From Languages to Geographies: Towards Evaluating Cultural Bias in Hate Speech Datasets

Manuel Tonneau[1, 2, 3], **Diyi Liu**[1], **Samuel Fraiberger**[2, 3, 4], **Ralph Schroeder**[1], **Scott A. Hale**[1,5], **Paul Röttger**[6]
[1]University of Oxford, [2]World Bank, [3]New York University,
[4]Massachusetts Institute of Technology, [5]Meedan, [6]Bocconi University

Abstract

Perceptions of hate can vary greatly across cultural contexts. Hate speech (HS) datasets, however, have traditionally been developed by language. This hides potential cultural biases, as one language may be spoken in different countries home to different cultures. In this work, we evaluate cultural bias in HS datasets by leveraging two interrelated cultural proxies: language and geography. We conduct a systematic survey of HS datasets in eight languages and confirm past findings on their English-language bias, but also show that this bias has been steadily decreasing in the past few years. For three geographically-widespread languages—English, Arabic and Spanish—we then leverage geographical metadata from tweets to approximate geo-cultural contexts by pairing language and country information. We find that HS datasets for these languages exhibit a strong geo-cultural bias, largely overrepresenting a handful of countries (e.g., US and UK for English) relative to their prominence in both the broader social media population and the general population speaking these languages. Based on these findings, we formulate recommendations for the creation of future HS datasets.

1 Introduction

Far from the idyllic image of social media connecting people, increasing social cohesion, or letting everyone have an equal say, harmful content including hate speech (HS) has become rampant online (Vidgen et al., 2019) and has been linked to social unrest, hate crimes, and even deaths (Banaji et al., 2019; Müller and Schwarz, 2021).

To counter this phenomenon, a mature body of research has developed annotated datasets for automatic HS detection (Vidgen and Derczynski, 2020). Past work, however, has highlighted systematic gaps and biases in HS datasets (Park et al., 2018; Davidson et al., 2019; Wiegand et al., 2019; Nejadgholi and Kiritchenko, 2020; Wich et al., 2020).

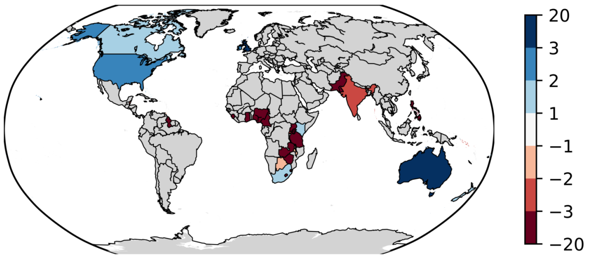

Figure 1: Geographical representativeness of author population of English hate speech datasets. A positive value N (negative value $-N$) indicates that a country is N times more (less) represented in English hate speech datasets relative to the global English-speaking population.

In particular, HS datasets exhibit a strong language bias, with the vast majority of datasets developed for English (Poletto et al., 2021). This focus on English, and more generally on languages, when developing HS datasets creates a risk of cultural blindness. Indeed, while certain languages, such as Basque, Icelandic or Yoruba, are highly indicative of a certain cultural context, others, such as English, are present across cultures. Yet, understanding the cultural context of a statement is crucial to determine whether it is hateful (Aroyo et al., 2019). Statements may be perceived as hateful in one culture but not in another (Lee et al., 2023b), even within the same language (Lee et al., 2023a). For instance, the term "Paki" is used as a neutral abbreviation for Pakistani in Pakistan whereas it is a racial slur in the UK. Despite the importance of the cultural context in the study of HS, the cultural origin of HS datasets remains largely unclear.

In this work, we aim to bridge this gap by answering the following research question: **To what extent are HS datasets culturally biased?** We operationalize cultural bias by measuring the representation of two cultural proxies in HS datasets: (a) language, and (b) geo-cultural contexts (de Rosa et al., 2018), defined as the combination of a language and a country. We first conduct a systematic survey of HS datasets in eight widely-spoken languages: Arabic, English, French, German, Indonesian, Portuguese, Spanish and Turkish. We confirm past findings on their English-language bias but also show that this dominance has been steadily decreasing in the past few years, with other languages such as Arabic catching up. We then depart from the traditional language-level analysis and situate our analysis in geo-cultural contexts. We focus on three geographically-widespread languages—English, Arabic and Spanish—and on Twitter, the main data source for HS datasets. We leverage geographical metadata from the annotated tweets in the datasets to infer the locations of their authors and find that HS datasets for these languages predominantly represent authors from a handful of countries (the US and UK for English, Chile and Spain for Spanish, and Jordan for Arabic). We also find that such countries are largely overrepresented in HS datasets compared to their prominence in both the broader social media population and the general population speaking these languages. We identify two main factors to explain the lack of representativeness of HS datasets: the lack of representativeness of Twitter itself as well as the sampling decisions made by authors. For the latter, we observe that non-uniform geographic sampling is typically intentional for Arabic and Spanish, motivated by a focus on specific geo-cultural contexts. In contrast, we find that such non-uniform sampling is commonly disregarded when compiling English HS datasets, which systematically lack information on the geographical origin of both data and annotators, hiding potential mismatches and ignoring the diversity of English speakers online. Based on these findings, we formulate recommendations for the creation of future HS datasets. Overall, our main contributions are:

1. A systematic survey of 75 HS datasets in eight languages (Arabic, English, French, German, Indonesian, Portuguese, Spanish and Turkish), revealing a persistent, though diminishing, dominance of English (§3).
2. Evidence of geo-cultural bias in existing HS datasets for three geographically-widespread languages: English, Arabic and Spanish (§4).
3. Preprocessed HS corpora for the eight surveyed languages and code for geocoding to stimulate research in this area.[1]

2 Background

2.1 Languages and Geographies as Interrelated Cultural Proxies

Language has historically played a pivotal role in cultural identity (Collins, 1999) and can be a good proxy for culture when a certain language is spoken only by a specific cultural group (e.g., Basque). Yet, some languages, such as English, Arabic or Spanish, have transcended cultural boundaries through human mobility, colonization, and imperialism. Such global adoption means that people who share a common language may come from diverse cultural backgrounds. These cultural differences also have online implications, whereby social media communities tend to form around both a common language and geography rather than just a common language (Mekacher et al., 2024). To take into account such differences, we use both language and geo-cultural contexts in our analysis of cultural bias. Cross-language bias measures how well different languages are represented, while geo-cultural contexts capture the representation of geographic locations, taking into account the cultural characteristics of a population, such as a common language (de Rosa et al., 2018).

2.2 Cultural Biases in NLP

The drastic progress in NLP tasks over the past decade can be partially attributed to the growing availability of large text corpora (Raffel et al., 2020), used to train language models. Yet, past work shows that these corpora are largely composed of English-language content (Joshi et al., 2020; Holtermann et al., 2024; Zhao et al., 2024), containing smaller amounts and lower-quality content for other widely spoken languages (Kreutzer et al., 2022). Adding to such language biases, past work has uncovered geographic biases in NLP corpora, where represented dialects and topics disproportionately originate from the Minority World (Graham et al., 2014b, 2015; Dodge et al., 2021). Driven by the necessity to include social factors in language modeling (Hovy and Yang, 2021), an emerging body of scholarship has developed approaches to include geographical information in language representation (Bamman et al., 2014; Rahimi et al., 2017; Hovy and Purschke, 2018; Kulkarni et al., 2021; Hofmann et al., 2022). Despite these efforts, recent language models still suffer from cultural biases, mirroring views largely aligned with Western, Educated, Industrialized,

[1] https://github.com/manueltonneau/hs-survey-cultural-bias

Rich and Democratic (WEIRD) individuals (Atari et al., 2023; Naous et al., 2023; Manvi et al., 2024). In order to mitigate such biases, it is crucial to document their presence in training and evaluation corpora, especially for culturally-sensitive tasks like HS detection (Baider, 2020).

2.3 Biases in Hate Speech Datasets

Past work has highlighted several biases in HS datasets. Many such biases can be linked to the subjectivity and demographics of annotators (Al Kuwatly et al., 2020), including racial bias (Davidson et al., 2019; Sap et al., 2019), gender bias (Park et al., 2018), and political bias (Wich et al., 2020). Other biases are related to the way such datasets are constructed, resulting in a large overrepresentation of the hateful class as well as certain topics and users (Dixon et al., 2018; Davidson et al., 2019; Wiegand et al., 2019; Nejadgholi and Kiritchenko, 2020). Despite the extent of this scholarship, little attention has been given to cultural bias in HS corpora. The most recent widely-cited and large-scale survey of HS resources does point to an English-language bias (Poletto et al., 2021) and a dominance of Twitter as a data source, which is known to be skewed towards certain geo-cultural contexts.[2] Also, Arango Monnar et al. (2022) point out that Spanish HS datasets are largely developed in the national context of Spain, motivating tailored approaches to other Spanish-speaking contexts such as Chile. Finally, past work highlights the cultural sensitivity of HS, uncovering country-specific offensive words (Ghosh et al., 2021) as well as disparities in cross-cultural HS annotations (Lee et al., 2023a), stereotype definition (Bhutani et al., 2024) and cross-dialect HS detection performance (Castillo-lópez et al., 2023) for a given language. To the best of our knowledge, our work is the first to systematically investigate cultural bias in HS datasets.

3 Language Bias in Hate Speech Datasets

We start our analysis of cultural bias at the language-level, as some languages are specific to single cultural contexts. We conduct a systematic survey of HS datasets in eight languages with a large presence on social media platforms: Arabic, English, French, German, Indonesian, Portuguese, Spanish and Turkish.

Language	Twitter only	Twitter + other	Other	Synthetic	**Total**
English	12	3	10	4	**29**
Arabic	11	0	0	1	**12**
Spanish	6	0	0	1	**7**
German	2	1	2	2	**7**
Turkish	5	0	1	0	**6**
French	3	0	1	2	**6**
Portuguese	3	0	1	1	**5**
Indonesian	2	0	1	0	**3**

Table 1: Number of available hate speech datasets by language and data source

3.1 Survey Approach

To identify HS datasets, we rely on three data sources. First, we inspect the Hate Speech Data Catalogue[3] (Vidgen and Derczynski, 2020) and find 80 candidate datasets for our languages of interest. Second, we inspect the datasets listed in the latest survey of HS datasets (Poletto et al., 2021) and find 20 additional candidate datasets that are not listed in the HS Data Catalogue. Finally, we conduct a Google search for each language and inspect the links of the first three result pages in each case, adding 43 candidate datasets that are neither in the HS Data Catalogue nor listed by Poletto et al. (2021). From those 143 unique datasets, we keep only the datasets that fit the following three criteria:

1. The dataset is documented, meaning it is attached to a research paper or a README file describing its construction.
2. The dataset is either publicly available or could be retrieved after contacting the authors.
3. The dataset focuses on HS, defined broadly as “any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor” (UN, 2019).

We provide additional details on the surveying in the Appendix (§A).

3.2 Results

Out of the 143 aforementioned datasets, we identify 75 available datasets that meet our three criteria for the eight languages of interest. We provide a breakdown in terms of language and data source in Table 1 as well as the number of datapoints by language (Table 4) and a complete list of the datasets for each language (§A.2) in the Appendix.

[2] https://datareportal.com/essential-twitter-stats

[3] https://hatespeechdata.com/

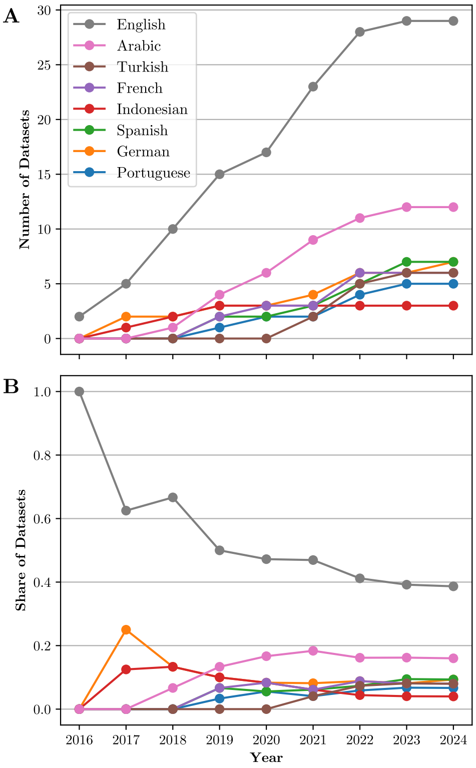

Figure 2: (A) Number of hate speech datasets per language over time (B) Share of hate speech datasets for the 8 languages of interest over time

Language and data source We find that English is the most common language in terms of HS detection resources, representing 39% of all available corpora and 41% of all annotated datapoints for our eight languages of interest. We also find that Twitter is by far the most common data source across languages. This is particularly the case for Arabic, with 92% of corpora originating from Twitter, followed by Spanish (86%) and Turkish (83%). Additionally, we find that some languages are particularly affected by a lack of data availability. For instance, 50% of identified Indonesian datasets and 38% of identified Portuguese datasets could not be retrieved (see Appendix §A.3 for more details).

Temporal dynamics To understand the dynamics of HS detection resource creation across languages, we further present the number of datasets per language over time as well as the language-level share of all datasets over time (Figure 2). We find that while English has dominated other languages in terms of the number of datasets over time, its dominance in terms of share of all HS datasets has steadily declined over the years, going from 100% of all datasets for the eight languages of interest in 2016 to 39% in 2023. In parallel, languages such as Arabic have been catching up. Such growth in corpus availability points towards a broadening of research that aims to address the multilingual nature of HS.

4 Geo-Cultural Bias in Hate Speech Datasets

While such language-level analysis is crucial to uncover gaps in existing resources and motivate the development of resources for under-served languages, it cannot account for and may hide potential large differences in resources between countries with a common language. In this section, we investigate the extent of geo-cultural bias in HS datasets, approximating geo-cultural contexts as a combination of one language and one country. For this purpose, we leverage the rich geographical metadata of tweets to map posts and their authors to a country location. We focus on three geographically widespread languages—English, Arabic and Spanish—for which the HS detection resources mostly emanate from Twitter (Table 1).

4.1 Author Location Inference

We use tweet geographical metadata to infer the country location of tweets' authors.

Information sources While there is a plethora of available information to infer user location from, from self-reported location to geocoordinates, timezone and linguistic features of tweets, each of these features has weaknesses. Profile locations are only available for a fraction of users, may contain vague locations (e.g., Planet Earth) or non-geographic text (Hecht et al., 2011) and may not always match with the device location (Graham et al., 2014a). Geo-coordinates are even rarer (1–2% of all tweets according to Twitter[4]) and may point to locations other than a user's home location, for instance if the user is travelling. Further, linguistic features have proven to not be a good proxy for location (Graham et al., 2014a) and while dialectal variability may inform on a user's location (Jurgens et al., 2017), language identification methods incorporating this variability are scarce beyond English. Finally, timezones of different countries with a common language may overlap. While acknowledging these limitations, we decide to use exclusively the two features that are equally available across languages to infer user country location: the geocoordinates of tweets and the self-reported user profile location.

[4] https://developer.twitter.com/en/docs/tutorials/advanced-filtering-for-geo-data

	English	Arabic	Spanish
Share of all Twitter datasets with retrieved tweet IDs	9/15	6/11	4/6
# unique tweets with tweet IDs	155,974	456,892	24,752
# tweets with tweet IDs and retrieved geographical metadata	64,057	251,178	14,684
# tweets with inferred author country location	50,116	247,408	13,273

Table 2: Summary statistics of data collection and author location inference

Geographical data collection Tweet geocoordinates and user profile location are usually not shared in public HS datasets for privacy reasons. In this context, we first attempt to retrieve the tweet IDs of all Twitter datasets for English, Arabic and Spanish by either collecting them when they are publicly available or contacting the authors to request access. We are able to retrieve tweet IDs for 9 English (Waseem, 2016; Waseem and Hovy, 2016; Jha and Mamidi, 2017; ElSherief et al., 2018a,b; Vidgen et al., 2020; Mathew et al., 2021; Samory et al., 2021; Toraman et al., 2022), 6 Arabic (Albadi et al., 2018; Alsafari et al., 2020; Alshaalan and Al-Khalifa, 2020; Mulki and Ghanem, 2021b; Ameur and Aliane, 2021; Ahmad et al., 2023) and 4 Spanish (Pereira-Kohatsu et al., 2019; García-Díaz et al., 2021; Arango Monnar et al., 2022; Vásquez et al., 2023) Twitter HS datasets. We then use the Twitter API to retrieve the tweet author self-reported location and the tweet geocoordinates if available. Out of all tweet IDs, we are able to retrieve some geographical information, that is either the tweet's author self-reported location, geocoordinates or both, for 64,057 (41%) English, 251,178 (55%) Arabic and 14,684 (59%) Spanish tweets. We report the main statistics of data collection in Table 2.

Country inference We infer the country of origin of a tweet author in two ways. First, in case a tweet is geotagged, we assign the country location of the geotag to its author. In cases where a user has no geotagged tweets but has a self-reported location, we use geocoding to convert the reported location to a country location. Specifically, we use the Google Geocoding API as Graham et al. (2014a) demonstrate it performs better than other geocoding tools. In case a tweet has no available geographical metadata, we are not able to infer its author country location and do not analyse it further.

Geocoding evaluation For each language, we sample 50 unique user locations geocoded within a country and have one author annotate whether this country match is correct. We also sample 50 unique user locations that could not be associated with a country and annotate whether they could have been associated from the information they contained. We find that the Google Geocoding API is able to associate approximately two thirds of unique user locations to a country, a value that is relatively constant across languages. We also find that this geocoding method exhibits a very high precision (92–96% across languages), with the few errors happening for ambiguous location strings containing multiple locations and which are therefore not geocodable. Also, the share of non-geocoded user locations that could have been geocoded from the provided information is relatively low (12–16%). These instances typically involve the use of emojis, such as national flags, and nicknames for locations (e.g., "Down Under" for Australia), which the Geocoding API fails to recognize. We provide more information on the geocoding evaluation in the Appendix (§B).

Inference In total, we are able to infer the country location of 50,116 English tweets, representing 8% of all posts from the surveyed English HS datasets, 247,408 Arabic tweets (52%) and 13,273 Spanish tweets (27%).

4.2 Reference Points for Representativeness

For each language L, we aim to assess the geo-cultural representativeness of Twitter HS datasets relative to three larger groups: the general Twitter user population speaking language L, the general social media population speaking L, and the general population of speakers of L.

Twitter user population In the absence of reliable information on country share of Twitter users by language, we derive this statistic by using a large Twitter dataset stemming from a recent collaborative project (Pfeffer et al., 2023) that collected all tweets posted within a 24-hour period starting on September 21, 2022, including the geographical metadata. This so-called **Twitter Day** dataset amounts to approximately 116 million English tweets, 27 million Spanish tweets and 19 million Arabic tweets posted by 17, 5 and 2 million users respectively.

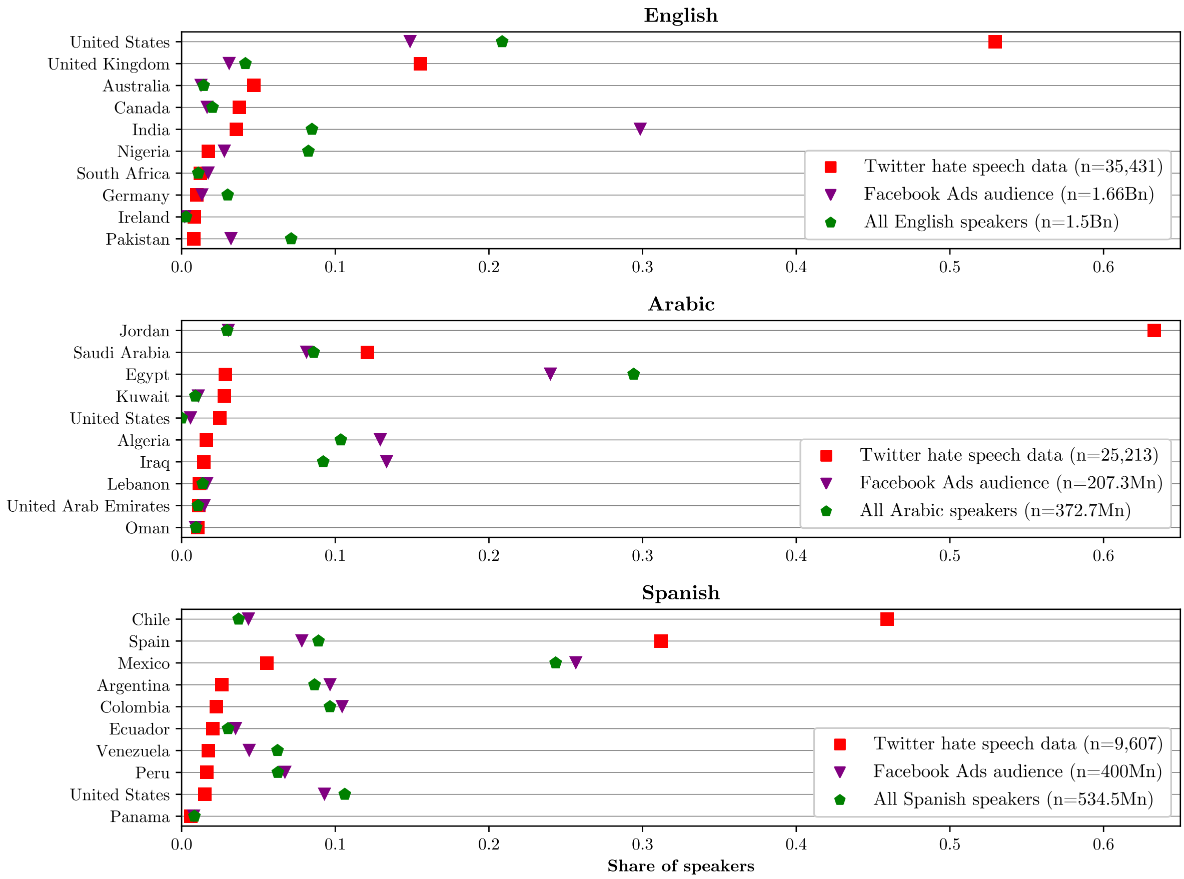

Figure 3: Share of speakers by country location in three reference populations: Twitter users who authored the posts in the Twitter public hate speech datasets (Twitter hate speech data); Facebook and Instagram users (Facebook Ads audience) and all speakers of a language (All [language] speakers).

General social media population Given their large user population and geographical coverage,[5] we use the Facebook and Instagram user populations as a proxy for the general social media population. Specifically, we use the audience measurement tool of **Facebook Ads**. This tool, which has been used in past demographic research (Zagheni et al., 2017; Palotti et al., 2020; Rama et al., 2020), provides the number of Facebook and Instagram users in a given country aged 13 and older that are using these platforms in each of our languages of interest. We then compute the country-level share of the overall Facebook Ads audience for each language.

General population Finally, we use official statistics on the country-level number of speakers of each language of interest. We provide further details on the data sources for each language in the Appendix (§A.4).

4.3 Results

We compute the country share of users speaking each language of interest from four different populations: (i) the Twitter users who authored the posts of the public Twitter HS datasets, (ii) the Twitter user population from the Twitter Day dataset, (iii) the broader social media population using Facebook and Instagram user populations as a proxy, and (iv) the full population of speakers of the language of interest. We report the comparison between (i), (iii), and (iv) in Figure 3 and between (i) and (ii) in Figure 6 in the Appendix.

Bias and lack of representativeness We observe that the majority of Twitter users who authored the posts from the HS datasets originate from a handful of countries for each language, namely the United States and the United Kingdom for English, Jordan for Arabic, and Chile and Spain for Spanish. We also find that the Twitter user population who authored the posts from the public HS datasets is a highly skewed subset of both the broader social

[5] https://datareportal.com/social-media-users

media population and the general speaker population in terms of country location. We further observe a general trend where countries with higher economic development are overrepresented in HS datasets compared to both the social media population and the general population of speakers (notably the US, UK, Australia, and Canada for English, Spain and Chile for Spanish and to a lesser extent, Saudi Arabia and Kuwait for Arabic). In contrast, countries with lower economic development tend to be under-represented in the HS datasets (e.g., India, Nigeria, and Pakistan for English, Egypt, Algeria and Iraq for Arabic and Colombia, Venezuela and Peru for Spanish).

Factors affecting representation Several factors could explain such lack of representativeness. First, the country representation in the Twitter HS data generally aligns with the country representation in the general Twitter population, which is also not representative of the broader social media population nor the total population of speakers. This is particularly the case for English (Pearson correlation of 0.99) but less the case for Spanish (0.43) and Arabic (0.21). Second, this misalignment can also be explained by sampling decisions made when creating the HS datasets. We observe that these decisions are largely intentional for Arabic and Spanish, motivated by the focus on a specific geocultural context. For instance, Jordan's dominance for Arabic is largely explained by the focus on users with a location in Jordan in the sampling of the largest Arabic HS dataset (Ahmad et al., 2023). Similarly, the importance of Chile for Spanish is driven by the choice of Chilean Spanish keywords used for sampling in Arango Monnar et al. (2022). In the case of English, sampling also appears to affect representation as we observe large gaps between the country representation in the HS datasets and in the general Twitter population (Figure 6). Yet, such decisions appear to be either implicit or unintentional as a country focus is almost never mentioned in English HS datasets.

Data and annotator origin Cultural misalignment between data and annotator origin creates a risk of annotation error, due to a lack of cultural understanding. Using the information provided by the dataset authors, we measure the alignment between data and annotator origin for all non-synthetic English, Arabic and Spanish datasets. We report the results in Figure 4.

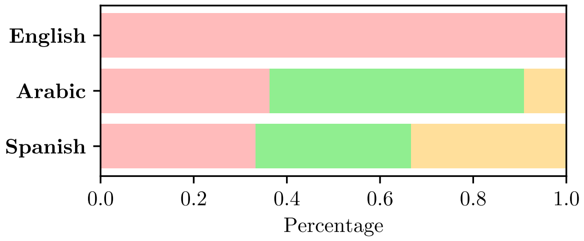

Figure 4: Percentage share (%) of each scenario when comparing data and annotator geographical origin: no information if either the origin of the data or of the annotator is not provided; partial alignment if data and annotator origin partly overlap (e.g., Spanish annotators annotate tweets from Spain and Mexico) and full alignment if data and annotator origin perfectly overlap.

Our most striking result is the lack of information provided by English HS dataset creators about potential cultural misalignment. Indeed, whereas both the data and annotator origin are provided and partially or fully align in 66% of cases for Spanish and 63% for Arabic, none of the surveyed English datasets provide both pieces of information. Specifically, the vast majority of English HS datasets report only the data source (e.g., Twitter) but no precise geographical origin. Similarly, annotator origin is provided in most cases but usually only contains the name of the crowdsourcing platform used (e.g., MTurk, Crowdflower), whose workers originate from a variety of geographies (Difallah et al., 2018).

5 Discussion and Recommendations

Bias evaluation In this work, we evaluated cultural bias in HS datasets in two steps: at the language level and at the geo-cultural level, approximated as a combination of one language and one country. At the language-level, we observe a dominance of English in the number of HS datasets but find that this dominance has been decreasing, with other languages such as Arabic catching up. We also observe that the vast majority of HS corpora originate from Twitter. This is in line and complements the most recent widely-cited and large-scale survey of HS resources (Poletto et al., 2021). Focusing on three geographically widespread languages, namely English, Arabic and Spanish, we then uncover large disparities in country representation, with the majority of data originating from a handful of countries. For each language, we also find that such countries are largely overrepresented in the

HS datasets compared to their prominence both in the broader social media population and the general population of speakers. While the cross-geographic disparities in resources for certain languages had been discussed in past work (e.g., Arango Monnar et al., 2022), our work is the first to quantify such disparities and expose the lack of representativeness of existing resources.

Reasons for bias An important reason for the lack of representativeness of HS datasets comes from their primary data source, Twitter, which itself is a highly non-uniform sample of the broader social media population and the general population (Mislove et al., 2011; Lasri et al., 2023). In this regard, while our analysis exclusively focuses on Twitter, our findings are likely applicable beyond Twitter, as other data sources, such as Reddit, suffer from the same lack of representativeness.[6] Beyond the data source, we observe that sampling decisions made by dataset creators are crucial in reducing representativeness. For instance, seed words are sometimes specific to certain countries, such as Chile for Spanish (Arango Monnar et al., 2022).

Implications The primary implication of our work is the higher risk for less represented cultural contexts to face HS detection errors, due to several factors. First, HS often manifests in culturally specific forms, from its targets (Ousidhoum, 2021) to country-specific offensive words (Ghosh et al., 2021). For instance, the Fulani ethnic group is an important target of online HS in Nigeria (Aliyu et al., 2022; Tonneau et al., 2024) whereas it is not in the US or the UK. The fact that such terms are likely to be less encountered during training may contribute to more false negatives and therefore less protection from HS in under-represented contexts (Dixon et al., 2018). Further, the same words could have different meanings across cultural contexts. For instance, Castillo-lópez et al. (2023) highlight the diverse connotations of the word “fregar” across Spanish-speaking regions, potentially carrying a misogynistic undertone in Spain but not in Ecuadorean Spanish. This discrepancy can lead to false positives and excessive moderation in under-represented contexts resulting from the application of cultural norms from over-represented contexts to under-represented contexts.

Moreover, this performance gap is compounded by a potential misalignment between the origins of data and annotators, resulting in a higher risk of annotation errors for less-represented countries in the annotation workforce. In this regard, we show that creators of English HS datasets seem less aware of this problem compared to Spanish or Arabic, as they consistently fail to provide information on the cultural contexts both the data and annotators originate from. A possible explanation for this difference is that contrary to English, dialects in some languages such as Arabic are not mutually intelligible (e.g., Moroccan and Syrian) rendering the match between data and annotator origin particularly relevant to ensure that the annotator understands the content they are supposed to annotate. Another possible explanation is the tendency to equate English with US-centric data as the majority of English tweets and researchers working on English HS originate from the United States, thereby overlooking the diversity of English speakers online. This lack of information on data and annotator origins may hide a misalignment. For instance, 48% of the crowdworkers employed by Founta et al. (2018) to annotate English tweets are from Venezuela. Lastly, we find that less developed countries tend to be under-represented in HS datasets, potentially reinforcing the marginalization of the same populations HS detection systems are built to protect. While this phenomenon has been documented within the US context for African Americans (Davidson et al., 2019), our findings suggest it can be extended globally.

Recommendations Based on our results, we formulate three recommendations for the development of future HS datasets.

Recommendation 1

Situate datasets in language and geography

When possible, we argue that such a step is necessary to reduce cross-cultural errors in HS detection, especially for culturally-widespread languages such as English. This can be operationalized by using context-specific seed words for sampling or restricting the analysis to users with a specific location. It will allow practitioners to use data that corresponds to the cultural context they want to apply their models in. This additional information will also help better quantify the cultural bias in HS datasets and identify low-resource contexts that require more annotated data.

[6] https://worldpopulationreview.com/country-rankings/reddit-users-by-country

Recommendation 2

Work with annotators that share the same origin as the data to annotate and specify their demographics

This second step will help further reduce detection errors, by ensuring that cultural nuances are well understood. Again, this is especially relevant for culturally-widespread languages and we acknowledge that this recommendation only holds in cases where the data's geographical origin is available or can be inferred. This is in line with prior work advocating for the inclusion of affected communities in determining what is hateful (Maronikolakis et al., 2022) and also echoes the necessity of well-documented data statements (Bender and Friedman, 2018).

Recommendation 3

Ensure data availability while protecting user privacy

We find that a non-trivial amount of datasets cannot be retrieved. While it is crucial to protect the privacy of users on such a sensitive topic, ensuring data access is also crucial to maximize HS detection performance. In line with prior work (Assenmacher et al., 2023), we recommend to publicly release an anonymized version of the dataset and provide full data upon request, under conditions that protect users.

6 Conclusion

This work presented the first evaluation of cultural bias in HS datasets. We confirm past findings on the English-language bias of HS datasets, but also show that this bias has been steadily decreasing in the past few years. We also find evidence of geo-cultural bias for English, Arabic and Spanish, with HS datasets overrepresenting more developed countries and underrepresenting less developed countries. We finally uncover a relative lack of awareness of the possibility of such bias among English HS dataset creators, who systematically fail to provide information about data and annotator origin, hiding potential mismatches. Based on our results, we call for a more nuanced approach to HS detection that takes into account the specific cultural contexts in which speech occurs. We highlight that both language and geography are imperfect representations of culture on their own and discuss the importance of situating datasets using both features and resort to annotators sharing the same origin as the data to limit cross-cultural errors. Still, we are aware that what constitutes "culture" is debated (e.g., Kuper, 2000), as are the rights of minority cultures vis-à-vis larger ones. We advocate for more inclusive representation of different cultures in resources like HS datasets, while recognizing the limitations of language and geography as cultural proxies.

Limitations

Missing data An important limitation of our work is the sole focus on Twitter for the evaluation of geo-cultural bias. While we believe that our conclusions extend to other geo-culturally biased data sources of HS datasets (e.g., Reddit), we cannot empirically verify this claim. Further, we are only able to retrieve geographical information for a subset of all tweets and Twitter users. For instance, we cannot retrieve information for tweets with unavailable IDs, that were deleted or that do not have any geographical metadata. This data is likely not missing at random and thus represents a source of bias in our analysis. For instance, there may be a selection bias where users from some countries are more likely to share their location.

Location and geography do not equate culture While we discuss the importance of using language and geography to define the origin of HS datasets, we are aware that both are imperfect proxies for culture. Diaspora communities illustrate this well: they often have a cultural mix from their origin and current countries. Also, users may provide incorrect location information.

Code-mixing In our analysis, we only focus on single languages (e.g., English, Spanish). Yet, we are aware that code-mixing, that is the combined use of several languages, is prevalent in many English-speaking Majority World countries such as India and Nigeria. We are also aware that a few HS datasets exist for such contexts (e.g., Mathur et al., 2018; Tonneau et al., 2024) and encourage future work to include them in their analysis, in order to get a better estimate of cultural bias in HS datasets.

Ethical Considerations

Data Privacy Owing to the sensitivity of the topic and to protect user privacy, we only provide aggregate results on user location.

Acknowledgements

We thank all the dataset authors who responded to our data requests and made this work possible. We also thank the reviewers for their feedback, which helped improve the paper. PR is a member of the Data and Marketing Insights research unit of the Bocconi Institute for Data Science and Analysis, and is supported by a MUR FARE 2020 initiative under grant agreement Prot. R20YSMBZ8S (INDOMITA).

References

Ashraf Ahmad, Mohammad Azzeh, Eman Alnagi, Qasem Abu Al-Haija, Dana Halabi, Abdullah Aref, and Yousef AbuHour. 2023. Hate speech detection in the arabic language: Corpus design, construction and evaluation.

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

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76. IEEE.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 international conference on advanced computer science and information systems (ICACSIS)*, pages 233–238. IEEE.

Saminu Mohammad Aliyu, Gregory Maksha Wajiga, Muhammad Murtala, Shamsuddeen Hassan Muhammad, Idris Abdulmumin, and Ibrahim Said Ahmad. 2022. Herdphobia: A dataset for hate speech against fulani in nigeria. *arXiv preprint arXiv:2211.15262*.

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

Raghad Alshaalan and Hend Al-Khalifa. 2020. Hate speech detection in saudi twittersphere: A deep learning approach. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 12–23, Barcelona, Spain (Online). Association for Computational Linguistics.

Mohamed Seghir Hadj Ameur and Hassina Aliane. 2021. Aracovid19-mfh: Arabic covid-19 multi-label fake news & hate speech detection dataset. *Procedia Computer Science*, 189:232–241.

Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. 2022. Resources for multilingual hate speech detection. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 122–130, Seattle, Washington (Hybrid). Association for Computational Linguistics.

İnanç Arın, Zeynep Işık, Seçilay Kutal, Somaiyeh Dehghan, Arzucan Özgür, and Berrin Yanikoğlu. 2023. Siu2023-nst-hate speech detection contest. In *2023 31st Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In *Companion proceedings of the 2019 world wide web conference*, pages 1100–1105.

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 *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Dennis Assenmacher, Indira Sen, Leon Fröhling, and Claudia Wagner. 2023. The end of the rehydration era the problem of sharing harmful twitter research data.

Ajeng Dwi Asti, Indra Budi, and Muhammad Okky Ibrohim. 2021. Multi-label classification for hate speech and abusive language in indonesian-local languages. In *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 1–6. IEEE.

Mohammad Atari, Mona J Xue, Peter S Park, Damián Blasi, and Joseph Henrich. 2023. Which humans?

Nofa Aulia and Indra Budi. 2019. Hate speech detection on indonesian long text documents using machine learning approach. In *Proceedings of the 2019 5th international conference on computing and artificial intelligence*, pages 164–169.

Fabienne Baider. 2020. Pragmatics lost? overview, synthesis and proposition in defining online hate speech. *Pragmatics and Society*, 11(2):196–218.

David Bamman, Chris Dyer, and Noah A. Smith. 2014. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, Baltimore, Maryland. Association for Computational Linguistics.

Shakuntala Banaji, Ramnath Bhat, Anushi Agarwal, Nihal Passanha, and Mukti Sadhana Pravin. 2019. Whatsapp vigilantes: An exploration of citizen reception and circulation of whatsapp misinformation linked to mob violence in india.

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 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.

Fatih Beyhan, Buse Çarık, İnanç Arın, Ayşecan Terzioğlu, Berrin Yanikoglu, and Reyyan Yeniterzi. 2022. A Turkish hate speech dataset and detection system. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4177–4185, Marseille, France. European Language Resources Association.

Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. Seegull multilingual: a dataset of geo-culturally situated stereotypes. *arXiv preprint arXiv:2403.05696*.

Uwe Bretschneider and Ralf Peters. 2017. Detecting offensive statements towards foreigners in social media.

Paula Carvalho, Danielle Caled, Cláudia Silva, Fernando Batista, and Ricardo Ribeiro. 2023. The expression of hate speech against afro-descendant, roma, and lgbtq+ communities in youtube comments. *Journal of Language Aggression and Conflict*.

Paula Carvalho, Bernardo Cunha, Raquel Santos, Fernando Batista, and Ricardo Ribeiro. 2022. Hate speech dynamics against African descent, Roma and LGBTQI communities in Portugal. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2362–2370, Marseille, France. European Language Resources Association.

Galo Castillo-lópez, Arij Riabi, and Djamé Seddah. 2023. Analyzing zero-shot transfer scenarios across Spanish variants for hate speech detection. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 1–13, Dubrovnik, Croatia. Association for Computational Linguistics.

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 Twelfth Language Resources and Evaluation Conference*, pages 1397–1403, Marseille, France. European Language Resources Association.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Randall Collins. 1999. *Macrohistory: Essays in sociology of the long run*. Stanford University Press.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Annamaria Silvana de Rosa, Laura Dryjanska, and Elena Bocci. 2018. Evaluative dimensions of urban tourism in capital cities by first-time visitors. In *Encyclopedia of Information Science and Technology, Fourth Edition*, pages 4064–4076. IGI Global.

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. Detox: A comprehensive dataset for German offensive language and conversation analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 143–153, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 135–143.

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

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018a. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018b. Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

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

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

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.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 313–328, Online. Association for Computational Linguistics.

Janis Goldzycher, Paul Röttger, and Gerold Schneider. 2024. Improving adversarial data collection by supporting annotators: Lessons from gahd, a german hate speech dataset. *arXiv preprint arXiv:2403.19559*.

Mark Graham, Scott A Hale, and Devin Gaffney. 2014a. Where in the world are you? geolocation and language identification in twitter. *The Professional Geographer*, 66(4):568–578.

Mark Graham, Bernie Hogan, Ralph K Straumann, and Ahmed Medhat. 2014b. Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers*, 104(4):746–764.

Mark Graham, Ralph K Straumann, and Bernie Hogan. 2015. Digital divisions of labor and informational magnetism: Mapping participation in wikipedia. *Annals of the Association of American Geographers*, 105(6):1158–1178.

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.

Hatem Haddad, Hala Mulki, and Asma Oueslati. 2019. T-hsab: A tunisian hate speech and abusive dataset. In *International conference on Arabic language processing*, pages 251–263. Springer.

Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. 2011. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 237–246.

Valentin Hofmann, Goran Glavaš, Nikola Ljubešić, Janet B Pierrehumbert, and Hinrich Schütze. 2022. Geographic adaptation of pretrained language models. *arXiv preprint arXiv:2203.08565*.

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the elementary multilingual capabilities of large language models with multiq. *arXiv preprint arXiv:2403.03814*.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

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.

Abraham Israeli and Oren Tsur. 2022. Free speech or free hate speech? analyzing the proliferation of hate speech in parler. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 109–121, Seattle, Washington (Hybrid). 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.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada. Association for Computational Linguistics.

Habibe Karayiğit, Ali Akdagli, and Çiğdem İnan Aci. 2022. Homophobic and hate speech detection using multilingual-bert model on turkish social media. *Information Technology and Control*, 51(2):356–375.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, pages 1–30.

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. *arXiv preprint arXiv:2009.10277*.

Hannah Kirk, Bertie Vidgen, Paul Rottger, Tristan Thrush, and Scott Hale. 2022. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1352–1368, Seattle, United States. Association for Computational Linguistics.

Hannah 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)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Vivek Kulkarni, Shubhanshu Mishra, and Aria Haghighi. 2021. LMSOC: An approach for socially sensitive pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2967–2975, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adam Kuper. 2000. *Culture: The anthropologists' account*. Harvard University Press.

Karim Lasri, Manuel Tonneau, Haaya Naushan, Niyati Malhotra, Ibrahim Farouq, Victor Orozco-Olvera, and Samuel Fraiberger. 2023. Large-scale demographic inference of social media users in a low-resource scenario. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 519–529.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Juho Kim, and Alice Oh. 2023a. Crehate: Cross-cultural re-annotation of english hate speech dataset. *arXiv preprint arXiv:2308.16705*.

Nayeon Lee, Chani Jung, and Alice Oh. 2023b. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for 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.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and

Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation*, pages 14–17.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680*.

Antonis Maronikolakis, Axel Wisiorek, Leah Nann, Haris Jabbar, Sahana Udupa, and Hinrich Schuetze. 2022. Listening to affected communities to define extreme speech: Dataset and experiments. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1089–1104, Dublin, Ireland. Association for Computational Linguistics.

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

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018. Did you offend me? classification of offensive tweets in Hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148, Brussels, Belgium. Association for Computational Linguistics.

İslam Mayda, DİRİ Banu, and Tuğba YILDIZ. 2021a. Türkçe tweetler üzerinde makine öğrenmesi ile nefret söylemi tespiti. *Avrupa Bilim ve Teknoloji Dergisi*, (24):328–334.

İslam Mayda, Yunus Emre Demir, Tuğba Dalyan, and Banu Diri. 2021b. Hate speech dataset from turkish tweets. In *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6. IEEE.

Amin Mekacher, Max Falkenberg, and Andrea Baronchelli. 2024. How language, culture, and geography shape online dialogue: Insights from koo. *arXiv preprint arXiv:2403.07531*.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and James Rosenquist. 2011. Understanding the demographics of twitter users. In *Proceedings of the international AAAI conference on web and social media*, volume 5, pages 554–557.

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

Hamdy Mubarak, Hend Al-Khalifa, and Abdulmohsen Al-Thubaity. 2022. Overview of OSACT5 shared task on Arabic offensive language and hate speech detection. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 162–166, Marseille, France. European Language Resources Association.

Hamdy Mubarak, Sabit Hassan, and Shammur Absar Chowdhury. 2023. Emojis as anchors to detect arabic offensive language and hate speech. *Natural Language Engineering*, 29(6):1436–1457.

Hala Mulki and Bilal Ghanem. 2021a. 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.

Hala Mulki and Bilal Ghanem. 2021b. Working notes of the workshop arabic misogyny identification (armi-2021). In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 7–8.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy. Association for Computational Linguistics.

Karsten Müller and Carlo Schwarz. 2021. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4):2131–2167.

Tarek Naous, Michael J Ryan, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

Isar Nejadgholi and Svetlana Kiritchenko. 2020. On cross-dataset generalization in automatic detection of online abuse. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 173–183, Online. Association for Computational Linguistics.

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

Felipe Oliveira, Victoria Reis, and Nelson Ebecken. 2023. Tupy-e: detecting hate speech in brazilian portuguese social media with a novel dataset and comprehensive analysis of models. *arXiv preprint arXiv:2312.17704*.

Anaïs Ollagnier, Elena Cabrio, Serena Villata, and Catherine Blaya. 2022. CyberAgressionAdo-v1: a dataset of annotated online aggressions in French collected through a role-playing game. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 867–875, Marseille, France. European Language Resources Association.

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.

Nedjma Djouhra Ousidhoum. 2021. *On the Importance and Challenges of the Experimental Design of Multilingual Toxic Content Detection*. Hong Kong University of Science and Technology (Hong Kong).

Joao Palotti, Natalia Adler, Alfredo Morales-Guzman, Jeffrey Villaveces, Vedran Sekara, Manuel Garcia Herranz, Musa Al-Asad, and Ingmar Weber. 2020. Monitoring of the venezuelan exodus through facebook's advertising platform. *Plos one*, 15(2):e0229175.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. Detecting and monitoring hate speech in twitter. *Sensors*, 19(21):4654.

Juergen Pfeffer, Daniel Matter, Kokil Jaidka, Onur Varol, Afra Mashhadi, Jana Lasser, Dennis Assenmacher, Siqi Wu, Diyi Yang, Cornelia Brantner, et al. 2023. Just another day on twitter: a complete 24 hours of twitter data. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1073–1081.

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*, 55:477–523.

Nur Indah Pratiwi, Indra Budi, and Ika Alfina. 2018. Hate speech detection on indonesian instagram comments using fasttext approach. In *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 447–450. IEEE.

Nur Indah Pratiwi, Indra Budi, and Meganingrum Arista Jiwanggi. 2019. Hate speech identification using the hate codes for indonesian tweets. In *Proceedings of the 2019 2nd international conference on data science and information technology*, pages 128–133.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. 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 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Afshin Rahimi, Timothy Baldwin, and Trevor Cohn. 2017. Continuous representation of location for geolocation and lexical dialectology using mixture density networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 167–176, Copenhagen, Denmark. Association for Computational Linguistics.

Daniele Rama, Yelena Mejova, Michele Tizzoni, Kyriaki Kalimeri, and Ingmar Weber. 2020. Facebook ads as a demographic tool to measure the urban-rural divide. In *Proceedings of The Web Conference 2020*, pages 327–338.

Mohammadreza Rezvan, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie L Shalin, and Amit Sheth. 2018. A quality type-aware annotated corpus and lexicon for harassment research. In *Proceedings of the 10th acm conference on web science*, pages 33–36.

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

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard Jansen. 2018. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

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 *Proceedings of the international AAAI conference on web and social media*, volume 15, pages 573–584.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Rupak Sarkar and Ashiqur R KhudaBukhsh. 2021. Are chess discussions racist? an adversarial hate speech data set (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15881–15882.

Manuel Tonneau, Pedro Vitor Quinta de Castro, Karim Lasri, Ibrahim Farouq, Lakshminarayanan Subramanian, Victor Orozco-Olvera, and Samuel Fraiberger. 2024. Naijahate: Evaluating hate speech detection on nigerian twitter using representative data. *arXiv preprint arXiv:2403.19260*.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

UN. 2019. Plan of action on hate speech.(2019). *Technical report*.

Natalia Vanetik and Elisheva Mimoun. 2022. Detection of racist language in french tweets. *Information*, 13(7):318.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

Juan Vásquez, Scott Andersen, Gemma Bel-enguix, Helena Gómez-adorno, and Sergio-luis Ojeda-trueba. 2023. HOMO-MEX: A Mexican Spanish annotated corpus for LGBT+phobia detection on Twitter. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

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

Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.

Bertie Vidgen, Helen Margetts, and Alex Harris. 2019. How much online abuse is there. *Alan Turing Institute*, 11.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021a. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021b. 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.

Bertie Vidgen and Taha Yasseri. 2020. Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1):66–78.

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

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

Maximilian Wich, Jan Bauer, and Georg Groh. 2020. Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 54–64, Online. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Emilio Zagheni, Ingmar Weber, and Krishna Gummadi. 2017. Leveraging facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, pages 721–734.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *arXiv preprint arXiv:2402.18815*.

A Data Sources

A.1 Additional Descriptive Statistics

We report the number of datasets by language and survey source in Table 3. The main reason for dropping datasets from the analysis is that a lot of datasets do not focus specifically on hate speech but rather toxicity or offensiveness. The second main reason is the lack of availability of some datasets, as further detailed in §A.3

We also provide additional information in Table 4 on the total number of data points annotated for hate speech as well as the share of all data points by language.

A.2 Retained Hate Speech Datasets

We list below the retained datasets for each language, including six datasets under a "Multilingual" heading.

Arabic

1. *Are They Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere* (Albadi et al., 2018): 6,136 annotated Arabic tweets sampled using names of religious groups. Tweets are annotated as containing religious hate or not and for the hateful ones, which religious group is targeted. Religious hate is defined as "speech that is insulting, offensive or hurtful and is intended to incite hate, discrimination, or violence against an individual or a group of people on the basis of religious beliefs or lack of any religious beliefs". The annotators are CrowdFlower Arabic-speaking crowdworkers with an IP address in the Middle East. The inter-annotator agreement rate is 81% for the first question and 55% for the second question.

2. *T-HSAB: A Tunisian Hate Speech and Abusive Dataset* (Haddad et al., 2019): 6,039 Tunisian Arabic social media posts sampled using hate-related keywords. The comments were annotated as either hateful, abusive or normal by three Tunisian native speakers with a higher education level. Hate comments are defined as instances that "(a) contain an abusive language, (b) dedicate the offensive, insulting, aggressive speech towards a person or a specific group of people and (c) demean or dehumanize that person or that group of people based on their descriptive identity (race, gender, religion, disability, skin color, belief)". The reported Krippendorff α is 0.75.

3. *L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language* (Mulki et al., 2019): 5,846 Levantine tweets sampled using hate-related keywords. The comments were annotated as either hateful, abusive or normal by three Levantine native speakers with a higher education level. Hate comments are defined as instances that "(a) contain an abusive language, (b) dedicate the offensive, insulting, aggressive speech towards a person or a specific group of people and (c) demean or dehumanize that person or that group of people based on their descriptive identity (race, gender, religion, disability, skin color, belief)". The reported Krippendorff α is 0.765.

4. *Hate and offensive speech detection on Arabic social media* (Alsafari et al., 2020): 5,361 Gulf and Modern Standard Arabic tweets sampled through keyword-based, hashtag-based and profile-based approaches. The tweets are annotated in terms of hatefulness, aggressiveness, offensiveness, irony, stereotype and intensity. Hate speech is defined as "possessing one or more of the following characteristics: 1. Insulting or defaming a specific group by using derogatory adjectives words or slurs.; 2. Defending or justifying hate crime.; 3. Promoting and encouraging hate.; 4. Advocating superiority of one group over the other.; 5. Threatening and inciting violence.; 6. Negative and disparaging stereotypes.; 7. Irony and jokes to humiliate and ridicule the target based on their protected characteristic.; 8. Special cases: a) Self-attacking, where the speaker attacks his own protected characteristic with hateful words. b) Re-posting or quoting hateful content". The annotators are three Gulf native speakers with a high educational level. The Cohen κ ranges from 0.77 to 0.9 across annotation levels.

5. *Hate Speech Detection in Saudi Twittersphere: A Deep Learning Approach* (Alshaalan and Al-Khalifa, 2020): 9,316 Saudi Arabic tweets sampled using keyword and hashtags. The tweets were annotated as hateful or not in batches by Figure Eight crowdworkers, Saudi

Language	HS Data Catalogue	Poletto et al. (2021)	Google Search	Total found	Total kept
English	52	16	7	**75**	**29**
Arabic	7	1	8	**16**	**12**
Spanish	3	0	6	**9**	**7**
German	6	1	3	**10**	**7**
Turkish	2	0	5	**7**	**6**
French	3	1	4	**8**	**6**
Portuguese	4	1	6	**11**	**5**
Indonesian	3	0	4	**7**	**3**

Table 3: Number of available hate speech datasets by language and data source

Language	# datapoints in HS datasets	Share of all HS datapoints
English	623,272	41%
Arabic	478,326	32%
Turkish	151,921	10%
German	120,085	8%
Spanish	48,861	3%
Portuguese	46,914	3%
French	25,486	2%
Indonesian	14,904	1%

Table 4: Number and share of datapoints by language for hate speech datasets

annotators and three freelancers familiar with the Saudi dialect. Hate speech is defined as “language that attack a person or a group based on some characteristic such as race, color, ethnicity, gender, religion, or other characteristic”. The inter-annotator agreement rate is not reported.

6. *AraCOVID19-MFH: Arabic COVID-19 Multi-label Fake News & Hate Speech Detection Dataset* (Ameur and Aliane, 2021): 10,828 Arabic tweets sampled using keywords in the context of COVID-19. The tweets are annotated as hateful or not, whether it gives advice, whether it is news or an opinion, whether it contains blame or other negative speech and whether it is worth fact-checking. It is annotated by only one expert annotator.

7. *Let-Mi: An Arabic Levantine Twitter Dataset for Misogynistic Language* (Mulki and Ghanem, 2021a) 6,550 Levantine Arabic tweets replying to popular female journalists during the October 17th 2019 in Lebanon. Tweets are annotated by three Levantine native speakers as non-misogynistic or as one of seven misogynistic categories (discredit, derailing, dominance, stereotyping and objectification, sexual harassment, threat of violence and damning). Unanimous agreement was found on 5,529 tweets, majority agreement on 1,021 tweets and conflicts on 53 tweets.

8. *Working Notes of the Workshop Arabic Misogyny Identification (ArMI-2021)* (Mulki and Ghanem, 2021b) 9,833 Arabic tweets for misogyny identification composed of Modern Standard Arabic and several Arabic dialects including Gulf, Egyptian and Levantine. The Levantine dataset corresponds to the Let-Mi dataset while the multi-dialectal tweets were collected using anti-women hashtags and scraping misogynists’ timelines. The annotation scheme is both binary (misogynystic or not) and multi-class, following the annotation scheme of the Let-Mi dataset. The Krippendorff α is 0.94 for the binary task and 0.67 for the multi-class task.

9. *Overview of OSACT5 Shared Task on Arabic Offensive Language and Hate Speech Detection* (Mubarak et al., 2022): Arabic tweets sampled from Mubarak et al. (2023). Each tweet was annotated by three Appen crowdworkers as 1) offensive or not and for offensive tweets 2) into fine-grained hate speech types. Hate speech is defined as “offensive language targeting individuals or groups based on common characteristics such as Race (including also ethnicity and nationality), Religion (including belief), Ideology (ex: political or sport affiliation), Disability (including diseases), Social Class, and Gender”. Cohen’s κ value is 0.82.

10. *Hate Speech Detection in the Arabic Language: Corpus Design, Construction and*

Evaluation (Ahmad et al., 2023): 403,688 Jordanian Arabic tweets sampled using language, keyword and location filters, focusing on users located in Jordan's main cities. The tweets were annotated by native Jordanian Arabic speakers as either positive, neutral, offensive but not hateful or hateful. Hate speech is defined as "as a form of discourse that targets individuals or groups on the basis of race, religion, gender, sexual orientation, or other characteristics". Fleiss' κ is 0.6.

English

1. *Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter* (Waseem and Hovy, 2016): 16,907 annotated English tweets using a decision list to identify offensive content, focusing on oppression of minorities. Labels include "Racism/Sexism/Neither". The tweets were first annotated by the two authors and later refined by an external annotator. Inter-annotator agreement is κ=0.84.

2. *Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter* (Waseem, 2016): 6,909 annotated English tweets as an extension of Waseem and Hovy (2016) dataset, with an overlap of 2,876 tweets. Labels include "Racism/Sexism/Neither/Both". Annotators are recruited from CrowdFlower without a background selection. The inter-annotator agreement is κ=0.57.

3. *Automated Hate Speech Detection and the Problem of Offensive Language* (Davidson et al., 2017): 24,802 annotated English tweets. Hate speech is defined as language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group, with an emphasis on context. Labels include "Hate speech/Offensive but not hate/Neither". Annotators are recruited from CrowdFlower and the inter-annotator agreement is 0.92.

4. *When Does a Compliment Become Sexist? Analysis and Classification of Ambivalent Sexism Using Twitter Data* (Jha and Mamidi, 2017): 7,205 annotated English tweets focusing on different types of *sexist* content. Original labels include "Benevolent sexism/Hostile sexism/Others". "Hostile sexism" (N=3,378) and "Others" (N=11,559) tweets were extracted from Waseem and Hovy (2016). "Benevolent sexism" content (N=7,205) was annotated by three experts with an interannotator agreement of 0.74.

5. *Detecting Online Hate Speech Using Context Aware Models* (Gao and Huang, 2017): 1,528 annotated comments of 678 users from the Fox News website. Hate speech is defined as language which explicitly or implicitly threatens or demeans a person or a group based upon a facet of their identity such as gender, ethnicity, or sexual orientation. Labels include "Hateful/Non-hateful", annotated by two native English speakers with an interannotator agreement of 0.98.

6. *Hate Speech Dataset from a White Supremacy Forum* (de Gibert et al., 2018): 10,568 annotated sentences from posts and threads from Stormfront. Hate speech is defined as (a) deliberate attack (b) directed towards a specific group of people while (c) motivated by aspects of the group's identity. Labels contain "Hate/No hate/Relation/Skip". "Relation" refers to a sentence that would be considered hateful when used together with other sentences. Three expert annotators achieved an agreement of 90.97%.

7. *Peer to Peer Hate: Hate Speech Instigators and Their Targets* (ElSherief et al., 2018b): 27,330 annotated English tweets identifying hate content, as well as hate instigator and target. Hate speech definition was in line with content guidelines of Facebook and Twitter. Each tweet was annotated (a) hateful or not and (b) as containing a direct attack towards the mentioned account or not, by three Crowdflower annotators. Inter-annotator agreement is 92.8% and 82.6% for the two classifications respectively.

8. *Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media* (ElSherief et al., 2018a): This dataset consists of 28,318 Twitter posts labeled as "directed" hate speech targeting specific individuals or entities, and 331 posts categorized as "generalized" hate speech directed towards broader groups with common protected characteristics like ethnic-

ity or sexual orientation. Each tweet was annotated by at least three independent annotators from Crowdflower, with a Krippendorff's α of 0.622.

9. *Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior* (Founta et al., 2018): 80,000 tweets annotated for various types of inappropriate speech. Initially classified into seven categories - offensive, abusive, hateful, aggressive, cyberbullying, spam, and normal - the final labels used were "Normal/Spam/Abusive/Hateful". Annotators were recruited from CrowdFlower with the largest group (48%) from Venezuela. Agreement of annotators was grouped in three categories, with approximately 55.9% of tweets receiving "overwhelming agreement" (at least 80% of the annotators agree).

10. *Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media* (Salminen et al., 2018): 5,143 comments annotated for hateful content from YouTube and Facebook videos published by news media. One author performed open coding to develop a taxonomy of four types of hateful language - "Accusations/Humiliation/Swearing/Promoting Violence" - and nine target categories (e.g., religion, political issues). Then two other researchers coded a random sample, achieving an overall agreement score of 75.3%.

11. *A Benchmark Dataset for Learning to Intervene in Online Hate Speech* (Qian et al., 2019): Two aggregated HS intervention datasets collected from Gab posts (N=21,747) and Reddit comments (N=7,641) respectively. Each conversation segment was annotated by three annotators who were recruited from Amazon Mechanical Turk (MTurk). The annotations include hate speech classification and suggested intervention responses.

12. *Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application* (Kennedy et al., 2020): 50,000 annotated social media comments from YouTube, Twitter, and Reddit written primarily in English. Annotations span eight categories from counterspeech to genocide. Annotators, recruited from MTurk, were evaluated using the (a) infit mean-squared statistic (0.37-1.9) to assess bias of favoring certain responses, and (b) the percentage of comments where the identity group of the hate target was flagged (no less than 20%).

13. *Detecting East Asian Prejudice on Social media* (Vidgen et al., 2020): 40,000 English tweets aimed at detecting content targeting the East Asian community during Covid-19. Tweets were categorized into five primary groups: "hostility/criticism/counterspeech/discussions of prejudice/unrelated". 20,000 of these tweets were further annotated with secondary labels such as threatening language, interpersonal abuse, and dehumanization. Trained annotators specializing in hate speech performed the annotations. Each tweet was annotated by two annotators with a Fleiss' κ of 0.54.

14. *HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection* (Mathew et al., 2021): A total of 20,148 annotated posts sourced from Twitter (N=9,055) and Reddit (N=11,093). Data were annotated by three annotators from three different perspectives: the basic ("hate/offensive/normal"), the target community, and the rationales (specific post components considered hateful). Each tweet was annotated by three annotators recruited from MTurk with a Krippendorff's α of 0.46.

15. *Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection* (Vidgen et al., 2021b): This synthetic dataset contains 41,255 entries annotated for hate speech and non-hate speech. Specific types of hate identified include derogation, animosity, threatening language, support for hateful entities, and dehumanization, with targets of hate also noted. Annotation was performed on an open-source web platform with each case labeled by 3-5 trained annotators, primarily British (60%), with expert oversights.

16. *"Call me sexist, but..." : Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples* (Samory et al., 2021): This re-annotated dataset comprises 4,078 entries from existing Twitter samples focused on sex-

ism. Annotations cover overall sexism, four specific sexist content categories including behavioral expectations, stereotypes and comparisons, endorsements and denials of inequality, and rejection of feminism, plus three phrasing categories: “uncivil and sexist/uncivil but not sexist/civil“. All annotators were U.S.-based MTurkers. Five annotators rate each entry and the majority agreement rates were 81% for content, 98.8% for phrasing, and 100% for overall sexism.

17. *HateCheck: Functional Tests for Hate Speech Detection Models* (Röttger et al., 2021): This synthetic dataset consists of 3,728 entries designed for hate speech detection, featuring 29 functionalities across 11 classes, such as profanity usage and pronoun reference. A team of ten trained annotators were recruited to ensure data quality, achieving a high inter-annotator agreement with a Fleiss' κ score of 0.93.

18. *An Expert Annotated Dataset for the Detection of Online Misogyny* (Guest et al., 2021): This dataset includes 6,383 Reddit posts and comments labeled for misogyny using a hierarchical taxonomy with four misogynistic categories (e.g., Pejoratives, Treatment, Derogation, Gendered Attacks) and three non-misogynistic categories (e.g., Counterspeech, Non-misogynistic Attacks, None). Secondary and third-level labels were also included. UK-based native English speakers annotated the dataset. Each data entry was annotated by 2-3 annotators. Inter-annotator agreement varied, with Fleiss' κ ranging from 0.145 to 0.559 for categories and 0.484 for the binary task (misogynistic/non-misognistic).

19. *Introducing CAD: the Contextual Abuse Dataset* (Vidgen et al., 2021a): This dataset features 25,000 annotated Reddit entries for classifying online abuse into six primary categories: “Identity-directed/Person-directed/Affiliation-directed/Counter Speech/Non-hateful Slurs/Neutral”, along with subcategories. Annotations also noted whether contextual information was necessary and included corresponding rationales. Instead of crowdsourcing, trained institutional annotators were recruited. Inter-annotator agreement for the primary categories, measured by Fleiss' κ, averaged 0.583.

20. *ETHOS: an Online Hate Speech Detection Dataset* (Mollas et al., 2022): Two datasets comprising 998 binary-labeled hateful comments and 433 messages with detailed labels were collected from YouTube (via Hatebusters) and Reddit. Annotations were conducted on the Figure-Eight platform, assessing whether comments contained hate speech, incited violence, or targeted specific groups. Further, comments were categorized based on hate speech related to gender, race, national origin, disability, religion, and sexual orientation. Almost each comment was annotated by five different annotators. Fleiss' κ scores varied, reaching 0.814 for the binary variable and up to 0.977 for disability-related hate speech.

21. *Hatemoji: A Test Suite and Adversarially-Generated Dataset for Benchmarking and Detecting Emoji-based Hate* (Kirk et al., 2022): The study presented two datasets examining hateful online emojis. The first dataset contains 3,930 hand-crafted test cases, annotated as hateful or non-hateful by three trained annotators, achieving a Randolph's κ of 0.85. The annotators represented three nationalities—Argentinian, British, and Iraqi—with one being a native English speaker. The second dataset includes 5,912 entries annotated by a team of 11 (including one quality control annotator). Each entry was initially classified by three annotators, with hateful entries further categorized into four types and targets of hate. The annotator team included seven British, and one each from Jordanian, Irish, Polish, and Spanish backgrounds, with nine being native English speakers. Randolph's κ scores for three rounds ranged from 0.902 to 0.938.

22. *Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale* (Kennedy et al., 2022): This dataset comprises 27,665 posts from Gab, annotated for hate speech using a hierarchical typology that distinguishes between high-level hate-based rhetoric, defined as “Language that intends to — through rhetorical devices and contextual references — attack the dignity of a group of people, either through an incitement to violence, encouragement of the incitement to violence, or the incitement to hatred”, tar-

geted populations (e.g., race or ethnicity), differentiation between mere vulgarity or aggression and hate speech, and between implicit and explicit rhetoric. Undergraduate research assistants based in the US were trained to annotate the data. Inter-annotator agreement was measured using Fleiss's κ and Prevalence-Adjusted, Bias-Adjusted κ. Agreement scores for top-level categories are human degradation (0.23, adjusted 0.67), calls for violence (0.28, adjusted 0.97), and vulgar/offensive content (0.30, adjusted 0.79).

23. *Free speech or Free Hate Speech? Analyzing the Proliferation of Hate Speech in Parler* (Israeli and Tsur, 2022): This dataset consists of 10,000 annotated posts from Parler, scored on a Likert scale from 1 (not hate) to 5 (extreme or explicit hate). A group of 112 student annotators achieved a satisfactory agreement level of 72% and a Cohen's κ of 0.44.

24. *SemEval-2023 Task 10: Explainable Detection of Online Sexism* (Kirk et al., 2023): This dataset includes 20,000 social media comments from Reddit and Gab to identify online sexism. Sexism was categorized on three levels: binary (sexist or not sexist), detailed subcategories (threats, harm plans and incitement, derogation, animosity, and prejudiced discussion), and 11 specific manifestations. Each social media entry was reviewed by three trained annotators who all self-identified as women. The annotator team included seven British, as well as Swedish, Swiss, Italian, and Argentinian annotators, with eight being native English speakers. For cases lacking unanimous agreement in binary judgments, or less than two-thirds consensus in sub-categories and detailed manifestations, expert reviewers were consulted to provide final labels.

French

1. *An Annotated Corpus for Sexism Detection in French Tweets* (Chiril et al., 2020): 11,834 tweets for detecting sexism. Sexist content was defined as directed/descriptive/reported assertions to the addressee. Each tweet was annotated by five student annotators with an average Cohen's κ of 0.72 for sexist content/non sexist/no decision categories, and 0.71 for direct/descriptive/reporting/non sexist/no decision.

2. *CyberAgressionAdo-v1: a Dataset of Annotated Online Aggressions in French Collected through a Role-playing Game* (Ollagnier et al., 2022): 19 multiparty chat conversations from a role-playing game for high-school students were collected and annotated to determine the presence of hate speech, type of verbal abuse, and humor. Hate speech was defined as content that mocks, insults, or discriminates based on characteristics like color, ethnicity, gender, sexual orientation, nationality, religion, or others. The dataset was fully annotated by one expert, with a second annotator reviewing four conversations. Inter-coder agreement reached Cohen's Kappa scores of 98.4% for hate speech, 91.5% for verbal abuse, and 96.3% for humor.

3. *Detection of Racist Language in French Tweets* (Vanetik and Mimoun, 2022): 2,856 annotated tweets for racist content detection. The dataset was annotated by two French native speakers with a κ agreement of 0.66. In the case of disagreement, a third annotator assigned the final label.

German

1. *Detecting Offensive Statements Towards Foreigners in Social Media* (Bretschneider and Peters, 2017): Three datasets sourced from Facebook (with sample sizes of 2,649; 2,641; and 546) and focused on cyberhate and offensive language, particularly hostility towards foreigners. Offensive statements, their severity, and targets were annotated by two human experts. The intercoder agreement Cohen's κ yielded scores of 0.78, 0.68, and 0.73 for the respective datasets

2. *Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis* (Ross et al., 2017): 541 annotated original tweets containing only textual content, specifically to detect hate speech related to the refugee crisis. Each part was annotated by two annotators with a Krippendorff's α of 0.38.

3. *RP-Mod & RP-Crowd: Moderator-and Crowd-Annotated German News Comment Datasets* (Assenmacher et al., 2021): 85,000 annotated comments from a German newspaper *Rheinische Post*. Comments were an-

notated for various types of hate speech including sexism, racism, threats, insults, and profane language, as well as for organizational content and advertisements. Annotations were conducted by crowdworkers from the Crowd Guru platform. Each comment was reviewed by five (close to) native German annotators, resulting in a Krippendorff's α interannotator agreement score of 0.19.

4. *DeTox: A Comprehensive Dataset for German Offensive Language and Conversation Analysis* (Demus et al., 2022): This dataset consists of 10,278 German annotated tweets, defined as hate speech if they "attack or disparage persons or groups based on characteristics such as political attitudes, religious affiliation, or sexual identity", and distinct from toxicity. Each comment was evaluated by three student annotators. Interannotator agreement, assessed using Gwet's Agreement Coefficient, ranged from 0.75 to 0.95 across different categories.

5. *Improving Adversarial Data Collection by Supporting Annotators: Lessons from GAHD, a German Hate Speech Dataset* (Goldzycher et al., 2024): This adversarial synthetic HS dataset includes approximately 10,966 examples. Hate speech was defined as abusive or discriminatory language targeting protected groups or individuals as members of such groups, with "poor people" also recognized as a protected category. All annotators are native or highly competent German speakers. The interannotator agreement across various rounds ranged from 0.83 to 0.99.

Indonesian

1. *Hate speech detection in the Indonesian language: A dataset and preliminary study* (Alfina et al., 2017): This dataset comprises 713 tweets related to the 2017 Jakarta Governor Election, annotated as hate speech or non-hate speech. Hate speech categories was defined as hatred of religion/ethnicity/race/gender. Each tweet was annotated by three student annotators, each from different religious, racial, and gender backgrounds. Tweets subject to disagreements were excluded, resulting in a 100% interannotator agreement for the included tweets.

2. *Hate Speech Detection on Indonesian Instagram Comments using FastText Approach* (Pratiwi et al., 2018): The dataset consists of 572 annotated Indonesian Instagram comments, with 286 labeled as "HS" (presumably indicating hate speech) and 286 labeled as "not HS" (non-hate speech). The annotations were done manually by three Indonesian annotators from diverse age and gender backgrounds. Comments with disagreement among annotators were removed, ensuring 100% inter-annotator agreement for the included samples.

3. *Multi-Label Hate Speech and Abusive Language Detection in Indonesian Twitter* (Ibrohim and Budi, 2019): 13,169 Indonesian tweets with 7,608 labeled as non-hate and 5,561 labeled as hate speech. The annotations cover abusive language, hate speech detection, identification of the target, category, and level of hate speech. The annotations were performed by crowdsourced native Indonesian annotators with diverse religious, racial/ethnic, and residential backgrounds. Each tweet was annotated by 3 annotators, and only tweets with 100% inter-annotator agreement on the final label were included.

Multilingual

1. *SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter* (Basile et al., 2019): The dataset contains 19,600 annotated tweets, with 13,000 in English and 6,600 in Spanish, focused on hate speech against immigrants and women. The annotations identify the presence of hate speech, the level of aggressiveness, and the targeted group. Three annotators labeled the data. For the English dataset, the reported average confidence scores (combining inter-rater agreement and reliability) are 0.83 for hate speech detection, 0.70 for identifying the target group, and 0.73 for aggressiveness level. For the Spanish dataset, the average confidence scores are 0.89, 0.47, and 0.47 respectively.

2. *CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech* (Chung et al., 2019): The dataset contains 4,078 pairs

of hate speech and counter-narrative text, with 1,288 pairs in English, 1,719 in French, and 1,071 in Italian. The synthetic dataset was created by crowdsourcing to NGOs in the UK, France, and Italy. Two annotators per language independently annotated all the counter-narratives. The inter-annotator agreement, measured by Cohen's κ, is 0.92 across the three languages for annotating the hate speech sub-topic.

3. *Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages* (Mandl et al., 2019): The datasets contain annotated Twitter and Facebook data for hate speech detection in Hindi (N=4,665), German (N=3,819), and English (N=5,852). The labels include binary hate speech detection, types of hate speech, and the targeted group (for English and Hindi only). Several junior annotators were recruited, and the overlap percentages between annotators for hate speech detection on a subset annotated twice were 72% for English, 83% for Hindi, and 96% for German.

4. *Multilingual and Multi-Aspect Hate Speech Analysis* (Ousidhoum et al., 2019): The dataset comprises 13,014 tweets in Arabic (N=3,353), English (N=5,647), and French (N=4,014), labeled via crowdsourced annotators from MTurk using a multi-level scheme. The annotations capture directness, hostility level, target, group, and the annotator's feeling aroused by the tweet. Each tweet was annotated by five annotators and the interannotator agreement is measured using Krippendorff's α with 0.153 for English, 0.244 for French, and 0.202 for Arabic.

5. *Multilingual HateCheck: Functional Tests for Multilingual Hate Speech Detection Models* (Röttger et al., 2022): The dataset contains synthetic test cases for detecting hateful speech across ten languages: Arabic, Dutch, French, German, Hindi, Italian, Mandarin, Polish, Portuguese, and Spanish. It comprises 36,582 test cases, out of which 25,511 (69.7%) are labeled as hateful, and 11,071 (30.2%) as non-hateful. Hate speech was defined as abuse targeted at a protected group based on age, disability, gender identity, race, national or ethnic origin, religion, sex, or sexual orientation. Each test case was reviewed by three native-speaking annotators. Annotator agreement was measured by the portion of disagreement where at least 2 out of 3 annotators disagreed with the expert gold label, ranging from 0.73% for Italian to 21.22% for French.

6. *Large-Scale Hate Speech Detection with Cross-Domain Transfer* (Toraman et al., 2022): 200,000 human-labeled tweets, covering both English (N=100,000) and Turkish (N=100,000) languages. Hate speech was defined including not only hateful behavior but also frequently observed domains based on target groups (religion, gender, race, politics, and sports). The labels include "hate speech/offensive/normal". Each tweet was annotated by five student annotators. The inter-annotator agreement, measured by Krippendorff's α coefficient, is 0.395 for the English data and 0.417 for the Turkish data.

Portuguese

1. *A Hierarchically-Labeled Portuguese Hate Speech Dataset* (Fortuna et al., 2019): 5,668 Portuguese tweets sampled using hate-related keywords and profiles. The annotators are Portuguese native speakers who are Information Science students. Each tweet is annotated by three students as hateful or not, and if hateful, the type of hate speech is also annotated (e.g., sexism). Hate speech is defined as "language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used". Fleiss' κ is 0.17.

2. *Toxic Language Dataset for Brazilian Portuguese (ToLD-Br)* (Leite et al., 2020): 20,818 Brazilian Portuguese tweets sampled using keywords, hashtags as well certain user profiles (e.g., Bolsonaro). Each tweet was annotated by three Brazilian university students as either LGBTQ+phobia, obscene, insult, racism, misogyny, xenophobia or neutral. The average Krippendorff's α is 0.55.

3. *HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection* (Vargas et al., 2022): 7,000 Brazilian Instagram posts commenting content from major Brazilian politicians. Each comment was annotated by three annotators in three steps: 1) offensive or not and 2) intensity of offensiveness and 3) hate speech type. Following Fortuna et al. (2019), hate speech is defined as "a kind of language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, or others, and it may occur with different linguistic styles, even in subtle forms or when humor is used. Therefore, hate speech is a type of language used against groups target of discrimination (e.g., sexism, racism, homophobia)." The annotators are Brazilians with a high education level. The average Cohen's κ is 0.75 for offensiveness and 0.47 for intensity of offensiveness.

4. *TuPy-E: detecting hate speech in Brazilian Portuguese social media with a novel dataset and comprehensive analysis of models* (Oliveira et al., 2023): 9,367 Brazilian Portuguese tweets sampled using hate-related keywords and random sampling. Each tweet was annotated by three individuals in two steps: 1) as aggressive or not, 2) if aggressive, assign to one hate speech category among ageism, aporophobia, body shame, capacitism, LGBTphobia, political, racism, religious intolerance, misogyny and xenophobia. Hate speech is defined as "the use of language that attacks or degrades, incites violence, or promotes hatred against groups based on specific characteristics such as physical appearance, religion, national or ethnic origin, sexual orientation". Annotators are Brazilian with a high level of education. The agreement rate is not reported.

Spanish

1. *Detecting and Monitoring Hate Speech in Twitter* (Pereira-Kohatsu et al., 2019): 6,000 annotated tweets from Spain selected using hate keywords. The tweets were annotated by four annotators (one public servant and three graduates) as hateful or not and a fifth annotation was sought in case of disagreements (Cohen κ: 0.588). Hate speech is defined as "a kind of speech that denigrates a person or multiple persons based on their membership to a group, usually defined by race, ethnicity, sexual orientation, gender identity, disability, religion, political affiliation, or views".

2. *Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings* (García-Díaz et al., 2021): 7,682 Spanish tweets from both Spain and Latin America, annotated as either misogynous or not. The tweets were annotated by two annotators (Krippendorff α: 0.69).

3. *Multilingual Resources for Offensive Language Detection* (Arango Monnar et al., 2022): 9,834 annotated Chilean Spanish tweets sampled using hate-related Chilean keywords. Tweets were annotated by three native Chileans as either hate speech, insult, unintended or intentional profanity. Hate speech is defined as "stereotypical language to offend minority groups such as women, immigrants, sexual or racial minorities". The authors report an agreement rate higher than 90% and a Krippendorff α higher than 0.7 for all labels.

4. *Analyzing Zero-Shot transfer Scenarios across Spanish variants for Hate Speech Detection* (Castillo-lópez et al., 2023): 4,000 Spanish tweets from both Spain and Latin America sampled using geolocation and hate-related keywords. The tweets were annotated by three Latin American native Spanish speakers as xenophobic, non-xenophobic or ambiguous (Cohen κ: 0.44, agreement rate: 88%). A tweet is xenophobic if (i) "The content of the tweet primarily targets immigrants as a group, or even a single individual, if they are considered to be a member of that group (and NOT because of their individual characteristics)" and (ii) "The content of the tweet propagates, incites, promotes, or justifies hatred or violence towards the target or a message that aims to dehumanize, hurt or intimidate the target".

5. *HOMO-MEX: A Mexican Spanish Annotated Corpus for LGBT+phobia Detection on Twitter* (Vásquez et al., 2023): 11,000 Mexican tweets sampled using nouns indicative of the LGBTQ+ community. The annotators were

composed of 11 Mexican and 1 Colombian individuals. Each tweet were annotated by four annotators as either "LGBTQ+phobic", "not LGBTQ+phobic" or "irrelevant to the LGBTQ+ community" (Cohen κ: 0.43). If annotated as LGBTQ+phobic, the tweets were further annotated by type of LGBTQ+phobia.

Turkish

1. *Hate Speech Detection with Machine Learning on Turkish Tweets* (Mayda et al., 2021a): 1,000 annotated Turkish tweets, sampled using names of target groups. Labels include *hate speech, offensive expression, none of the two*. Annotated by two evaluators and disagreements are annotated by a third annotator (agreement rate of 83.4%).

2. *Hate Speech Dataset from Turkish Tweets* (Mayda et al., 2021b): 10,224 annotated Turkish tweets, sampled using name of target groups (e.g., jews). Labels include hate speech, offensive speech, or neutral. The tweets classified as hate were further annotated into subclasses, including ethnic, religious, sexist, and political tags. Two annotators labeled tweets separately, reaching a 92.5% agreement rate, later increased to 98.4% after discussion. A third evaluator resolved remaining disagreements.

3. *A Turkish Hate Speech Dataset and Detection System* (Beyhan et al., 2022): This work contributes two hate speech datasets: the Istanbul Convention Dataset and the Refugee dataset. Hate speech is defined as "language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group". The annotation scheme has four parts: (1) whether the tweet has no, weak or strong offensive language, (2) stance towards the Istanbul Convention or Refugees (pro, against or neutral), (3) target group and (4) hate speech type (e.g., insult, exclusion). The Istanbul Convention Dataset is composed of 1,206 tweets selected using hashtags and keywords. It was annotated by three senior undergraduate students (Krippendorff α: 0.84 for binary task and 0.82 for multi-class task). The Refugee Dataset is composed of 1,278 tweets selected using immigrant-related keywords. Part of it was annotated by the undergraduate students and another part was annotated by employees of the Hrant Dink Foundation.

4. *Homophobic and Hate Speech Detection Using Multilingual-BERT Model* (Karayiğit et al., 2022): 31,290 Turkish Instagram comments sampled from accounts often posting homophobic and more generally hateful comments. The comments are annotated as either homophobic, hateful or neutral. The posts were annotated by two researchers.

5. *SIU2023-NST - Hate Speech Detection Contest* (Arın et al., 2023): Shared task contributing two Turkish hate speech datasets: 2,240 tweets on the Israel-Palestine conflict annotated by hate speech type, as well as how severe hateful cases are; 4,683 tweets on refugees annotated as hate speech or not, as well as how severe hateful cases are.

A.3 Unavailable Datasets

We were not able to retrieve 5 English (Nobata et al., 2016; Fersini et al., 2018; Rezvan et al., 2018; Sarkar and KhudaBukhsh, 2021; Vidgen and Yasseri, 2020), 3 Indonesian (Aulia and Budi, 2019; Pratiwi et al., 2019; Asti et al., 2021), 3 Portuguese (Maronikolakis et al., 2022; Carvalho et al., 2022, 2023), 1 Spanish (Fersini et al., 2018) and 1 German (Maronikolakis et al., 2022) datasets.

A.4 Official Statistics

For English, we use data on the number of speakers as a first or second language[7]. In the absence of such detailed data for other languages, we use data on the number of native speakers by country for Spanish[8] and Arabic[9].

B Geocoding Evaluation

We provide the full results of the geocoding evaluation in Table 5.

C Comparison with Twitter Day

Post-level We provide a comparison between the country shares for posts in the Twitter hate speech

[7] https://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population

[8] https://cvc.cervantes.es/lengua/espanol_lengua_viva/pdf/espanol_lengua_viva_2022.pdf

[9] https://www.worlddata.info/languages/arabic.php

	English	Arabic	Spanish
Share of geocoded user locations	59%	71%	66%
Share of correct geocoding	92%	94%	96%
Share of non-geocoded user locations that could have been geocoded from the provided information	14%	12%	16%

Table 5: Geocoding evaluation

data and the Twitter Day dataset in Figure 5.

User-level We provide a comparison between the country shares for users in the Twitter hate speech data and in the Twitter Day datasets across languages (Figure 6).

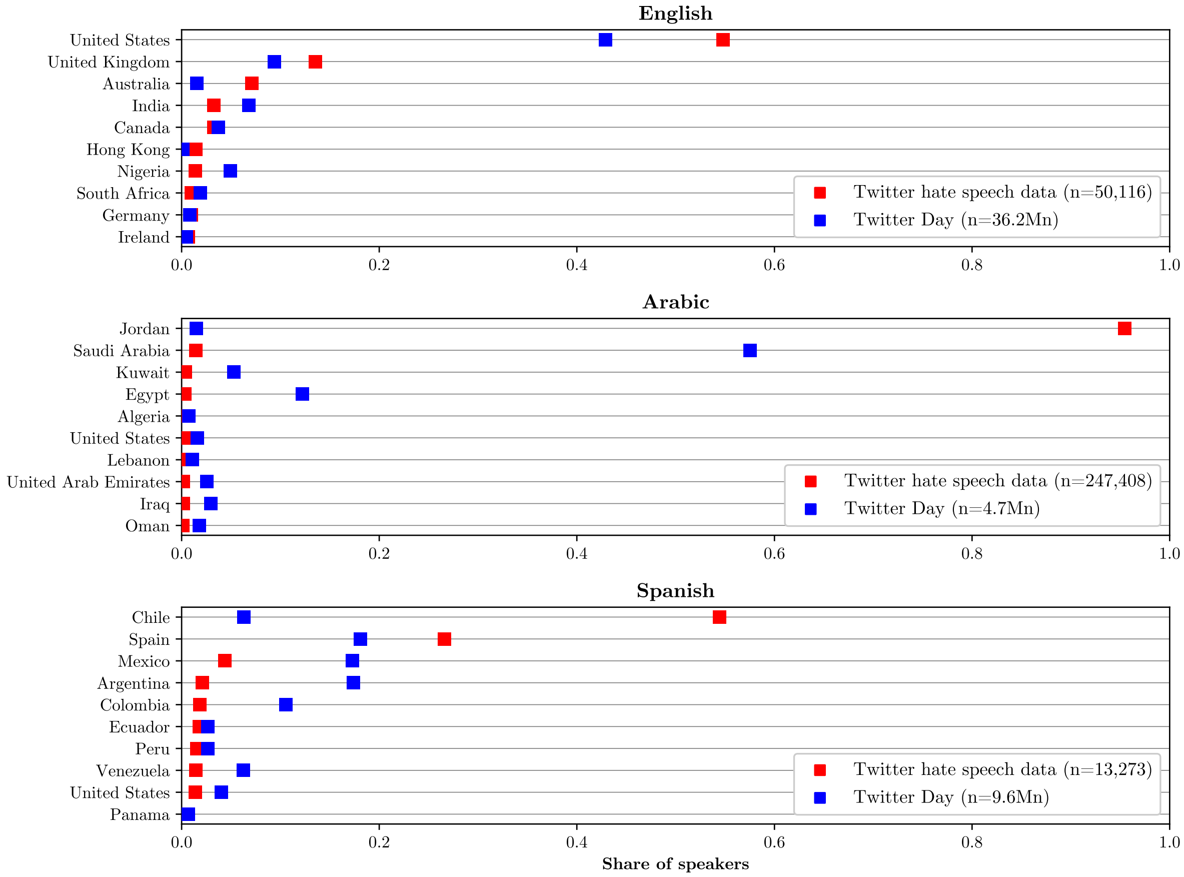

Figure 5: Share of posts by country location in two reference populations: posts in the Twitter public hate speech datasets (Twitter hate speech data) and all Twitter posts, using the Twitter Day dataset as a proxy (Twitter Day)

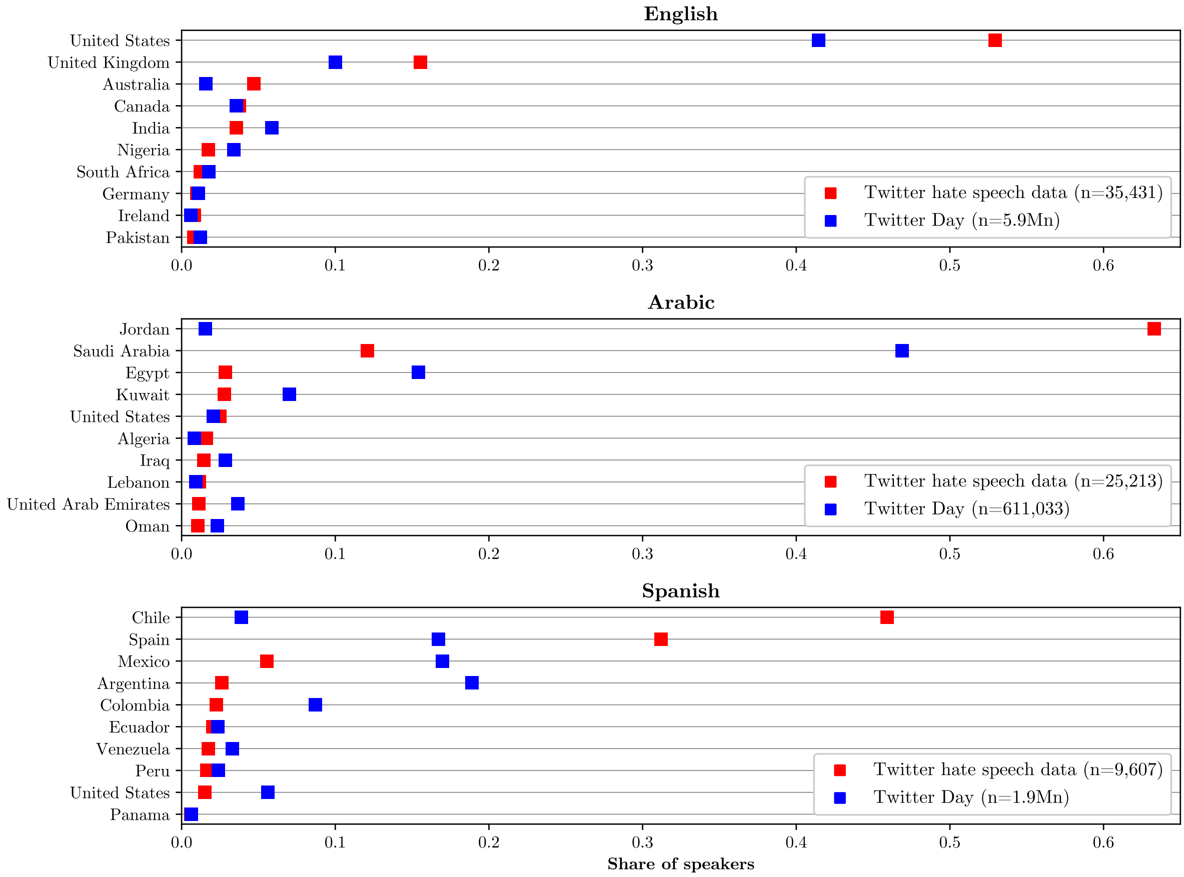

Figure 6: Share of speakers by country location in two reference populations: Twitter users who authored the posts in the Twitter public hate speech datasets (Twitter hate speech data) and Twitter user population, using the Twitter Day data as a proxy (Twitter Day)