Automatic Identification of Harmful, Aggressive, Abusive, and Offensive Language on the Web
A Survey of Technical Biases Informed by Psychology Literature

Balayn, Agathe; Yang, Jie; Szlávik, Zoltán; Bozzon, Alessandro

DOI
10.1145/3479158

Publication date
2021

Document Version
Final published version

Published in
ACM Transactions on Social Computing

Citation (APA)
Balayn, A., Yang, J., Szlávik, Z., & Bozzon, A. (2021). Automatic Identification of Harmful, Aggressive, Abusive, and Offensive Language on the Web: A Survey of Technical Biases Informed by Psychology Literature. *ACM Transactions on Social Computing*, *4*(3), 11:1 - 11:56. Article 11. https://doi.org/10.1145/3479158

Automatic Identification of Harmful, Aggressive, Abusive, and Offensive Language on the Web: A Survey of Technical Biases Informed by Psychology Literature

AGATHE BALAYN and JIE YANG, Delft University of Technology, Netherlands
ZOLTAN SZLAVIK, myTomorrows, Netherlands
ALESSANDRO BOZZON, Delft University of Technology, Netherlands

11

The automatic detection of conflictual languages (harmful, aggressive, abusive, and offensive languages) is essential to provide a healthy conversation environment on the Web. To design and develop detection systems that are capable of achieving satisfactory performance, a thorough understanding of the nature and properties of the targeted type of conflictual language is of great importance. The scientific communities investigating human psychology and social behavior have studied these languages in details, but their insights have only partially reached the computer science community.

In this survey, we aim both at systematically characterizing the conceptual properties of online conflictual languages, and at investigating the extent to which they are reflected in state-of-the-art automatic detection systems. Through an analysis of psychology literature, we provide a reconciled taxonomy that denotes the ensemble of conflictual languages typically studied in computer science. We then characterize the conceptual mismatches that can be observed in the main semantic and contextual properties of these languages and their treatment in computer science works; and systematically uncover resulting technical biases in the design of machine learning classification models and the dataset created for their training. Finally, we discuss diverse research opportunities for the computer science community and reflect on broader technical and structural issues.

CCS Concepts: • **Computing methodologies → Natural language processing**; **Information extraction**; **Machine learning algorithms**; • **Information systems** → *Web mining*; *Crowdsourcing*; Social networks; • **Social and professional topics → Hate speech**; *User characteristics;*

Additional Key Words and Phrases: Bias, discrimination, cyberbullying, offensive language, abusive language, harassment, toxic language, harmful language

ACM Reference format:
Agathe Balayn, Jie Yang, Zoltan Szlavik, and Alessandro Bozzon. 2021. Automatic Identification of Harmful, Aggressive, Abusive, and Offensive Language on the Web: A Survey of Technical Biases Informed by Psychology Literature. *ACM Trans. Soc. Comput.* 4, 3, Article 11 (October 2021), 56 pages.
https://doi.org/10.1145/3479158

Authors' addresses: A. Balayn, J. Yang, and A. Bozzon, Delft University of Technology, Netherlands; emails: {a.m.a.balayn, J.Yang-3, A.Bozzon}@tudelft.nl; Z. Szlavik, myTomorrows, Netherlands; email: zoltan.szlavik@mytomorrows.com.

2469-7818/2021/10-ART11 $15.00
https://doi.org/10.1145/3479158

1 INTRODUCTION

Harmful, aggressive, abusive, and offensive languages in online communications are a growing concern [115, 187, 287]. They constitute a threat to Freedom of Speech [268], damage the dignity of the targeted individuals [280], and prevent healthy and fruitful conversations [169]. The recent hearings [154] of the biggest social network's platform (Facebook) CEO also testify of the growing public attention on the issue.

Manual moderation is still the most reliable method for content filtering [142, 153, 158, 201], but it suffers from several issues. Content moderators cannot handle the deluge of user-generated content fast enough not to endanger anyone. Moreover, they are continuously exposed to hurtful content, which induces mental issues and can lead to self-harm acts [252].

Under the societal and political pressure [101, 134], online platforms are urged to find computational solutions to detect conflictual languages [105]. Machine learning approaches are considered the best solutions [101], due to their promise to achieve reasonable detection performance at scale. In practice, error rates still demand for extensive manual moderation. For instance, Arango et al. [14] show the frequent drop of performance for machine learning models evaluated on deployment data (e.g., a model that achieves 70 F1-score on its test dataset can only achieve 21.1 F1-score on another dataset).

Classification errors also raise concerns of discrimination [260]. For example, models might systematically misclassify certain populations more often than others, for instance more often associating tweets written in African-American English to negative classes than tweets written in Standard American English [175], or misrepresent their identities due to stereotypical associations between certain concepts and sensitive attributes [40]. The causes of these errors can be summarized under the broad term of *bias*. When the training dataset is biased towards certain (latent) characteristics, the model is implicitly taught a biased representation of the conflictual languages. While these biases are *technical* artifacts, we argue that their root causes and solutions cannot only be found in the technical realm. Issues at the *conceptual* level induce these biases and the challenges in tackling them. Through this survey, we show the existence of several *mismatches* between the typical formalization of conflictual languages in the computer science literature and how people perceive and experience such languages in reality. Mismatches first manifest at a *terminological* level, as publications often use an incorrect term to refer to the conflictual language they study; but they further deepen into *semantic and contextual* levels. For instance, psychology literature highlights that the perception of conflictual languages depends on various contextual factors [60], such as one's prior experiences (e.g., someone who is frequently subject to racial prejudice might perceive sentences as hate speech more strongly), or the direct context of a sentence, e.g., its author and target. Failing to acknowledge such rich characterization has obvious implications for the correctness and effectiveness of the deployed system. Consider, for instance, the widely used practice of keyword-based sampling in training data construction, i.e., collecting conflictual text based on certain keywords. This method implicitly teaches a model that conflictual languages contain specific words, and leaves out offensive texts with more subtle—or "coded" language that, in practice, makes the resulting system ineffective.

In this survey, we aim at surfacing and systematically characterizing these mismatches and the technical biases that reinforce them to highlight relevant research challenges. Figure 1 summarizes the research fields and technical aspects addressed in our survey. By interrogating psychology literature, we drive an informed analysis of trends in computer science papers and propose a consolidate taxonomy for conflictual languages. Then, we identify the biases that arise from prior conceptual mismatches. By adopting a data-centered view, we show that many issues in the outputs of the systems originate from problematic choices in the design of data engineering pipelines.

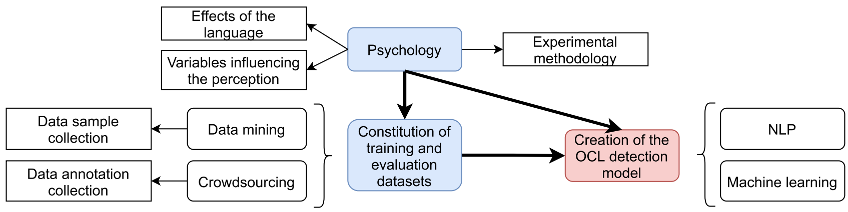


Fig. 1. Dependencies that influence the design of online conflictual language detection systems. Technical works in NLP and machine learning, and possibly works from psychology and politics, determine the inference task. Datasets are then developed (or selected) according to the task, in a way that is also informed by data mining and crowdsourcing literature.

1.1 Scope and Terminology

The computer science literature on the automatic detection of online conflictual languages focuses on a few languages: hate speech [101, 240], cyberbullying [6, 8, 135, 232], flaming [163], offensive [234], and aggressive language [152]. We compare all these languages and focus on their most common manifestation, text. We believe that research on one language might benefit research for another language, and that a precise and organized terminology is needed to improve the quality and applicability of automatic solutions [278].

We employ the term **Online Conflictual Language (*OCL*)** to refer to the overarching category of online language that subsumes all these types. We use the term "language" instead of "speech" (used in computer science to refer to hate speech [83, 101]), because the latter implies the spoken nature of the sentence [237]. In contrast to terms with specific meanings (e.g., "aggression" implies the intention to harm), we use the term "conflict," defined as *"the occurrence of mutually antagonistic or opposing forces, including events, behaviors, desires, attitudes, and emotions."* We use "Online Conflictual Language" also to avoid ambiguity and confusion, as the term has not been previously used in psychology, linguistics, or computer science.

Multiple social sciences such as psychology, sociology, media studies, political science, law and history, discuss online conflictual languages from perspectives such as manifestation, dynamics, and impact. This article primarily focuses on psychology, as it provides clear definitions and a diverse set of actionable information. When relevant, discussions from those other social sciences are also included.

In this survey, we discuss the creation of datasets for online conflictual language detection. We do not provide a list of open source projects or a list of common datasets, as previous works (Schmidt et al. [240], Vidgen and Derczynski [277], and Fortuna et al. [101]) provide an adequate overview. We also refrain from focusing on the political aspects of online conflictual languages, their definitions in laws and regulations, or the ethical concerns raised by their study (e.g., impact on researchers involved in the topics). These topics are, respectively, addressed in Fortuna et al. [101], and in Vidgen et al. [278]. While several challenges identified in these papers are also addressed in our work, our analysis based on social science literature enables us to provide complementary recommendations and directions for future work.

1.2 Comparison to Previous Works

Several surveys analyze literature on *OCL* ([101, 232, 240, 258, 278]) and the technical challenges for the development of accurate detection systems. Recent surveys [101, 278] also recognize and address the difficulties in understanding the object of study. For instance, Fortuna et al. [101] show terminological confusions in the definitions of hate speech by various social media platforms.

However, their analysis often refers to only few types of languages and only partially explores related disciplines.

Our survey substantially departs from previous works precisely by engaging in an analysis of the problem of *OCL* informed by psychology research. It surfaces new conceptual aspects important for automatic detection tasks, including consolidated definitions of *OCL*, and factors that influence their perceptions. It also highlights limitations in the current setup of detection tasks, especially to account for context and subjectivity of *OCL*. While these constitute technical challenges, their presence also hints at structural challenges in the organization of the research field, such as the development of collaborations with other domains and the acknowledgment of well-constructed datasets as valuable scientific contributions.

1.3 Original Contributions

In details, this manuscript provides the following five contributions:

(1) A set of definitions and properties, and a taxonomy to reconcile the *OCL* terminology (Section 3). This reconciliation speaks to an increasingly advocated need for conceptual clarity [278].
(2) A discussion of the psychological aspects related to *OCL*s (Section 4) that uncovers conceptual mismatches with automatic detection works and a reflection on the experimental practices that could contribute to computer science research.
(3) A comprehensive review of the typical data engineering pipelines used for building datasets (Section 5) and of their technical biases (e.g., usage of disagreement metrics for evaluating the annotation quality of subjective *OCL*) that can be harmful and participate to the low generalization abilities of the systems.
(4) A quantitative review of conflictual language detection models (Section 6) and an analysis of their limitations in terms of performance, leading to the identification of additional biases. Guided by our *OCL* taxonomy, our work offers a principled characterization of differences, similarities, limitations, and opportunities in computer science approaches. The lack of features relevant to individual *OCL* and the integration of social biases are pressing issues, for which future research could draw inspiration from psychology literature and machine learning fairness and explainability literature.
(5) An extensive discussion of open, technical and structural, research challenges, with clear and actionable suggestions for future work inspired by various psychology and computer science domains and informed by our systematic literature analysis (Section 7).

2 METHODOLOGY AND PAPER COLLECTION

In this section, we introduce the methodology employed to achieve the aforementioned contributions, and we explain the procedures followed to collect the computer science and social science papers that we analyze.

2.1 Methodology

We take a multi-step approach including, (1) retrieving relevant terms about *OCL*, (2) literature search and analysis, (3) taxonomy creation, and (4) analysis of the research challenges. Details of these steps and their connections are summarized in Figure 2.

2.2 Paper Collection

2.2.1 Retrieval of the List of Terms. Starting from the most comprehensive (to date) *hate speech* survey [101] and other related surveys, we iteratively gathered relevant terms by searching

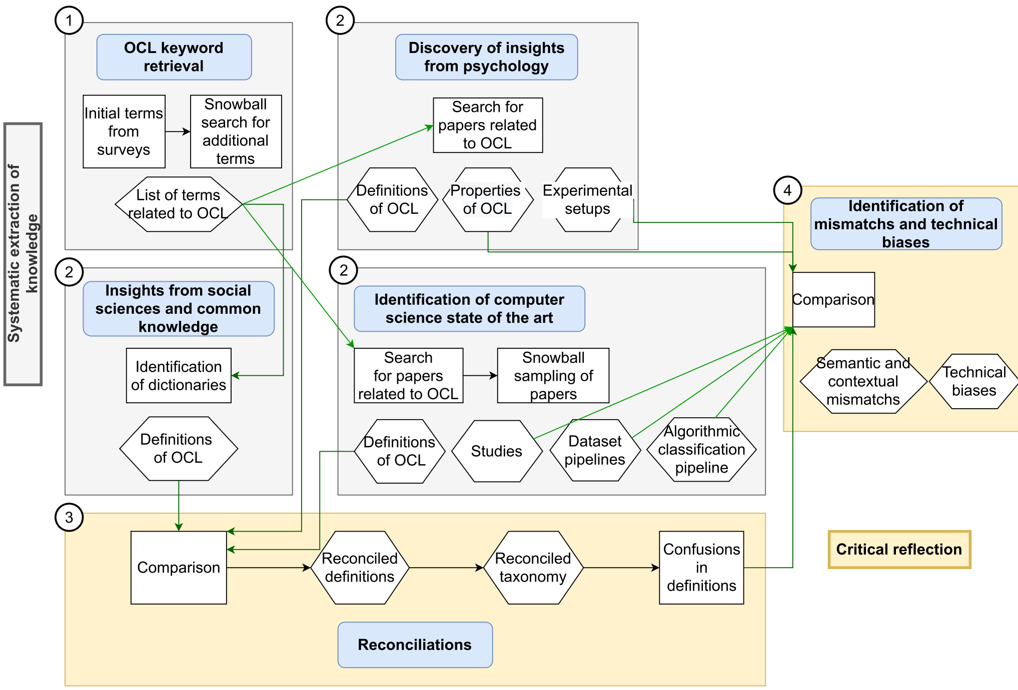


Fig. 2. Methodology employed to develop the surveyActions taken are represented in rectangles and the resulting artifacts in diamond shapes. (1) After retrieving the *OCL* terms, (2) we identify the main knowledge developed by—and experimental processes employed in—our different fields of interest, i.e., psychology and social science, computer science, and common knowledge. (3) From such information, we reconcile the terminological and conceptual mismatches in-between the fields. (4) Finally, we reflect on the mismatches to identify the technical biases they create or reinforce.

referenced literature and identified the following: *hate, hateful, toxic, aggressive, abusive, offensive and harmful speeches, profanity, cyberbullying, cyberaggression, flaming, harassment, denigration, impersonation, outing, trickery, exclusion, cyberstalking, flooding, trolling, discrimination.* In the rest of the survey, we refer to the sum of all these concepts with our proposed expression *online conflictual language* (abbreviated to *OCL*).

2.2.2 Retrieval of Psychology Papers. We searched for papers published in psychology venues, that explain the different languages (definitions) or study the variables that influence their perceptions. We did not include papers for which the major focus is to understand the impact of the language or speech; how the feeling (e.g., hate) develops in an individual; or the extent of the spread of the speech or language. We focused on the field of *social psychology* to avoid non-relevant literature (e.g., consumption habits of people when related to toxicity). After retrieving the initial documents, we used a snowball approach to identify additional papers. Without loss of generality, we cite only a subset of the considered papers, striving for complete topical coverage and not for completeness of literature works.

We inputted the following query in Google Scholar: (OCL keyword)`AND` (((variable)`OR` (perception)`OR`(definition)`OR`(judgement)) `AND`((web)`OR`(online)`OR`(internet)) `AND` (source:“social Psychology”) from which we later removed (web)`OR`(online)`OR`(internet), because there are very few papers about these languages online. Retrieved papers are from the *Journal of Applied Social Psychology*, the *Journal of Experimental Social Psychology,* and the *Personality, Psychology, Public Policy, and Law and Social Psychology Review*, but also from other venues such as *Psychology of Women Quarterly, Journal of Social Issues, Language Sciences, Computers in Human Behavior, Group Dynamics: Theory, Research, and Practice.*

2.2.3 Retrieval of Computer Science Papers. We conducted a systematic literature review following the steps listed below:

(1) *Query formulation*: We are interested in all papers about *OCL detection tasks*, *creation of dataset*, or *collection of data annotations*. Hence, we chose the keywords: "filtering," "crowd," "crowdsourcing," "annotation," "dataset," "detection," "prediction," "classification." We formulated the query by combining these keywords and the ones listed in Section 2.2.1 with `OR` query clauses; `AND` clauses are used to create the final queries, e.g., "((cyberbullying)`OR`(hate speech))`AND`((detection)`OR`(annotation))."
(2) *Document search*: We retrieved the documents from several libraries (Scopus, ACM, IEEE, DBLP, Google Scholar) by matching the title, abstract, and keywords of the documents with our query. As Scopus covers diverse research fields (we retrieved plenty of papers from Chemical Engineering due to the "toxic" keyword), we limited the search to computer science papers. The papers were collected at the end of 2018, complemented with works from 2019 and 2020 during the paper revision process.
(3) *Document filtering*: We removed the duplicates and limited the retrieved documents to computer science papers. We manually removed documents about toxic behaviors in online games, when the behaviors did not consist in the use of *OCL*s or the documents did not tackle *OCL*s effects or causes, e.g., Kwak et al. [149] about how people report toxic behaviors while gaming. We removed works related to *trolling* [235], because it is a very broad topic, where most of the papers study the phenomenon, but do not propose automatic methods for detection, and *spamming*, which is not characterized as conflictual language. We refer the interested reader to Berghel et al. [32] and Fornacciari et al. [100] for more information about trolling.
(4) *Search extension*: From the selected documents, we retrieved their list of references and performed the document filtering step again on these additional documents.

We retrieved $N = 219$ relevant and accessible computer science papers. The classification of all the retained papers in terms of meta information (authors, year, publishers) and technical artifacts is available on the companion page.[1]

3 TERMINOLOGICAL MISMATCH: ENTANGLED DEFINITIONS

In this section, we analyze how *OCL* languages are defined and studied in social sciences, and particularly in psychology. We reconcile the definitions of the *OCL* terms and create an informed taxonomy. Later, we will discuss how this taxonomy poses new challenges for the creation of automatic detection systems.

3.1 Definitions of *OCL*

3.1.1 Definitions from Psychology Literature. Table 6 in Appendix A.1 lists the definitions of the *OCL*s that we retrieved from a psychology dictionary and psychology literature. These definitions highlight properties of the concepts (e.g., intent, effect, target) that are necessary pillars to reconcile the terminologies in the next subsections. Note that we could not find a definition for all concepts

[1] https://sites.google.com/view/survey-on-ocl.

due to their recency in the online context, and that certain concepts do not yet have a single, commonly agreed upon definition.

Hate. Hate has a multitude of definitions that share many similarities [257]. For instance, the most comprehensive and broadly adopted definition of *hate crime* [202] is *"a hate crime can be defined as one in which the victim is selected because of his or her actual or perceived race, religion, disability, sexual orientation, or ethnicity/national origin (U.S. Department of Justice, 1999"* [206, 259], which is very similar to *"the violence of intolerance and bigotry, intended to hurt and intimidate someone because of their race, ethnicity, national origin, religion, sexual orientation, or disability. [...] Hate crimes differ from other crimes in that they typically involve use of explosives, arson, weapons, vandalism, physical violence, and verbal threats of violence to instill fear in their victims, and the community to which they belong"* [207]. The definitions of *hate online* also bear common properties between each other and with hate crime, e.g., *"cyberhate—namely, online messages demeaning people on the basis of their race/ethnicity, gender, national origin, or sexual preferenc"* [157]. These definitions clearly define the type of **targets** of the language.

Aggression. Similarly, the concept of aggression remains in discussion [144]. For instance, Burbank et al. [44] raise the following question: *"Assuming that we define 'aggression' as behavior that results in physical or psychological harm, we must question whether or not an act that results in the harm of another was indeed intended to do so."* Verbal social aggression has reached a consensus *"these forms of aggression are intended to cause harm by using others, spreading rumors, gossiping, and excluding others from the group or ignoring them,"* but its categorization into subconcepts is still discussed [16]. Here, both the **effect** (harm) and **intent** to cause the harm are highlighted.

Bullying. Bullying has an agreed upon definition: *"physical, verbal, or psychological intimidation that is intended to cause fear, distress, or harm to the victim" [219] with "the repetition of the behaviour over a period of time and the relational asymmetry between bully and victim"* [20]. Some works further categorize bullying behaviors into different groups.[2]

Discrimination. Definitions of discrimination also seem to converge: *"harmful actions toward members of historically subordinated groups because of their membership in a particular group. [...] Discriminatory behaviors are carried out based on personal prejudices or stereotypes about members of a specific group"* [177, 194].

Harassment. Harassment presents a precise definition, e.g., for verbal sexual harassment—*"judgments of appearance, obscene and euphemistic statements about sexual receptivity, and remarks belittling the competency of one's gender"* [114]—the definition points out to specific **types of natural language** (e.g., euphemism).

The above definitions present common properties across languages that could be identified and exploited for developing automatic detection methods. Interestingly, certain publications even distinguish explicitly different languages by pointing out different dimensions, e.g., *bullying* is

[2]E.g., "threat to professional status (e.g., belittling opinion, public professional humiliation, accusation regarding lack of effort); threat to personal standing (e.g., name-calling, insults, intimidation, devaluing with reference to age); isolation (e.g., preventing access to opportunities, physical or social isolation, withholding of information); overwork (e.g., undue pressure, impossible deadlines, unnecessary disruptions); and destabilization (e.g., failure to give credit when due, meaningless tasks, removal of responsibility, repeated reminders of blunders [...]" [219].

a repeated *aggression* over time [24, 251] (**time** dimension), *bullying* and *discrimination* differ by the type of entities they target [194][3] (**target dimension**).

3.1.2 Reconciled Definitions.

Motivation. In computer science publications, the terms related to **online conflictual languages (*OCLs*)** are not always defined, or with definitions that remain ambiguous. Besides, they are often used interchangeably, as we show next in Section 3.3.

A few noticeable exceptions exist. Certain works survey the definitions of hate speech from various companies, legal frameworks, and scientific publications [101]; study properties of abusive languages or harassment in details referring to psychology works [116, 284]; discuss in depth the differences between different languages (e.g., hate speech, offensive language, and other harassing languages [103, 262]). Yet, these works do not address all types of *OCL*, they sometimes do not align across publications, and some of the definitions are not directly actionable for distinguishing the various languages. For instance, the following definitions of *offensive* and *abusive* languages "*Profanity, strongly impolite, rude or vulgar language expressed with fighting or hurtful words in order to insult a targeted individual or group*," "*Any strongly impolite, rude or hurtful language using profanity, that can show a debasement of someone or something, or show intense emotion*" [102] hold many similarities, that prevent from clearly identifying their differences. Also, Golbeck et al. [116] consider "*jokes with poor taste*" offensive, while they do not necessarily imply a specific type of language, conflicting with the above definitions.

As *OCL* terms do not all have precise and consistent definitions, we attempted to find definitions (Table 6 Appendix A.1) in a general dictionary,[4] a specialized psychology dictionary,[5] and a dictionary from other social sciences.[6] Because these definitions were again neither clear or consistent, nor specific to the online context, we decided to define reconciled definitions and taxonomies as described below.

Methodology. In cases where both a social science and a computer science definition are found, we opt for the social science definition, as the field(s) has been studying such concepts more extensively until now. For terms where only computer science provides definitions, we select a single definition based on the frequency at which they all appear in papers covered by our survey. In case of ties, we choose the definition most similar to our intuition about the term.

Results. The reconciled definitions are summarized in Appendix A.2 Table 7. Most papers use the term *hate speech* with the meaning in Table 7 and define sub-categories based on the target of the language [12, 83, 136, 270, 283, 309]. *Hateful speech* was defined in computer science to solve ambiguities in the definitions of *hate speech* and focuses on the expression of hate without specifying any intent from the author of the speech [233]. *Hate* does not specify the group "affiliation" of the target. Conversely, *offensive* language does not imply a specific intention (only two computer science papers mentioned it [200, 285]), but a notion of perception from the target of the language.

Cyberaggression and *cyberbullying (traces)* are often confused: *Cyberbullying* is a specific case of repeated *cyberaggression*, and *cyberbullying traces* correspond to *cyberbullying* and its

[3]"The major difference between bullying and discrimination lies in the characteristics of victimized targets; that is, in target specificity. Discriminatory acts are directed narrowly toward members of specific, socially subordinated groups (e.g., gays, the obese); whereas bullying acts are directed toward broader, more heterogeneous targets that may include socially subordinated groups, but also people who wear strange clothes or are socially withdrawn."
[4]https://dictionary.cambridge.org/.
[5]https://dictionary.apa.org/.
[6]http://bitbucket.icaap.org/.

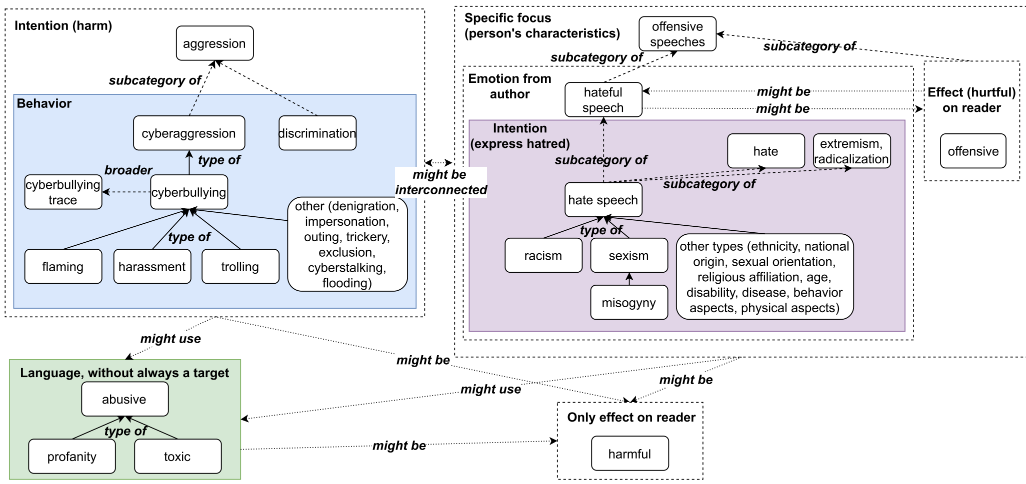


Fig. 3. Taxonomy of the Online Conflictual Languages (*OCLs*). The boxes correspond to one or more properties—specified in bold—which are common to the set of concepts contained in them. The arrows specify the relationships between languages (in italic).

responses [295, 306], requiring different methods for detection. Although some definitions of *harassment* [221, 298] do not mention repetition, we consider it is a *cyberbullying* category because it appears mostly in *cyberbullying* publications [8, 29, 48, 54, 65, 87, 171, 178, 185, 238, 247, 256]. We do not distinguish between *aggression* and *cyberaggression* in the survey, as their only difference lies in that *aggression* is not specific to the Web.

Harmful [91, 156, 245] and *toxic languages* [113, 166, 224] are not defined precisely in literature, except for toxicity in video games. However, *toxic speech* is described in a crowdsourcing task to collect a dataset of toxic comments [292][7] and insists on the type of language used, which motivated our characterization choice.

3.2 Reconciled Taxonomy

Motivation. The reconciled definitions highlight the differences between *OCL* concepts, but do not make their relations explicit. For instance, the definitions of *abusive* and *offensive* languages of Founta et al. [102] are precise, but the language properties remain implicit (*abuse*'s main property is the type of language used, and *offensive* language is characterized by the focus on someone's characteristics, without necessarily employing rude language—as shown by our selected definitions). Thus, we propose common properties to categorize the concepts and their sub-categories (Appendix Table 8) and derive a taxonomy in Figure 3.

Methodology. We define seven binary properties based on computer and social science works in an effort to build independent categories of concepts. We map the concepts to the categories based on the descriptions and examples of *OCLs* in computer science. We resolve disagreements between papers using the frequency to which the properties are mentioned. A positive attribution of one concept to one dimension (i.e., a "yes" in Table 8) means that the instances of the concept necessarily contain this element, while a "no" means that it is not necessary.

[7] https://github.com/ewulczyn/wiki-detox/blob/master/src/modeling/toxicity_question.png.

Because no concept was fully independent or entirely derived from other concepts, we refined the classification by dividing certain properties into non-binary sub-properties (second line of headers in Table 8). For example, *hate* is more general than *hate speech* (*hate speech* always focuses on stereotypes), so the definitions share a common set of properties (i.e., same "yes" in the table), but *hate speech* also has more constraints.

This categorization highlights clearly interpretable clusters of concepts with different relationships: sub-categories (additional mandatory properties), sub-types (more precise elements specifying the properties), broader relationships (concepts that might use several sub-concepts). This is the base for the final taxonomy.

Results. The final seven properties are the following:

- ***Intention***: The author of the language has a negative intention (hatred[8] or harm[9]).
- ***Behavior***: The language is defined by a specific type of behavior of the author.
- ***Specific focus***: The language deals with a particular topic of interest (a characteristic of a person or another interest such as a rumor).
- ***Emotion of the author***: Its author feels a specific emotion when writing.
- ***Language***: The language contains a specific type of natural language (e.g., euphemism).
- ***Target***: The author of the language is targeting a defined entity.
- ***Effect***: The language has a specific effect on the reader.

The mapping of the different *OCL* concepts into these properties can be found in Table 8 in Appendix A.3. The clusters of *OCL* are indicated with the cell colors. The final taxonomy is represented in Figure 3. Four main groups of Online Conflictual Languages appear:

- ***Aggression***: characterized primarily by the intention of the speaker to harm.
- ***Offensive languages***: characterized by the focus on a person's characteristics.
- ***Abusive language***: characterized jointly by the use of a specific language style and the non-specification of a target.
- ***Harmful languages***: characterized solely by their effect on the reader while none of the other properties has to be specified.

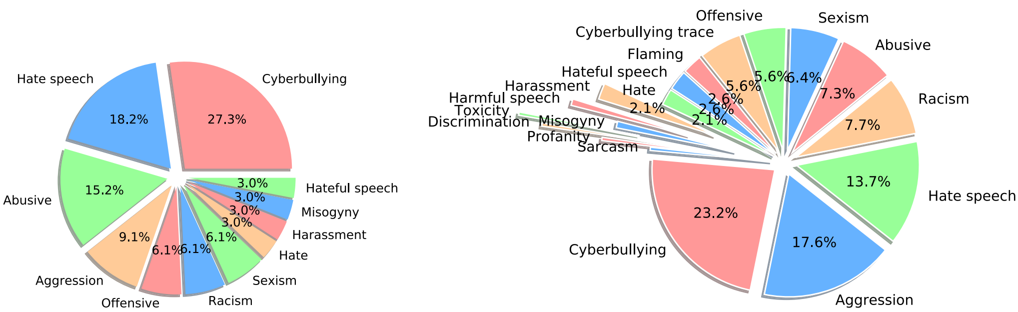
These groups are not entirely independent, as *OCL* languages inherently span overlapping properties. This intersectionality can be ultimately exploited by re-purposing research focused on one language to other languages sharing a common property. We recommend to investigate how to automatically understand each property separately and combine the findings of this research when addressing a language characterized by multiple properties.

3.3 Mapping of the Computer Science Literature into the Revised Taxonomy

In this section, we analyze the mapping (see Appendix A.4 Figure 12) between the way terms were originally used in computer science papers and their definitions in the new taxonomy for the 184 papers from which a description of the concepts could be traced back —*16%* of publications do not refer to a definition. We find ambiguity in *58%* of the papers.

Hate, *hate speech,* and *hateful speech* are used interchangeably. *Profanity, aggression, and cyberbullying* are not defined, probably because they have simple dictionary definitions. *Abusive*, *offensive,* and *neural* [234] languages are often associated with incorrect terms. In 10 papers, the word *abusive* is used to refer to *aggression*, while *offensive* is confused, respectively, with sexism, racism, cyberbullying, aggression (1 time), hate, hateful speech (2 times), hate speech (5 times), abusive

[8]"An extremely strong feeling of dislike." - definition from https://dictionary.cambridge.org/dictionary/english/hatred.
[9]"Physical or other injury or damage." - definition from https://dictionary.cambridge.org/dictionary/english/harm.



(a) *OCL* concepts in studies.

(b) *OCL* concepts in publications that develop solutions for classification tasks.

Fig. 4. Distributions of *OCL* concepts in computer science works. We differentiate between works that study the use of *OCL* and works that develop technical solutions to automatically classify *OCL*.

language (4 times). The confusions between *aggression*, *harassment*, *cyberbullying*, *cyberbullying traces,* and *cyberaggression* are also highlighted. Thirteen papers mentioning cyberbullying investigate cyberbullying traces, and 19 other papers actually study aggression. Finally, two sub-types of *hate speech*, *racism,* and *sexism* are confused with more general categories such as offensive, hateful, and hate speech languages. This is possibly due to the lack of datasets that would support studying broader concepts.

The terminological mismatch shows that there is no consensus on the definitions of *OCL*s in computer science. Authors often mention a certain *OCL*, but might actually only identify a subcategory of it (e.g., hate speech and racism) or identify a broader set of languages (e.g., cyberbullying traces and cyberbullying) or a totally different type of *OCL* (e.g., hate speech and abusive language). Depending on the application at hand and its specifications, such mismatch could easily lead to systems that do not fit their requirements.

3.4 Distribution of Works on *OCL* Concepts

This terminological reconciliation shows that certain types of languages have attracted less attention from computer scientists, as shown in Figure 4(a). Aggregated into the coarse-grain categories, 45.5% of papers tackle offensive languages, 39.4% of papers investigate aggression, and 15.2% work on the use of abusive language. No paper was found specifically on harmful language. Similarly, the distribution of concepts investigated in classification tasks (Figure 4(b)) presents an imbalance between coarse-grained concepts such as *cyberbullying, aggression, abusive, and hate speeches* and certain finer-grained concepts (*toxic speech, misogyny, ...*) are seldom studied. This might be explained by the "popularity" of cyberbullying and hate speech in the media, but also by the lack of understanding of finer-grain concepts. Yet, detecting each finer-grain concept could enable a more precise and modular filtering of concepts.

4 CONCEPTUAL MISMATCHES TOWARDS TECHNICAL BIASES

In this section, we argue for the existence of profound *conceptual mismatches* pertaining to the focus of computer science literature on the development of algorithmic pipelines, mostly due to lack of consideration for the application context —*contextual mismatch*—or for the specific properties of the targeted *OCL* —*semantic mismatch*. We provide an overview of the insights about *OCL* that can be found outside computer science research and compare them to high-level findings from our systematic survey of computer science literature.

Table 1. The Factors Identified in Psychology Literature that Influence *OCL* Perception, Organised in 3 Categories (*Internal* Characteristics of the Observer, Characteristics of the *Sentence Content* and of the *Sentence Context*), and the Approach taken to Measure these Variables

Category	Variable	Measure	Paper
Observer	Gender	Question	[66, 67, 93, 120]
Observer	Ethnicity	Question	[66, 67, 289]
Observer	Education	Question	[66, 67]
Observer	Age	Question	[66, 67]
Observer	Liberalism inclination	Question (scale)	[93]
Observer	"Individuals' attributions of intent", angry and anxious dispositions	Not investigated	[120]
Observer	Sense of mastery, self-esteem	Question	[204]
Observer	Frequency to which people are subject to racial prejudice, "beliefs about the appropriateness of expressing racial prejudice"	Question (scale)	[186, 289]
Observer	Membership esteem to the offended group	Question (scales)	[37]
Context/Content	Targeted group or person	Scenario	[37, 66, 67, 126]
Content	Category of hate speech	Info in dataset	[126]
Content	Prejudice, sentence properties	In the dataset	[66, 86]
Context	Public or private sentence	Scenario	[66, 67]
Context	Received response to the language	Scenario	[66–68]
Context	Author, its characteristics, race, gender	Scenario	[70, 207]
Context	Hierarchical level of perpetrator and victim	Question	[265]
Context	Internet community	Info in dataset	[253]
Context	Social status of a group	Question	[126]

4.1 External Insights on Online Conflictual Languages

4.1.1 Semantic Knowledge from Psychology. Researchers in psychology have extensively studied conflictual languages, beyond the context of Web communication platforms. We summarize here the major insights relevant for the prospect of detecting these *OCL*.

Three main types of variables influence how *OCL* is perceived by external *observers* (see Table 1): the *language content*, including the properties of a *person or group targeted* by *OCL*; the *language context*; and characteristics of the *observer*.

Internal characteristics of the observer. The perception of certain *OCL* depends on the internal characteristics of someone who observes the language. This hints at the subjective nature of many online conflictual languages. For instance, Guberman et al. [120] observe a difference in *aggressiveness* ratings of tweets depending on *gender* (women rate tweets more often as aggressive than men) and mention the tendency that some people have "to interpret ambiguous stimuli as being intentionally aggressive" and the dispositions of people to become angry and anxious. Downs et al. [93] identify that *gender and liberalism inclination* influence how harmful a hate speech is perceived. Similarly, Cowan et al. [66, 67] point out that the *ethnicity, gender, education, and age* of the observer influence the perceived offensiveness of hate speech. Besides, attention is called on the distinction between the perceived *offensiveness* and *harmfulness* [68], with for example ethnicity being a main factor in the perceived harmfulness. This highlights the importance of clearly and precisely defining the *OCL* to detect, in order to account for the correct variables of importance.

Works focused on racial hate speech also pinpoint *the frequency to which people are subject to racial prejudice* and *people's "beliefs about the appropriateness of expressing racial prejudice"* [186], and *ethnicity* [289] (e.g. people of color who are more often subject of racial aggression perceive Web memes as more offensive, unlike White people). This speech triggers various emotional responses (fear, anger, sadness, outrage), and people with high membership esteem react more strongly to threats to their group than low identifiers [37].

Sentence content and context. The syntactic and semantic *properties of the sentence*, e.g. length, usage of profanity, and its *context* –author [70] and how its direct target behaved and felt [68], targeted group, whether it is public or private, and whether it received a response [66, 67]– influence how offensive it is perceived [66, 70, 126]. For instance, the perception of profanity depends on the *community* [253] as different communities use profanity with different frequencies and contexts and judge the words differently. Besides, a speech toward a single individual is seen as more offensive than a speech toward a group of people [37]. Also, a speech is offensive when it presents a property of an individual ("personal characteristic, belief", etc.) in a certain way which does not need to be hateful [15], as the wrongfulness comes solely from the aim of its author: "attempt to denigrate, humiliate, diminish, dishonour, or disrespect the other". The context is particularly relevant when distinguishing between languages that are *harmful* – which damages someone's interests – from languages that are *hurtful (offensive)* – which causes mental distress.

These three types of variables implicitly include finer-grained characteristics of the language: *the focus towards certain types of population and specific targets, the type of language used, the author, its intent and the effect on the targets.*

4.1.2 Contextual Information around OCL Detection Systems.

Context of application of the systems. The application domain of an *OCL* detection system determines its *context* of operation (e.g., a social media primarily used by children within a single country using a single language or used by a specific political community to discuss political opinions on specific subjects). Context consists in the type of platform (e.g., social media, conversational agent) on which *OCL* should be detected, the type of end-users and their backgrounds, the type of communities and populations that are present on the platform or interact with this agent, the topics that are frequently tackled, and the natural language typically employed (which can be different from offline language). These characteristics might impact how someone perceives *OCL* [278]. Understanding this impact would allow to scope the context in which systems can be used and would determine how to collect datasets for training and how to develop and test algorithms.

Laws and regulations, either governmental or from social media platforms, further constrain the type of online conflictual languages to be detected. They focus on certain properties of language, such as intent or targets (identified in the previous section) that are often more specific or nuanced [35]. For instance, the British government decided after many debates on "protections only against intentionally threatening expressions of religious hatred, not against those that were merely abusive or insulting, nor those that are reckless and likely to stir up hatred." Philosophy also studies when *OCL* should be limited and similarly defines criteria to make a decision, by analyzing case-by-case past events of *OCL* on social media [121]. Especially, it should be limited when "it is reasonable and feasible to assume that an act of Internet speech will cause harm to others," and more specifically when "targeted hate speech that carries with it immediate harm (capability to carry out the violence), individualized harm (capability to assault the target), and capability to carry out the threat (actualized means of committing the violence)." As our investigation in the remaining of the article shows, such nuances are not necessarily reflected in the ways datasets

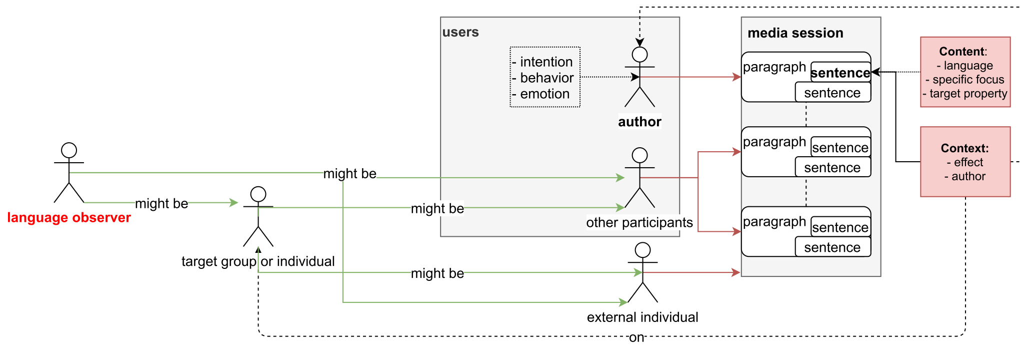


Fig. 5. Summary of the entities of importance in the understanding of *OCL*, as identified by computer science studies.

and models are developed, yet would be of importance, for instance, not to unintentionally restrict freedom of expression.

Hard technical requirements for the applications. The applications in which *OCL* detection systems are implemented also impose hard technical requirements (e.g., *OCL* posts should be removed from a platform within a certain amount of time). While these requirements do not necessarily impact the nature of the *OCL* to detect, they might impose constraints on the detection pipelines (e.g., cost of data collection, speed of machine learning inferences with or without the possibility to involve humans-in-the-loop), and tradeoffs with the system accuracy (e.g., scalability vs. accuracy). However, these requirements are not often accounted for in the literature, which instead focuses on accuracy. Only 4% of surveyed publications mention the *scalability* of their system, mainly the time efficiency to detect *OCL*, and only 6% tackle the creation of a full system in opposition to a detection method. These numbers are small, considering the need for efficient solutions, since leaving *OCL* public for too long might have psychological consequences for the readers.

The systems are ought to perform well *continuously over time*. Yet, only few systems continuously collect datasets, whereas this would shed light on the evolution of *OCL* along time, the changes in the users of platforms, how they impact a model's performance, and so on. Efforts to develop systems such as MANDOLA [195] or the Online Hate Index[10] would greatly contribute to progress in the field.

4.2 Computer Science Studies on *OCL*

Researchers in computer science have conducted studies on the use and spread of *OCL*s on the Web. They perform both manual analysis and statistical observations on datasets collected for the studies and discover properties of the languages that could be used to tune the features employed by automatic detection methods. These studies serve as a source to identify the entities studied in computer science literature, their relations, and their properties (summarized in Figure 5). These entities are primarily the *author* of a language—its behavior, intentions and emotions—the *language content* itself, be it a single sentence or an entire paragraph—the language used, the targeted property of a person or group, and implicitly the focus of the language since only sentences containing expressions of hate are studied—and its *context*—how it affects the *target person or group.*

Hateful behaviors are characterized with perpetrators' *internal characteristics*—their account creation dates, e.g., *hateful* users might be often banned; the amount of the users' activity on

[10]https://www.adl.org/resources/reports/the-online-hate-index.

Table 2. Type of Entity per *OCL*, Accounted for in Computer Science Classification Tasks

	Aggression	Offensive	Abusive	Harmful language
Media sessions	6	0	0	0
Sentence	83	75	12	1
User	13	1	1	0
Words	3	0	1	0

the media; the position of the users in the network graph; whether the users are identified as spammers—and the *characteristics of the sentences they write*—the lexical content and sentiment of their posts and hashtags [48, 52, 222]. ElSherief et al. [96] also identify various personality traits of both authors and targets of hate speech.

Other studies target *media sessions*, i.e., a conversation between several individuals. This is the case for cyberaggression, where both text, images, and possibly users are studied—e.g., the role of the author in the cyberbullying—sometimes with a temporal dimension [31, 129].

Certain studies [65, 80, 122, 127, 143, 173, 248, 253, 266, 281, 303] characterize the *language* itself, through the *sentence content* (i.e., the used vocabulary); the *targets*; the *context* (how the language is perceived); the relation between the type of target and the type of content employed [95]; and the effect of users' anonymity and users' geography. These properties are compared across platforms [148]. One study focuses on why and with which intensity a language is perceived as conflictual by an observer, using questionnaires: a sentence is seen as cyberbullying when it contains threats of physical violence, harassment, and profanity terms [87].

4.3 Computer Science Framing of *OCL*

In the remaining of this section, we identify conceptual mismatches that translate into technical biases in the design of automatic *OCL* detection systems. To do so, we compare the formulation of detection tasks in computer science publications to the above insights. We also provide an outline of the works on biases and contrast them with our previous insights.

4.3.1 Framing of Automatic Detection Tasks. Here, we present how classification tasks are generally framed and show the diversity of the classes and entities used across tasks.

Entities. We find a strong imbalance across entities targeted by classification tasks (Table 2). Sentences are the most studied. A few works also detect single words corresponding to a specific *OCL*, or identify users, public accounts, and media sessions that comport *OCL*, based on the detection of sentences and words. Retrieving data for media sessions or users is technically more challenging than for words or sentences. Media sessions are only studied for *aggression,* because they allow to analyze the users' behaviors that emphasize user intention, a characteristic specific to aggression. Studying sentences allows to access certain properties of *OCL* (e.g., language type, focus, and possibly intention), but leaves out information relevant for certain types of languages, such as the effect on the reader for offensive languages or possibly the intention of the author.

Classes. The number of classes targeted in the classification tasks is also imbalanced. Most tasks use 2 classes (77.7%) (e.g., is hate, is not hate language) or 3 classes (15.6%) (e.g., is positive, is neutral, is hate language), which corresponds to the basic requirement of the systems. The tasks with more classes (4 to 13) reflect the intensity of an *OCL* language, which is more challenging to detect. As we discuss in the next subsection, binary classes do not necessarily reflect the understanding of *OCL* obtained from our previous analysis. For instance, psychology pointed out to the dependency of certain *OCL* perception on various contextual factors, left out when binary classes are predicted for bare sentences.

4.3.2 Main Bias Concerns. We report here the types of biases studied explicitly in relation to automatic *OCL* detection. These mainly relate to certain inherent contextual properties of *OCL* identified by psychology literature, and to a few properties specific to the online context—in certain cases using the term "bias" directly—but also to the potential discriminatory impact of *OCL* detection systems. We also investigate how these bias concerns compare to the semantic and contextual information identified in the previous subsection.

Inherent contextual biases. Works on cyberbullying detection have shown how different *authors* of *OCL* —difference based on gender [71], age, profanity history [74], or intent [3]—shape differently their sentences. A few properties of the target or *observer* of the language have also been indirectly studied, mostly through the properties (especially the gender) of the employed dataset annotators (e.g., workers from crowdsourcing platforms) [236]. Yet, the actual observers (e.g., social media users) do not necessarily resemble the annotators of a crowdsourcing platform, and hence studies might not fit the perceptions of actual users. The *conversation context*—specifically, *replies* to *OCL*—has also been investigated in a few works [159, 199].

Biases related to the online context of the systems. The contextual characteristics identified in the previous subsections are often not mentioned in papers developing detection methods, except for the *platforms* from which datasets are collected. The similarities and differences in the natural language written across platforms is sometimes investigated by measuring the generalizability performance of models trained on one platform and one dataset across platforms and across datasets [4, 119] as a proxy for the intensity of the differences. Besides, no work was found to study the diverse perceptions of *OCL* of users across platforms.

Similarly, only few works discuss the end-user related information that should drive the development of a system. Arango et al. [14] show that many datasets suffer from *user biases*. Few users constitute the authors of the majority of *OCL* in common datasets, thus identifying *OCL* could translate into identifying the author of a text sample, leading to overestimating models' performance. Besides, only the user social network [138] is investigated as user contextual cue, while it is shown to increase detection accuracy of models relying on it.

Discrimination-related biases. Recent papers often employ the term "bias" to study system artifacts that might create discriminatory harms. Such harms are identified by comparing the performance of a system for different subpopulations of users, e.g., based on gender [193] or other sensitive information [21], e.g., sexual orientation [61]; and possibly on intersectional attributes of the users, e.g., gender and political orientation [141]; or racial biases based on dialects [78, 236]. These biases all rely on properties of the end-users and their translation into natural language in the applications (e.g., the background of the end-users imply a dialect). These harms are often explained by imbalances of various nature in training datasets (e.g., more sentences written by male authors than by authors of other genders). Sun et al. [260] provide an extensive review of the formalization of these biases in natural language processing tasks, not specifically related to *OCL* detection.

Computer science works that account for biases do not yet encompass all kinds of relevant contextual and semantic information. We take a systematic approach in the remaining of this article to identify the technical biases that occur from the non-consideration of this information. That is what we discuss in greater extent in the next subsection.

4.4 Towards the Technical Mismatches

We identified the main properties of online conflictual languages as defined by social sciences and the applications' context and the ones integrated into computer science works. We now synthesize these properties to surface mismatches in computer science research. These mismatches relate to the inherent properties of *OCL* and to the subjectivity of certain *OCL*left out from both datasets and machine learning models.

4.4.1 Mismatches and Challenges in the Exploitation of the Characteristics of OCL.

Mismatches in the selection of variables. The three types of variables that influence the perceptions of *OCL*s identified from social science (Section 4.1), i.e., the internal characteristics of the observer, the sentence context, and its content, are similar to the ones found in computer science studies (Section 4.2). However, the exact characteristics investigated vary. Computer science studies focus on properties directly measurable or that can be inferred from information available on the online platforms, while psychology works rely on additional individual questionnaires.

Besides, only few detection methods use these specific characteristics of the languages. For instance, it is recommended to use a sentence context in a media session, and possibly the interactions of the sentence author with other users. It was also shown that the aggregation of hate messages from multiple sources creates stronger harms than a single message from one unique source [157]. However, only individual sentences are usually collected, without any metadata on context. Psychology also points out to specific language uses, such as euphemism in harassment [114] or humor for hate speech [291], e.g., humor affects the perception of offensiveness for certain types of hate speech (here, racism or sexism). However, these are often cited as future work in computer science, except for Magu et Luo [162] who study euphemisms within hate speech, or the recent works on sarcasm in ACL workshops.

Mismatch in the choice of target entity to detect. Psychology and computer science studies highlight the importance of looking beyond sentences, and at single user's behaviors or at entire scenarios, and of distinguishing between certain specific *OCL*. However, current setups do not focus on these factors (Section 4.3.1), which could lead computer science researchers to target research objects that are ill-defined. Hence, we recommend to refer to the social science literature around the targeted *OCL* to identify the important elements to include in datasets or algorithms for automatic classification of each *OCL*.

Challenges in data collection. The above gaps constitute socio-technical challenges: The social science insights need to be translated into accurate quantities measurable in practice in the technical systems. For instance, considering context in computer science is challenging due to the difficulty in scoping and collecting it, e.g., links in posts are often outdated, finding characteristics of the authors or receivers might be intractable and privacy infringing. This could—ideally—be solved when building training datasets by interrogating users on their perceptions and intentions, but it would be impossible in deployment where users could not be solicited for each post. This shows again the necessity to identify requirements of applications precisely, as they shape the constraints for training and deployment.

The relevant variables that impact the perceptions of *OCL* need to be identified more exhaustively as psychology studies do not necessarily tackle *OCL*s on the Web, but also in real-life scenarios. Also, certain *OCL*s are rarely addressed in psychology research, certainly due to their exclusive online nature (e.g., flaming).

The validity and importance of certain properties about the context of the language used only in computer science (e.g., user account creation date, amount of her activity on the media, her position in the network graph) could be further explored by adopting the methodology followed in psychology. Certain properties might be proxies for some of the psychology variables, e.g., they could help to identify the intent of the author of a post. This leaves the opportunity for computer

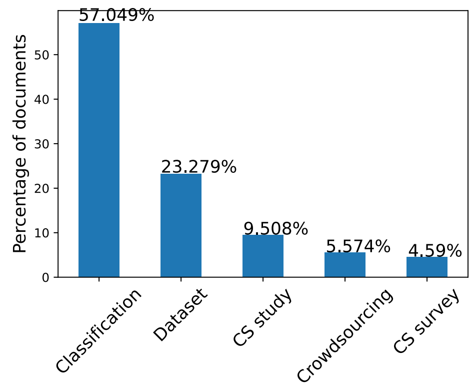


Fig. 6. Distribution of computer science literature focusing on *OCL*.

scientists to work with psychologists to bridge the gap between these domains and to more precisely define the concepts they study.

4.4.2 Spread of the Mismatches into the Classification Pipelines. The development of *OCL* detection systems follows the general development of machine learning applications [10, 261]. First, requirements are defined and specified into characteristics for the data, machine learning model, and its evaluation. Then, data are collected, cleaned, and labeled by annotators. Features are extracted, a machine learning algorithm is developed and trained. The resulting model is evaluated and later deployed and monitored. Certain steps might be iterated over to approach closer the initial requirements and possibly to revise these requirements.

Shortcomings in the systems arise from these steps. Under-defined requirements (mentioned in previous subsections) propagate into the next data-oriented and algorithm-oriented steps of the pipelines. Tuning pipeline components even for well-defined requirements is challenging. For instance, a system might be asked to perform equally well for children and adult users. However, with the subjectivity of certain *OCL*, building datasets with single, binary labels for each data record, and models that predict single labels, does not fit this requirement.

We identified five research directions in the computer science literature that integrate the different steps of the pipelines (literature surveys, statistical studies, classification methods, creation of datasets, and crowdsourcing tasks to collect labels), with a strong bias towards classification methods (Figure 6). There are especially few papers interested in crowdsourcing methods despite the challenge of obtaining high-quality *OCL* labels with such ambiguous and subjective *OCL* [120]. This hints at many research opportunities, especially around the biases contained in datasets, and studies to better understand *OCL*.

Next, we investigate the biases in detection pipelines. We pass current practices through the new requirements coming from the semantic and contextual mismatches to identify limitations, challenges, and potential solutions. To further substantiate our critical analysis, we situate literature on machine learning biases and unfairness [261] in the present pipelines.

5 DATASET CONSTRUCTION FOR THE DETECTION OF *OCL*

We now analyze the datasets and data engineering pipelines used in *OCL*detection systems. While the process of creating a dataset is long and costly, out of the 194 publications for which experiments have been conducted, only 33% of them use an already-existing dataset (5 do not specify the dataset used). Such numbers motivate the need to understand the specificities of data

Table 3. Dataset Sources Distribution

Data source	Count
Twitter	98
Formspring	18
News site	16
YouTube	14
MySpace	14
Forum	13
Wikipedia	12
Facebook, individual or group conversations	11
Instagram	9
Yahoo	8
Other content-sharing social media	7
AskFM	7
Website (non social media, e.g., Tumblr, Whisper)	6

Table 4. Datasets Language Distribution

Sample language	Count
English	157
Indonesian	6
Japanese	6
Dutch	5
Spanish	4
Portuguese	4
German	4
Arabic	3
Hindi	3
English-Hindi	3
French	2
Korean	2
Greek	2
Italian	2
Bengali	1
Russian	1
Turkish	1

pipelines, which do not seem standardized. We critically reflect on the pipelines and their biases. In light of the recent research on data excellence [69, 196, 286], this surfaces new challenges to adapt the pipelines to the types of *OCL* targeted and the various applications in which the systems might be applied.

5.1 Data Sample Collection

5.1.1 Data Retrieval.

Data sources. Data samples are collected from various sources on the Web (Table 3). Twitter is used in majority due to its popularity and the easiness to get data, while other social media (Formspring, YouTube, MySpace, Wikipedia, and Facebook) are used less [167]. Various sites such as the news website Gazzetta.it [198] usually specialized in one topic such as sport or politics and discussion forums such as voat, 4chan, or reddit are also investigated. Table 4 shows the distribution of languages in the publications and highlights a strong unbalance between English (74.4%) and the other languages present only in 1 to 6 papers.

Yet, recent works exhibit efforts towards the diversification of the objects of study. Datasets are created for less-studied languages such as Hinglish [61, 139], Bengali [147], and Arabic [62, 123], revealing new challenges pertaining to the particular language structures (e.g., in Hinglish, the grammar is not fixed, the written words use Roman script for spoken works in Hindi [139], a list of challenges for Arabic is proposed in Al-Hassan et al. [5]); and for less-common social media platforms (e.g., YouTube comments [62, 147]).

Following these works, we consider worth building new datasets to investigate more sources and languages and increasing the research on cross-sources for more adaptability of the models [119]. Machine translation models in conjunction with English-based classifiers could also be investigated, especially for datasets that mix multiple languages.

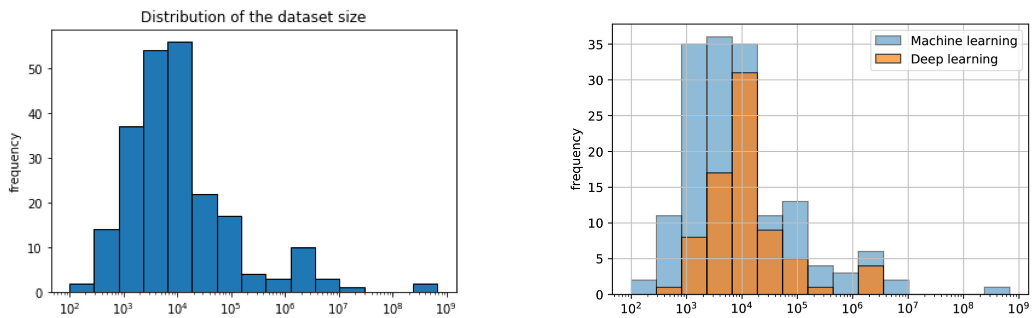
Data mining methods. Most datasets are collected by retrieving samples that contain specific elements, such as abusive words [133], hashtags, and keywords from controversial politics sites [38] or offensiveness dictionaries [221]. Several papers use snowball sampling [130, 216] or variations such as first retrieving tweets based on hashtags and then all the other tweets from their authors [264]. Others are retrieved by crawling entire pages selected for their likeliness to contain *OCL*(e.g., anti-Islam pages [270], offensive blog posts [83], public celebrity pages [97]), or by crawling and randomly sampling social media feeds [182, 248]. Additional filtering based on keywords or negative vocabulary is sometimes applied to maximize the number of *OCL* samples [209]. Similarly to psychology studies, the authors of Reference [228] manually create cyberbullying scenarios from which students write an entire discussion used as dataset.

Fifteen percent of the classification papers simplify the detection task by distinguishing smaller tasks of sub-topics that share similar properties. Researchers use datasets for specific *OCL* subtype (e.g., datasets on sexism and racism for hate speech [106, 192, 205, 214, 283, 285, 303], on hateful speech towards black people, plus-sized individuals, and women [233], or towards refugees and Muslims [42, 304]), or domains (e.g., news, politics, entertainment, business for insult detection [255] or disability, race, and sexual orientation for hate speech [47]).

Introduction of biases. Each parameter setup for data collection biases the dataset. The choice of data source, keyword for retrieving initial sets of samples, and languages for these queries directly impact the type of users for which the subsequent trained model will show good performance. Less obvious choices also skew the data distribution; for instance, through the selection of random samples from a forum history or by selecting only the first posts. In both cases, the topics discussed might be more or less detailed, or the authors of posts might use more or less strong *OCL*. Skews are also introduced by a crawler's (human or automatic) browser setting, e.g., due to the geographical region or search habits. Poletto et al. [208] discuss further certain of these biases in their survey. The period of time when the dataset is collected is also of importance. This concern is highlighted in computer vision, such as for the Pascal VOC dataset [128], reportedly collected in January, and composed of an above-average number of Christmas trees, as images in Flickr (the media they used) were ordered by recency. Machine learning models for *OCL* detection are especially sensitive to the events contained in the data [98], as these events shape the type of language and topics the models can interpret. Ptaszynskia et al. [211] recommend regularly collecting samples to update datasets with the most recent vocabulary. Sampling per keyword also introduces biases in the datasets [102]. The samples retrieved often contain words considered rude, while more subtle forms of *OCL* might not be accounted for. Founta et al. [102] instead propose to collect data by combining random sampling and tweets retrieved using keywords.

These biases become harmful when they skew the data distribution away from the expected distribution or enforce discriminatory associations between attributes. According to the bias framework of Suresh et al. [261], *representation biases* manifest when the training data distributions integrate few information around underrepresented populations, leading to low model performance. This definition could be expanded to over-represented populations, for which a model might learn spurious correlations, and to "population" as either individuals or other kinds of concepts such as conversation topics.

Various fields (e.g., linguistics) study the different strategies employed to express *OCL*; for instance, when expressing hate [18]: othering, stereotyping, conceptual metaphors, implicitness, constructive and fictive dialogues. Linguistics identifies these strategies for individual topics— e.g., "conceptual metaphors in comments related to migrants in Cyprus"; or media studies—e.g., "in the case of racism, it was found the use of vicarious observation, racist humor, negative racial stereotyping, racist online media, and racist online hate groups. The online hate against women tends to



(a) General distribution. (b) Distribution per classification method.

Fig. 7. Distribution of the number of training data employed in classification tasks.

use shaming. [...] flaming, trolling, hostility, obscenity, high incidence of insults, aggressive lexis, suspicion, demasculinization, and dehumanization can inflict harm" [51]. This information could be exploited to verify the diversity and representativeness of the samples collected in a dataset.

Dataset collection parameters are not always aligned with insights from psychology. While psychology puts forward context as important for classifying *OCL*, most posts are stripped down from their metadata and conversational context. Pavlopoulos [199] did not find any interaction with the title and the previous sentence of a post, yet context can be broader, e.g., the whole discussion, and merits further investigation. Multiple challenges in reference to this mismatch are discussed in Section 4.4.

5.1.2 Data Processing.

Data augmentation methods. Figure 7(a) shows the distribution of the number of training data employed in the classification tasks, with a majority of datasets around 1,000 and 11,000 samples. As expected, deep learning approaches make use of larger datasets (about 10,000 samples) than traditional machine learning approaches (about 5,000 samples)—Figure 7(b).

Despite needing large datasets, only 14% of the classification papers mention explicitly data augmentation techniques, mainly to balance datasets. This is common, as Web platforms contain a majority of non-*OCL* text (e.g., abusive tweets only represent 0.1% to 3% of tweets [102]). This extreme unbalance explains why certain papers further retrieve data using *OCL* seed words, instead of performing synthetic data augmentation. Out of the 69 papers whose figures are available, 39% have a balanced dataset.

Data augmentation is performed either by over-sampling or by under-sampling certain classes or both. Nine papers randomly duplicate the minority class samples and 8 remove samples from the majority class. Six papers employ the **Synthetic Minority Over-sampling Technique (SMOTE)** for over-sampling by creating artificial data samples in the feature space. Two create synthetic data with two-way sample translation and sliding windows [229] or with random sample generation with a character encoding and introduction of known *OCL*words in these sequences [243].

The different data augmentation methods do not all perform well for each classification task [57]. Thus, we not only recommend to investigate data augmentation further, but we also propose to create a list of large datasets for each type of *OCL* so researchers have common benchmark datasets for evaluation, as suggested for abuse detection by Jurgens et al. [137]. Poletto et al. [208] propose a review of existing benchmark corpora that supports the identification of missing text corpus. Existing datasets could be merged together to augment their size. Deep generative models are also recently investigated to synthesize new data samples automatically, with promising results [293].

Further investigation of their conditions of applications, and of the choice of hyperparameters, would be beneficial.

Next to balancing a dataset, Park et al. [193] augment their dataset by substituting female entities to males ones and vice versa to reduce gender bias. The validity of the synthesized data samples would merit being further investigated in relation to the specific types of *OCL* of each use-case, especially when studying multiple sub-categories of *OCL*.

Pre-processing data samples. Most papers employ a standard form of data pre-processing for the English language (stop words removal, tokenization, stemming, lemmatization) [258], with few variations when the language varies. One paper for the Indonesian language additionally uses a dictionary to transform informal words into formal ones [133]; another for English removes the rarest words from the samples [102], and researchers tackling Japanese use methods specific to this language (e.g., Japanese POS [190]).

Introduction of biases. As a sign of representational biases, Grondahl et al. [119] show that models performing well on a dataset with the same distribution as the training dataset, perform poorly on other datasets, but perform equally well when they are retrained on a dataset with this other distribution. These results suggest that the architecture of the model is not the primary factor for the resulting performance, but that the datasets themselves all contain their own biases, hindering generalization to other datasets.

Data augmentation and processing reinforce or introduce representational biases. For instance, most data instances that are representative of a certain *OCL* might deal primarily with a certain topic. Augmenting the dataset for the *OCL* class would then reinforce the presence of this topic in association with the *OCL* label. Also, basic pre-processing activities such as stemming and lemmatization can remove useful indications, e.g., gender word endings in gendered languages, skewing the data towards one single type of representation. The curation of misspellings might skew the representation of populations that frequently use such spelling. Grondahl et al. [119] experimented with natural-looking adversarial perturbations—which could be misspellings—and showed that models are not robust to those. Besides, misspellings are not all spelling mistakes, but can be meaningful, and vary the interpretation of a sentence from the "clean sentence." Curating the data then prevents a model to learn such new types of interpretations.

In other domains such as computer vision [125, 176, 218], models are made less brittle by augmenting the datasets with natural or adversarial perturbations that could arise at deployment time. We suggest to test similar solutions in the context of *OCL*. Especially, brittleness to natural perturbations such as voluntary or unintentional misspellings might be partly due to the ways data are processed: When misspellings are resolved, the models are not trained on such diverse, possibly adversarial inputs, increasing their brittleness.

5.1.3 Data Splitting. Dataset splitting is not standardized in the *OCL* detection pipelines. Arango et al. [14] showed it can lead to overestimation of models' performance. When it is done after feature engineering (or after data augmentation and curation), information from the test data is leaked into the training data, as the feature extraction methods might rely on data distributions, resulting in obtaining high performance in laboratory settings but low performance in deployment.

This highlights general issues with the management of data in research settings. If the data are studied along time, then it is important not to sample them randomly but follow this temporal sequence to observe how generalizable a dataset from one time window is to another time window. These and more issues are also identified in the general data management literature for machine learning [239]. The implementation of common benchmark structures respecting these data management rules would support the propagation of good practices in the preparation of datasets for the training and evaluation of models.

5.2 Data Annotation Collection

Here, we discuss how dataset annotations are collected. Annotation refers to the labeling of data instances (e.g., a sentence or a tweet) that might contain *OCL*. These annotations are usually collected by aggregating the inputs of multiple annotators into a single label to ensure its quality. Ninety-five percent of the 80 papers with available information go through this human annotation phase. A few papers instead use machine learning [48], inference from data context [233, 264], or semi-supervised learning [109] to infer labels.

Notably, some works mentioned by Fortuna et al. [101], build lexicons of *OCL* [94, 288] to train better classification algorithms. We do not include them here, as they do not correspond to the annotation of evaluation datasets and do not detail their crowdsourcing setup.

5.2.1 Set-up of the Annotation Process.

Instructions to the annotators. A binary question is typically asked to the annotators (the answer "undecided" is sometimes added), potentially with a rating [45, 56]. However, it is argued in psychology literature that rating comments on a valence scale is too vague for the annotators, who prefer binary questions [254, 255]. Closer to psychology that asks annotators to rate several propositions, Guberman et al. [120] investigate perceived violence of tweets through an adapted version of the multiple proposition **Buss-Perry Aggression Questionnaire (BPAQ)**. Using six annotators on Amazon Mechanical Turk and 14 gold questions (12 correct answers required), they still found 30% disagreement that they partly explain with the non-adaptation of the questionnaire to tweet violence.

Out of the 74 papers using crowdsourcing, only 32% mention giving a definition of the concept to annotate to the annotators, such as detailed offensiveness criteria[11, 12] and hate speech definition.[13] Gamback et al. [106] through several crowdsourcing tests provide a detailed question to the annotators.[14] Not providing clear definitions is an issue, because the annotators might have different definitions of *OCL*in mind, leading to collected data labels that would not be suited to the goal of the application.

Data annotators. The annotation tasks are conducted on crowdsourcing platforms or programs created by the authors of the publications. Certain papers show that the type of annotators employed influences the quality of the annotations. CrowdFlower (now Appen.com), expert and manually recruited annotators are equally used (23.7% each), while students of universities (13.8%) and

[11] "A tweet is offensive if it (1) uses a sexist or racial slur; (2) attacks a minority; (3) seeks to silence a minority; (4) criticizes a minority (without a well-founded argument); (5) promotes, but does not directly use, hate speech or violent crime; (6) criticizes a minority and uses a straw man argument; (7) blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims; (8) shows support of problematic hashtags. E.g., "#BanIslam," "#whoriental," "#whitegenocide"; (9) negatively stereotypes a minority; (10) defends xenophobia or sexism; (11) contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria." [285].

[12] "tweets that explicitly or implicitly propagate stereotypes targeting a specific group whether it is the initial expression or a meta-expression discussing the hate speech itself" [109].

[13] "the language which explicitly or implicitly threatens or demeans a person or a group based upon a facet of their identity such as gender, ethnicity, or sexual orientation" [108].

[14] "Does the comment contain a personal attack or harassment? Targeted at the recipient of the message (i.e., you suck). Targeted at a third party (i.e., Bob sucks). Being reported or quoted (i.e., Bob said Henri sucks). Another kind of attack or harassment. This is not an attack or harassment."

Amazon Mechanical Turk (15%) are less. The expert category comprehends authors themselves, researchers of similar fields, specialists in gender studies, and "non-activist feminist" for sexism annotations, persons with linguistic background, trained raters, educators working with middle-school children, and people with cyberbullying experience.

Annotation aggregation. Among the 50 papers for which the information is available (out of the 74 papers using crowdsourcing), 49 papers aggregate the annotations from multiple annotators into binary labels. Seventy-eight percent use majority-voting, 10% filter out samples for which there is no full agreement between the annotators, 8% create rules that define how to aggregate according to different scenarios of annotations (e.g., majority-voting and removal of the samples with the highest disagreement rates and the samples for which the annotators agreed they are undecided [46]). One paper uses a weighted majority-vote scheme [130]. Only Wulczyn et al. [292] derive percentage from the annotations.

Annotation quality control. 32.4% of the papers mention techniques to obtain high-quality labels. Within the annotation task, they investigate using precise definitions and clear questions to remove ambiguities [227]. After the task, annotations are aggregated to resolve disparities between annotators' opinions, and low-quality annotations or annotators are filtered, with quality scores computed over the history of the annotators, the time they take to answer each question, or their answers to gold questions [129].

Half of the tasks have 3 annotators, 15% make use of 5 annotators and 22% of 2 annotators. Using an odd number of annotators enables to break ties in annotations with majority voting, while using 2 annotators is cheap and fast. The rest of the tasks employ 1, 4, 6, or 10 annotators. The papers using more than 5 annotators per sample are rare, most probably because of the cost. Using only the cases of full agreement among amateur annotators produces relatively good annotations compared to expert annotators, and they suggest to use experts only to break the ties of the amateur annotators [283].

Different metrics are employed to evaluate the annotation quality by measuring the agreement between annotators (Figure 8). Most papers use Cohen's Kappa for 2 annotators and Fleiss' Kappa for more. 22.9% of the papers mention "inter-annotator agreement" or "kappa" scores without further precision. Krippendorff's alpha and the percentage agreement are less adopted, the second one making a possibly wrong assumption that the majority is correct [170]. In the publications, we notice a high proportion of low Cohen's Kappa and Fleiss' Kappa scores (under 0.6) for tasks with 3 or 5 annotators, which proves the difficulty to design unambiguous tasks and hint at the subjectivity of the concepts to rate.

5.2.2 Biases in the Annotation Process. The data annotation process introduces various types of biases with each of the design choices.

Identification of mismatches. Here, we take the hypothetical scenario of developing a dataset for aggression language. Certain definitions of aggression highlight the need for looking at the context of a sentence, at the behavior of its author, and at the person judging this language, to understand how a sentence would be perceived, e.g., aggression is "neither descriptive nor neutral. It deals much more with a judgmental attribute" [177]. Psychology identified the variables that influence this judgment, mostly "cultural background" [44], the role of the judge, i.e., aggressor, target, observer, and so on, "norm deviation, intent, and injury," but also "the form and extent of injuries actually occurring" [161]. To obtain a controlled and realistic dataset and reduce ambiguity, these pieces of information around the annotators of the language would be needed, the annotator's role (e.g., victim or observer) should be decided, and the context of the sentence (e.g., harm caused by a sentence) displayed.

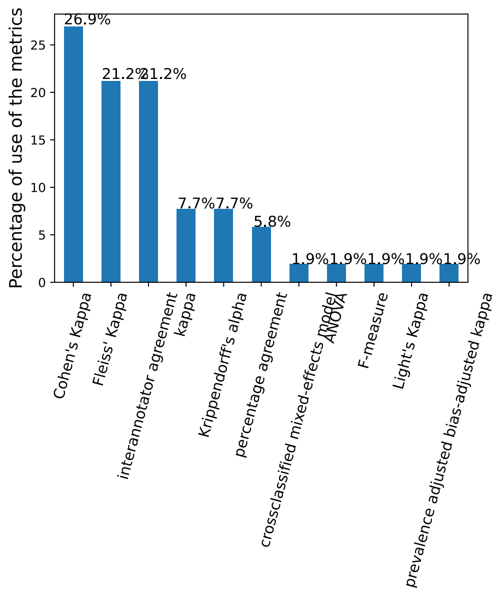


Fig. 8. Distribution of the metrics used to evaluate the annotations.

A similar example is the perceived offensiveness of group-based slurs, which depends on the perception of the status of the target group [126]. In this case, both the context and observer are of importance, since the social status of a target group could be uncovered from context knowledge but can also depend on the perception of the observer.

These issues resonate with the historical biases in machine learning ethics literature [261]. In the dataset, there is a mismatch between the judgments of the annotators, the judgments of the actual targets of an *OCL*, and the judgments from external observers. Consequently, the dataset is not aligned with what the machine learning model is expected to learn.

Missing context information. Psychology literature showed that for many conflictual languages, the sample context influences the perception of a sample. Most crowdsourcing tasks, however, do not specify it, neither in the instructions nor within the sample presented to the annotator [53, 240]. Guberman et al. [120] put forward the insufficient context that leaves many aspects of the text to interpretation as a reason for disagreement in harassment annotations. Golbeck et al. [116], while not including any context in their corpus, acknowledge this limitation and develop precise annotation guidelines that aim at removing ambiguities stemming from the absence of context. Ross et al. [227] provide a definition of the *OCL* to annotate and find that the task remains ambiguous, suggesting that even for objective tasks, context information might be missing to provide an objective rating.

The type of context to include and its framing (e.g., a conversation, structured information about multiple characteristics) remain to be investigated to address ambiguities, while controlling the cost of the annotations. Pavlopoulos et al. [199] have already shown that annotations with conversational context (post and its parent comment, as well as the discussion title) significantly differ from annotations without it. Sap et al. [236] have primed annotators with dialect and race information explicitly to reduce racial biases in annotations (more samples written in African American English than in general American English are labeled as offensive). Creating datasets that tackle single specific contexts such as "hate speech against immigrants and women" is also a direction to investigate [28].

Lack of annotator control and information. Psychology highlights that many *OCL* are subjective. Linguistics also shows the diversity of interpretation of *OCL* by different communities or within a same population [18]. For instance, a study shows that in Malta, participants typically identify homophobic comments as hate speech, but not necessarily xenophobic ones, and explains it with the recent acceptance of the LGBTQ community in the Maltese society, while "migrants are still very much left on the periphery." Similar studies in other regions of the world would probably lead to different conclusions, illustrating the importance of the annotator background. Hence, choices in the crowdsourcing task design that impact the pool of annotators (country of origin of the annotators, language, expertise, educational background, and how they are filtered) integrate implicitly biases in a dataset.

Psychology indicates characteristics of an individual that impact one's perception of a sentence relative to an *OCL*. Some of these characteristics are also observed in computer science papers, such as the differences of annotations based on gender [120]. Communication studies also investigate the characteristics of an individual that impact their willingness to censor hate speech and identify age (e.g., "older people are less willing to censor hate speech than younger people"), neuroticism, commitment to democratic principles, level of authoritarianism, level of religiosity, and gender [151]. Such factors could possibly also impact one's attitude toward annotating hate speech. While the design choices do not map to these characteristics, creating schemes to control, or at least measure them, is a valuable research direction. Certain crowdsourcing frameworks [27] are a first step towards this control. Verifying that the same characteristics apply in the online and offline contexts is also important following previous contradictions, e.g., one computer science study observed that annotators from both genders usually agree for clear cases of misogyny and disagree for cases of general hate speech [290], contradicting findings in psychology literature.

Additional properties of the annotators, not investigated in psychology, can bias the datasets. For instance, annotators from crowdsourcing platforms, who have no training on what hate speech is, are biased towards the hate label, contrary to expert annotators [283]. Research is hence also needed in assessing the level of education around *OCL* that annotators have, in educating them, and in maintaining them engaged for more annotation tasks.

Simplification of the annotations. The way the annotations are processed creates biases. Aggregating the annotations into single labels does not allow for subjectivity and skews datasets towards certain types of perceptions, generally the majority opinions [22]. This might raise issues of unfairness—non-inclusion of certain opinions—and reinforce filter bubbles. For instance, Binns et al. [34] show that a toxicity detection algorithm performs better on annotations from male users than from female ones and is consequently unfair to women. This reflects *aggregation biases* [261]: A single dataset to train a single machine learning model for a whole platform is collected, whereas different populations need adaptation.

Subjectivity brings new challenges in measuring and obtaining "high-quality" annotations. Measures of quality are now centered around agreement—the lowest the disagreement, the highest the quality—and post-processing methods use the majority opinion, yet the majority is only one perception of a subjective *OCL*. Instead, methods should filter out annotations that are obviously incorrect—often due to spams—or erroneous for different individuals, while accounting for the existence of multiple relevant and disagreeing judgments. For that, works from the human computation community, such as CrowdTruth [17], which provides metrics for the quality of annotations and annotators without assuming the existence of a unique ground truth, could be investigated. More annotators might be needed, and schemes to infer relevant clusters of annotators could be

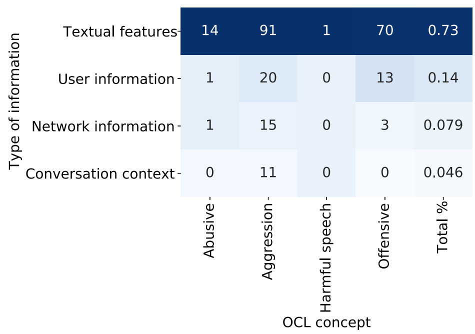


Fig. 9. Type of information used by the classification methods according to the *OCL* concepts.

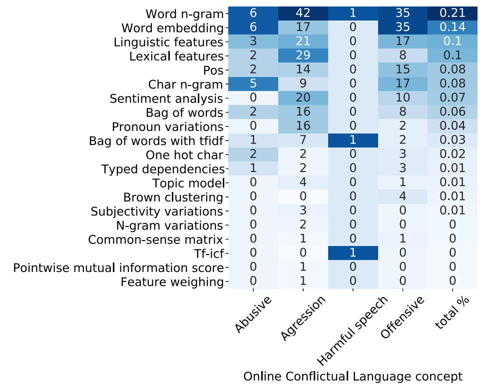


Fig. 10. The textual features per *OCL* coarse-grained concept used in the classification papers.

investigated to trade off between quality and cost considerations. Mishra et al. [171] noted that in digital media, a small amount of users frequently give their opinions, ranking positively highly offensive posts—a form of bias towards the opinion of these few users. The researchers propose a semi-supervised method to identify these biased users and correct the ratings.

Leveraging psychology and human computation methods. Research from other fields could be adapted to improve *OCL* annotation pipelines, as recommendations from crowdsourcing literature or psychology are not necessarily followed for now. Only 32% of papers mention methods to ensure a level of quality (e.g., golden questions, annotator quality score, precise definitions of the terms) and few papers employ more than five annotators per sample, whereas crowdsourcing literature encourages that. Taking inspiration from psychology and judgment collection methods can also be a promising direction. Psychology studies use multiple questions with scales, whose answers are aggregated to collect the perception of each person (e.g., 10, 6, 3 propositions on [1; 9], [1; 6], [1; 12] scales [37, 68, 186]). To measure offensiveness, participants rate images visualizing a scenario along how comfortable, acceptable, offensive, hurtful, and annoying they are on a 7-point Likert scale [289]. Cunningham et al. [70] show scenarios with four situations to participants, who select the most offensive one. Example scenario and situation are, respectively, attending a men's basketball game and "A Caucasian, female said: 'Of course we lost. We played like a bunch of girls."' While these studies are not specific to *OCL*s, the general method could be used, and the specific questions investigated. The challenge of asking such questions while maintaining the cost low would become important.

6 CLASSIFICATION MODELS FOR THE DETECTION OF *OCL*

In this section, we discuss the algorithmic methods used for *OCL*detection. We focus on the features extracted from data, on the algorithms, and on the selected evaluation procedures. We aim at identifying implicit biases integrated into the design choices of the detection pipelines.

6.1 Features for Classification

6.1.1 Types of Features Extracted from the Data. Features employed in the classification models use four main types of information, detailed below and summarized in Figure 9.[15]

[15]Interested readers can refer to Schmidt et al. [240] and Fortuna et al. [101] for an extensive explanation of the properties of each feature.

Textual features. Advantages and disadvantages of the features are explained in Reference [101], we briefly mention their variants in Appendix A.5. Textual information is represented differently, depending on the classification methods. Word n-grams, **bag of words (BoW)**, and embeddings are employed in majority, because they are adapted inputs to machine learning classifiers. Word n-grams represent more information (order of the words) than BoW, which improves the classification performance, while word embeddings are recently developed for deep learning. Certain features are rarely investigated (common-sense matrix [88], **tf-icf (Inverse Category Frequency)** [156], pointwise mutual information score [181]), and merit more research in the future. The distributions of the textual features used across *OCL*coarse-grained concepts (Figure 10) are mostly similar, which indicates a potential lack of adaptation of the individual features to each task at hand.

Information about the users (emitter and reader). This is the second most used information for classification. It includes the user popularity in the social media based on the number of followers and friends, the user activity based on the number of posted and liked tweets [73, 102, 308], her gender [283], age [73] and location [124, 285], the subscribed lists and the age of the account [102], and information extracted from the conversation history such as the frequently used terms [283], the tendency to use *OCL* [205] or the Second Order Attributes representation of the link between documents and users [13]. These characteristics might be studied for a user across social media platforms [72].

Information about the network of the users. Often it consists in measuring how much a user reciprocates the follower connections she receives, "the power difference between a user and his mentions, the user's position in his network (hub, authority, eigenvector, and closeness centrality), as well as a user's tendency to cluster with others" [102], but also graph metrics computed over the combined social networks of the sender and receiver [132, 256].

Conversation context. This is the conversation [29] or the set of questions and answers [183, 226] surrounding the data samples, the images found with the textual samples in the social media [130] and their captions [131], information about the parent-child relationships of the samples in the conversation [159], or information about the samples themselves such as the popularity of a post among its social media [131, 263] or its publication time [131].

6.1.2 Feature Selection. Certain papers start with a large amount of input features and then decrease the dimensionality to improve the classification performance.

For this, 12% of papers use feature selection methods: Chi-square [38] (5), Singular Value Decomposition [88] (5), information gain [191] (3) or mutual information [241] (2) based selection, Fisher score [306], recursive elimination with logistic regression (training a classifier with all the features but one, and eliminating the one leading to the worst performance) [241] or simply evaluating a classifier on different subsets of features and selecting the one with the best performance [122], backward selection (removing variables with high correlation) [131], test statistic (Student t-test) [241], PCA [64], Latent Semantic Analysis [130].

Feature weighting is used with SVM scores [210], logistic regression weights [241], or by computing a score that represents the easiness to falsify the outputs of the classifier with one feature and selecting features based on this score [110].

Yoshida et al. [299] compute an entropy score indicative of whether a word corresponds to a sentiment and define a set of rules to select the words to keep, and Lee et al. [156] compute the less common words in a set of documents.

6.1.3 *Introduction of Biases.*

Measurement bias. The choice of features automatically biases the model towards using certain types of information and biases its outputs towards specific types of errors. This is a *measurement bias* [261], where the choice of features might leave out factors that are relevant for inference. In the following, we identify various measurement biases.

Mismatch with psychology. We identify measurement biases in the way features are engineered. The inputs to the classification methods are mostly textual information. Although psychology shows that the context surrounding text also impacts *OCL* perception, only 23% of papers use additional information (Figure 9). Non-textual features are mostly used for the classification of aggression language, possibly because it is characterized by the behavior of users, however, the other types of languages are also impacted by context. The way the feature dimensionality is reduced also impacts the type of information used by a model.

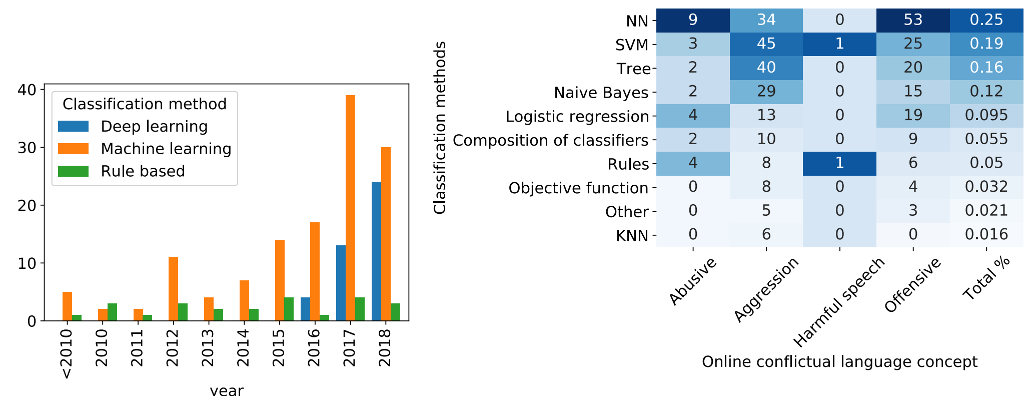
The information used often does not correspond to the variables identified by psychology, which might explain performance issues [132, 199, 304]. Measurement biases also reflect the non-consideration of subjectivity. Adding to the common features other features describing users would allow to personalize inferences, which would render the models more inclusive of various opinions. One main challenge here would be to define precisely which information should be extracted from the datasets into features, and how to represent it effectively.

Lack of OCL-dependent features. Several experimental studies show the difficulty for machine learning models to distinguish between different *OCL* [164, 279] (e.g., difficulty to differentiate between hate speech and profanity [164]). Also, our systematic survey shows a lack of adaptation of the features to each specific *OCL*. While feature engineering might not seem entirely relevant with deep learning, we suggest to study the introduction of hand-crafted features to differentiate between these *OCL*, inspired from the psychology literature and our categories in Table 8. For example, someone interested in offensive language could explicitly integrate the identification of the targeted individual or community in a language sample, instead of letting the machine learning model eventually discover these characteristics. This comes hand-in-hand with creating more adapted datasets where the different types of *OCL* have to be well-represented and the necessary information present.

Recent works show promising results in this direction. Training word embeddings on a specific hate corpus and appending manually crafted features specific to the target class achieves higher accuracy performance than pre-trained embeddings or more traditional features (e.g., n-grams), for the classification of various intensities of Islamophobic hate speech [279]. Zhang and Luo [303] extract more informative features than classic ones like n-grams by using deep learning structures that learn relations between words.

Low classification performance also comes from the lack of adaptation of the features to the specific ways people use *OCL*in different social media, such as making spelling "mistakes," mixing languages in informal language [133, 150], using language that follows evolving trends over time [150, 182], using implicit *OCL* [155]. We recommend to specifically investigate how to integrate these characteristics into future models. For instance, Alorainy et al. [9] extract features specifically to identify othering language, Bansal et al. [26] and recent publications in ACL workshops [19] focus on humor and sarcasm.

Discriminatory features. Recent concerns have been voiced around the discriminatory character of certain features, especially those ones coming from word embeddings. Caliskan et al. [49] adapted a psychology test (Implicit Association Test) to measure biases in word embeddings and



(a) Evolution of the classification methods over years.

(b) Distribution of classification methods per *OCL* coarse-grained concepts. Colours represent the frequencies of methods per concept (white to blue, lowest to highest frequency).

Fig. 11. Quantitative analysis of the classification methods.

showed that these embeddings reproduce historical human biases. Garg et al. [111] showed that training embeddings on text corpora from different time periods incorporates in these embeddings the job-related biases from the various periods.

Methods exist to debias such embeddings [40, 43, 305]. Although not focused on *OCL*, they could be investigated, as some of them rely on training word embeddings to extract adapted features. One might search for the biases introduced when word embeddings are trained on *OCL* corpora, instead of general natural language processing corpora.

6.2 Methods for Classification

6.2.1 Overview of the Classifiers. We note three main trends in the classification methods: rule-based models, machine learning models—that we define as simple classifiers—and deep learning models. 4.7% of the papers combine several models with ensemble and boosting methods. Although computer science papers report performance measures, it is difficult to tell which are the "best" methods, as the measures are not obtained from the same datasets.

The use of machine learning methods has increased over years since 2012 (Figure 11(a)), following the general increase of *OCL* research. Research on deep learning for *OCL* started in 2016 with the general increase in deep learning research, and its amount increased quickly, almost catching up with machine learning research. Research on rule-based methods has been constant over years and rarely adopted.

Among these three categories, various methods are used. A majority of machine learning papers use **Support Vector Machines (SVM)**, tree-based classifiers (decision trees and random forests), **Naive Bayes classifiers (NB)**, **Multi-Layer Perceptron (MLP)**, and **Logistic Regression (LR)**. Deep learning papers mainly investigate **Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN)**, and their combinations. These methods are further explained with their variants in Appendix A.6. Figure 11(b) shows that regular deep learning, SVM, tree-based, and rule-based classifiers concern every type of *OCL*, while research on naive Bayes classifiers, composition of classifiers, and optimization of application-tuned objective functions has been sparsely conducted especially for harmful and abusive languages.

6.2.2 Training Process. Most publications follow the same pipeline: dataset collection and model creation. However, 5% of the papers diverge. During the training process, they perform active learning [171] or semi-supervised learning where part of the training data samples do not have labels but these samples are still used (often by label inference) [171, 217, 299]. They perform feature selection and classifier learning simultaneously [310]. Certain papers employ transfer learning by incorporating a learned word probability distribution in the target domain to the classifier for training efficiency [4, 190, 243] or to reduce gender biases [193].

Besides, a few papers compare the performance of models trained on the whole dataset or trained by cutting the dataset into domains and by learning a multi-class classifier (one class per domain) (e.g., cyberbullying related to race, sexuality, and intelligence [88, 89]). Other papers detect the sub-types of the concept instead of simply detecting the coarse-grain concept (e.g., detecting cyberbullying by classifying curse, defamation, defense, encouragement, insult, threat, and sexual talk [273], detecting misogyny by classifying discredit, sexual harassment, threats of violence, stereotype and objectification, dominance, derailing [12]).

6.2.3 Introduction of Biases. The choice of classification algorithm and its hyperparameters participates in the introduction of various biases in the outputs of classification models.

Aggregation bias. Such bias is defined by the development and application of a single machine learning model on various distinct populations [261]. This practice is problematic for subjective *OCL*. A solution could be to learn distinct models on sets of annotations from different populations, possibly also taking into account the context of application and learning distinct models for different platforms for instance. Sharing some information across models while fine-tuning them for specific context remains to be investigated in order not to require too large amount of data and too large computational resources.

Mitigating discriminatory biases. A large body of literature on machine learning for structured data highlights unfairness issues for decision-making systems, propose metrics [276], mitigation methods [104], and toolkits [30] to explore the causes of unfairness and to support industry practitioners in integrating these formalizations of fairness into their practices. Recent works have introduced different methods to debias the outputs of NLP models, e.g., by transforming the features employed, by modifying the optimization objective employed to train a classifier (e.g., adversarial training of deep learning models with a regularization term corresponding to the protected attributes at hand [294]), or possibly by transforming the outputs of the classifier [260]. A more extensive account of such works is given in Reference [260]. In certain cases, the training process is also modified to involve a bias expert [61]. Few recent works propose sample weighing methods to account for dataset biases, respectively, in toxicity or hate speech detection tasks [175, 301], and integrate knowledge bases to correct datasets from biases by substituting words indicator of identity by more general entities [21].

Most works are not specific to *OCL* and need adaptation. For example, some works do not easily translate to classification tasks of more than two classes, but this becomes necessary for *OCL*. Tradeoffs between discriminatory biases and performance measures [193] nudge for works at the intersection of natural language processing and human-computer interaction to understand how to set acceptable thresholds for the metrics. Toolkits could also be developed. Besides these more usual notions of unfairness, a new type of unfairness with regard to the social network centrality of a potential victim of cyberbullying is also exposed in Singh et al. [249] and would merit further investigation.

Debugging biases and other errors. Investigating how to apply interpretability methods to *OCL* classification tasks could enable to understand specific causes of the low performance or unfairness of the classifiers for specific samples. Little effort has investigated such direction until now: Risch et al. [223] with usual interpretability methods and Cheng et al. [59] with the causality angle for performance, and Kennedy et al. [140] for biases.

Human-in-the-loop methods could be developed to identify the shortcomings of trained models by asking humans to generate samples that lead the model to a wrong prediction. This could serve to identify more social biases or simply to make the model more robust to tricky samples. In this direction, Dinan et al. [90] asked crowdworkers to generate sentences that would break their offensiveness detector and noted that crowdworkers identify samples of a nature, which is rare in the original dataset, with less obvious profanity but more figurative language and language that requires background knowledge to be interpreted.

6.3 Performance Evaluation

6.3.1 Evaluation Dataset.

Data samples. To evaluate the models, the dataset is divided into training and test set, and performance metrics are computed on the test set. Some works now also evaluate their models on other datasets that have different distributions, to understand how generalizable the models are. This emulates the production setup, where new data samples are continuously inputted, for which the distribution might differ from the training one when new users and new context are added. Few works [182] evaluate the classification performance along time.

Ground truth. While most papers consider binary labels as ground truth, some aggregate the crowdsourced labels into continuous scores to investigate whether a model learned the distribution of judgments or the majority labels. A distinction between the data samples whose labels received full consensus and the data samples of lower consensus is also sometimes made [155] for explanation's sake, i.e., better understanding where errors come from.

6.3.2 Evaluation Metric. A small number of metrics is used: F1 score (macro, micro, or average) (23.8%), recall (22.9%), precision (20.5%), ROC-AUC (7.9%), accuracy score (14.3%), true negative, false negative, and false positive rates (4%). Accuracy is discouraged, because its measure is impacted by unbalanced datasets. Accuracy, precision, and recall are calculated on average for all the classes or for the different classes separately.

Few papers use the Cohen's Kappa score [89, 233], the Kappa statistic [53, 88, 231], the Spearman correlation [197, 292], the precision-recall curve with the precision-recall breakeven point [110, 160, 255], and the Hamming loss [229] as an evaluation metric. Others use error calculation-based metrics such as the mean squared error [53, 81, 171, 180], the root mean square forecasting error, and the mean absolute percent error [210]. Park et al. [193] use the False Positive and False Negative Equality Differences to quantify gender biases.

Some publications assess the time taken to train the models or the time to detect the *OCL* [165, 215, 224, 300]. Some papers further study the performance of the models by investigating in more detail the types of sentences usually missclassified.

6.3.3 Accountability and Transparency. There is generally no common dataset and evaluation metric to compare models. Benchmark datasets would ideally include context information and information about the annotators and state clearly the scope of the dataset. Using the same metrics across publications that target the same goal would be helpful. The advantages of the less-frequent metrics should be investigated.

Reporting the pipeline used to build the datasets would allow to better understand their limitations and biases. As suggested by literature on transparency, datasheets [112] could support the controlled use of the datasets, both in research and industry. This relates to *deployment bias* [261], when a model is used for an application it was not built for.

Table 5. Summary of Biases Introduced in the Online Conflictual Language Detection Systems through the Design of the Data Collection Pipelines and of the Classification Models

Data Collection	**Sample retrieval**	Source & time → contextual bias; Keyword and rank biases; Topic & language biases; Representation bias; Collection of context information
	Dataset processing	Data augmentation bias; Pre-processing biases
	Dataset splitting	Information leakage → Evaluation bias
	Sample annotation	Annotator *OCL* knowledge; Annotator background; Annotation instruction; Presentation of context; Annotation aggregation
Model	**Feature engineering**	Measurement bias (context, psychology); Discriminatory features
	Classification algorithms	Aggregation bias; Discrimination bias
	Performance evaluation	Evaluation bias; Data representativeness; Metric relevance

6.3.4 Refinement of the Metrics. Most frequent metrics reflect the accuracy of a model, which is not necessarily aligned with what end-users deem important. For subjective *OCL*, evaluations could be personalized to the different perceptions of users, depending on their background [188]. To measure user satisfaction, metrics inspired from the machine learning fairness literature [276] could be adopted, e.g., measuring the accuracy of the model inferences for groups of users and computing their ratio. These issues are termed *evaluation bias* [261], where the metrics employed or the scope of the evaluation dataset do not correspond to the type of samples or the goals for which a model would be used in practice.

Unfairness issues in datasets and classification outputs also need systematic investigation, for instance, using existing fairness metrics. Yet, it is important to accurately interpret these metrics, as they might simplify too much the actual discrimination issues, and optimizing for them might not lead to fair results in practice [189, 242].

Critical studies [33, 269] have been published recently in computer vision, evaluating benchmark datasets and issues with performance metrics (e.g., top-1 accuracy might underestimate the performance of a model, while multiple labels could be relevant for a same image), showing how they lead to correct or wrong conclusions. Inspiration could also be taken to develop better mental models of the functioning of the *OCL* detection systems.

7 SUMMARY AND BROADER CHALLENGES AROUND *OCL* RESEARCH

Throughout the survey, we have identified biases integrated into computer systems through their development pipelines and the ways used to tackle those biases. Here, we summarize these biases and reflect at a higher level on the causes of these errors and the issues they reinforce. We identify additional challenges both of technical and structural nature.

7.1 Biases

In Table 5, we summarize the technical biases identified along the survey. These biases often arise from under-defined online conflictual languages in terms of semantic properties and contextual properties or from technical difficulties in accounting for these properties. While the biases arise from different parts of the data and model pipelines, their harmful impact generally stems from the outputs of the machine learning models applied to real use-cases.

7.2 Technical Challenges

7.2.1 Issues Stemming from the Technical Biases. The biases identified resonate with multiple domains of machine learning research, especially unfairness, robustness to natural perturbations and to adversarial attacks, and model failures that come from the distribution mismatch between the training data and the data in deployment. Most issues are ultimately questions of ill-defined requirements. Developing methods to better identify the requirements of the systems prior to their development, and to test for such requirements, would allow to foresee such issues and possibly correct for them [23]. A recent study (not from the *OCL* domain) refers to adjacent problems as underspecification of machine learning models [76], i.e., models trained on the same dataset with the same architecture but various seemingly "unimportant" hyperparameters (e.g., initialization seed) provide similar performance on a test set, but diverging performance on the deployment data.

As for natural perturbations, it remains to be defined what the nature of such perturbations is in the context of *OCL*. In computer vision, natural perturbations are generated artificially on images with prior knowledge of usual transformations of the data samples, and a model is trained and evaluated with the worst-case perturbation or the average perturbation [125]. The equivalent in natural language could be spelling mistakes or intentional misspellings, variations of languages within a sentence, grammatical mistakes, and so on.

As for model failures, identification methods exist especially in computer vision and rely on a human-in-the-loop approach to make sense of data samples and cluster them into meaningful groups [25]. Similarly, designing tasks that crowd workers could perform in large scale for *OCL* needs attention, especially if their subjectivity is taken into account while attributing labels. Besides, a redefinition of model error formalization might be needed to adhere to this subjectivity. For computer vision and tabular data, bias mitigation methods are developed, often transforming the latent representations learned by the models [117] once the biases are identified. These methods could be similarly applied to *OCL* detection.

7.2.2 Other Issues. Similarly to other machine learning-heavy fields, *OCL* detection might be concerned with issues of privacy, explainability, and accountability. Studying them for *OCL* might present new challenges. For instance, concerning explainability, an author might want to know why their text was flagged (local explanation), while a platform user would want to know about the general types of content flagged for them (global explanation). An unintentional author of *OCL* might need indications to express their ideas in a non-problematic way (to the extent this is), which could be inspired from works on recourse in machine learning. Few works answer these challenges in natural language processing.

As for privacy, issues could arise from the need for large datasets or from the use of machine learning models. The sources of the datasets and the way they are stored might raise privacy issues if, for instance, posts are collected from social media users—even though these posts are made public [41]. The annotation activities might also create privacy issues in cases where the data samples contain private information that the data annotator would be exposed to. A model trained on a dataset containing posts from specific individuals might also be "attacked" to identify which individuals were contained in the training set [275].

7.3 Adjacent Challenges in the Field

7.3.1 Adjacent Research. In this survey, we focused on *OCL*s. However, other types of Web content, such as images and memes, require automatic moderation, as they can also be harmful. Only few works have addressed this problem [107, 230].

Counterspeech is a way to answer to *OCL* in an attempt to diminish it, while not reducing freedom of speech [63, 168, 212, 267]. While our survey does not target counterspeech, this is a new trend that merits further investigation. Especially, investigating the psychology of counterspeech to identify the type of language that is the most effective, depending on latent context variables, is a promising research direction.

7.3.2 Handling OCL. *OCL* content can be handled in various ways, with various pros and cons. Besides filtering out the content—which might infringe freedom of expression—or countering it, another recent avenue is to provide a warning to the recipient of *OCL* [271]. This could prevent harm of waiting for verification and removal, while not infringing freedom of expression. Gorwal et al. [118] list additional political issues with content removal, such as the opacity of the procedure, that could be handled by making transparent each decision.

7.3.3 Reproducibility. A lot of papers do not report important figures and methodological information, although these are needed to understand the validity and domain of application of the dataset and to reproduce the results. Inspired from Timnit et al. [112] and Mitchell et al. [172], it could benefit the community to develop a set of specifications on the datasets and machine learning models that should be reported in each research paper.

7.4 Structural Challenges

Many of the technical, contextual, and semantic challenges identified all along the survey find their underlying causes in the ways research and development on *OCL* have been structured. While structural issues are not changed easily, it is worth enumerating some of them.

Disconnection between machine learning and social science research. While setting up interdisciplinary collaborations is difficult, the survey showed research opportunities for each discipline. For instance, while computer science would benefit introducing contextual information from psychology works in datasets and models, psychology research has not yet studied all variations of *OCL* , and computer science tools could facilitate this work [244].

Disconnection between research and real-world scenarios. Datasets often remain large-grain on the context of *OCL* and on the annotations. However, delving into specific *OCL*, possibly engaging with the communities involved, especially with the authors of *OCL* and their targets, would allow to better understand the requirements that a system should verify. Participatory design, recently raising in machine learning works [146], while not being the entire solution [250], would benefit the area of *OCL*and the comprehension of human-aligned requirements. Yet, an obstacle might generally be the stronger interest for algorithmic works than for dataset works in computer science conferences.

Finally, computer science research can benefit from the tradition of social science work that usually begins with the definition of the concepts studied. For instance, psychology researchers who identify the individual and group targets of hate speech point out categories of people with similar socio-demographic attributes (race, religion, disability, sexual orientation, ethnicity, class, gender, behavioral and physical aspects [248], as well as moral [184] and mental status [126]). Clarification as such can help scope the work and avoid conceptual confusions even with disagreement on the definition. Similarly, computer science works on biases and unfairness can benefit from a clear statement about the biases and harms they study. Blodgett et al. [36] provide an extensive review of the study of biases in natural language processing publications and provide recommendations on that end.

8 CONCLUSION

In this work, we used *online conflictual languages* (*OCL*) to refer to the multitude of hate-related languages, and we explained the ones targeted in the survey. We gave an overview of these concepts from a psychology and a computer science point of view. We proposed a unified set of definitions of the *OCL* and of properties that characterize *OCL*, and we organized them into a taxonomy to distinguish them. This is a first attempt to reconcile the literature, but it is not meant to be the final way to characterize the different *OCL*, as further investigation in social science literature might increase precision and formality.

We then proceeded to a systematic survey of the classification methods and dataset collection methods used in computer science. We identified the main trends in the design of these methods and reflected on the main biases that are incorporated into the detection systems by drawing on the new insights from psychology literature and the consideration around the online context. We highlighted numerous implicit biases related to the semantic and contextual nature of many *OCL*, but also simply to the importance of a language's content in its interpretation. The identification of these biases led us to discuss various socio-technical research opportunities for the future and to consider and question the structures that developed these biases within computer science research.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers and the editor of TSC for their in-depth comments that helped us improve the quality of our manuscript.

A APPENDIX

A.1 Clarification of *OCL* Definitions

See Table 6.

Table 6. Definitions of the *OCL* concepts taken from regular (https://dictionary.cambridge.org/), Psychology (https://dictionary.apa.org/) (in italic) and other Social Sciences (http://bitbucket.icaap.org/) (with SoSc) dictionaries

Language	Definition
Offensive	(1) Causing someone to be upset or to have hurt feelings. (2) Offensive can be used more generally to mean unpleasant.
Hateful	(1) Filled with or causing strong dislike. (2) Very unpleasant.
Hate speech	Public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation.
Hate	An extremely strong dislike. *A hostile emotion combining intense feelings of detestation, anger, and often a desire to do harm. Also called hatred.*
Aggression	Spoken or physical behaviour that is threatening or involves harm to someone or something. *Behavior aimed at harming others physically or psychologically.*
Cyberbullying	The activity of using the internet to harm or frighten another person, especially by sending them unpleasant messages. *Cyberbullying is verbally threatening or harassing behavior conducted through such electronic technology as cell phones, e-mail, and text messaging.*
Flaming	The act of sending an angry or insulting email.
Harassment	(1) Behaviour that annoys or upsets someone. (2) Illegal behaviour towards a person that causes mental or emotional suffering, which includes repeated unwanted contacts without a reasonable purpose, insults, threats, touching, or offensive language.
Denigration	Saying that someone or something is not good or important.
Impersonation	(1) The act of intentionally copying another person's characteristics, such as their behavior, speech, appearance, or expressions, especially to make people laugh. (2) The act of attempting to deceive someone by pretending that you are another person. *(1) The deliberate assumption of another person's identity, usually as a means of gaining status or other advantage. (2) The imitation of another person's behavior or mannerisms, which is sometimes done for its corrective or therapeutic effect on one's own behavior (e.g., to gain insight).*
Trickery	The activity of using tricks to deceive or cheat people.
Exclusion	Intentionally not including something.
Flooding	*A technique in behavior therapy in which the individual is exposed directly to a maximum-intensity anxiety-producing situation or stimulus, either described or real, without any attempt made to lessen or avoid anxiety or fear during the exposure.*
Trolling	The act of leaving an insulting message on the internet in order to annoy someone.
Abusive	(1) Using rude and offensive words. (2) Treating someone badly or cruelly. *Interactions in which one person behaves in a cruel, violent, demeaning, or invasive manner toward another person. The term most commonly implies physical mistreatment but also encompasses sexual and psychological (emotional) mistreatment.*
Discrimination	Treating a person or particular group of people differently, especially in a worse way from the way in which you treat other people, because of their skin colour, sexuality, and so on. *Differential treatment of the members of different ethnic, religious, or other groups. Discrimination is usually the behavioral manifestation of prejudice and therefore involves negative, hostile, and injurious treatment of the members of rejected groups. So Sc: The unequal treatment of individuals on the basis of their personal characteristics, which may include age, sex, sexual orientation, ethnic or physical identity. Discrimination usually refers to negative treatment, but discrimination in favour of particular groups can also occur.*
Profanity	1) (An example of) showing no respect for a god or a religion, especially through language. 2) An offensive or obscene word or phrase.
Harmful	Hurting someone or damaging something.

Cyberaggression, outing, cyberstalking, and toxic speech were not associated with relevant definitions in the three dictionaries.

A.2 Reconciled Definitions

See Table 7.

Table 7. Selected Definitions of *OCL*s

Language	Definition
Offensive	Communication which attacks persons on some of their characteristics, most often with rude language. (combination of References [58, 209, 220, 285, 296])
Hateful speech	Speech which contains an expression of hatred on the part of the speaker/author, against a person or people, based on their group identity. [233]
Hate speech	Language used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group. (from Reference [145], similar to References [7, 79, 102, 173, 174, 209, 248, 282, 304])
Aggression	Intention to harm. [81, 148, 209]
Cyberaggression	Online aggressive behavior with intention to harm. [53, 102, 130, 216]
Cyberbullying	Willful and repeated harm inflicted to an individual through the medium of electronic text. [6, 71, 72, 77, 85, 178, 179, 185, 233, 256, 258]
Flaming	Online fights using electronic messages with angry and vulgar language. [247]
Harassment	Repeatedly sending nasty, mean, and insulting messages to intentionally annoy others. [298]
Denigration	Dissing someone online. Sending or posting gossip or rumors about a person to damage his or her reputation or friendships. [247]
Impersonation	Pretending to be someone else and sending or posting material to get that person in trouble or danger or to damage that person's reputation or friendships. [247]
Outing	Sharing someone's secrets or embarrassing information or images online. [247]
Trickery	Talking someone into revealing secrets or embarrassing information or images online. [247]
Exclusion	Intentionally and cruelly excluding someone from an online group. [247]
Cyberstalking	Repeated, intense harassment and denigration that includes threats or creates significant fear. [247]
Flooding	Repeatedly entering the same comment, nonsense comments, or holding down the enter key for the purpose of not allowing the victim to contribute to the conversation. [29]
Trolling (baiting)	Intentionally posting comments that disagree with other posts in the thread for the purpose of provoking a fight, even if the comments don't necessarily reflect the poster's actual opinion. [29]
Abusive	Any strongly impolite, rude or hurtful language using profanity, that can show a debasement of someone or something, or show intense emotion. (Reference [103], close to References [1, 133, 155])
Toxic	Rude, disrespectful, aggressive comment likely to make somebody leave a discussion. [292]
Hate	Expression of hostility without any stated explanation for it. [101]
Discrimination	Process through which a difference is identified and then used as the basis of unfair treatment. [101]
Profanity	Offensive or obscene word or phrase. [101]
Harmful	Text which has a negative effect on somebody. (proposed based on dictionaries)

A.3 Detailed Overview of *OCL*s Individual Characteristics

See Table 8.

Table 8. Analysis of the *OCL*

Concept	Intention		Behavior	Specific focus		Emotion (hatred)	Language	Target	Effect
	Hatred	***Harm***		***Other***	***Character.***				
offensive	N	N	N	N	Y	N	N	Y ((type) person)	Y
hateful speech	N	N	N	N	Y (stereo)	Y	N	Y (person, group)	N
hate speech	Y	N	N	N	Y (stereo)	Y	N	Y (person, group)	N
aggression	N	Y	N	N	N	N	N	Y	N
cyberaggression	N	Y	Y	N	N	N	Y	N	
cyberbullying	N	Y	Y (repetitive aggression)	N	N	N	N (often aggressive)	Y (power imbalance)	N
flaming	N	Y	Y (fight)	N	N	N	Y (abusive)	Y	N
harassment	N	Y	Y (repetition)	Y (offense target)	N	N	Y (abusive)	Y	Y (annoy)
denigration	N	Y	Y (damage reputation)	Y (gossip, rumor)	N	N	N	Y	N
impersonation	N	Y	Y (pretend to be the target)	N	N	N	N	Y	N
outing	N	Y	Y (sharing target's secret)	Y (secret)	N	N	N	Y	N
trickery	N	Y	Y (forced info. sharing)	Y (private info.)	N	N	N	Y	N
exclusion	N	Y	Y	N	N	N	N	Y	Y (exclusion)
cyberstalking	N	Y	Y (repeated)	Y (offense target)	N	N	N	Y	Y (threat, fear)
flooding	N	Y	Y (repetitive posting behavior)	N	N	N	N	Y	Y (exclusion)
trolling	N	Y	Y (disagreeing posts)	Y (disagree with posts)	N	N	N	Y	N
abusive	N	N	N	N	N	N	Y	N	N
toxic	N	N	N	N	N	N	Y (disre-spectful)	N	Y (leave the discussion)
hate	Y	N	N	N	Y	Y	N	Y	N
discrimination	N	Y	Y	N	Y (difference)	N	N	Y	Y (unfair treatment)
profanity	N	N	N	N	N	N	Y (rude)	N	Y (offense)
harmful	N	N	N	N	N	N	N	N	Y

The different *OCLs* are classified along the dimensions of studies identified earlier. They are later clustered (colors) in groups that share similar characteristics, to organize them into a taxonomy. "Stereo" refers to stereotype. ("Y" stands for "yes" and "N" for "no.")

A.4 Adjacency Matrix Illustrating the Confusions of the Different *OCL* According to our Reconciliation

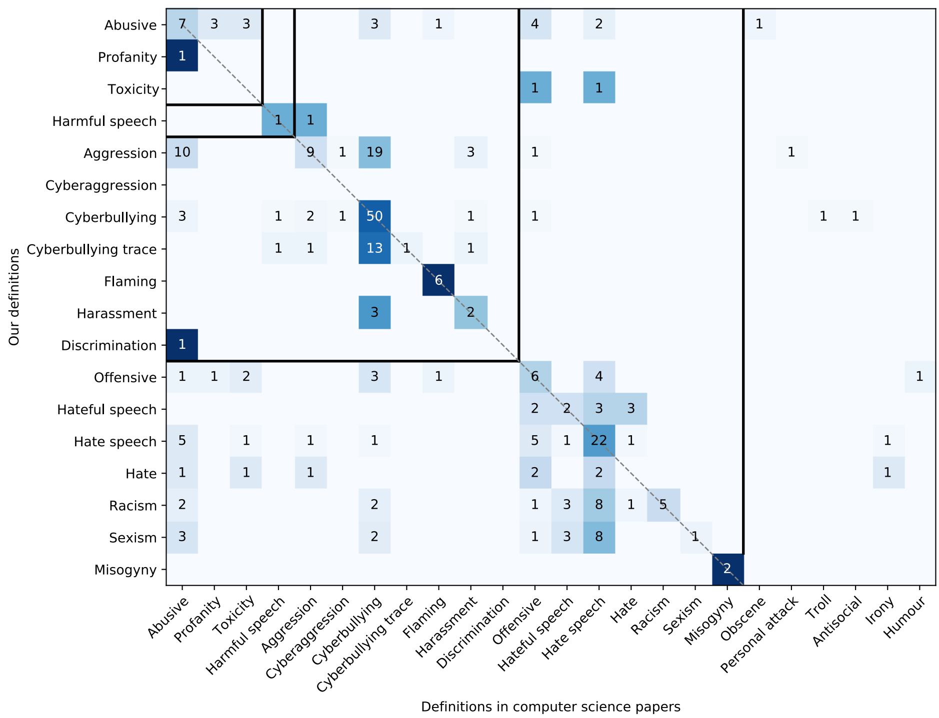


Fig. 12. Adjacency matrix of the different *OCL*s terms in the computer science (CS) literature. Computer science papers are counted based on the initial definition of the *OCL*s term they use—vertically—and on the new term associated to their definition through our taxonomy—horizontally. The colors visualize the distribution of papers that would be classified as a certain concept with our definitions and that were classified with a (same or different) concept in computer science papers (i.e., integrating the values in a row adds up to 1, the entire distribution). The darker the color is, the higher the percentage of papers that fit one concept and are denoted by one concept is. We see that out of the 219 papers we reviewed, hate speech is used in 50 (23%) of them; yet only 22 (44%) of these 50 papers actually address the problem of hate speech.

A.5 Summary of Common Features

Here, we list the features that are often employed for *OCL* detection.

- N-gram features: Word and character n-grams are mostly used to encode the samples, but other n-grams can also be used such as skip-gram, lemma n-grams and lemma sentiment polarity n-grams, dependency type n-grams, and repetitions of n-grams [82].
- BoW: BoW are encoded with binary elements, with the tf-idf score or the frequency of the words in a group of text data. They are sometimes extended for example by using word embeddings to find words close to the initial set of words in the embedding space and add them to the BoW model (embedding enhanced BoW [307]).
- Embeddings: They are mostly used for Deep Learning and are usually learned while training the neural networks or pre-trained. These are mainly Word2vec, GloVe, FastText [272] with sub-words embeddings and **continuous BoW (CBOW)** for word and character

levels, paragraph2vec for paragraph-level [92]. Certain publications investigate different initializations of the word embeddings to train, such as random initialization, GloVe or Sentiment Specific Word Embedding [4]. Zhang et al. [302] use a phonetic representation of words to avoid misspellings and learn their embedding during the training process. Machine learning models also encode sentences with word embeddings averaged over the whole sentence [55, 182]. Hasanuzzaman et al. [124] investigate the combination of word embedding with demographic information. More recently, new types of embeddings which incorporate context within the embeddings of the words, specifically ELMo [203]- and BERT [84]-based embeddings, have proven to allow for the best performance in various *OCL*-related challenges such as the task 5 at SemEval-2019 [39] on hate speech detection, the TRAC shared task on aggression identification and gendered aggression identification [147] at LREC 2020, and the task 12 at SemEval-2020 [75] on offensive language detection.

- Lexical features: They reflect the vocabulary of the samples. Words representative of *OCL* in the samples are identified and possibly counted, often by matching each word to a dictionary of words such as negative and hate [38], offensive [81], blacklisted [155], or swear words with **Linguistic Inquiry and Word Count (LIWC)** [99].
- Linguistic features: They reflect the construction of the sentences and words. These can be the count of specific punctuation marks such as question [38] and exclamation marks, the number of uncommon capital letters [99], the data sample length [81] or the length of the longest word, the average and median word lengths, the number of long words [12], or characteristics potentially indicative of *OCL*, such as the number of abbreviations and of words using special characters [155], the number of smileys [132], the number of hashtags, Flesch-Kincaid Grade Level and Flesch Reading Ease scores to measure the readability of a document [122], word similarity with a training set [246].
- Sentiment analysis: Various methods are used to identify the sentiment of words or sentences (possibly by averaging the sentiment of each word [274]) and to compute a sentiment score or a binary value, such as matching words to a sentiment dictionary [81] or using sentiment analysis tools. Certain papers also encode emotions [99] as valence, dominance, and arousal scores [310] or the tone of the samples [2].
- **Part-of-speech (POS)**, typed dependencies: POS tagging or n-gram [12] and typed dependencies [46, 99], often based on the Stanford Dependency Parser, are used to encode grammatical relations between words, as it is assumed they characterize *OCL*.
- Pronoun variations: The use of pronouns is related to the use of *OCL* that often target people. Certain papers identify or count the number of occurrences of the second pronoun [81], while others take into account all the pronouns or only the ones associated with a negative noun or a noun from a specific dictionary [185] or profanity windows (association of a pronoun and a profane word) [74].
- Topic model: Topics are retrieved by topic extraction, mostly **Latent Dirichlet Allocation (LDA)** [91, 308] and **Latent Semantic Analysis (LSA)** [307], but also using text summarization with a centroid-based method [160].
- Subjectivity analysis: a few papers investigate whether the data samples are subjective, because they assume that subjectivity is a sign for *OCL* [273].

A.6 Summary of Common Classification Algorithms

Rule-based classification. The design of rules is often done in two steps. Dictionaries expressing *OCL* are prepared [11, 224, 296] (e.g., subjectivity lexicon, hate lexicon, and list of hate-representative grammatical relations [83], lists of profane words augmented with genomics-inspired techniques [225]). Then, lists of rules are defined to attribute a score to samples based on their use of the dictionary vocabulary, often using pattern or word matching. The patterns to match with the samples are defined manually or automatically [42, 248]. The score enables the final classification of the samples. For example, Yadav et al. [297] use the AHO-Corasick String Pattern Matching algorithm to find the words in the sentences contained in a dictionary of offensive words, while Bayzick et al. [29] check for the existence of words from a cyberbullying dictionary and the presence of a second person pronoun.

Machine learning. The most used algorithm is **Support Vector Machine (SVM)** with several variants. Most papers use non-linear kernels when dealing with complex tasks, possibly with cost-sensitive SVM to circumvent dataset imbalance [99, 160, 255, 273, 302]. Experiments on the design of the classifiers obtained diverging results that merit being investigated. E.g., Warner et al. [282] suggest to use distinct SVM for different categories of hate speech which use stereotypes of distinct lexical fields, since it should be an easier learning task. This was tested by Dinakar et al. [89], which shows that SVM trained on topic-specific datasets achieves higher performance than SVM trained on the whole dataset for detecting three cyberbullying topics (sexuality, race and culture, and intelligence). However, Sood et al. [255] found out that classifying insults in a general domain or training separate SVM for different categories of comments (politics, news, entertainment, business, world) might not change performance. The results might depend on the categories, certain employing more specific language than others.

The **Naive Bayes (NB)** classifier is used in its original version or variants such as Bernoulli or multinomial NB [97], Complement or Multinomial Updatable NB, or Decision Table NB [220]. Wulczyn et al. [292] use a Multi-Layer Perceptron and Logistic Regression in the only paper addressing the subjectivity of judgments. They show that comments with high annotator agreement are different from the ones with lower agreement and that empirical distributions better represent the labels.

Deep learning. Convolutional Neural Networks, autoencoders [213], **Recurrent Neural Networks (RNN)** or the variants Long-Short Term Memory and Gated Recurrent Units with or without attention and uni or bi-directional are trained or several networks combined. Few papers experiment with other methods (RNN and reinforcement learning [214], unsupervised deep learning like growing hierarchical self-organizing map [85]). The deep learning methods are claimed to achieve better performance than traditional machine learning.

REFERENCES

[1] E. A. Abozinadah and J. H. Jones. 2017. A statistical learning approach to detect abusive Twitter accounts. In *Proceedings of the International Conference on Compute and Data Analysis*. ACM Press, New York. https://doi.org/10.1145/3093241.3093281

[2] S. Agarwal and A. Sureka. 2016. But I did not mean it! —Intent classification of racist posts on Tumblr. In *Proceedings of the European Intelligence and Security Informatics Conference (EISIC)*. IEEE. https://doi.org/10.1109/EISIC.2016.032

[3] Swati Agarwal and Ashish Sureka. 2016. But I did not mean it!—Intent classification of racist posts on Tumblr. In *Proceedings of the European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 124–127.

[4] Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26–29, 2018, Proceedings (Lecture Notes in Computer Science)*, Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury (Eds.), Vol. 10772. Springer, 141–153. https://doi.org/10.1007/978-3-319-76941-7_11

[5] Areej Al-Hassan and Hmood Al-Dossari. 2019. Detection of hate speech in social networks: A survey on multilingual corpus. In *Proceedings of the 6th International Conference on Computer Science and Information Technology*.

[6] A. H. Alduailej and M. B. Khan. 2017. The challenge of cyberbullying and its automatic detection in Arabic text. In *Proceedings of the International Conference on Computer and Applications (ICCA)*. IEEE.

[7] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata. 2017. Hate speech detection in the Indonesian language: A dataset and preliminary study. In *Proceedings of the International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE, 233–238. https://doi.org/10.1109/ICACSIS.2017.8355039

[8] Ashraf Alhujailli and Waldemar Karwowski. 2018. Emotional and stress responses to cyberbullying. In *Proceedings of the International Conference on Applied Human Factors and Ergonomics*. Springer, 33–43.

[9] Wafa Alorainy, Pete Burnap, Han Liu, and Matthew L. Williams. 2019. "The enemy among us" Detecting cyber hate speech with threats-based othering language embeddings. *ACM Trans. Web* 13, 3 (2019), 1–26.

[10] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software engineering for machine learning: A case study. In *Proceedings of the IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 291–300.

[11] S. Ando, Y. Fujii, and T. Ito. 2010. Filtering harmful sentences based on multiple word co-occurrence. In *Proceedings of the IEEE/ACIS 9th International Conference on Computer and Information Science*. IEEE. https://doi.org/10.1109/ICIS.2010.96

[12] Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on Twitter. In *Proceedings of the International Conference on Applications of Natural Language to Information Systems*. Springer, 57–64.

[13] M. E. Aragón and A. P. López-Monroy. 2018. Author profiling and aggressiveness detection in Spanish tweets: MEX-A3T 2018. In *Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval)*.

[14] Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 45–54.

[15] D. Archard. 2014. Insults, free speech and offensiveness. *J. Appl. Philos.* 31, 2 (2014).

[16] John Archer and Sarah M. Coyne. 2005. An integrated review of indirect, relational, and social aggression. *Personal. Soc. Psychol. Rev.* 9, 3 (2005), 212–230.

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

[18] Stavros Assimakopoulos, Fabienne H. Baider, and Sharon Millar. 2017. *Online Hate Speech in the European Union: A Discourse-analytic Perspective*. Springer Nature.

[19] Adithya Avvaru, Sanath Vobilisetty, and Radhika Mamidi. 2020. Detecting sarcasm in conversation context using transformer-based models. In *Proceedings of the 2nd Workshop on Figurative Language Processing*. 98–103.

[20] Dario Bacchini, Giovanna Esposito, and Gaetana Affuso. 2009. Social experience and school bullying. *J. Communit. Appl. Soc. Psychol.* 19, 1 (2009), 17–32.

[21] Pinkesh Badjatiya, M. Gupta, and V. Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *Proceedings of the World Wide Web Conference*. 49–59.

[22] A. Balayn, P. Mavridis, A. Bozzon, B. Timmermans, and Z. Szlávik. 2018. Characterising and mitigating aggregation-bias in crowdsourced toxicity annotations. In *Proceedings of the 1st Workshop on Disentangling the Relation between Crowdsourcing and Bias Management*. CEUR.

[23] Agathe Balayn, Panagiotis Soilis, Christoph Lofi, Jie Yang, and Alessandro Bozzon. 2021. What do you mean? Interpreting image classification with crowdsourced concept extraction and analysis. In *Proceedings of the World Wide Web Conference (WWW)*. Association for Computing Machinery, New York, NY, 1937–1948. https://doi.org/10.1145/3442381.3450069

[24] Anna C. Baldry and David P. Farrington. 2000. Bullies and delinquents: Personal characteristics and parental styles. *J. Communit. Appl. Soc. Psychol.* 10, 1 (2000), 17–31.

[25] Gagan Bansal and Daniel S. Weld. 2018. A coverage-based utility model for identifying unknown unknowns. In *Proceedings of the AAAI Conference*.

[26] Srijan Bansal, Vishal Garimella, Ayush Suhane, Jasabanta Patro, and Animesh Mukherjee. 2020. Code-switching patterns can be an effective route to improve performance of downstream NLP applications: A case study of humour, sarcasm and hate speech detection. *arXiv preprint arXiv:2005.02295* (2020).

[27] Natã M. Barbosa and Monchu Chen. 2019. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–12.

[28] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 54–63.

[29] J. Bayzick. 2011. Detecting the presence of cyberbullying using computer software. In *Proceedings of the 3rd International Web Science Conference*. 1–2.

[30] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).

[31] A. Bellmore, J. Calvin, J.-M. Xu, and X. Zhu. 2015. The five W's of "bullying" on Twitter: Who, what, why, where, and when. *Comput. Hum. Behav.* 44 (Mar. 2015). https://doi.org/10.1016/J.CHB.2014.11.052
[32] H. Berghel and D. Berleant. 2018. The online trolling ecosystem. *IEEE Comput.* 51, 8 (2018). https://doi.org/10.1109/MC.2018.3191256
[33] Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. 2020. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159* (2020).
[34] R. Binns, M. Veale, M. Van Kleek, and N. Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *Proceedings of the International Conference on Social Informatics*. Springer.
[35] Erik Bleich. 2011. The rise of hate speech and hate crime laws in liberal democracies. *J. Ethn. Migrat. Stud.* 37, 6 (2011), 917–934. https://doi.org/10.1080/1369183X.2011.576195 arXiv:https://doi.org/10.1080/1369183X.2011.576195
[36] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. *arXiv preprint arXiv:2005.14050* (2020).
[37] R. J. Boeckmann and J. Liew. 2002. Hate speech: Asian American students' justice judgments and psychological responses. *J. Soc. Iss.* 58, 2 (2002), 363–381.
[38] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the 2nd Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*. Association for Computational Linguistics, 36–41. https://doi.org/10.18653/v1/W18-1105
[39] Michal Bojkovský and Matúš Pikuliak. 2019. STUFIIT at SemEval-2019 task 5: Multilingual hate speech detection on Twitter with MUSE and ELMo embeddings. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 464–468.
[40] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 4349–4357.
[41] Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Inf., Commun. Societ.* 15, 5 (2012), 662–679.
[42] Uwe Bretschneider and Ralf Peters. 2017. Detecting offensive statements towards foreigners in social media. In *50th Hawaii International Conference on System Sciences, HICSS 2017, Hilton Waikoloa Village, Hawaii, USA, January 4-7, 2017*, Tung Bui (Ed.). AIS Electronic Library (AISeL), 1–10. http://hdl.handle.net/10125/41423.
[43] Marc-Etienne Brunet, C. Alkalay-Houlihan, A. Anderson, and R. Zemel. 2019. Understanding the origins of bias in word embeddings. In *Proceedings of the International Conference on Machine Learning*. 803–811.
[44] Victoria K. Burbank. 1994. Cross-cultural perspectives on aggression in women and girls: An introduction. *Sex Roles* 30, 3–4 (1994), 169–176.
[45] P. Burnap and M. L. Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making. *Internet, Polic. Polit.* (2014).
[46] P. Burnap and M. L. Williams. 2015. Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Polic. Internet* 7 (2015). https://doi.org/10.1002/poi3.85
[47] P. Burnap and M. L. Williams. 2016. Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Sci.* 5, 1 (Dec. 2016), 11. https://doi.org/10.1140/epjds/s13688-016-0072-6
[48] C. Chelmis, D.-S. Zois, and M. Yao. 2017. Mining patterns of cyberbullying on Twitter. In *Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 126–133. https://doi.org/10.1109/ICDMW.2017.22
[49] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.
[50] Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (Eds.). 2018. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA). http://www.lrec-conf.org/lrec2018.
[51] Sergio Andrés Castaño-Pulgarín, Natalia Suárez-Betancur, Luz Magnolia Tilano Vega, and Harvey Mauricio Herrera López. 2021. Internet, social media and online hate speech. Systematic review. *Aggress. Viol. Behav.* 58 (2021), 101608. https://prohic.nl/wp-content/uploads/2021/05/213-17mei2021-InternetOnlineHateSpeechtSystematicReview.pdf
[52] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali. 2017. Measuring #GamerGate: A tale of hate, sexism, and bullying. In *Proceedings of the 26th International Conference on World Wide Web*. ACM Press, New York, 1285–1290. https://doi.org/10.1145/3041021.3053890
[53] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali. 2017. Mean birds: Detecting aggression and bullying on Twitter. In *Proceedings of the ACM on Web Science Conference (WebSci)*. ACM, New York, NY, 13–22. https://doi.org/10.1145/3091478.3091487

[54] V. S. Chavan and S. S. Shylaja. 2015. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2354–2358. https://doi.org/10.1109/ICACCI.2015.7275970

[55] Hao Chen, Susan McKeever, and Sarah Jane Delany. 2017. Abusive text detection using neural networks. In *Proceedings of the 25th Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Ireland, December 7–8, 2017 (CEUR Workshop Proceedings)*, John McAuley and Susan McKeever (Eds.), Vol. 2086. CEUR-WS.org, 258–260. http://ceur-ws.org/Vol-2086/AICS2017_paper_44.pdf.

[56] H. Chen, S. Mckeever, and S. J. Delany. 2017. Presenting a labelled dataset for real-time detection of abusive user posts. In *Proceedings of the International Conference on Web Intelligence (WI)*. ACM, New York, NY, 884–890. https://doi.org/10.1145/3106426.3106456

[57] Hao Chen, Susan McKeever, and Sarah Jane Delany. 2018. A comparison of classical versus deep learning techniques for abusive content detection on social media sites. In *Proceedings of the International Conference on Social Informatics*. Springer, 117–133.

[58] Y. Chen, Y. Zhou, S. Zhu, and H. Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of the International Conference on Privacy, Security, Risk and Trust (PASSAT), and International Conference on Social Computing (SocialCom)*. IEEE, 71–80.

[59] Lu Cheng, Ruocheng Guo, and Huan Liu. 2019. Robust cyberbullying detection with causal interpretation. In *Proceedings of the World Wide Web Conference*. 169–175.

[60] Naganna Chetty and Sreejith Alathur. 2018. Hate speech review in the context of online social networks. *Aggress. Viol. Behav.* 40 (2018), 108–118.

[61] Shivang Chopra, Ramit Sawhney, Puneet Mathur, and Rajiv Ratn Shah. 2020. Hindi-English hate speech detection: Author profiling, debiasing, and practical perspectives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 386–393.

[62] Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J. Jansen, and Joni Salminen. 2020. A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 6203–6212.

[63] Yi-Ling Chung, E. Kuzmenko, S. S. Tekiroglu, and M. Guerini. 2019. CONAN-counter narratives through nichesourcing: A multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2819–2829.

[64] Sebastián Correa and Alberto Martin. 2018. Linguistic generalization of slang used in Mexican tweets, applied in aggressiveness detection. In *Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018 (CEUR Workshop Proceedings)*, Paolo Rosso, Julio Gonzalo, Raquel Martínez, Soto Montalvo, and Jorge Carrillo de Albornoz (Eds.), Vol. 2150. CEUR-WS.org, 119–127. http://ceur-ws.org/Vol-2150/MEX-A3T_paper5.pdf.

[65] K. Cortis and S. Handschuh. 2015. Analysis of cyberbullying tweets in trending world events. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*. ACM Press, New York, 1–8. https://doi.org/10.1145/2809563.2809605

[66] G. Cowan and C. Hodge. 1996. Judgments of hate speech: The effects of target group, publicness, and behavioral responses of the target. *J. Appl. Soc. Psychol.* 26, 4 (1996), 355–374.

[67] G. Cowan and D. Khatchadourian. 2003. Empathy, ways of knowing, and interdependence as mediators of gender differences in attitudes toward hate speech and freedom of speech. *Psychol. Wom. Quart.* 27, 4 (2003), 300–308.

[68] G. Cowan and J. Mettrick. 2002. The effects of target variables and setting on perceptions of hate speech. *J. Appl. Soc. Psychol.* 32, 2 (2002), 277–299.

[69] Kate Crawford and Trevor Paglen. 2019. Excavating AI: The politics of images in machine learning training sets. *Excavating AI* (2019), 1–12.

[70] G. B. Cunningham, M. Ferreira, and J. S. Fink. 2009. Reactions to prejudicial statements: The influence of statement content and characteristics of the commenter. *Group Dynam.: Theor., Res. Pract.* 13, 1 (2009), 59.

[71] M. Dadvar, F. M. G. de Jong, R. Ordelman, and D. Trieschnigg. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the 12th Dutch-Belgian Information Retrieval Workshop (DIR)*.

[72] Maral Dadvar, R. Ordelman, F. de Jong, and D. Trieschnigg. 2012. Towards user modelling in the combat against cyberbullying. In *Proceedings of the International Conference on Application of Natural Language to Information Systems*. Springer, 277–283.

[73] M. Dadvar, D. Trieschnigg, and F. de Jong. 2014. Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Proceedings of the Canadian Conference on Artificial Intelligence*. Springer, 275–281.

[74] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong. 2013. Improving cyberbullying detection with user context. In *Proceedings of the European Conference on Information Retrieval*. Springer, 693–696.

[75] Wenliang Dai, Tiezheng Yu, Zihan Liu, and Pascale Fung. 2020. Kungfupanda at SemEval-2020 Task 12: BERT-based multi-task learning for offensive language detection. *arXiv* (2020), arXiv–2004.

[76] Alexander D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395* (2020).

[77] Harsh Dani, Jundong Li, and Huan Liu. 2017. Sentiment informed cyberbullying detection in social media. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer, 52–67.

[78] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the 3rd Workshop on Abusive Language Online.* 25–35.

[79] Ona de Gibert, N. Perez, A. García-Pablos, and M. Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW).* 11–20.

[80] T. De Smedt, G. De Pauw, and P. Van Ostaeyen. 2018. Automatic detection of online jihadist hate speech. *Computational Linguistics & Psycholinguistics Technical Report Series, CTRS-007, FEBRUARY 2018.* university of Antwerpen

[81] Laura P. Del Bosque and Sara Elena Garza. 2014. Aggressive text detection for cyberbullying. In *Proceedings of the Mexican International Conference on Artificial Intelligence.* Springer, 221–232.

[82] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the 1st Italian Conference on Cybersecurity (ITASEC).* 86–95.

[83] N. Dennis-Gitari, Z. Zuping, H. Damien, and J. Long. 2015. A lexicon-based approach for hate speech detection. *Int. J. Multimed. Ubiq. Eng.* 10, 4 (2015). https://doi.org/10.14257/ijmue.2015.10.4.21

[84] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* 4171–4186.

[85] M. Di Capua, E. Di Nardo, and A. Petrosino. 2016. Unsupervised cyber bullying detection in social networks. In *Proceedings of the 23rd International Conference on Pattern Recognition (ICPR).* IEEE, 432–437. https://doi.org/10.1109/ICPR.2016.7899672

[86] K. R. Dickson. 2012. All prejudices are not created equal: Different responses to subtle versus blatant expressions of prejudice. *Electronic Thesis and Dissertation Repository.* Retrieved from *https://ir.lib.uwo.ca/etd/704*.

[87] Edward Dillon, Jamie Macbeth, Robin Kowalski, Elizabeth Whittaker, and Juan E. Gilbert. 2016. "Is this cyberbullying or not?": Intertwining computational detection with human perception (a case study). In *Advances in Human Factors in Cybersecurity.* Springer, 337–345.

[88] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Trans. Interact. Intell. Syst.* 2, 3 (Sep. 2012), 1–30. https://doi.org/10.1145/2362394.2362400

[89] K. Dinakar, R. Reichart, and H. Lieberman. 2011. Modeling the detection of textual cyberbullying. *Soc. Mob. Web* 11, 02 (2011). http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/3841/4384.

[90] Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build It Break It Fix It for Dialogue Safety: Robustness from Adversarial Human Attack. arXiv:cs.CL/1908.06083.

[91] Vinutha H. Divyashree and N. S. Deepashree. 2016. An effective approach for cyberbullying detection and avoidance. *Int. J. Innov. Res. Comput. Commun. Eng.* 14 (2016).

[92] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web.* ACM Press, New York, 29–30. https://doi.org/10.1145/2740908.2742760

[93] D. M. Downs and G. Cowan. 2012. Predicting the importance of freedom of speech and the perceived harm of hate speech. *J. Appl. Soc. Psychol.* 42, 6 (2012), 1353–1375.

[94] Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2019. At the lower end of language—Exploring the vulgar and obscene side of German. In *Proceedings of the 3rd Workshop on Abusive Language Online.* 119–128.

[95] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. *arXiv preprint arXiv:1804.04257* (2018).

[96] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to peer hate: Hate speech instigators and their targets. *arXiv preprint arXiv:1804.04649* (2018).

[97] S. C. Eshan and M. S. Hasan. 2017. An application of machine learning to detect abusive Bengali text. In *Proceedings of the 20th International Conference of Computer and Information Technology (ICCIT).* IEEE. https://doi.org/10.1109/ICCITECHN.2017.8281787

[98] Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Appl. Sci.* 10, 12 (2020), 4180.

[99] Y. J. Foong and M. Oussalah. 2017. Cyberbullying system detection and analysis. In *Proceedings of the European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 40–46. https://doi.org/10.1109/EISIC.2017.43

[100] Paolo Fornacciari, Monica Mordonini, Agostino Poggi, Laura Sani, and Michele Tomaiuolo. 2018. A holistic system for troll detection on Twitter. *Comput. Hum. Behav.* 89 (2018), 258–268. https://doi.org/10.1016/j.chb.2018.08.008

[101] P. Fortuna and S. Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.* 51, 4 (July 2018). https://doi.org/10.1145/3232676

[102] Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM Conference on Web Science*. ACM, 105–114.

[103] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*.

[104] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 329–338.

[105] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering Online Hate Speech*. Unesco Publishing.

[106] B. Gambäck and U. K. Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the 1st Workshop on Abusive Language Online*. 85–90.

[107] Shreyansh Gandhi, Samrat Kokkula, Abon Chaudhuri, Alessandro Magnani, Theban Stanley, Behzad Ahmadi, Venkatesh Kandaswamy, Omer Ovenc, and Shie Mannor. 2020. Scalable detection of offensive and non-compliant content/logo in product images. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. 2247–2256.

[108] L. Gao and R. Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*.

[109] L. Gao, A. Kuppersmith, and R. Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, Vol. 1. 774–782.

[110] Álvaro García-Recuero, Jeffrey Burdges, and Christian Grothoff. 2016. Privacy-preserving abuse detection in future decentralised online social networks. In *Data Privacy Management and Security Assurance*. Springer, 78–93.

[111] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Nat. Acad. Sci.* 115, 16 (2018), E3635–E3644.

[112] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

[113] S. V. Georgakopoulos, S. K. Tasoulis, A. G. Vrahatis, and V. P. Plagianakos. 2018. Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*. ACM Press, New York, 1–6. https://doi.org/10.1145/3200947.3208069

[114] Amy Herstein Gervasio and Katy Ruckdeschel. 1992. College students' judgments of verbal sexual harassment. *J. Appl. Soc. Psychol.* 22, 3 (1992), 190–211.

[115] Fabio Giglietto and Yenn Lee. 2015. To be or not to be Charlie: Twitter hashtags as a discourse and counter-discourse in the aftermath of the 2015 Charlie Hebdo shooting in France. In *Proceedings of the 5th Workshop on Making Sense of Microposts co-located with the 24th International World Wide Web Conference*. 33–37.

[116] Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, et al. 2017. A large labeled corpus for online harassment research. In *Proceedings of the ACM on Web Science Conference*. 229–233.

[117] Sixue Gong, Xiaoming Liu, and Anil K. Jain. 2020. Jointly de-biasing face recognition and demographic attribute estimation. In *Proceedings of the European Conference on Computer Vision*. Springer, 330–347.

[118] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data Societ.* 7, 1 (2020), 2053951719897945.

[119] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is "love" evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. 2–12.

[120] J. Guberman and L. Hemphill. 2017. Challenges in modifying existing scales for detecting harassment in individual tweets. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

[121] Amos Guiora and Elizabeth A. Park. 2017. Hate speech on social media. *Philosophia* 45, 3 (2017), 957–971.

[122] I. Guy and B. Shapira. 2018. From royals to vegans: Characterizing question trolling on a community question answering website. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM Press, New York, 835–844. DOI: https://doi.org/10.1145/3209978.3210058

[123] Bushr Haddad, Zoher Orabe, Anas Al-Abood, and Nada Ghneim. 2020. Arabic offensive language detection with attention-based deep neural networks. In *Proceedings of the 4th Workshop on Open-source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*.

[124] M. Hasanuzzaman, G. Dias, and A. Way. 2017. Demographic word embeddings for racism detection on Twitter. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*. 926–936. https://aclanthology.info/papers/I17-1093/i17-1093.

[125] Dan Hendrycks and Thomas Dietterich. 2018. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the International Conference on Learning Representations*.

[126] P. J. Henry, S. E. Butler, and M. J. Brandt. 2014. The influence of target group status on the perception of the offensiveness of group-based slurs. *J. Experim. Soc. Psychol.* 53 (2014).

[127] Sarah Hewitt, Thanassis Tiropanis, and C. Bokhove. 2016. The problem of identifying misogynist language on Twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science, WebSci 2016, Hanover, Germany, May 22–25, 2016*, Wolfgang Nejdl, Wendy Hall, Paolo Parigi, and Steffen Staab (Eds.). ACM, 333–335. https://doi.org/10.1145/2908131.2908183

[128] Derek Hoiem, Santosh K. Divvala, and James H. Hays. 2009. Pascal VOC 2008 challenge. In *Proceedings of the PASCAL Challenge Workshop in ECCV*. Citeseer.

[129] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing labeled cyberbullying incidents on the Instagram social network. In *Proceedings of the International Conference on Social Informatics*. Springer, 49–66.

[130] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra. 2015. Detection of cyberbullying incidents on the Instagram social network. (Mar. 2015). arXiv:1503.03909. http://arxiv.org/abs/1503.03909.

[131] H. Hosseinmardi, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra. 2016. Prediction of cyberbullying incidents in a media-based social network. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 186–192. https://doi.org/10.1109/ASONAM.2016.7752233

[132] Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially Aware Multimedia*. ACM, 3–6.

[133] M. O. Ibrohim and I. Budi. 2018. A dataset and preliminaries study for abusive language detection in Indonesian social media. *Procedia Comput. Sci.* 135 (2018), 222–229. https://doi.org/10.1016/j.procs.2018.08.169

[134] I. Iglezakis. 2017. The legal regulation of hate speech on the internet. In *EU Internet Law*. Springer.

[135] A. Ioannou, J. Blackburn, G. Siringhini, E. De Chrisiofaro, N. Kouriellis, M. Sirivianos, and P. Zaphiris. 2017. From risk factors to detection and intervention: A metareview and practical proposal for research on cyberbullying. In *Proceedings of the IST-Africa Week Conference (IST-Africa)*. IEEE, 1–8.

[136] A. Jha and R. Mamidi. 2017. When does a compliment become sexist? Analysis and classification of ambivalent sexism using Twitter data. In *Proceedings of the 2nd Workshop on NLP and Computational Social Science*. Association for Computational Linguistics, 7–16. https://doi.org/10.18653/v1/W17-2902

[137] David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 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*. 3658–3666.

[138] V. K. Singh, Q. Huang, and P. K. Atrey. 2016. Cyberbullying detection using probabilistic socio-textual information fusion. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 884–887. https://doi.org/10.1109/ASONAM.2016.7752342

[139] Raghav Kapoor, Yaman Kumar, Kshitij Rajput, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. 2019. Mind your language: Abuse and offense detection for code-switched languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9951–9952.

[140] Brendan Kennedy, X. Jin, A. M. Davani, M. Dehghani, and X. Ren. 2020. Contextualizing hate speech classifiers with post hoc explanation. *arXiv preprint arXiv:2005.02439* (2020).

[141] Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets. *arXiv preprint arXiv:2005.05921* (2020).

[142] Kate Klonick. 2018. The new governors: The people, rules and processes governing online speech. *Harv. Law Rev.* 131 (2018), 1598.

[143] F. Klubička and R. Fernández. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. In *Proceedings of the 4REAL Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language*.

[144] Barbara Krahé. 2013. *The Social Psychology of Aggression*. Psychology Press.

[145] Rohan Kshirsagar, Tyus Cukuvac, Kathy McKeown, and Susan McGregor. 2018. Predictive embeddings for hate speech detection on Twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online, ALW@EMNLP 2018, Brussels, Belgium, October 31, 2018.* Darja Fiser, Ruihong Huang, Vinodkumar Prabhakaran, Rob Voigt, Zeerak Waseem, and Jacqueline Wernimont (Eds.). Association for Computational Linguistics. 26–32. https://doi.org/10.18653/v1/w18-5104

[146] Bogdan Kulynych, David Madras, Smitha Milli, Inioluwa Deborah Raji, Angela Zhou, and Richard Zemel. 2020. Participatory approaches to machine learning. In *Proceedings of the International Conference on Machine Learning Workshop.*

[147] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of the 2nd Workshop on Trolling, Aggression and Cyberbullying*. 1–5.

[148] Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the 11th International Conference on Language Resources and Evaluation.* European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2018/summaries/861.html.

[149] H. Kwak, J. Blackburn, and S. Han. 2015. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* ACM, 3739–3748.

[150] Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence.* AAAI Press, 1621–1622.

[151] Jennifer L. Lambe. 2004. Who wants to censor pornography and hate speech? *Mass Commun. Societ.* 7, 3 (2004), 279–299.

[152] J. Langham and K. Gosha. 2018. The classification of aggressive dialogue in social media platforms. In *Proceedings of the ACM SIGMIS Conference on Computers and People Research.* ACM.

[153] Kyle Langvardt. 2018. Regulating online content moderation. *Georget. Law J.* 106, 5 (2018), 1353–1389.

[154] Issie Lapowsky. 2018. Mark Zuckerberg and the Tale of Two Hearings. Retrieved from https://www.wired.com/story/mark-zuckerberg-congress-day-two/.

[155] H.-S. Lee, H.-R. Lee, J.-U. Park, and Y.-S. Han. 2018. An abusive text detection system based on enhanced abusive and non-abusive word lists. *Decis. Supp. Syst.* 113 (2018), 22–31.

[156] Wonhee Lee, Samuel Sangkon Lee, Seungjong Chung, and Dongun An. 2007. Harmful contents classification using the harmful word filtering and SVM. In *Proceedings of the International Conference on Computational Science.* Springer, 18–25.

[157] Roselyn J. Lee-Won, Tiffany N. White, Hyunjin Song, Ji Young Lee, and Mikhail R. Smith. 2019. Source magnification of cyberhate: Affective and cognitive effects of multiple-source hate messages on target group members. *Media Psychol.* 23, 5 (2020), 603–624.

[158] Sam Levin. 2017. Google to hire thousands of moderators after outcry over YouTube abuse videos. Retrieved from https://www.theguardian.com/technology/2017/dec/04/google-youtube-hire-moderators-child-abuse-videos.

[159] Z. Li, J. Kawamoto, Y. Feng, and K. Sakurai. 2016. Cyberbullying detection using parent-child relationship between comments. In *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services.* ACM Press, New York. https://doi.org/10.1145/3011141.3011182

[160] S. Liu and T. Forss. 2015. Text classification models for web content filtering and online safety. In *Proceedings of the IEEE International Conference on Data Mining Workshop (ICDMW).* IEEE, 961–968. https://doi.org/10.1109/ICDMW.2015.143

[161] Gabi Löschper, Amélie Mummendey, Volker Linneweber, and Manfred Bornewasser. 1984. The judgement of behaviour as aggressive and sanctionable. *Eur. J. Soc. Psychol.* 14, 4 (1984), 391–404.

[162] Rijul Magu and Jiebo Luo. 2018. Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW).*

[163] V. Mal and A. J. Agrawal. 2016. Removing flaming problems from social networking sites using semi-supervised learning approach. In *Proceedings of the 2nd International Conference on Information and Communication Technology for Competitive Strategies (ICTCS).* ACM, New York. https://doi.org/10.1145/2905055.2905338

[164] Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *J. Experim. Theoret. Artif. Intell.* 30, 2 (2018), 187–202.

[165] A. Mangaonkar, A. Hayrapetian, and R. Raje. 2015. Collaborative detection of cyberbullying behavior in Twitter data. In *Proceedings of the IEEE International Conference on Electro/Information Technology (EIT).* https://doi.org/10.1109/EIT.2015.7293405

[166] Marcus Märtens, Siqi Shen, Alexandru Iosup, and Fernando A. Kuipers. 2015. Toxicity detection in multiplayer online games. In *Proceedings of the International Workshop on Network and Systems Support for Games, NetGames.* IEEE, 1–6. https://doi.org/10.1109/NetGames.2015.7382991

[167] Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Telev. New Media* 22, 2 (2021), 205–224.

[168] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13.

[169] J. Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proc. Nat. Acad. Sci.* 116, 20 (2019), 9785–9789.

[170] M. L. McHugh. 2012. Interrater reliability: The kappa statistic. *Biochem. Medic.* 22, 3 (2012), 276–282.

[171] A. Mishra and R. Rastogi. 2012. Semi-supervised correction of biased comment ratings. In *Proceedings of the 21st International Conference on World Wide Web*. ACM Press, New York. https://doi.org/10.1145/2187836.2187862

[172] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 220–229. https://doi.org/10.1145/3287560.3287596

[173] M. Mondal, L. A. Silva, and F. Benevenuto. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. ACM Press, New York, 85–94. https://doi.org/10.1145/3078714.3078723

[174] M. Mondal, L. A. Silva, D. Correa, and F. Benevenuto. 2018. Characterizing usage of explicit hate expressions in social media. *New Rev. Hypermed. Multimed.* 24, 2 (Apr. 2018), 110–130. https://doi.org/10.1080/13614568.2018.1489001

[175] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS One* 15, 8 (2020), e0237861.

[176] Norman Mu and Justin Gilmer. 2019. MNIST-C: A robustness benchmark for computer vision. *arXiv preprint arXiv:1906.02337* (2019).

[177] Amelie Mummendey and Michael Wenzel. 1999. Social discrimination and tolerance in intergroup relations: Reactions to intergroup difference. *Personal. Soc. Psychol. Rev.* 3, 2 (1999), 158–174.

[178] S. Nadali, M. A. A. Murad, N. M. Sharef, A. Mustapha, and S. Shojaee. 2013. A review of cyberbullying detection: An overview. In *Proceedings of the 13th International Conference on Intelligent Systems Design and Applications*. IEEE, 325–330. https://doi.org/10.1109/ISDA.2013.6920758

[179] B. S. Nandhini and J. I. Sheeba. 2015. Online social network bullying detection using intelligence techniques. *Procedia Comput. Sci.* 45 (Jan. 2015), 485–492. https://www.sciencedirect.com/science/article/pii/S187705091500321X.

[180] B. S. Nandhini and J. I. Sheeba. 2015. Cyberbullying detection and classification using information retrieval algorithm. In *Proceedings of the International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET)*. ACM Press, New York, 1–5. https://doi.org/10.1145/2743065.2743085

[181] T. Nitta, F. Masui, M. Ptaszynski, Y. Kimura, R. Rzepka, and K. Araki. 2013. Detecting cyberbullying entries on informal school websites based on category relevance maximization. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*.

[182] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*. ACM Press, New York. https://doi.org/10.1145/2872427.2883062

[183] Noviantho, S. M. Isa, and L. Ashianti. 2017. Cyberbullying classification using text mining. In *Proceedings of the 1st International Conference on Informatics and Computational Sciences (ICICoS)*. IEEE, 241–246. https://doi.org/10.1109/ICICOS.2017.8276369

[184] Ika Nurfarida and Laudetta Dianne Fitri. [n. d.]. Mapping and defining hate speech in Instagram's comments: A study of language use in social media. *Prosiding: Seminar Bahasa dan Sastra Nasional 8 (SENABASTRA 8) di Prodi Sastra Inggris Fakultas Ilmu Sosial dan Ilmu Budaya Universitas Trunojoyo Madura*.

[185] H. Nurrahmi and D. Nurjanah. 2018. Indonesian Twitter cyberbullying detection using text classification and user credibility. In *Proceedings of the International Conference on Information and Communications Technology (ICOIACT)*. IEEE, 543–548. https://doi.org/10.1109/ICOIACT.2018.8350758

[186] C. J. O'Dea, S. S. Miller, E. B. Andres, M. H. Ray, D. F. Till, and D. A. Saucier. 2015. Out of bounds: Factors affecting the perceived offensiveness of racial slurs. *Lang. Sci.* 52 (2015), 155–164.

[187] Council of Europe: European Commission against Racism and Intolerance (ECRI). [n. d.]. Hate speech and violence. Retrieved from https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/hate-speech-and-violence.

[188] A. Olteanu, K. Talamadupula, and K. R. Varshney. 2017. The limits of abstract evaluation metrics: The case of hate speech detection. In *Proceedings of the ACM on Web Science Conference (WebSci)*. ACM, New York, NY, 405–406. https://doi.org/10.1145/3091478.3098871

[189] Rebekah Overdorf, Bogdan Kulynych, Ero Balsa, Carmela Troncoso, and Seda Gürses. 2018. Questioning the assumptions behind fairness solutions. *arXiv preprint arXiv:1811.11293* (2018).

[190] S. Ozawa, S. Yoshida, J. Kitazono, T. Sugawara, and T. Haga. 2016. A sentiment polarity prediction model using transfer learning and its application to SNS flaming event detection. In *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 1–7. https://doi.org/10.1109/SSCI.2016.7849868

[191] S. A. Ozel, E. Sarac, S. Akdemir, and H. Aksu. 2017. Detection of cyberbullying on social media messages in Turkish. In *Proceedings of the International Conference on Computer Science and Engineering*. IEEE. https://doi.org/10.1109/UBMK.2017.8093411

[192] Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on Twitter. In *Proceedings of the 1st Workshop on Abusive Language Online, ALW@ACL 2017, Vancouver, BC, Canada, August 4, 2017*, Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 41–45. https://doi.org/10.18653/v1/w17-3006

[193] J. H. Park, J. Shin, and P. Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2799–2804.

[194] Irina Sumajin Parkins, Harold D. Fishbein, and P. Neal Ritchey. 2006. The influence of personality on workplace bullying and discrimination. *J. Appl. Soc. Psychol.* 36, 10 (2006), 2554–2577.

[195] Demetris Paschalides, Dimosthenis Stephanidis, Andreas Andreou, Kalia Orphanou, George Pallis, Marios D. Dikaiakos, and Evangelos Markatos. 2020. MANDOLA: A big-data processing and visualization platform for monitoring and detecting online hate speech. *ACM Trans. Internet Technol.* 20, 2 (2020), 1–21.

[196] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2020. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv preprint arXiv:2012.05345* (2020).

[197] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper attention to abusive user content moderation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1125–1135.

[198] John Pavlopoulos, Prodromos Malakasiotis, Juli Bakagianni, and Ion Androutsopoulos. 2017. Improved abusive comment moderation with user embeddings. In *Proceedings of the 2017 Workshop: Natural Language Processing meets Journalism, NLPmJ@EMNLP, Copenhagen, Denmark, September 7, 2017*, Octavian Popescu and Carlo Strapparava (Eds.). Association for Computational Linguistics, 51–55. https://doi.org/10.18653/v1/w17-4209

[199] John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? *arXiv preprint arXiv:2006.00998* (2020).

[200] R. Pelle, C. Alcântara, and V. P. Moreira. 2018. A classifier ensemble for offensive text detection. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*. ACM Press, New York, 237–243. https://doi.org/10.1145/3243082.3243111

[201] Billy Perrigo. 2019. Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch. Retrieved from https://time.com/5739688/facebook-hate-speech-languages/.

[202] Barbara Perry et al. 2001. *In the Name of Hate: Understanding Hate Crimes*. Psychology Press.

[203] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*. 2227–2237.

[204] Jean S. Phinney, Tanya Madden, and Lorena J. Santos. 1998. Psychological variables as predictors of perceived ethnic discrimination among minority and immigrant adolescents. *J. Appl. Soc. Psychol.* 28, 11 (1998), 937–953.

[205] G. K. Pitsilis, H. Ramampiaro, and H. Langseth. 2018. Effective hate-speech detection in Twitter data using recurrent neural networks. *Appl. Intell.* 48, 12 (2018), 4730–4742.

[206] Karyn M. Plumm and Cheryl A. Terrance. 2013. Gender-bias hate crimes: What constitutes a hate crime from a potential juror's perspective? *J. Appl. Soc. Psychol.* 43, 7 (2013), 1468–1479.

[207] Karyn M. Plumm, Cheryl A. Terrance, and Adam Austin. 2014. Not all hate crimes are created equal: An examination of the roles of ambiguity and expectations in perceptions of hate crimes. *Curr. Psychol.* 33, 3 (2014), 321–364.

[208] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: A systematic review. *Lang. Resour. Eval.* 55, 2 (2021), 477–523.

[209] Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate speech annotation: Analysis of an Italian Twitter corpus. In *Proceedings of the 4th Italian Conference on Computational Linguistics*, Vol. 2006. CEUR-WS, 1–6.

[210] N. Potha and M. Maragoudakis. 2014. Cyberbullying detection using time series modeling. In *Proceedings of the IEEE International Conference on Data Mining Workshop*. IEEE, 373–382. https://doi.org/10.1109/ICDMW.2014.170

[211] Michal Ptaszynski, Pawel Dybala, Tatsuaki Matsuba, Fumito Masui, Rafal Rzepka, and Kenji Araki. 2010. Machine learning and affect analysis against cyber-bullying. In *Proceedings of the 36th Artificial Intelligence and Simulation of Behaviour Conference*. 7–16.

[212] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4757–4766.

[213] Jing Qian, Mai ElSherief, Elizabeth M. Belding, and William Yang Wang. 2018. Hierarchical CVAE for fine-grained hate speech classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31–November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 3550–3559. https://doi.org/10.18653/v1/d18-1391

[214] Jing Qian, Mai ElSherief, Elizabeth M. Belding, and William Yang Wang. 2018. Leveraging intra-user and inter-user representation learning for automated hate speech detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1–6, 2018, Volume 2 (Short Papers)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 118–123. https://doi.org/10.18653/v1/n18-2019

[215] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, and S. Mishra. 2018. Scalable and timely detection of cyberbullying in online social networks. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. ACM Press, New York, 1738–1747. https://doi.org/10.1145/3167132.3167317

[216] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson. 2015. Careful what you share in six seconds: Detecting cyberbullying instances in Vine. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM Press, New York, 617–622. https://doi.org/10.1145/2808797.2809381

[217] E. Raisi and B. Huang. 2018. Weakly supervised cyberbullying detection with participant-vocabulary consistency. *Soc. Netw. Anal. Mining* 8, 1 (Dec. 2018), 38. https://doi.org/10.1007/s13278-018-0517-y

[218] Arvind Ramanathan, Laura Pullum, Zubir Husein, Sunny Raj, Neslisah Torosdagli, Sumanta Pattanaik, and Sumit K. Jha. 2017. Adversarial attacks on computer vision algorithms using natural perturbations. In *Proceedings of the 10th International Conference on Contemporary Computing (IC3)*. IEEE, 1–6.

[219] Charlotte Rayner and Helge Hoel. 1997. A summary review of literature relating to workplace bullying. *J. Communit. Appl. Soc. Psychol.* 7, 3 (1997), 181–191.

[220] Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Proceedings of the Canadian Conference on Artificial Intelligence*. Springer, 16–27.

[221] M. Rezvan, S. Shekarpour, L. Balasuriya, K. Thirunarayan, V. L. Shalin, and A. Sheth. 2018. A quality type-aware annotated corpus and lexicon for harassment research. In *Proceedings of the 10th ACM Conference on Web Science*. ACM Press, New York, 33–36. https://doi.org/10.1145/3201064.3201103

[222] M. H. Ribeiro, P. H. Calais, Y. A. Santos, V. A. F. Almeida, and W. Meira. 2018. "Like sheep among wolves": Characterizing hateful users on Twitter. In *Proceedings of WSDM Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*. ACM. https://doi.org/10.475/123_4.

[223] Julian Risch, Robin Ruff, and Ralf Krestel. 2020. Offensive language detection explained. In *Proceedings of the 2nd Workshop on Trolling, Aggression and Cyberbullying*. 137–143.

[224] Nestor Rodriguez and Sergio Rojas-Galeano. 2018. Fighting adversarial attacks on online abusive language moderation. In *Proceedings of the Workshop on Engineering Applications*. Springer, 480–493.

[225] S. Rojas-Galeano. 2017. On obstructing obscenity obfuscation. *ACM Trans. Web* 11, 2 (2017), 12.

[226] H. Rosa, J. P. Carvalho, P. Calado, B. Martins, R. Ribeiro, and L. Coheur. 2018. Using fuzzy fingerprints for cyberbullying detection in social networks. In *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 1–7. https://doi.org/10.1109/FUZZ-IEEE.2018.8491557

[227] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer Mediated Communication, vol. 17*, Beisswenger M., Wojatzki M., and Zesch T. (Eds.). Bochumer Linguistischer Arbeitsberichte, 6–9.

[228] T. Roy, J. McClendon, and L. Hodges. 2018. Analyzing abusive text messages to detect digital dating abuse. In *Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 284–293. https://doi.org/10.1109/ICHI.2018.00039

[229] Maciej Rybinski, William Miller, Javier Del Ser, Miren Nekane Bilbao, and José F. Aldana-Montes. 2018. On the design and tuning of machine learning models for language toxicity classification in online platforms. In *Proceedings of the International Symposium on Intelligent and Distributed Computing*. Springer, 329–343.

[230] Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation. *arXiv preprint arXiv:1910.02334* (2019).

[231] Aastha Sahni and Naveen Raja. 2017. Analyzation and detection of cyberbullying: A Twitter based Indian case study. In *Proceedings of the International Conference on Recent Developments in Science, Engineering and Technology*. Springer, 484–497.

[232] S. Salawu, Y. He, and J. Lumsden. 2017. Approaches to automated detection of cyberbullying: A survey. *IEEE Trans. Affect. Comput.* 1 (2017), 1–1.

[233] H. M. Saleem, K. P. Dillon, S. Benesch, and D. Ruths. 2016. A web of hate: Tackling hateful speech in online social spaces. In *Proceedings of the 1st Workshop on Text Analytics for Cybersecurity and Online Safety at LREC 2016*.

[234] P. Salunkhe, S. Bharne, and P. Padiya. 2016. Filtering unwanted messages from OSN walls. In *Proceedings of the International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH).* IEEE. https://doi.org/10.1109/ICICCS.2016.7542319

[235] M. Samory and E. Peserico. 2017. Sizing up the troll: A quantitative characterization of moderator-identified trolling in an online forum. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI).* ACM, New York, NY, 6943–6947. https://doi.org/10.1145/3025453.3026007

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

[237] Edward Sapir. 2004. *Language: An Introduction to the Study of Speech.* Courier Corporation.

[238] A. Saravanaraj, J. I. Sheeba, and S. Pradeep Devaneyan. 2016. Automatic detection of cyberbullying from Twitter. *Int. J. Comput. Sci. Info. Technol. Secur* 6 (2016).

[239] Sebastian Schelter, Yuxuan He, Jatin Khilnani, and Julia Stoyanovich. 2019. FairPrep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions. *arXiv preprint arXiv:1911.12587* (2019).

[240] A. Schmidt and M. Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media.* Association for Computational Linguistics, 1–10.

[241] Tina Schuh and Stephan Dreiseitl. 2018. Evaluating novel features for aggressive language detection. In *Proceedings of the International Conference on Speech and Computer.* Springer, 585–595.

[242] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* 59–68.

[243] Suin Seo and Sung-Bea Cho. 2017. Offensive sentence classification using character-level CNN and transfer learning with fake sentences. In *Proceedings of the International Conference on Neural Information Processing.* Springer, 532–539.

[244] Sima Sharifirad and Stan Matwin. 2019. Using attention-based bidirectional LSTM to identify different categories of offensive language directed toward female celebrities. In *Proceedings of the Workshop on Widening NLP.* 46–48.

[245] Sanjana Sharma, Saksham Agrawal, and Manish Shrivastava. 2018. Degree based classification of harmful speech using Twitter data. In *Proceedings of the 1st Workshop on Trolling, Aggression and Cyberbullying, TRAC@COLING 2018, Santa Fe, New Mexico, USA, August 25, 2018*, Ritesh Kumar, Atul Kr. Ojha, Marcos Zampieri, and Shervin Malmasi (Eds.). Association for Computational Linguistics, 106–112. https://www.aclweb.org/anthology/W18-4413/.

[246] J. I. Sheeba and K. Vivekanandan. 2013. Low frequency keyword extraction with sentiment classification and cyberbully detection using fuzzy logic technique. In *Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research.* IEEE, 1–5. https://doi.org/10.1109/ICCIC.2013.6724124

[247] Del Siegle. 2010. Cyberbullying and sexting: Technology abuses of the 21st century. *Gifted Child Today Mag.* 33, 2 (2010), 14–16.

[248] Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Proceedings of the 10th International AAAI Conference on Web and Social Media.* AAAI, 687–690.

[249] Vivek K. Singh and Connor Hofenbitzer. 2019. Fairness across network positions in cyberbullying detection algorithms. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.* 557–559.

[250] M. Sloane, E. Moss, O. Awomolo, and L. Forlano. 2020. Participation is not a design fix for machine learning. In *Proceedings of the 37th International Conference on Machine Learning*, Vienna, Austria, PMLR 119, 2020.

[251] Peter K. Smith and Georges Steffgen. 2013. *Cyberbullying through the New Media: Findings from an International Network.* Psychology Press.

[252] Olivia Solon. 2017. Facebook is hiring moderators. But is the job too gruesome to handle? Retrieved from https://www.theguardian.com/technology/2017/may/04/facebook-content-moderators-ptsd-psychological-dangers.

[253] Sara Owsley Sood, Judd Antin, and Elizabeth F. Churchill. 2012. Profanity use in online communities. In *CHI Conference on Human Factors in Computing Systems, CHI'12, Austin, TX, USA - May 05–10, 2012*, Joseph A. Konstan, Ed H. Chi, and Kristina Höök (Eds.). ACM, 1481–1490. https://doi.org/10.1145/2207676.2208610

[254] S. O. Sood, J. Antin, and E. F. Churchill. 2012. Using crowdsourcing to improve profanity detection. In *Proceedings of the AAAI Spring Symposium: Wisdom of the Crowd*, Vol. 12. 06.

[255] S. O. Sood, E. F. Churchill, and J. Antin. 2012. Automatic identification of personal insults on social news sites. *J. Assoc. Inf. Sci. Technol.* 63, 2 (2012), 270–285.

[256] Anna Squicciarini, Sarah Rajtmajer, Y. Liu, and Christopher Griffin. 2015. Identification and characterization of cyberbullying dynamics in an online social network. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.* 280–285.

[257] Robert J. Sternberg. 2018. FLOTSAM: A model for the development and transmission of hate. *J. Theoret. Soc. Psychol.* 2, 4 (2018), 97–106.
[258] R. Sugandhi, A. Pande, S. Chawla, A. Agrawal, and H. Bhagat. 2015. Methods for detection of cyberbullying: A survey. In *Proceedings of the 15th International Conference on Intelligent Systems Design and Applications (ISDA)*. IEEE, 173–177. https://doi.org/10.1109/ISDA.2015.7489220
[259] Megan Sullaway. 2004. Psychological perspectives on hate crime laws. *Psychol., Pub. Polic., Law* 10, 3 (2004), 250.
[260] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1630–1640.
[261] Harini Suresh and John V. Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002* (2019).
[262] T. Davidson, D. Warmsley, M. Macy, and I. Weber. 2017. Automated hate speech detection and the problem of offensive language. (Mar. 2017). arXiv:1703.04009. http://arxiv.org/abs/1703.04009.
[263] N. Tahmasbi and A. Fuchsberger. 2018. Challenges and future directions of automated cyberbullying detection. *AMCIS 2018 Proc.* (Aug. 2018). https://aisel.aisnet.org/amcis2018/SocialComputing/Presentations/10.
[264] N. Tahmasbi and E. Rastegari. 2018. A socio-contextual approach in automated detection of cyberbullying. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*. https://aisel.aisnet.org/hicss-51/dsm/social{_}media{_}culture/3.
[265] Jasmine Tata. 1993. The structure and phenomenon of sexual harassment: Impact of category of sexually harassing behavior, gender, and hierarchical level. *J. Appl. Soc. Psychol.* 23, 3 (1993), 199–211.
[266] P. L. Teh, C.-B. Cheng, and W. M. Chee. 2018. Identifying and categorising profane words in hate speech. In *Proceedings of the 2nd International Conference on Compute and Data Analysis*. ACM Press, New York. https://doi.org/10.1145/3193077.3193078
[267] Serra Sinem Tekiroglu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. *arXiv preprint arXiv:2004.04216* (2020).
[268] A. Tsesis. 2001. Hate in cyberspace: Regulating hate speech on the Internet. *San Diego L. Rev.* 38 (2001), 817.
[269] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. 2020. From ImageNet to image classification: Contextualizing progress on benchmarks. *arXiv preprint arXiv:2005.11295* (2020).
[270] Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based approach to racism detection in Dutch social media. In *Proceedings of the 1st Workshop on Text Analytics for Cybersecurity and Online Safety*. 1–7.
[271] Stefanie Ullmann and Marcus Tomalin. 2020. Quarantining online hate speech: Technical and ethical perspectives. *Ethics Inf. Technol.* 22, 1 (2020), 69–80.
[272] Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online, ALW@EMNLP 2018, Brussels, Belgium, October 31, 2018*. Darja Fiser, Ruihong Huang, Vinodkumar Prabhakaran, Rob Voigt, Zeerak Waseem, and Jacqueline Wernimont (Eds.) Association for Computational Linguistics, 33–42. https://doi.org/10.18653/v1/w18-5105
[273] Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *PloS One* 13, 10 (2018).
[274] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*.
[275] Michael Veale, Reuben Binns, and Lilian Edwards. 2018. Algorithms that remember: Model inversion attacks and data protection law. *Philos. Trans. Roy. Societ. A: Math., Phys. Eng. Sci.* 376, 2133 (2018), 20180083.
[276] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.
[277] Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *arXiv preprint arXiv:2004.01670* (2020).
[278] Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the 3rd Workshop on Abusive Language Online*. 80–93.
[279] Bertie Vidgen and Taha Yasseri. 2020. Detecting weak and strong Islamophobic hate speech on social media. *J. Inf. Technol. Polit.* 17, 1 (2020), 66–78.
[280] Jeremy Waldron. 2012. *The Harm in Hate Speech*. Harvard University Press.
[281] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2014. Cursing in English on Twitter. In *Computer Supported Cooperative Work, CSCW'14, Baltimore, MD, February 15–19, 2014*, Susan R. Fussell, Wayne G. Lutters, Meredith Ringel Morris, and Madhu Reddy (Eds.). ACM, 415–425. https://doi.org/10.1145/2531602.2531734

[282] W. Warner and J. Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the 2nd Workshop on Language in Social Media*. Association for Computational Linguistics, 19–26.

[283] Z. Waseem. 2016. Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the 1st Workshop on NLP and Computational Social Science*.

[284] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks.. arXiv:1705.09899. http://aclweb.org/anthology/W17-3012. http://arxiv.org/abs/1705.09899.

[285] Z. Waseem and D. Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the Student Research Workshop@ the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. 88–93.

[286] Chris Welty, Praveen Paritosh, and Lora Aroyo. 2019. Metrology for AI: From benchmarks to instruments. *arXiv preprint arXiv:1911.01875* (2019).

[287] Mike Wendling. 2015. 2015: The year that angry won the internet. Retrieved from https://www.bbc.com/news/blogs-trending-35111707.

[288] Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words—A feature-based approach. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1046–1056.

[289] A. Williams, C. Oliver, K. Aumer, and C. Meyers. 2016. Racial microaggressions and perceptions of Internet memes. *Comput. Hum. Behav.* 63 (2016), 424–432.

[290] Michael Wojatzki, Tobias Horsmann, Darina Gold, and Torsten Zesch. 2018. Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments. In *Proceedings of the 14th Conference on Natural Language Processing, KONVENS 2018, Vienna, Austria, September 19–21, 2018*, Adrien Barbaresi, Hanno Biber, Friedrich Neubarth, and Rainer Osswald (Eds.). Österreichische Akademie der Wissenschaften, 110–120. https://www.oeaw.ac.at/fileadmin/subsites/academiaecorpora/PDF/konvens18_13.pdf.

[291] Julie A. Woodzicka, Robyn K. Mallett, Shelbi Hendricks, and Astrid V. Pruitt. 2015. It's just a (sexist) joke: Comparing reactions to sexist versus racist communications. *Humor* 28, 2 (2015), 289–309.

[292] E. Wulczyn, N. Thain, and L. Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1391–1399.

[293] Tomer Wullach, Amir Adler, and Einat Minkov. 2020. Towards hate speech detection at large via deep generative modeling. *arXiv preprint arXiv:2005.06370* (2020).

[294] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, 7–14. DOI: https://aclanthology.org/2020.socialnlp-1.2

[295] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore. 2012. *Learning from Bullying Traces in Social Media*. Association for Computational Linguistics. https://dl.acm.org/citation.cfm?id=2382139.

[296] Z. Xu and S. Zhu. 2010. Filtering offensive language in online communities using grammatical relations. In *Proceedings of the 7th Annual Collaboration, Electronic Messaging, Anti-abuse and Spam Conference*. 1–10.

[297] S. H. Yadav and P. M. Manwatkar. 2015. An approach for offensive text detection and prevention in Social Networks. In *Proceedings of the International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*. IEEE, 1–4. https://doi.org/10.1109/ICIIECS.2015.7193018

[298] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D. Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proc. Content Anal. WEB* 2 (2009).

[299] S. Yoshida, J. Kitazono, S. Ozawa, T. Sugawara, T. Haga, and S. Nakamura. 2014. Sentiment analysis for various SNS media using naive Bayes classifier and its application to flaming detection. In *Proceedings of the IEEE Symposium on Computational Intelligence in Big Data (CIBD)*. IEEE, 1–6. https://doi.org/10.1109/CIBD.2014.7011523

[300] W. D. Yu, M. Gole, N. Prabhuswamy, S. Prakash, and V. G. Shankaramurthy. 2016. An approach to design and analyze the framework for preventing cyberbullying. In *Proceedings of the IEEE International Conference on Services Computing (SCC)*. IEEE, 864–867. https://doi.org/10.1109/SCC.2016.125

[301] Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. *arXiv preprint arXiv:2004.14088* (2020).

[302] X. Zhang, J. Tong, N. Vishwamitra, E. Whittaker, J. P. Mazer, R. Kowalski, H. Hu, F. Luo, J. Macbeth, and E. Dillon. 2016. Cyberbullying detection with a pronunciation based convolutional neural network. In *Proceedings of the 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. https://doi.org/10.1109/ICMLA.2016.0132

[303] Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? The challenging case of long tail on Twitter. *Seman. Web* 10, 5 (2019), 925–945. https://doi.org/10.3233/SW-180338

[304] Z. Zhang, D. Robinson, and J. Tepper. 2018. Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In *The Semantic Web*, Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam (Eds.). Springer International Publishing, Cham, 745–760.

[305] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 629–634.

[306] R. Zhao and K. Mao. 2017. Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. *IEEE Trans. Affect. Comput.* 8, 3 (July 2017), 328–339. https://doi.org/10.1109/TAFFC.2016.2531682

[307] R. Zhao, A. Zhou, and K. Mao. 2016. Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th International Conference on Distributed Computing and Networking*. ACM Press, New York, 1–6. https://doi.org/10.1145/2833312.2849567

[308] H. Zhong, H. Li, A. Squicciarini, S. Rajtmajer, C. Griffin, D. Miller, and C. Caragea. 2016. Content-driven detection of cyberbullying on the Instagram social network. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. AAAI Press, 3952–3958.

[309] Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2018/summaries/292.html.

[310] D.-S. Zois, A. Kapodistria, M. Yao, and C. Chelmis. 2018. Optimal online cyberbullying detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. https://doi.org/10.1109/ICASSP.2018.8462092

Received June 2020; revised July 2021; accepted August 2021