



# What Is Abusive Language?

## Integrating Different Views on Abusive Language for Machine Learning

Marco Niemann<sup>(✉)</sup>, Dennis M. Riehle, Jens Brunk, and Jörg Becker

University of Münster – ERCIS, Leonardo-Campus 3, 48149 Münster, Germany  
{marco.niemann,dennis.riehle,jens.brunk,  
joerg.becker}@ercis.uni-muenster.de  
<http://www.ercis.org>

**Abstract.** Abusive language has been corrupting online conversations since the inception of the internet. Substantial research efforts have been put into the investigation and algorithmic resolution of the problem. Different aspects such as “cyberbullying”, “hate speech” or “profanity” have undergone ample amounts of investigation, however, often using inconsistent vocabulary such as “offensive language” or “harassment”. This led to a state of confusion within the research community. The inconsistency can be considered an inhibitor for the domain: It increases the risk of unintentional redundant work and leads to undifferentiated and thus hard to use and justifiable machine learning classifiers. To remedy this effect, this paper introduces a novel configurable, multi-view approach to define abusive language concepts.

**Keywords:** Abusive language · Hate speech · Offensive language · Harassment · Machine learning

## 1 The Issue of Abusive Online User Comments

Online debates are getting out of control. Hidden behind the anonymity of the internet, people are posting content in a style of speech, which is unlikely to be used in the offline world. The result can, for instance, be seen in a much-noted article published in The Guardian. Many comments received by The Guardian were “crude, bigoted, or just vile”, showing of xenophobia, racism, sexism, and homophobia. The authors refer to these comments as “the dark side of Guardian comments” [38]. A similar observation could be made during the German refugee crisis in 2016, which triggered a national debate on hate against refugees and made German authorities build a special task force [35]. As a consequence, (German) news outlets have to do more intense filtering of user-generated content on their websites, which—when done manually—is a challenging task [69].

With methods from the domain of machine learning (ML) being on the rise, it is not surprising that researchers started to apply ML techniques to detect hateful comments or abusive language in general [e.g., 50, 51]. A reliable method

for the detection of abusive language would significantly reduce the currently required manual work for the moderation of user-generated content [see, e.g., 13, 64]. Unquestionably, news outlets need to moderate user-generated content, since having abusive content on their websites may not only reduce their amount of visitors but may also lower the income social media providers make with advertisement [51]. In an industry where advertisement marks a large percentage of the profit, (semi-)automatic detection of abusive language is crucial. Abstracting a bit further, we are currently facing a situation in which computational systems excel more and more in supporting the human need for communication, but lack the flexibility to deal with detrimental users disturbing the established online communities [39]. While it is intriguing to directly opt for computational solutions given the advances in machine learning and artificial intelligence, we argue to include human decisions in the moderation process, which also ensures no algorithmic censorship is taking place—an important factor for user acceptance of such systems [12]. Consequently, we want to approach this socio-technical problem [2, 41] from the underlying linguistic, legal, and academic perspective – plus potential instance-specific adaptations for individual platforms.

A core concept of supervised ML is that an ML algorithm learns data from a so-called training data set. Therefore, training data needs to be collected, which reflects the artifact based on which the algorithm should be trained. This, however, implies that we have a common understanding of the term abusive language. Nevertheless, even after more than 10 years of research in this area, there is no unequivocal vocabulary available. Through this paper, we want to illustrate the existing gap and will present an approach to create definitions that account for the different perspectives linked to abusive language.

## 2 Prior Work and Definitional Approaches

Even though online abusive language has already been around for more than 20 years, the first attempts to systematically define and identify it were conducted by [73] in 2009. At that time, their focus has been on the detection of “(online) harassment”, broadly defined as the intentional annoyance of a target, including intentional offensiveness and personal insults. However, only five years later [10] picked up the term “harassment” for one of the first abusive language publications targeting the German language. Doing so they also redefined the term to refer to electronic messages causing psychological harm to a targeted victim, also including profanity and cyber-bullying (as a repeated form). Similarly, [42, 54] altered the existing term further to include “hate speech”, “self-harm”, “sexual violence”, and “reputation damaging rumors”. In the end, both [42, 54] even agree that “harassment” might be too complex to define in a format that might serve as an annotation schema.

However, “harassment” has not remained the only term used to characterize and detect abusive language. So for example, [62] and [63] put a focus on “insults” and “profanity” while, e.g., [16, 46] focus what they term “offensive language”. Looking deeper into the single publications, clear-cut definitions are

typically still lacking, as, e.g., [63, p. 270] rather broadly summarize the detection of “insults” and “profanity” as “identifying negative content that is offered with malicious intent”. Similarly, “offensive language” according to [16, 46] can contain different types of language ranging from “vulgar” over “pornographic” to “hateful”. Interestingly, “hateful speech” or “hate speech” has also evolved to one of the core constructs of analysis for several authors such as [37, 51, 59, 66, 69]. While some of them [e.g., 66] still align it as a sub-concept of “offensive language”, others either treat it as an independent form of abusive language [e.g., 51] whereas a third group even understands “hate speech” as the primary concept and “offensive language” just as one of its sub-concepts [59, 69].

A further concept that is getting more commonly used is the term “abusive language” which has, among others, been coined by [51] as a more integrative term to summarize the already existing concepts under a larger umbrella<sup>1</sup>. As such, it has been accepted well by the community as can, e.g., be seen by the Workshop on Abusive Language Online [5, 6] now regularly taking place<sup>2</sup>. Yet, even though “abusive language” is becoming an established general concept, the exact definitions of its sub-concepts are still rather unclear. As this section has shown, there is no lack of definitions for the specific concepts, however, these often either contain large amounts of ambiguity or different definitions even contradict each other.

### 3 A Configurable, Multi-view Approach to Abusive Language Definitions

Looking back at Sect. 2 from a purely academic perspective, there appears to be little to no problem since debate and discussion are fundamental elements of academic work. However, as we pointed out in the introduction, the underlying social problem is leading to the shut-down of public discussion fora and even the suicide of attacked individuals—with no existing wide-spread computer-supported moderation support tool available<sup>3</sup>. While this is partly attributed to lacking computational intelligence in the media [see, e.g., 40], the fault might not only be with the machine but also the data it learns from. Without clear, consistent and suitable definitions of what is considered abusive, it is hard to create human-made training data sets that can be used to train effective machine classifiers [cf., e.g., 36, 68].

Hence, we want to present a novel, configurable, multi-view approach to elicit abusive language definitions (see Fig. 1) in a meaningful and consistent respectively highly reproducible manner. The approach is based on a careful analysis of potential views that should be part of a definition of “abusive language”.

<sup>1</sup> With the term “Socially Unacceptable Discourse” [36] introduced another umbrella term, which, however, so far has not received a similar uptake as “Abusive Language”.

<sup>2</sup> The third iteration in the year 2019 is already scheduled [7].

<sup>3</sup> For example, Facebook is still opening ever new moderation centers [61] and there is a growing amount of reports on how the moderation of content gets ever more unhandable [40].

A first view that comes to mind when discussing the meaning of a linguistic term is the linguistic one. It will not only help to get a deeper understanding of the meaning of the respective abusive language-related terms but also help to identify linked concepts and synonyms.

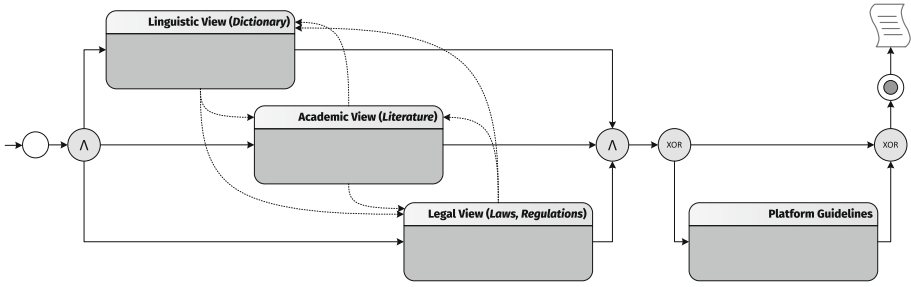
The second view of the approach has a more pragmatic background: In many countries, all web services (incl. discussion fora and comment section) are subject to certain legal restrictions concerning publishable content (cf., e.g., [36]). Hence, it will be inevitable to include a legal view to obtain an “abusive language” definition if subsequent research outcomes should ever be usable in a practice setting.

Last but not least, it will be undoubtedly helpful to reconsider prior academic work to make use of already conducted analyses and to support the alignment of a new definition with existing prior ones.

A visualization of the underlying model can be found in Fig. 1. The presented model is structured into two major parts: The first and major part describes the creation of the abusive language definitions as discussed above. In the first step (as indicated by the leftmost node) an assessment of the general linguistic notion of abusiveness will be carried out. Once a general understanding is reached, the academic literature will be assessed next to get a broader context of abusiveness notions which have been subject to academic (and practice) assessments so far. Finally, the legal perspective will be checked to identify those concepts that are justiciable and require persecution respectively deletion. However, the analysis of the three views is not conducted in a strictly iterative fashion. As indicated by the dotted arrows in Fig. 1, each view is meant to inform the other. So, for example, insights gained from the legal analysis can be double-checked against academic and dictionary sources to, e.g., account for different naming rules and conventions. To model this implicit parallelism the three views are enclosed by two AND operators.

The second part of the presented model is an extension added after the presentation at the MISDOOM 2019. It accounts for the feedback received after the presentation and subsequent talks with practitioners who outlined the need to have the ability to not rely on uniform standard definitions but to adjust them to the audiences present. The underlying issue is that the different outlets might be willing to accept different styles of language used, since, e.g., sensational outlets might go with rather loose rules whereas very traditional and sophisticated outlets might filter even beyond the legal standards (cf. also [47, 54]). However, this view/step – differing from the others – is only meant to be optional since it lacks any form of generalizability given the differences in platform terms of use [54].

Furthermore, including the linguistic and the legal perspective (and also the optional platform perspective) prohibits the achievement of a single and unified abusive language definition. Not only do the different language systems (e.g., English vs. German vs. Chinese) imply different notions of “abusiveness” – also the legal requirements will differ for different nations affected. To account for this, we decided to make the model configurable (see the dark gray boxes in Fig. 1), so that even though no unified definitions will be possible, at least the approach itself can be applied consistently.



**Fig. 1.** Configurable, multi-view approach to abusive language definitions

### 3.1 Extracting Definitional Information from Dictionaries

Even though the linguistic perspective might be one of the simpler views, it still demands careful analysis. For example, one crucial point is the selection of queried dictionaries. Here, nowadays one has to choose between the reputable paper-based versions and their more regularly updated but less controlled online counterparts [1, 3]. Furthermore, researchers have to decide to go for either monolingual or multi-lingual works, the language of choice (will typically be the language of the target country) as well as the depth of analysis to circumvent the symbol grounding problem [44, 45].

### 3.2 Extracting Definitional Information from Literature

The most common view in the extant abusive language literature is the academic one. In general, researchers can follow the guidelines for a thorough literature review as postulated by [18, 71] or [11]. Aside of the methodological approach to reviewing the literature, most other configurative options include the period to be analyzed, as well as the potential keywords (might be different for different languages) and outlets to be searched.

### 3.3 Extracting Definitional Information from Legal Texts

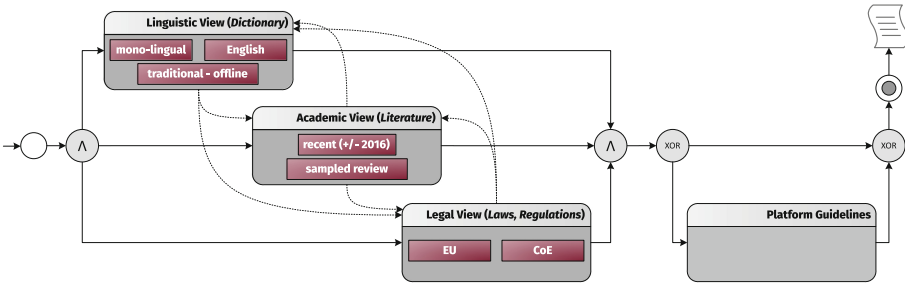
Another, supposedly rather straightforward view is the legal one. Here the obvious area of analysis is the respective national legislation of the target country. However, for many countries and areas worldwide this might be too limited as super-national organizations and institutions might have released further (non-)binding legislation to be taken into account. One classical example of this is the European Union which through its regulations and directives can have a direct impact on [24, 57]. Furthermore, it might be necessary to go beyond restrictive legislation, since many countries also have specific laws guaranteeing free speech [e.g., 31] within certain boundaries.

### 3.4 Extracting Definitional Information from Guidelines

For the adaption to platform-specific needs, relevant policy documents, community guidelines, and rules regarding content moderation need to be checked. As an example, one can refer to the paper of Pater et al. [54] who conducted such a similar analysis for major social media platforms. This configuration on a per-instance level is, however, not part of this research.

## 4 Test-Case: Europe

The European Union (EU) with its currently 28 participating nations is one of the largest political and economic unions in the world. Both the European Commission (EC) and the Council of Europe (CoE) are important players involved in the development of laws and publication of directives and resolutions. As these documents have a significant impact on the large number of member states, this paper focuses on abusive language and hate speech in Europe. While we will be able to assess our newly developed approach and demonstrate its applicability in the domain of abusive language research, this super-national focus also enhances the relevance of this publication laying an easy to use and adjust foundation for nation-specific definitions of more than two dozen European countries.



**Fig. 2.** Configurable, multi-view approach applied to the European case

The corresponding configuration of our model is depicted in Fig. 2. Given the European focus of our test case, the legal perspective will take into considerations official publications of the corresponding European bodies (CoE, EU, EC). For the linguistic view, we will focus on English as the most commonly spoken official language in Europe and restrict the assessment to traditional offline dictionaries given their higher credibility. Regarding the academic perspective, the focus will be on a selected sample of recent publications representing the current understanding of abusive language.

**Table 1.** Dictionary definitions of *Abusive* (Language)

	Coarse	Cruel	Harsh	Illegality	Injustice	Insulting	Maltreatment	Offensive	Rude	Scolding	Scurrilous	Violence
[15]		✓						✓	✓			
[17]	✓	✓	✓			✓	✓		✓	✓	✓	✓
[48]		✓				✓		✓				✓
[49]		✓	✓			✓						
[52]		✓		✓	✓	✓		✓				
[55]		✓										

#### 4.1 Dictionary-Based View

For the initial linguistic contemplation of the concept of “abusive language” six major, mono-lingual English dictionaries have been analyzed. The selection ranges from old, established (and partially academically rooted) dictionaries such as Merriam-Webster [49], Collins English Dictionary [17], Cambridge [15] and Oxford Dictionary [52] towards rather modern ones like the Macmillan Dictionary [48] and the Longman Dictionary of Contemporary English [55].

Since none of these publications defines “abusive language”, the search has been restricted to the keyword “abusive”<sup>4</sup>. Identified synonyms and related concepts have been mapped in Table 1.

Even though the analysis shows no full consensus between dictionaries, most to all of them agree on “abusive” being related to concepts such as cruelty<sup>5</sup>, “violence” (in terms of communicative harmfulness towards others), “insults” and offensiveness. Abstracting from these concepts, one could state that “abusive (language)” refers to language that is intentionally used to inflict harm on others. Given the presence of the less intentional concepts “coarse” and “harsh”, it is debatable if “intention” is a mandatory characteristic of “abusive (language)”.

To further illuminate the concept of “abusive language” from a more problem-centric perspective, the academic view will be assessed next.

#### 4.2 Literature-Based View

Following the dictionary-based view, the literature-based view adds concepts found in state-of-the-art research. Unlike dictionaries, the existing literature knows and also defines abusive language; especially in cases where Natural Language Processing (NLP) is applied to detect abusive language through computational methods. Here, annotated data sets are necessary to train machine learning classifiers. In the annotation process, clear definitions and guidelines

<sup>4</sup> Given the context of the paper, “language” is assumed to refer to written online comments.

<sup>5</sup> The theoretical need to obtain a fully-grounded definition for “cruel” as postulated through [44,45]’s symbol grounding problem is acknowledged. However, a full grounding is beyond the scope of this study and is hence left for future work of a more apt linguist.

on what constitutes abusive language and possibly its sub-concepts are necessary. The following section summarizes the most used and relevant academic definitions of abusive language. Generally, there are two different ways to approach data annotations for NLP tasks. Either, the goal is a binary classification of a text, e.g., as *abusive* or *non-abusive*, or there are multiple labels that can be applied to a text. Additionally, there are also combinations of the two. Here, often a binary main classification is used, which is then subdivided on the second level into more detailed concepts (e.g., [51]).

In the early days of abusive language detection, [58] approached the task to automatically detect internet and cellular-based text messages that contain *flame*. Accordingly, they worked on binary data sets (flame/no flame). However, their understanding of *flame* also included concepts such as *attacks*, *abusive* or *hostile* words. This kind of a binary understanding of an okay/not okay text is one of the most used approaches in the domain. In many cases ([8, 51, 58, 65, 72]) binary data sets are applied which distinguish between *abusive/flame/inappropriate* and *non-abusive/okay/clean* texts. Similarly, [4, 14, 43, 59, 60] focus on the distinction of whether something is *hateful*, *violent*, *offensive*, *sexist* or not. Additionally, there are some cases, where more than two labels are used as a scale from *okay* to *not okay* content [22, 23, 70]. Even though the labels might differ, many researchers have adopted this binary classification, which we from here on refer to as *abusive* or *clean*. When a data set is not annotated in a binary form, then it usually includes multiple (exclusive) labels. There are many studies that work with multiple labels in different combinations. The concepts that stand out the most in these works are labels regarding *sexism* (e.g., [9, 23, 66, 67]), *racism* (e.g., [9, 23, 56, 67]), *threat* (e.g., [4, 9, 56]), *insult* (e.g., [14, 56, 65]), and *profane language* (e.g., [8, 51, 65]).

### 4.3 Law-Based View

After narrowing down the concept of abusive language through a dictionary and literature analysis in the prior subsections, this part is meant to these insights with the existing regulatory framework in the European Union.

One of the first things that become apparent when assessing the different legal texts is the massive significance that is attributed to free speech by both the EU [31] and the CoE [21]. Even though not directly concerned with *abusive language* or any of its related concepts, permissive fundamental statements like these set the bar high for any valid definition since they only provide very vague statements with regard to things that may or have to be legally withheld (e.g., “subject to such formalities, conditions, restrictions or penalties [...] prescribed by law [...] for the protection of health or morals [...]”, [21]).

Unfortunately, on this supra-national level, clear-cut definitions are rare and many of the terms, e.g., more commonly used in academic literature (e.g., *profanity*, *abusive language*, *offensive language*, ...) find limited to no uptake in the legal domain so far. However, both the EU and the CoE are not completely “blind” with regard to abusive speech in general—and in online settings as a special form. Taking a step back, the assessment reveals several aspects and



concepts exhibiting a strong legal relevance. Different from the literature-based approach, the binary distinction into acceptable and in-acceptable usually is not explicated in the legal context—even though always implicitly present, since the above-stated Article 10 and Article 11 make everything legal which is not rendered illegal by further restrictions. Given Europe’s history of nationalism and racism-induced wars, the first restrictions of Articles 10 and 11 were made regarding all forms of “hatred, xenophobia, [...] or other forms of hatred” [19] which are based on for example “aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin” [19]. Hence, looking at the concepts deemed problematic on the European level, the majority of relevant documents [19, 20, 25–29] specify *racist* and *xenophobic* utterances<sup>6</sup> as abusive and hence punishable offenses. In several of the stated cases, the documents give further indications on what they consider to be problematic, e.g., explicitly stating skin color, personal descent as well as the origin from a national or ethnic perspective as subsumed characteristics [29]. In recent years the gender debate and associated forms of discrimination also have found their way into legal considerations [20, 29, 30, 32, 34] making *sexism* a category that is not only present in academic considerations. Similar to academic literature, a concept that is often wrapped around and hence often present in the analyzed documents is the so-called *hate speech* [19]. It is often used as an umbrella term to subsume *racist* and *sexist* offenses following a similar style of speech but targeting a very different set of victims. Hence, we refrain from using the rather imprecise aggregate term and stick with the more precisely described concepts of *racism* and *sexism*. Aside from the denigrating talk linked to the previously stated abusiveness concepts both CoE and EU documents also repeatedly list *threats* and *insults* as further strictly prohibitive offenses [28, 29, 33, 53], specifying them as proposed attacks on the physical integrity of the victims [53] respectively omitting further specifications. Further concepts such as *offensive language* or *profanity* only get few mentions [29], which is, however, understandable as these concepts are morally debatable but are far from being justiciable.

#### 4.4 Synopsis

After completing the assessment of all three views, the causal reason for our undertaking is reaffirmed: Abusive language is indeed a diverse and broadly defined topic. However, the analysis of the literature and the legislative view indicate four concepts that have to be distinguished:

---

<sup>6</sup> We subsume *anti-semitism*, *anti-muslim*, and other *religious* utterance at this point.

<b>sexism</b>	Attacks on people based on their gender (identity), often with a focus on women
<b>racism</b>	Attacks on people based on their origin, ethnicity, nation - typically meant to incite some form of hatred
<b>threats</b>	Announcements of the violation of the physical integrity of the victim
<b>insults</b>	Denigrating, insolent or contemptuous statements (usually left without further specification)

Even though not legally required, the re-occurrence within academic texts and the high likelihood of profane content being removed at last through community guidelines [54] made us further include:


**profane language** Usage of sexually explicit and inappropriate language

## 5 Conclusion and Outlook for the Testcase of Germany

As we outlined at the beginning of this paper, in prior work on abusive language and related constructs, there is considerable ambiguity regarding the exact definition and relationship between the used concepts. To remedy the situation, we propose a new configurable approach to abusive language definitions including a linguistic, legal and academic view. In addition, our model is capable of including a platform-specific point of view, which allows customization on a per-instance level. We think that this is an important feature to adopt platform-specific needs and to foster a common understanding of abusive language. However, as this adoption refers to concrete platforms and happens after the aggregation of the other three views (c.f. Fig. 1), we have not further discussed this process during the creation of our model.

Subsequently, we demonstrated the applicability of our newly developed model through the creation of a definition for the European level. Based on the assessment of English dictionaries, recent academic literature, and legal documents from the European level we were able to successfully elicit five abusive language concepts that need to be treated and defined. While this European approach provides an ample basis to label and categorize comments, it also provides an easy opportunity to fine-tune the definitions for a national level by re-configuring the legal and linguistic view. This may, for instance, be the case when our approach is adopted for a European country with a language other than English.

Given the design-orientation of the presented artifact, a step for future research would be the application of an abusive language definition created by our configurable model in a practice setting, to assess its suitability and to further evaluate our approach.

**Acknowledgments.** The research leading to these results received funding from the federal state of North Rhine-Westphalia and the European Regional Development Fund (EFRE.NRW 2014–2020), Project:  **DERATI** (No. CM-2-2-036a).

## References

1. Abel, A., Meyer, C.M.: The dynamics outside the paper: user contributions to online dictionaries. In: Proceedings of the 3rd eLex Conference 'Electronic Lexicography in the 21st Century: Thinking Outside the Paper', pp. 179–194. eLex, Tallinn (2013)
2. Ackerman, M.S.: The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Hum. Comput. Interact.* **15**(2–3), 179–203 (2000)
3. Al Sohibani, M., Al Osaimi, N., Al Ehaidib, R., Al Muhanna, S., Dahanayake, A.: Factors that influence the quality of crowdsourcing. In: New Trends Database Information Systems II: Selected Papers 18th East European Conference on Advances in Databases and Information Systems and Associated Satellite Events, ADBIS 2014, Ohrid, Macedonia, pp. 287–300 (2015)
4. Anzovino, M., Fersini, E., Rosso, P.: Automatic identification and classification of misogynistic language on Twitter. In: Silberstein, M., Atigui, F., Kornysheva, E., Métais, E., Meziane, F. (eds.) NLDB 2018. LNCS, vol. 10859, pp. 57–64. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-91947-8\\_6](https://doi.org/10.1007/978-3-319-91947-8_6)
5. Association of Computational Linguistics: ALW1: 1st Workshop on Abusive Language Online (2017). <https://sites.google.com/site/abusivelanguageworkshop2017/home>
6. Association of Computational Linguistics: ALW2: 2nd Workshop on Abusive Language Online (2018). <https://sites.google.com/view/alw2018>
7. Association of Computational Linguistics: ALW3: 3rd Workshop on Abusive Language Online (2019). <https://sites.google.com/view/alw3/home>
8. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in Tweet. In: Proceedings 26th International Conference World Wide Web Companion, WWW 2017 Companion, pp. 759–760. International World Wide Web Conferences Steering Committee, Perth, Australia (2017)
9. Bourgonje, P., Moreno-Schneider, J., Srivastava, A., Rehm, G.: Automatic classification of abusive language and personal attacks in various forms of online communication. In: Rehm, G., Declerck, T. (eds.) GSCL 2017. LNCS (LNAI), vol. 10713, pp. 180–191. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-73706-5\\_15](https://doi.org/10.1007/978-3-319-73706-5_15)
10. Bretschneider, U., Wöhner, T., Peters, R.: Detecting online harassment in social networks. In: Proceedings International Conference on Information Systems - Building a Better World Through Information Systems, ICIS 2014, pp. 1–14. Association for Information Systems, Auckland, New Zealand (2014)
11. vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., Cleven, A.: Reconstructing the giant: on the importance of rigour in documenting the literature search process. In: Proceedings 17th European Conference on Information Systems, ECIS 2009, Verona, Italy, pp. 2206–2217 (2009)
12. Brunk, J., Mattern, J., Riehle, D.M.: Effect of transparency and trust on acceptance of automatic online comment moderation systems. In: Proceedings 21st IEEE Conference on Business Informatics, CBI 2019. IEEE, Moscow, Russia (2019)
13. Brunk, J., Niemann, M., Riehle, D.M.: Can analytics as a service save the media industry? - The case of online comment moderation. In: Proceedings 21st IEEE Conference on Business Informatics, CBI 2019. IEEE, Moscow (2019)
14. Burnap, P., Williams, M.L.: Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Sci.* **5**(1), 11 (2016)

15. Cambridge University Press: abusive (2017). <http://dictionary.cambridge.org/dictionary/english/abusive>
16. Chen, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting offensive language in social media to protect adolescent online safety. In: Proceedings 2012 ASE/IEEE International Conference on Social Computing, 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, Amsterdam, Netherlands, pp. 71–80 (2012)
17. Collins: abusive definition and meaning (2017). <https://www.collinsdictionary.com/dictionary/english/abusive>
18. Cooper, H.M.: Organizing knowledge syntheses: a taxonomy of literature reviews. *Knowl. Soc.* **1**(1), 104–126 (1988)
19. Council of Europe: Recommendation No. R (97) 20 of the Committee of Ministers to Member States on “Hate Speech” (1997)
20. Council of Europe: Recommendation No. R (97) 21 of the Committee of Ministers to Member States on the Media and the Promotion of a Culture of Tolerance (1997)
21. Council of Europe: European Convention on Human Rights (2010)
22. Davidson, T., Warmley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Eleventh International AAAI Conference on Web and Social Media, Montreal, Canada (2017)
23. Del Vigna, F., Cimino, A., Dell’Orletta, F., Petrocchi, M., Tesconi, M.: Hate me, hate me not: hate speech detection on Facebook. In: 1st Italian Conference on Cybersecurity, Venice, Italy (2017)
24. European Commission: Applying EU law (2017). [https://ec.europa.eu/info/law/law-making-process/overview-law-making-process/applying-eu-law\\_en](https://ec.europa.eu/info/law/law-making-process/overview-law-making-process/applying-eu-law_en)
25. European Commission against Racism and Intolerance: ECRI General Policy Recommendation No. 1 on Combating Racism, Xenophobia, Antisemitism and Intolerance (1996)
26. European Commission against Racism and Intolerance: ECRI General Policy Recommendation No. 2 on Specialised Bodies to Combat Racism, Xenophobia, Antisemitism and Intolerance at National Level (1997)
27. European Commission against Racism and Intolerance: ECRI General Policy Recommendation No. 6 on Combating the Dissemination of Racist, Xenophobic and Antisemitic Material via the Internet (2000)
28. European Commission against Racism and Intolerance: ECRI General Policy Recommendation No. 7 on National Legislation to Combat Racism and Racial Discrimination (2002)
29. European Commission against Racism and Intolerance: ECRI General Policy Recommendation No. 15 on Combating Hate Speech (2015)
30. European Union: Council directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin. *Off. J. Eur. Communities* **L 180**, 22–26 (2000)
31. European Union: The charter of fundamental rights of the European union. *Off. J. Eur. Communities* **C 364**, 1–22 (2000)
32. European Union: Treaty of Lisbon - amending the Treaty on European Union and the Treaty establishing the European community. *Off. J. Eur. Union* **C 306**, 1–271 (2007)
33. European Union: Council framework decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law. *Off. J. Eur. Union* **L 328**, 55–58 (2008)
34. European Union: Consolidated version of the treaty on the functioning of the European union. *Off. J. Eur. Union* **C 326**, 47–390 (2012)

35. Faiola, A.: Germany springs to action over hate speech against migrants (2016). [https://www.washingtonpost.com/world/europe/germany-springs-to-action-over-hate-speech-against-migrants/2016/01/06/6031218e-b315-11e5-8abc-d09392edc612\\_story.html?utm\\_term=.737b4d4453d3](https://www.washingtonpost.com/world/europe/germany-springs-to-action-over-hate-speech-against-migrants/2016/01/06/6031218e-b315-11e5-8abc-d09392edc612_story.html?utm_term=.737b4d4453d3)
36. Fišer, D., Erjavec, T., Ljubešić, N.: Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In: Proceedings First Workshop on Abusive Language Online, Vancouver, Canada, pp. 46–51 (2017)
37. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. *ACM Comput. Surv.* **51**(4), 1–30 (2018)
38. Gardiner, B., Mansfield, M., Anderson, I., Holder, J., Louter, D., Ulmanu, M.: The dark side of Guardian comments (2016). <https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments>
39. Gilbert, E., Lampe, C., Leavitt, A., Lo, K., Yarosh, L.: Conceptualizing, creating, & controlling constructive and controversial comments. In: Companion 2017 ACM Conference Computer Supported Cooperative Work, Social Computing, Portland, OR, USA, pp. 425–430 (2017)
40. Gillespie, T.: The scale is just unfathomable (2018). <https://logicmag.io/04-the-scale-is-just-unfathomable/>
41. Grudin, J.: Computer-supported cooperative work: history and focus. *Computer* **27**(5), 19–26 (1994)
42. Guberman, J., Hemphill, L.: Challenges in modifying existing scales for detecting harassment in individual Tweets. In: Proceedings 50th Hawaii International Conference System Sciences, HICSS 2017, pp. 2203–2212. Association for Information Systems, Waikoloa Village, Hawaii, USA (2017)
43. Hammer, H.L.: Automatic detection of hateful comments in online discussion. In: Maglaras, L.A., Janicke, H., Jones, K. (eds.) *INISCOM 2016*. LNCS, vol. 188, pp. 164–173. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-52569-3\\_15](https://doi.org/10.1007/978-3-319-52569-3_15)
44. Harnad, S.: The symbol grounding problem. *Physica D* **42**(1–3), 335–346 (1990)
45. Harnad, S.: Symbol-grounding problem. In: *Encyclopedia of Cognitive Science*, vol. 42, pp. 335–346. Wiley, Chichester (2006)
46. Jay, T., Janschewitz, K.: The pragmatics of swearing. *J. Politeness Res. Lang. Behav. Cult.* **4**(2), 267–288 (2008)
47. Köffer, S., Riehle, D.M., Höhenberger, S., Becker, J.: Discussing the value of automatic hate speech detection in online debates. In: Drews, P., Funk, B., Niemeyer, P., Xie, L. (eds.) *MKWI 2018*, Lüneburg, Germany (2018)
48. Macmillan Publishers Limited: abusive (adjective) definition and synonyms (2017). <http://www.macmillandictionary.com/dictionary/british/abusive>
49. Merriam-Webster: Abusive (2017). <https://www.merriam-webster.com/dictionary/abusive>
50. Niemann, M.: Abusiveness is non-binary: five shades of gray in German online news-comments. In: Proceedings 21st IEEE Conference Business Informatics, CBI 2019. IEEE, Moscow, Russia (2019)
51. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proceedings 25th International Conference World Wide Web, pp. 145–153, Montreal, Canada (2016)
52. Oxford University Press: Abusive (2017). <https://en.oxforddictionaries.com/definition/abusive>
53. Parliamentary Assembly: Recommendation 1805 (2007): Blasphemy, religious insults and hate speech against persons on grounds of their religion (2007)

54. Pater, J.A., Kim, M.K., Mynatt, E.D., Fiesler, C.: Characterizations of online harassment: comparing policies across social media platforms. In: Proceedings 19th International Conference Supporting Group Work, GROUP 2016, pp. 369–374. ACM Press, Sanibel Island, Florida, USA (2016)
55. Pearson: Abusive (2017). <http://www.ldoceonline.com/dictionary/abusive>
56. Poletto, F., Stranisci, M., Sanguinetti, M., Patti, V., Bosco, C.: Hate speech annotation: analysis of an Italian Twitter corpus. In: 4th Italian Conference on Computational Linguistics, CLiC-it 2017, vol. 2006, pp. 1–6. CEUR-WS (2017)
57. Raviševičius, P.: The enforcement of the primacy of the European Union law-legal doctrine and practice. *Jurisprudence* **18**(4), 1369–1388 (2011)
58. Razavi, A.H., Inkpen, D., Uritsky, S., Matwin, S.: Offensive language detection using multi-level classification. In: Farzindar, A., Kešelj, V. (eds.) AI 2010. LNCS (LNAI), vol. 6085, pp. 16–27. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-13059-5\\_5](https://doi.org/10.1007/978-3-642-13059-5_5)
59. Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., Wojatzki, M.: Measuring the reliability of hate speech annotations: the case of the European refugee crisis. In: Proceedings 3rd Workshop on Natural Language Processing for Computer-Mediated Communication, Bochum, Germany, pp. 6–9 (2016)
60. Seo, S., Cho, S.B.: Offensive sentence classification using character-level CNN and transfer learning with fake sentences. In: Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, E.S. (eds.) International Conference on Neural Information Processing, pp. 532–539. Springer, Cham (2017)
61. Solon, O.: Underpaid and overburdened: the life of a Facebook moderator (2017). <https://www.theguardian.com/news/2017/may/25/facebook-moderator-underpaid-overburdened-extreme-content>
62. Sood, S.O., Antin, J., Churchill, E.F.: Using crowdsourcing to improve profanity detection. In: AAAI Spring Symposium Series, Palo Alto, CA, USA, pp. 69–74 (2012)
63. Sood, S.O., Churchill, E.F., Antin, J.: Automatic identification of personal insults on social news sites. *J. Am. Soc. Inf. Sci. Technol.* **63**(2), 270–285 (2012)
64. Švec, A., Pikuliak, M., Šimko, M., Bielíková, M.: Improving moderation of online discussions via interpretable neural models. In: Proceedings Second Workshop on Abusive Language Online, ALW2, Brussels, Belgium (2018)
65. Tuarob, S., Mitranont, J.L.: Automatic discovery of abusive thai language usages in social networks. In: Choemprayong, S., Crestani, F., Cunningham, S.J. (eds.) ICADL 2017. LNCS, vol. 10647, pp. 267–278. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-70232-2\\_23](https://doi.org/10.1007/978-3-319-70232-2_23)
66. Warner, W., Hirschberg, J.: Detecting hate speech on the world wide web. In: Proceedings Second Workshop on Language in Social Media, Montreal, Canada, pp. 19–26 (2012)
67. Waseem, Z.: Are you a racist or Am I seeing things? Annotator influence on hate speech detection on Twitter. In: Proceedings First Workshop on NLP and Computational Social Science, Austin, Texas, USA, pp. 138–142 (2016)
68. Waseem, Z., Davidson, T., Warmesley, D., Weber, I.: Understanding abuse: a typology of abusive language detection subtasks. In: Proceedings First Workshop Abusive Language Online, Vancouver, Canada, pp. 78–84 (2017)
69. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: Proceedings NAACL Student Research Workshop, Stroudsburg, PA, USA, pp. 88–93 (2016)

70. Watanabe, H., Bouazizi, M., Ohtsuki, T.: Hate speech on Twitter: a pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access* **6**, 13825–13835 (2018)
71. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future: writing a literature review. *MIS Q.* **26**(2), xiii–xxiii (2002)
72. Yenala, H., Jhanwar, A., Chinnakotla, M.K., Goyal, J.: Deep learning for detecting inappropriate content in text. *Int. J. Data Sci. Anal.* **6**(4), 273–286 (2018)
73. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., Edwards, L.: Detection of harassment on Web 2.0. In: *Proceedings Content Analysis WEB, CAW2.0*, Madrid, Spain, pp. 1–7 (2009)