Toward Disambiguating the Definitions of Abusive, Offensive, Toxic, and Uncivil Comments

Pia Pachinger[1]
Research Unit
Data Science
TU Wien

Allan Hanbury
Research Unit
Data Science
TU Wien

Julia Neidhardt
CD Lab for Recommender Systems
TU Wien

Anna Planitzer
PolCom
Research Group
U of Vienna

Abstract

The definitions of *abusive, offensive, toxic* and *uncivil* comments used for annotating corpora for automated content moderation are highly intersected and researchers call for their disambiguation. We summarize the definitions of these terms as they appear in 23 papers across different fields. We compare examples given for *uncivil, offensive*, and *toxic* comments, attempting to foster more unified scientific resources. Additionally, we stress that the term *incivility* that frequently appears in social science literature has hardly been mentioned in the literature we analyzed that focuses on computational linguistics and natural language processing.

1 Introduction

The current low to toxic quality of online discussions and the massive amount of user-generated content lead to the need of automatic content moderation (Su et al., 2018). But the definitions of which comments are actually in need of moderation are not standardized, resulting in a clutter of inconsistent annotated data sets which makes it difficult to build models using multiple data sources (Poletto et al., 2021). Phenomena such as hate speech and offensiveness cannot be distinguished by classification models and rare or subtle forms of abusive language are not detected (Davidson et al. 2017, Jurgens et al. 2019).

Fortuna et al. (2020) analyzed the similarity of classes of six distinct hate speech data sets and compared the predicted labels for these data sets with the Perspective API Toxicity Classifier. They came to the conclusion that many definitions are used for equivalent concepts. They called for avoidance of creating new categories and for referring to categories already existing in the literature. Furthermore, they stated that if a new category is defined it should be justified and clearly defined.

Khurana et al. (2022a) proposed a framework consisting of the aspects *target group, dominance of target group, perpetrator characteristics, type of negative group reference*, and *potential consequences*. This framework should provide the means to classify data sets on *hate speech* in a unified manner, but for now it has not been expanded on more subtle forms of abuse such as *toxic* speech.

We analyze and compare prominent papers across languages and fields focusing on online *abusiveness, incivility, offensiveness* and *toxicity*. Concretely, we contribute the following insights:

- An overview of the definitions of *abusiveness, incivility, offensiveness* and *toxicity* in the context of content moderation as they appear in 23 prominent papers across fields
- A comparison of examples given for *incivility, offensiveness* and *toxicity* in these papers
- Pointers to potentially relevant contents on *incivility* originating from the field of communication science

These efforts should inspire future work on how to merge already existing but non unified valuable data sources and on how to build annotated corpora which are compatible with existing corpora.

2 Related Work

Madukwe et al. (2020) compared the attributes of existing data sets for hate speech detection. They outlined their limitations, called for a benchmark data set and recommend approaches for improving quality of research in this field.

Risch et al. (2021a) provided code to automatically merge the labels of 43 data sets, resulting in 57 sub classes of toxicity. Yet, they did not provide detailed information on the meaning of the labels.

In order to be able to detect nuances of abusive language and to provide well-defined classes for

[1]Contact: pia.pachinger@tuwien.ac.at

Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), pages 107–113

classification models, more fine grained annotations were proposed:

Directed towards an individual / a generalized group	Waseem et al. 2017
Targeted (to an individual or a group), Not targeted	Zampieri et al. 2019a
Explicit, Implicit	Waseem et al. 2017, Ousidhoum et al. 2019, Caselli et al. 2020, Demus et al. 2022b
Target group	Basile et al. 2019, Ousidhoum et al. 2019, Shvets et al. 2021, Khurana et al. 2022b, Demus et al. 2022b
Attribute based on which post discriminates	Ousidhoum et al. 2019, Shvets et al. 2021
Annotators' feelings	Ousidhoum et al. 2019
Criminal relevance	Demus et al. 2022b

3 Definitions of *Abusive, Offensive, Toxic,* and *Uncivil Talk*

We analyzed prominent papers across fields and languages treating the terms *abusiveness* (*abusive language / speech*), *offensiveness* (*offensive language / speech*), *toxicity* (*toxic language / speech*) and *incivility* (*uncivil language / speech*).

The analyzed sources contain six overviews of shared tasks (Germeval and Semeval) on *abusive, offensive* or *toxic* comment classification in German and English, two toxicity classification challenges by Google Jigsaw, a survey paper on hate speech detection, two resource papers on annotated hate speech corpora, one resource paper on an annotated corpus on offensive comments, five papers on different aspects of hate speech and toxic comment detection and six papers from the social science domain. Only three of the analyzed papers have less than 30 citations (they are all from 2021). Only Risch et al. (2021b) referred to annotation guidelines which were not entirely documented in the paper. We analyzed the annotation guidelines documented in the papers.

We summarized the definitions for the concepts in Table 1. The definitions vary notably in their length and scope for all concepts. Furthermore, we can observe a difference in the publication venues where definitions for the distinct concepts appear.

4 Relations of *Abusive, Offensive, Toxic,* and *Uncivil Talk*

We summarized the verbally expressed statements of how the concepts relate to each other in the papers (Table 2). The analyzed papers are the same as in Table 1. $A = B$ means that concepts A and B were used as synonyms. $A \subset B$ expresses that B was understood as a broader concept than A and that all instances of A are also instances of B. To give an example, Pavlopoulos et al. (2021b) stated that "*[...] the majority of the short spans comprises common cuss or clearly abusive words, which can be directly classified as toxic*" in their error analysis. From this sentence we extracted the relation Abusive $\subset$ Toxic. Another example is the relation depicted in Fortuna et al. (2020): "*[...] Scientific publications focused on the automatic detection of different types of offensive speech, among them, e.g., toxicity, hate, abuse [...]*".

Implications such as $B \supset A$ were not added to the table for readability. $A \subseteq B$ expresses the same as $A \subset B$, additionally there is the possibility that A and B are the same concept, but this is not explicitly stated. $A \not\subset B$ depicts that the authors implicitly state that there exist instances which are examples of concept A but not of concept B.

The implications of all these statements clearly lead to several contradictions, which point once more to the fact that there do not exist generally accepted definitions of these concepts.

5 Instances of *Offensive, Toxic,* and *Uncivil Talk*

We manually extracted examples given for the distinct concepts in the analyzed papers. We will henceforth call these examples *instances*. For instance, a *hurtful* comment is an *instance* of an *offensive* comment according to Wiegand et al. (2019) (Table 1). The extracted instances can be found in Figures 1 and 2. The instances were extracted from the papers appearing in Table 1. We either found the instances as examples given for the definitions of the concepts or from the annotation guidelines appearing in the papers. We fused the following terms which we considered to be very similar:

Degrading	→ Aspersion
Derogatory	→ Pejorative
Disrespectful	→ Rude
Identity attack	→ Personal attack
Vulgarity, swearing	→ Profanity

We found few instances for *abusiveness*, therefore we did not depict them in the figures.

Paper / Shared task	**Toxic talk / Toxicity**
Jigsaw 2018, Jigsaw 2019, Risch et al. 2021a	Likely to make someone leave a discussion (Disrespect, rudeness)
Poletto et al. 2021	(Aggressiveness, hate speech, homophobia, misogyny, racism)
SemEval 2021 (Pavlopoulos et al.)	Somewhat likely to make a user leave a discussion or give up on sharing their perspective (Disrespect, identity attacks, insults, obscenity, rudeness, threats, unreasonableness)
Germeval 2021 (Risch et al.)	Uncivil forms of communication (Accusation of lying, attacks on democracy, discrimination or discreditation of participants, implied volume via capital letters, insults of participants, vulgarity, sarcasm, making it difficult for others to participate, threats of violence)
Demus et al. 2022a	Potential of a comment to "poison" a conversation. Encourages aggressive responses or triggers other participants to leave the conversation.
	Offensive talk / Offensiveness
Davidson et al. 2017	Targets disadvantaged social groups in a potentially harmful manner
Germeval 2018 (Wiegand et al.) Germeval 2019 (Struß et al.)	Abusive language, insults, profanity
Semeval 2019 (Zampieri et al.)	Any form of non-acceptable language, or a targeted offense, veiled or direct. This consists of insult/threat to an individual or a group or profanity and swearing.
Wiegand et al. 2019	Hurtful, derogatory or obscene utterances to another person (Cyberbullying, hate speech)
Semeval 2020 (Zampieri et al.)	Targeted insult or threat towards a group or an individual, or text containing untargeted profanity or swearing
Paasch-Colberg et al. 2021	Insults, degrading metaphors, degrading wordplays, slurs
Quandt et al. 2022	Attacks against single individuals that violate norms of politeness (Cyberbullying, trolling)
	Abusive talk / Abusiveness
Germeval 2018 (Wiegand et al.) Germeval 2019 (Struß et al.)	Ascribing a social identity to a person that is judged negatively by a (perceived) majority of society. This identity is seen as a shameful, unworthy, morally objectionable or marginal identity. The target of judgment is seen as a representative of a group and it is ascribed negative qualities that are taken to be universal.
Ousidhoum et al. 2019	A tweet sounding dangerous
	Uncivil talk / Incivility
Coe et al. 2014	Unnecessarily disrespectful tone toward the discussion forum, its participants, or its topics. Key forms: Aspersion, name-calling, lying, pejorative speech, vulgarity
Muddiman 2017	Rudeness, emotion, name-calling, extreme partisan attacks (e.g. calling the political opposition Nazis), norm violations (e.g. misinformation)
Rossini 2019	Mockery, disdain, pejorative language, profanity, personal attacks focused on demeaning characteristics, personality, ideas, or arguments
Otto et al. 2020	Violation of norms of interpersonal interaction (Eye-rolling, exaggeration, ignoring the opponent, insults, name calling)
Germeval 2021 (Risch et al.)	Violation of democratic discourse values (Attacking basic democratic principles, complicating participation of others)
Rossini 2022	Violation of discussion and social norms. Sub types: Attacks on arguments or perspective, lying and aspersion, personal attack, profanity or vulgarity. (Shouting)

Table 1: Definitions of *abusive, offensive, toxic* and *uncivil* speech according to distinct sources. Pink lines represent papers published in venues mainly covering computational linguistics and NLP, blue lines represent venues mainly covering other fields. Terms in brackets are examples given for the respective concept.

Paper	**Toxic**	**Offense**	**Abuse**	**Uncivil**
Germeval 2018			$\subset$ Offense	
Germeval 2019			$\subset$ Offense	
Wiegand et al. 2019		$=$ Abuse	$=$ Offense	
Fortuna et al. 2020	$\subseteq$ Offense		$\subseteq$ Offense	
Semeval 2020		$\subseteq$ Abuse		
Germeval 2021c	$\subseteq$ Uncivil	$=$ Toxic	$=$ Toxic	
Poletto et al. 2021	= Abuse	$\nsubseteq$Toxic	= Toxic	
Risch et al. 2021a		$\subseteq$ Toxic	$\subseteq$ Toxic	$\subseteq$ Toxic
SemEval 2021a			$\subset$ Toxic	
Shvets et al. 2021		$\subset$Abuse		
Gevers et al. 2022		$\subseteq$ Toxic	$\subseteq$ Toxic	
Rossini 2022		$\subset$ Abuse		$\not\subset$ Toxic, $\not\subset$ Offense
Quandt et al. 2022		$\subset$ Uncivil		

Table 2: Subcategories of *abusive, offensive, toxic* and *uncivil* speech as expressed in the papers we analyzed. Some relations we extracted were only briefly mentioned in the paper. See Section 4 for details.

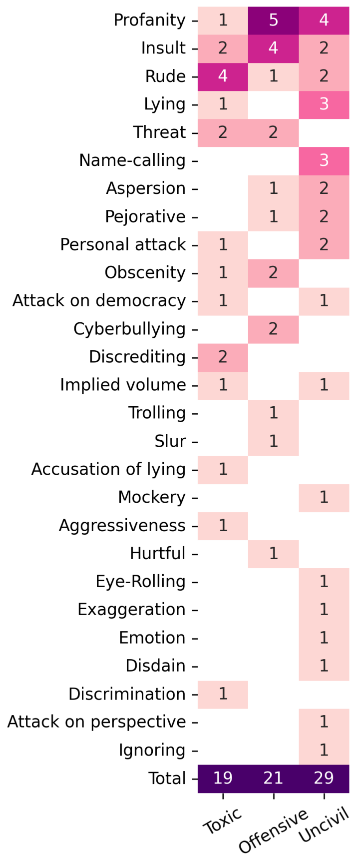

Figure 1: Instances of *incivility, offensiveness* and *toxicity*. The numbers represent the counts of the instances appearing in distinct papers.

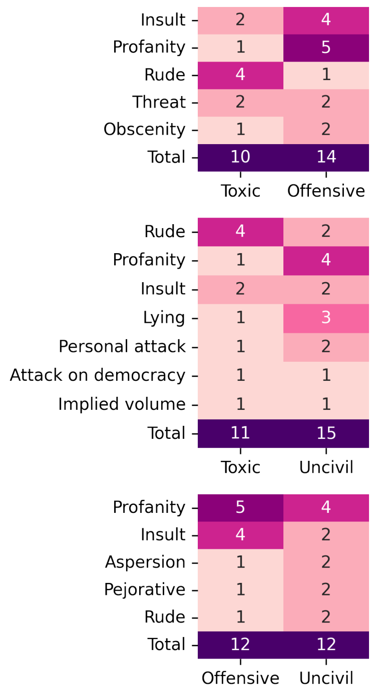

Figure 2: Common instances of *incivility, offensiveness* and *toxicity* in the literature we analyzed

6 Discussion

We found considerable overlap of instances considered as *offensive, toxic* and *uncivil* in distinct papers (Figure 2). Additionally, we verified inconsistencies regarding the percieved relations of *abusive, offensive, toxic* or *uncivil* speech (Table 2). Therefore, we propose that literature and annotated data sets on all four concepts should be taken into account when working with one of them. Tables 1 and 2 serve as initial pointers to distinct sources. The research community would benefit from exact working definitions and from listings of data and models with compatible concepts and labels.

Fortuna et al. (2020) point out that fine grained labels representing distinct aspects of a broader phenomenon such as *abusive, offensive* and *toxic speech* inherently allow for the classification model to learn more nuanced appearances of this phenomenon. They furthermore state that future annotations should be based on existing annotation guidelines in order to make data sets compatible. This is not a trivial task given that existing annotations are based on distinct perceptions of related phenomena (Table 2). We expanded the framework for developing annotation guidelines for hate speech by Khurana et al. (2022b) with suggestions for aspects which could be taken into account for annotating data sets of *abusive, offensive, uncivil* or *toxic* comments based on our findings of the previous sections (Figure 3).

7 *Incivility* from Communication Scientists' Perspectives

We noticed a considerable overlap of instances considered as *uncivil* and instances considered as *offensive* or *toxic* (Figure 2). At the same time, the term *incivility* did not appear in most of the papers published at venues for natural language processing and computational linguistics we screened (Tables 1 and 2). We provide examples of works originating from communication science exhibiting potential relevance for automated classification of *abusive, offensive, toxic* and *uncivil speech*.

Coe et al. (2014) found that incivility is associated with contextual factors such as the topic of the article and the sources quoted within the article. Moreover, they state that frequent users are more civil than infrequent users.

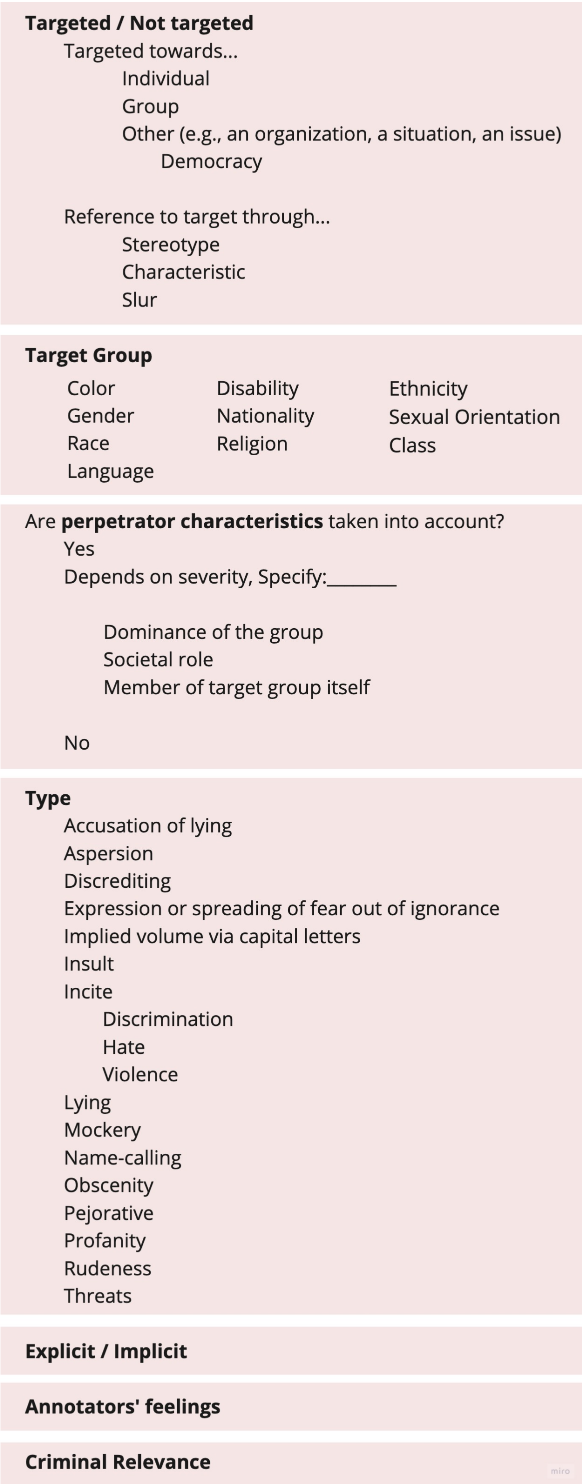

Figure 3: Aspects which can be taken into account when annotating *abusive, offensive, toxic* or *uncivil* comments. The scheme is an expansion of a proposed scheme for *hate speech* annotation by Khurana et al. (2022b). Aspects proposed in referenced papers in the table of Section 2 and instances found in the analyzed papers (Section 5) were used for expanding the framework. Note that it does not guarantee to cover all cases of *offensive, toxic* and *uncivil* language, it rather presents a summary of the 23 papers we scanned.

Muddiman (2017) found that personal-level incivility (impoliteness) is perceived as more uncivil than public-level incivility (e.g. lack of deliberativeness).

Otto et al. (2020) showed that political conflict has negative effects on political participation intention in a homogeneous manner across the Netherlands, UK and Spain. Classification models across certain languages could rely on similar annotation guidelines. Furthermore, they show that people with low tolerance for disagreement are more affected by uncivil conflict. These insights can be related to approaches where distinct classification models are trained for distinct groups of people (Akhtar et al., 2020).

8 Conclusion and Future Work

We provided an overview of definitions of the terms *abusiveness, incivility, offensiveness* and *toxicity* as they appear in the context of (automated) content moderation in 23 papers across fields. Furthermore, we compared examples given for these concepts and reflected on a more unified usage of these terms in the scientific literature on automated content moderation. Based on existing annotation guidelines, we proposed aspects which can be taken into account when designing annotation guidelines for one of the four concepts. Lastly, we introduced some examples of scientific literature on *incivility* from communication scientists' perspectives.

This paper should provoke initial thoughts on a framework for designing annotation guidelines for classifying *abusive, offensive, toxic* and *uncivil* comments that can be tailored to different tasks. There are more concepts similar to these four terms such as *intolerant speech / talk* and *dark participation* which could be analyzed as well.

Limitations

This work should serve as a pointer to awareness according to terms used in the automatic classification of *abusive, offensive, toxic* and *uncivil* online comments. It does not represent a structured review paper, therefore, we cannot guarantee to depict all usages of these terms in the context of automated content moderation.

Acknowledgements

This research has been funded by the Vienna Science and Technology Fund (WWTF) [10.47379/ICT20015].

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 151–154.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th language resources and evaluation conference*, pages 6193–6202.

Kevin Coe, Kate Kenski, and Stephen A Rains. 2014. Online and uncivil? patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, 64(4):658–679.

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

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022a. A comprehensive dataset for german offensive language and conversation analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 143–153.

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

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th language resources and evaluation conference*, pages 6786–6794.

Ine Gevers, Ilia Markov, and Walter Daelemans. 2022. Linguistic analysis of toxic language on social media. *Computational linguistics in the Netherlands journal*, 12:33–48.

Google Jigsaw. 2018. Toxic comment classification challenge.

Google Jigsaw. 2019. Multilingual toxic comment classification challenge.

David Jurgens, Eshwar Chandrasekharan, and Libby Hemphill. 2019. A just and comprehensive strategy for using nlp to address online abuse. *arXiv preprint arXiv:1906.01738*.

Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022a. Hate speech criteria: A modular approach to task-specific hate speech definitions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022b. Hate speech criteria: A modular approach to task-specific hate speech definitions. *arXiv preprint arXiv:2206.15455*.

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161.

Ashley Muddiman. 2017. Personal and public levels of political incivility. *International Journal of Communication*, 11:21.

Lukas P Otto, Sophie Lecheler, and Andreas RT Schuck. 2020. Is context the key? the (non-) differential effects of mediated incivility in three european countries. *Political Communication*, 37(1):88–107.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

Sünje Paasch-Colberg, Christian Strippel, Joachim Trebbe, and Martin Emmer. 2021. From insult to hate speech: Mapping offensive language in german user comments on immigration. *Media and Communication*, 9(1):171–180.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021a. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021b. Semeval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*, pages 59–69.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Thorsten Quandt, Johanna Klapproth, and Lena Frischlich. 2022. Dark social media participation and well-being. *Current Opinion in Psychology*, 45:101284.

Julian Risch, Philipp Schmidt, and Ralf Krestel. 2021a. Data integration for toxic comment classification: Making more than 40 datasets easily accessible in one unified format. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 157–163.

Julian Risch, Philipp Schmidt, and Ralf Krestel. 2021b. Data integration for toxic comment classification: Making more than 40 datasets easily accessible in one unified format. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 157–163, Online. Association for Computational Linguistics.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021c. Overview of the germeval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the Germeval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12.

Patricia Rossini. 2019. Toxic for whom? examining the targets of uncivil and intolerant discourse in online political talk.

Patrícia Rossini. 2022. Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, 49(3):399–425.

Alexander Shvets, Paula Fortuna, Juan Soler, and Leo Wanner. 2021. Targets and aspects in social media hate speech. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 179–190.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, Manfred Klenner, et al. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language.

Leona Yi-Fan Su, Michael A Xenos, Kathleen M Rose, Christopher Wirz, Dietram A Scheufele, and Dominique Brossard. 2018. Uncivil and personal? comparing patterns of incivility in comments on the facebook pages of news outlets. *New Media & Society*, 20(10):3678–3699.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)*, pages 602–608.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.