Lang Resources & Evaluation
https://doi.org/10.1007/s10579-020-09502-8

SURVEY

Resources and benchmark corpora for hate speech detection: a systematic review

Fabio Poletto[1] · Valerio Basile[1] · Manuela Sanguinetti[1] · Cristina Bosco[1] · Viviana Patti[1]

Abstract Hate Speech in social media is a complex phenomenon, whose detection has recently gained significant traction in the Natural Language Processing community, as attested by several recent review works. Annotated corpora and benchmarks are key resources, considering the vast number of supervised approaches that have been proposed. Lexica play an important role as well for the development of hate speech detection systems. In this review, we systematically analyze the resources made available by the community at large, including their development methodology, topical focus, language coverage, and other factors. The results of our analysis highlight a heterogeneous, growing landscape, marked by several issues and venues for improvement.

The work of F. Poletto is funded by Fondazione Giovanni Goria and Fondazione Cassa di Risparmio di Torino (*Talenti della Società Civile 2018*). The work of V. Basile, C. Bosco, V. Patti and M. Sanguinetti is partially funded by Progetto di Ateneo/Compagnia di San Paolo 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01) .

✉ Fabio Poletto
poletto@di.unito.it

Valerio Basile
basile@di.unito.it

Manuela Sanguinetti
msanguin@di.unito.it

Cristina Bosco
bosco@di.unito.it

Viviana Patti
patti@di.unito.it

[1] University of Turin, Turin, Italy

Published online: 30 September 2020

Keywords Hate speech detection · Benchmark corpora · Natural Language Processing shared tasks · Systematic review

1 Introduction

Within the field of AI, and Natural Language Processing (NLP) in particular, techniques for tasks related to Sentiment Analysis and Opinion Mining (SA&OM) grew in relevance over the past decades. Such techniques are typically motivated by purposes such as extracting users' opinion on a given product or polling political stance. Robust and effective approaches are made possible by the rapid progress in supervised learning technologies and by the huge amount of user-generated contents available online, especially on social media. More recently the NLP community witnesses a growing interest in tasks related to social and ethical issues, also encouraged by the global commitment to fighting extremism, violence, fake news and other plagues affecting the online environment. One such phenomenon is hate speech, a toxic discourse which stems from prejudices and intolerance and which can lead to episodes, and even structured policies, of violence, discrimination and persecution.

Hate Speech (HS), lying at the intersection of multiple tensions as expression of conflicts between different groups within and across societies, is a phenomenon that can easily proliferate on social media. It is a vivid example of how technologies with a transformative potential are loaded with both opportunities and challenges. Implying a complex balance between freedom of expression and defense of human dignity, HS is hotly debated and has recently gained traction in the AI community, that can play a leading role in developing tools to confront pervasive dangerous trends such as the escalation of violence and hatred in online communication, or the spread of fake news.

The motivation to study HS from a computational perspective is manifold. On the one hand, as a linguistic and pragmatic phenomenon, computational linguistic techniques enable the scholar to gain insights and empirical evidence on its intrinsic characteristics. On the other hand, several actors—including institutions and ICT companies to comply to governments' demands for counteracting the HS phenomenon[1]—have an increasing need for automatic support to moderation or for monitoring and mapping the dynamics and the diffusion of HS dynamics over a territory (Capozzi et al. 2019), which is only possible at a large scale by employing computational methods.

HS is a complex and multi-faceted notion that has proven difficult to recognize, both by humans and machines. Researchers who recently started tackling this issue from an NLP perspective are designing operational frameworks for HS, annotating corpora with several semantic frameworks, figuring out the most representative features, and testing automatic classifiers. Moreover, the involvement of the scientific community resulted in a number of evaluation tasks organized in different

[1] See for instance the Code of Conduct on countering illegal hate speech online issued by EU commission (EU Commission 2016).

languages, releasing benchmark corpora and encouraging participants to develop their own classification systems.

Being the subject in a yet recent stage, it suffers from several weaknesses, related to both the specific targets and nuances of HS and the nature of the classification task at large, that prevent systems from reaching optimal results. One of the major issues consists in the intrinsic complexity in defining HS and in a widespread vagueness in the use of related terms (such as abusive, toxic, dangerous, offensive or aggressive language), that often overlap and are prone to strongly subjective interpretations. As we will also show in the present survey, this results in a sparsity of heterogeneous resources each reflecting a subjective perception, and in a variety of systems each trained on a different resource.

Given the considerable amount of research produced in recent years, we undertook the task of writing a systematic and up-to-date review on the subject, focusing on shared tasks organized and resources released so far for HS detection. Purposes of a systematic survey include summarizing existing work, helping identify gaps and weaknesses in current research, suggesting areas for further investigation, and providing a solid framework for improving NLP research on HS detection.

This contribution aims at complementing other surveys proposed in this field, in particular by Lucas (2014), Schmidt and Wiegand (2017) and Fortuna and Nunes (2018). In fact, we analyzed their work, bearing in mind a number of objective questions meant to help point out their strengths and weaknesses. In doing so, we focused in particular on the main reviews' objectives, the sources and depth of the search of the reviewed studies, the inclusion/exclusion criteria adopted to select these studies, how data were extracted, synthesized and combined, and whether conclusions flow from the evidence.

These reviews mention either explicit research questions, open issues or suggestions about future work, and are conducted with varying degrees of systematicity. Overall, their main objective is to provide an overview of the approaches proposed in literature for automatic HS detection, focusing either on high-level descriptions of methods (Lucas 2014) or on specific computational approaches, with a special emphasis on NLP (Fortuna and Nunes 2018; Schmidt and Wiegand 2017), thus analyzing models, features and algorithms.

As regards the sources and depth of the search, in Schmidt and Wiegand (2017) there is no explicit mention of how sources were explored, and in Lucas (2014) potential sources have been admittedly overlooked, while in Fortuna and Nunes (2018) the methodology was meant to be systematic and aimed at finding as many documents as possible in the areas of interest (computer science and engineering). Among these three surveys, the latter is also the only one that states explicit inclusion/exclusion criteria to select the studies and that reports numerical results from the surveyed papers. The conclusions drawn from such results are that it is not clear which approaches perform better, also due to differences in the datasets used (among other factors). The need for benchmark datasets that allow comparative studies is also highlighted in Schmidt and Wiegand (2017). However, it must be noted that many of the resources included in this survey had not yet been released when the previous surveys were published (or, at least, when their search was

carried out), especially those released for shared tasks—which proves, once again, how dynamic and fast—growing the field is. What is more important, a large proportion of HS resources developed in the recent past includes data in languages other than English, thus broadening the HS detection scenario to a multiplicity of linguistic—as well as cultural—perspectives. Such linguistic diversity, on the other hand, also confirms the need to provide a complete picture of the resources available to the research community, especially for those aiming to adopt multilingual approaches. In this respect, it is worth mentioning a repository[2] that attempts to gather all the corpora on HS and related phenomena that have been released so far, cataloguing them according to the language involved. Such repository, however, just provides a list with concise information on the datasets to those interested in using the data for computational purposes. To the best of our knowledge, a complete overview of such resources that would also take into account of different viewpoints and dimensions is still missing. This work aims therefore at providing a more comprehensive view of the datasets, lexica and evaluation campaigns that are centered on the notion of HS.

Furthermore, similarly to what has been done in Fortuna and Nunes (2018) with respect to papers on HS detection, we apply a systematic approach based on explicit research and evaluation criteria, in order to draw conclusions on the state of the art and suggestions for future work that can only emerge from a comprehensive analysis of the subject.

This paper describes first how the research was conducted, analyzing the criteria adopted and the search results (Sect. 2). It then provides an overview of the resources found (Sects. 3 and 4), also proposing a lexical analysis of some of them (Sect. 5), aiming to highlight how topic biases can be pervasive in such kind of resources. Some concluding remarks (Sect. 6), drawn from the survey findings, close the paper.

2 Methodology

In compiling this survey, we relied on the guidelines provided by Kitchenham (2004) for writing systematic reviews on the subject of software engineering, adapting them to the peculiarities of our field. In this section, we will mention the main steps we followed in the research process. A set of keywords was set up and used to browse search engines and repositories. We picked English keywords since English is used worldwide as working language among scholars; however, we did not restrict our search to works based on English data alone, instead including as many languages as possible.

[2] https://hatespeechdata.com/.

2.1 Sources

We collected any peer-reviewed academic work found on Google Scholar[3] and Google Books[4], limiting our query to the first ten pages for each keyword and sorting results by relevance, without time filter. The systematic search was conducted in two occasions: the main search was carried between June 2018 and April 2019, and subsequently the results were updated with a new search by the same parameters, conducted between March and April 2020. We also collected resources for which references to the used methodology or the implemented system were provided on public version control repositories on Github[5], Gitlab[6] and Bitbucket[7]. Finally, the first two pages of results of the general Web search by Google[8] have been accessed. We furthermore scanned the proceedings of workshops and shared tasks found on these sources with the same keywords (see Sect. 4.2 for a complete list).

We carefully read each work and labeled it with a set of specifically-designed labels, sorting our list by research field (e.g., *field-socialsciences*, *field-NLP*, etc.), main focus (e.g., *content-resource*, *content-system*, etc.), methodology (e.g., *method-nn* for neural nets, *method-ml* for machine learning, etc.), specific phenomena investigated (e.g., *topic-hs* for HS at large, *topic-racism* when the topical focus is on racist speech, etc.) and language (e.g., *lang-en*, *lang-it*, etc.). Although we collected a much larger number of works, the present review only describes those labeled as resources or shared task overviews.

2.2 Inclusion and exclusion criteria

All works not related to HS (and similar subjects), not presenting a NLP approach or not peer-reviewed were discarded, with the exception of a few datasets only published on the Web. A major issue we had to deal with are the fuzzy boundaries between HS and broader concepts such as abusive language, offensive language and toxic language on one hand, and between HS and more specific focus-driven labels such as racism, anti-semitism, sexism, misogyny and homophobia on the other hand. The lack of a common framework among scholars from a variety of disciplines leaves room for subjective interpretations, so that the same linguistic phenomenon can be given different names, or conversely the same label used for different phenomena.

In order to ground our study in a methodologically sound foundation, we rely on the definition of HS given by Sanguinetti et al. (2018), here rephrased and summarized: a content defined by its action—generally spreading hatred or inciting violence, or threatening by any means people's freedom, dignity and safety—and by

[3] https://scholar.google.com.

[4] https://books.google.com.

[5] https://github.com.

[6] https://about.gitlab.com/.

[7] https://bitbucket.org/.

[8] https://www.google.com.

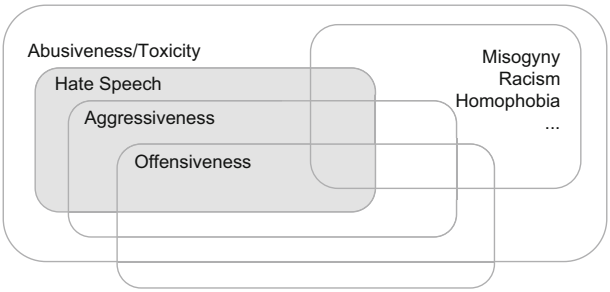

Fig. 1 Relations between HS and related concepts

its target—which must be a protected group, or an individual targeted for belonging to such a group and not for his/her individual characteristics. This definition is in turn based on a thorough investigation of definitions proposed in a variety of fields, including computational linguistics, pragmatics, law and social sciences, and is the result of an attempt to merge some key points into a structured framework apt for computational purposes. Different definitions may in fact stress different aspects of HS: some focus on the linguistic form, others on the writer's intention, others yet on the potential effect on the victim. In compiling a survey, we are not called to propose our own original definition; but it is of primary importance to recognize those works and resources that are related to the concept, even though some of them call it with a different name.

Figure 1 shows a depiction of our working framework, and an attempt to clarify the matter, based also on the reviewed literature. While we consider HS an instance of abusive language, not all manifestations of hatred towards certain targets are categorized as HS under our definition. For instance, racial *microaggressions* (Sue et al. 2007) are definitely expressions of racism, but they do not necessarily contain a call to violent action that would put them in the HS class of our framework.

Below we show some examples of the various concepts related to HS in Fig. 1, that is texts extracted by the benchmark corpora and HS detection resources for different languages we reviewed, that were labeled as representative samples of such phenomena:

altro che profughi? sono zavorre e tutti uomini (refugees? They are deadweights and all men)
Source: (Bosco et al. 2018) **Label:** hateful **Language:** Italian

tutto tempo danaro e sacrificio umano sprecato senza eliminazione fisica dei talebani e dei radicali musulmani e tutto inutile (it's all a waste of time, money and human lives without the extermination of Taliban and radical Muslims it's all useless)
Source: (Sanguinetti et al. 2018) **Label:** aggressive **Language:** Italian

@USER Figures! What is wrong with these idiots? Thank God for @USER
Source: (Zampieri et al. 2019b) **Label:** offensive **Language:** English

Table 1 Glossary of terms relevant to the present survey, with their definitions from the literature

Term and definitions	Source
Hate Speech	Warner and Hirschberg (2012)
Any communication that disparages a person or a group on the basis of some characteristic such as race, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic	
Use of a sexist or racial slur, attack a minority, promotes hate speech or violent crime, blatantly misrepresents truth, shows support of problematic hashtags, defends xenophobia or sexism, or contains a screen name that is offensive	Waseem and Hovy (2016)
Act of offending, insulting or threatening a person or a group of similar people on the basis of religion, race, caste, sexual orientation, gender or belongingness to a specific stereotyped community	Schmidt and Wiegand (2017)
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	Davidson et al. (2017)
Any communication that disparages a target group of people based on some characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic	Nockleby (2000)
Aggressiveness	Sanguinetti et al. (2018)
Intention to be aggressive, harmful, or even to incite, in various forms, to violent acts against a given target	
Offensiveness	Zampieri et al. (2019a)
Any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct	
Profanity, strongly impolite, rude or vulgar language expressed with fighting or hurtful words in order to insult a targeted individual or group	Fortuna and Nunes (2018)
Abusiveness/ toxicity	Founta et al. (2018)
Hurtful language, including hate speech, derogatory language and also profanity	
Any strongly impolite, rude or hurtful language using profanity, that can show a debasement of someone or something, or show intense emotion	Fortuna and Nunes (2018)
Extremely offensive and insulting; engaging in or characterized by habitual violence and cruelty	Oxford English Dictionary (2019)
Using harsh, insulting language an angry and abusive crowd; harsh and insulting abusive language; using or involving physical violence or emotional cruelty	Merriam-Webster Online (2009)
Misogyny	Poland (2016)
Hate speech whose targets are women.	

Table 1 continued

Term and definitions	Source
A hatred of women	Merriam-Webster Online (2009)
Racism	Oxford English Dictionary (2019)
Prejudice, discrimination, or antagonism directed against someone of a different race based on the belief that one's own race is superior	
A belief that race is the primary determinant of human traits and capacities and that racial differences produce an inherent superiority of a particular race	Merriam-Webster Online (2009)
Homophobia	Merriam-Webster Online (2009)
Irrational fear of, aversion to, or discrimination against homosexuality or homosexuals	

You should be fired, you're a moronic wimp who is too lazy to do research. It makes me sick that people like you exist in this world
Source: Hate Speech Hackathon **Label:** toxic **Language:** English

*I've yet to come across a nice girl. They all end up being bit**es in the end*
Source: (Fersini et al. 2018a) **Label:** misogynous **Language:** English

*These savages invade Our Country, disrupt cities, turn many into sh***es like where they came from and WE THE PEOPLE are paying for this SH*T. [...]*
Source: (Basile et al. 2019) **Label:** hate speech **Target:** migrants **Language:** English

oltre 2300 miliardi diuro. Il P.D. va a caccia , ora, dei soli voti di ricchioni, omosessuali, trans, naziskin , ... URL
(over 2300 billionuros. PD is now hunting only votes from fags, homosexuals, trans, skinheads, ... URL)
Source: Akhtar et al. (2019) **Label:** homophobic **Language:** Italian

To further clarify the concepts under study and their relationships with each other, we compiled a glossary of the terms in Fig. 1 and their definitions according to several sources from recent literature, shown in Table 1. Partial attempts to precisely classify overlapping abusive phenomena are found in the literature, such as Malmasi and Zampieri (2018) exploring the distinction between HS and profanity. Davidson et al. (2017) further distinguish HS from offensive language, citing examples such as:

- *Stupid f*cking n*gger LeBron. You flipping jun- gle bunny monkey f*ggot* (Hate Speech)
- *Why you worried bout that other h*e? Cuz that other h*e aint worried bout another h*e* (Offensive)

Moreover, (Waseem et al. 2017) contributes to the critical reflection on the relationships between different phenomena that have been grouped under the "abusive language" label, by introducing a two-fold typology that considers (i) whether the abuse is directed at a specific target or towards a generalized group, and (ii) the degree to which it is explicit or implicit. Authors argue about the implications for annotation of the proposed classification, which inspired the multi-layer annotation scheme proposed for the dataset of the OffensEval2019 shared task (Zampieri et al. 2019b) and other works, including the target-aware annotation in Basile et al. (2019) and the implicit-explicit distinction in the annotation of Caselli et al. (2020).

The present survey wants to draw attention to the recent efforts towards a structured NLP community concerned with hateful language recognition, efforts that necessarily include not only systems implementation but also, and primarily, the development of solid resources from different sources and in different languages. Unlike HS detection systems, resources and tasks in this field have received little or no coverage by previous review works (see Sect. 1), also due to their very recent spread: this, too, is why we chose to focus on this subject.

2.3 Analysis of search results

The works retrieved by our systematic search are critically analyzed and compared according to five dimensions:

- TYPE: what is the structure of the resource;
- TOPICAL FOCUS: how HS and related phenomena are distinguished according to their topical focus or targets, and to what extent such topics or targets are studied;
- DATA SOURCE: where data have been collected from;
- ANNOTATION: how and by whom data have been labeled, according to what framework, and how quality has been assessed;
- LANGUAGE: how different languages are covered, and how resources and definitions vary across languages.

Note that we deliberately excluded the high-level motivation for building a resource (e.g., automatic moderation, or monitoring and mapping the HS dynamics in a territory) from the dimensions used for their categorization. While some works explicitly mention their end goal, e.g., Sanguinetti et al. (2018) for monitoring, most do so implicitly at best, or do not indicate a motivation at all.

Overall, we have found 64 original resources, described in 60 papers published in journals or in conference proceedings (four papers present both a dataset and one or more lexica). Among these, 11 are resources specifically released as benchmark datasets for shared tasks, and are all available on request or by a public URL. As for the remainder, 23 are publicly available resources; 1 is available on request[9]; 29 resources are not available inasmuch as no valid URL is provided nor any other ways to access data is suggested. We have not performed further research in the attempt to find these latter resources; yet, since they are described in detail, we included them in this review.

We located 54 papers browsing Google or Google Scholar with the keywords *hate speech nlp*, *hate speech detection*, *dataset hate speech*, *hate speech lexicon*, *hate speech shared task* and *hate speech detection syntax*; 3 were found on GitHub and 3 on the ACL Anthology, both browsed with the keywords *hate speech*. Several entries appeared as results of more than one search string, but we associated them only with the first string that returned them.

In a few cases, more than one resource is described in one paper: some authors have built different corpora for comparison purposes, others extract one or multiple lexica from a dataset and describe all of them, others yet describe non-novel resources from which they derive a novel one. In all these cases, we count all items of the same type presented in a paper as one, and provide detailed explanations when they are mentioned.

It is interesting to point out that all the material we found is dated from 2016 onward; more precisely, 5 resources were published in 2016, 13 in 2017, 24 in 2018,

[9] For all the available resources, see the URLs provided in Tables 11, 12 and 13 in "Appendix".

20 in 2019 and 2 in 2020[10]. This confirms how the task is in a very recent stage of development yet, but is at the same time growing popular in the NLP community.

Some resources will be mentioned more than once along the paper, according to the focus determined by each dimension, as we want to offer multiple perspectives on the present scenario and provide examples. For the sake of completeness, though, Sect. 4 gives an overview of all the resources and tasks included in our research.

3 Comparative analysis along five main dimensions

In this section, we describe the different strategies used to design and build resources for HS detection, according to the five dimensions of comparison introduced in Sect. 2.3, and will draw general observations on their characteristics.

3.1 Type

A primary distinction is to be made between **annotated corpora**, meant as collection of textual instances from various sources, each labeled across one or more dimensions, and **lexica**, i.e. lists of words or phrases related to a common semantic field. 56 of our resources are corpora, while 8 are lexica and four papers contain both a corpus and one or more lexica. Among corpora, 11 are benchmark datasets released for shared tasks.

3.2 Topical focus

The most relevant factor of diversity among resources is the **topical focus**, i.e., the specific topics and abusive phenomena addressed, which also may depend on the exact target towards which hate is directed. This may vary according to the reach of the key concept and to its definition. Not only there is a number of overlapping concepts, as shown in Fig. 1, but each of these is prone to subjectivity and can be defined by more or less fuzzy boundaries, depending on the cultural background, individual perception and so on.

Coherently with our search criteria, HS is the most frequently investigated topic, often combined with other related phenomena (see Fig. 2).

That HS is an extremely complex notion is well known to those familiar with the topic, and the variety of definitions proposed in the papers we found proves it. HS is often conveyed by means of rhetoric devices such as aggressive language, threats, slurs, obscenity, offenses and even sarcasm; yet, it can be expressed just as well without any of these devices. Furthermore, depending on the group it targets, it can be known as racism, misogyny or sexism, homophobia, islamophobia, anti-semitism, anti-gypsism, and more; yet, all these terms express phenomena that exist as well outside the boundaries of HS.

Such complexity explains the many attempts to investigate not only HS itself but also some of its characteristics, related either to the way of expressing hate or to the

[10] Our research is last updated on 2020, April 28.

Topical focus: Abusiveness (5); Aggressiveness (2); Anti-Roma (1); Child sexual abuse (1); Cyberbullying (2); Flames (1); Harassment (1); Homophobia (4); HS (36); Islamophobia (2); Obscenity, Profanity (3); Offensiveness (13); Personal Attacks (1); Racism (6); Sexism, Misogyny (9); Threats, Violence (1); Toxicity (1); White supremacy (1).

Fig. 2 Number of resources focusing on HS and/or other related phenomena

targeted group. Yet, a certain confusion lingers around this melting pot: some authors do not provide a clear definition of the phenomenon they propose to investigate, and take their meaning for granted. As also shown in Table 5, not all the papers surveyed in this work provide a definition or illustrative examples of the notions and categories adopted for the corpus annotation. This "I-know-it-when-I-see-it" approach allows quick progress on a task, but may compromise precision. For each of these notions there are prototypical instances on which everyone would agree on, and controversial ones that seem to match more than one definition, or none at all: this results in blurred lines between concepts, "twilight zones" where most of the disagreement lies. Such complexity explains the many attempts to leave behind binary "black and white" definitions and investigate finer shades of HS and similar concepts, be they related to the way of expressing hate or to the targeted group.

3.3 Data source

A second key distinction concerns the source from which data are retrieved. The **microblogging** platform Twitter[11] is by far the most exploited source, due to the relatively reduced length of texts and to a friendly policy on making data publicly available: 32 resources contain tweets, one of which (Olteanu et al. 2018) also features posts from the social aggregator Reddit[12], one (Nascimento et al. 2019) also retrieves comments from the 55chan[13] imageboard, while in two works (Bosco et al. 2018; Mandl et al. 2019) Facebook[14] comments are collected along with tweets. Other resources include as main source several other social media such as Facebook (Del Vigna et al. 2017; Ishmam and Sharmin 2019; Mossie and Wang 2020; Vu et al. 2019), Reddit (Nithyanand et al. 2017; Schäfer and Burtenshaw 2019; Sabat et al. 2019; Qian et al. 2019a), Gab (Qian et al. 2019a), and Instagram (Corazza et al. 2019). Users' comments to newspaper articles are collected in de Pelle and Moreira (2016), Kolhatkar et al. (2019), Nobata et al. (2016) Pavlopoulos et al. (2017), and Steinberger et al. (2017); de Gibert et al. (2018) use sentences from the well-known white-supremacist forum Stormfront; the dataset released for the Hate Speech Hackathon[15] contains posts from the Wikipedia

[11] https://twitter.com.

[12] https://reddit.com.

[13] http://www.55chan.org.

[14] https://facebook.com.

[15] https://www.swisstext.org/2018/workshops/Hackathon.html.

discussion forum; Hammer (2017) and Kumar Sharma et al. (2018) use comments from controversial Youtube videos[16].

Nearly all the resources feature user-generated public contents, mostly micro-blog posts, often retrieved with a keyword-based approach and mostly using words with a negative polarity. To address the problem of the biases introduced keyword-based data collection approaches in corpora development, which will be better discussed in Sect. 5, some authors have embraced alternative approaches or combined collection strategies, moving beyond the simple lexicon-based approaches. In some cases the keyword-based strategy is combined with retrieving the whole timeline from users or pages considered hateful, i.e., where it is likely to find hateful contents (Mubarak et al. 2017; Kumar et al. 2018a), or from discussion threads about controversial topics that can easily trigger a certain language (Hammer 2017), taking into account the caveat of collecting contents from a large variety of users. In (Basile et al. 2019; Fersini et al. 2018a) a combined approach has been applied to collect the hateful and misogynous tweets, by monitoring potential victims of hate accounts, downloading the history of identified haters and filtering Twitter streams with keywords. In few other cases (see Nascimento et al. (2019)), a sort of *a priori* classification is attributed to the texts according to the retrieval source, assuming that all the items collected from a given source can be considered hateful. Quite uniquely, Fišer et al. (2017) use a corpus extracted from an online platform that collects spontaneous reports by the Internet users of any material containing HS or child sexual abuse: the corpus is then checked by experts validation, assessing that more than 40% is not actually disturbing content and that only 3% can be considered illegal content.

An overall count of the number of resources by source is available in Fig. 3.

3.4 Annotation

We found that data annotation may be a relevant source of variability. For each resource, we considered the annotation framework, the labels used and the number and type of annotators involved. Due to space limitations, we will not describe each work in detail, but only the major trends we observed.

As for the **annotation scheme** and the label inventory, there are three main strategies. The first is a binary scheme: two mutually-exclusive values, (typically *yes/no*) to mark the presence or absence of a given phenomenon. The second is a non-binary scheme: more than two mutually exclusive or non-exclusive values, accounting either for different shades of a given phenomenon, such as *strong hate, weak hate, no hate* (Del Vigna et al. 2017), *overtly aggressive, covertly aggressive, not aggressive* (Kumar et al. 2018a), *hate speech, abusive but not hateful, non-offensive* (Mathur et al. 2018); or for several phenomena at the same time, such as *hate speech, aggressiveness, offensiveness, irony, stereotype* (Sanguinetti et al. 2018), *racism, sexism, both, neither* (Waseem and Hovy 2016), *toxic, severe toxic, obscene, threat, insult, identity hate* for the Hate Speech Hackathon dataset.

[16] https://youtube.com.

Source: Facebook (8); Fora (1); Gab (1); Google Image (1); Instagram (1); News websites (6); Other (6); Reddit (5); Twitter (32); Wikipedia (1); YouTube (2).

Fig. 3 Number of resources by data source. Lexica are not included in the count as they are not directly extracted from an external source. Resources with multiple sources are mentioned multiple times

The third strategy features multi-level annotation, with finer-grained schemes accounting for different phenomena. This is the most complex annotation scheme and typically involves both a number of different traits and a scale of variation. For example, Fišer et al. (2017) use a complex scheme that accounts for *typology*, *target* and *metadata* of Socially Unacceptable Discourse, where each dimension has one or two layers of labels; Nobata et al. (2016) distinguish between *clean* and *abusive* language, where the latter can be labeled as *hate speech*, *derogatory* or *profane*. Fersini et al. (2018a, b) distinguish different behaviors within the class *misogyny*, namely *stereotyping and objectification, dominance, derailing, harassment and threat, discredit*. Olteanu et al. (2018) use a complex non-binary, multi-level annotation scheme with several labels for each one of four dimensions, namely stance, target, severity and framing, while Basile et al. (2019) adopt a three-layer binary annotation for HS, aggressiveness and nature of the target (individual or group).

Researchers adopt a wide range of strategies also with respect to the **number and background of the annotators**. Again, we traced three main options: having data annotated by experts (be they developers themselves or other judges with knowledge of the subject), having them annotated by amateur/non-expert annotators recruited either as volunteers (often among students) or on a crowdsourcing platform—those used are FigureEight (now acquired by Appen[17] and previously known as Crowdflower) and Amazon Mechanical Turk[18]—or, finally, using an automatic classifier to assign labels.

While 15 works rely only on expert judges, 9 on crowdsourced annotation and 5 on a classifier, the remaining works use a combined annotation: some start by having a small sample annotated by experts and then obtain a larger corpus by crowdsourcing, others use a classifier but rely on experts or on crowdsourcing for validation. Nobata et al. (2016), for example, use news comments reported as "abusive" by users, but rely as well on both expert judges and crowdsourcing for validation. In some cases, it is not clear what "expert judge" means, whether someone who has a long experience in that specific subject or someone who has been briefly trained for performing the task, and whether judges have been provided detailed instructions and guidelines or just a generic definition of the labels. An interesting case is that of Waseem (2016), who recruited feminist and anti-racist activist as trained and experienced annotators.

[17] https://appen.com.

[18] https://www.mturk.com/.

Not all authors give detailed information about the annotation process[19]. Most of them mention how many annotators have been involved: numbers range from a few expert annotators up to a unrestrained community of non-experts or contributors on a crowdsourcing platform. Individual judges may annotate only part of the dataset, or partially overlapping subsets.

Overall, we report wide variability and sparsity among different approaches: each resource is built referring to *ad hoc* definitions of the phenomena addressed, shaped so as to be suitable for a specific purpose, but what often lacks is a wider view on the topic and an eye towards interoperability of resources.

Similar problems of sparsity and lack of data affects the measurement of **inter-annotator agreement**: again, 21 papers do not provide information about this, while those who do it adopt different measures according to the number of judges and labels. The measures mostly adopted are Cohen's κ, Fleiss' κ, Krippendorf's α or a plain numerical or percentage value. Values range from extremely high, as in Bohra et al. (2018) (Cohen's $\kappa = 0.982$ between two expert judges on a binary classification task) and in Hammer (2017) (two annotators agree on 98% of the binary labels on a small sample of the data), to extremely poor, as in Del Vigna et al. (2017) (Fleiss' $\kappa = 0.19$ among 5 trained judges on a non-binary scheme with 3 labels) and in Kolhatkar et al. (2019) (Krippendorff's $\alpha = 0.18$ among CrowdFlower contributors on a non-binary scheme with 4 labels). Such variability may depend on a number of factors: how complex the annotation scheme is, how many judges are involved and how well they have been trained, and more.

Generally speaking, we highlight two opposing trends. Some authors opt for more straightforward schemes and few annotators, trading off multiple annotation and computing inter-annotator agreement only on a small sample, with the aim of obtaining a large labeled corpus in a short time and be able to use it for training classifiers or extracting lexica. Others try to design complex schemes that account for different dimensions and hues, and involve more than two annotators in an attempt to smooth individual biases; they might be more interested in modeling what certainly is a complex phenomenon, or to train sophisticated systems able to distinguish shades in natural languages.

In the case of shared tasks, even when the original dataset was annotated with complex and fine-grained scheme, a trade-off has been sought between the richness of the description and the data usability.

3.5 Language

Being English the *de facto* common language among scholars worldwide, we expected to find a great number of English resources. Indeed, 37 out of 64 are English corpora or lexica: yet, many other languages are represented too, and this certainly is of great value to an international community that seeks to tackle a worldwide social issue spread in many languages. An important role in releasing non-English resources is played by national evaluation campaigns and shared tasks ,

[19] Due to this, we are not able to provide a summarizing figure as the ones proposed for the other dimensions.

whose aim is exactly encouraging researchers to work on national languages. An effort emerges from Indian researchers to create baseline datasets in Hindi and promote research on dangerous contents on social media at large: the predominance of Hindi–English code–mixed data could be explained by the large spread of mixed forms and of Hindi words written in Latin script in non-formal online communication among Indians.

4 Overview by resource type

In the previous section, we outlined the main factors and issues related to building resources for HS detection, along five main axes of comparison, citing examples at need. In this section, we provide a synthetic overview of all the resources included in our review, based on their type: corpora, resources released for shared tasks, and lexica.

4.1 Hate speech corpora

The largest typology by number is that of annotated corpora, often specifically developed for training an automatic system and presented jointly, with observations on the performance and, sometimes, an error analysis. A classifier for HS (or any related phenomenon) is often, in fact, the paper's main focus—which is no surprise, as the development of solid classifiers outperforming the state of the art is the most lively area of this field. Our interest here remains nonetheless the linguistic resource, as we want to stress the importance of quality data for training quality systems. Among those works that train a classifier on a dataset built *ad hoc* by the authors themselves, the room left to the resource description and to the process that brought it into being vary considerably: in some cases it is little more than a section of the paper, in other cases it is broader and reports in details the important decision behind the final product. Essential information are almost always present: the most neglected piece of information concerns inter-annotator agreement, that is missing in 15 out of 44 corpora. Guidelines that clearly define the concept to be annotated, provide examples and suggest how to deal with difficult cases are also not always present.

Table 2 provides an overview of the resources along with their main characteristics. The label "*no*" in the column "Available" simply means that no URL to the resource is provided in the paper. For all the remaining resources, a link to the data is provided in Table 11.

As for the number of citations in the right-most column, we relied on Google Scholar for this information, but we opted for not reporting the exact number measured on a given day, as such number is volatile and may not be the most reliable indicator of the actual impact of a resource. Instead, we mapped each number to an interval, as we believe that the reader can get a clearer first-sight understanding of the order of magnitude of each resource's impact. Such intervals are as follows: < 10, < 50, < 100, < 250, < 500, where the upper bound of each class is the lower bound of the next class.

Table 2 Essential information of all the annotated corpora included in the review and briefly described in Sect. 4.1

Refs.	Focus	Language	Size	Av.	Cit.
ABP	Homophobia	ita	1859	No	< 10
AKM	HS	ara	6136	Yes	< 50
AMFE	HS	ind	1100	Yes	< 50
BVSAS	HS	hin-eng	4575	Yes	< 50
CKTG	HS	eng, fre, ita	15,024	Yes	< 10
CMCTV	HS	ita	6710	No	< 10
DCDPT	HS	ita	6502	No	< 100
DWMW	HS, racism, sexism, homophobia	eng	24,802	Yes	< 500
ENNVB	HS, personal attack	eng	27,330	Yes	< 50
FEL	HS, child sexual abuse	slv	13,000	Yes	< 50
FLKA	HS	swe	3056	No	< 10
GH	HS	eng	1528	Yes	< 50
GKH	HS	eng	62M	No	< 50
GPGC	White supremacy	eng	10,568	Yes	< 50
H	Threats, violence	eng	24,840	No	< 100
HUO	HS, abusiveness	ara	6039	No	< 10
IS	HS	ben	5126	No	< 10
KKS	Cyberbullying	eng	2235	No	< 10
KRBM	Aggression	eng, hin	39,000	No	< 50
KWCFST	HS	eng	1043	No	< 10
MDM	Obscenity, profanity, offensiveness	ara	33,100	No	< 100
MGANH	HS, racism, sexism, homophobia	eng	975	No	<10
MSSM	HS, abusiveness	hin-eng	3679	No	< 50
MW	HS	amh	491,424	No	< 10

Table 2 continued

Refs.	Focus	Language	Size	Av.	Cit.
NCCVG	Offensiveness	por	7672	Yes	< 10
NSG	Offensiveness	eng	168M	No	< 10
NTTMC	Abusiveness	eng	3,1M	No	< 500
OCBV	HS	eng	+150M	No	< 50
OLZSY	HS	ara, eng, fre	13,014	Yes	< 10
PBBPS	HS	ita	4000	No	< 10
PM	Offensiveness	por	2283	Yes	< 50
PMBA	Abusiveness	gre	1,5M	Yes	< 10
QBLBW	HS	eng	56,100	Yes	< 10
QEBW	HS	eng	3.5M	No	< 50
QEBW2	HS	eng	18,667	No	< 10
RRCCKW	HS	ger	541	Yes	< 250
SB	Offensiveness	eng	11M+	No	< 10
SBHK	Flames	cze, eng, fre, ita, ger	5077	Yes	< 10
SCG	hs	eng	5020	No	< 10
SPBPS	HS, Islamophobia, racism, anti-Roma	ita	6009	Yes	< 50
VY	HS	eng	1364	No	< 10
W	HS, racism, sexism	eng	6909	Yes	< 250
WH	HS, racism, sexism	eng	16,907	Yes	< 500
HSH	Toxicity, HS	eng	322,022	Yes	/
KTHS	HS	eng	49,161	on request	/

Due to space constraints, we adopted some shortening devices. As for column names, "Ref." = "Reference", "Av." = "Availability", "Cit." = "Number of Citations". The column "Reference" only reports the initial letter of the authors' last names or of the resource: each acronym is associated to the full-length citation in the resource description below. Language names have been shortened using the ISO 639-2/B standardized nomenclature for language classification. The size of the corpora is reported in terms of number of instances (e.g. tweets)

Table 3 Distribution of corpora for each language

Language	ISO	Reference	Count
Amharic	amh	MW	1
Arabic (all varieties)	ara	AKM, HUO, MDM	3
Bengali	ben	IS	1
Czech	cze	SBHK	1
English	eng	BVSAS, CKTG, DWMW, ENNVB, GKH, GPGC, H, HSH, KWCFST, KKS, KTHS, MGANH, MSSM, NSG, NTTMC, OCBV, QBLBW, QEBW, QEBW2, SB, SBHK, VY, W, WH	24
French	fre	CKTG, SBHK	2
German	ger	RRCCKW, SBHK	2
Greek	gre	PMBA	1
Hindi	hin	BVSAS, MSSM	2
Indonesian	ind	AMFE	1
Italian	ita	ABP, CKTG, CMCTV, DCDPT, PBBPS, SBHK, SPBPS,	7
Portuguese (all varieties)	por	NCCVG, PM	2
Slovenian	slv	FEL	1
Swedish	swe	FLKA	1

Languages full names are reported here next to their standardized code, in order to make abbreviations easier to understand across the tables, while resources are only cited by their acronym introduced in Table 2. Resources including multiple languages appear multiple times

We also summarized some of the salient features of the surveyed corpora along four dimensions of comparison, also described in Sect. 3, i.e. language, data source, annotation strategy and the presence in the relative paper of annotation guidelines.

Regarding the languages, as expected, most of the resources use English data, although in some cases they are collected along with texts in Hindi (Bohra et al. 2018; Kumar et al. 2018a; Mathur et al. 2018) or they are part of even larger multilingual collections (Chung et al. 2019; Ousidhoum et al. 2019; Steinberger et al. 2017). It is also worth pointing out that less-resourced languages such as Amharic, Bengali, Slovene and Swedish, are also represented in the corpora we found, thus enabling a greater linguistic diversity in this field. Table 3 shows the distribution of corpora for each of the represented languages.

As for data sources, the distribution shown in Table 4 confirms the general trend observed in Sect. 3.3, with Twitter establishing itself as by far the most exploited source. An interesting and promising effort is that by Sabat et al. (2019) and, partly, by Corazza et al. (2019), who mix up textual and visual data: although still at an early stage, this path could be explored further, given the amount of image-based

Table 4 Distribution of corpora for each source. Resources having multiple sources appear multiple times

Source	Reference	Count
Facebook	DCDPT, IS, KRBM, MW	4
Fora	FLKA, GPGC	2
Gab	QBLBW	1
Google Image	SCG	1
Instagram	CMCTV	1
News websites	GH, KWCFST, NTTMC, PMBA, PM, SBHK	6
Other	CKTG, FEL, HUO, NCCVG	4
Reddit	NSG, OCBV, QBLBW, SB, SCG	5
Twitter	ABP, AKM, AMFE, BVSAS, DWMW, ENNVB, GKH, KTHS, KRBM, MGANH, MSSM, MDM, NCCVG, OCBV, PBBPS, OLZSY, PBBPS, QEBW, QEBW2, RRCCKW, SPBPS, VY, W, WH	24
Youtube	H, KKS	2
Wikipedia	HSH	1

online communication that takes place everyday—including, of course, hateful language and violent propaganda by organized groups.

From Table 2 it can be observed that the resources size spans from a few hundreds to several million items: this information correlates with the collection and annotation procedure inasmuch as automatic methods allow for much larger data collection, while human labeling, especially if performed by a few experts, results in smaller dataset and require a greater effort. On the other hand, if many authors prefer to collect finer-grained and higher-quality annotation on smaller samples, this suggests a commitment to creating resources of higher quality, to exploring more complex nuances and to better understand how HS can be framed with NLP techniques. It is not rare that the two methods are combined: either starting from a manually annotated corpus, or a manually compiled list of terms, used as a seed to obtain a larger corpus or list by implementing a classifier; or, conversely, starting by automatically classifying a large dataset and then having a small subset annotated by experts for validation.

Overall, information provided by papers about the number, typology and characteristics of annotators is not homogeneous enough to aggregate data in a table effectively. Yet, we could aggregate corpora by the type of annotation strategy (or of classification, in case of automated labeling) and by whether each paper describes or at least mentions any guidelines developed for the annotation.

In Table 5 we refer to the same three main strategies described in Sect. 3.4, but we add four sub-types for the non-binary strategy. The sub-type "*no, low, high*" uses three labels to indicate a clean or neutral content (in other words, the absence

Table 5 Distribution of corpora for each annotation strategy

Strategy	Sub-type	Reference	Count
Binary		ABP, AMFE, BVSAS, CMCTV, GH, GKH, GPGC, H, KKS, KTHS, NCCVG, NSG, PBBPS, PMBA, RRCCKW, SB, SBHK, SCG	18
Non-binary	No, low, high	DCDPT, VY	2
	No, A, B	DWMW, HUO, MDM, MSSM, W, WH	6
	A, B, C +	IS	1
	scale	FLKA, PBBPS	2
Multi-level		AKM, CKTG, ENNVB, FEL, HSH, KRBM, KWCFST, MGANH, MW, NTTMC, OCBV, OLZSY, PM, QBLBW, QEBW, QEBW2, SPBPS	17
Other		PBBPS	1

For non-binary schemes, a farther distinction is proposed, based on the number and type of labels applied. Resources using multiple strategies appear multiple times

of the phenomenon), a weak intensity and a strong intensity. The sub-type "*no, A, B*" uses three labels to indicate a clean content, and the presence of one of the two phenomena considered. The distinction between these two sub-types emerged from the observation of our database: in the first case two different phenomena, e.g. abuse and hate, are considered as shades of the same concept, so that the stronger (hate) implies and contains the weaker (abuse) and they only differ quantitatively; in the second case, the two phenomena are qualitatively different and represent two separate concepts, so that they are mutually exclusive and do not overlap. This distinction does not depend on the concepts themselves, but only on the interpretation given by the authors, and despite being theoretically sound it was not always straightforward to apply. The sub-type "*A, B, C +*" is similar to the previous one, but makes use of more than two labels (plus a *clean* label). The last sub-type "*scale*" is somehow similar to the first one, but explicitly asks to rate the intensity of a phenomenon on a numeric scale of varying length, where numbers may be associated to short definitions. The only work in the type "*other*" uses a Best-Worst Scale, which is not comparable to other strategies.

Table 6, finally, shows that little more than half of the corpora we have found come with by guidelines that support the annotation process and provide explicit definitions of the concepts and instructions about how to label data. Among those that do provide guidelines, cases range from terse definitions to long and detailed descriptions for every class furnished with examples. It is likely that many of the works that provide no guidelines actually used some operational definitions or rules

Table 6 Distribution of corpora by presence of guidelines, meant as any kind of instructions for the human annotators: this may include a definition of the concepts and/or some examples for the classes to be annotated

Guidelines	Reference	Count
Yes	BVSAS, CKTG, CMCTV, DWMW, ENNVB, FEL, GH, GKH, GPGC, HUO, IS, KKS, KRBM, KWCFST, MSSM, NTTMC, OCBV, OLZSY, QBLBW, PBBPS, SBHK, SPBPS, VY, WH	24
No	ABP, AKM, AMFE, DCDPT, FLKA, H, MDM, MGANH, MW, NCCVG, NSG, PM, PMBA, QEBW, QEBW2, RRCCKW, SB, SCG, W	19
NA	HSH, KTHS	2

"NA" includes those papers from which it was not possible to determine whether any guidelines was provided to the annotators

for annotation: perhaps, especially for in-house labeling, they have not been formalized, or they may have left out for space constraints.

Akhtar et al. (2019) (marked as ABP in Table 2)—1859 tweets in Italian annotated as "*homophobic/ not homophobic*" by 5 trained volunteers. This dataset is used together with existing English datasets, reannotated for racism and sexism for the specific purpose of the research. Inter-annotator agreement for the novel dataset is measured with a Fleiss' $\kappa = 0.35$.

Albadi et al. (2018) (AKM)—about 6000 tweets in Arabic, annotated with crowdsourcing for religious hatred ("*hateful/ not hateful/ unclear or unrelated*") and for religious group (6 groups plus an "*other*" label). Agreement is measured as 81% for the first class and 55% for the second group. Three polarity lexicon for Arabic are released along with the dataset.

Alfina et al. (2017) (AMFE)—1100 tweets in Indonesian, annotated as "*HS/ no HS*" by 30 students. 100% agreement is reached on 713 tweets, then reduced to 520 in order to obtain a balanced dataset.

Bohra et al. (2018) (BVSAS)—4575 tweets in Hindi-English code-mixed variety, annotated as "*HS/ normal speech*" by two annotators. Agreement results in a Cohen's $\kappa = 0.982$.

Chung et al. (2019) (CKTG)—15,024 short text in English, French and Italian, consisting of HS–counterspeech (CS) pairs created *ad hoc* by experts. These pairs have been paraphrased, annotated by non-experts with multiple labels for HS type, HS sub-topic, CS type, and then translated from Italian and French to English so as to get parallel data across languages. This is one of the only two corpora built for the purpose of automatically generating CS.

Corazza et al. (2019) (CMCTV)—6710 Instagram posts in Italian, annotated as "*hateful/ not hateful*" by expert judges. This novel dataset is combined with existing Italian datasets from other sources for cross-genre analyses.

Del Vigna et al. (2017) (DCDPT)—6502 Facebook comments in Italian, sorted by target ("*religion/ physical or mental handicap/ socio-economical status/ politics/ race/ sex and gender issues/ other*") and annotated by five trained judges with the labels "*strong hate/ weak hate/ no hate*". Agreement is measured with a Fleiss' κ = 0.19 on comments with five annotations.

Davidson et al. (2017) (DWMW)—24,802 tweets in English, annotated with crowdsourcing as *HS; offensive but not HS; none*. Only 5% of tweets are annotated as *HS* by the majority. Authors propose a thorough error analysis on both human annotation and the performance of a classifier, distinguishing different topical focuses (racism, sexism, homophobia).

ElSherief et al. (2018) (ENNVB)—27,330 tweets in English, annotated with crowdsourcing as "*hateful [personal attack/ no]/ not hateful*". Agreement is measured as 92% for the hate class and 82% for the personal attack class.

Fišer et al. (2017) (FEL)—13,000 instances of online contents in Slovene reported by web users as hateful or containing child sexual abuse. Data are annotated by experts with a complex scheme that allows for coarse–, medium– and fine–grained annotation, and is based on the concept of Socially Unacceptable Discorse, which includes legally prosecutable expressions such as HS, threats, abuse and defamation, and non prosecutable expressions such as immoral insults and obscenities.

Fernquist et al. (2019) (FLKA)—3056 comments from Swedish web fora, annotated by trained students with a scalar scheme summed up as follows: "–3: aggression/–2: insult/–1: dislike/0: neutral". Agreement is measured with a Krippendorf's α = 0.9.

Gao and Huang (2017) (GH)—1528 comments in English posted on 10 discussion threads on the Fox News website. Comments are annotated as "*HS/no HS*" by two experts, with a very high agreement expressed as Cohen's κ = 0.98.

Gao et al. (2017) (GKH)—62 millions tweets automatically classified with a weakly supervised system trained on existing corpora, with a small sample of 1000 tweets annotated manually by two trained judges to evaluate accuracy. Agreement between the annotators is measured as Cohen's κ = 85%. The process include a seed list of slurs, manually compiled from existing lexica, which is shown in the paper; this list is then automatically expanded and exploited for the automated detection of hateful tweets.

de Gibert et al. (2018)(GPGC)—10,568 English sentences extracted from the right-wing forum Stormfront and manually annotated by three experts as "*HS/no HS*"; the labels "*skip*" and "*relation*" (meaning that the sentence can only be understood in relation to its context) are also used. Average percentage agreement among annotators on the four labels is 90.97%.

Hammer (2017) (H)—24,840 English sentences from YouTube comments posted under videos related to controversial topics. Sentences are labeled as "*threatening or violent/ clean*" by one judge, except a small subset of 120 sentences annotated by a second judge in order for agreement raating purposes, resulting in a 98% agreement.

Haddad et al. (2019) (HUO)—6039 social media comments in Tunisian Arabic, annotated by three trained judges as "*hateful/ abusive/ normal*", with an observed agreement of 81%.

Ishmam and Sharmin (2019) (IS)—5126 Facebook comments in Bengali, annotated by three trained judges into six classes, namely "*HS/ inciteful/ religious hatred/ communal hatred/ religious comment/ political comment*", where the first four labels identify overall hateful comments while the other two identify non-hateful comments. Inter-annotator agreement is given for each class, averaging a percentage of 0.78%.

Kumar Sharma et al. (2018) (KKS)—2235 Youtube comments in English posted below controversial videos, annotated as "*insulting/ not insulting*" in relation to cyberbullyism detection (used in a broad sense).

Kumar et al. (2018b) (KRBM)—39,000 texts between tweets and Facebook comments in Hindi-English code-mixed variety, annotated by with a multi-level scheme based on verbal aggression. The first level identifies "*overtly aggressive/ covertly aggressive/ not aggressive*"; the second level, which applies only to aggressive texts, identifies the discursive role '*attack/ defend/ abet*" and the discursive effect (ten categories based on the reason of the aggression). The annotation develops in two stages: a first exploratory annotation is performed by experts, and results in a few minor changes to the scheme; the second stage is done with crowdsourcing, and reaches an agreement of 72% for the first level and of 57% for the discursive effect.

Kolhatkar et al. (2019) (KWCFST)—1043 English comments from a Canadian news website, annotated with regard to four dimensions: constructiveness and toxicity (annotated with crowdsourcing), negation and appraisal (annotated by experts). As for the toxicity, four scale-like labels were available: "*very toxic/ toxic/ mildly toxic/ not toxic*".

Mubarak et al. (2017) (MDM)—three resources for Arabic language including: a lexicon of 288 obscene words; a test set of 1100 tweets for manual validation; a dataset of 32,000 comments that have been removed from the popular news website AlJazeera. The test set is annotated with crowdsourcing as "*obscene/ offensive but not obscene/ clean*", reaching a 87% agreement rate.

Martins et al. (2018) (MGANH)—975 tweets in English labeled with a complex multi-level scheme. Starting from the dataset released by Davidson et al. (2017), authors first perform statistical analysis to assess its reliability for HS detection; then extract a subset of 975 tweets, already labeled as "*HS/offensive but not HS/none*", and automatically assign to each tweet an emotion (using the model created by Plutchik (1980)), a score for the intensity of the emotion "*anger*" on a 0-1 scale, a score for polarity on a 0–1 scale, and a flag if the tweet matches any offensive word included in the *HateBase* lexicon.

Mathur et al. (2018) (MSSM)—3679 tweets in Hindi-English code-mixed variety, annotated by 10 experts as "*HS/abusive/ not offensive*".

Mossie and Wang (2020) (MW)—5876 Facebook posts along with 485,548 Facebook comments in Amharic, annotated by trained students as "*HS/no HS*" and then as the intent of "*ethnic/ religious/ political/ economic*" status.

Nascimento et al. (2019) (NCCVG)—7672 posts from Twitter and 55chan (an imageboard website) in Brasilian Portuguese. Data are automatically classified as "*offensive/not offensive*" during on the collection process, combining their source and some filters based on the emotional categories in the LIWC lexicon for Brasilian Portuguese.

Nithyanand et al. (2017) (NSG)—168 millions offensive Reddit comments in English, retrieved by a classifier that was trained on an existing dataset and two lists of offensive words.

Nobata et al. (2016) (NTTMC)—three corpora of comments in English from the news websites Yahoo!News and Yahoo!Finance. The primary dataset contains 2 millions comments annotated as "*abusive/clean*" by Yahoo's internal staff, and is used to train a classifier which in turn is used to retrieve a second dataset of 1,1 million comments covering a broader time span. A third, smaller dataset of a few thousands comments is built for evaluation, and annotated by three trained raters as "*abusive/ clean*" and for the sub-category of abuse ("*hate/ derogatory language/ profanity*"). Agreement rate is 0.922 and Fleiss' κ is 0.843.

Olteanu et al. (2018) (OCBV)—150+ millions items from Twitter and Reddit, plus a list of 1,890 unique terms contained in the data. Such terms are annotated with crowdsourcing using a complex scheme that includes for dimensions: stance ("*favorable/unfavorable/commentary/neutral*"), target ("*Muslims/other religious groups/Arabs/ethnic groups/immigrants/other groups*"), severity ("*promotes violence/ intimidates/offends or discriminates*") and framing ("*diagnoses causes/suggests solutions/both*").

Ousidhoum et al. (2019) (OLZSY)—13,014 tweets in Arabic, English and French, annotated with crowdsourcing using a multi-level scheme that accounts for directness ("*direct/indirect*"), hostility ("*abusive/hateful/offensive/disrespectful/ fearful/normal*"), target ("*origin/gender/sexual orientation/religion/disability/ other*"), group ("*individual/woman/special needs/African descent/other*") and the feeling aroused in the annotator by the tweet ("*disgust/shock/anger/sadness/ fear/confusion/indifference*"). Agreement is measured for each language as Krippendorf's $\alpha = 0.153$ (English), 0.244 (French), 0.202 (Arabic).

Poletto et al. (2019) (PBBPS)—4000 tweets in Italian, to which three different schemes are applied with crowdsourcing. The first scheme is a binary choice ("*HS/ no HS*"); the second is an unbalanced rating scale ("*– 3/– 2/– 1/0/1*") that encompasses content, tone and intention of the tweet; the third is a Best-Worst Scale, where annotators are presented with randomized sets of four tweets at a time and are asked to pick the most and the least hateful.

de Pelle and Moreira (2016) (PM)—10,336 comments in Brasilian Portuguese from a news website, 1250 of which are annotated by three judges as "*offensive/not offensive*" and for the target or reason of the offense ("*racism/sexism/homophobia/ xenophobia/religious intolerance/cursing*"). Two different dataset are obtained by computing the agreement, one with majority agreement (2/3) and one with full agreement.

Pavlopoulos et al. (2017) (PMBA)— 1,5 million comments in Greek from news portal, retrieved along with a label "*accept/ reject*" referring to the website comment moderation.

Qian et al. (2019a) (QBLBW)—56,100 posts in English from Gab and Reddit, arranged in dialogical structure as retrieved from the source, plus 41,730 counterspeech (CS) responses. The annotations collected with crowdsourcing include labeling which turns in the conversation are HS, and for each of them an instance of CS freely proposed by the contributor. This is one of the only two corpora built for the purpose of automatically generating CS.

Qian et al. (2018) (QEBW)—3,5 millions hateful tweets in English, associated to 40 U.S.-based hate groups and referencing 13 hate ideologies. Tweets are automatically labeled as for group and ideologies on the basis of the retrieval process.

Qian et al. (2019b) (QEBW2)—18,667 hateful tweets in English, retrieved from a starting list of 2,105 hate symbols used by hate groups, which is in turn collected from Urban Dictionary. Symbols in the list come from the source associated to one of the following tags: "*hate/racism/racist/sexism/sexist/nazi*".

Ross et al. (2017) (RRCCKW)—541 tweets in German, annotated with the labels "*HS/ no HS*" and with a discrete value for offensiveness on a 1–6 rating scale. Annotation is performed in two rounds: first by six experts, then by two separate groups of non-expert, only one of whom is showed a definition of HS. Agreement is admittedly low, with a Krippendorf's α ranging between 0.18 to 0.29.

Schäfer and Burtenshaw (2019) (SB)—more than 11 millions Reddit posts and comments in English, organized in a dialogical structure. Every post or comment is automatically assigned an offensiveness probability by an algorithm trained on a dataset annotated as "*offensive/not offensive*".

Steinberger et al. (2017) (SBHK)—5077 comments from news websites in Czech, English, French, Italian and German, annotated as "*flames/no flames*". Annotation was performed by three experts for English and Czech, and by one expert for the other languages. Agreement is measured for English and Czech with different metrics, all scoring little below 0.6.

Sabat et al. (2019) (SCG)—5020 memes, containing images and words (in English), collected from Google Images and from Reddit. Classification is based on the collection process: all memes obtained from Google Images (distinguished between "*racist/jew/muslims*" are assumed to be hateful, while all memes retrieved from Reddit are assumed to be non-hateful.

Sanguinetti et al. (2018) (SPBPS)—6009 tweets in Italian, annotated partly by experts and partly with crowdsourcing. A multi-level scheme is applied, accounting for HS, stereotype, irony (labeled as "*yes/no*"), aggressiveness and offensiveness (labeled as "*no/weak/strong*"), plus the intensity of HS when present (labeled with a rating scale from "*1—mildest*" to "*4—strongest*"). Agreement is measured with a Kohen's κ = 0.45 between experts and with a Krippendorf's α = 0.38 among crowdsource contributors.

Vidgen and Yasseri (2020) (VY)—4000 tweets in English, annotated by experts as "*not islamophobic/weakly islamophobic/strongly islamophobic*". Agreement is measured with different metrics: percentage = 89.9%, Fleiss' κ = 0.837, Krippendorf's α = 0.895. The final dataset is reduced to 1364 in order to have a balanced distribution.

Waseem (2016) (W)—6909 tweets in English, expanding the dataset presented in Waseem and Hovy (2016). Tweets are labeled as "*sexist/racist/neither*", first by expert judges, then by crowdsource contributors. Agreement is measured as κ = 0.57.

Waseem and Hovy (2016) (WH)—16,907 tweets in English, annotated as "*sexist/racist/both/neither*" by expert judges. Agreement is measured as κ = 0.85.

Two resources are presented separately because they differ in nature from all the resources described so far. In fact, they are not associated to a scientific paper that describes their features and gives details about their creation or usage. Nonetheless, since they are made publicly available for research competition purpose and they appear among the results of our systematic query, we decided to include them in this review. Yet, considering that such competitions were organized in a slightly different way compared to traditional shared tasks—no information on participating systems, nor on their results, was given—-, we decided to classify them as generic (not benchmark) corpora.

Hate Speech Hackathon (HSH) is a workshop held within SwissText 2018, the 3rd Swiss Text Analytics Conference, where participants where invited to train and test supervised classifiers for HS detection. The resource includes about 300,000 comments from English Wikipedia discussions and is annotated with the labels "*toxic/ severe toxic/ obscene/ insult/ threat/ identity hate*".

The Kaggle Twitter Hate Speech (KTHS) dataset is a resources released in 2018 on the Kaggle platform with the purpose of training supervised systems for HS detection. It includes about 49,000 tweets in English annotated as "*hateful/not hateful*". It is not possible to assess its impact in terms of citations, but some statistics can be found on the Kaggle webpage of the resource: from its release on July 2018 it collected 8994 views and 1527 downloads, with a quite constant trend (verified on May, 5th 2020).

4.2 Shared tasks

Several corpora found in our systematic search have been developed with the purpose of organizing *shared tasks*, i.e., open scientific competitions where benchmark data are made available and participants are invited to submit the prediction of their systems and a discussion of their methods.

Eleven shared tasks were organized in the context of international (SemEval) and national[20] evaluation campaigns of NLP technologies, while one was organized as part of the Workshop on Trolling, Aggression and Cyberbullying (TRAC-1). In all instances, the original data was collected from social media (Twitter and Facebook), and annotated manually by experts but integrating in two cases crowdsourced annotations. The tasks, with their main focus, are summarized in Table 7.

HS (against multiple targets) is the main topic in **HaSpeeDe** (Bosco et al. 2018), one of the tasks organized at EVALITA 2018; while, more specifically, HS against women is addressed to in the two editions of **AMI** (Fersini et al. 2018a, b) and in **HatEval** (Basile et al. 2019) (which, in turn, included data also on HS against

[20] Namely EVALITA, FIRE, GermEval, IberEval, PolEval, and VLSP.

Table 7 Shared Tasks on HS detection (HS), aggressiveness (AG) and offensiveness (OF) identification as main task with specific focuses, languages involved, size of datasets, number of participating teams and number of citations of the overview paper

Name	Event	Task	Focus	Lang.	Size	Teams	Cit.
AMI	IberEval 2018	HS	Misogyny	eng spa	8115	11	< 50
AMI	EVALITA 2018	HS	Misogyny	eng, ita	10,000	16	< 50
HASOC	FIRE 2019	HS, OF	–	eng, ger, hin	17,657	37	< 50
HaSpeeDe	EVALITA 2018	HS	Racism, generic	ita	8000	9	< 50
HatEval	SemEval 2019	HS	Misogyny, racism	eng, spa	19,600	74	< 100
HSD	VLSP 2019	HS, OF	–	vie	25,431	14	<10
–	GermEval 2018	OF	–	ger	8541	20	< 100
task 6	PolEval 2019	HS	Cyberbullying, generic	pol	11,041	9	<10
TRAC-1	TRAC 2018	AG	–	eng, hin	15,000	30	< 100
OffensEval	SemEval 2019	OF	–	eng	14,100	115	< 100

In this table, we adopt the same conventions as in Table 2

immigrants), and a focus on cyberbullying is proposed in **Task 6 at PolEval** (Ptaszynski et al. 2019).

Despite our focus being HS, we retrieved shared tasks on related phenomena such as aggressive identification (AG) and offensive language detection (OF). Among these, **TRAC-1** (Kumar et al. 2018a) deals with online aggression, trolling, cyberbullying and other related phenomena, while in **MEX-A3T** (Álvarez-Carmona et al. 2018), aggressive language detection is one of the two tracks set for the competition. Offensive language is the main track of **OffensEval** (Zampieri et al. 2019b, a) and the corresponding task at GermEval campaign in 2018 (Wiegand et al. 2018b).

Finally, two competitions explicitly focused on the identification of both HS and offensive language, i.e. **HASOC** at FIRE 2019 (Mandl et al. 2019) and **HSD**, the HS detection task on Vietnamese at VLSP campaign in 2019 (Vu et al. 2019).

In some cases, the need to account for the complexity of the phenomena dealt with is reflected in the type of predictions required to participating systems, often going beyond the simple binary classification: this is done either by proposing a non-binary classification or by introducing finer-grained sub-tasks aiming at detecting even more specific aspects.

The former scheme was followed in TRAC-1, where a distinction between overtly and covertly aggressive is drawn, and in the HS detection task at VLSP

2019, where a three-way classification was proposed to distinguish among hateful, non-hateful but offensive and neither hateful nor offensive content.

With the exception of HaSpeeDe 2018, the remaining competitions were rather organized around a first binary-classification task and one or more additional sub-tasks aimed at further specifying the binary scheme. In HatEval, systems were asked to classify hateful tweets as aggressive or non aggressive, and to determine whether the target was a single person or a whole group; the latter aspect was included also in both editions of AMI (task B), along with the detection of the type of misogynistic behavior, and in task C of OffensEval: here, the posts classified as targeted insults in task B (in contrast to generic insults) were to be further distinguished as targeted to individuals, groups or other (events, organizations, etc.). In **GermEval** 2018, the fine-graned sub-task consisted in the classification of the type of offense detected in the main task , which can be a profanity, an insult or the strongest type of offense, defined as abuse.

In task 6 at PolEval 2019, harmful tweets had to be classified as either examples of cyberbullying or of HS. Finally, in HASOC two additional sub-tasks aimed at labeling non-neutral content in posts as either hateful, offensive or profane (sub-task B) and to distinguish whether posts contained generic, non-acceptable language or rather insults or threats towards specific individuals or groups (sub-task C).

The high participation recorded by most of the shared tasks, also considering the short span of time they took place in, not only is indicative of the interest of the international community towards the problem of HS detection, but also encouraged the organizers to propose new editions of such competitions: at the time of writing, the second edition of OffensEval[21] and the TRAC shared task[22] have recently closed (see Sect. 4.4), while the second editions of HaSpeeDe[23] and AMI[24] have just been launched. Interestingly, in the rerun of HaSpeeDe, the Hate Speech Detection shared task for Italian proposed for EVALITA 2020, the organizers chose to go beyond the simple binary classification (hateful vs not-hateful), giving space also to a pilot task on finer-grained aspects related, albeit indirectly, to HS, namely the presence of stereotypes referring to one of the targets identified within the task dataset (Muslims, Roma and immigrants). In fact, an error analysis of the best performing systems participating to the HaSpeeDe 2018 dataset (Francesconi et al. 2019) pointed out that the occurrence of these elements constitutes a common source of error in HS identification. Moreover, a second pilot task related to the syntactic realisation of HS is proposed, as a sequence labeling task aimed at recognizing nominal utterances in hateful tweets. The more systematic exploration of the relations between the presence of nominal utterance and populist rhetoric in hateful tweets was inspired by the preliminary investigations in (Comandini and Patti

[21] https://sites.google.com/site/offensevalsharedtask/.

[22] https://sites.google.com/view/trac2/shared-task.

[23] HaSpeeDe 2020: http://di.unito.it/haspeede20.

[24] AMI 2020: https://amievalita2020.github.io/.

2019), suggesting that the most hateful part of hateful tweets are often verbless sentences or verbless fragments[25].

The rerun of the Automatic Misogyny Identification proposed at EVALITA 2020 (AMI 2020) is featured, among other things, by a very interesting novelty related to the important issue of guaranteeing the fairness of the misogyny detection models and, therefore, to reduce the error due to unintended bias, a problem that was initially addressed in (Nozza et al. 2019). On this line, a dedicated subtask of AMI 2020 has been devoted to ask systems to discriminate misogynistic contents from the non-misogynistic ones, while guaranteeing the fairness of the model in terms of unintended bias, relying on an *ad hoc* synthetic dataset released next to the standard dataset including raw data[26].

4.3 Hate speech lexica

We found 8 lexica of HS published as resources (Table 8). However, a number of approaches to HS detection are based on the development of *ad-hoc* lexica that are not given the status of standalone resources by their authors. The user-generated lexicon from the project Hatebase[27] provides a small-sized English lexicon of HS-related terms, employed, among others, by Davidson et al. (2017), who present a list of 179 English words derived from HateBase. Wiegand et al. (2018a) propose two lexica of English abusive words, a base one of 1,650 entries and one of 8,478 expanded with a classifier, where each word is annotated as abusive or not abusive. Another, slightly larger, monolingual lexicon is distributed as part of the contribution of the approach to HS detection on Arabic social media by Mubarak et al. (2017). Three Arabic lexica are also automatically generated in Albadi et al. (2018), using different feature selection methods, i.e. Bi-Normal Separation, Chi-square test and Pointwise Mutual Information, thus resulting in the AraHate-CHI, AraHate-BNS and AraHate-PMI. Each resource consists of words and their relative score expressing its association to HS, and all of them are publicly available along with the resource they were extracted from (also included in our survey, see 4.1).

Olteanu et al. (2018) mentions a list of 163 hateful terms created indirectly from the lexicon presented in Davidson et al. (2017): they collect the most frequent words that co-occurr with those listed by Davidson, assuming that the latters are certainly a sign that the tweet is hateful, and that frequent words in hateful tweets are themselves likely to be hateful. The ONG *PeaceTech Lab* has distributed, as part of their humanitarian effort in central Africa, a report containing a lexicon of HS terms in several languages, including English, Fulani, Hausa, Igbo, Pidgin, and Yoruba[28]. In the report containing the lexicon, alternative words and spellings are provided for the hateful expressions. Qian et al. (2019b) mention 2,105 list of

[25] See details about the datasets being released in the task guidelines available here https://github.com/msang/haspeede/blob/master/2020/HaSpeeDe2020_Task_guidelines.pdf.

[26] See details about the datasets being released in the task guidelines available here: https://amievalita2020.github.io/how_to_partecipate/.

[27] https://hatebase.org/.

[28] https://www.peacetechlab.org/nigeria-hate-speech-lexicon.

Table 8 Summary of HS lexica found in our search. Where an explicit name for the resource has not been provided, we included in the table its corresponding reference. In this table, we adopt the same conventions as in Table 2. The size of the resources is reported in terms of number of lexical entries

Name/Reference	Focus	Language	Size	Av.	Cit.
AraHate-BNS/CHI/PMI (Albadi et al. 2018)	HS	ara	1523	Yes	< 50
(Davidson et al. 2017)	HS, racism, sexism, homophobia	eng	179	Yes	< 500
HurtLex (Bassignana et al. 2018)	abusiveness, offensiveness	53 languages	< 100,000	Yes	< 50
(Mubarak et al. 2017)	obscenity, profanity, offensiveness	ara	288	No	< 100
(Olteanu et al. 2018)	HS	eng	163	No	< 50
PeaceTechLab lexicon (Ferroggiaro et al. 2018)	HS	multilingual	< 1000	Yes	n.a.
(Qian et al. 2019b)	HS	eng	2105	No	< 10
(Wiegand et al. 2018a)	abusiveness	eng	1651/8479	Yes	< 50

hateful symbols—meant as acronyms, numbers, slang words and any other sign used by hate groups to convey hateful messages in a sort of coded language. The starting point is the Urban Dictionary, from where they collect 1,590 words which they expand adding alternative forms for the same symbol. Finally, HurtLex is a multilingual (53 languages) lexicon of offensive and hateful words, built semi-automatically from an originally handcrafted Italian lexicon (Bassignana et al. 2018), counting roughly 1000 to 10,000 word per language. The words in HurtLex are divided into 17 overlapping categories and marked for the presence of stereotype.

4.4 Resources beyond systematic search

During the systematic process of searching and reading papers, we often found multiple references to other resources. Many are cited in the "Related Work" Section as examples of similar outputs in the field, while some are directly exploited as a starting point for building a larger dataset, developing a classifier or extract a lexicon. Whatever the purpose, in most of these cases the reference paper for these resources either had already been included in our database or it would be included later, because it was found with our systematic search. Yet six of these papers did not appear in any of the searches we carried out. Sticking to the criteria we adopted, such works should be excluded by this survey, as they were not found with the only method we allowed ourselves to use. Still, after having stumbled upon them in

papers found systematically, and having verified that these six papers are regularly peer-reviewed and published and describe novel resources for HS, we could not simply ignore them.

We intend the rigorous approach of this survey as a guarantee for inclusivity and reproducibility, but it should not turn into a limit that prevents us from offering a picture of the current situation as exhaustive and up-to-date as possible. For this reason we decided to present these six resources in a separate paragraph, so to make clear that they fall outside the results of our systematic search, but also that they are no less important contributions to the field than all the others. We acknowledge that, despite our effort, it is very hard to include every existing work, and something may still go missing—especially in such a young and lively field. A systematic approach can at least limit losses and provide explanations for them. Here we briefly describe these resources, which anyway are not included in the previous Tables.

Founta et al. (2018) (FDCLBSVSK)—80,000 tweets in English annotated with crowdsourcing. In a preliminary round of annotation several labels are used, then merged into the following four: "*HS/abusive/spam/normal*".

Golbeck et al. (2017) (GAB)—35,000 tweets in English annotated as "*harassing/not harassing*" by 2 judges, plus a third one to settle cases in disagreement. Agreement is measured with a Cohen's $\kappa = 0.84$.

Ibrohim and Budi (2018) (IB)—2016 tweets in Indonesian, annotated with crowdsourcing as "*not abusive/abusive but not offensive/offensive*", with a minimum of three annotations per tweet.

Ibrohim and Budi (2019) (IB2)—13,169 tweets in Indonesian, annotated with crowdsourcing using a multi-level scheme, where the first level distinguishes "*HS/abusive/not HS*" and the second level, which only applies to hateful tweets, specifies the intensity ("*weak/moderate/strong*") and the category or target ("*religion/race/physical/gender/other*").

Mulki et al. (2019) (MHBA)—5846 tweets in Levantine Arabic, annotated by three trained judges as "*hateful/ abusive/ normal*", with an observed percentage agreement of 81%.

Zampieri et al. (2019a) (ZMNRFK)—14,100 tweets in English, annotated with crowdsourcing using a multi-level scheme. The first level distinguished "*offensive/not offensive*"; then offensive tweets are labeled as "*targeted insult/ untargeted insult*"; eventually, targeted insults can be labeled as "*individual/group/other*". Agreement is found between two annotators in about 60% of the cases, while a third judge intervened for the remainder. The paper describes in detail the "Offensive Language Identification Dataset" (OLID) used in the OffensEval shared task "Identifying and Categorizing Offensive Language in Social Media" (Zampieri et al. 2019b).

The same rationale explained above motivates the decision to include in this Section four recently held shared tasks, which did not appear in our search when it was conducted but whose existence can not be ignored. In a fast-developing field such as HS detection, the number of shared tasks is constantly growing: we describe the resources used in the two following tasks with the will to provide a complete and up-to-date list.

All four shared tasks are new editions of previously experimented formats. **MEX-A3T** (Aragón et al. 2019), held at IberLEF2019, focuses on authorship and aggressiveness detection in Mexican Spanish: the dataset is the same as 2018 edition's (see Table 7). The **GermEval 2019** Shared Task on the Identification of Offensive Language (Struß et al. 2019) is similar to the previous year's, with the adding of a third level of annotation. The dataset consists of 7025 tweets annotated as "*offensive/ not offensive*" and then, if offensive, as "*profanity/insult/abuse/ other*" according to the type of offense and as "*implicit/ explicit*" according to the language used. **OffensEval2020**, Multilingual Offensive Language Identification in Social Media (Zampieri et al. 2020) is the second edition of a shared task on offensive language organized at SemEval 2020. The task features corpora in five languages (Arabic, Danish, English, Greek, Turkish) annotated for offensiveness ("*offensive/non-offensive*"), type of offense ("*targeted/untargeted*") and target ("*individual/group/other*"). **TRAC-2** is the second Workshop on Trolling, Aggression and Cyberbullying, which proposed a rerun of the shared task on Aggression identification (Kumar et al. 2020). Participants were provided with a multilingual dataset of 5,000 texts from YouTube comments in English, Bangla and Hindi, annotated at two-levels for two different sub-tasks: "*overtly aggressive/covertly aggressive/non-aggressive*" (*Sub-task A: Aggression Identification Task*), "*gendered/ non-gendered*" (*Sub-task B: Misogynistic Aggression Identification Task)*. A description of the development of the multilingual annotated corpus can be found in (Bhattacharya et al. 2020).

5 Lexical analysis

Most corpora surveyed in this work are collected by querying social media APIs with lists of keywords. Such keywords are not necessarily explicitly abusive or offensive terms. In fact, they are often chosen to be neutral with respect to negative connotations, in order to collect both positive and negative instances of HS or otherwise abusive language—see for instance Sanguinetti et al. (2018). However, the keyword-based data collection process still introduces a bias in the data, in terms of the topics they cover, and therefore it impacts the representativity of the corpora.

Wiegand et al. (2019) analyze the topic bias in several abusive language corpora collected with keyword querying. They extract lists of words having strong correlation with abusive microposts by computing their Pointwise Mutual Information. The experiment shows that some datasets contain a degree of topic bias, with negative implications for their application in machine learning: a supervised system could learn that words related, e.g., to football, are indicative of HS.

We perform a similar analysis of the lexical content of the datasets subject of this work. Rather than PMI, we compute the *Weirdness index* (WI) of the words in each dataset, in order to extract the most characteristic words of each dataset. The WI was introduced by Ahmad et al. (1999) as an automatic metric to retrieve words characteristic of a *special language* with respect to their common usage in general language. According to this metric, a word is highly *weird* in a specific collection of documents if it occurs significantly more often in that context than in a general

language corpus. In practice, given a *specialist* text corpus and a *general* text corpus, the weirdness index of a word is the ratio of its relative frequencies in the respective corpora. Calling w_s the frequency of the word w in the specialist language corpus, w_g the frequency of the word w in the general language corpus, and t_s and t_g the total count of words the specialist and general language corpora respectively, the weirdness index of w is computed as:

$$Weirdness(w) = \frac{w_s/t_s}{w_g/t_g}$$

When applied to an annotated corpus of HS (treated as the specialized corpus), we expect that the words with high WI will reflect the most characteristic concepts in that corpus, those who distinguish it most from generic language.

We also postulate a variant of WI that takes the labels of the messages into account. We refer to such variant as *Polarized Weirdness Index* (PWI). In this variant, we compare the relative frequencies of a word as it occurs in the subset of a labeled dataset identified by one value of the label against its complement. Consider a labeled corpus $C = \{(e_1, l_1), (e_2, l_2), ...\}$ where $e_i = \{w_1, w_2, ...\}$ is an instance of text, and l_i is the label associated with the text where e_i occurs, belonging to a fixed set L (e.g., $\{HS, not - HS\}$). The *polarized weirdness* of w with respect to the label $l*$ is the ratio of the relative frequency of w in the subset $\{e_i \in C : l_i = l*\}$ over the relative frequency of w in the subset $\{e_i \in C : l_i \neq l*\}$ We hypothesize that high-PWI words from a class will give a strong indication of the most characteristic words to distinguish that class (e.g. hate speech) from its complement (e.g. not hate speech).

We compute the WI of all the words in the shared task datasets described in Sect. 4.2, in five languages: English, Italian, Spanish, Hindi, and German. For Italian and German, we use the frequency counts for general language from the ItWaC and DeWaC corpora (Baroni et al. 2009); for English, we compute the word frequencies from the British National Corpus (Clear 1993); for Spanish we compute the word frequencies from the Spanish Billion Word corpus (Cardellino 2016); for Hindi we use the Leipzig corpora collection (Goldhahn et al. 2012). For the sake of this analysis, we only performed a standard, light preprocessing involving tokenization and ignoring cases. We also do not apply a smoothing scheme, effectively assuming that every word in the specialized corpus is also present in the general corpus, and simply setting $WI = 0$ when this is not the case.

For illustrative purposes, we report the 20 highest ranking words according to their WI and PWI on both classes in the HatEval dataset (English subset), as an example, in Table 9. From the first column, it is evident that this dataset has a strong topic bias towards politics, with high-WI words related to such topic, e.g. *maga* (the popular *Make America Great Again* pro-Trump slogan), *obama* (Democrat U.S. President), *salvini* (rightwing Italian politician), *gop* (the Republican Party). Looking at the high-PWI words, the most characteristic words in the HS-labeled tweets of HatEval are, as expected, related to negative connotations of the targets, e.g., *womensuck*, *nomorerefugees*, *invading*, and so on. However, the analysis reveals a bias where concepts related to immigrants are more represented than

Table 9 List of words from the English HatEval datasets with highest Weirdness Index (WI, left column), and highest Polarized Weirdness Index (PWI) for the HS class (center) and not-HS class (right column)

WI	PWI(HS)	PWI (not-HS)
maga	womensuck	ram
obama	:sweat_drops:	@refugees
wanna	indians	OSC
nigga	carolina	relief
skank	invasion	rohingya
niggas	Nomorerefugees	palestinian
kunt	invading	center
illegals	assimilate	latest
daca	invaded	blog
tweets	@diamondandsilk	@unmigration
tryna	detain	worldrefugeeday
salvini	invaders	provide
idk	peoples	sessions
cuz	cheated	shelters
tweeting	nerve	director
gop	@senategop	–
fuckin	pl	focus
wtf	deportillegals	lead
yall	skinny	inhumane
hoes	prosecuted	arrived

concepts related to women, while the two targets are supposed to be represented equally in the corpus. This kind of unbalance is a reflection of the strategies adopted to collect the data. In the HatEval English set, for instance, the number of keywords used for the two targets differ, and therefore the word distributions in the resulting corpora will be less natural. More in general, the use of keywords to retrieve potentially abusive messages is prone to introduce topic bias. To this effect, recent work is exploring the alternative route of collecting data for HS detection from "hateful" users (Ribeiro et al. 2018; Mishra et al. 2018).

We repeated the analysis on a selection of the corpora subject of this paper, in particular those pertaining to shared tasks. We computed the list of top-WI and PWI words according to the method described earlier in this section, inspected the resulting ranked lists of words, and manually assign a label to the most prominent semantic categories of the concepts found among the top-WI and top-PWI words. The results, presented in Table 10, summarize the topic bias emerging from this analysis. While some of the emerging topics are directly related to the datasets (e.g., misogyny and homophobia in the MEX-A3T data, collected for a shared task on the identification of such phenomena in test), others are orthogonal to the intended modeling goal of the corpora. Politics, in particular, is a highly represented topic in many datasets. Biases of this kind can be detrimental when corpora are used to benchmark HS detection systems (all the corpora examined in this section are from shared tasks), since they could reward systems that model HS in a specific, narrow domain.

Table 10 Topic bias emerging from the list of top-WI and PWI words in the shared task datasets

Dataset	Language	Topic bias
HatEval	English	U.S. politics
HatEval	Spanish	Immigrants
HaSpeeDe-TW	Italian	Italian Politics
HaSpeeDe-FB	Italian	Insults, TV
MEX-A3T	Spanish	Misogyny, homophobia
StackOverflow	English	Swear words, software development
GermEval	German	Politics
OffensEval	English	U.S. and world politics
AMI EVALITA	English	U.S. politics
AMI EVALITA	Italian	Misogyny, adult content, football
AMI IberEval	English	African American Vernacular
AMI IberEval	Spanish	Misogyny
TRAC-1	English	Religion
TRAC-1	Hindi	Religion

6 Discussion and conclusions

The high number of resources and benchmark corpora for many different languages developed in a very narrow time span, from 2016 onward, confirms the growing interest of the community around abusive language in social media and HS detection in particular. Being the subject in a yet recent stage, it suffers from several weaknesses, related to both the specific targets and nuances of HS and the nature of the classification task at large, that represent an obstacle toward reaching optimal results. It should be indeed observed that the features of the involved phenomena make them especially hard to model, and increase the risk of creating data that are biased or too much related to a specific resource (overfitting).

Some of these issues have been highlighted also by the previous surveys in the field (Lucas 2014; Schmidt and Wiegand 2017; Fortuna and Nunes 2018), whose *leitmotiv* revolves around the need for a common operational framework and benchmark resources. This recommendation is still valid, but recently steps forward have been taken, some issues are being tackled while others are emerging. For example, our survey captures a great availability of benchmark datasets for the evaluation of abusive language and hate speech detection systems, in several languages and with several topical focuses. This adds to the challenge of investigating architectures which are stable and well-performing across different languages and abusive domains, making it a more and more promising topic to research (Corazza et al. 2020; Pamungkas and Patti 2019; Ousidhoum et al. 2019).

As this survey shows, there are several interconnected phenomena at stake, but often only a specific aspect is dealt with. The field would highly benefit from a shared, data-driven taxonomy that highlights how all these concepts are linked and how they differ from one another. This would provide a common framework for researchers

who want to investigate either the phenomenon at large or one of its many facets. This direction is explored, for example, in a recent work by Fortuna et al. (2019).

Another major issue are biases in the design and annotation of corpora. For example, Sap et al. (2019) point out how annotated data may carry racial biases, and how widespread HS detection models can learn such biases. They show how some typical African American English, used with no derogatory intent, are mistaken for abusive language (like the word *"nigga"* used by African Americans): when a classifier is trained on such biased data, it will end up showing a negative bias towards content posted by African Americans. Topic bias is another factor to consider when developing resources for hate speech detection, as the results of our lexical analysis shows in Sect. 5. Recent studies are showing how the volatile nature of topics, especially on social media, can hinder the predictive capability of supervised models trained on data collected with particular keyword sets (Wiegand et al. 2019), or in restricted time spans (Florio et al. 2020).

With respect to this, an in-depth error analysis on the results of the systems trained on a given dataset can be an effective tool to highlight limits and biases in the data. Among the papers described in this review, this aspect is stressed in Davidson et al. (2017), who propose an error analysis on both human annotations and performance of a classifier, pointing out that offensive language is often mislabeled as hateful due to unclear definitions, and that human coders tend to consider racist or homophobic terms as hateful more frequently than they do with sexist terms. Another common source of errors is the one related to the presence of swear words, which in social media are often used in casual contexts, also with positive social functions. The lack of understanding of the different functions of swearing and pragmatic aspects related to vulgarity often lead to false positives in abusive language automatic identification, when swear words occur in non abusive contexts. Some recent studies started to address the problem, by proposing specific annotated resources to go towards a deeper investigation of these phenomena (Pamungkas et al. 2020; Holgate et al. 2018).

Especially in the context of shared tasks, where multiple systems are trained and tested on the same dataset, a thorough error analysis should be encouraged by the organizers, not just for the purposes of the system evaluation, but also to highlight any critical issue in the dataset scheme and its annotation. *A posteriori* analysis of the results of shared tasks are also helpful in gaining insights on the quality of the data, as done for instance for sentiment analysis in Basile et al. (2018). This, in turn, would contribute constructively to the debate on good practices to be adopted in the creation of high-quality corpora, when relating to such complex topics.

As for annotation schemes, in the surveyed works different perspectives and levels of granularity are assumed. Even if a standard form of annotation is still far, it often seems possible to recognize a common broad scheme beyond those implemented in existing resources. Fine–grained or multi–level annotation schemes start to be widely used in benchmark corpora for shared tasks, as they can be helpful, also for annotators, in order to better understand the dimensions of the observed phenomena during the development of the resources.

In addition, we noted that very few are the authors who give a detailed account of the guidelines used for annotation. More often only the labels of the scheme are

provided, with no further instruction on how to interpret them. This mostly happens when plain and straightforward labels are used, such as *"hateful/ not hateful"* or *"abusive/ not abusive"*, probably assuming that they do not need further explanation. Another possible reason might be the fact that sometimes the dataset description is framed within the broader description of the system used to perform a given task; more emphasis is therefore given to the experiment setups and the results obtained by the system, rather than to the theoretical issues related to the creation of the corpus. Yet, our research has shown that even apparently simple terms such as *"hateful"* or *"abusive"* convey complex and ambiguous concepts, which can be subject to various interpretations. Therefore, even though it is clear that detailed guidelines alone are not a solution to the many issues involved, an effort to clarify all the concepts and definitions used in the annotation scheme can still be useful to obtain high quality and comparable resources.

More boldly, Jurgens et al. (2019) call for a paradigm shift in the use of NLP technologies to address abusive language. Authors point out that only some phenomena along the spectrum of abusive content are actually addressed, while others are neglected for being either too subtle or quite rare. Their claim is that the whole range of toxic or abusive language should be dealt with, including common instances such as microaggressions and insults, because they too contribute to a negative environment. Furthermore, they encourage the community to adopt a proactive approach oriented to justice, claiming that the present attitude is reactive (it only tackles abusive content that has already been published) and oriented to moderation and censorship (it simply aims at the absence of explicit abuse, rather than to a positive environment). Chung et al. (2019) take a similar stand by creating a large corpus of HS and counter-speech pairs, thus focusing on positive responses rather than simply on the negative side. An added value of this work lies in the fact that annotators are NGOs activists, trained and experienced in contrasting and preventing HS: their insight might be especially valuable for building such resources.

The need for a new paradigm in the detection of HS and negative content at large develops from the awareness of the delicate social implications of such phenomenon. In fact, HS detection deals with an actual and serious problem that affects our society and is spreading fast, especially on the web (Gelber and McNamara 2016). With this respect, besides developing effective computational tools that tackle portions of the problem, it is of utmost importance to understand the phenomenon in its complexity and to work towards solutions that are positive for the society. A proactive, prevention-oriented attitude is then much needed, as is cooperation between academy, social platforms and public institutions.

Awareness of these issues and a comprehensive overview on the results achieved so far can certainly help researchers to gain a deeper understanding of the subject. Furthermore, it will allow the community to effectively take into account the specificities related to language and culture, and work towards counteracting HS and reducing unintended bias and stereotypes underlining the phenomenon.

Funding Open access funding provided by Università degli Studi di Torino within the CRUI-CARE Agreement.

Appendix

Below we provide the URLs to the available resources surveyed in this paper, specifically corpora (in Table 11), benchmark datasets (Table 12) and lexica (Table 13).

Table 11 List of the available corpora mentioned in Sects. 4.1 and 4.4 along with the link where they can be found or requested

Reference	URL
AKM	https://github.com/nuhaalbadi/Arabic_hatespeech
AMFE	https://github.com/ialfina/id-hatespeech-detection
CKTG	https://github.com/marcoguerini/CONAN
DWMW	https://github.com/t-davidson/hate-speech-and-offensive-language
ENNVB	https://github.com/mayelsherif/hate_speech_icwsm18
FDCLBSVSK[NS]	https://dataverse.mpi-sws.org/dataset.xhtml?persistentId=doi:10.5072/FK2/ZDTEMN&version=1.0
FEL	http://www.spletno-oko.si/english/
GAB[NS]	Request to http://jgolbeck@umd.edu
GH	https://github.com/sjtuprog/fox-news-comments
GPGC	https://github.com/aitor-garcia-p/hate-speech-dataset
HSH	https://drive.google.com/uc?id=1nKuo8wN0a1tAsaCB_6IrNYVOwhaSX3jw
	https://drive.google.com/file/d/1jcJ7BqwK7HAoDpX0jE5io0QQcE567rR3/edit
IB[NS] and IB2[NS]	https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection
KTHS	https://www.kaggle.com/vkrahul/twitter-hate-speech
MHBA[NS]	https://github.com/Hala-Mulki/L-HSAB-First-Arabic-Levantine-HateSpeech-Dataset
NCCVG	https://github.com/LaCAfe/Dataset-Hatespeech
OLZSY	https://github.com/HKUST-KnowComp/MLMA_hate_speech
PM	http://inf.ufrgs.br/~rppelle/hatedetector/
PMBA	http://nlp.cs.aueb.gr/software.html

Table 11 continued

Reference	URL
QBLBW	https://github.com/jing-qian/A-Benchmark-Dataset-for-Learning-to-Intervene-in-Online-Hate-Speech
RRCCKW	https://github.com/UCSM-DUE/IWG_hatespeech_public
SPBPS	https://github.com/msang/hate-speech-corpus
W and WH	https://github.com/zeerakw/hatespeech
ZMNRFK[NS]	https://competitions.codalab.org/competitions/20011

The [NS] superscript is used to mark the resources that were not found with our systematic search

Table 12 List of the shared task datasets mentioned in Sects. 4.2 and 4.4 along with the link where they can be found or requested (some URLs have been shortened due to space constraints)

Name	Event	URL
AMI	IberEval 2018	https://amiibereval2018.wordpress.com/important-dates/data/
AMI	EVALITA 2018	https://amievalita2018.wordpress.com/data/
		(For registered participants)
HASOC	FIRE 2019	https://hasocfire.github.io/hasoc/2019/dataset.html
HaSpeeDe	EVALITA 2018	https://shorturl.at/uvHQ0
		(Google form to get the data)
HatEval	SemEval 2019	https://competitions.codalab.org/competitions/19935
		(For registered participants)
HSD	VLSP 2019	https://vlsp.org.vn/vlsp2019/eval/hsd
MEX-A3T	IberEval 2018	https://mexa3t.wixsite.com/home/contact
		(For registered participants)
MEX-A3T[NS]	IberLef 2019	https://sites.google.com/view/mex-a3t2019/registration
		(For registered participants)
OffensEval	SemEval 2019	https://competitions.codalab.org/competitions/20011
		(For registered participants)
OffensEval[NS]	SemEval 2020	https://sites.google.com/site/offensevalsharedtask/results-and-paper-submission
–	GermEval 2018	https://github.com/uds-lsv/GermEval-2018-Data
task 2[NS]	GermEval 2019	https://projects.fzai.h-da.de/iggsa/data-2019/
task 6	PolEval 2019	https://github.com/ptaszynski/cyberbullying-Polish
TRAC-1	TRAC 2018	https://github.com/kmi-linguistics/trac-1
TRAC-2[NS]	TRAC 2020	https://sites.google.com/view/trac2/home

The [NS] superscript is used to mark the shared task reports that were not found with our systematic search

Table 13 List of the available lexica mentioned in Sect. 4.3 along with the link where they can be found or requested (some URLs have been shortened due to space constraints)

Name/Reference	URL
AraHate-BNS/CHI/PMI (Albadi et al. 2018)	https://github.com/nuhaalbadi/Arabic_hatespeech
(Davidson et al. 2017)	https://github.com/t-davidson/hate-speech-and-offensive-language
HurtLex (Bassignana et al. 2018)	http://hatespeech.di.unito.it/resources.html
PeaceTech Lab (Ferroggiaro et al. 2018)	http://shorturl.at/cjszS
(Wiegand et al. 2018a)	https://github.com/uds-lsv/lexicon-of-abusive-words

References

Ahmad, K., Gillam, L., & Tostevin, L. (1999). University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER). In: The Eighth Text REtrieval Conference (TREC-8), National Institute of Standards and Technology (NIST).

Akhtar, S., Basile, V., & Patti, V. (2019). A New Measure of Polarization in the Annotation of Hate Speech. In *Proceedings of the international conference of the Italian association for artificial intelligence* (pp. 588–603).

Albadi, N., Kurdi, M., & Mishra, S. (2018). Are they our brothers? analysis and detection of religious hate speech in the Arabic Twittersphere. In *Proceedings of the 2018 IEEE/ACM international conference on advances in social networks analysis and mining*, ASONAM 2018, IEEE (pp. 69–76).

Alfina, I., Mulia, R., Fanany, M.I., & Ekanata, Y. (2017). hate speech detection in the Indonesian Language: A dataset and preliminary study. In *Proceedings of 2017 international conference on advanced computer science and information systems (ICACSIS)*, IEEE.

Álvarez-Carmona, M., Guzmán-Falcón, E., Montes-y Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., Reyes-Meza, V., & Rico-Sulayes, A. (2018). Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish Tweets. In *Proceedings of the third workshop on evaluation of human language technologies for Iberian Languages (IberEval 2018)*, CEUR.org (pp. 74–96).

Aragón, M.E., Álvarez-Carmona, M.Á., Montes-Y-Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., & Moctezuma, D. (2019). Overview of MEX-A3T at IberLEF 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets. In *1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), CEUR-WS.org, CEUR Workshop Proceedings* (vol. 2421, pp. 478–494).

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel, F., Rosso, P., & Sanguinetti, M. (2019). SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of SemEval 2019, Association for Computational Linguistics (ACL)* (pp. 54–63).

Basile, V., Novielli, N., Croce, D., Barbieri, F., Nissim, M., & Patti, V. (2018). Sentiment polarity classification at EVALITA: Lessons learned and open challenges. *IEEE Transactions on Affective Computing*,. https://doi.org/10.1109/TAFFC.2018.2884015.

Bassignana, E., Basile, V., & Patti, V. (2018). Hurtlex: A Multilingual Lexicon of Words to Hurt. In *Proceedings of the fifth Italian conference on computational linguistics (CLiC-it 2018)*, CEUR.org (pp. 1–6).

Bhattacharya, S., Singh, S., Kumar, R., Bansal, A., Bhagat, A., Dawer, Y., Lahiri, B., & Ojha, A.K. (2020). Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of*

the second workshop on trolling, aggression and cyberbullying, European Language Resources Association (ELRA), Marseille, France (pp. 158–168), https://www.aclweb.org/anthology/2020.trac-1.25/.

Bohra, A., Vijay, D., Singh, V., Akhtar, S.S., & Shrivastava, M. (2018). A dataset of Hindi–English code-mixed social media text for hate speech detection. In *Proceedings of the second worskshop on computational modeling of people's opinions, personality, and emotions in social media, Association for Computational Linguistics (ACL)* (pp. 36–41).

Bosco, C., Dell'Orletta, F., Poletto, F., Sanguinetti, M., & Tesconi, M. (2018). Overview of the evalita 2018 hate speech detection task. In *Proceedings of the sixth evaluation campaign of natural language processing and speech tools for Italian*. Final Workshop (EVALITA 2018), CEUR.org.

Capozzi, A.T., Lai, M., Basile, V., Poletto, F., Sanguinetti, M., Bosco, C., Patti, V., Ruffo, G., Musto, C., & Polignano, M., et al. (2019). Computational linguistics against hate: Hate speech detection and visualization on social media in the" contro l'odio" project. In *6th Italian conference on computational linguistics*, CLiC-it 2019, CEUR-WS, Bari, Italy (vol. 2481, pp. 1–6).

Cardellino, C. (2016). Spanish billion words corpus and embeddings. https://crscardellino.github.io/SBWCE/.

Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., & Granitzer, M. (2020). I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of The 12th language resources and evaluation conference*, European Language Resources Association, Marseille, France (pp. 6195–6204), https://www.aclweb.org/anthology/2020.lrec-1.760.

Chung, Y. L,. Kuzmenko, E., Tekiroglu, S. S., & Guerini, M. (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*, Association for Computational Linguistics (pp. 2819–2829).

Clear, J. H. (1993). The British National Corpus. In P. Delany (Ed.), *Landow GP* (pp. 163–187). MIT Press: The Digital Word.

Comandini, G., & Patti, V. (2019). An impossible dialogue! nominal utterances and populist rhetoric in an Italian twitter corpus of hate speech against immigrants. In *Proceedings of the third workshop on abusive language online*. Association for Computational Linguistics, Florence, Italy (pp. 163–171), 10.18653/v1/W19-3518, https://www.aclweb.org/anthology/W19-3518.

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2019). Cross-platform evaluation for Italian hate speech detection. In *Proceedings of the sixth Italian conference on computational linguistics*.

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2020). A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology, Special Section on Emotions in Conflictual Social Interactions*,. https://doi.org/10.1145/3377323.

Davidson, T., Warmsley, D., Macy, M., Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the eleventh international conference on web and social media, AAAI* (pp 512–515).

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd workshop on abusive language online (ALW2), Association for Computational Linguistics (ACL)*, (pp. 11–20).

de Pelle, R., & Moreira, V. P. (2016). Offensive comments in the Brazilian Web: A dataset and baseline results. In *Proceedings of the fifth Brazilian workshop on social network analysis and mining (BraSNAM 2016)* (pp. 510–519).

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian conference on cybersecurity (ITASEC17)*, CEUR.org (pp. 86–95).

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the twelfth international conference on web and social media, AAAI* (pp. 52–61).

EU Commission. (2016). Code of conduct on countering illegal hate speech online. https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en#theeucodeofconduct.

Fernquist, J., Lindholm, O., Kaati, L., & Akrami, N. (2019). A study on the feasibility to detect hate speech in Swedish. In *2019 IEEE international conference on big data (Big Data)*, 2019, IEEE, (pp. 4724–4729).

Ferroggiaro, W., Dolan, T., Gichuhi, C., Ya'u, Y., Hamza, I., Abdulsalam, L.U., Ibrahim, J., Murad, N., & Creed, D. (2018). Social media and conflict in South Sudan: A lexicon of hate speech terms. v4, PeaceTech Lab.

Fersini, E., Nozza, D., & Rosso, P. (2018a). Overview of the EVALITA 2018 task on automatic misogyny identification (AMI). In *Proceedings of the sixth evaluation campaign of natural language processing and speech tools for Italian*. Final Workshop (EVALITA 2018), CEUR.org.

Fersini, E., Rosso, P., & Anzovino, M. (2018b). Overview of the task on automatic misogyny identification at IberEval 2018. In *Proceedings of the third workshop on evaluation of human language technologies for Iberian Languages (IberEval 2018)*, CEUR.org, (pp. 1–15).

Fišer D, Erjavec, T., & Ljubešić, N. (2017). Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In *Proceedings of the first workshop on abusive language online, Association for Computational Linguistics (ACL)* (pp. 46–51).

Florio, K., Basile, V., Polignano, M., Basile, P., & Patti, V. (2020). Time of your hate: The challenge of time in hate speech detection on social media. Applied Sciences 10(12), 10.3390/app10124180, https://www.mdpi.com/2076-3417/10/12/4180.

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., & Nunes, S. (2019). A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*. Association for Computational Linguistics, Florence, Italy (pp. 94–104), https://www.aclweb.org/anthology/W19-3510.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4), 85:1–85:30.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proceedings of the twelfth international conference on web and social media*, ICWSM 2018, Stanford, California, USA, June 25–28, 2018, AAAI Press (pp. 491–500), https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17909.

Francesconi, C., Bosco, C., Poletto, F., & Sanguinetti, M. (2019). Error analysis in a hate speech detection task: The case of HaSpeeDe-TW at EVALITA 2018. In *CLiC-it, CEUR-WS.org, CEUR Workshop Proceedings*, (vol. 2481).

Gao, L., & Huang, R. (2017). Detecting online hate speech using context aware models. In *Proceedings of the international conference recent advances in natural language processing*, RANLP 2017, INCOMA Ltd., (pp. 260–266).

Gao, L., Kuppersmith, A., & Huang, R. (2017). Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the eighth international joint conference on natural language processing (Volume 1: Long Papers)*, Asian Federation of Natural Language Processing, Taipei, Taiwan (pp. 774–782).

Gelber, K., & McNamara, L. (2016). Evidencing the harms of hate speech. *Social Identities*, *22*(3), 324–341.

Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A. A., Gergory, Q., Gnanasekaran, R. K., Gunasekaran, R. R., Hoffman, K. M., Hottle, J., Jienjitlert, V., Khare, S., Lau, R., Martindale, M. J., Naik, S., Nixon, H. L., Ramachandran, P., Rogers, K. M., Rogers, L., Sarin, M. S., Shahane, G., Thanki, J., Vengataraman, P., Wan, Z., & Wu, D. M. (2017). A large human-labeled corpus for online harassment research. In *WebSci 2017—proceedings of the 2017 ACM web science conference* (pp. 229–233).

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the eight international conference on language resources and evaluation (LREC'12)*, European Language Resources Association (ELRA).

Haddad, H., Mulki, H., & Oueslati, A. (2019). T-HSAB: A Tunisian hate speech and abusive dataset. In *7th international conference on Arabic language processing* (pp. 251–263).

Hammer, H. L. (2017). Automatic Detection of Hateful Comments in Online Discussion. In: Maglaras, L., Janicke, H., Jones, K. (eds) Industrial networks and intelligent systems. INISCOM,. (2016). *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering* (Vol. 188, pp. 164–173). Cham: Springer.

Holgate, E., Cachola, I., Preoţiuc-Pietro, D., & Li, JJ. (2018). Why swear? Analyzing and inferring the intentions of vulgar expressions. In *Proceedings of the 2018 conference on empirical methods in*

natural language processing, Association for Computational Linguistics, Brussels, Belgium pp. 4405–4414, 10.18653/v1/D18-1471, https://www.aclweb.org/anthology/D18-1471.

Ibrohim, M. O., & Budi, I. (2018). A dataset and preliminaries study for abusive language detection in Indonesian social media. In *3rd international conference on computer science and computational intelligence 2018* (pp. 222–229).

Ibrohim, M.O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the 3rd workshop on abusive language online* (pp. 46–57).

Ishmam, A. M., & Sharmin, S. (2019). Hateful speech detection in public Facebook pages for the Bengali language. In: Wani, M.A., Khoshgoftaar, T.M., Wang, D., Wang, H., Seliya, N. (eds) *18th IEEE international conference on machine learning and applications, ICMLA 2019*, Boca Raton, FL, USA, December 16–19, 2019, IEEE (pp. 555–560).

Jurgens, D., Chandrasekharan, E., & Hemphill, L. (2019). A just and comprehensive strategy for using nlp to address online abuse. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, Association for Computational Linguistics (ACL) (pp. 3658–3666).

Kitchenham, B. (2004). *Procedures for performing systematic reviews*. Tech. rep.: Keele University, Department of Computer Science.

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., & Taboada, M. (2019). *The SFU opinion and comments corpus: A corpus for the analysis of online news comments*. Springer International Publishing

Kumar Sharma, H., Kshitiz, K., & Shailendra. (2018). NLP and machine learning techniques for detecting insulting comments on social networking platforms. In *Proceedings on 2018 international conference on advances in computing and communication engineering, ICACCE 2018*, IEEE (pp. 265–272).

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018a). Benchmarking aggression identification in social media. In *Proceedings of the first workshop on trolling, aggression and cyberbullying, Association for Computational Linguistics (ACL)* (pp. 1–11).

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2020). Evaluating aggression identification in social media. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*. European Language Resources Association (ELRA), Marseille, France (pp. 1–5), https://www.aclweb.org/anthology/2020.trac-1.1/.

Kumar, R., Reganti, A. N., Bhatia, A., & Maheshwari, T. (2018b). Aggression-annotated corpus of Hindi–English code-mixed data. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*

Lucas, B. (2014). *Methods for monitoring and mapping online hate speech*. Tech. rep.: University of Birmingham.

Malmasi, S., & Zampieri, M. (2018). Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, *30*(2), 187–202.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in Indo-European languages. In *Proceedings of the 11th forum for information retrieval evaluation*. Association for Computing Machinery, New York, NY, USA, FIRE '19 (pp. 14–17), 10.1145/3368567.3368584, https://doi.org/10.1145/3368567.3368584.

Martins, R., Gomes, M., Almeida, J. J., Novais, P., & Henriques, P. (2018). Hate speech classification in social media using emotional analysis. In *Proceedings of the 2018 Brazilian conference on intelligent systems, BRACIS 2018* (pp. 61–66).

Mathur, P., Shah, R., Sawhney, R., & Mahata, D. (2018). Detecting offensive tweets in Hindi–English code-switched language. In *Proceedings of the sixth international workshop on natural language processing for social media, Association for Computational Linguistics (ACL)* (pp. 1–9).

Merriam-Webster Online. (2009). Merriam-Webster Online Dictionary. http://www.merriam-webster.com.

Mishra, P., Del Tredici, M., Yannakoudakis, H., & Shutova, E. (2018). Author profiling for abuse detection. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1088–1098).

Mossie, Z., & Wang, J. H. (2020). Vulnerable community identification using hate speech detection on social media. *Information Processing and Management*, *57*(3), 102087.

Mubarak, H., Darwish, K., & Magdy, W. (2017). Abusive language detection on Arabic social media. In *Proceedings of the first workshop on abusive language online, Association for Computational Linguistics (ACL)* (pp. 52–56).

Mulki, H., Haddad, H., Bechikh Ali, C., & Alshabani, H. (2019). L-HSAB: A Levantine twitter dataset for hate speech and abusive language. In *Proceedings of the third workshop on abusive language online* (pp. 111–118).

Nascimento, G., Carvalho, F., Da Cunha, A. M., Viana, C. R., & Guedes, G. P. (2019). Hate speech detection using Brazilian imageboards. In *Proceedings of the 25th Brazillian symposium on multimedia and the web, WebMedia 2019* (pp. 325–328).

Nithyanand, R., Schaffner, B., & Gill, P. (2017). Measuring offensive speech in online political discourse. In *7th USENIX workshop on free and open communications on the internet (FOCI 17)*.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web (WWW'16), international world wide web conferences steering committee* (pp. 145–153).

Nockleby, J. T. (2000). Hate speech., Macmillan Reference USA, New York, NY (pp. 1277–1279).

Nozza, D., Volpetti, C., & Fersini, E. (2019). Unintended bias in misogyny detection. In *IEEE/WIC/ACM international conference on web intelligence*. Association for Computing Machinery, New York, NY, USA, WI '19 (pp. 149–155), 10.1145/3350546.3352512, https://doi.org/10.1145/3350546.3352512.

Olteanu, A., Castillo, C., Boy, J., & Varshney, K. R. (2018). The effect of extremist violence on hateful speech online. In *Twelfth international AAAI conference on web and social media, AAAI* (pp. 221–230).

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., & Yeung, D. Y. (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), Association for Computational Linguistics* (pp. 4675–4684).

Oxford English Dictionary. (2019). Oxford English Dictionary Online. https://www.oed.com.

Pamungkas, E. W., & Patti, V. (2019). Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics: Student Research Workshop*, Association for Computational Linguistics, Florence, Italy (pp. 363–370), 10.18653/v1/P19-2051, https://www.aclweb.org/anthology/P19-2051.

Pamungkas, E. W., Basile, V., & Patti, V. (2020). Do you really want to hurt me? Predicting abusive swearing in social media. In *Proceedings of the 12th language resources and evaluation conference*. European Language Resources Association, Marseille, France (pp. 6237–6246), https://www.aclweb.org/anthology/2020.lrec-1.765.

Pavlopoulos, J., Malakasiotis, P., Bakagianni, J., & Androutsopoulos, I. (2017). Improved abusive comment moderation with user embeddings. In *Proceedings of the 2017 conference on empirical methods in natural language processing*.

Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis*. Manhattan: Harper & Row.

Poland, B. (2016). Haters: Harassment, abuse, and violence online. Potomac Books, https://books.google.it/books?id=Jd4nDwAAQBAJ.

Poletto, F., Basile, V., Bosco, C., Patti, V., & Stranisci, M. (2019). Annotating hate speech: Three schemes at comparison. In *Proceedings of the sixth Italian conference on computational linguistics*.

Ptaszynski, M., Pieciukiewicz, A., & Dybała, P. (2019). Results of the PolEval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in Polish Twitter. In *Proceedings of the PolEval 2019 Workshop*.

Qian, J., Bethke, A., Liu, Y., Belding, E., & Wang, W. Y. (2019a). 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), Association for Computational Linguistics* (pp. 4755–4764).

Qian, J., ElSherief, M., Belding, E., & Wang, W. Y. (2018). Hierarchical CVAE for fine-grained hate speech classification. In *Proceedings of the 2018 conference on empirical methods in natural language processing, Association for Computational Linguistics*.

Qian, J., ElSherief, M., Belding, E., & Wang, W. Y. (2019b). Learning to decipher hate symbols. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics* (pp. 3006–3015).

Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A., & Meira Jr, W. (2018). Characterizing and detecting hateful users on Twitter. In *Twelfth international AAAI conference on web and social media*.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In *NLP4CMC III: 3rd workshop on natural language processing for computer-mediated communication*.

Sabat, B. O., Ferrer, C. C., & Giro-i Nieto. X. (2019). Hate speech in pixels: Detection of offensive memes towards automatic moderation. In *Proceedings of NeurIPS joint workshop on AI for social good*, 1910.02334.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018). An Italian twitter corpus of hate speech against immigrants. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC'18)*. European Language Resources Association (ELRA) (pp. 2798–2895).

Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N. A., Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., et al. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics, association for computational linguistics (ACL)* (pp. 1668–1678).

Schäfer, J., & Burtenshaw, B. (2019). Offence in dialogues: A corpus-based study. In *International conference recent advances in natural language processing, RANLP* (pp. 1085–1093).

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media, Association for Computational Linguistics (ACL)* pp. 1–10.

Steinberger J, Brychcín, T., Hercig, T., & Krejzl, P. (2017). Cross-lingual flames detection in news discussions. In *Volume: Proceedings of the international conference recent advances in natural language processing (RANLP 2017)*. INCOMA Ltd., (pp. 694–700).

Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., Klenner, M. (2019). Overview of GermEval Task 2, 2019 shared task on the identification of offensive language. In *Proceedings of the 15th conference on natural language processing (KONVENS 2019)*.

Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A., Nadal, K. L., et al. (2007). Racial microaggressions in everyday life: Implications for clinical practice. *American psychologist*, *62*(4), 271.

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

Vu, X. S., Vu, T., Tran, M. V., Le-Cong, T.&, Nguyen, H. T. M. (2019). HSD shared task in VLSP campaign 2019: Hate speech detection for social good. In *Proceedings of VLSP 2019*.

Warner, W., & Hirschberg, J. (2012) Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*. Association for Computational Linguistics, Montréal, Canada (pp. 19–26), https://www.aclweb.org/anthology/W12-2103.

Waseem, Z. (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, Association for Computational Linguistics (ACL)* (pp. 138–142).

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *Proceedings of the North American chapter of the association for computational linguistics: Human language technologies 2016, Association for Computational Linguistics (ACL)* (pp. 88–93).

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the first workshop on abusive language online*. Association for Computational Linguistics, Vancouver, BC, Canada (pp. 78–84), 10.18653/v1/W17-3012, https://www.aclweb.org/anthology/W17-3012.

Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (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*. (Long and Short Papers), Association for Computational Linguistics (ACL), (Vol. 1, pp. 602–608).

Wiegand, M., Ruppenhofer, J., Schmidt, A., & Greenberg, C. (2018a). Inducing a Lexicon of abusive words—A feature-based approach. In *Proceedings of the North American chapter of the association for computational linguistics: Human Language Technologies, Association for Computational Linguistics (ACL)* (pp. 1046–1056).

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018b). Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of the GermEval 2018 workshop, 14th conference on natural language processing (KONVENS 2018)* (pp. 1–10).

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019a). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies*. (Long and Short Papers) (Vol. 1, pp. 1415–1420).

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019b). SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th international workshop on semantic evaluation (SemEval-2019), Association for Computational Linguistics (ACL)* (pp. 75–86).

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, C. (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the 14th international workshop on semantic evaluation*. https://arxiv.org/abs/2006.07235, to appear. Preprint available on arXiv.