Abusive Comments in Online Media and How to Fight Them

State of the Domain and a Call to Action

Marco Niemann(✉), Jens Welsing, Dennis M. Riehle, Jens Brunk, Dennis Assenmacher, and Jörg Becker

University of Münster – ERCIS, Leonardo-Campus 3, 48149 Münster, Germany
marco.niemann@ercis.uni-muenster.de

Abstract. While abusive language in online contexts is a long-known problem, algorithmic detection and moderation support are only recently experiencing rising interest. This survey provides a structured overview of the latest academic publications in the domain. Assessed concepts include the used datasets, their language, annotation origins and quality, as well as applied machine learning approaches. It is rounded off by an assessment of meta aspects such as author collaborations and networks as well as extant funding opportunities. Despite all progress, the domain still has the potential to improve on many aspects: (international) collaboration, diversifying and increasing available datasets, careful annotations, and transparency. Furthermore, abusive language detection is a topic of high societal relevance and requires increased funding from public authorities.

Keywords: Abusive language · Comment moderation · Machine learning · Review

1 Introduction

Abusive language[1] (and especially hate speech as one of its most extreme manifestations) in online communities is becoming more and more prevalent. What has been a fringe phenomenon in the early days of the web, is now affecting the lives of millions of individuals [17,20,24]. These phenomena, which are often discussed under trivializing names such as "(hate) speech" and "(abusive) language", are not only mere inconveniences but issues that have been proven to be detrimental to the mental health of individuals and even societies at large scale [4,23]. Consequentially, these forms of inadequate communication are typically subject to legal regulations prohibiting their utterance as well as the display in

[1] We are aware that there are multiple terms and concepts, such as "abusive language", "hate speech", "offensive language", and many more. For this publication, we will use the term "abusive language", as it receives increasing acceptance in the domain (cf., the "Workshop on Abusive Language Online" conducted annually) and is sufficiently generic to account for a multitude of equally problematic types of language.

M. van Duijn et al. (Eds.): MISDOOM 2020, LNCS 12259, pp. 122–137, 2020.
https://doi.org/10.1007/978-3-030-61841-4_9

public (e.g., in Germany it is illegal to post any form of hate speech as well as to tolerate such comments on your platform once you are made aware of their existence [13]). Hence, platform operators have multiple incentives to keep their discussion spaces clean, as they otherwise risk getting sued and/or lose visitors and as a consequence traffic as well. The initial response of many outlets has been to close down their discussion spaces (in Germany up to 50% of the newspapers took this step [30]), as manual moderation results in considerable personnel cost that is not linked to any direct income [18]. Aside from the apparent issues for open societies and democracies, which arise from silencing people, journalists and media companies struggle with these radical decisions, as they limit user engagement [8] and therefore have the same detrimental potential as the abusive comments they try to avoid [18].

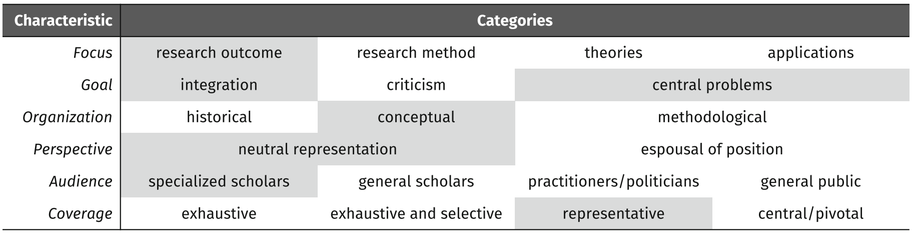
To address this issue, people from different computer science domains such as machine learning (ML) and natural language processing (NLP) have started working on (semi-)automated solutions. They are supposed to reduce the workload of journalists and community managers through both automated pre-filtering as well as moderation support (e.g., highlighting problematic parts of comments). Work is done by academics as well as practitioners, and over the last decade, a lively stream of research developed around the topic of abusive language detection. With the constant growth of the domain, it gets increasingly difficult for the individual researcher to keep track of the progress. Hence, as with any other rapidly growing domain, review papers are getting more and more important to streamline ongoing research. Prior survey and review papers typically addressed issues such as used definitions of abusive language, applied ML algorithms, conducted preprocessing, and sources for datasets [14,29]. While we will follow up on most of these aspects, we introduce several additional meta aspects such as extant author networks, funding parties, and annotator qualifications.

The remainder of this work is structured as follows: In Sect. 2, we outline our survey approach, including search and analysis strategies. The subsequent Sect. 3 is used to analyze the identified literature for aspects such as Author Analysis, Datasets and ML Approaches. As a wrap-up of this publication, we summarize our findings in Sect. 4 with a call for future action.

2 Research Approach

To understand the current state of research on computer-assisted detection of abusive speech online, it is mandatory to review the recent developments in that area carefully. The method of choice is a structured literature review according to the principles of Webster and Watson [33] and vom Brocke et al. [7]. According to the taxonomy of Cooper [10] (see Fig. 1), we focus on the synthesis and integration of central findings and problems of the abusive language detection domain. Given the conference format, our coverage is representative in nature and mainly addresses the scholars active in the domain.

To conduct the search, we composed the search string depicted below. It combines the central concept of our survey (“abusive language”) with the often



Characteristic	Categories			
Focus	research outcome	research method	theories	applications
Goal	integration	criticism	central problems	
Organization	historical	conceptual	methodological	
Perspective	neutral representation		espousal of position	
Audience	specialized scholars	general scholars	practitioners/politicians	general public
Coverage	exhaustive	exhaustive and selective	representative	central/pivotal

Fig. 1. Categorization of our research in the taxonomy of [10]

synonymously used concept of "hate speech". As we want to focus on works that propose novel approaches to tackle the problem of abusive language online, we add the keywords "detection" and "classification" to limit results to papers working on (semi-)automated solutions. To avoid duplicating work already conducted by, e.g., Fortuna et al. [15], we restrict the search to the years 2018 and 2019 by using the following search criterion:

("abusive language" OR "hate speech") AND ("detection" OR "classification") AND PUBYEAR $\in \{2018, 2019\}$

As search engines, we considered *Scopus*, *Microsoft Academic*, and the *Web of Science*, which have been found to excel through broad coverage in general and for the topic at hand. *Google Scholar* was excluded as despite its massive portfolio it still lacks a lot of search and filter functionality plus its Google heritage subjects it to the Google algorithm which performs context-specific optimizations and hence does not allow reproducibility [5]. Beyond this initial search, we additionally conducted a forward- and backward-search on works standing out through their constant appearance in the papers we found using a structured search.

3 State of the Domain

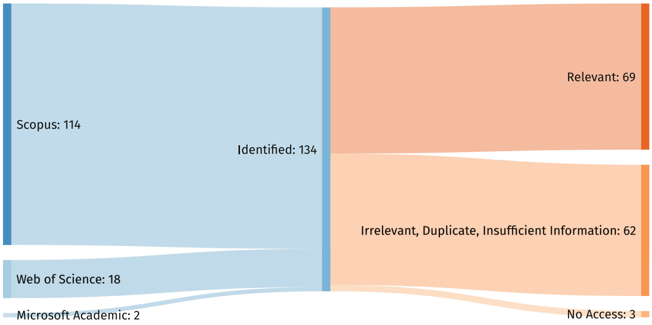
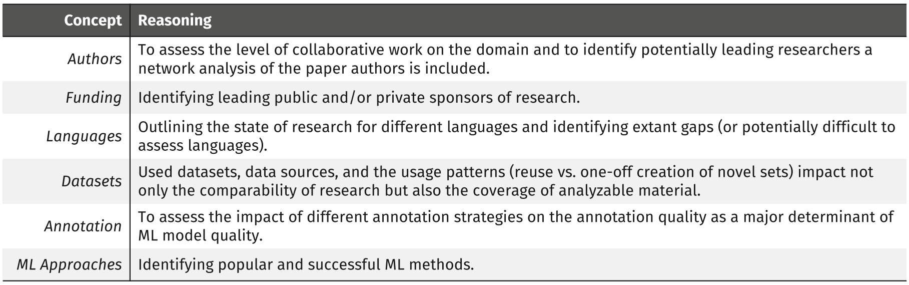


Fig. 2. Literature search results breakdown

As depicted in Fig. 2, our structured literature search returned 134 results, of which 69 were identified as relevant for our research. 62 publications were excluded as they are either duplicates or without an individual approach to automated abusive language detection (i.e., other literature reviews, . . .).

The traditional approach of Webster and Watson [33] and vom Brocke et al. [7] suggests the use of concept matrices to assess the identified material. Figure 3 presents the concepts evaluated within this chapter and the reasoning for their selection. Based on the chosen concepts, we decided against using traditional table-based concept matrices and will instead make use of more visual means as, e.g., networks are hard to depict in text-based formats.



Concept	Reasoning
Authors	To assess the level of collaborative work on the domain and to identify potentially leading researchers a network analysis of the paper authors is included.
Funding	Identifying leading public and/or private sponsors of research.
Languages	Outlining the state of research for different languages and identifying extant gaps (or potentially difficult to assess languages).
Datasets	Used datasets, data sources, and the usage patterns (reuse vs. one-off creation of novel sets) impact not only the comparability of research but also the coverage of analyzable material.
Annotation	To assess the impact of different annotation strategies on the annotation quality as a major determinant of ML model quality.
ML Approaches	Identifying popular and successful ML methods.

Fig. 3. Selected concepts for analysis

3.1 Author Analysis

To conduct the analysis, we used the VOSviewer [12], which is one of the standard tools for the creation and visualization of bibliographic networks. At first, we created a map for the overall authorscape of the identified publications, which is depicted in Fig. 4.

Looking at Fig. 4, the most striking feature is the abundance of mostly disconnected author groups. The only larger cluster can be found at the center of the figure and revolves around the four Italian authors Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. Taking a closer look at the cluster (see Fig. 5) and the associated publications, the underlying bond can be identified as the "Evalita 2018 Hate Speech Detection Task" (EVALITA). Most of the collaborating authors are—as to be expected—of Italian origin, respectively, working for Italian research institutions [6]. Nevertheless, the proceedings also contain several international authors from countries with similar languages [14], respectively, several who contributed through mixed tasks (detection of English and Italian) [16]. Since the only two remaining larger clusters only contain Italian[2] respectively Indonesian authors, this is indeed the only identified occasion of large-scale international collaboration and knowledge exchange.

[2] The authors are interestingly not linked to the other Italian author cluster, indicating a somewhat lacking national collaboration.

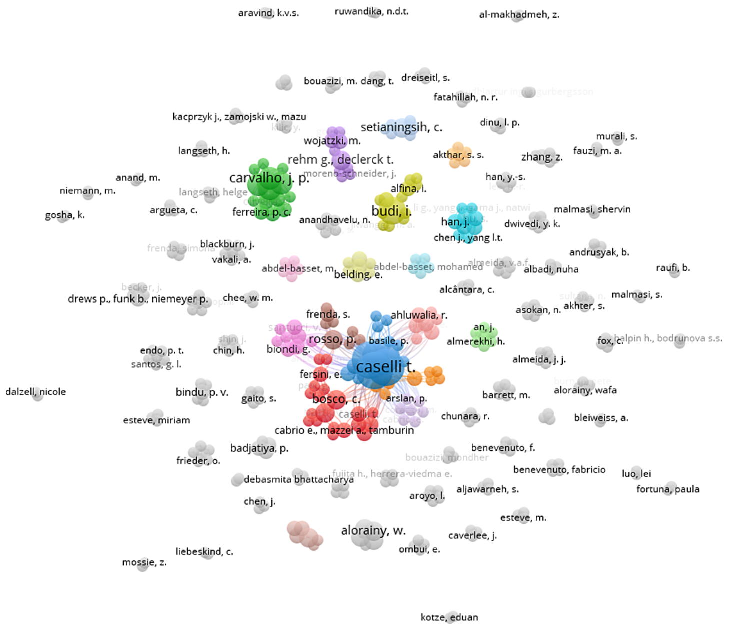


Fig. 4. VOSviewer distance-based label view of the co-author relationships. Node size depends on the number of co-authors. Distance (and edges) between nodes indicate the strength of co-author relationships; colors the automatically identified clusters [12]. (Color figure online)

Given the prevalence of the issue of abusive language and the maturing stage of the domain, it is surprising to observe this level of disconnection. Even though partially separated by language, the structural approaches and issues in the domain are typically shared and should consequently provide ample opportunity for collaboration.

3.2 Funding

Given the massive social, legal, and economic impacts of abusive language, it is receiving attention from both public bodies as well as private institutions. Hence, it is interesting to see whether this interest mirrors into corresponding funding programs and whether certain institutions are taking significant influence on the research carried out.

Out of the assessed 69 publications, 38 (approx. 55%) state that they received some form of funding/financial support. Despite the unambiguous relevance for media companies, only one of these 38 publications officially received support from a private entity: Fortuna et al. [14] have been supported by Google's DNI grant. The remaining 37 funded publications received public money from various ministries or societies. Amongst these, no dominant organization could be

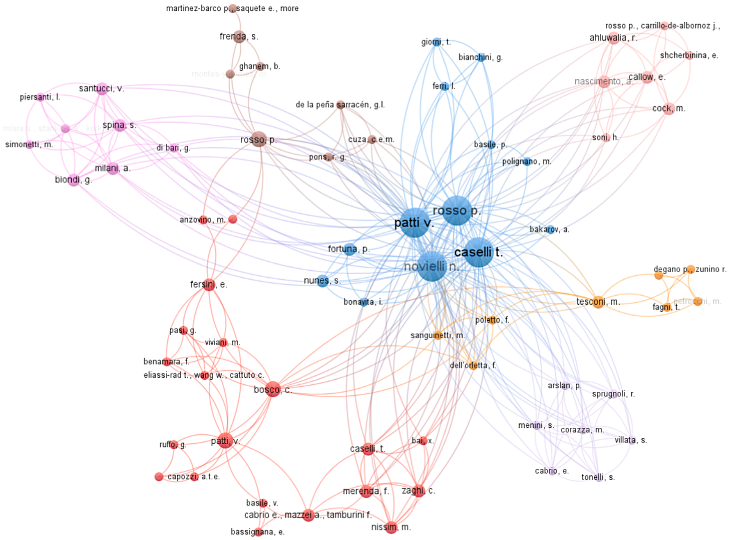


Fig. 5. Literature search results breakdown

identified as only a few sponsors are linked to more than one publication. Interestingly the "Directorate Research and Community Services of the Universitas Indonesia", a Spanish, and a Portuguese ministry, are amongst the most stated sponsors (cf., Fig. 6). This is interesting, since Indonesian, Spanish, Portuguese have not been among the most assessed languages in the domain so far. The EU—while heavily investing in AI and related technology—is only supporting three publications through Horizon 2020 grants.

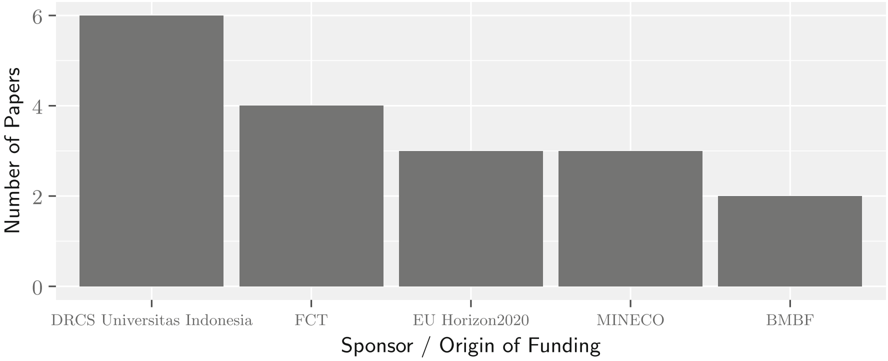


Fig. 6. Most influential sponsors of abusive language detection research

The assessed data only represents a limited snapshot of the funding situation; however, it can be acknowledged that the issue of abusive language online is having a sufficient societal impact to receive substantial public funding. So far, no public or private organization appears to be taking any noticeable influence.

3.3 Languages

Undoubtedly, languages differ a lot in their vocabulary, grammar, usage, and at times even in their alphabet. Consequently, it is unlikely to see an OSFA (one size fits all) solution to abusive language detection for all languages or even for language families.

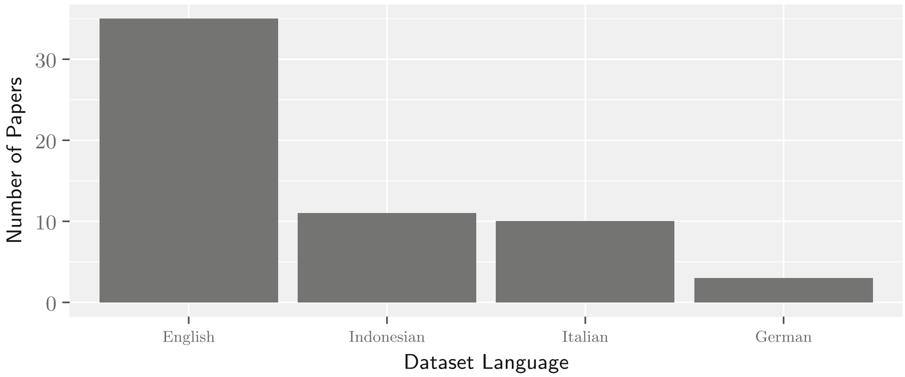


Fig. 7. Distribution of publications among most common languages

Amongst the 69 assessed publications—as in the majority of extant papers—English is the predominant working language solutions are developed for (50.72% = 35 papers; cf., Fig. 7). Considering that English is the most common language online [31] and that many of the leading universities for ML and NLP (e.g., Stanford) are located in the US, this comes as no surprise. However, differing from the early days of abusive language detection research, approx. 50% of the publications are already working on different languages—which again indicates the globality of the problem as well as the increasing promise of ML and NLP given their spread to less common and often more complex languages (e.g., German; cf. [25]).

As all the assessed research aims at finding automated solutions to uncover abusive language, we are furthermore interested in potential differences regarding the quality of those solutions. Conducting an in-depth assessment of the approaches and the underlying linguistic structure of the languages would be beyond the scope of this review. Instead, we focus on the achieved $F1$-score.

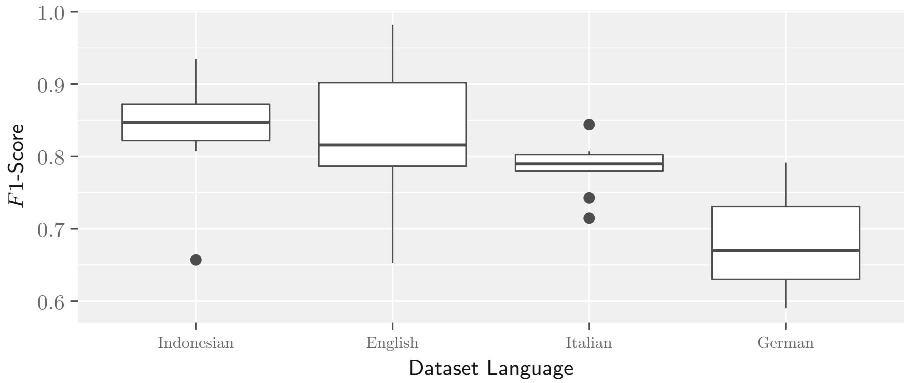


Fig. 8. $F1$-scores achieved for different languages

As a proxy, it captures both the derived automated solution's quality and the "simplicity" of the underlying language. In Fig. 8, we plotted the $F1$-scores of the most commonly assessed languages—and surprisingly, the so far rather under-researched, Indonesian turns out to perform best. English as the most popular language has a slightly worse median performance; however, the variance is higher with individual publications reaching $F1$-scores of up to 0.9 [27,35]. The performance of abusive language detection for German is substantially worse, with a median of only 0.675 and a maximum of under 0.8. This is in line with the observations of individual authors, who observed that German tends to be comparatively complex to analyze and work with [25]. A final interesting observation can be made considering the performance of Italian: The majority of the publications there worked on the ELVITA dataset and stand out through their low variance. This is an indication that not only the language itself makes a difference but also the dataset and the type of language used.

3.4 Datasets

The traditional approach towards the detection of abusive language through ML is based on the sub-class of supervised machine-learning. Training and performance of this class of algorithms substantially depend on the kind and quality of the employed training data [2]. Hence, we take a look at both data sources as well as datasets[3] used in recent publications.

The first observation that can be made is that the majority of the datasets used are still curated based on Twitter data (see Fig. 9; in our sample in 40 papers out of 69 use datasets made up of Tweets), as already observed by prior reviews [15]. Even though most data in the web is more or less freely accessible, Twitter offers a comparatively easy to use and unrestricted API making it rather

[3] Each recombined dataset (combined of $n \geq 2$ already existent ones) is considered a novel dataset of its own right.

promising for creating larger collections of data—plus most of the texts are similar in structure (prev. limit to 140 characters). All other social networks typically exercise more rigid control over their user-generated content (UGC), making them only minor sources of research data until today. The same holds for online newspapers.

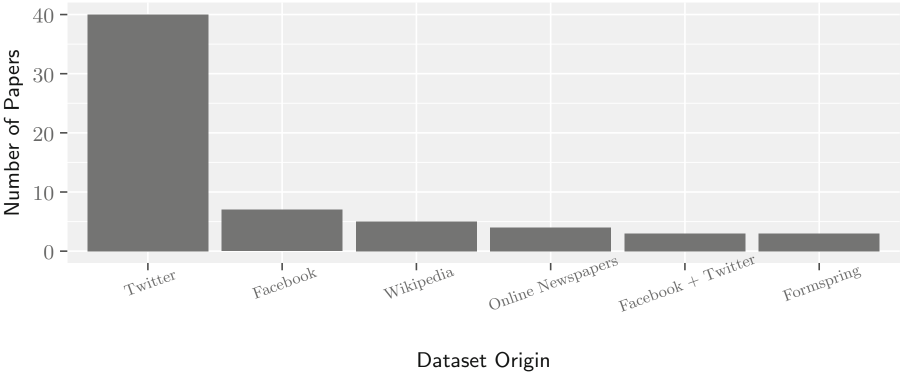


Fig. 9. Most common dataset origins (occurrences >1)

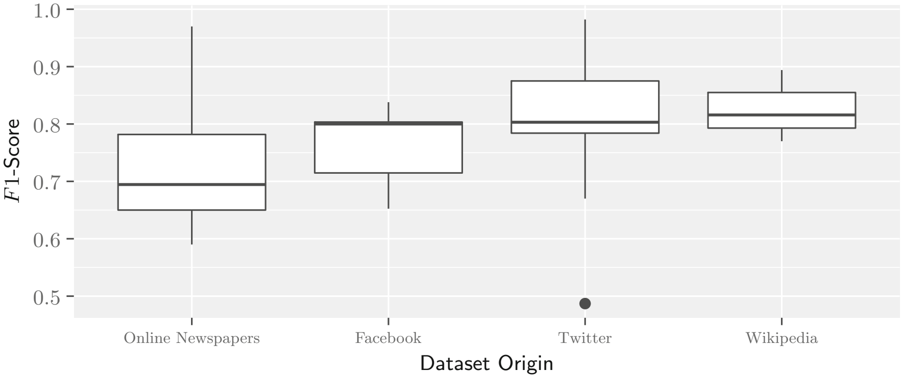


Fig. 10. $F1$-scores for common dataset origins

Assessing the data origins further and comparing the classification results which are achieved based on them, especially Twitter but also Wikipedia stand out (see Fig. 10): In both cases, the best performing half of the publications performs better than at least 75% of the publications using sources such as Facebook or online newspapers. Overall, it appears to be most complicated to correctly classify comments from online newspapers, which is the only source with a median $F1$-score being lower than 0.8 (with $F1 = 0.69$). However, given the smaller sample ($n = 4$), these results are to be treated with caution.

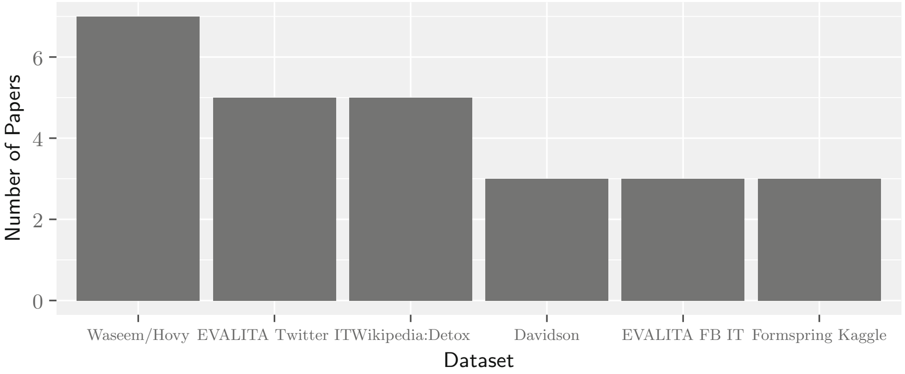


Fig. 11. Distribution of datasets across publications (occurrences >1)

Differing from prior reviews such as the one by Fortuna and Nunes [15] (one reused dataset in 17 assessed publications; 6%), the share of research reusing published datasets has been rising to ∼56% (39 out of 69) so that only 28 publications use a novel, self-created dataset. Three of the largest datasets (the ones by [32,34], and [11]; cf. Fig.11) are already 2–3 years old, which corresponds to classical academic publishing cycles and might explain their slow uptake. While this development is good for comparability between different publications, a certain amount of scrutiny should be kept. Considering that, e.g., Twitter data is not representative for UGC on the web, an excess reliance on these kinds of datasets might turn out to be problematic—especially considering that other data sources appear harder to work with (cf. above and Fig. 10).

3.5 Annotation

Another side-effect that comes with the use of supervised machine-learning is the necessity to use so-called labeled or annotated ground-truth data: To train suitable classifiers, it is insufficient to use raw comments. Instead, for each comment, an annotation indicating its type (e.g., abusive vs. clean) is needed. However, comments are not labeled by their original creators; hence, this task is typically done by a third party—often crowd workers, researchers, or student assistants. As such, annotations are subjective and laden with emotions [21], obtaining an agreed-upon annotation is complex with potential repercussions to the final classification [26].

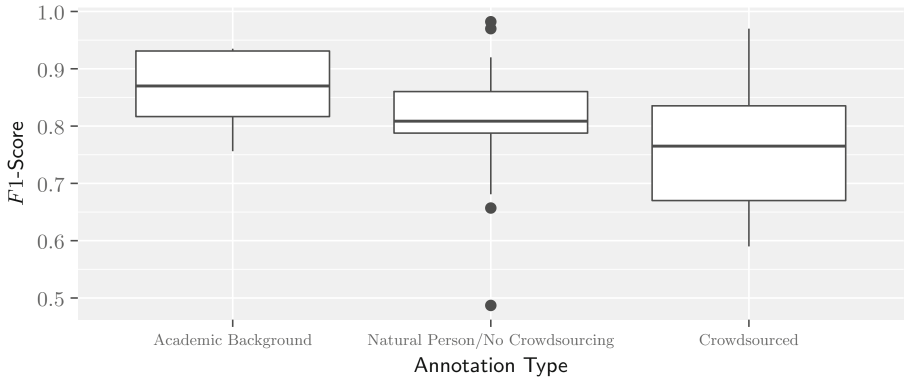


Fig. 12. $F1$-scores for different annotation types

Given the high investments (of time or money) in annotated datasets, it is vital to understand which investment delivers the best value. In Fig. 12 we depict the $F1$-scores reached with datasets annotated by a) people with a stated academic background, b) crowd workers, and c) natural persons without further specification of their educational background. The primary finding from this assessment is that a controlled group of annotators with a scientific background delivers more usable annotations than crowd workers from diverse backgrounds[4]. However, the larger variance on the crowdsourced datasets indicates that they hold potential: For example, Albadi et al. [1] used upfront quizzes and continuous control questions to ensure that each partaking crowd worker had a correct understanding of the concepts in question.

Another metric that is usually put under scrutiny is the agreement between annotators or inter-rater agreement[5], which is an indication of annotation reliability given coders' shared understanding of the labels [3]. For the assessed 69 papers, we found varying agreement scores, which, however, did not show any substantial correlation with the final $F1$-scores ($\rho = 0.2088$). This is in line with observations made by Koltsova [21]. Consequently, alternative approaches to label data (e.g., using multiple labels per comment) (cf. [25]) might help to overcome the inherent subjectivity of traditional approaches while improving classification quality.

3.6 ML Approaches

One of the most diverse areas is the selection of classification algorithms. As many of the extant approaches such as recurrent neural networks (RNNs) or decision trees (DTs) exist in various flavors and configurations, we decided to

[4] Datasets with all annotation strategies would have to be subjected to a testbed of multiple classifiers to ensure the improved performance is due to the chosen strategy.

[5] For commonly used measures such as Krippendorff's alpha, this should be around 0.8 or higher [22]. However, other measures can have different scales.

group algorithmically similar approaches ending up with six larger groups: SVM, Ensembles of Classifiers, RNNs (incl. LSTMs, GRUs, ...), DTs (incl. random forests), Logistic Regression (LR), and vanilla Neural Networks (NNs).

Despite their extensive coverage in public media [28], NNs are not the most common approach to classify comments in our selection (cf., Fig. 13). Instead, SVMs are still taking the first place as the most used algorithmic approach, closely followed by the classifier ensembles. Only in places 3 and 6, we can find RNNs and NNs, split by DTs and LR-based solutions.

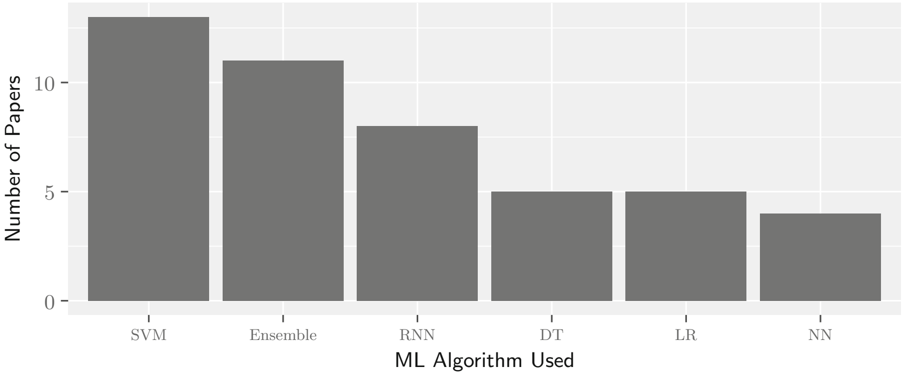


Fig. 13. Distribution of ML approaches

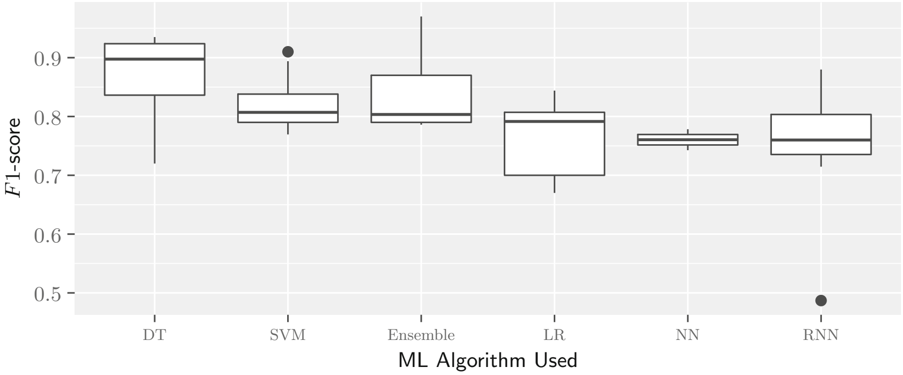


Fig. 14. $F1$-scores for different ML approaches

Based on the performance assessment plotted in Fig. 14, there are two algorithmic winners: The first is the DTs with a median $F1$-score of approx. 0.9 (with the limitation of being computed based on only four publications). This is somewhat surprising, as DTs can be considered as a promising approach to classification, as they avoid the often introduced algorithmic black-box. The second winning group is the classifier ensembles, which have a slightly reduced median performance, but the best overall performance. From a theoretical perspective, this should not be surprising since ensemble classifiers have the benefit of combining multiple algorithms simultaneously, giving them the ability to smoothen potential weaknesses of individual algorithms [2,19].

So far, there appears to be no set of dominating algorithmic approaches. The vivid, ongoing experimentation, however, indicates that classic ML algorithms and ensembles are especially promising. Impact factors such as used features or dataset sizes were not considered for this analysis.

4 Conclusion and Call to Action

On the preceding pages, we presented a short overview of recent developments in the area of (semi-)automated detection of abusive language in online spaces. We found that currently, there is still very collaborative work: There are heterogeneous micro-clusters of authors who, however, rarely work with people from different clusters, which limits knowledge exchange. One of the few exceptions has been an Italian competition bringing together a larger group of experts. From the language perspective, the research horizon is broadening, increasingly including languages other than English. Furthermore, we found additional evidence that languages differ in complexity and hence require different levels of investigation. Public funding is only applied (or at least acknowledged) sparsely, also from administrations claiming to work heavily on AI-supported systems (e.g., Germany and the EU). Regarding the data perspective, researchers continue to use Twitter data primarily. While classification turns out to work comparatively well on this kind of data, other more diverse types such as comments from online newspapers still require additional work. Related to the quality of training and test data, we found that the agreement on comment labels might not be the quality-determining factor while selecting competent, educated, and well-briefed annotators has a higher impact. For the final classification, traditional ML approaches are still the most common ones, with inherently transparent models like DTs outperforming complex black box models. Be aware that these results have a propositional state, not considering, e.g., potential confounding factors such as the influence of annotations on the language's $F1$-score.

To move beyond a purely retrospective view on the domain, we want to propose and discuss a set of steps that should help to advance the domain:

- Inclusion of competitions (like the ELVITA one) or paper-a-thons on domain conferences to further (international) collaboration and exchange. This would help to transfer extant knowledge to junior researchers as well as those working on so far under-researched languages.
- Additional focus on complex languages (such as German).
- Publish more datasets from diverse sources (other than Twitter). This goes beyond the data-related aspects such as collection, storage, curation, and release, but has a legal (adjust copyrights to enable releases for research)

and social (create awareness researchers are not stealing property but reusing published items to reduce abusive content) component. As a starting point, all publicly funded research that generates new datasets should be obliged to release the data.

- Ensure careful annotation. This ranges from the provision of proper labels to the thorough annotation through professional annotators. Furthermore, the awareness for accurate annotation should be risen, as they are the baseline for “algorithmic moderators” (to avoid issues such as censorship claims).
- Incentivize further research on models that trade-off between predictive quality (already comparatively high) and transparency. Academic outlets should also start assessing algorithmic transparency as one criterion of eligibility for publishing to avoid a shift towards highly optimized black box models.
- Increase the amount of public funding for abusive language research. SME media companies otherwise lack the resources to gain access to competitive ML-assisted moderating solutions (cf. [9]). Beyond that, many of the above-stated goals also require a substantial financial commitment that is otherwise hardly bearable for public research institutions.

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: MODERAT! (No. CM-2-2-036a).

References

1. Albadi, N., Kurdi, M., Mishra, S.: Investigating the effect of combining GRU neural networks with handcrafted features for religious hatred detection on Arabic Twitter space. Soc. Netw. Anal. Min. **9**(1), 1–19 (2019). https://doi.org/10.1007/s13278-019-0587-5
2. Alpaydin, E.: Introduction to Machine Learning, 3rd edn. The MIT Press, Cambridge (2014)
3. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Comput. Linguist. **34**(4), 555–596 (2008)
4. Benesch, S.: Countering dangerous speech to prevent mass violence during Kenya’s 2013 elections. Tech. rep., Dangerous Speech Project (2013)
5. Boeker, M., Vach, W., Motschall, E.: Google scholar as replacement for systematic literature searches: good relative recall and precision are not enough. BMC Med. Res. Methodol. **13**(1), 131 (2013)
6. Bosco, C., Sanguinetti, M., Dell’Orletta, F., Poletto, F., Tesconi, M.: Overview of the EVALITA 2018 hate speech detection task. In: Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, EVALITA 2018, Turin, Italy, pp. 1–8 (2018)
7. 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 of 17th European Conference on Information Systems, ECIS 2009, Verona, Italy, pp. 2206–2217 (2009)
8. Brunk, J., Mattern, J., Riehle, D.M.: Effect of Transparency and Trust on Acceptance of Automatic Online Comment Moderation Systems. In: Proceedings of 21st IEEE Conference on Business Informatics, Moscow, Russia, pp. 429–435 (2019)

9. Brunk, J., Niemann, M., Riehle, D.M.: Can analytics as a service save the online discussion culture? - The case of comment moderation in the media industry. In: Proceedings of 21st IEEE Conference on Business Informatics, CBI 2019, Moscow, Russia, pp. 472–481 (2019)
10. Cooper, H.M.: Organizing knowledge syntheses: a taxonomy of literature reviews. Knowl. Soc. **1**(1), 104–126 (1988)
11. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of Eleventh International Conference on Web and Social Media, ICWSM-2017, Montreal, Canada, pp. 512–515 (2017)
12. van Eck, N.J., Waltman, L.: Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics **84**(2), 523–538 (2010)
13. Ellis, J.: What happened after 7 news sites got rid of reader comments (2015). https://www.niemanlab.org/2015/09/what-happened-after-7-news-sites-got-rid-of-reader-comments/
14. Fortuna, P., Bonavita, I., Nunes, S.: Merging datasets for hate speech classification in Italian. In: Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, EVALITA 2018, Turin, Italy, pp. 1–6 (2018)
15. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput. Surv. **51**(4), 1–30 (2018)
16. Frenda, S., Ghanem, B., Guzmán-Falcón, E., Montes-y Gómez, M., Villaseñor-Pineda, L.: Automatic expansion of lexicons for multilingual misogyny detection. In: Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, EVALITA 2018, Turin, Italy, pp. 1–6 (2018)
17. 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
18. Green, M.: No comment! why more news sites are dumping their comment sections (2018). https://www.kqed.org/lowdown/29720/no-comment-why-a-growing-number-of-news-sites-are-dumping-their-comment-sections
19. Hastie, T., Tibshirani, R., Friedman, J.: Additive models, trees, and related methods. The Elements of Statistical Learning. SSS, pp. 295–336. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7_9
20. Hine, G.E., et al.: Kek, cucks, and god emperor trump: a measurement study of 4chan's politically incorrect forum and its effects on the web. In: Proceedings of 11th International Conference on Web and Social Media, ICWSM-2017, Montral, Canada, pp. 92–101 (2017)
21. Koltsova, O.: Methodological challenges for detecting interethnic hostility on social media. In: Bodrunova, S.S., et al. (eds.) INSCI 2018. LNCS, vol. 11551, pp. 7–18. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-17705-8_1
22. Krippendorff, K.: Content Analysis: An Introduction to its Methodology, 2nd edn. SAGE Publications Inc., Thousand Oaks (2004)
23. Langham, J., Gosha, K.: The classification of aggressive dialogue in social media platforms. In: Proceedings of 2018 ACM SIGMIS Conference on Computers and People Research, SIGMIS-CPR 2018, Buffalo-Niagara Falls, NY, USA, pp. 60–63 (2018)
24. Mansfield, M.: How we analysed 70m comments on the guardian website (2016). https://www.theguardian.com/technology/2016/apr/12/how-we-analysed-70m-comments-guardian-website

25. Niemann, M.: Abusiveness is non-binary: five shades of gray in German online news-comments. In: Proceedings of 21st IEEE Conference on Business Informatics, CBI 2019, Moscow, Russia, pp. 11–20 (2019)
26. Niemann, M., Riehle, D.M., Brunk, J., Becker, J.: What is abusive language? In: Grimme, C., Preuss, M., Takes, F.W., Waldherr, A. (eds.) MISDOOM 2019. LNCS, vol. 12021, pp. 59–73. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-39627-5_6
27. Pitsilis, G.K., Ramampiaro, H., Langseth, H.: Effective hate-speech detection in Twitter data using recurrent neural networks. Appl. Intell. **48**(12), 4730–4742 (2018). https://doi.org/10.1007/s10489-018-1242-y
28. Sample, I.: 'It's able to create knowledge itself': Google unveils AI that learns on its own (2017). https://www.theguardian.com/science/2017/oct/18/its-able-to-create-knowledge-itself-google-unveils-ai-learns-all-on-its-own
29. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, SocialNLP 2017, Valencia, Spain, pp. 1–10 (2017)
30. Siegert, S.: Nahezu jede zweite Zeitungsredaktion schränkt Online-Kommentare ein (2016). http://www.journalist.de/aktuelles/meldungen/journalist-umfrage-nahezu-jede-2-zeitungsredaktion-schraenkt-onlinekommentare-ein.html
31. W3Techs: Historical trends in the usage statistics of content languages for websites, February 2020. https://w3techs.com/technologies/history_overview/content_language
32. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: Proceedings of NAACL Student Research Workshop, Stroudsburg, PA, USA, pp. 88–93 (2016)
33. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future: writing a literature review. MIS Q. **26**(2), xiii–xxiii (2002)
34. Wulczyn, E., Thain, N., Dixon, L.: Ex machina. In: Proceedings of 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, pp. 1391–1399 (2017)
35. Zhang, Z., Luo, L.: Hate speech detection: a solved problem? The challenging case of long tail on Twitter. Semant. Web **10**(5), 925–945 (2019)