2021; Vidgen and Derczynski, 2021). While some of these have focused on sexist and misogynistic abuse (e.g. Jiang et al., 2022; Zeinert et al., 2021), overall, systems still perform worse at detecting such instances, with high failure rates (Nozza et al., 2019).

In this review, we examine efforts at producing resources for automated content moderation through the lens of Gender-Based Violence (GBV). We particularly focus on the extent to which stakeholders, namely GBV experts and people with lived experience of GBV have been included in the design and production of these resources.

The GBV framework While there is a growing body of natural language processing (NLP) work purporting to address *sexism* and *misogyny*, these terms are often used imprecisely in the literature and dataset taxonomies. We advocate for the use of the term 'gender-based violence', which was first used by the United Nations to promote a comprehensive, umbrella theorisation of endemic violence and abuse (United Nations, 2021) arising from a gender stereotypic society of unequal gender orders and gender stratification (UN General Assembly, 1993). GBV is often non-linear[1] and overlapping, entailing hybrid behaviours of physical, digital, verbal, psychological, and sexual violence; implicit and explicit forms; and spanning multiple spaces, actors, and events–inclusive of numerous types of abuse and specialist focuses, such as coercive control, domestic violence, intimate partner violence, sexual harassment and stalking.

The concept has been broadened by the European Union to include online abuse (Dominique, 2021; Lomba et al., 2021) as GBV has come to be understood as affecting both online and offline life, manifesting in victims/survivors' communities, domestic, and occupational lives. Conceptualising GBV in a modern context shows how the framework has adapted to a digitised and globalised world, expanding and diversifying to contemporary types. Online forms of GBV, with a particular focus on 'cybersexism' and 'cybermisogny' include taking photographs and

[*]Now at Google DeepMind.

[1]'Non-linearity' refers to how the realities of GBV do not follow isolated incident trajectories of 'not victim'/victim/recovery. Victimisation is episodic, always mixing different forms, and happens multiple times across lifespans (it cross-cuts 'time and space') (Lindgren and Renck, 2008; Mouffe, 2013).

The 7th Workshop on Online Abuse and Harms (WOAH), pages 170–186
July 13, 2023

videos without consent, so-called 'revenge pornography' (or 'image-based abuse'), deepfakes, rape-supportive jokes and memes, cyberflashing, cyberstalking (including 'creeping'), cyberbullying, trolling, anti-feminist forums and bots targeting feminist content, social media-based harassment, grooming, threatening private messages, the dissemination of private information, catfishing and doxing (Get Safe Online, 2023; Glitch, 2022). As phenomena that are morphing, multi-pronged, and crossing the boundaries of multiple social worlds, modern GBV is more complex than ever and more challenging to regulate. Online GBV is of specific interest because it has distinct characteristics, namely that it is rising sharply and is mostly perpetrated by strangers (Amnesty International, 2017).

The GBV concept recognises that people of all genders are victimised by, perpetrate, uphold, and enable (gender) stereotypes and the systematic violence and abuse arising from them, occurring at the point(s) of situational power differentials and axes of difference. Spectrum-based and pluralistic, GBV is perpetrated by numerous people across boundaries of time and cultural sites, experienced in every level of social life, combining macro factors, such as patriarchal belief systems, meso factors such as institutional dismissal, and micro factors such as interpersonal relations (Public Health Scotland, 2021). The GBV framework has been recognised and strategically adopted by organisations such as the World Bank (2019), the World Health Organization (2020), and the Scottish Government (2016), among others. Its increasing take-up in policy-making at both supranational and national levels relates to the framework's exhaustive and inclusive approach, considering age, class, disability, geography, history, race, and socioeconomics.

Terminology As the framework is widely encompassing, GBV accounts for terms that are often used loosely and interchangeably in NLP literature, annotation schema and guidelines, which we clarify here. According to Manne (2017), **Sexism** 'consists in ideology that has the overall function of rationalising and justifying patriarchal social relations'. Sexism provides the underlying assumptions, beliefs, and stereotypes, as well as theories and narratives concerning gender differences that cause people to 'support and participate in patriarchal social arrangements'—and engage in misogynistic behaviour. **Misogyny**, on the other hand, consists of actions that serve to police and enforce those sexist norms and assumptions. As Manne (2017) puts it, misogyny is the "'law enforcement" branch of a patriarchal order'.

Our contributions In this paper, we reassess resources for automated abusive language identification through the GBV framework, paying particular attention to the conceptual strand dedicated to violence against women and girls (VAWG) in the form of (online) sexism and misogyny. We conduct a systematic review considering factors that are pertinent to stakeholders (i.e. people with lived experience of GBV and organisations that support them), such as stakeholder representation and data selection. We highlight gaps in currently available resources, and make recommendations for future dataset creation. Specifically, we address the following **Research questions**:

R1. How is GBV characterised?

R2. Who is represented in annotation of the data?

R3. From which platforms have the data been sourced?

R4. How has the data been sampled?

R5. Which languages are represented?

R6. During which time periods were the data created?

For motivation of these questions and analysis of the findings, see section 4. We create a new repository of resources for computational identification of GBV structured around the issues highlighted here. This is available at `https://github.com/HWU-NLP/GBV-Resources`.

2 Related work

In addition to the sociological and policy literature outlined in section 1, our methodology and research aims are informed by work from NLP and human-computer interaction in a number of areas.

GBV online A number of studies address computational analysis of aspects of GBV, such as the tone of news reports on incidents of rape and femicide (De La Paz et al., 2017; Minnema et al., 2022) and user engagement with GBV stories on social media (ElSherief et al., 2017; Purohit et al., 2016). However, we are not aware of prior work applying the framework to abusive language detection.

Abusive, hateful, and toxic language detection There are several reviews summarising work on detection of related but broader phenomena such as hate speech (e.g. Vidgen et al., 2019). In a survey of ethical issues surrounding automated content moderation, Kiritchenko et al. (2021) highlight the importance of engaging with stakeholders, considering annotator welfare and labelling disagreement—factors we also analyse in this online GBV review.

For hate speech detection resources, Poletto et al. (2021) present a systematic review of hate speech benchmark datasets, finding that the field lacks a common framework, that annotation schema and taxonomies are not systematically described, and that targeted sampling methodologies result in neglect of prevalent forms of abuse—issues we further examine and make recommendations on.

We draw heavily on Vidgen and Derczynski (2021), who systematically reviewed abusive language datasets and provide the hatespeechdata.com repository. While this comprehensive resource provides one of our search sources and many of the resources we review, we examine a number of factors it does not touch upon, such as the correspondence of annotation schemes to the GBV framework, and the levels of stakeholder participation.

Sexism and misogyny detection In recent years, there has been growing interest in developing datasets for the identification of phenomena related to sexism and misogyny as a separate task from more general abusive, hateful, offensive, or toxic language detection. This has included a number of shared tasks, such as EXIST (Rodríguez-Sánchez et al., 2021, 2022; Plaza et al., 2023), AMI (Fersini et al., 2018, 2022), SemEval-2019 Task 5 (Basile et al., 2019), and EDOS (Kirk et al., 2023).

For an earlier overview, Shushkevich and Cardiff (2019) surveyed the detection of misogynistic text, primarily on Twitter. They focus on approaches to technical aspects of automatic classification and performance measured on benchmark datasets. We are not aware of prior work that situates computational resources within a cohesive framework rooted in social science and policy, as we provide.

Stakeholder participation In this review, we focus on the extent to which stakeholders such as experts in and victims of GBV are included and consulted in the production of resources for its identification. Participatory design has a long history of being incorporated into projects in the field of human-computer interaction (e.g Muller and Kuhn, 1993). However, despite a handful of successful projects (e.g Birhane et al., 2022), the inclusion of stakeholders in NLP and AI design tends to remain superficial at best (Delgado et al., 2021).

3 Review methodology

In order to form a comprehensive picture of the available resources and to conduct a replicable and transparent review, we follow the systematic methodology of Moher et al. (2009). The search protocol is shown in Figure 1, and outlined below.

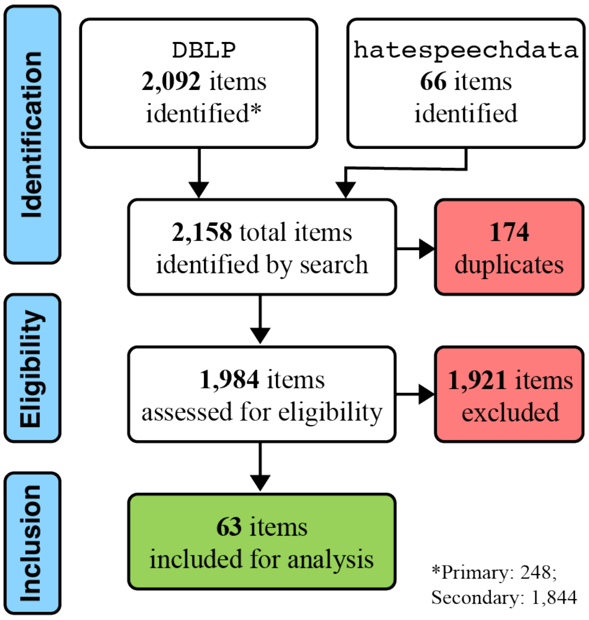

Figure 1: Flow diagram showing the phases of the selection of research items analysed in this review.

Databases Following a scoping study to establish coverage of GBV-related publications and datasets, we searched two databases: the DBLP Computer Science Bibliography[2] and hatespeechdata.com.[3] We found that these were sufficient to cover all papers published at typical NLP venues such as the ACL Anthology.[4]

Keyword selection We used the primary search keywords *misogyn**, *sexis**, and *"gender based violence"*. For DBLP, to capture publications that concern hate speech and abusive language more generally, but that include categories relevant to GBV, we also search using the secondary keywords *hate speech* | *detection* | *rhetoric*, *abuse*,

[2]https://dblp.org/
[3]https://hatespeechdata.com/
[4]https://aclanthology.org/

and *abusive* | *offensive* | *toxic language* | *speech*, which we developed from the results of our scoping study. Search using secondary terms is unnecessary in hatespeechdata.com, where all included entries concern hate speech and abusive language. To filter out irrelevant publications, we then search within the whole text results for our primary keywords. We also perform a manual search of hatespeechdata.com, adding items that describe general hate speech and abusive or toxic language datasets which include sexism, misogyny, or gender-based abuse as categories in their taxonomies. We conducted all searches on April 21^{st} 2023.

Eligibility criteria Table 1 shows the inclusion and exclusion criteria we applied. Two authors of this paper read the identified items applying the criteria, and cross checking agreement.

Include	Exclude
Describes a dataset designed and manually annotated for text classification of toxic language, hate speech, or related phenomena.	Describes a previously released dataset with no modifications (e.g. shared task system paper).
Data is from online sources such as social media and website comments.	Data is from other sources such as scripted TV shows.
GBV specified as target phenomena (e.g. 'misogyny', 'sexism').	Describes general toxic language dataset without fine-grained GBV concepts.

Table 1: Inclusion/exclusion criteria.

For items found in hatespeechdata.com, we directly apply the inclusion/exclusion criteria. For items retrieved from the DBLP, we first automatically select two groups of items for the first round of eligibility assessment: i) dataset description papers with keywords '*dataset*' / '*corpus*' in the title; ii) GBV-related papers with primary keywords mentioned in the whole text content. We then apply the criteria to manually check the remaining items.

Summary of included resources Following the systematic search process, we eventually include 63 relevant items for analysis in the review. These are shown in Table 2 along with summary statistics describing the resources. Of these, all but eight of the described datasets are currently available to download, while those described by Fersini et al. (2022) and Zeinert et al. (2021) require sign-up or email request to obtain access. Due to licensing and privacy issues, the majority of the resources sourced from Twitter include only the ID numbers of posts, which is likely to result in difficulties in retrieving their contents given elapsed time and changes in the accessibility of the platform's API.

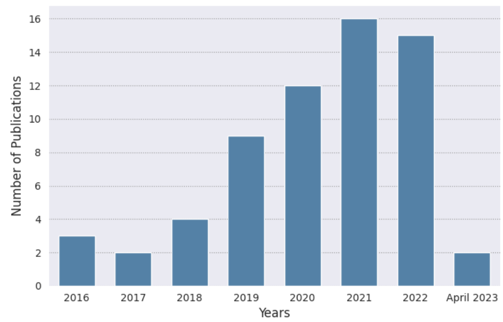

Figure 2: Publications per year up to April 2023.

Figure 2 shows the number of GBV detection resources over time, with relevant work first appearing in 2016 and increasing in number until 2022.[5]

4 Research questions and analysis

With this review, we synthesise information on the following aspects of the available resources:[6]

Characterisation of GBV Given the framework outlined in section 1, we investigate how GBV is characterised in the resources: what terminology is used to describe GBV (e.g. 'sexism', 'misogyny'), how these concepts are theorised, and how GBV fits into the datasets' taxonomies. Overall, we find that use of terminology is confused, and limited engagement with sociological theory.

We find that a large number of resources (28, 41.8%) name '*sexism*' as their target phenomena of interest. The majority of these describe this only superficially as, for example 'hate against women' (Guellil et al., 2021b) or 'hate speech including sexism' (Yadav et al., 2023). However, several 'sexism' resources are grounded—to greater or lesser extents—in sociological theory. Sharifirad and Jacovi (2019) cite Mills (2008)' definitions of sexism, concluding that 'sexism seems to be a relatively complex concept which is [not] easy to define', while Jha and Mamidi (2017) contrasts 'benevolent' and 'hostile' forms of sexism as described by Glick and Fiske (1997). The most comprehensive grounding of sexism in theory is provided by Samory et al. (2021), who compile a 'sexism codebook' based on nearly 30 psychological

[5]For further statistics and visualisations, see Appendix A.

[6]Detailed notes on the resources with respect to these dimensions are provided in the repository at `https://github.com/HWU-NLP/GBV-Resources.git`.

Publication/source reference	Conceptualisation of target phenomena	Media platform	Level of analysis	Language	Size	Avail-ability
Al-Hassan and Al-Dossari (2022)	*Sexism* as category	Twitter	Post	Arabic	11,000	✗
Almanea and Poesio (2022)	*Misogyny, Sexism*	Twitter	Post	Arabic	964	✓
Alsafari et al. (2020)	*Gender-based hate* as category	Twitter	Post	Arabic	5361	✓
Anzovino et al. (2018)	*Misogyny*	Twitter	Post	English	4,454	✓
Assenmacher et al. (2021)	*Sexism*	Rheinische Post	Post	German	85,000	✓
Basile et al. (2019)	*Women* as target	Twitter	Post	English, Spanish	19,600	✓
Bhattacharya et al. (2020)	*Misogyny*	Facebook, Twitter, YouTube	Post	Bangla, English, Hindi	25,000+	✓
Borkan et al. (2019)	*Gender identity (female, male, transgender, non-binary)*	Online comment forums	Comment	English	450,000	✓
Bosco et al. (2018)	*Gender issues* as category	Facebook, Twitter	Post	Italian	8,000	✓
Cercas Curry et al. (2021)	*Sexism, Sexual harassment*	Dialogue systems, Facebook	Conversation	English	4,185	✓
Chiril et al. (2021)	*Sexism*	Twitter	Post	French	9,282	✓
Chiril et al. (2019)	*Sexism*	Twitter	Post	French	3,085	✗
Chiril et al. (2020)	*Sexism*	Twitter	Post	French	12,000	✓
Chung and Lin (2021)	*Sex (gender, sexual orientation, or gender identity)* as category	PTT (Taiwanese bulletin board)	Post, comment	Chinese	1000 posts, 121,344 com.	✓
Das et al. (2022)	*Gender as target*	Twitter	Post	Bengali	10,178	✓
El Ansari et al. (2020)	*Discrimination and Violence Against Women*	Twitter	Post	Arabic	1,690	✗
Fanton et al. (2021)	*Women* as target	Semi-synthetic text	Post	English	5,003	✓
Fersini et al. (2018)	*Misogyny*	Twitter	Post	English, Spanish	8,115	✓
Fersini et al. (2020)	*Misogyny*	Twitter	Post	Italian	7,961	✓
Fersini et al. (2022)	*Misogyny*	9GaG, Imgur, Knowyourmeme, Reddit, Twitter	Meme	English	15,000	✓
García-Díaz et al. (2021)	*Misogyny, Violence against Women*	Twitter	Post	Speanish	7,682	✓
Gomez et al. (2020)	*Sexism*	Twitter	Post	English	149,823	✓
Gong et al. (2021)	*Gender* as target	YouTube	Comment, sentence	English	11,540	✗
Grosz and Conde-Cespedes (2020)	*Sexism*	Twitter, related quotes collection	Post, quote	English	1,100+	✓
Guellil et al. (2021a)	*Sexism*	YouTube	Comment, reply	Arabic	3,798	✗
Guest et al. (2021)	*Misogyny*	Reddit	Post (header and body)	English	6,567	✓
Hewitt et al. (2016)	*Misogyny*	Twitter	Post	English	5,500	✗
Hoefels et al. (2022)	*Sexism*	Twitter		Romanian	39,245	✓
Ibrohim and Budi (2019)	*Gender* as category	Twitter	Post	Indonesian	13,169	✓
Jha and Mamidi (2017)	*Sexism (benevolent vs hostile)*	Twitter	Post	English	712	✓
Jiang et al. (2022)	*Sexism*	Sina Weibo	Post, comment	Chinese	8,969	✓
Jeong et al. (2022)	*Gender & sexual orientation* as target	NAVER news, YouTube	Post	Korean	40,429	✓
Kennedy et al. (2020)	*Gender identity* as target, *Sexist speech*	Twitter, Reddit, YouTube	Comment	English	39,565	✓
Kennedy et al. (2022)	*Gender identity* as target	Gab	Post	English	27,665	✓
Kirk et al. (2023)	*Sexism*	Gab; Reddit	Post, comment	English	20,000	✓
Kumar et al. (2018)	*Gendered Aggression*	Facebook, Twitter	Post, comment	Hindi-English	39,000	✓
Kwarteng et al. (2022)	*Misogyny (misogynoir)*	Twitter	Post	English	4,532	✓
Lee et al. (2022)	*Gender* as category	Korean news site	Comment	Korean	109,692	✓
Leite et al. (2020)	*Misogyny*	Twitter	Post	(Brazilian) Portuguese	21,000	✓
Lynn et al. (2019)	*Misogyny*	Urban Dictionary	Post	English	2,285	✓
Mathew et al. (2021)	*Women* as target	Twitter, Gab	Words, phrases, posts	English	20,148	✓
Mulki and Ghanem (2021)	*Misogyny*	Twitter	Post	Arabic (Levantine)	6,550	✓
Mollas et al. (2022)	*Gender* as category	Reddit, Youtube	Post, comment	English	1,072	✓
Moon et al. (2020)	*Gender bias* as category	NAVER entertainment news	Comment	Korean	9,381	✓
Ousidhoum et al. (2019)	*Gender* as target	Twitter	Post	Arabic, English, French	13,000	✓
Petrak and Krenn (2022)	*Misogyny*	Austrian news	Comment	German	6,600	✗
Plaza et al. (2023)	*Sexism*	Twitter, Gab,	Post	English, spanish	9,400	✓
de Pelle and Moreira (2017)	*Sexism*	Globo (news)	Post	(Brazilian) Portuguese	1,250	✓
Rizwan et al. (2020)	*Sexism*	Twitter	Post	Roman Urdu	10,041	✓
Rodríguez-Sánchez et al. (2020)	*Sexism*	Twitter	Post	Spanish	3,600	✓
Rodríguez-Sánchez et al. (2021)	*Sexism*	Gab, Twitter	Post	English, Spanish	11,345	✓
Rodríguez-Sánchez et al. (2022)	*Sexism*	Gab, Twitter	Post	English, Spanish	12,403	✓
Romim et al. (2022)	*Gender* as category	Facebook, TikTok, YouTube	Post, comment	Bangla	50,281	✓
Samory et al. (2021)	*Sexism*	Twitter	Post	English	91	✓
Sharifirad and Jacovi (2019)	*Sexism*	Twitter	Post	English	3,240	✓
Sharifirad and Matwin (2019)	*Sexism*	Twitter	Post	English		✓
Strathern and Pfeffer (2022)	*Misogyny*	Twitter	Post	English	266,579	✓
Talat (2016)	*Sexism*	Twitter	Post	English	4,033	✓
Talat and Hovy (2016)	*Sexism*	Twitter	Post	English	16,000	✓
Toosi (2019)	*Sexism*	Twitter	Post	English	31,961	✓
Vidgen et al. (2021)	*Gender: women & Gender: minorities* as targets	Synthetic text	Post	English	41,255	✓
Yadav et al. (2023)	*Sexism* as a category	Twitter	Post	Arabic, English, French, German, Hindi, Spanish	497,660	✗
Zeinert et al. (2021)	*Misogyny*	Twitter, Facebook, Reddit	Post	Danish	279,000	✓

Table 2: Summary of included resources for automated identification of GBV-related phenomena.

scales including *Attitudes toward Women* (Spence and Helmreich, 1972), *Neosexism* (Tougas et al., 1995), and *Gender-Roles Attitudes* (García-Cueto et al., 2015). They also bemoan the 'lack of definitional clarity' in prior work on automated sexism detection.

19 (28.4%) of the resources are constructed with *gender*-based abuse as one of several *categories* or *targets* of more general hate speech. These are variously described as 'gender bias' (Moon et al., 2020), 'gender issues' (Bosco et al., 2018), or to include female, male, transgender, and non-binary genders (Borkan et al., 2019). The latter is similar in approach to the eight resources in which gender is conceived as one of various *targets*. Inclusion in gender as a target ranges from 'women' (Basile et al., 2019; Fanton et al., 2021; Mathew et al., 2021); to separation of 'gender: women' from 'gender: minorities' (Vidgen et al., 2021); to 'women, men, non-binary or third gender, transgender women, transgender men, transgender (unspecified)' (Kennedy et al., 2020), the latter identifying these groups as those protected in U.S. law.

16 (23.9%) of the resources characterise the target phenomenon as '*misogyny*'. Almanea and Poesio (2022) ground this only in prior computer science literature, describing misogynistic language as 'a category which overlap[s] with sexism towards women'. Petrak and Krenn (2022) explicitly conflate sexism and misogyny, but provide the disclaimer that their guidelines 'are not meant as an accurate abstract definition', but rather to assist annotators in making judgements. García-Díaz et al. (2021) delineate online misogyny into several categories including 'violence against relevant women', where 'relevant' signifies known targets of abuse. Anzovino et al. (2018) and Mulki and Ghanem (2021) consider language used in 'cybermisogyny', as outlined by Poland (2016). The latter also characterises misogyny as 'hatred of or contempt for women', citing feminist sociology and media studies (Moloney and Love, 2018) and the U.S. Constitution (Nockleby, 2000). Strathern and Pfeffer (2022) provide the most comprehensive overview of misogyny, comparing, among other sources, definitions from feminist philosophy (Allen, 2022), digital media studies (Ostini and Hopkins, 2015), and gender studies (Megarry, 2014), and devise a taxonomy based on these as well as computer science resources.

Despite its widespread adoption in policymaking (see section 1), we do not find any existing resources rooted in the GBV framework.

Annotators Most datasets for supervised machine learning are annotated by small numbers of anonymous crowdworkers (Vidgen and Derczynski, 2021), biasing the labelled data towards the opinions, world views, and lived experiences of those people who happen to work on the crowdsourcing platforms. Rottger et al. (2022) describe a scale of annotation scenarios ranging from highly *prescriptive* to *descriptive*, where the former attempts to induce annotators to follow a defined schema, while the latter seeks to elicit their individual and potentially conflicting points of view. There is a growing movement to recognise, that for many tasks, there may be no single 'ground truth', different judgements may be equally valid, or preservation of minority perspectives should be facilitated (Abercrombie et al., 2022; Aroyo and Welty, 2015; Plank, 2022). In the following, We report on who and how many annotators are represented, their expert or stakeholder knowledge, the level of training and/or supervision, and the guidelines and instructions with which they work. We examine these resources through the lenses of data *perspectivism* (Cabitza et al., 2023),[7] *participatory design* (Delgado et al., 2021; Muller et al., 2021) and *design justice* (Costanza-Chock, 2020), reporting on the extent to which different points of view are represented and the levels at which stakeholders are included as participants in decision making.

Due to the psychological harm working with abusive language can cause and its potential to traumatise victims (Kirk et al., 2022; Shmueli et al., 2021), we also assess the annotator welfare measures reportedly taken in constructing these resources.

Overall, we find that engagement with stakeholders is limited, minority annotator perspectives are usually not preserved, and comprehensive annotator welfare measures are unusual.

Representation: Reporting of *who* undertook dataset annotation is patchy, with only nine resources accompanied by a full data statement or annotator information to a similar degree of detail (Assenmacher et al., 2021; Cercas Curry et al., 2021; Das et al., 2022; Guest et al., 2021; Ibrohim and Budi, 2019; Leite et al., 2020; Kirk et al., 2023;

[7]See also the Perspectivist Data Manifesto: `https://pdai.info/`

Zeinert et al., 2021). From the information that *is* provided, we find that 16 (25%) of the datasets were annotated by crowdworkers, and 19 (30%) by people at various levels of academia ranging from the authors and other researchers to undergraduate students. The term 'expert' is used loosely, and refers variously to Gender Studies students (Cercas Curry et al., 2021; Chiril et al., 2020, 2021), people the authors provided some form of training to (Guest et al., 2021), 'experienced moderators' (Petrak and Krenn, 2022), or is not explicitly defined at all (Rodríguez-Sánchez et al., 2022; Vidgen et al., 2021). Where we understand the 'experts' in question to potentially be stakeholders, they are described as 'non-activist feminists' (Jha and Mamidi, 2017), 'feminist and anti-racism activists' Talat (2016), or Gender Studies students. Only Vidgen et al. (2021) report on whether their annotators have themselves been victims of online abuse, and we do not find evidence of the authors engaging with GBV-focused organisations to ensure victims are represented.

Data perspectivism: We find only six datasets (10%) released with multiple labels preserved (Cercas Curry et al., 2021; Hoefels et al., 2022; Kennedy et al., 2020; Kirk et al., 2023; Leite et al., 2020; Talat, 2016), with the others providing only aggregated labels, hence losing any potentially informative minority judgements.

Annotator welfare: Very few publications report any measures taken to ensure annotator welfare. Those that do follow welfare guidelines by Kennedy et al. (2020) (Strathern and Pfeffer, 2022); Vidgen et al. (2019) (Vidgen et al., 2021); the ACL Code of Ethics (Lee et al., 2022); Kirk et al. (2022) (Kirk et al., 2023); and Rivers and Lewis (2014) (Das et al., 2022). Despite the fact that any research with human subjects (including annotators) requires approval by an Institutional Review Board (IRB) (particularly when dealing with potentially upsetting material) (Shmueli et al., 2021), only two papers reports their studies having passed ethical review (Cercas Curry et al., 2021; Jeong et al., 2022).

Platforms While GBV is prevalent in all online spaces, most NLP research tends to collect data from freely accessible social media sources such as Twitter and Facebook. We ask: for which platforms are datasets available, and what is the modality of the data (i.e. text or multi-modal)? We find that the resources are very heavily skewed towards textual data from Twitter.

The majority of GBV resources are sourced from social media such as Twitter, Reddit, and Gab (a platform known for its right-wing user base). Twitter is by far the most accessible platform that provides an API and more lenient policies for gathering and disseminating data, with almost half of the available datasets (51.8%) being obtained exclusively or in combination with other sources from it. Reddit (7.1%) and Gab (7.1%) are also widely sourced with relatively lax moderation policies for user-generated content. Other popular platforms for procuring GBV datasets include Youtube (8.2%), Facebook (5.9%), and news website (7.1%) And around 34.9% of resources collect data from mixed sources.

Almost all the resources directly collect user-generated content online, except for Vidgen et al. (2021)'s set of human-generated synthetic data that mimics real-world social media posts, and another employing a semi-synthetic collection approach by iteratively refining a generative language model to create new samples that experts review and/or post-edit (Fanton et al., 2021). The only multi-modal datasets are those of Fersini et al. (2022), who released a set of misogynistic memes, and Gomez et al. (2020), who collected and labelled tweets that include text and images for attacks on different communities including the label '*sexist*'.

Overall, we find no evidence that researchers' choices of which media platforms to target are driven by stakeholders' requirements.

Data sampling A strong motivation for engaging stakeholders in annotation is that, following *standpoint theory* (Harding, 1991), in many cases, those with relevant lived experience are the only people capable of recognising subtle, implicit abuse such as stereotypes and micro-aggressions. However, it is recognised that commonly used data sampling techniques do not account for this type of language, meaning that it is sparsely represented in datasets (Vidgen and Derczynski, 2021).

Indeed, we find that, where reported, nearly all the resources (20) have been sampled using keyword search. Those that have not, were generally gathered from specific sources known to consist predominantly of text espousing hateful ideologies such as Gab (Kennedy et al., 2022; Mathew et al., 2021; Plaza et al., 2023; Rodríguez-Sánchez et al., 2022) or particular forums on Reddit (Fersini et al., 2022; Guest et al., 2021; Kennedy et al., 2020; Kirk et al., 2023; Mollas et al., 2022). Alternative strategies are to collect items on topics that attract toxic comments (Bhattacharya et al., 2020), items already flagged by community moderators (Assenmacher et al., 2021), or those addressed to people known to be victims of online abuse (Basile et al., 2019; Fersini et al., 2022; García-Díaz et al., 2021; Mulki and Ghanem, 2021; Strathern and Pfeffer, 2022; Yadav et al., 2023). Only Lee et al. (2022) rely on random selection to produce a more realistic but sparse data representation, while Zeinert et al. (2021) explore a range of sampling techniques in an effort to obtain a balanced representation of positively labelled (i.e. misogynistic) examples.

Languages As NLP research is heavily skewed towards English (Bender, 2009; Hovy and Prabhumoye, 2021), negatively affecting its ability to benefit diverse communities, we report on the languages represented in the available resources.

The resources cover a total of 16 languages, the vast majority of which are Indo-European (49 datasets, 77.8%). Specifically, most available resources are exclusively in English (26, 41.3%), followed by Spanish (8, 12.7%), Arabic (8, 12.7%), and French (5, 7.9%). There are also nine multilingual datasets covering a variety of languages including Arabic, French, German, Hindi, Italian, and Spanish, all of which include English as one of the languages. Overall, coverage of non-English languages is poor, with only one dataset even for a language as widely spoken as Chinese (Jiang et al., 2022).

Temporality While language use evolves, new societal events occur, and abusers use creative ways to circumvent content moderation (Talat et al., 2017), NLP datasets are usually collected over a specific time frame, limiting the ability of systems to make correct predictions on new instances (Kiela et al., 2021). We report on the time frames and scales over which the datasets were collected and whether they are static or dynamic.

25 (39.7%) of the datasets do not report collection dates. Time spans of those that do are presented in Figure 3[8]. The majority were collected in the past five years. The variation in the time frames covered by GBV datasets could be due to a variety of factors, such as the release of new platforms or tools for data collection, the emergence of new GBV-related topics, and changes in the policy or accessibility of social media platforms. The fact that Twitter is the most commonly used platform for data collection, as previously mentioned in the analysis of platforms, could be one factor in the time spans distribution. Twitter's popularity, user activity, and high volume of user-generated content may make it easier for researchers to collect data over shorter time frames. And the distribution of time frames is also likely influenced by factors such as the scope of GBV data and the size of the datasets.

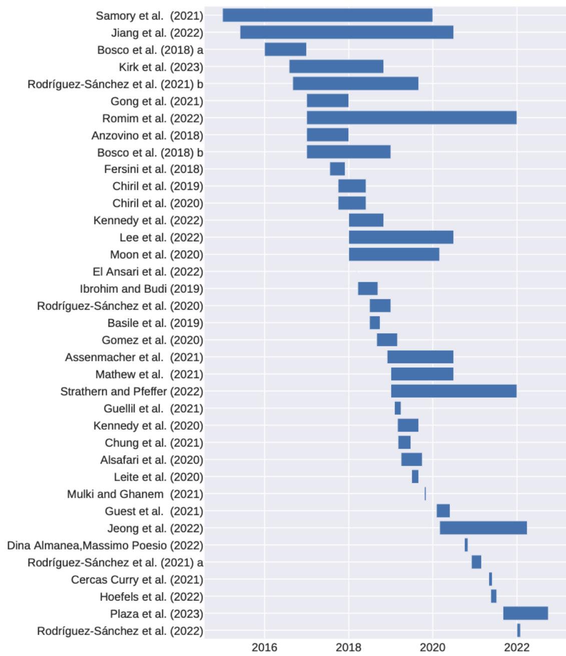

Figure 3: Data time spans. Those labeled *a*/*b* are data subsets from the same resource but different platforms and periods.

All but one of the resources are collected on a static time scale, with only one gathered dynamically in a human-in-the-loop setting (Vidgen et al., 2021). Current classification systems are commonly trained on these static datasets over fixed time frames, which has negative implications for their effectiveness, generalisability, and robustness in identifying instances of GBV in real-time.

5 Discussion and recommendations

This review has uncovered several limitations in the available resources and the approaches of NLP researchers towards constructing them. We summarise these and make future recommendations.

[8]For space, we exclude Lynn et al. (2019) (collected 1999-2006) and show Samory et al. (2021) (2008-2019) from 2015.

Conceptualisation With a couple of exceptions (e.g. Samory et al., 2021; Strathern and Pfeffer, 2022), the phenomena targeted in the reviewed resources are not clearly defined or strongly rooted in theory or expertise from outside computer science. Similar observations have been made for operationalisation of related concepts, such as bias and stereotypes (Blodgett et al., 2021), and value alignment (Irving and Askell, 2019).
Recommendation: Resource creators should collaborate with social scientists to ground them in expert knowledge of the target phenomena. We advocate for the use of GBV as a framework, which encompasses several facets currently operationalised in different ways by computer science researchers. It recognises how all forms of online abuse affect people of every gender both online and off, and has been widely adopted by policymakers.

Stakeholder participation Parker and Ruths (2023) propose that computer scientists should:

> *stop thinking about online hate speech as something requiring methods, and start thinking about it as something that demands solutions. This change — treating hate speech less like a task and more like the real-world problem it is — would orient CS research towards the concerns of other stakeholders, and thus begin the collaborative pursuit toward a safe Internet.*

However, we find little evidence of such a paradigm shift having occurred when it comes to designing these resources, with stakeholder participation limited to the recruitment of loosely defined 'expert' annotators—where it occurs at all.
Recommendations: Resource development projects should, as far as possible, strive to include stakeholders from the outset by including representatives in research teams. Stakeholder participation should be integrated throughout development, and is especially important in the design of taxonomies, guidelines, and at annotation, when judgements about what constitutes GBV are made. Due to the risks involved, annotator welfare should be prioritised by following guidelines such as those of Kirk et al. (2022), and IRB approval sought before any data collection. In documenting resources, authors should provide full data statements or similar (e.g. Bender and Friedman, 2018; Díaz et al., 2022), and, to preserve minority voices, dataset releases should includenon-aggregated labels (Prabhakaran et al., 2021).

Data collection Media data for these resources is not sourced from diverse sources, with the majority from Twitter, the choice of which does not appear to be driven by stakeholders. Furthermore, as the datasets are static in nature, their relevance as reference sources for automated classification decays over time; and, due to data sampling methods, positively labelled (i.e. abusive) examples are skewed towards the more explicit forms of online GBV.
Recommendation: There is a great need for the development of new methods to surface the diversity of GBV found online. One solution is to create platforms to which victims of abuse and bystanders can submit examples. This could facilitate creation of improved resources on many of the limiting dimensions we outline in this review: dynamic datasets to which new examples are regularly added; stakeholder participation in data and platform selection and labelling; and inclusion of implicit and subtle examples of GBV, as well as multimedia data.

Limitations and ethical considerations

We use a systematic review methodology in order to provide a reproducible and objective snapshot of the current research situation. However, we acknowledge that the choices made (such as search repositories and eligibility criteria) may not have captured every existing relevant resource. We aim to regularly update the repository of GBV resources at `https://github.com/HWU-NLP/GBV-Resources` and open it to submissions via push requests in order to provide a dynamic and comprehensive record.

Following D'Ignazio and Klein (2020), we acknowledge that this research is influenced by the positionalities of its authors. To situate our perspective, we are four Computer Science and one Social Science academic researchers working in public institutions in Europe. Three of us identify as women and two as men, and we are of European and Asian nationalities. This work forms part of a project conducted in partnership with charitable organisations that work on combating GBV and supporting its victims.

In this paper, we make a number of recommendations that complicate typical NLP resource creation workflows, and could have the unintended consequence of dissuading researchers from working on these problems. However, we appreciate that interdisciplinary work is difficult to instigate, organise, and carry out, and that it is not usually motivated by typical academic or industry reward structures. Our intention is to point out practical ways in which resource development can be improved and to encourage researchers to move towards more participatory solutions.

Acknowledgements

We would like to thank the anonymous reviewers for all their valuable comments and suggestions. Gavin Abercrombie, Aiqi Jiang, Poppy Gerrard-Abbott, Ioannis Konstas, and Verena Rieser were supported by the EPSRC project 'Equally Safe Online' (EP/W025272/1 and EP/W025493/1), and Gavin Abercrombie and Verena Rieser were also supported by the EPSRC project 'Gender Bias in Conversational AI' (EP/T023767/1). Verena Rieser was also supported by a Leverhulme Trust Senior Research Fellowship (SRF/R1/201100). Aiqi Jiang was supported by China Scholarship Council (CSC Funding, No. 201908510140).

References

Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma, editors. 2022. *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*. European Language Resources Association, Marseille, France.

Areej Al-Hassan and Hmood Al-Dossari. 2022. Detection of hate speech in Arabic tweets using deep learning. *Multimedia systems*, 28(6):1963–1974.

Amy Allen. 2022. Feminist Perspectives on Power. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2022 edition. Metaphysics Research Lab, Stanford University.

Dina Almanea and Massimo Poesio. 2022. ArMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291, Marseille, France. European Language Resources Association.

Safa Alsafari, Samira Sadaoui, and Malek Mouhoub. 2020. Hate and offensive speech detection on Arabic social media. *Online Social Networks and Media*, 19:100096.

Amnesty International. 2017. Amnesty reveals alarming impact of online abuse against women. Accessed: 2023-05-09.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on Twitter. In *Natural Language Processing and Information Systems*, pages 57–64, Cham. Springer International Publishing.

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

Dennis Assenmacher, Marco Niemann, Kilian Müller, Moritz Seiler, Dennis M Riehle, and Heike Trautmann. 2021. Rp-mod&RP-crowd: Moderator-and crowd-annotated German news comment datasets. In *NeurIPS Datasets and Benchmarks*.

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

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

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

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168, Marseille, France. European Language Resources Association (ELRA).

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? Opportunities and challenges for participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of*

The 2019 World Wide Web Conference, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the EVALITA 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9. CEUR.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2021. "be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2833–2844, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patricia Chiril, Farah Benamara Zitoune, Véronique Moriceau, Marlène Coulomb-Gully, and Abhishek Kumar. 2019. Multilingual and multitarget hate speech detection in tweets. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, pages 351–360, Toulouse, France. ATALA.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. An annotated corpus for sexism detection in French tweets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1397–1403, Marseille, France. European Language Resources Association.

I Chung and Chuan-Jie Lin. 2021. TOCAB: A dataset for Chinese abusive language processing. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 445–452.

Sasha Costanza-Chock. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. MIT Press.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022. Hate speech and offensive language detection in Bengali. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 286–296, Online only. Association for Computational Linguistics.

Meliza De La Paz, Maria Regina Estuar, and John Noel Victorino. 2017. Discovering conversation spaces in the public discourse of gender violence: a comparative between two different contexts. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 376–383. The National University (Phillippines).

Rogers Prates de Pelle and Viviane P Moreira. 2017. Offensive comments in the Brazilian web: A dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2021. Stakeholder participation in AI: Beyond "add diverse stakeholders and stir". In *Proceedings of the Human-Centered AI workshop at NeurIPS 2021*.

Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2342–2351, New York, NY, USA. Association for Computing Machinery.

Catherine D'Ignazio and Lauren Klein. 2020. *Data Feminism*. The MIT Press.

Fontaine-Lepage Dominique. 2021. Combating Gender-Based Violence: Cyber violence. European added value assessment study, EESC: European Economic and Social Committee.

Oumayma El Ansari, Zahir Jihad, and Mousannif Hajar. 2020. A dataset to support sexist content detection in Arabic text. In *Image and Signal Processing*, pages 130–137, Cham. Springer International Publishing.

Mai ElSherief, Elizabeth Belding, and Dana Nguyen. 2017. #NotOkay: Understanding Gender-Based Violence in social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):52–61.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. AMI@ EVALITA2020: Automatic misogyny identification. In *EVALITA*.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at IberEval 2018. *Ibereval@ sepln*, 2150:214–228.

Eduardo García-Cueto, Francisco Javier Rodríguez-Díaz, Carolina Bringas-Molleda, Javier López-Cepero, Susana Paíno-Quesada, and Luis Rodríguez-Franco. 2015. Development of the gender role attitudes scale (GRAS) amongst young Spanish people. *International Journal of Clinical and Health Psychology*, 15(1):61–68.

José Antonio García-Díaz, Mar Cánovas-García, Ricardo Colomo-Palacios, and Rafael Valencia-García. 2021. Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114:506–518.

Get Safe Online. 2023. Online abuse. https://www.getsafeonline.org/personal/articles/online-abuse/. Accessed 2023-05-07.

Peter Glick and Susan T. Fiske. 1997. Hostile and benevolent sexism. *Psychology of Women Quarterly*, 21(1):119–135.

Glitch. 2022. Violence Against Women & Girls code of practice. Accessed: 2023-05-09.

Glitch UK and EVAW. 2020. The ripple effect: COVID-19 and the epidemic of online abuse.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Hongyu Gong, Alberto Valido, Katherine M. Ingram, Giulia Fanti, Suma Bhat, and Dorothy L. Espelage. 2021. Abusive language detection in heterogeneous contexts: Dataset collection and the role of supervised attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14804–14812.

Dylan Grosz and Patricia Conde-Cespedes. 2020. Automatic detection of sexist statements commonly used at the workplace. In *Pacific Asian Conference on Knowledge Discovery and Data Mining (PAKDD), Wokshop (Learning Data Representation for Clustering) LDRC*, Singapour, Singapore.

Imane Guellil, Ahsan Adeel, Faiçal Azouaou, Mohamed Boubred, Yousra Houichi, and Akram Abdelhaq Moumna. 2021a. Sexism detection: The first corpus in Algerian dialect with a code-switching in Arabic/ French and English. *CoRR*, abs/2104.01443.

Imane Guellil, Faical Azouaou, Fodil Benali, and Hachani Ala-Eddine. 2021b. ONE: Toward ONE model, ONE algorithm, ONE corpus dedicated to sentiment analysis of Arabic/Arabizi and its dialects. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 236–249, Online. Association for Computational Linguistics.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Sandra Harding. 1991. *Whose science? Whose knowledge?: Thinking from women's lives*. Cornell University Press.

Sarah Hewitt, T. Tiropanis, and C. Bokhove. 2016. The problem of identifying misogynist language on Twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, page 333–335, New York, NY, USA. Association for Computing Machinery.

Diana Constantina Hoefels, Çağrı Çöltekin, and Irina Diana Mădroane. 2022. CoRoSeOf - an annotated corpus of Romanian sexist and offensive tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2269–2281, Marseille, France. European Language Resources Association.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

Geoffrey Irving and Amanda Askell. 2019. AI safety needs social scientists. *Distill*.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada. Association for Computational Linguistics.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. SWSR: A Chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs Jr., Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Cardenas, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, Yong Zhang, Beth Meyerowitz, and Morteza Dehghani. 2022. Introducing the Gab hate corpus: Defining and applying hate-based rhetoric to social media posts at scale. *Language Resources & Evaluation*, 56:79–108.

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

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

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

Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and presenting harmful text in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 497–510, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Joseph Kwarteng, Serena Coppolino Perfumi, Tracie Farrell, Aisling Third, and Miriam Fernandez. 2022. Misogynoir: Challenges in detecting intersectional hate. *Social Network Analysis and Mining*, 12(1):166.

Jean Lee, Taejun Lim, Heejun Lee, Bogeun Jo, Yangsok Kim, Heegeun Yoon, and Soyeon Caren Han. 2022. K-MHaS: A multi-label hate speech detection dataset in Korean online news comment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3530–3538, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Maria Scheffer Lindgren and Barbro Renck. 2008. Intimate partner violence and the leaving process: Interviews with abused women. *International Journal of Qualitative Studies on Health and Well-being*, 3(2):113–124.

N. Lomba, C. Navarra, and M. Fernandes. 2021. Combating Gender–Based Violence: Cyber violence. European added value assessment study, European Parliament.

Theo Lynn, Patricia Takako Endo, Pierangelo Rosati, Ivanovitch Silva, Guto Leoni Santos, and Debbie Ging. 2019. Urban dictionary definitions dataset for misogyny speech detection.

Kate Manne. 2017. *Down Girl: The Logic of Misogyny*. Oxford University Press.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Jessica Megarry. 2014. Online incivility or sexual harassment? Conceptualising women's experiences in the digital age. *Women's Studies International Forum*, 47:46–55.

Sara Mills. 2008. *Language and Sexism*. Cambridge University Press.

Gosse Minnema, Sara Gemelli, Chiara Zanchi, Tommaso Caselli, and Malvina Nissim. 2022. Dead or murdered? predicting responsibility perception in femicide news reports. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1078–1090, Online only. Association for Computational Linguistics.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G. Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine*, 151(4):264–269. PMID: 19622511.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*.

Mairead Eastin Moloney and Tony P. Love. 2018. Assessing online misogyny: Perspectives from sociology and feminist media studies. *Sociology Compass*, 12(5):e12577. E12577 SOCO-1299.R2.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Chantal Mouffe. 2013. Feminism, citizenship, and radical democratic politics. In *Feminists Theorize the Political*, pages 387–402. Routledge.

Hala Mulki and Bilal Ghanem. 2021. Let-mi: An Arabic Levantine Twitter dataset for misogynistic language. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 154–163, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Michael Muller, Christine T. Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, and Casey Dugan. 2021. Designing ground truth and the social life of labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Communications of the ACM*, 36(6):24–28.

John T. Nockleby. 2000. Hate speech, volume 1. In Leonard W. Levy and Kenneth L. Karst, editors, *Encyclopedia of the American Constitution (2nd ed.)*. Macmillan.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, page 149–155, New York, NY, USA. Association for Computing Machinery.

Jenny Ostini and Susan Hopkins. 2015. Online harassment is a form of violence. *The Conversation*, 8:1–4.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. 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)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Sara Parker and Derek Ruths. 2023. Is hate speech detection the solution the world wants? *Proceedings of the National Academy of Sciences*, 120(10):e2209384120.

Johann Petrak and Brigitte Krenn. 2022. Misogyny classification of German newspaper forum comments.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Laura Plaza, Jorge Carrillo-de Albornoz, Roser Morante, Enrique Amigó, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2023. Overview of EXIST 2023: sEXism Identification in Social NeTworks. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*, pages 593–599. Springer.

Bailey Poland. 2016. *Haters: Harassment, abuse, and violence online*. University of Nebraska Press.

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.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Public Health Scotland. 2021. Public Health Scotland.

Hemant Purohit, Tanvi Banerjee, Andrew Hampton, Valerie L. Shalin, Nayanesh Bhandutia, and Amit Sheth. 2016. Gender-based violence in 140 characters or fewer: A #BigData case study of Twitter. *First Monday*, 21(1).

Caitlin M. Rivers and Bryan L. Lewis. 2014. Ethical research standards in a world of big data. *F1000Research*, 3:38.

Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. 2020. Hate-speech and offensive language detection in Roman Urdu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2512–2522, Online. Association for Computational Linguistics.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on Twitter data. *IEEE Access*, 8:219563–219576.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of EXIST 2021: sEXism Identification in Social NeTworks. *Procesamiento del Lenguaje Natural*, 67:195–207.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Adrián Mendieta-Aragón, Guillermo Marco-Remón, Maryna Makeienko, María Plaza, Julio Gonzalo, Damiano Spina, and Paolo Rosso. 2022. Overview of EXIST 2022: sEXism Identification in Social NeTworks. *Procesamiento del Lenguaje Natural*, 69:229–240.

Nauros Romim, Mosahed Ahmed, Md Saiful Islam, Arnab Sen Sharma, Hriteshwar Talukder, and Mohammad Ruhul Amin. 2022. BD-SHS: A benchmark dataset for learning to detect online Bangla hate speech in different social contexts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5153–5162, Marseille, France. European Language Resources Association.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. "Call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In *ICWSM*, pages 573–584.

Scottish Government. 2016. Equally Safe: Scotland's strategy for preventing and eradicating violence against women and girls. Strategy plan, Scottish Government.

Sima Sharifirad and Alon Jacovi. 2019. Learning and understanding different categories of sexism using convolutional neural network's filters. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 21–23, Florence, Italy. Association for Computational Linguistics.

Sima Sharifirad and Stan Matwin. 2019. When a tweet is actually sexist. A more comprehensive classification of different online harassment categories and the challenges in NLP. *arXiv preprint arXiv:1902.10584*.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, Online. Association for Computational Linguistics.

Elena Shushkevich and John Cardiff. 2019. Automatic misogyny detection in social media: A survey. *Computación y Sistemas*.

Janet T. Spence and Robert Helmreich. 1972. The attitudes toward women scale: An objective instrument to measure attitudes toward the rights and roles of women in contemporary society. *Catalog of Selected Documents in Psychology*, 2.

Wienke Strathern and Juergen Pfeffer. 2022. Identifying different layers of online misogyny.

Zeerak Talat. 2016. Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zeerak Talat, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Talat and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Ali Toosi. 2019. Twitter sentiment analysis. https://www.kaggle.com/datasets/arkhoshghalb/twitter-sentiment-analysis-hatred-speech. Accessed: 2023-04-28.

Francine Tougas, Rupert Brown, Ann M Beaton, and Stéphane Joly. 1995. Neosexism: Plus ça change, plus c'est pareil. *Personality and social psychology bulletin*, 21(8):842–849.

UN General Assembly. 1993. Declaration on the elimination of violence against women. un general assembly resolution 48/104 assembly. Resolution, United Nations.

United Nations. 2021. 'endemic violence against women cannot be stopped with a vaccine' says who chief. `https://news.un.org/en/story/2021/03/1086812`. Accessed: 2023-06-07.

Bertie Vidgen and Leon Derczynski. 2021. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15:1–32.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019.

Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

World Bank. 2019. Gender-Based Violence (Violence Against Women and Girls). Accessed: 2023-05-09.

World Health Organization. 2020. WHO announced as a global leader of the generation equality action coalition on ending gender-based violence. https://www.who.int/news/item/01-07-2020-Equality-Action-Coalition-endinggender-based-violence.

Ankit Yadav, Shubham Chandel, Sushant Chatufale, and Anil Bandhakavi. 2023. LAHM : Large annotated dataset for multi-domain and multilingual hate speech identification.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197, Online. Association for Computational Linguistics.

A Figures of Analysis

We present visualisations of resource statistics in Figures 4, 5, 6, 7, and 8.

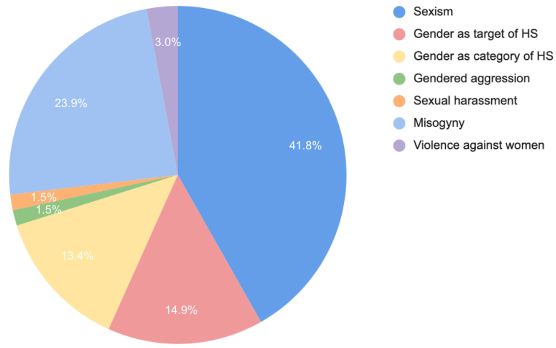

Figure 5: The distribution of characterisation of GBV.

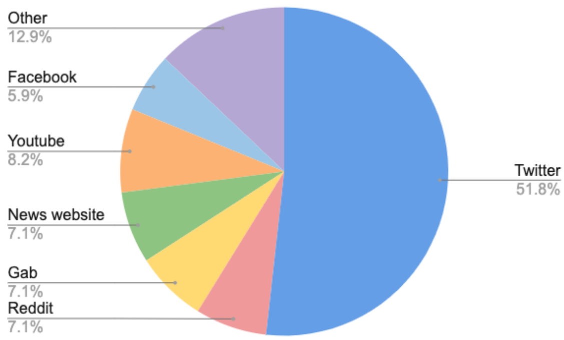

Figure 6: The distribution of platforms for GBV data collection.

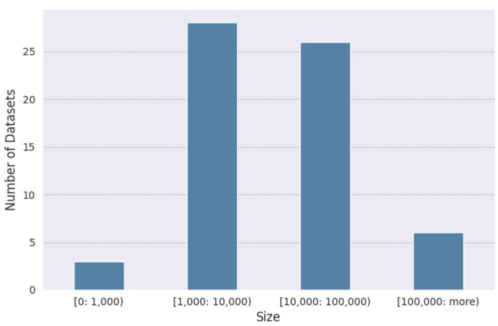

Figure 4: The distribution of GBV dataset sizes.

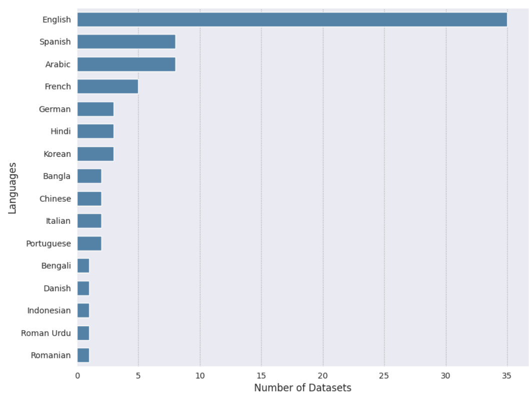

Figure 7: Number of GBV datasets across languages, including numbers if the language in multilingual datasets.

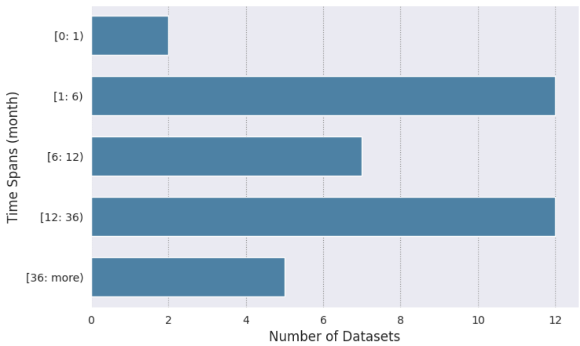

Figure 8: The distribution of time spans in GBV resources, excluding resources that are not reported collection time.