# CYLENCE: Strategies and Tools for Cross-Media Reporting, Detection, and Treatment of Cyberbullying and Hatespeech in Law Enforcement Agencies

Marc-André Kaufhold
Markus Bayer
Julian Bäumler
Christian Reuter
Technical University of Darmstadt, Science and Technology for Peace and Security (PEASEC)
Darmstadt, Germany
{kaufhold,bayer,reuter}@peasec.tu-darmstadt.de
julian.baeumler@stud.tu-darmstadt.de

Stefan Stieglitz
Ali Sercan Basyurt
University of Potsdam, Information Systems and Digital Transformation (DIGICAT)
Potsdam, Germany
stefan.stieglitz@uni-potsdam.de
ali.basyurt@uni-potsdam.de

Milad Mirbabaie
University of Paderborn, Information Systems and Digital Society (DISCIETY)
Paderborn, Germany
milad.mirbabaie@uni-paderborn.de

Christoph Fuchß
Kaan Eyilmez
Virtimo AG
Berlin, Germany
{fuchss,kaan.eyilmez}@virtimo.de

## ABSTRACT

Despite the merits of public and social media in private and professional spaces, citizens and professionals are increasingly exposed to cyberabuse, such as cyberbullying and hate speech. Thus, Law Enforcement Agencies (LEA) are deployed in many countries and organisations to enhance the preventive and reactive capabilities against cyberabuse. However, their tasks are getting more complex by the increasing amount and varying quality of information disseminated into public channels. Adopting the perspectives of Crisis Informatics and safety-critical Human-Computer Interaction (HCI) and based on both a narrative literature review and group discussions, this paper first outlines the research agenda of the CYLENCE project, which seeks to design strategies and tools for cross-media reporting, detection, and treatment of cyberbullying and hatespeech in investigative and law enforcement agencies. Second, it identifies and elaborates seven research challenges with regard to the monitoring, analysis and communication of cyberabuse in LEAs, which serve as a starting point for in-depth research within the project.

## KEYWORDS

cyberbullying, hate speech, law enforcement agencies, situational awareness, human-computer interaction


*Veröffentlicht durch die Gesellschaft für Informatik e.V.*
*in P. Fröhlich & V. Cobus (Hrsg.):*
*Mensch und Computer 2023 – Workshopband, 03.-06. September 2023, Rapperswil (SG)*

https://doi.org/10.18420/muc2023-mci-ws01-211

## 1 INTRODUCTION

In the last 20 years, social media has not only established itself as an integral part of everyday social life [85], but also as a platform for exchange and a source of information in acute crisis situations arising from real or virtual space [81, 83]. Despite these potentials, abuse phenomena also increasingly arise from digital space, including cyberbullying and hate speech. Cyberbullying means "insulting, threatening, exposing or harassing people using communication media, such as smartphones, emails, websites, forums, chats and communities" [7]. While cyberbullying is mostly directed against individuals, hate speech usually refers to groups of people. According to the European Commission against Racism and Intolerance [26], hate speech includes all forms of expression that denigrate, belittle, insult, stigmatise, threaten or attack people or groups of people on the basis of perceived group-related characteristics and status characteristics attributed to them. Against the background of an increasingly complex information space, special framework conditions arise with regard to civil security.

Although the internet has now produced a variety of cyber-abuse awareness, reporting and prevention campaigns for end-users, the resulting information is not integrated into everyday applications and technologies that can be considered multipliers for the reach of this information. For example, recent studies suggest that providing awareness and prevention information via mobile information and alert apps (e.g., NINA) is desired by the German population [38, 53]. Supporting these communication and prevention activities requires establishing professional analysis strategies for LEAs to strengthen public communication on how to deal with cyber abuse.

In addition, LEAs face the challenge of obtaining an accurate overview of the situation regarding the spread of cyberbullying and hate speech. While they do set up reporting points for hate speech (e.g., in the context of the campaign "Hessen gegen Hetze"), only a

part of them can be captured by active reporting through the population. However, the challenge facing the additional monitoring of (partially) public data sources by LEAs is that the number of sources that require analysis and verification, such as blogs, feeds, photo and video portals, social networks, and websites, is significantly increasing. This is compounded by the targeted publication of abusive content by automated botnets, which adds to the workload of the analysis of manually distributed content. This requires the development of innovative machine detection algorithms that are integrated into novel detection, reporting and visualisation tools.

This paper sets up the research agenda of the CYLENCE project and seeks to identify and elaborate these challenges, such as the gathering and analysis of a multitude of confusing information in complex cyberabuse situations, inter-organisational collaboration with other LEAs for effective incident management, actor-specific communication to affected stakeholders as well as the protection of sensitive data and compliance with data protection regulations. First, it will present related work (Section 2) and the method (Section 3) of this paper. Second, it will present the research agenda of the CYLENCE project including goals, the security scenario and planned innovations (Section 4). Thereafter, the paper discusses the identified research challenges (Section 5) and finishes with a short conclusion (Section 6).

## 2 RELATED WORK

In order to inform the scope of our research project, we reviewed existing strategies in Germany as well as technologies for the reporting, detection, analysis, and visualization of cyberbullying and hate speech. In the following, we provide a short overview of the initial findings.

### 2.1 Strategies for Dealing With Cyberbullying and Hate Speech

The literature on dealing with hate speech includes the strategies of education and deletion [20]. According to this literature, educational measures can help raising citizens' awareness, offering support and developing creative solutions against hate speech. The information portal DAS NETTZ, for example, sees itself as a networking centre against hate speech and offers a search for initiatives from German-speaking countries that can be filtered by topics such as de-escalation, counter-speech, support or reporting hate speech. In the research field about fake news and misinformation, various technologies have been developed to tackle the issue. For instance, the smartphone app "Fake News Check" enables students to answer topic-specific questions and obtain tips on how to manage misinformation [24]. Similarly, the browser plugin TrustyTweet highlights Twitter messages as potential fake news using specific indicators [40]. CYLENCE will test the extent to which a plug-in can be used for the recognition and reporting of cyberbullying and hate speech.

On a legal level, the removal of hate speech in Germany is primarily defined by the NetzDG, which requires social network operators to remove or block "obviously illegal content within 24 hours" of receiving a complaint (§ 3 Abs. 2 Nr. 2 NetzDG). As part of the "Hessen gegen Hetze" (Hesse against hate) initiative, the state has established a reporting office for citizens. This office serves to provide counselling and support services to those affected by hate comments, while also forwarding these comments to platform operators with the aim of "promptly removing hate speech from public perception" [23]. The voluntary initiative Hassmelden (Hate Reporting) also sees itself as the first and only central reporting office for hate speech, which also offers a smartphone app for reporting hate speech [41]. In the research field of crisis informatics, it has been shown that German citizens prefer installing a warning app that provides police information (e.g., fraud offences, cybercrime, behavioural and preventive offers) in addition to classic disaster warnings [53]. Thus, CYLENCE aims for the integration of a reporting function for cyberbullying and hate speech in the warning app hessenWARN.

### 2.2 Technologies for Detection and Visualization of a Situation Picture

In principle, many algorithms have already been tested and datasets published that enable automatic detection of cyberbullying [28, 86] and hate speech [32, 79] in social media using AI, especially artificial neural networks. Current research suggests that classification quality can be improved by using large language models [19] (e.g., GPT-3 [11]). Flexibility can also be improved by adapting those models with Few-Shot Learning, i.e., using a small domain-specific training data set. As quantity and quality of data become increasingly important to further improve the classification quality of models [3, 84], the research area of data augmentation investigates the artificial generation of training data [31]. However, uncritical data annotation and model building can lead to cyberbullying [36] and hate speech [69, 89] detection algorithms actually reinforcing social biases [97]. Nonetheless, research shows that interpolation-based approaches can mitigate this effect [18, 93]. For this, it is essential that users can understand the decisions made by the algorithm. The use of model-agnostic white-box approaches (e.g., LIME [82], SHAP [62]) seems promising to explain and visualise these decisions. CYLENCE therefore investigates to what extent an adaptable, fair and comprehensible detection of cyberbullying and hate speech can be realised by the innovative combination of few-short learning, data augmentation and whitebox approaches.

After the classification of the data, an appealing and target-oriented visualisation of the situation is still required in order to establish appropriate situational awareness and to support the decision-making based on it [27, 107]. The sheer amount of data, also called Big Social Data [72], that is generated in everyday life and during major events across platforms, for example on Facebook, Telegram or Twitter, can lead to information overload, which implies that technical support solutions must have very good usability as well as configurable filter mechanisms and classifiers in order to reduce the amount of data [54]. While crisis informatics has already explored a variety of interactive interfaces for the collection and analysis of public data for crisis management [52, 73], there are only a few research approaches for the visualisation of cyberbullying [63] and hate speech [12, 76], which are not tailored to the requirements and needs of LEAs. In CYLENCE, the insights of crisis informatics will be combined with the domain requirements of LEAs to support the detection and analysis of, as well as communication about, cyberbullying and hate speech through a customised interface.

# 3 METHOD

We conducted a narrative literature review and group discussions to develop the research agenda of the CYLENCE project and identify research challenges for developing LEA-focused strategies and technologies for the reporting, detection, and treatment of cyberbullying and hatespeech. First, narrative literature reviews aim to summarise prior knowledge, address a broad scope of questions or topics, usually deploy a selective search strategy and integrate both conceptual and empirical work [75]. We used Google Scholar to search for domain-specific (i.e. crisis informatics, cyber situational awareness and cyber threat communication) and method-specific (i.e. supervised machine learning, visual analytics and technology assessment) literature, focussing on method applications within the present domain. Second, following the search process, we conducted multiple group discussions among the authors. These discussions did not follow a predefined structure but were designed to achieve a first sketch of the research idea and then iteratively integrate and revise our findings into the final research agenda (Section 4) and distinct research challenges (Section 5). Both results serve as a starting point for more comprehensive and rigor research within the three-year CYLENCE project, including systematic literature reviews, qualitative and quantitative empirical research, design science research, usability and user experience research as well as technology assessment.

# 4 RESULTS I: THE RESEARCH AGENDA

The aim of CYLENCE is to develop strategies and tools for cross-media reporting, detection and treatment of cyberbullying and hate speech. To this end, organisational strategies and tools for collecting and analysing (partially) public, social data sources (e.g., Facebook, Telegram, Twitter) based on a participatory development process shall be used to enable LEA to improve early detection and treatment of cyber abuse cases. A training strategy geared towards this will be complemented by an interactive tutorial for learning the use of the developed tools, which will use Artificial Intelligence (AI) and Visual Analytics (VA) to support customisable, fair and traceable AI detection and real-time-based dashboard processing of cyber-abuse content. To enhance civilian security, the detection and reporting of cyberbullying and hate speech by the public shall be further strengthened. This includes a strategy to improve communication between citizens, victims and LEA, which will be supported by empirical field research (e.g., representative surveys) and tested within the framework of the campaign "Hessen gegen Hetze". For this purpose, a browser plug-in and the extension of the smartphone app hessenWARN will be conceptualized. These tools are designed to detect and report instances of cyber abuse.

## 4.1 Security Scenario

According to a comparative study by the Bündnis gegen Cybermobbing e.V. [4], around 11.5% of people in Germany were affected by cyberbullying in 2021. While slightly more than 53% of cyberbullying incidents occur in the private sphere, 38% still occur in a work environment. In addition to depression, addiction risk or physical complaints, around 15% of those affected by bullying and cyberbullying classified themselves as suicidal. While over a third of those affected had communicated with friends or family in response to (cyber)bullying, another third said they had taken no action and only 15% said they had looked for information and help on the internet. From an economic point of view, the willingness of bullying victims to quit is 40% higher, those affected have almost twice as many sick days as the average and the annual costs of lost production in the German economy are estimated at around 8 billion euros. A regular survey by the Media Authority of North Rhine-Westphalia (Landesanstalt für Medien NRW) [35] shows that the number of internet users in Germany who are frequently confronted with hate speech has risen in recent years from 27% (2017) to 39% (2021). Although more than two-thirds of the respondents in 2021 have already noticed hate comments, only 28% of them have reported a hate comment to the respective portal. Nevertheless, internet users see prosecution (87%) or deletion of hate comments (73%) as more effective than behavioural guidelines (42%) or active counter-speech (17%).

As part of the "Hessen gegen Hetze" campaign, the Hessen3C has set up a reporting office that allows those affected to report hate messages by providing a brief description, the source, the time when the comment was noticed, a hyperlink (URL) and optionally a screenshot. On this basis, the Hessen3C carries out a preliminary assessment of the hate messages - sometimes including cases of cyberbullying - and forwards the information to other authorities and institutions if necessary. In the case of a direct threat situation, the Hessian State Criminal Police Office (Hessisches Landeskriminalamt, HLKA) is called in, in the case of criminal relevance the Central Office for Combating Internet and Computer Crime (Zentralstelle zur Bekämpfung der Internet- und Computerkriminalität, ZIT), in the case of supra-regional relevance the Federal Criminal Police Office (Bundeskriminalamt, BKA) and in cases of extremism additionally the State Office for the Protection of the Constitution (Landesamt für Verfassungsschutz, LfV). Furthermore, content is reported by form to the original platform operators (e.g., Facebook, Twitter) within the framework of the Network Enforcement Act (NetzDG) in order to remove it from the public as soon as possible. At present, however, the reporting office is not able to provide a comprehensive picture of the situation regarding cyberbullying and hate speech in the area of responsibility of the state of Hesse. On one hand, the range of services and the visibility of the reporting office could be improved through technological multipliers (e.g., via a smartphone app), on the other hand, the detection and reporting of cyberbullying and hate messages could be supported by easy-to-use tools for citizens (e.g., as a browser plug-in). In order to establish the situation picture, it would also be necessary to monitor the relevant channels, especially social media, which cannot be achieved manually due to the sheer amount of social data and the limited personnel resources available. Here, it seems logical to use novel AI approaches to enable the most accurate possible pre-classification of the data and to visualise it in an interactive dashboard, which enables a quick sifting, prioritisation, case handling and forwarding of information. However, as the significance and legal relevance of cyber abuse is permanently renegotiated and reassessed, possibilities of intervention and subsequent modification of classification criteria and calculation bases of AI have to be considered.

## 4.2 Planned Innovations

CYLENCE creates strategic and technological solutions that optimise and future-proof the processes of collecting and analysing cyberbullying and hate speech for LEA. It relates to the funding policy objectives of awareness raising by investigating (1) tools and measures to detect or prosecute cyber abuse (esp. cyberbullying and hate messages), (2) awareness raising and training of LEA, and (3) the role of botnets in spreading hate. The project is characterised by the following objectives:

- Development of a taxonomy for the classification and treatment of cyber abuse posted in (partially) public social media (e.g., blogs, instant messengers, photo and video portals, social networks, website forums). Although the focus in the project is on cyberbullying and hate speech, a cross-domain, transferable concept is aimed at through literature research and involvement of associated partners.
- Collecting empirical evidence on the (inter-)organizational analysis of cyberbullying and hate messages in LEA, as well as on expectations and current practices of authorities and citizens regarding cyber abuse communication. In CYLENCE, these findings form the basis for the user-oriented development of novel strategies and tools for LEA to enhance civil security, expand the state of the art through scientific publications, and are communicated to citizens and LEA.
- New methods for adaptable, fair and comprehensible classification of cyberbullying and hate messages. Based on a baseline model (e.g., GPT-3), Few-Shot Learning (model building with few training data) allows for fast adaptation of the model, Data Augmentation (artificial generation of training data with interpolation that avoids social biases) allows for fair improvement of the model, and Whitebox methods (e.g., LIME) allow for explanation of algorithmic decisions.
- A novel demonstrator for cross-media integration of citizen-reported incidents via hessenWARN, browser plugin or web form (reporting module), for real-time-based collection of (partially) public social data sources (collection module), and for configurable and visual analysis for early detection and prioritization of cyberbullying and hate messages based on a dashboard (analysis module) for LEA.
- Organizational strategies for LEA for systematic intra- and interorganizational analysis of and communication about cyberbullying and hate messages, taking into account advancing digitalization, networking, and constant change in the technology landscape. This includes a training strategy with interactive tutorials on the use of developed tools and a communication strategy for the exchange between those affected and LEA.

## 4.3 Results II: Identified Research Challenges

Based on the overall research agenda, we elaborated seven distinct research challenges combined with information on how CYLENCE will attempt to overcome these challenges.

*4.3.1 Adapting Content Moderation Research to the Domain of Law Enforcement Agencies (C1).* Particularly in recent years, the research fields of Computer Supported Cooperative Work (CSCW), Human-Computer Interaction (HCI), and Information Systems (IS) have investigated and developed various technical and strategic measures against abusive online content such as hate speech and cyberbullying. However, this extensive and empirically grounded research landscape focuses almost exclusively on practices and strategies [15, 16, 21, 25, 37, 49, 50, 56, 66, 90, 91, 105], technologies [17, 47, 48, 59, 71], and issues of explainability [13, 58, 67, 98], transparency [70, 74], and contestability [100, 101] of AI systems in content moderation. In contrast, there is little research on the specific work practices, requirements, and challenges of detecting and handling abusive Internet content by LEAs such as reporting centers, police departments, and prosecutors' offices. The few existing works on this particular application domain lack a systematic empirical foundation, have a methodological or technical approach, and focus specifically on the detection of hate speech [39, 43].

CYLENCE addresses this gap by drawing on theories, methods, and discourses from content moderation research in HCI, CSCW, and IS to empirically identify existing strategies, practices, and challenges of the detection and handling of hate speech and cyberbullying cases by LEAs. This lays the foundation for a subsequent development and evaluation of user-centered technology artifacts specifically for this application domain.

*4.3.2 Conceptualization and Differentiation of Hatespeech and Cyberbullying (C2).* In research, there has been no established definition of the concept of Hatespeech [64, 95], which results in a situation where research datasets are often underpinned by significantly different understandings of the concept [2]. This complicates the reliable evaluation of detection algorithms and limits their generalizability [64, 106]. Furthermore, hate speech is often insufficiently differentiated from broader concepts, e.g., abusive language, as well as from specific forms of group-related hostility, e.g., antisemitism, which may lead to the use of different terms for the same phenomenon [79]. Against this background, research has already developed taxonomies of abusive Internet content [2]. There are also approaches to differentiate hate speech according to the respective targeted group or criminal relevance [22, 96]. However, to the best of the authors' knowledge, there exists no taxonomy of abusive Internet content tailored to the particular requirements of LEAs that systematically distinguishes hate speech and cyberbullying from other content categories relevant to law enforcement and, if reasonable, differentiates subtypes for the individual categories.

CYLENCE approaches this challenge by developing a comprehensive taxonomy of criminally relevant abusive online content and possible countermeasures on the basis of a systematic literature study. The taxonomy will subsequently be evaluated and further refined with practitioners from LEA to establish a basis for the user-centered development of detection algorithms (cf. C3), classification algorithms (cf. C4), and visual analytics approaches (cf. C6).

*4.3.3 Explainable Detection of Hatespeech and Cyberbullying in Multi-Modal Multi-Language Data (C3).* Numerous algorithms and datasets enable an AI-based detection of cyberbullying [28, 86] and hate speech [32, 79] in textual social media data of different languages. Other AI models aim to detect non-textual expressions of hate speech and cyberbullying, for example in visual [55, 87] or audiovisual data [9]. While so far most AI models for recognizing such content have been implemented as black boxes [67], i.e., without

revealing the logic behind the algorithmic classification decisions to end users, in recent research novel approaches have been explored to improve the explainability of such models [1, 67, 98]. Since AI research has so far considered issues of explainable, multilingual, and multimodal hate speech and cyberbullying detection in isolation, there is a gap with respect to the design of AI systems and datasets that both enable the detection of this content in data of different types and languages and are comprehensible to non-experts.

CYLENCE investigates to what extent a novel combination of few-shot learning, data augmentation and XAI approaches enables the detection of cyberbullying and hate speech in multi-modal and multi-lingual data streams. The datasets and algorithms developed for this purpose will be evaluated not only on the basis of technical metrics like performance, but also with respect to their transparency and comprehensibility by involving practitioners in LEAs.

*4.3.4 Strategies and Techologies for the Classification and Prioritization of Hatespeech and Cyberbullying in Law Enforcement Agencies (C4).* Technical research on the classification of hate speech and cyberbullying has so far been primarily concerned with the binary detection of these types of content or of individual subtypes [32, 86]. Only sporadic work has additionally examined multi-label classification, in particular for identifying subtypes of hate speech based on the targeted groups [14, 88]. Furthermore, initial hate speech datasets have been developed which, in addition to differentiating between targeted groups, also distinguish content by general criminal relevance and individual criminal norms under German law [22]. However, evaluated models for multi-label classification of hate speech [14, 88] are so far limited to a few targeted groups, were developed without considering the requirements of LEAs, and do not allow for a classification according to criminal relevance. In addition, multi-label classification models for cyberbullying have not yet been created, nor has the possibility of an algorithmic prioritization of content, such as by severity or urgency, been explored.

To support decision-making in LEAs, CYLENCE will develop strategies for differentiating and prioritizing different subtypes of hate speech and cyberbullying based on a taxonomy of abusive Internet content (cf. C2) and empirical research on the working context (cf. C1). Building on this, it will explore how multi-label classification and prioritization algorithms can contribute to the establishment of a differentiated situational picture.

*4.3.5 Usable and Accessible Reporting Technologies (C5).* The detection and reporting of hate speech and cyberbullying on the part of users can also be supported with technologies [61, 94]. First, apps to report such content have been developed for mobile devices [42, 80]. However, no empirical design case studies or evaluation studies with the respective target groups have been conducted for these applications, meaning that to date there is no generalizable design knowledge on such technologies. In addition, the possibility of integrating reporting functionalities into widely used applications such as warning apps, as well as associated advantages and disadvantages, has not yet been investigated. Second, browser plugins have also been developed that leverage AI to detect and visually highlight certain forms of hate speech in Facebook and Twitter timelines in real time [68] or to filter out or hide such content [10, 46]. However, such tools do not support the documentation and subsequent reporting of the respective content to LEAs.

Based on these research gaps, CYLENCE involves potential users in the development and evaluation of usable and accessible reporting technologies, in particular apps and browser plugins, which at the same time satisfy the specific requirements of LEAs by ensuring the transmission of all information relevant to criminal prosecution as well as compliance with data protection regulations.

*4.3.6 Visual analytics for Situational Awareness on Cyberbullying and Hate Speech (C6).* One key component of improving early detection and treatment of cyberbullying and hate speech cases by LEAs lies in enhancing situational awareness. Social data sources in particular can provide a valuable contribution to this if their content includes real-time descriptions, they have a large and active user base, and a public application programming interface (API) is available [60]. Moreover, a (semi-)automated evaluation of these sources also offers significant potential for detecting, analyzing, and predicting hate crimes beyond the Internet [78]. The concept of situational awareness [29, 30] has already been applied to the domain of cybersecurity [44, 57] and considered in the development of corresponding technology artifacts [45, 51]. However, in the development of artifacts for analyzing hate speech and cyberbullying, there has been little linkage to the concept of situational awareness [60]. Nonetheless, dashboards for visualizing and analyzing content collected from social media in particular have been developed [77, 78], but these are only partly real-time capable and do not allow for the integration of citizen-reported content. Both of these functionalities, however, appear particularly central to establishing comprehensive situational awareness in LEAs.

Against this background, CYLENCE explores to what extent the concept of situational awareness could be adapted for the governmental treatment of hate speech and cyberbullying. Based on this, we will investigate how visual analytics approaches can contribute to the joint analysis, prioritization, and handling of both proactively collected and citizen-reported content by LEAs.

*4.3.7 Value-Sensitive Design and Technology Assessment (C7).* The early assessment of ethical, legal, and social implications (ELSI) of technology development is particularly important in safety-critical contexts to ensure the acceptance of technical artifacts, minimize risks and unintended consequences of the development and use of artifacts, and thus sustainably ensure the success of a research project [8]. The ethical implications of using artifacts to detect and analyze abusive Internet content have already been explored with respect to private sector companies [92] and LEAs [99, 103]. Different ELSI considerations have also been highlighted with respect to the development of such artifacts [65, 102, 104]. Systematic biases of algorithmic models for detecting such content were identified as a particularly crucial issue [6, 102]. However, the specific domain requirements and challenges of LEAs have not yet received increased attention during the development of novel detection and analysis technologies for hate speech and cyberbullying.

CYLENCE ties in with this by examining different forms of bias [34] in the context of ML-based detection of hate speech and cyberbullying by involving affected stakeholder groups, looking at possible consequences, and deriving mitigation measures. In particular, this includes a reflection on the development of training datasets using data statements [5]. In order to ensure that ELSI considerations are taken into account throughout the project, a

Value Sensitive Design approach [33] that involves various stakeholders and a scientific advisory board with (inter-)national experts in the development of the technology in structured workshops is followed, identifying potential challenges and risks at an early stage and deriving appropriate solutions. With this innovative approach, the project contributes to the further methodological development of existing co-design approaches in safety-critical HCI.

## 5 CONCLUSION

In this paper, we presented the research agenda of the CYLENCE project and identified seven challenges for developing strategies and technologies in order to enhance the reporting, detection, and treatment of cyberbullying and hatespeech in LEAs. Using the lens of HCI, these challenges discuss a meaningful integration of humans and technology, including the adoption of content moderation technologies and strategies by LEAs (C1), semi-automatic and transparent data collection and analysis (C3, C4), support for citizen reporting (C5), integrated, configurable and usable data analytics (C6), and ethical, legal, and social implications (C7). However, the conducted narrative literature review does not "involve a systematic and comprehensive search of all relevant literature" [75] and the group discussions followed an open structure, lacking explicit and reproducible methods. Thus, based on these initial challenges, the project has started conducting systematic literature reviews on cyber situational awareness and cyber threat communication, performing qualitative expert interviews with German LEA personnel and conceptualising a representative citizen survey. The theoretical and empirical insights will be used to design, implement and evaluate supportive strategies and technologies. Finally, the project will explore the transfer of results to cyber abuse advisory services, reporting offices, and other related organisations.

## ACKNOWLEDGMENTS

This work has been funded by the German Federal Ministry of Education and Research (BMBF) in the project CYLENCE (13N16636-13N16639).

## REFERENCES

[1] Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, and Idowu Ademola Osinuga. 2020. Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review* 38 (2020), 100311. https://doi.org/10.1016/j.cosrev.2020.100311

[2] Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A Unified Taxonomy of Harmful Content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Association for Computational Linguistics, 125–137. https://doi.org/10.18653/v1/2020.alw-1.16

[3] Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A Survey on Data Augmentation for Text Classification. *Comput. Surveys* (7 2022). https://doi.org/10.1145/3544558

[4] Franz Beitzinger and Uwe Leest. 2021. *Mobbing und Cybermobbing bei Erwachsenen: Eine empirische Bestandsaufnahme in Deutschland, Österreich und der deutschsprachigen Schweiz.* Technical Report.

[5] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (12 2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[6] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In *Social Informatics. 9th International Conference*, Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri (Eds.). Springer Cham, Oxford, UK, 405–415. https://doi.org/10.1007/978-3-319-67256-4_32

[7] BMFSFJ. 2022. Was ist Cybermobbing? https://www.bmfsfj.de/bmfsfj/themen/kinder-und-jugend/medienkompetenz/was-ist-cybermobbing--86484 [Online; accessed 2022-03-14].

[8] Alexander Boden, Michael Liegl, and Monika Büscher. 2021. Ethische, rechtliche und soziale Implikationen (ELSI). In *Sicherheitskritische Mensch-Computer-Interaktion*, Christian Reuter (Ed.). Springer Fachmedien Wiesbaden, Wiesbaden, 185–205. https://doi.org/10.1007/978-3-658-32795-8_9

[9] Fariha Tahosin Boishakhi, Ponkoj Chandra Shill, and Md. Golam Rabiul Alam. 2021. Multi-modal Hate Speech Detection using Machine Learning. In *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 4496–4499. https://doi.org/10.1109/BigData52589.2021.9671955

[10] Jack Bowker and Jacques Ophoff. 2022. Reducing Exposure to Hateful Speech Online. In *Intelligent Computing.SAI 2022*, Kohei Arai (Ed.). Springer International Publishing, Cham, 630–645. https://doi.org/10.1007/978-3-031-10467-1_38

[11] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020).

[12] Enrico Bunde. 2021. AI-assisted and explainable hate speech detection for social media moderators - A design science approach. *Proceedings of the Annual Hawaii International Conference on System Sciences* 2020-Janua (2021), 1264–1273. https://doi.org/10.24251/hicss.2021.154

[13] Enrico Bunde. 2021. AI-Assisted and Explainable Hate Speech Detection for Social Media Moderators – A Design Science Approach. In *Proceedings of the 54th Hawaii International Conference on System Sciences*. AIS Electronic Library (AISeL), Kauai, Hawaii, USA, 1264–1273. https://aisel.aisnet.org/hicss-54/da/xai/2/

[14] Pete Burnap and Matthew Williams. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science* 5 (2016), 1–15. https://doi.org/10.1140/epjds/s13688-016-0072-6

[15] Jie Cai and Donghee Yvette Wohn. 2021. After Violation But Before Sanction: Understanding Volunteer Moderators' Profiling Processes Toward Violators in Live Streaming Communities. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25. https://doi.org/10.1145/3479554

[16] Jie Cai and Donghee Yvette Wohn. 2022. Coordination and Collaboration: How do Volunteer Moderators Work as a Team in Live Streaming Communities?. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3491102.3517628

[17] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30. https://doi.org/10.1145/3359276

[18] Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification. *arXiv* (2020). https://doi.org/10.18653/v1/2020.acl-main.194

[19] Ke-Li Chiu and Rohan Alexander. 2021. Detecting Hate Speech with GPT-3. *arXiv* (2021).

[20] Danielle Keats Citron and Helen Norton. 2011. Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review* 91 (2011), 1435.

[21] Amanda L. L. Cullen and Sanjay R. Kairam. 2022. Practicing Moderation: Community Moderation as Reflective Practice. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–32. https://doi.org/10.1145/3512958

[22] Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. DeTox: A Comprehensive Dataset for German Offensive Language and Conversation Analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 143–153. https://doi.org/10.18653/v1/2022.woah-1.14

[23] Hessisches Ministerium des Innern und für Sport. 2022. Hessen gegen Hetze. https://hessengegenhetze.de/node/59 [Online; accessed 2022-02-18].

[24] Neue Wege des Lernens e.V. 2017. Fake News Check. https://www.neue-wege-des-lernens.de/apps/ [Online; accessed 2018-09-20].

[25] Bryan Dosono and Bryan Semaan. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300372

[26] ECRI. 2015. ECRI General Policy Recommendation N°15. https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/recommendation-no.15 [Online; accessed 2022-03-14].

[27] Kathrin Eismann, Oliver Posegga, and Kai Fischbach. 2018. Decision Making in Emergency Management: The Role of Social Media. *Proceedings of the 26th European Conference on Information Systems (ECIS)*, 1–20.

[28] Fatma Elsafoury, Stamos Katsigiannis, Zeeshan Pervez, and Naeem Ramzan. 2021. When the Timeline Meets the Pipeline: A Survey on Automated Cyberbullying Detection. *IEEE Access* 9 (2021), 103541–103563. https://doi.org/10.1109/ACCESS.2021.3098979

[29] M.R. Endsley. 1995. Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37, 1 (mar 1995), 32–64. https://doi.org/10.1518/001872095779049543

[30] Mica R. Endsley. 2000. Theoretical Underpinnings of Situation Awareness: A Critical Review. In *Situation Awareness Analysis and Measurement*, Mica R. Endsley and Daniel J. Garland (Eds.). Lawrence Erlbaum Associates, Mahwah, New Jersey, USA, 3–32.

[31] Steven Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A Survey of Data Augmentation Approaches for NLP. *59t Annual Meeting of the Association for Computational Linguistcs and the 10th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, 968–988. https://doi.org/10.18653/v1/2021.findings-acl.84

[32] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *Comput. Surveys* 51, 4 (2018). https://doi.org/10.1145/3232676

[33] Batya Friedman, Peter H. Kahn Jr., Alan Borning, and Alina Huldtgren. 2013. Value Sensitive Design and Information Systems. In *Early engagement and new technologies: Opening up the laboratory. Philosophy of Engineering and Technology*, N. Doorn, D. Schuurbiers, van de I. Poel, and M. Gorman (Eds.). Springer, Dordrecht, 55–95.

[34] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems* 14, 3 (jul 1996), 330–347. https://doi.org/10.1145/230538.230561

[35] Landesanstalt für Medien NRW. 2021. *Forsa-Befragung zur Wahrnehmung von Hassrede.* Technical Report.

[36] Oguzhan Gencoglu. 2021. Cyberbullying Detection With Fairness Constraints. *IEEE Internet Computing* 25, 1 (2021), 20–29. https://doi.org/10.1109/MIC.2020.3032461

[37] Sarah A. Gilbert. 2020. "I run the world's largest historical outreach project and it's on a cesspool of a website." Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–27. https://doi.org/10.1145/3392822

[38] Margarita Grinko, Marc-André Kaufhold, and Christian Reuter. 2019. Adoption, Use and Diffusion of Crisis Apps in Germany: A Representative Survey, Florian Alt, Andreas Bulling, and Tanja Döring (Eds.). *Mensch und Computer 2019*, 263–274.

[39] Oren Halvani. 2023. Möglichkeiten zur Erkennung von Hate Speech. *Datenschutz und Datensicherheit - DuD* (2023), 209–214. Issue 47. https://doi.org/10.1007/s11623-023-1747-3

[40] Katrin Hartwig and Christian Reuter. 2019. TrustyTweet: An Indicator-based Browser-Plugin to Assist Users in Dealing with Fake News on Twitter. *Proceedings of the International Conference on Wirtschaftsinformatik (WI).*

[41] Hassmelden. 2022. Melde Hatespeech. Unterstütze Betroffene. Sorge für Strafverfolgung. Verpflichte die Politik. https://hassmelden.de/ [Online; accessed 2022-02-18].

[42] HateAid. 2020. App gegen Hass – Mach mit und werde MeldeHeld*in. https://hateaid.org/meldehelden-app/

[43] Julian Romeo Hildebrandt, Martina Ziefle, and André Calero Valdez. 2022. Entscheidungsautonomie und KI - Methodische Hinweise zur Untersuchung von KI-Nutzung in Sicherheitsbehörden. In *Mensch und Computer 2022 - Workshopband*, Karola Marky, Uwe Grünefeld, and Thomas Kosch (Eds.). Gesellschaft für Informatik e.V., Bonn. https://doi.org/10.18420/muc2022-mci-ws10-230

[44] Martin Husák, Tomáš Jirsík, and Shanchieh Jay Yang. 2020. SoK: contemporary issues and challenges to enable cyber situational awareness for network security. In *Proceedings of the 15th International Conference on Availability, Reliability and Security*. ACM, New York, NY, USA, 1–10. https://doi.org/10.1145/3407023.3407062

[45] Martin Husák, Lukáš Sadlek, Stanislav Špaček, Martin Laštovička, Michal Javorník, and Jana Komárková. 2022. CRUSOE: A toolset for cyber situational awareness and decision support in incident handling. *Computers & Security* 115 (apr 2022), 102609. https://doi.org/10.1016/j.cose.2022.102609

[46] Shreyans Jain and Deepali Kamthania. 2020. Hate Speech Detector: Negator. In *Proceedings of the International Conference on Innovative Computing & Communications (ICICC) 2020*. SSRN, 1–4. https://doi.org/10.2139/ssrn.3563563

[47] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction* 26, 5 (2019), 1–35. https://doi.org/10.1145/3338243

[48] Shagun Jhaver, Quan Ze Chen, Detlef Knauss, and Amy X. Zhang. 2022. Designing Word Filter Tools for Creator-led Comment Moderation. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–21. https://doi.org/10.1145/3491102.3517505

[49] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. Moderation Challenges in Voice-based Online Communities on Discord. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23. https://doi.org/10.1145/3359157

[50] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices. *Proceedings of the ACM on Human-Computer Interaction* 4, GROUP (2020), 1–35. https://doi.org/10.1145/3375197

[51] Marc-André Kaufhold, Ali Sercan Basyurt, Kaan Eyilmez, Marc Stöttinger, and Christian Reuter. 2022. Cyber Threat Observatory: Design and Evaluation of an Interactive Dashboard for Computer Emergency Response Teams. In *Proceedings of the European Conference on Information Systems (ECIS)*. Timisoara, Romania, 1–17.

[52] Marc-André Kaufhold, Markus Bayer, and Christian Reuter. 2020. Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning. *Information Processing & Management* 57, 1 (2020), 1–32. https://doi.org/10.1016/j.ipm.2019.102132

[53] Marc-André Kaufhold, Jasmin Haunschild, and Christian Reuter. 2020. Warning the Public: A Survey on Attitudes, Expectations and Use of Mobile Crisis Apps in Germany. *Proceedings of the European Conference on Information Systems (ECIS).*

[54] Marc-André Kaufhold, Nicola Rupp, Christian Reuter, and Matthias Habdank. 2020. Mitigating Information Overload in Social Media during Conflicts and Crises: Design and Evaluation of a Cross-Platform Alerting System. *Behaviour & Information Technology (BIT)* 39, 3 (2020), 319–342. https://doi.org/10.1080/0144929X.2019.1620334

[55] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 2611–2624. https://proceedings.neurips.cc/paper/2020/file/1b84c4cee2b8b3d823b30e2d604b1878-Paper.pdf

[56] Charles Kiene, Jialun Aaron Jiang, and Benjamin Mako Hill. 2019. Technological Frames and User Innovation: Exploring Technological Change in Community Moderation Teams. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23. https://doi.org/10.1145/3359146

[57] Alexander Kott, Cliff Wang, and Robert F. Erbacher. 2014. *Cyber Defense and Situational Awareness*. Advances in Information Security, Vol. 62. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-11391-3

[58] Yubo Kou and Xinning Gui. 2020. Mediating Community-AI Interaction through Situated Explanation: The Case of AI-Led Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27. https://doi.org/10.1145/3415173

[59] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18. https://doi.org/10.1145/3491102.3501999

[60] Rabindra Lamsal, Aaron Harwood, and Maria Rodriguez Read. 2023. Socially Enhanced Situation Awareness from Microblogs Using Artificial Intelligence: A Survey. *Comput. Surveys* 55, 4 (2023), 1–38. https://doi.org/10.1145/3524498

[61] Alana Lentin and Justine Humphry. 2017. Antiracism apps: framing understandings and approaches to antiracism education and intervention. *Information, Communication & Society* 20, 10 (2017), 1539–1553. https://doi.org/10.1080/1369118X.2016.1240824

[62] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* 30 (2017).

[63] Antonio López-Martínez, José Antonio García-Díaz, Rafael Valencia-García, and Antonio Ruiz-Martínez. 2019. CyberDect. A novel approach for cyberbullying detection on twitter. *International Conference on Technologies and Innovation*, 109–121.

[64] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PLOS ONE* 14, 8 (2019), e0221152. https://doi.org/10.1371/journal.pone.0221152

[65] Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, Hate Speech, and Social Media: A Systematic Review and Critique. *Television & New Media* 22, 2 (feb 2021), 205–224. https://doi.org/10.1177/1527476420982230

[66] Aiden R. McGillicuddy, Jean-Gregoire Bernard, and Jocelyn Ann Cranefield. 2016. Controlling Bad Behavior in Online Communities: An Examination of Moderation Work. In *ICIS 2016 Proceedings*. AIS Electronic Library (AISeL). https://aisel.aisnet.org/icis2016/SocialMedia/Presentations/23

[67] Christian Meske and Enrico Bunde. 2022. Design Principles for User Interfaces in AI-Based Decision Support Systems: The Case of Explainable Hate Speech Detection. *Information Systems Frontiers* (2022). https://doi.org/10.1007/s10796-021-10234-5

[68] Sandip Modha, Prasenjit Majumder, Thomas Mandl, and Chintak Mandalia. 2020. Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance. *Expert Systems with Applications* 161 (2020), 113725. https://doi.org/10.1016/j.eswa.2020.113725

[69] Guanyi Mou and Kyumin Lee. 2021. An Effective, Robust and Fairness-aware Hate Speech Detection Framework. *IEEE International Conference on Big Data*, 687–697. https://doi.org/10.1109/bigdata52589.2021.9672022

[70] Kilian Müller, Holger Koelmann, Marco Niemann, and Jörg Becker. 2022. Exploring Audience's Attitudes Towards Machine Learning-based Automation in Comment ModerationAutomation in Comment Moderation. In *Wirtschaftsinformatik 2022 Proceedings*. AIS Electronic Library (AISeL), Nürnberg, Germany. https://aisel.aisnet.org/wi2022/human_rights/human_rights/1

[71] Marco Niemann. 2021. Elicitation of Requirements for an AI-Enhanced Comment Moderation Support System for Non-tech Media Companies. In *HCI International 2021 - Posters. HCII 2021* (communicat ed.), Constantine Stephanidis, Antona Margherita, and Stavroula Ntoa (Eds.). Springer, 573–581. https://doi.org/10.1007/978-3-030-78635-9_73

[72] Ekaterina Olshannikova, Thomas Olsson, Jukka Huhtamäki, and Hannu Kärkkäinen. 2017. Conceptualizing Big Social Data. *Journal of Big Data* 4, 1 (2017), 1–19. https://doi.org/10.1186/s40537-017-0063-x

[73] Teresa Onorati, Paloma Díaz, and Belen Carrion. 2018. From social networks to emergency operation centers: A semantic visualization approach. *Future Generation Computer Systems* (2018). https://doi.org/10.1016/j.future.2018.01.052

[74] Marie Ozanne, Aparajita Bhandari, Natalya N Bazarova, and Dominic DiFranzo. 2022. Shall AI moderators be made visible? Perception of accountability and trust in moderation systems on social media platforms. *Big Data & Society* 9, 2 (2022). https://doi.org/10.1177/20539517221115666

[75] Guy Paré, Marie-Claude Trudel, Mirou Jaana, and Spyros Kitsiou. 2015. Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management* 52, 2 (2015), 183–199.

[76] Demetris Paschalides, Dimosthenis Stephanidis, Andreas Andreou, Kalia Orphanou, George Pallis, Marios D. Dikaiakos, and Evangelos Markatos. 2020. Mandola: A Big-Data Processing and Visualization Platform for Monitoring and Detecting Online Hate Speech. *ACM Transactions on Internet Technology* 20, 2 (2020), 1–21. https://doi.org/10.1145/3371276

[77] Demetris Paschalides, Dimosthenis Stephanidis, Andreas Andreou, Kalia Orphanou, George Pallis, Marios D. Dikaiakos, and Evangelos Markatos. 2020. MANDOLA: A Big-Data Processing and Visualization Platform for Monitoring and Detecting Online Hate Speech. *ACM Transactions on Internet Technology* 20, 2 (2020), 1–21. https://doi.org/10.1145/3371276

[78] Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. Detecting and Monitoring Hate Speech in Twitter. *Sensors* 19, 21 (2019), 4654. https://doi.org/10.3390/s19214654

[79] 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 (2021), 477–523. https://doi.org/10.1007/s10579-020-09502-8

[80] Research center on security and crime. 2019. eMORE Project. Monitoring and Reporting Online Hate Speech in Europe. https://www.rissc.it/homepage/our-projects/emore-project/

[81] Christian Reuter and Marc-André Kaufhold. 2018. Fifteen Years of Social Media in Emergencies: A Retrospective Review and Future Directions for Crisis Informatics. *Journal of Contingencies and Crisis Management (JCCM)* 26, 1 (2018), 41–57. https://doi.org/10.1111/1468-5973.12196

[82] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

[83] Thea Riebe, Marc-André Kaufhold, and Christian Reuter. 2021. The Impact of Organizational Structure and Technology Use on Collaborative Practices in Computer Emergency Response Teams: An Empirical Study. *Proceedings of the ACM: Human Computer Interaction (PACM): Computer-Supported Cooperative Work and Social Computing* CSCW (2021), 1–26. https://doi.org/10.1145/3479865

[84] Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification. *International Conference on Information and Knowledge Management (CIKM)*. https://doi.org/10.1145/3357384.3358040

[85] Tom Robinson, Clark Callahan, Kristoffer Boyle, Erica Rivera, and Janice K Cho. 2017. I like FB: A Q-Methodology Analysis of Why People 'Like' Facebook. *International Journal of Virtual Communities and Social Networking (IJVCSN)* 9, 2 (2017), 46–61. https://doi.org/10.4018/IJVCSN.2017040103

[86] H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. M. Veiga Simão, and I. Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior* 93, December 2018 (2019), 333–345. https://doi.org/10.1016/j.chb.2018.12.021

[87] Siva Sai, Naman Deep Srivastava, and Yashvardhan Sharma. 2022. Explorative application of fusion techniques for multimodal hate speech detection. *SN Computer Science* 3, 2 (2022), 122. https://doi.org/10.1007/s42979-021-01007-7

[88] Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard J. Jansen. 2018. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Twelfth International AAAI Conference on Web and Social Media*. Palo Alto, California, USA, 330–339. https://doi.org/10.1609/icwsm.v12i1.15028

[89] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2020. The risk of racial bias in hate speech detection. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 1668–1678. https://doi.org/10.18653/v1/p19-1163

[90] Charlotte Schluger, Jonathan P. Chang, Cristian Danescu-Niculescu-Mizil, and Karen Levy. 2022. Proactive Moderation of Online Discussions: Existing Practices and the Potential for Algorithmic Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–27. https://doi.org/10.1145/3555095

[91] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, New York, NY, USA, 111–125. https://doi.org/10.1145/2998181.2998277

[92] Leslie Regan Shade and Rianka Singh. 2016. "Honestly, We're Not Spying on Kids": School Surveillance of Young People's Social Media. *Social Media + Society* 2, 4 (oct 2016), 205630511668000. https://doi.org/10.1177/2056305116680005

[93] Haoyue Shi, Karen Livescu, and Kevin Gimpel. 2021. Substructure Substitution: Structured Data Augmentation for NLP. *arXiv*.

[94] Eugenia Siapera, Elena Moreo, and Jiang Zhou. 2018. *Hate track: Tracking and monitoring online racist speech*. Technical Report. Irish Human Rights and Equality Commission. https://www.ihrec.ie/documents/hatetrack-tracking-and-monitoring-racist-hate-speech-online/

[95] Alexandra A. Siegel. 2020. Online Hate Speech. In *Social Media and Democracy: The State of the Field, Prospects for Reform*, Nathaniel Persily and Joshua A. Tucker (Eds.). Cambridge University Press, Cambridge, 56–88.

[96] Leandro Silva, Mainack Mondal, Denzil Correa, Fabricio Benevenuto, and Ingmar Weber. 2016. Analyzing the Targets of Hate in Online Social Media. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*. Association for the Advancement of Artificial Intelligence, 687–690. https://doi.org/10.1609/icwsm.v10i1.14811

[97] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *arXiv*.

[98] Lukas Sontheimer, Johannes Schäfer, and Thomas Mandl. 2022. Enabling Informational Autonomy through Explanation of Content Moderation: UI Design for Hate Speech Detection. In *Mensch und Computer 2022 - Workshopband*. Gesellschaft für Informatik e.V., Darmstadt, Germany. https://doi.org/10.18420/muc2022-mci-ws12-260

[99] Daniel Trottier. 2015. Open source intelligence, social media and law enforcement: Visions, constraints and critiques. *European Journal of Cultural Studies* 18, 4-5 (aug 2015), 530–547. https://doi.org/10.1177/1367549415577396

[100] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. "At the End of the Day Facebook Does What ItWants": How Users Experience Contesting Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–22. https://doi.org/10.1145/3415238

[101] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability For Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28. https://doi.org/10.1145/3476059

[102] Bertie Vidgen, Dong Nguyen, Rebekah Tromble, Alex Harris, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy, 80–93.

[103] James P. Walsh and Christopher O'Connor. 2019. Social media and policing: A review of recent research. *Sociology Compass* 13, 1 (jan 2019), e12648. https://doi.org/10.1111/soc4.12648

[104] Helena Webb, Marina Jirotka, Bernd Carsten Stahl, William Housley, Adam Edwards, Matthew Williams, Rob Procter, Omer Rana, and Pete Burnap. 2017. The Ethical Challenges of Publishing Twitter Data for Research Dissemination. In *Proceedings of the 2017 ACM on Web Science Conference*. ACM, New York, NY, USA, 339–348. https://doi.org/10.1145/3091478.3091489

[105] Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300390

[106] Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science* 7 (2021), e598. https://doi.org/10.7717/peerj-cs.598

[107] Himanshu Zade, Kushal Shah, Vaibhavi Rangarajan, Priyanka Kshirsagar, Muhammad Imran, and Kate Starbird. 2018. From Situational Awareness to Actionability: Towards Improving the Utility of Social Media Data for Crisis Response. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW, Article 195 (nov 2018), 18 pages. https://doi.org/10.1145/3274464