Dataset annotation in abusive language detection

Fortuna, Paula; Soler-Company, Juan; Wanner, Leo

Erstveröffentlichung / Primary Publication
Sammelwerksbeitrag / collection article

Digital
Communication
Research.de

Recommended citation: Fortuna, P., Soler-Company, J., & Wanner, L. (2023). Dataset annotation in abusive language detection. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 443–464). Digital Communication Research. https://doi.org/10.48541/dcr.v12.26

Abstract: The last decade saw the rise of research in the area of hate speech and abusive language detection. A lot of research has been conducted, with further datasets being introduced and new models put forward. However, contrastive studies of the annotation of different datasets also revealed that some problematic issues remain. Theoretically ambiguous and misleading definitions between different studies make it more difficult to evaluate model reproducibility and generalizability and require additional steps for dataset standardization. To overcome these challenges, the field needs a common understanding of concepts and problems such that standard datasets and different compatible approaches can be developed, avoiding inefficient and redundant research. This article attempts to identify persistent challenges and develop guidelines to help future annotation tasks. Some of the challenges and guidelines identified and discussed in the article relate to concept subjectivity, focus on overt hate speech, dataset integrity and lack of ethical considerations.

DOI 10.48541/dcr.v12.26

Paula Fortuna, Juan Soler-Company & Leo Wanner

Dataset Annotation in Abusive Language Detection

1 Why is the sharing of concepts and datasets important?

The last decade saw a rise in research in the area of hate speech and abusive language detection. The early period of this research was thereby characterized by a low number of publicly available datasets. A corresponding survey (Fortuna & Nunes, 2018), in which works on the topic until mid-2017 are reviewed, points out that the majority of the studies describe the collection and annotation of new datasets, but that only a few of those datasets were made available to the community. This is certainly a problem since progress in a field depends to a large extent on a critical comparison between different approaches and thus requires the sharing of resources. In the years following this survey, a lot of research has been conducted, with more datasets being introduced and new models put forward, and, as shown by a more recent survey (Vidgen & Derczynski, 2021), fortunately, the tendency to keep datasets locked away has changed: By 2020, 63 datasets were publicly available, making research more comparable and the types of data available for the detection of online hate speech, abuse, and harm more diverse. However, contrastive studies on the annotation of different datasets also revealed that other issues remain (Fortuna et al., 2020). In particular, the excessive amount of idiosyncratic interpretations of common terminology in the context

of abusive language analysis has been identified as problematic since multiple, often ambiguous, definitions and different interpretations of the same terms led to fragmented research and difficulties in data reuse. In more generic terms, ambiguous definitions make it more difficult to evaluate model reproducibility and generalizability and require additional steps for dataset standardization (Fortuna et al., 2021; Kumar et al., 2018; Vidgen et al., 2019). To overcome these challenges, the field needs a common understanding of concepts and problems, such that standard datasets and different compatible approaches can be developed to avoid inefficient and redundant research.

Sharing concept definitions and datasets is an essential component of mature research areas, and there are several reasons for this. *First*, without a common conceptual framework it is not possible to establish a dialog between the different research contributions. *Second*, annotating new data and sharing them promotes the study of new phenomena or different aspects of the same phenomena. Usually, gathering new data implies new annotation schemes and guidelines. These should be carefully constructed and documented when annotating broad concepts, such as hate speech. *Third*, and as already mentioned, shared datasets are essential for comparing results between different experiments and models. Hence, it is common that benchmarking datasets are established, which serve as a baseline for comparison and model evaluation. However, in contrast to other research areas, so far, no datasets have been established as standard in the field of abusive language detection. *Fourth*, data quality is of primary relevance and should not be taken for granted, so, along with shared data, it is necessary to provide evidence of its quality evaluation. Thus, it has been observed that several hate speech datasets pose issues regarding bias (Sap et al., 2019). *Finally*, from a pragmatic point of view, resource sharing avoids repeated work, namely in the form of concept definition and data annotation.

In view of the fact that for further advances in the field of abusive language detection, we urgently need to establish a common understanding of the basic concepts we are working with and share datasets that are based on these concepts, this article attempts to identify persistent intra- and inter-dataset challenges and develop guidelines to help future annotation tasks.

2 Challenges related to the annotation of abusive language

In what follows, the central challenges related to the annotation of abusive language datasets are analyzed. In this context, it is useful to distinguish between *intra*-dataset and *inter*-dataset quality challenges. The first concerns topics related to the annotation criteria and annotation procedure within one dataset, while the second concerns topics related to the coherence and compatibility of the annotation across different datasets. Note that this distinction also implies that a dataset that covers all the intra-dataset quality requirements may still be problematic when used and analyzed in comparison with other datasets. Table 1 lists these challenges and the guidelines aligned with them. The intra-dataset challenges concern concept definition, bias, data sharing, and ethics. As concept definition challenges, concept subjectivity, unclear definitions, and the varying generalization potential of coarse-grained categories can be identified. Bias related to dataset composition may originate in the focus on overt hate speech, in a reduced number of authors of the posts in the dataset, and in a lack of information on the annotation procedure. Data sharing is often an important challenge because of, for example, author privacy or copyright issues (the authored statements may not be shared because this would violate the rights of the author or infringe the copyright terms) or dataset integrity (the data may have undergone unwanted alterations, may not be accessible via the provided link, or may simply have been removed from the repository). The last of the intra-dataset challenges concerns ethics in the sense that the composition of a dataset may be used for harmful actions. The inter-dataset challenges concern the introduction of redundant and contradicting features across datasets.

In what follows, we address both the intra- and inter-dataset challenges (with a particular emphasis on concept definition, bias, data sharing, and ethics) in relation to abusive language.

2.1 Challenges in intra-dataset quality

In this section, the challenges related to intra-dataset quality in hate speech, as identified in Table 1, are discussed.

Table 1: Abusive language annotation challenges and the guidelines aligned with them

Level	Topic	Challenges	Guidelines
Intra-dataset	Concept definition	Concept subjectivity	Adopt a problem-driven approach
		Unclear definitions	Motivate definitions, tasks, and datasets socially
		Varying generalization potential of coarse-grained categories	Use clear category definitions
			Use coarse-grained categories with caution
			Prioritize fine-grained categories
			Match collected and targeted data
	Bias	Focus on overt hate speech	Increase versatile message search
		Reduced number of authors	Control communities, message threads, and author distribution
		Lack of information on annotation	Control covered time spans
			Provide information on the annotator profiles
			Define precise guidelines for annotation
	Data Sharing	Author privacy and copyrights	Protect the identity of the authors of the data and comply with copyright legislation
		Dataset integrity	Ensure data preservation and availability
			Include a data statement
	Ethics	Lack of ethical consideration	Follow ethical principles

Inter-dataset	Redundant data features	Diversify data characteristics
	Redundant and contradicting labels	Avoid redundant labels
		Provide definitions, examples, and justifications
		Position new concepts on the map of standardized categories

Challenges related to concept definition

Concept subjectivity. Defining the meaning of online hate speech, abuse, or harm is not a trivial task. The difficulty in providing definitions for these concepts arises from their subjectivity and the dependence of the connotation in the context in which a statement is made (see Litvinenko in this volume). For instance, according to some authors, the "N-word" is a slur no matter the context, while its intra-group usage may be considered harmless (Weir-Reeves, 2010); "*You son of a b****" is offensive in a neutral context but can be meant as an expression of admiration between friends; and the mention of the cultural background of an individual can be interpreted as racist or hate speech in some contexts. Furthermore, we cannot ignore that the public interpretation of the concepts of hate speech (or abusive language, in general) is also predetermined by legal regulations, such as the European Union Code of Conduct on Countering Illegal Hate Speech (European Commission, 2017), or the terms of use of social media platforms. That is, the determination of the interpretation and scope of the concepts of abusive language, offensive language, or hate speech underlying the research on their detection requires a thorough assessment of the different points of view and different contexts in which these may occur.

Unclear definitions. The worst lack of clarity with respect to central concepts or categories in the field is when data and annotation schemata do not provide any definitions at all, leaving what a certain data category actually represents open to interpretation. However, even when present, definition characteristics may not comply with the best standards. Low-quality definitions are vague, suffer from

being too generic, are defined in terms of negation of other categories (e.g., when "covert aggression"[1] is simply defined as the negation of "overt aggression"), or make an assumption with respect to the sensibility of the audience. This is, for example, pointed out by Vidgen et al. (2019) with respect to the concept of "offensiveness" by Davidson et al. (2017), which implies the question, "Offensive for whom?", because what is considered offensive by one audience, or in one context, might not be offensive elsewhere.

Varying generalization potential of coarse-grained categories. Categories are coarse-grained if they contain other subcategories (e.g., "hate speech" is a coarse-grained category when compared to "sexism" as a subcategory). In previous works, coarse-grained categories such as "hate speech" proved to be difficult to be generalized across datasets, while others like "toxicity" generalized well (Fortuna et al., 2021).

Challenges concerning bias

Online abuse is a rather sparse phenomenon if we consider the total volume of data on social media. This makes data collection a laborious task. Different strategies are applied to overcome this problem. However, these strategies may imply biases that are discussed in the next paragraphs.

Abusive message collection using keywords. The sparsity of online abuse data has led researchers to develop specific sampling techniques to increase their chances of retrieving abusive messages. The most common technique is to apply specific keywords for abusive message searches (e.g., derived from the Hatebase resource[2]). However, the use of specific keywords (and thus training on explicit abusive language posts) leads to a poor identification of posts with covert abusive language messages (Fortuna et al., 2021). More generally, the use of keywords, the focus on messages in communities or threads with a likely high percentage of abusive content, or sampling over relatively short time periods (Poletto et al., 2020) will necessarily generate datasets that have very specific characteristics, such that the modules trained on them are likely to perform less well on datasets with other characteristics.

1 Please note that we use single quotes for the names of abusive language categories.

2 https://hatebase.org/

Reduced number of message authors. Datasets related to hate speech or abusive language are often collected from a limited set of authors (Arango et al., 2019). If this fact is not taken into account during the partitioning of the information into training, development, and test data, messages from the same author can be randomly divided and may appear in both the training and the test data, which distorts the evaluation of the quality of the classification task. In other words, in this case, it is not possible to distinguish whether a model is capable of classifying hate speech or the content of particular authors.

Biased annotation. Guidelines for annotation are central when creating a new dataset as they will condition the data classification. Apart from the fact that when a hate speech dataset is published, the corresponding guidelines, if provided at all, are not always sufficiently clear and rigorous, the annotation will reflect the socio-cultural backgrounds of the annotators (see Kim in this volume). In the case that this is inevitable or tolerated, the socio-cultural bias characteristics should be well documented.

Challenges for data sharing

Once the data have been collected and annotated, nowadays, it is common practice to share the obtained dataset. However, dataset sharing may also put at risk the viability of the dataset.

Violation of authors' privacy and copyright legislation. In the majority of cases when social media text data are collected, no permission can be granted by the authors of the messages. This constitutes a potential violation of the authors' privacy and copyright issues. In order to comply with the legal regulations on data protection and privacy and not to invalidate the dataset as a whole, it is of utmost importance to strictly observe the data sharing and data use policies of the corresponding social media platforms. It is also essential to comply with the legislation related to copyrights for digital content. It is important to note that there are differences between the European and US legislation in this respect.

Loss of dataset integrity. A common practice is to provide only annotations and original IDs of the messages on the platforms where they have been spotted, with no direct access to the posted content of the dataset. However, in this case, there is a substantial risk that the content will be removed from the platform sooner or later, and thus not be accessible anymore, which is often the case when dealing

with abusive and harmful content. When this happens, the proportion of positive and negative classes changes, a new version of the dataset has to be created, and the advantages and purpose of sharing datasets get lost.

Ethical concerns

Another challenge that the area of hate speech automatic detection faces is that researchers do not always address *ethical concerns* related to their resources.

Lack of ethical considerations and data statements. Technological solutions need to adhere to ethical principles in order to ensure that harmful side effects are avoided. The main issues related to ethical principles are user privacy, bias, and dual use. User privacy and bias have already been discussed above. Let us thus focus on dual use. Here, dual use refers to the repurposing of abusive language detection technologies such that they cause harm. In the case of automatic hate speech and abusive language detection, the deployment of such technologies has already resulted in mistakenly flagging non-hateful discourses (Sarkar & KhudaBukhsh, 2021) and, even worse, the marginalization of some minority groups (see, e.g., Oliva et al., 2021).

To avoid pitfalls, the authors are morally obliged to anticipate how a new annotated dataset could be repurposed in a negative way and to design their data model in a way that does not cause harm. In the case of abusive language detection, research has not always been accompanied by proper ethical reflections, considerations, and terms of use. We discuss some possible solutions in the corresponding guidelines section.

2.2 Challenges of coherent annotation across different datasets

Abusive language occurs in different forms, thus potentially in different styles, and in different languages. Therefore, it is crucial that the research community can count on diverse datasets so that a representative sample of the spectrum of abusive language is covered. Furthermore, when annotating a dataset, it is important to be consistent with existing research in the field in order to avoid research duplication and contradiction. This results in at least two challenges.

Diversity of collected data across datasets

The majority of the available collected datasets in the field share a data modality, language, and platform. Thus, data are shared in text format, are mostly in English, and in their majority, stem from Twitter (see, e.g., Davidson et al., 2017; Waseem & Hovy, 2016). This reduces the variability of the available data and, if the data overlap, also results in data redundancy with respect to repetitive data.

Redundant and contradicting labels

There has been confusion in terms of concept definition and the usage of the terms related to hate speech, abuse, and harm. This is critical since the use of different terms for equivalent categories hampers the reuse of resources. For instance, it is not clear what the differences between generic concepts, such as "toxicity," "offensive" and "abusive" are. Moreover, almost equivalent concepts such as "sexism" and "misogyny" are not always used in the same way. Detailed analyses of the diverging terms in abusive language dataset compilation and the consequences of this divergence are discussed in detail in Fortuna et al. (2020, 2021).

3 Towards transparent dataset construction and annotation guidelines

As seen in the previous section, there are a series of challenges that may undermine the quality of resources in the field of abusive language detection and analysis. Inspired by the study of these challenges, we propose, in what follows, a set of guidelines for leveraging quality resources. As in the previous section, we distinguish between intra-dataset and inter-dataset aspects.

3.1 Intra-dataset quality guidelines

As already pointed out above, in order to reduce the subjectivity and ambiguity of *concept definitions* used in the field it is important to follow certain guidelines.

Guidelines for concept definition

Adopt a problem-driven approach. Task definition should follow, as much as possible, a holistic, problem-driven approach, rather than a data-motivated approach. In other words, the task formulation should motivate data collection, instead of the task being defined based on the available data (Gudivada et al., 2015).

Motivate definitions, tasks, and datasets socially. Online hate speech, cyberbullying, abuse, and harm infliction are inter-personal behaviors with a strong social component. As these behaviors are within the scope of different academic disciplines, researchers in the field of abusive language detection should reach out to other relevant disciplines before defining the task and annotating data material. Literature from humanities and social science (including, for example, law, sociology, and anthropology) may become an important source of insight, together with existing surveys in the field (e.g., Fortuna & Nunes, 2018; Poletto et al., 2020; Schmidt & Wiegand, 2017; Vidgen & Derczynski, 2021).

Use clear category definitions. One of the goals of future annotation tasks should be to establish a clear taxonomy with meaningful and theoretically sound categories. Several theoretical studies already outline possible procedures concerning how this can be done (Vidgen & Derczynski, 2021). In this context, we should aim for explicit, precise, and universal category definitions. Such clear category definitions are instrumental in high-quality annotation.

Use coarse-grained categories with caution. In view of the challenges of using coarse-grained categories, we may conclude that such categories should be used with caution. In the case that they serve a given task or purpose well, they need to be clearly defined (see above), and the phenomena that they are supposed to cover should be clearly delimited. Along with each coarse-grained category, more specific categories, which further detail this category, should be spelled out and annotated such that an error analysis on the model performance can be conducted in order to assess whether it equally detects all the subcategories of a generic class (Fortuna et al., 2020, 2021; Pamungkas & Patti, 2019).

Prioritize fine-grained categories. Irrespective of the guideline above, previous research suggests that in the case of hate speech, fine-grained categories are better suited than coarse-grained categories. Experiments show that when a model is trained and tested on fine-grained categories, such as "sexism" or "racism," better levels of generalization are achieved. This also further buttresses the argumentation

in favor of more fine-grained taxonomies of abusive language categories. Despite this general rule, it is also important to note that excessively detailed taxonomies may lead to an unbalanced distribution of the data across categories, such that certain categories may end up with only a few samples. This would, obviously, also be problematic for standard supervised machine learning models. A compromise between taxonomy detail and data availability is thus necessary, and the granularity of the taxonomy should be carefully analyzed and justified from the perspective of the goals of a given experiment or study (Fortuna et al., 2021).

Match training and target data. For machine learning models to work, data collected for training and data to which the model is then applied (either for testing or, in the case of practical applications, during routine use) should share properties. One basic requirement is to observe that both share similar features, such as text length, style, and topic. Otherwise, the model generalizability capacities are put at risk.

Guidelines for bias mitigation

Data sampling techniques may involve decisions and the application of strategies for data collection that imply bias. If such decisions or strategies cannot be avoided, the focus should be on the minimization of their negative consequences and on the documentation of the data collection procedure, such that researchers in the field can select the datasets and procedures that best fit the application they are targeting. In what follows, we outline some guidelines for bias mitigation.

Increase versatile message searches. As already mentioned above, a common practice during data collection is the use of explicit keywords for the identification of relevant messages. While this practice has the advantage that it ensures the presence of abusive content, it has the drawback of introducing vocabulary bias into the collected material. The use of explicit keywords should thus be avoided and replaced by more versatile methods of message identification. Should keyword-based searches be necessary, a list of the keywords used should be provided in the data statement.

Control communities, message threads, and author distribution. Another strategy for data collection is to gather data from specific threads or from profiles that belong to authors previously identified as posting a higher rate of abusive content. This type of sampling procedure also has limitations since, if not controlled, the number of message authors will be small, and, as a result, the dataset

will be biased with respect to their writing style. A possible way to control this problem is to make sure that a reduced number of messages per author is collected. Another alternative is to ensure that the distribution of the authors in the data collection is balanced. In other words, text from the same author should not be present in both training and testing sets (see also, e.g., Arango et al., 2019). Controlling this type of bias will improve the model generalization as the model will be less tuned to a very reduced number of authors.

Control covered time spans. The data should cover a broad time span. Thus, while obtaining, for instance, feedback on an election, it makes sense to only gather data over a short period of time, to obtain a realistic picture of the use of abusive language in a social medium, the data should cover a longer time period since samples with narrower timeframes will be more affected by exceptional events. Again, the covered time span should be protocolled in the data statement, and in the case that any societal events influence the tenor and content of the data, this should be recorded as well.

Provide information on the annotator profiles. For the annotation of abuse language datasets, annotators with different backgrounds can be drawn upon (see Vidgen & Derczynski, 2021):

- trained specialists (one of the most common options that, however, usually provides little information on the type of annotators' expertise);
- crowdworkers (an option that is prone to trade quality for quantity);
- professional moderators (usually employees of a social medium platform who annotate following the platform's policies);
- a mix of crowdworkers and experts; and
- synthetic data creation (less representative of real-world data).

The profile of any of these types of annotators will necessarily influence the way an annotation will be carried out (and thus what the final annotated dataset will look like). Therefore, the profiles of the annotators involved should be properly described, such that biases can be measured and counter-balanced. The main information to be recorded in a strictly anonymized way concerns:

- demographic features (age, gender, nationality),
- annotation expertise, and

- personal experience with abuse (i.e., whether the annotator was a victim of online abuse).

It is furthermore important to add relevant attitudes and beliefs. Thus, attitudes toward discrimination and political orientation are closely related to the capacity of evaluating online abuse; annotating racist material in research should not, for example, be left to the discretion of a prejudiced individual.

Define precise guidelines for annotation. The annotation guidelines should be transparent and comprehensive. Rules should account for difficult or counter-intuitive cases, and a set of shared practices should be developed. The rules should be enriched with clear and easy-to-understand examples. Ideally, experienced annotators will be involved in the development of the guidelines since only they know the language used in the material and can thus ensure that it is captured in appropriate consistent categories (Vidgen & Derczynski, 2021).

Guidelines for data sharing

Let us now have a look at the guidelines for data sharing to ensure data preservation over time.

Protect the identity of the authors of the data. The identity of the authors of the data related to abuse language research must be protected, if not strictly anonymized, during data collection and also during the training, evaluation, and sharing of the material. With regard to a published dataset, IDs or user names that allow for the direct retrieval of the material from social media should not be freely published in an open repository. When it is necessary to share this type of information, the data should be kept private and only be accessed strictly in accordance with the terms of use of the data of the social medium in question and the relevant legal regulations.

Ensure data preservation and availability. As already mentioned, sharing or making data publicly available risks violating terms and conditions of social media platforms. On the other hand, using IDs instead of providing actual data poses data integrity risks. If both types of risks cannot be discarded in a concrete case, a possible solution is to use synthetic data, which would also solve the issue of data privacy and offers the advantage of allowing a better control of data quality. The disadvantage synthetic data brings is the loss of variability.

Data donations by social media platforms are another alternative, as are data trusts, which also provide a framework for storing and accessing data and respective terms and contracts for data access (Vidgen & Derczynski, 2021).

Include a data statement. When making a dataset available, it is important to provide detailed information on all stages of the dataset creation. This includes information on the following:

- task definition (concept definition, taxonomy, related concepts, targeted groups);
- decisions taken with respect to the data collection;
- data sampling procedure (social network, socio-historic data context, e.g., comments on news about politics or sports, the time and location of the data collection);
- researchers' and annotators' backgrounds;
- annotation guidelines (interviews, steps, task design on platforms); and
- class-balancing procedures.

Only with proper data annotation and dataset documentation will it be possible to achieve more standardization in the field.

Guidelines ensuring ethical principles

Last but not least, dataset creation must follow guidelines that ensure that the compilation procedures and the obtained dataset are in line with ethical principles.

Follow ethical principles. As already pointed out, technological solutions need to adhere to ethical principles and ensure that the harm done when developing a technology is minimized. These principles also apply to dataset collection and annotation. In this case, the main issue concerns bias, as discussed in the previous paragraphs (Bender et al., 2020; Tomašev et al., 2020), and the privacy of the message author and target. Datasets should be accompanied by a data statement in which the procedure followed to compile the dataset, the introduced bias, and the dataset purpose are described. It is only recently that some researchers in the field have started to adopt this practice of automatic hate speech detection (e.g., Sap et al., 2020).

3.2 Guidelines for inter-dataset coherence

The guidelines related to inter-dataset coherence concern, first of all, four aspects that are discussed below.

Diversify data characteristics. English is the most common *language* for the analysis of hate speech, but since hate in social media is a global phenomenon, other languages have to be considered as well, prioritizing under-represented languages. Due to the increased popularity of multilingual approaches, it would also be valuable to annotate equivalent phenomena in different languages at the same time. Code-mixed textual material has been collected in the community as well, which is adequate to represent online communication using more than one language at the same time.

The most common *source* of hate speech material with which the community works is Twitter. However, this also raises the question of platform diversification—especially in view of the specific characteristics of Twitter messages. In the future, platforms other than Twitter should be studied such that abusive language of the communities that use other platforms is also captured.

Regarding the *modality*, the majority of datasets only contain textual material, while image, audio, or multimodal data can also be relevant. Furthermore, it is necessary to keep in mind that the context of the collected material provides essential clues for the assessment of whether certain data are abusive. In the case of texts, this can be achieved by collecting complete conversation threads, including the main stimulus invoking a thread (e.g., news, a comment, a video) and replies to it. For instance, certain communities use slurs as a sign of identity. Multimodal context information can help to identify them.

As far as the *dataset size* is concerned, supervised machine learning-based techniques require large amounts of high-quality annotated data. Automatically annotated datasets may help to create bigger data collections. However, even if quantity matters, it is important to ensure annotation quality—for instance, by contrasting a manually annotated data sample with the corresponding automatically annotated one. Finally, it would be advantageous to also annotate other dataset characteristics in terms of linguistic features, including, for example, "overture" ("covert abuse" vs. "overt abuse"), "irony," "sarcasm," "adversarial," etc. From the previous literature (Caselli et al., 2020; Fortuna et al., 2021; Sanguinetti et al., 2018), we know that online abuse correlates with these characteristics, and their annotation would help to better understand this correlation.

Avoid redundant labels. Given the amount of ongoing research and available datasets in the field of abusive language detection and classification, it is also necessary to position a new dataset in the context of the datasets that already exist. The community should avoid creating new categories to refer to concepts already present in the literature and move toward dataset standardization. Previous work has shown that categories such as "toxicity," "offensive," and "abusive" correlate well with each other and lead to good cross-dataset generalization when used as training categories (Fortuna et al., 2021). With this in mind, it is appropriate to introduce a generic category term, "abuse and harms" to replace "toxicity," "offensive," and "abusive." This term also captures the recent insight into the community reflected by the change in the title of the most popular workshop in the area from *Abusive Language Workshop* to *Workshop on Online Abuse and Harms*. In Fortuna et al. (2021), it was also observed that classifiers trained on the categories of "sexism" and "misogyny" achieved a cross-dataset generalization between both concepts, indicating that using the label of "sexism" to refer to both would avoid the need for an extra label.

Provide definitions, examples, and justifications. In the case that a new category is identified, clear examples and justification of why a new category is needed and in what way it enriches the field should be provided. Due to its importance, again, the process of the definition of a new category should be well documented and grounded, based on the insights from social sciences.

Position new concepts on the map of standardized categories. Previous research provides standardized categories that allow for the conversion between different datasets (Fortuna et al., 2021). In this study, different publicly available datasets on abuse in English are annotated with respect to their similarity and compatibility. In the future, studies on other, new, datasets should conduct the same type of analysis. However, the question of how to ensure dataset standardization may remain, and there is no simple answer to it.

In any case, existing dataset definitions and surveys of existing datasets should be taken into account, and already introduced notions and categories should be adopted whenever possible. For instance, if a dataset contains and is annotated with "sexism" and "racism" categories, the creators of the dataset may compare these categories with more generic categories, such as "hate speech," "abuse and harms" and assess to what extent the targeted phenomena relate to one of these categories (obviously, this does not mean that coarse-grained categories are to be preferred (see also our discussion above).

Careful data analysis can help detect similarity between datasets. This can consist of a comparison of dataset feature descriptions, application algorithms for text similarity detection, topic extraction, and class comparison or cross-dataset classification (Fortuna et al., 2020, 2021).

4 Prospects of uniform annotation across abusive language detection applications

The evolution from the early to the latest research in the field of abusive language detection shows that it is difficult to predict in advance all the problems and nuances related to defining tasks and collecting and annotating data in this field. However, the field has also advanced considerably. While in the early era, proprietary datasets were created, and rarely generalizable models were developed, this tendency has changed in recent years. Now it is time to identify the remaining challenges and to agree collectively on strategies aimed at achieving a more mature research area.

In this article, we enumerated what we consider to be the central challenges of the field, which include the need for better and clearer concept definitions; the lack of data diversity in terms of languages and the platforms analyzed; the introduced bias when collecting, annotating, and publishing data; and the creation of new data resources that are compatible with the previous research in the field. To address these challenges, guidelines, which are summarized in the following set of instructions, were discussed and proposed:

- find solid theoretical ground (from social sciences and previous research in the field) and prefer clear fine-grained definitions;
- diversify data (e.g., find new data source languages and provide the data context);
- mitigate bias by controlling the message search, data properties, and data annotation (e.g., provide information on authors, topics, dates, annotation procedures);
- ensure data availability, but at same time, protect the authors of the data;

- Well document the data and the methodologies followed to compile them (e.g., include a data statement); and
- follow ethical guidelines.

Steps toward more maturity with respect to dataset collection and annotation can be observed. Datasets are becoming more diverse, with new languages and modalities being annotated (de Gibert et al., 2018; Suryawanshi et al., 2020). Data quality is being discussed (Vidgen & Derczynski, 2021), and datasets and annotation schemas (Fortuna et al., 2020, 2021) are being compared in search of good practices. Platforms that ensure data availability while observing content author privacy are also beginning to emerge.[3]

Another tendency that can be observed involves gathering and merging existing resources and building new annotation schemes based on this material, instead of always collecting and annotating new datasets, as was done in earlier research. This leads to more extensive and alternative collections of data (Sap et al., 2020).

Paula Fortuna is a final year PhD student at the Department of Information and Communication Technologies of the Universitat Pompeu Fabra of Barcelona, Spain. https://orcid.org/0000-0002-2306-9276

Juan Soler-Company is a data scientist in Pepsico. Previously he was a postdoctoral researcher at the Natural Language Processing group of Universitat Pompeu Fabra of Barcelona, Spain. https://orcid.org/0000-0002-8645-0162

Leo Wanner is ICREA Research Professor at the Department of Information and Communication Technologies of the Universitat Pompeu Fabra of Barcelona, Spain. https://orcid.org/0000-0002-9446-3748

Acknowledgments

The first author is supported by the research grant SFRH/BD/143623/2019, provided by the Portuguese National Funding Agency for Science, Research, and Technology, Fundação para a Ciência e a Tecnologia (FCT), within the scope of the *Human Capital* Operational Program (POCH), supported by the European Social

3 https://hatespeechdata.com/

Fund and by national funds from MCTES. The work of the second and third authors has been supported by the European Commission in the context of the H2020 Research Program under contract numbers 700024 and 786731.

References

Arango, A., Pérez, J., & Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 45–54).

Bender, E. M., Hovy, D., & Schofield, A. (2020). Integrating ethics into the NLP curriculum. In A. Savary & Y. Zhang (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6–9). https://www.aclweb.org/anthology/2020.acl-tutorials.2/

Caselli, T., Basile, V., Mitrovic, J., Kartoziya, I., & Granitzer, M. (2020). I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive language. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, ..., & S. Piperidis (Eds.), *Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020* (pp. 6193–6202). https://www.aclweb.org/anthology/2020.lrec-1.760/

Davidson, T., Warmsley, D., Macy, M. W., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the 11th International Conference on Web and Social Media* (pp. 512–515). https://doi.org/10.48550/arXiv.1703.04009

de Gibert, O., Pérez, N., Pablos, A. G., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. In D. Fiser, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, & J. Wernimont (Eds.), *Proceedings of the 2nd Workshop on Abusive Language Online* (pp. 11–20). https://doi.org/10.18653/v1/w18-5102

European Commission. (2017). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Tackling illegal content online, towards an enhanced responsibility of online platforms.* Reference: *COM* (2017)555.

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

Fortuna, P., Soler-Company, J., & Wanner, L. (2020). Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 6788–6796).

Fortuna, P., Soler-Company, J., & Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing Management*, *58*(3), 102524. https://doi.org/10.1016/j.ipm.2021.102524

Gudivada, V. N., Baeza-Yates, R., & Raghavan, V. V. (2015). Big data: Promises and problems. *Computer*, *48*(3), 20–23. https://doi.org/10.1109/MC.2015.62

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018). Benchmarking aggression identification in social media. *Proceedings of the 1st Workshop on Trolling, Aggression and Cyberbulling (TRAC)*, 1-11.

Le, T., Wang, S., & Lee, D. (2020). Malcom: Generating malicious comments to attack neural fake news detection models. *Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM)* (pp. 282–291).

Oliva, T. D., Antonialli, D. M., & Gomes, A. (2021). Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LQBTQ voices online. *Sexuality & Culture*, *25*(2), 700–732.

Pamungkas, E. W., & Patti, V. (2019). Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* (pp. 363–370).

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, *55(2),* 1–47.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018). An Italian Twitter corpus of hate speech against immigrants. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, ..., & T. Tokunaga (Eds.), *Proceedings of the 11th International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7–12, 2018*. http://www.lrec-conf.org/proceedings/lrec2018/summaries/710.html

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics* (pp. 1668–1678). https://doi.org/10.18653/v1/p19-1163

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2020). Social bias frames: Reasoning about social and power implications of language. In D. Jurafsky, J. Chai, N. Schluter, & J. R. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5477–5490). https://doi.org/10.18653/v1/2020.acl-main.486

Sarkar, R., & KhudaBukhsh, A. R. (2021). Are chess discussions racist? An adversarial hate speech data set. *In Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35 (pp. 15881–15882).

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In L.-W. Ku & C.-T. Li (Eds.), *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media* (pp. 1–10). https://doi.org/10.18653/v1/w17-1101

Suryawanshi, S., Chakravarthi, B. R., Arcan, M., & Buitelaar, P. (2020). Multimodal meme dataset (multioff) for identifying offensive content in image and text. In R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, & D. Kadar (Eds.), *Proceedings of the 2nd Workshop on Trolling, Aggression and Cyberbullying* (pp. 32–41). https://www.aclweb.org/anthology/2020.trac-1.6/

Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D. C., Ezer, D., van der Haert, F. C., Mugisha, F., Haert F.C., Mugisha F., Abila G. (2020). AI for social good: Unlocking the opportunity for positive impact. *Nature Communications*, *11*(1), 1–6.

Vidgen, B., & Derczynski, L. (2021). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, *15*(12), 1–32. https://doi.org/10.1371/journal.pone.0243300

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. *Proceedings of the 3rd Workshop on Abusive Language Online* (pp. 80–93).

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. *In Proceedings of the NAACL Student Research Workshop* (pp. 88–93).

Weir-Reeves, J. (2010). Is the N-word acceptable? *The Temple News*. https://temple-news.com/is-the-n-word-acceptable/