AI & SOCIETY
https://doi.org/10.1007/s00146-025-02424-5

REVIEW

A multidisciplinary analysis of transparent AI-driven toxicity detection tools for civic engagement platforms

Maria Zangl[1] · Iliana Loi[2] · Panagiotis Zachos[2] · Michael Bedek[1] · Emmanouil Dimogerontakis[3] · Charikleia-Eleni Nikolaou[3] · Dietrich Albert[1] · Konstantinos Moustakas[4]

Received: 13 March 2025 / Accepted: 5 June 2025
© The Author(s) 2025

Abstract
Toxic speech on online civic engagement platforms (CEPs) disproportionately affects marginalized groups and threatens the diversity of citizen voices. However, the deployment of AI-driven toxic speech detection (TSD) tools for CEPs faces complex challenges from legal, psychological, and technical perspectives that remain insufficiently explored. We present a first-of-its-kind interdisciplinary review of these challenges, focusing on the explainability of TSD systems, their compliance with European legal standards and offer a roadmap for ethical deployment. Our review reveals three main findings. First, although transparency in AI decision-making is necessary from both legal and psychological perspectives, assessing the explainability of AI-driven TSD tools, and their compliance with legal regulations within Europe, remains a significant challenge. Second, current explainability approaches, ranging from toxic span identification to advanced explainable AI methods, lack standardized metrics. This makes it difficult to assess their reliability and appropriateness for CEPs. Third, despite the importance of TSD, frameworks and best practices for CEPs are still lacking in existing literature. This paper aims to fill this gap by providing a holistic perspective on the challenges and solutions for TSD deployment. It provides the foundation for collaborative efforts to develop and standardize metrics, evaluation protocols, and best practices that can ensure AI decisions in CEPs are transparent, accountable, and aligned with users' needs.

Keywords Toxic speech detection · Content moderation · Artificial intelligence (AI) · Online civic engagement · Legal regulations · Explainable AI

1 Introduction

We are living in an era, where digital platforms have become the new public square. Dialogs, discussions, and information exchanges increasingly occur online, overcoming traditional limitations of time and location. However, the ease of access and potential anonymity on social platforms often lead to destructive and uncivil interactions (Ziegele et al. 2018). Toxic speech on online platforms primarily targets individuals from marginalized or vulnerable groups, frequently resulting in reduced or restrained online activity among those affected (ADL 2019; Silva et al. 2021). Since

✉ Maria Zangl
maria.zangl@uni-graz.at

Iliana Loi
loi@ceid.upatras.gr

Panagiotis Zachos
p_zachos@ac.upatras.gr

Michael Bedek
michael.bedek@uni-graz.at

Emmanouil Dimogerontakis
m.dimogerontakis@legalshield365.com

Charikleia-Eleni Nikolaou
tmt@legalshield365.com

Dietrich Albert
dietrich.albert@uni-graz.at

Konstantinos Moustakas
moustakas@upatras.gr

[1] Institute of Psychology, University of Graz, Graz, Austria

[2] Department of Electrical and Computer Engineering, University of Patras, Pátrai, Greece

[3] School of Law, National and Kapodistrian University of Athens, Athens, Greece

[4] Communications and Information Technology Lab, University of Patras, Pátrai, Greece

Published online: 07 July 2025

individuals from marginalized groups are already underrepresented and thus less heard in political discourses (Kuran et al. 2020; Raisio et al. 2014), such verbal attacks on online platforms can exacerbate the inequalities, discriminating against individuals, such as those with disabilities, members of the LGBTQ-community, ethnic minorities, and other marginalized groups.

This review is structured as follows: we start with an overview from a psychological perspective on the necessity and the scope of Artificial Intelligence (AI) toxic speech detection (TSD) on online platforms, with a particular focus on civic engagement platforms (CEPs). The focus of Sect. 2 lies on the laws and regulations within the European Union (EU). This constitutes the legal prerequisites for applying TSD tools and content moderation on social platforms. This is followed by a comprehensive technical review of current AI TSD tools and relevant approaches to assess their explainability in Sect. 3. In Sect. 4, we then synthesize legal insights, psychological implications and technical gaps and deliver CEPs-specific recommendations. Subsequently, we present a research agenda where we announce both short- and long-term goals from each discipline to provide directives for future research.

1.1 The psychological impact of toxic speech

Toxic speech, encompassing various forms of harmful communication such as offensive, abusive, and hate speech (D'Sa et al. 2020), on online social platforms is a growing concern. Research has shown its significant impact on user behavior on online platforms. Smirnov et al. (2023), for example, found that toxic speech on Wikipedia is associated with decreased editor activity and an increased likelihood of editors permanently leaving the platform. Waseem et al. (2017) provide a typology of abusive language, distinguishing between explicit versus implicit and directed versus generalized abuse. Abusive language can be aimed at a specific individual or group, or it can be directed more broadly at a generalized 'other', such as individuals of a particular ethnicity or sexual orientation. The explicit versus implicit dimension differentiates between denotation, the literal meaning of a term, and connotation, which considers the sociocultural context (Waseem et al. 2017). Implicit abusive language often disguises its offensive nature through sarcasm or ambiguous terms, relies on stereotypes and indirect language, and has received limited attention (Hoang et al. 2024). Explicit abusive language clearly exhibits offensive characteristics, including racist or homophobic terms, and has been studied more widely (Schmidt and Wiegand 2017; Jahan and Oussalah 2023). For further classifications and relations between abusive and toxic, hate and aggressive speech, sexist/homophobic/religious and offensive speech concepts and a hierarchy, see Poletto et al. (2021). For this review, *toxic speech* is used as an umbrella term encompassing both explicit and implicit language aimed at harming others.

A 2018 survey of U.S. residents revealed that marginalized groups are the most frequent targets of online toxicity (ADL 2019). In one of three cases, online harassment is targeted at a characteristic of the victim that requires protection, including disability, religion, gender identity, sexual orientation, race, or ethnicity. Persons from the LGBTQ community, Muslims, Hispanics, and African Americans are particularly vulnerable to experience identity-based discrimination (ADL 2019). A study by Silva et al. (2021) found that terms related to racism were among the most frequently used hate words on major social platforms. Such toxic behavior has discouraged the participation of targeted individuals in online discourse, undermining diversity in digital discussions (ADL 2019). For example, a Swedish study reported how journalists expressed fear and thoughts on leaving their profession after being confronted with online harassment (Löfgren Nilsson and Örnebring 2016). Exposure to hate speech on platforms like Reddit has also been linked to increased stress and depressive symptoms among users (Saha et al. 2019; Wachs et al. 2022).

1.2 Toxic speech and online civic engagement

Considering the underrepresentation of marginalized groups in political activities and online civic engagement (see e.g., Guldvik et al. 2013; Davies and Procter 2020), TSD on platforms where citizens exchange opinions, suggest civic projects or vote and comment on them is an essential requirement for equal representation and inclusion on online CEPs.

Civic engagement is defined as the individual or collective participation of citizens in solving social problems in their communities (Gil de Zúñiga et al. 2012). Traditionally, this engagement apart from elections has been pursued through focus groups, surveys, and town hall meetings. Digital technologies have expanded these opportunities for citizen participation, allowing for engagement that is independent of time and place (Berntzen and Johannessen 2016).

CEPs, thus, play an increasingly important role in facilitating public participation in societal issues by making use of digital technologies to engage citizens with each other and with local governments (Berntzen and Johannessen 2016; Cho et al. 2021). They have recently shown to complement offline engagement and also directly influence civic attitudes and behavior (Zait and Andrei 2019). Digital participation platforms are a particular type of collaborative social media, providing the characteristics of popular social media (e.g., Facebook or X) and supporting user-generated content and sharing (Falco and Kleinhans 2018). Because of that, it is reasonable to apply TSD methods, which have been used in the context of social media platforms, for CEPs as well.

However, democracy thrives on diverse voices, and so it is also a high goal of CEPs that citizens from vulnerable and marginalized groups are represented and actively engaged in civic life. For these platforms to thrive, an inclusive and respectful online environment is essential, where toxic content is detected and removed immediately.

TSD, a form of content moderation, is defined as '*the screening, evaluation, categorization, approval or removal/hiding of online content according to relevant communications and publishing policies. It seeks to support and enforce positive communications behavior online, and to minimize aggression and anti-social behavior*' (Flew et al. 2019). As online discussion platforms are designed to enable easy and often anonymous posting, moderation is essential to prevent a flood of offensive or toxic text, image, or other material by users. Content moderation and the detection of toxic content can take place before (pre-moderation) and/or after (post-moderation) a post has been published (De Gregorio 2020). When the quantity of new posts or comments is high, pre-moderation is usually conducted automatically to avoid time delays in communication (Werthner et al. 2024). Post-moderation of platform content, dealing with posts or comments that are online already, is usually done by moderators who are in charge of a positive discussion environment, or by users who report them for different reasons (Werthner et al. 2024). Beyond the risk of users encountering toxic speech before it is removed by post-moderation methods, ElSherief et al. (2021) found that hate speech can increase the online visibility of its instigators. Thus, pre-moderation should be preferred to post-moderation, on the one hand to avoid the confrontation of toxic speech with users altogether and on the other hand not to give instigators more power through visibility.

As there is much debate about the extent to which online forums are moderated and where to draw the line between acceptable and unacceptable expressions and terms (Jakob et al. 2023; Abadi et al. 2016), it is important to consider the purpose of the platform and the goals it is pursuing. For CEPs, which should aim to gather the opinions, wishes, and suggestions of a diverse civic community, particularly targeting underrepresented groups, there should be no tolerance for toxic speech.

In addition, an environment in which respectful language without derogatory terms is used influences not only marginalized groups but also majority groups. This was shown in a study by Carnaghi and Maass (2007), where heterosexual people were more likely to have negative attitudes towards homosexual people when they are exposed to depreciating terms (i.e., slurs) compared to non-depreciating terms for homosexual people. Toxicity detection tools are, thus, not only capable of creating a safer online space by identifying and mitigating various forms of negative behavior, including hate speech and harassment (Aroyo et al. 2019) but also of avoiding majority groups being exposed to derogatory labels for minorities and can, therefore, lead to more positive attitudes towards them (Carnaghi and Maass 2007).

Implicit toxic speech, heavily context-dependent and less detected by predetermined linguistic rule-based approaches (Breitfeller et al. 2019), presents a challenge of detecting abusive language, which, in general, is the subjective nature of offense. For instance, it was found by Waseem et al. (2017) that humans moderating abusive content are more likely to identify racism as hate speech than content including sexism, highlighting the social biases in human moderation. The use of algorithms for content moderation as an ex ante approach that aims to enforce platform policies by removing toxic content as it is uploaded, before it can be viewed or shared by users (Cobbe 2021), can prevent prohibited content from spreading and reduce the workload on human moderators (Young 2022). However, as manual and semi-automated moderation methods have shown limited success, Gongane et al. (2024) advocate for more advanced approaches using Natural Language Processing (NLP) with Machine Learning (ML) algorithms.

1.3 The potential of (transparent) AI in TSD and content moderation

While the performance of AI-driven approaches is significant, the transparency of these systems is a key component for being well received and used by platform users. Research has shown the important role of transparency in AI systems for user trust and understanding. For example, Molina and Sundar (2022) showed that providing information about the basis on which a platform's content moderation system makes decisions improves users' perceptions of disclosure, agreement with the platform's policies, and trust in the moderation system. For instance, disclosing the keywords or decision criteria used for classification empowers users with a sense of agency, enabling them to understand and align with ML moderation practices. A user survey by Suzor et al. (2019) found that opaque moderation practices on social media platforms significantly hinder users' ability to understand and learn from content removal decisions. Often, users received little or no explanation, and consequently struggled to grasp the rules or adjust their behavior accordingly. Only about half of the surveyed users felt confident in understanding the reasons behind moderation actions. In response to vague or absent explanations, many developed their own informal theories, often attributing content removal to biased moderators or external influences. This aligns with Jhaver et al. (2019), who demonstrated the effect of providing explanations for content moderation on reduced likelihood of future post removals. Their analysis of 32 million posts showed that removal explanations educate users about community norms, helping them become

more productive members. Interestingly, explanations given by human moderators did not have a significant advantage over those provided by bots in reducing future removals. In the context of politically sensitive online debates on social media, a study by Abokhodair et al. (2024) found that opaque content moderation practices had a particularly harmful impact on political activists. The lack of transparency and perceived bias led to a profound loss of trust in platforms, reinforcing feelings of marginalization. This sense of being systematically silenced heightened frustration and anxiety among activists, contributing to more polarized and fragmented online discourse.

In addition to that, Cook et al. (2021) revealed that users who better understood content moderation practices were more likely to actively participate in toxicity management on platforms (e.g., flagging toxic content or communicating with platform managers). Thus, transparent AI systems for content moderation can improve perceived fairness by clearly communicating how toxic content is detected, reviewed, and removed, reducing potential misunderstanding (Abokhodair et al. 2024). Beyond that, compared to non-transparent AI systems for content moderation, transparent systems can consistently enhance the effectiveness of moderation through increased user support and feedback.

To date, there has been no systematic investigation into how technology addresses toxic speech and harmful interactions in online CEPs. To address this, we investigate the current state of TSD technologies, identify gaps in their explainability, and provide recommendations for future research. As such, this work sets the foundation for the development of a unified explainability evaluation framework where platform moderators would be able to compare available tools, and clearly observe the advantages and disadvantages of each available solution. To understand the requirements of constructive engagement in online platforms, it is crucial to examine existing European legal regulations and relevant ethical considerations before conducting a state-of-the-art analysis of AI-driven TSD tools.

2 Legal and ethical considerations

TSD tools have become integral to digital platforms, ensuring compliance with the Digital Services Act (DSA). The broader EU legal framework, including the General Data Protection Regulation (GDPR), and the EU Artificial Intelligence Act (AI Act), imposes stringent requirements on the development and implementation of such systems. TSD tools must adhere to key legal principles, including transparency, accountability, accuracy, human intervention, and the protection of fundamental rights, as outlined in these regulatory frameworks. While comparable initiatives exist outside Europe—including the U.S. Blueprint for an AI Bill of Rights,[1] a non-binding framework with diminished influence under the current administration, Canada's draft Artificial Intelligence and Data Act,[2] which stalled and was not enacted as of May 2025, and South Korea's 2023 AI Basic Act,[3] which has been enacted and will come into force in January 2026—none currently impose obligations as comprehensive, enforceable, or harmonized as the combined GDPR, DSA, and AI Act regime.

2.1 TSD and the DSA

The DSA defines content moderation as actions taken by platforms to detect, identify, and address illegal content or content that violates a platform's Terms of Service (ToS) (Article 3(t) DSA). TSD tools, as automated tools, are tasked with filtering harmful content, including hate speech and harassment. Their regulation differs based on whether the content is classified as illegal or as content violating a platform's ToS (Farinho 2024; Rojszczak 2023; Appelman 2021).

From a legal standpoint, it is essential to distinguish between these two categories:

- Illegal content includes speech that violates national or EU laws, such as incitement to violence or hate speech (Article 3(h) DSA). Platforms are obligated to remove this content swiftly under the DSA framework.
- In contrast, content that violates a platform's ToS may be harmful but not necessarily illegal. Platforms typically prohibit such content to maintain community standards, and they must ensure that these standards are transparently defined and applied consistently to prevent arbitrary restrictions on freedom of expression (Rojszczak 2023; Appelman 2021).

In CEPs, it is especially important that terms and conditions promote inclusivity while safeguarding vulnerable groups. These platforms must clearly define toxic speech, even if it is legal, and implement precise detection tools to manage it effectively (Appelman 2021).

The DSA imposes several obligations on platforms employing TSD tools:

- Transparency: Platforms must inform users about their content moderation policies, including the use of auto-

[1] https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

[2] https://www.parl.ca/documentviewer/en/44-1/bill/C-27/first-reading

[3] https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=1071&pageIndex=&sCode=eng&searchOpt=ALL&searchTxt=

mated tools for TSD on (Articles 14(1) and 15(1) DSA). In addition, platforms must publish reports detailing the accuracy and error rates of these systems.

- Appeal mechanisms: The DSA mandates that platforms provide users with mechanisms to appeal content moderation decisions. This guarantees human oversight in the moderation process, allowing users to challenge decisions made by automated TSD tools.
- Risk assessments: Platforms, especially very large online platforms, are required to conduct regular risk assessments to evaluate potential harms arising from content moderation practices, including those associated with AI-driven systems such as TSD.

These measures ensure that content moderation is transparent, fair, and accountable, thereby balancing the protection of free speech with the need to address harmful content. When assessing the impact of TSD tools on fundamental rights, especially in the context of CEPs, it is critical to consider their effects on democratic processes. While eliminating toxic speech can enhance inclusivity, excessive restrictions on free speech may undermine public dialog and influence public opinion in democratic societies (Rojszczak 2023; De Gregorio 2022).

In the Results and discussion section, we will dive further into this trade-off between preserving free speech and constraining toxic speech, while also examining the associated psychological implications of intervention in public discourses.

2.2 TSD and the GDPR

The GDPR is key to regulating AI systems used for content moderation, including TSD tools. Article 22 of the GDPR limits automated decision-making, including profiling, unless it is based on explicit consent or contractual necessity. This provision ensures that platforms using TSD allow users to challenge decisions and request human intervention, especially when these decisions significantly affect their rights (Farinho 2024).

A key principle of the GDPR is transparency, particularly under Articles 13 and 14, which require platforms, when they act as data controllers, to clearly inform users (data subjects) about the processing of their personal data, including the logic behind automated decisions. For platforms using TSD tools, this implies informing users about why their content has been flagged or removed (Farinho 2024). Transparency fosters user trust and ensures accountability in the content moderation process.

Moreover, the explainability principle, outlined in Article 13(2)(f), mandates that the data controllers shall provide 'meaningful information' on how decisions are made as well as the significance and the envisaged consequences of such processing for the data subject. This requirement is particularly relevant for TSD tools, where decisions may be based on complex algorithms that are not easily understood.

2.3 TSD and the AI act

The AI Act introduces additional requirements for platforms or tools that may be classified as AI systems, as defined in Article 3, Paragraph 1: *'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.'* These systems may be classified as high-risk AI systems if they significantly impact democratic processes or public discourse, as outlined in Annex III of the AI Act. Such a classification imposes stricter legal obligations, requiring platforms to implement measures for transparency, explainability, and human oversight (Farinho 2024; Appelman 2021).

Article 13 of the AI Act mandates that high-risk AI systems must provide detailed information regarding their functionality, accuracy, and limitations. This is crucial for TSD tools to ensure that platform deployers understand how decisions are made, how the AI operates, and acknowledge the potential for errors.

Another critical aspect of the AI Act is human oversight. Article 14 requires platforms using high-risk AI systems to ensure human intervention is possible in cases where AI decisions negatively affect fundamental rights, such as unjustified content removal. The inclusion of human oversight acts as a safeguard against overly aggressive or biased automated moderation, which could suppress free speech.

In addition to transparency and human oversight, Article 15 stipulates that high-risk AI systems must maintain high standards of accuracy, robustness, and cybersecurity throughout their lifecycle. This provision ensures that AI tools used for content moderation are reliable, resistant to manipulation, and secure from potential breaches. Accuracy in detecting toxic speech and robustness in maintaining high performance are key to mitigating the risks posed by AI driven content moderation systems.

Based on the legal frameworks presented in this section, it is made clear that there is a strong need for the development of qualitative criteria for evaluating the potential threats posed by the deployment of AI tools in CEPs. Such criteria are vital to ensure the tools are effectively integrated, operate ethically, and comply with legal standards, thereby maintaining the integrity and functionality of the platform. The effectiveness of AI tools hinges on their ability to deliver accurate and reliable outputs. By laying out these qualitative criteria, the groundwork is set for identifying quantitative

criteria that can offer more precise measures of the AI tools' performance and alignment with regulatory standards.

3 Overview on TSD mechanisms and tools

As mentioned above, identifying toxic speech on online platforms is an issue that requires urgent measures to create a safe space (as far as possible) for online users from all social groups. To address this, TSD tools based on ML/Deep Learning[4] (DL) have been widely employed on online social platforms, since automated moderation systems provide real-time detection capacities. Below, we provide a thorough review on such TSD mechanisms with a focus on the AI and explainability techniques they incorporate.

3.1 TSD mechanisms in social online platforms

With the goal of mitigating the adverse effects of toxic speech, various AI methods have been developed to assist human moderation (Alkomah and Ma 2022; Chhabra and Vishwakarma 2023; Jahan and Oussalah 2023; Malik et al. 2021; Dessi et al. 2021; Mahajan et al. 2021; Bonetti et al. 2023; Salminen et al. 2020; Aurpa et al. 2022; Anjum 2024) and are trained with hate-speech datasets (Alkomah and Ma 2022; Chhabra and Vishwakarma 2023). Malik et al. (2021) conducted two sets of experiments for TSD on datasets containing toxic content from Facebook and Twitter (now X) to determine the best method in terms of accuracy. The first experiment involved traditional ML techniques (e.g., Support Vector Machines (SVMs), Random Forest, etc.), where the datasets were pre-processed using standard NLP methods such as stop-word removal, part-of-speech tagging, lemmatization etc. The second tested DL models, such as Convolutional Neural Networks (CNNs) after applying Bidirectional Encoder Representations from Transformers (BERT) and fastText embeddings. BERT (Devlin et al. 2019; D'Sa et al. 2020), is a transformer-based neuronal network for NLP. It generates word embeddings by considering the context of each word within a sentence allowing it to capture correlations and contexts more effectively. Moreover, it is called 'bidirectional', since it can parse a text token from either of its two sides (left and right) during the training process (Malik et al. 2021). Different variations of BERT can be found in the literature. For example, MBert, a multilingual transformer, RoBERTa (Liu et al. 2019) a fine-tuned version of BERT that is more efficient than the latter in specific NLP tasks, and BERTweet (Nguyen et al. 2020), which can understand both text and emoticons and is thus, more suitable for TSD in social media (especially Tweets in English). In contrast to BERT, FastText (Joulin et al. 2017; D'Sa et al. 2020) is non-context-based and can effectively learn the structure of a word/phrase (i.e., suffixes, prefixes, and short words) by creating character n-gram representations of words (Malik et al. 2021). It is worth mentioning that these techniques and similar ones are large-scale models that are usually pre-trained and possibly later fine-tuned for specific applications.

Similarly, in Salminen et al. (2020) and Bonetti et al. (2023), a comparative analysis between various ML and DL models with NLP processing techniques for TSD in social media was made. Bonetti et al. (2023) showed that BERTweet achieves the best accuracy among all DL models, but is computationally expensive. Similarly, Salminen et al. (2020) demonstrated BERT's strong performance. In contrast to Salminen et al. (2020); Malik et al. (2021) and Bonetti et al. (2023), an evaluation by Dessi et al. (2021) focused exclusively on evaluating DL architectures using various word embeddings, including BERT, to identify different levels of toxicity in online content. Aurpa et al. (2022) employed both BERT and ELECTRA (Efficiency Learning an Encoder that Classifies Token Replacements Accurately) and other transformers to identify offensive comments in Bangla from Facebook. ELECTRA (Clark et al. 2020) understands the difference between 'real' input text sequences from those generated by another language model, operating more effectively than BERT-like methods, while in Aurpa et al. (2022), these two approaches yielded almost identical accuracy results. Similar to many of the previously reviewed studies, Mahajan et al. (2021) and Mehta and Passi (2022) employed ML and DL models with different pre-trained word embeddings (e.g., fastText in Mahajan et al. 2021 and BERT in Mehta and Passi 2022), for TSD in various datasets. In both studies, RNNs (Recurrent Neural Networks) were found to be the most accurate compared to the rest of the methods. Specifically, GRU (Gated Recurrent Units) were most effective according to Mahajan et al. (2021) and LSTM achieved the highest accuracy in Mehta and Passi (2022).

However, all aforementioned methods focus on detecting abusive text in social network platforms. In contrast, TSD on CEPs has gained limited to no attention, despite the fact that it is particularly essential in this context and requires a more fine-tuned approach. A few works address societal issues, such as Fan et al. (2021) and Grimminger and Klinger (2021), which cover TSD BERT-based methods for social media as well. In Fan et al. (2021), a toxicity detection approach was developed and trained upon Tweets commenting on UK Brexit, while the dataset used for training in Grimminger and Klinger (2021) contains Tweets from Biden and Trump supporters during the 2020 US Elections.

[4] Convolutional Neural Networks, Long Short Term Memory Networks (LSTMs), Recurrent Neural Networks (RNNs).

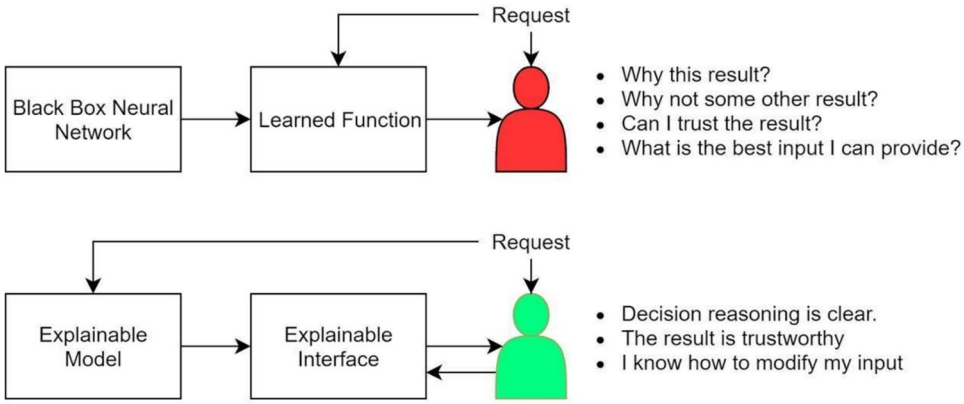

Fig. 1 Comparison between black box models and explainable models

3.2 Explainability in AI models

Content moderation on social media platforms is essential for raising international awareness of social movements and requires fair, transparent, and culturally sensitive policies. Transparency, which allows for regular fairness evaluations, is especially important during politically sensitive events. For example, Meta's Oversight Board criticized its automated moderation during the Israel–Hamas conflict due to several flaws in content removal (Abokhodair et al. 2024). Particularly, algorithmic content moderation can err on the side of caution, leading to the over-removal of content that is not genuinely harmful. This includes satire, political dissent, and expressions in minority dialects, all of which risk being suppressed by automated tools. Such over-removal disproportionately affects marginalized groups by documenting false positive bias in ML models used for abusive content detection (Yee et al. 2022). Studies show that identity-related terms (e.g., 'Muslim', 'gay'), as well as language associated with disability, gender, or African-American Vernacular English, are more likely to be misclassified as abusive (Hutchinson et al. 2020; Halevy et al. 2021). These errors are exacerbated by annotator bias and the penalization of linguistic features like non-standard vocabulary (Yee et al. 2022). The rise in toxic online speech, particularly targeting underprivileged groups, highlights the need for effective yet transparent tools for detecting toxic speech. Despite advances in ML, DL, and NLP, the most widely used TSD tools on social media often lack explainability, limiting their potential for employing them in CEPs. By addressing this limitation and utilizing Explainable AI (XAI) techniques, more inclusive and trustworthy, and thus safer online spaces where diverse voices can participate would be enabled. This would not only foster a more equitable digital environment but also contribute to strengthening democratic processes and promoting social cohesion.

When individuals, whether they be users, developers, or the individuals affected by the AI model's decisions (citizens on CEPs), understand how an AI system works and why it makes the decisions it does, they are more likely to trust it, as well as the decisions made by it (Cook et al. 2021; Abokhodair et al. 2024; Suzor et al. 2019). This is especially important for high-stakes applications of AI, such as civic engagement, healthcare, finance, and criminal justice. Figure 1 illustrates the contrast between traditional 'black-box' DL models and the goal of XAI.

Despite the benefits of explainable and transparent AI, there are also some challenges. One is that some AI models are very complex and difficult to explain. Another challenge is that there is no single definition of what it means for an AI system to be explainable or transparent.

ML models such as decision trees and Bayesian networks are easily interpreted by evaluating the contribution of each feature to the output due to their linear combination nature. However more complex models containing significant non-linearity do not provide any intuitive insight of their internal mechanisms. Thus, they are often referred to as 'Black-Box Systems' (Banerjee and Barnwal 2022).

By not disclosing their internal designs and decision-making processes, the predictions of such systems are especially hard to be justified to the end-users. The goal of XAI is to successfully communicate to its users why a decision was made, while also aiding the researchers in retracing the functionality of the models with the goal of deducing the inference mechanism employed.

3.3 Explainability in TSD AI tools

In the case of AI-driven TSD tools, some works develop methods that inherently provide a level of explainability (Hoang et al. 2024), while others focus on data augmentation methods (Ansari et al. 2024) and toxicity-annotated datasets (Mathew et al. 2021) to improve explainability. Hoang et al. (2024) developed a DL encoder–decoder framework, where the encoder detects implicit toxic speech and the decoder produces its explanation. In implicit toxic speech, the toxicity of the text is not always immediately apparent. Thus, an explanation of why a passage was deemed offensive must

Table 1 Overview of XAI models

XAI tool	Description
LIME (Local Interpretable Model-Agnostic Explanations) Ribeiro et al. (2016)	Provides local, interpretable explanations for individual predictions made by any ML model (model-agnostic). **LIME** explains the problem locally meaning that it does not generate an explanation for the whole model. Thus, it works by creating a simpler, interpretable model that approximates the behavior of the complex model. **LIME** also highlights words that render a text toxic, providing visual results
ERASER (Evaluating Rationales And Simple English Reasoning) DeYoung et al. (2020)	**ERASER** contains various explanation metrics to measure the plausibility and faithfulness of NLP model predictions compared with ground-truth human rationales. Plausibility indicates whether the NLP predictions are convincing to humans, while faithfulness is a measure of how accurate is the reasoning of the model i.e. in identifying toxic speech, and, thus, both these measures provide explainability
SHAP (SHapley Additive exPlanations[a])	**SHAP** calculates the contribution of each feature to the output of a ML model, thus being prediction-local like LIME. This can be helpful for understanding how each feature affects the model's predictions
T-EBANO Ventura et al. (2022)	**T-EBANO** produces both prediction-local and class-based model-global (across predictions) explanations in the form of human-interpretable text and numerical reports that are also human-interpretable. To provide explanations about the decision-making process of a model, **T-EBANO** uses objective metrics to measure the contribution of interpretable features during inference at the local and global levels. Subsequently, subjective measurements were applied through human surveys to evaluate the quality of the proposed explanations method

[a]https://christophm.github.io/interpretable-ml-book/shapley.html

be provided. To address this challenge, Ansari et al. (2024) applied data augmentation techniques, synthetically increasing the volume of a dataset to improve the explainability of a model. A representative dataset for explainable TSD is HateXplain (Mathew et al. 2021), which contains annotations for both the type of toxicity (e.g., hate, offensive, normal), the disadvantaged group that has been targeted, and an explanation as to why a text excerpt was deemed toxic.

Furthermore, there are numerous works, such as Mahajan et al. (2021), Mehta and Passi (2022), Shakil and Alam (2022), Ibrahim et al. (2022), Ventura et al. (2022), Ansari et al. (2024), and Kibriya et al. (2024), where XAI models as the ones described in Table 1, were utilized to evaluate the explainability of the implemented TSD tools. In Mehta and Passi (2022), one of the most prominent XAI tools, LIME (Local Interpretable Model-Agnostic Explanations, Ribeiro et al. 2016) as well as the ERASER (Evaluating Rationales and Simple English Reasoning, as in DeYoung et al. 2020) benchmark, were applied as explainability methods. More specifically, Mehta and Passi (2022) tested both of these XAI tools on the HateXplain dataset and concluded that an Artificial Neural Network (ANN) model with convolutional and feedforward layers with BERT embedding performed the best in terms of accuracy and explainability in HateXplain. In Shakil and Alam (2022) as well as in Ansari et al. (2024), LIME was also used to assess the explainability of their proposed explainable methods. Another prominent XAI model is SHAP (SHapley Additive exPlanations[5]), which was employed alongside LIME to provide explanations for a DL TSD network in Kibriya et al. (2024). Notably, Ventura et al. (2022) introduced a new XAI technique named T-EBANO, which was applied to DL NLP models for toxicity classification.

It is important to note, that some of the aforementioned methods (Hoang et al. 2024; Bonetti et al. 2023) included in Table 2 marked as explainable do not evaluate their self-reported explainability with any strict metrics. Specifically, they lack the application of standardized metrics, such as those reported by the OECD or evaluations based on established XAI methods (Gongane et al. 2024). Instead, these studies offer more task-specific explanation approaches (Pavlopoulos et al. 2021) or highlight toxicity spans inside a text (Sarker et al. 2023) to indicate the source of the model's decision. Others categorize toxicity into subclasses for multi-label/multi-class classification (Risch and Krestel 2020), or provide auxiliary information for the TSD process, such as the target group of abusive language (Hoang et al. 2024). Nevertheless, many works combine XAI techniques with non-strict explainability techniques, e.g., data augmentation, multi-class classification, etc., to provide better reasoning for the decision-making processes of TSD methods (Shakil and Alam 2022; Ansari et al. 2024; Kibriya et al. 2024).

Just like the non-explainable TSD AI methods described in Sect. 3.2, these methods use ML/DL models whose training data mainly comes from social media content (Gongane et al. 2024). However, explainable TSD tools that have been trained and tested on comments, posts, and conversations from CEPs are still scarce to date.

[5] https://christophm.github.io/interpretable-ml-book/shapley.html

Table 2 Overview of AI TSD Tools

Tool	Type of mode (embedding + TSD classifier)	Explainable (method)	Methods/metrics for explainability	Open source/open code	Domain of dataset, evaluation	Commercial
LLM Guard, 2025 Protect AI (2025)	Unbiased-toxic-RoBERTa	✘	N/A	✓	n.s., n.s	✘
Kibriya et al. (2024)	Keras embedding layer + CNN/RNN-based model	✓	LIME, SHAP	✘ (Using open-source models → LIME, SHAP)	Social media, n.s	✘
PerspectiveAPI (2023)	Multilingual BERT + single language CNN	✘	N/A	✘	Social media, CEP	✘
TrollWall AI (2025)	Custom embedding + transformer-based model	✘	N/A	✘	Social media, CEP	✓
Sarker et al. (2023)	RoBERTa + transformer-based models	✓	Highlighting techniques	✓	Software development platforms, n.s	✘
Pavlopoulos et al. (2021)	BERT, RoBERTa, FastText etc. + RNN-based models (e.g., LSTM, BiLSTM)	✓	Highlighting techniques	✘	News platforms, n.s	✘
Amazon transcribe toxicity detection A.W.S. (2024)	ML-based	✓	Highlighting techniques	✘	n.s., Social media	✓
One AI (2024)	GPT based (unspecified)	✘	N/A	✘	n.s., n.s	✓
Salminen et al. (2020)	BERT, etc. + ML models (e.g., SVM) and DL models (e.g., feed-forward NN)	✘	N/A	✓	Social media, n.s	✘
Malik et al. (2021)	BERT + ML models (e.g., SVMs, random forest, etc.) and DL models (e.g., CNNs)	✘	N/A	✘ (Using open-source models → BERT, fastText)	Social media	✘
Dessi et al. (2021)	BERT, etc. + DL models (CNNs, LSTMs, etc.)	✘	N/A	✓	Information portal, n.s	✘
Aurpa et al. (2022)	BERT, ELECTRA + transformer-based models	✘	N/A	✘ (Using open-source models → BERT, ELECTRA)	Social media, social media	✘
Bonetti et al. (2023)	BERTweet + ML/DL models	✓	N/A	✘ (Using open-source models → BERTweet)	Social media, n.s	✘
Hoang et al. (2024)	DL encoder–decoder model	✓	N/A	✘ (They will open-source their model)	Social media, n.s	✘
Mahajan et al. (2021)	fastText etc. + ML/DL models(e.g., GRUs)	✓	LIME	✘ (Using open-source models → fastText, LIME etc.)	Social media, n.s	✘
Mehta and Passi (2022)	BERT + ML/DL models (e.g., CNNs, LSTMs)	✓	LIME, ERASER	✘ (Using open-source models → BERT, LIME, ERASER)	Social media, n.s	✘
Shakil and Alam (2022)	FastText, etc. + CNN and ML models (e.g., random forest, etc.)	✓	LIME	✘(Using open-source models → LIME)	Social media, n.s	✘

Table 2 (continued)

Tool	Type of mode (embedding + TSD classifier)	Explainable (method)	Methods/metrics for explainability	Open source/open code	Domain of dataset, evaluation	Commercial
Ansari et al. (2024)	BERT + DL models (CNNs, LSTMs)	✓	LIME	✗ (Using open-source models → BERT, LIME, etc.)	Social media, n.s	✗
Ventura et al. (2022)	BERT, etc. + LSTM	✓	T-EBANO	✓	Information portal, n.s	✗

n.s. not stated; *N/A* not applicable

4 Results and discussion

In this section, the main results from a psychological, legal and implementation perspective are described and discussed, in particular, the differences and similarities between TSD on social media platforms and CEPs, the quality and clarity of ToS on content moderation, the concepts of effectiveness, reliability, and explainability of TSD tools, and finally, a closer look at two tools, Perspective API and TrollWallAI, comparing their benefits, limitations and ethical and legal shortcomings.

4.1 Quintessence of the psychological and legal perspective

As mentioned previously, it is crucial to recognize that the detection and potential filtering of toxic speech represent a double-edged sword. This is even more critical for CEPs, as these platforms, which may be managed by public entities, or private companies, play a direct role in democratic processes by facilitating dialog, voting on community issues, and interactions with municipal bodies. Consequently, decisions on content moderation and TSD directly impact democratic participation, public dialog-opinions. This makes content moderation policies highly sensitive, as they may unintentionally restrict public input and limit the free exchange of ideas, both of which are crucial for healthy or even deliberative democratic processes (Sunstein 1995). Attempts to introduce excessive 'political correctness' may negatively impact public discourse (Vaughan-Johnston et al. 2024). When individuals perceive threats to their freedom of expression, they often experience psychological reactance, a combination of anger and resistance (Brehm and Brehm 1981; Dillard and Shen 2005). This reactance can lead to backlash effects, negative attitudes towards protected groups, and active defense of perceived communicative autonomy. Furthermore, restricting communication opportunities for groups can decrease well-being and even motivates them to actively defend their place within their social hierarchy (Kachanoff et al. 2022).

Beyond psychological considerations, content moderation, particularly through TSD, has become a legal obligation for digital platforms. These platforms are required to take action not only against illegal content but also against content that violates their own terms of service. In the case of social online platforms, this requirement is particularly sensitive, as it allows private entities not only to interpret broad legal terms like 'hate speech' to identify illegal content but also to set standards for unacceptable material, such as toxic speech, based on their own ToS.

This raises complex questions about the power of private companies over public discourse to shape societal norms and public opinion.

Another challenge for content moderation is the meaning of a spoken act, which is never fixed by words but shaped through their interpretation by others (Butler 1997). Whether terms are perceived as toxic arises primarily from their context, the person delivering, the person receiving, and sometimes the relationship between them (Read 2018; Haythornthwaite 2023). This subjective nature of toxicity underlines the challenges of AI-driven tools for detecting toxic language that effectively account for contextual differences. As algorithmic systems are yet unable to capture complex social dynamics, the inclusion of human oversight remains important. At present, achieving fairness and non-discrimination in content moderation requires human intervention for a more context-sensitive approach, complementing the current limitations of AI-driven TSD.

4.2 Gaps between TSD tools on social media platforms and CEP-specific requirements

At first sight, there is no reason to assume that the underlying principles and intended objectives of TSD should be any different for social media platforms than for CEPs (e.g., to protect vulnerable groups and communities and to avoid discrimination). However, there are at least two significant differences between these two types of platforms. First, in terms of content diversity and second, in that the aim of CEPs should be to support collective and collaborative problem-solving and decision-making, which is not necessarily the case with social media platforms. While content on social media platforms can cover all kinds of topics, including less polarizing ones (e.g., posts about pets or vacations), CEPs are exclusively about one more polarizing topic: politics. In the ideal case of a CEP, people feel a kind of connectedness and relevance when it comes to a discourse about—for example—mobility plans in their neighborhood or city, otherwise the ambition of participatory democracy would be missed. The discourse and the resulting collaborative decisions can have a direct impact on individual users (for example, if the mobility plans under discussion would lead to increased traffic and noise pollution near their homes). This also leads to polarization, which in turn can reinforce toxic speech as well as a discussion behavior, that might be emotional or exaggerated, but not toxic speech in the narrower sense, and necessary for a democratic discourse and collaborative decision-making. From this perspective, it can be argued that TSD gains additional importance in a CEP and that a better distinction is needed between genuine toxic speech and lively, emotional contributions to the discussion.

From a legal standpoint, there is an important and highly relevant distinction worth noting. CEPs are usually operated by public bodies or organizations acting under a public mandate to encourage participatory democratic processes. Because of this, they are more likely to be considered high-risk systems under the AI Act. This designation brings stricter obligations, including greater transparency, thorough documentation, active human oversight, and the requirement to carry out a Fundamental Rights Impact Assessment. In contrast, these legal duties do not always extend to commercial social media platforms. In addition, since CEPs directly help citizens exercise their fundamental democratic rights, their operators shoulder extra responsibility. They must ensure that any automated moderation tools they adopt—including TSD systems—uphold essential values like non-discrimination, freedom of expression, and procedural fairness, as set out in the Charter of Fundamental Rights of the European Union.

Concluding with the prerequisites of an ethical deployment of AI-driven TSD, meaning prioritizing fairness and using AI for the good, transparent reasoning behind automated decisions of systems also allow users to understand the promotion of fairness and guidelines on a platform. This enables both users and developers a continuous improvement of the system's power and the mitigation of potential biases.

4.3 The practical implementation perspective

Given this context, the distinction between general online platforms, social online platforms, and CEPs (as a subcategory of the others) is crucial, with implications beyond legal classifications. While CEPs using AI-based TSD tools may be classified as high-risk under the EU AI Act—thus facing stricter regulatory obligations—their unique role in promoting democratic participation means that the impact of TSD implementation reaches far beyond compliance. The integration of TSD in CEPs profoundly influences users and can shape democratic values and societal norms.

As our review clearly shows, both non-explainable AI and XAI TSD methods reviewed in this work, focus on evaluating datasets originating from social media platforms, where the majority of social interactions take place and abusive, discriminatory, and offensive language lurks, revealing a literature gap in TSD tools and respective XAI tools employed in the context of CEPs.

In light of these considerations, we strongly recommend, from a multidisciplinary perspective, not only the adaptation of TSD tools commonly used on broader online platforms, recognizing that the vast majority of CEPs currently lack TSD tools, but also the establishment of essential minimum requirements for the effective adoption of content moderation mechanisms. These include (1) quality and clarity, (2) effectiveness and reliability and (3) explainability of AI TSD tools.

4.3.1 Quality and clarity of ToS on content moderation

- Platforms should clearly and unambiguously define toxic speech within their ToS, detailing what constitutes unacceptable content to: (a) avoid subjective interpretations, (b) evaluate the accuracy of TSD tools, (c) allow users the opportunity to challenge content moderation decisions, and (d) strengthen and improve the AI system's accuracy by learning from past results.
- The ToS should include transparent explanations and practical examples to help users understand the boundaries of acceptable content, even listing specific words or phrases that may be characterized as toxic speech.
- Terms should avoid overreach and not arbitrarily limit free expression, ensuring terms are reasonable, fair, and not overly restrictive.
- Safeguards should be in place during the creation and updating of ToS, establishing a clear, objective methodology for the operation and decisionmaking of content moderation mechanisms. This approach helps prevent potential manipulation of public opinion or distortion of community preferences and intentions by the operators of the CEPs. It should contain the methodology for drafting and reviewing the ToS, the body responsible for drafting, reviewing and officially accepting the ToS of the platform.

4.3.2 Effectiveness and reliability of TSD AI tools

- Accuracy: TSD tools must achieve a high level of precision in identifying toxic content to prevent the unjust suppression of legitimate discourse. From a technical standpoint, the accuracy of these AI tools often varies, influenced by factors such as the nature of the data used for training such methods, the optimal selection of hyperparameters (e.g., learning rate, type of optimizer, number of epochs, etc.), which is often a trial and error procedure, the architecture of the AI model (e.g., number of layers, number of neurons in each layer, etc.) as well as language, platform content, and context that may indicate the presence of implicit offensive language in text excerpts evaluated during inference. Our state-of-the-art review showed that the most commonly used word embedding prediction AI mechanism, BERT (Devlin et al. 2019; D'Sa et al. 2020), and its variations (e.g., BERTweet, RoBERTa, etc.) are found to have the best accuracy among similar mechanisms such as FastText (Liu et al. 2019; Salminen et al. 2020; Bonetti et al. 2023; Aurpa et al. 2022), mostly due to being context-aware, meaning that its decision-making process is based on the context of the under-evaluation text excerpt. Nevertheless, it should be underlined that high accuracy entails the increased need for computational resources (Bonetti et al. 2023), which poses an obstacle in the case of both reproducing/fine-tuning and applying these methods. As for the AI models combined with BERT and FastText embeddings to increase the effectiveness of toxic language processing, recent works on TSD focus on sophisticated ML [e.g., Random Forest, SVMs, (Malik et al. 2021; Bonetti et al. 2023)] and Deep Learning Neural Networks [e.g., CNNs, LSTMs, etc. (Dessì et al. 2021; Mehta and Passi 2022)], with DL methods such as RNNs (Mahajan et al. 2021; Mehta and Passi 2022) indicating the best performance regarding this task. Since accuracy depends on a multitude of factors and, thus, being problem-specific, the task of defining the optimal TSD based solely on accuracy is rather difficult. However, a direction to ensure increased accuracy based on our review findings is to create TSD tools that consist of an RNN as a base (e.g., LSTM) and utilize BERTtype embeddings, whose form will be decided upon the nature of the problem (e.g., MBERT to perform toxicity detection in multilingual text).
- Efficiency: For TSD systems to effectively maintain platform integrity, swift action is crucial. Rapidly identifying and removing harmful content is essential, as even brief exposure to toxic material can have significant consequences. Even when post-moderation quickly deletes such content, the short-term visibility can lead to lasting effects. This exposure can intensify negative attitudes within majority groups towards minority groups (Carnaghi and Maass 2007), increase online visibility for toxic speech instigators (ElSherief et al. 2018), and create a perceived unsafe environment for affected minority groups.

4.3.3 Explainability

- Transparency in how TSD tools make decisions
- Accountability to ensure alignment with legal standards and platform policies
- A robust right to object or users to challenge moderation decisions
- Effective human oversight to verify AI-driven actions.

With TSD being a task of many social and legal implications, the choice of explainability method is crucial for achieving the objectives above, to provide either the user, developer, or researcher with an insight into the AI decision-making processes (i.e., why a text was deemed toxic). The literature review performed for the purposes of this work resulted in a categorization of publications that focus on explainable TSD. However, while multiple tools presented in Table 2 are classified as explainable, there is a clear gap identified in this work, when it comes to comparing the explainability between different tools or methods.

Our findings demonstrate that while steps have been taken toward the right direction of providing explainable tools, current approaches are varied, ranging from the inclusion of an additional classification step after the decision of a model (Kirk et al. 2023), to highlighting words that drive the model to decide a text excerpt was toxic (Pavlopoulos et al. 2021), hence hindering the selection of the most appropriate explainability tool for each TSD model. The lack of strict explainable methods, i.e., metrics or public benchmark datasets employed on such tools, makes the task of selecting an automated toxicity sensor with the goal of moderating text-based content on a CEP even harder than it already is.

Moreover, it is crucial to emphasize that explainability alone does not guarantee regulatory compliance. Legal conformity of TSD tools must be assessed holistically, considering broader system integration, transparency, fairness, human oversight, and the platform's operational context. Under the DSA, the evaluation should take into account the tool's role within a complete content moderation system, including moderation criteria, Terms of Service alignment, and user appeal mechanisms. GDPR compliance depends not only on technical design but also on the structure of associated data processing activities, requiring human oversight, records of processing, and Data Protection Impact Assessments. The AI Act further requires systems to align with the provider's intended use, include effective human-in-the-loop mechanisms, and incorporate measures ensuring fairness and bias detection. In specific cases, a Fundamental Rights Impact Assessment may also be required. Therefore, the existence of an explicit explainability method does not, by itself, render a tool compliant; rather, a broader, context-sensitive legal analysis is necessary to ensure that TSD tools are implemented in a legally sound manner.

4.3.4 TSD on CEPs—a closer look on perspective API and TrollWallAI

Out of the CEPs that have been reviewed for the purposes of this work, only two (YourPriorities,[6] CEEDW[7]) state that a TSD is employed on their platform. Such an observation is problematic, due to the fact that these TSD tools, Perspective API, and TrollWallAI, respectively, are not explainable and Perspective API is not even open source, as stated in Table 2. Nevertheless, a more thorough look at these two tools, Perspective API and TrollWallAI, is more than necessary to further understand the benefits that render them suitable for CEP deployment as well as their limitations and ethical and legal implications.

[6] https://www.yrpri.org

[7] https://ceeddw.org

4.4 Benefits

Perspective API presents high performance as was concluded via holdout evaluations across 18 different languages. More precisely, before training, a percentage of the annotated (toxic/non-toxic) comments are omitted from the training dataset. Subsequently, the trained model's scores are compared against the labels of the left-out comments. Depending on the language, these scores vary, for instance, in languages with little data, the scores are lower. Out of 18 languages, English has the best performance results, while Indonesian has the lowest scores. Moreover, great effort has been made to conduct a toxicity bias evaluation over 18 languages, reporting very low bias scores on almost all languages. Overall, via the model's website, the developers of Perspective API are quite transparent regarding the toxicity rates achieved across different languages and tasks. While Perspective API's publication of toxicity rates by language is commendable and aligns with best practices in transparency, there is no specific EU legal provision that explicitly mandates vendors to publish exactly these metrics. Therefore, you might frame this as a voluntary transparency measure that supports compliance, rather than implying it is strictly required by law.

In the case of TrollWall AI, many benefits are reported in TrollWallAI's official website, including fast real-time content moderation providing notifications/recommendations for the user to take action in the case of toxic content. Among its functionalities, TrollWall AI provides statistics regarding the amount of hidden/non-hidden content on a platform and their sources, while it can learn online and adapt from human moderator comments.

4.5 Limitations

As for the limitations of Perspective AI, the model may contain unintended bias stemming from the data used for its training. Hence, biases in the training data may be reflected in the predictions of the model. As cited in Perspective API's website, the developers of the tool are working diligently to mitigate this issue. Furthermore, the tool does not incorporate any explainability mechanisms, which hinders the model's transparency regarding its decision-making processes. According to the above legal analysis, transparency in TSD models is more than necessary especially in the case of CEPs. Another limitation of Perspective API is that it is closed-source, rendering AI developers and researchers unable to modify/re-train it based on the needs of the platform (either social or civic engagement one) that this TSD tool will be employed. TrollWall AI is also closed-source and not explainable. However, unlike Perspective API, the developers of TrollWall AI do not expose any accuracy or

bias scores to be used for comparative analysis with other TSD tools.

4.6 Ethical and Legal shortcomings

DSA transparency obligations
Articles 14 and 15 of the DSA require platforms to publish clear content moderation policies and provide transparent reporting on the accuracy and error rates of automated systems. YourPriorities, which integrates Perspective API, provides general Terms of Service prohibiting 'unlawful, harmful, threatening, abusive, harassing […] or otherwise objectionable' content but fails to define 'toxic speech' with sufficient precision. As Perspective API only outputs a numerical toxicity score without contextual explanation, the platform must retrofit moderation justifications, complicating compliance with the DSA's 'statement of reasons' requirement. TrollWallAI, similarly, does not provide public criteria or explanations for its moderation decisions, leaving the deployer unable to fully meet the DSA's transparency expectations.

GDPR right to explanation and human review
Under Articles 13, 14, and 22 of the GDPR, data controllers must provide 'meaningful information' about the logic of automated decision-making and allow users to contest decisions and request human review. Neither Perspective API nor TrollWallAI offers built-in explainability mechanisms or discloses their use of automated moderation in a way that enables effective human oversight. For YourPriorities (Perspective API), the privacy policy does not mention the existence of automated moderation or any human intervention procedures. This severely limits users' ability to exercise their GDPR rights. TrollWallAI also does not publish information enabling platforms to satisfy these legal duties.

AI act documentation and oversight requirements
Articles 13 and 14 of the proposed AI Act impose documentation, risk management, and transparency obligations for high-risk AI systems used in public discourse. Based on the information publicly available on their websites, it is unclear whether either Perspective API or TrollWallAI provides sufficient technical documentation to meet the provider obligations set out by the AI Act.

Ethical implications: erosion of trust, accountability, and chilling effects
The closed-source nature of Perspective API precludes independent auditing and public scrutiny of its decision-making processes, which can fuel suspicion about hidden biases or inconsistent enforcement. In both cases, the absence of human-readable justifications for content removals prevent users from meaningfully challenging decisions. This undermines civic trust and risks chilling legitimate speech, particularly from vulnerable or marginalized groups, ultimately weakening democratic participation and the legitimacy of CEPs.

With our state-of-the-art review of XAI TSD tools, we have shown what steps have been taken so far towards automating the transparent AI-driven detection of toxic language on online platforms, to create a safe and inclusive environment for all users regardless of which societal group they belong to. Nevertheless, when adopting TSD mechanisms in CEPs, a combination of RNN-BERT-type tool with a mechanism of highlighting toxic words such as an XAI like LIME would be proved beneficial for such platforms, since this approach ensures both high accuracy and end-user transparency and explainability. This feature would allow users to see precisely which terms or phrases triggered the moderation. For CEPs, this combination of accuracy and explainability could enhance trust and democratic engagement, making it particularly relevant for a paper focused on the distinctive content moderation needs of CEPs.

5 Conclusion and outlook

Our multidisciplinary review highlights a critical misalignment between the current landscape of AI-driven TSD tools and the normative requirements of CEPs. Most existing tools have been developed for social media contexts and lack the transparency, explainability, and contextual adaptability necessary for the democratic and legally accountable environment of CEPs. Although some tools incorporate explainability methods such as LIME or SHAP, these remain the exception rather than the rule, and even among those labeled as 'explainable' rigorous evaluation standards are often missing. Particularly, the only TSD tools evaluated or used on CEPs—Perspective API and TrollWallAI—are proprietary, non-explainable. The legal addendum contributes with the remark that even explainability of tools does not guarantee compliance with legal frameworks, such as the DSA, GDPR, and AI Act. Rather, a holistic, context-sensitive integration of TSD tools into platform governance structures is essential. Based on the gap of legally compliant, explainable, and domain-evaluated TSD tools, we provided a research agenda with short- and long-term goals to be investigated within each discipline.

Building on our review of the legal, psychological, and technical challenges of TSD in CEPs, we call for future research directions, which address the gaps in explainability of AI TSD tools and their ethical and legally compliant deployment. We, therefore, propose short-term and long-term goals from each perspective.

Future psychological research must explore how users perceive and respond to explainable TSD tools in civic engagement contexts.

Short-term goal: Investigate how different user groups (e.g., marginalized vs. majority users) understand and trust various forms of XAI-generated explanations. In addition, examine how perceived fairness and degrees of automation in moderation (fully human, hybrid, or fully automated) influence civic participation and user engagement on CEPs.
Long-term goal: Assess the broader psychological implications of TSD tools applied on CEP on user diversity and the equal representation of social groups on the platform.

Future technical research aims to improve evaluation and standardization of explainability in TSD tools on CEP.

Short-term goal: Perform explainability evaluation, using existing objective metrics and correlate with the understanding of CEP platform users via questionnaires.
*Long-term goal***:** Propose strict standards that should be fulfilled via a unified explainability evaluation framework.

Legal research aims to address regulatory gaps and propose governance solutions for the ethical deployment of TSD systems on CEPs.

Short-term goal: Ensure user-facing transparency and meaningful explanations for moderation decisions, in line with GDPR, DSA, and AI Act requirements. Platforms must establish human-in-the-loop oversight policies, and providers of AI TSD tools should publish the technical documentation necessary for conducting a DPIA and a Fundamental Rights Impact Assessment under AI Act. A specific Fundamental Rights Impact Assessment template should be developed for the use of AI TSD tools in CEPs, aiming to identify and implement concrete technical and organizational measures that guarantee fair and transparent detection of toxic speech. These measures must also safeguard freedom of expression and ensure transparent processes for political discourse and fair public dialog.
*Long-term goal***:** Advance research on the development of open documentation and audit-friendly standards to enable regulatory compliance and enhance public accountability. Institutionalize civic transparency reporting as a legal obligation, in line with DSA and the principles of digital constitutionalism, to ensure that CEPs regularly disclose detailed information on moderation activities and automated decision-making processes. This reporting must also address the specific role of CEPs in safeguarding freedom of expression and ensuring fair and open political discourse. Further research should specify the potential role of participatory governance mechanisms—such as user panels, citizen advisory boards, or ombudspersons—in monitoring the platform's compliance with legal frameworks and the responsible use of AI TSD tools. Such mechanisms can promote democratic accountability and inclusiveness by providing structured opportunities for stakeholders to oversee and contest moderation practices.

Acknowledgements The authors of this work are funded by the European Union, as part of the project Artificial Intelligence to enhance civic participation (ITHACA; Horizon Grant Agreement No. 101094364).

Author contributions M.Z., I.L., P.Z., M.B., and M.D. wrote the main manuscript text and revised it accordingly. I.L. and P.Z. prepared Fig. 1 and Table 1. I.L, P.Z., and M.Z. contributed to and created Table 2. All authors (M.Z., I.L., P.Z., M.B., M.D., C.N., D.A., K.M.) reviewed the manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

A.W.S. (2024) Amazon transcribe toxicity detection. https://aws.amazon.com/transcribe/toxicity-detection/. Retrieved 10 September 2024

Abadi M, Chu A, Goodfellow I et al (2016) Deep learning with differential privacy. In: 2016 ACM SIGSAC conference, pp 308–318. https://doi.org/10.1145/2976749.2978318

Abokhodair N, Skop Y, Rüller S et al (2024) Opaque algorithms, transparent biases: automated content moderation during the Sheikh Jarrah crisis. First Monday. https://doi.org/10.5210/fm.v29i4.13620

ADL (2019) Online hate and harassment: the American experience | ADL. https://www.adl.org/resources/report/online-hate-and-harassment-american-experience

Alkomah F, Ma X (2022) A literature review of textual hate speech detection methods and datasets. Information 13(6):273

Anjum KR (2024) Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities. Int J Inf Secur 23(1):577–608

Ansari G, Kaur P, Saxena C (2024) Data augmentation for improving explainability of hate speech detection. Arab J Sci Eng 49(3):3609–3621

Appelman N (2021) Using terms and conditions to apply fundamental rights to content moderation. Verfassungsblog: On matters constitutional, Fachinformationsdienst für internationale und interdisziplinäre Rechtsforschung. https://doi.org/10.17176/20210901-233103-0

Aroyo L, Dixon L, Thain N et al (2019) Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In: Companion proceedings of the 2019 world wide web conference. WWW'19: the web conference. ACM, San Francisco, USA, pp 1100–1105. ISBN: 9781450366755. https://doi.org/10.1145/3308560.3317083

Aurpa TT, Sadik R, Ahmed MS (2022) Abusive Bangla comments detection on Facebook using transformer-based deep learning models. Soc Netw Anal Min 12(1):24

Banerjee P, Barnwal RP (2022) Methods and metrics for explaining artificial intelligence models: a review. Springer International Publishing, pp 61–88. https://doi.org/10.1007/978-3-031-12807-3

Berntzen L, Johannessen MR (2016) The role of citizen participation in municipal smart city projects: lessons learned from Norway. In: Gil-Garcia JR, Pardo TA, Nam T (eds) Smarter as the new urban agenda, vol 11. Springer International Publishing, New York, pp 299–314

Bonetti A, Martínez-Sober M, Torres JC et al (2023) Comparison between machine learning and deep learning approaches for the detection of toxic comments on social networks. Appl Sci 13(10):6038

Brehm SS, Brehm JW (1981) Chapter 1—introduction: freedom, control, and reactance theory. In: Brehm SS, Brehm JW (eds) Psychological reactance. Academic Press, USA, pp 1–7

Breitfeller L, Ahn E, Jurgens D et al (2019) Finding microaggressions in the wild: a case for locating elusive phenomena in social media posts. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp 1664–1674

Butler J (1997) The psychic life of power: theories in subjection. Standford UP, USA

Carnaghi A, Maass A (2007) In-group and out-group perspectives in the use of derogatory group labels: gay versus fag. J Lang Soc Psychol 26(2):142–156. https://doi.org/10.1177/0261927X07300077

Chhabra A, Vishwakarma DK (2023) A literature survey on multimodal and multilingual automatic hate speech identification. Multimedia Syst 29(3):1203–1230

Cho S, Mossberger K, Swindell D et al (2021) Experimenting with public engagement platforms in local government. Urban Affairs Rev 57(3):763–793. https://doi.org/10.1177/1078087419897821

Clark K, Luong MT, Le QV et al (2020) Electra: pre-training text encoders as discriminators rather than generators

Cobbe J (2021) Algorithmic censorship by social platforms: power and resistance. Philos Technol 34(4):739–766. https://doi.org/10.1007/s13347-020-00429-0

Cook CL, Patel A, Wohn DY (2021) Commercial versus volunteer: Comparing user perceptions of toxicity and transparency in content moderation across social media platforms. Front Hum Dyn 3:626409. https://doi.org/10.3389/fhumd.2021.626409

D'Sa AG, Illina I, Fohr D (2020) BERT and fast text embeddings for automatic detection of toxic speech. In: 2020 International multi-conference on: "organization of knowledge and advanced technologies" (OCTA), pp 1–5. https://doi.org/10.1109/OCTA49274.2020.9151853

Davies J, Procter R (2020) Online platforms of public participation—a deliberative democracy or a delusion? https://doi.org/10.48550/ARXIV.2009.14074

De Gregorio G (2020) Democratising online content moderation: a constitutional framework. Comput Law Secur Rev 36:105374. https://doi.org/10.1016/j.clsr.2019.105374

De Gregorio G (2022) Digital constitutionalism and freedom of expression. In: Cambridge studies in European law and policy. Cambridge University Press, pp 157–215

Dessi D, Recupero DR, Sack H (2021) An assessment of deep learning models and word embeddings for toxicity detection within online textual comments. Electronics 10(7):779

Devlin J, Chang MW, Lee K et al (2019) Bert: pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805

DeYoung J, Jain S, Rajani N et al (2020) Eraser: a benchmark to evaluate rationalized NLP models. arXiv:1911.03429

Dillard JP, Shen L (2005) On the nature of reactance and its role in persuasive health communication. Commun Monogr 72(2):144–168. https://doi.org/10.1080/03637750500111815

ElSherief M, Nilizadeh S, Nguyen D et al (2018) Peer to peer hate: hate speech instigators and their targets. Proc Inter AAAI Conf Web Soc Media. https://doi.org/10.1609/icwsm.v12i1.15038

ElSherief M, Ziems C, Muchlinski D et al (2021) Latent hatred: a benchmark for understanding implicit hate speech. In: Moens MF, Huang X, Specia L et al (eds) Proceedings of the 2021 conference on empirical methods in natural language processing. Association for Computational Linguistics, USA, pp 345–363

Falco E, Kleinhans R (2018) Beyond technology: Identifying local government challenges for using digital platforms for citizen engagement. Int J Inf Manage 40:17–20. https://doi.org/10.1016/j.ijinfomgt.2018.01.007

Fan H, Du W, Dahou A et al (2021) Social media toxicity classification using deep learning: real-world application UK Brexit. Electronics 10(11):1332

Farinho DS (2024) Personal data processing by online platforms and search engines: the case of the EU digital services act. Pub Govern Admin Fin L Rev 9:37

Flew T, Martin F, Suzor N (2019) Internet regulation as media policy: rethinking the question of digital communication platform governance. J Digit Media Policy 10(1):33–50. https://doi.org/10.1386/jdmp.10.1.331

Gil de Zúñiga H, Jung N, Valenzuela S (2012) Social media use for news and individuals' social capital, civic engagement and political participation. J Comput Mediat Commun 17(3):319–336. https://doi.org/10.1111/j.1083-6101.2012.01574.x

Gongane V, Munot M, Anuse A (2024) A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms. J Comput Soc Sc 7:587–623. https://doi.org/10.1007/s42001-024-00248-9

Grimminger L, Klinger R (2021) Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In: De Clercq O, Balahur A, Sedoc J et al (eds) Proceedings of the eleventh workshop on computational approaches to subjectivity, sentiment and social media analysis. Association for Computational Linguistics, USA, pp 171–180

Guldvik I, Askheim OP, Johansen V (2013) Political citizenship and local political participation for disabled people. Citizenship Stud 17(1):76–91. https://doi.org/10.1080/13621025.2013.764219

Halevy M, Harris C, Bruckman A et al (2021) Mitigating racial biases in toxic language detection with an equity-based ensemble framework. https://doi.org/10.48550/ARXIV.2109.13137

Haythornthwaite C (2023) Moderation, networks, and anti-social behavior online. Soc Media Soc. https://doi.org/10.1177/20563051231196874

Hoang NM, Do XL, Do DA et al (2024) Toxcl: a unified framework for toxic speech detection and explanation. https://arxiv.org/abs/2403.16685

Hutchinson B, Prabhakaran V, Denton E et al (2020) Social biases in NLP models as barriers for persons with disabilities. Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics, USA, pp 5491–5501

Ibrahim MA, Arifin S, Yudistira IGAA et al (2022) An explainable AI model for hate speech detection on Indonesian twitter. CommIT (Commun Inf Technol) J 16(2):175–182

Jahan MS, Oussalah M (2023) A systematic review of hate speech automatic detection using natural language processing. Neurocomputing 546:126232. https://doi.org/10.1016/j.neucom.2023.126232

Jakob J, Dobbrick T, Freudenthaler R et al (2023) Is constructive engagement online a lost cause? Toxic outrage in online user comments across democratic political systems and discussion arenas. Commun Res 50(4):508531. https://doi.org/10.1177/00936502211062773

Jhaver S, Bruckman A, Gilbert E (2019) Does transparency in moderation really matter?: user behavior after content removal explanations on reddit. Proc ACM Hum-Comput Interact 3:1–27. https://doi.org/10.1145/3359252

Joulin A, Grave E, Bojanowski P et al (2017) Bag of tricks for efficient text classification. In: Lapata M, Blunsom P, Koller A (eds) Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 2, short papers. Association for Computational Linguistics, Valencia, Spain, pp 427–431

Kachanoff FJ, Gray K, Koestner R et al (2022) Collective autonomy: why groups fight for power and status. Soc Pers Psychol Compass 16(1):e12652

Kibriya H, Siddiqa A, Khan WZ et al (2024) Towards safer online communities: deep learning and explainable AI for hate speech detection and classification. Comput Electr Eng 116:109153. https://doi.org/10.1016/j.compeleceng.2024.109153

Kirk HR, Yin W, Vidgen B et al (2023) Semeval-2023 task 10: explainable detection of online sexism. https://arxiv.org/abs/2303.04222

Kuran CHA, Morsut C, Kruke BI et al (2020) Vulnerability and vulnerable groups from an intersectionality perspective. Int J Disaster Risk Reduct 50:101826. https://doi.org/10.1016/j.ijdrr.2020.101826

Liu Y, Ott M, Goyal N, et al (2019) RoBERTta: a robustly optimized BERT pretraining approach. https://arxiv.org/abs/1907.11692

Löfgren Nilsson M, Örnebring H (2016) Journalism under threat: intimidation and harassment of Swedish journalists. J Pract 10(7):880–890. https://doi.org/10.1080/17512786.2016.1164614

Mahajan A, Shah D, Jafar G (2021) Explainable AI approach towards toxic comment classification. Emerging technologies in data mining and information security: proceedings of IEMIS 2020, vol 2. Springer, New York, pp 849–858

Malik P, Aggrawal A, Vishwakarma DK (2021) Toxic speech detection using traditional machine learning models and BERT and fastText embedding with deep neural networks. In: 2021 5th international conference on computing methodologies and communication (ICCMC). IEEE, pp 1254–1259

Mathew B, Saha P, Yimam SM et al (2021) Hatexplain: a benchmark dataset for explainable hate speech detection. In: Proceedings of the AAAI conference on artificial intelligence, pp 14867–14875

Mehta H, Passi K (2022) Social media hate speech detection using explainable artificial intelligence (xai). Algorithms 15(8):291

Molina MD, Sundar SS (2022) When AI moderates online content: effects of human collaboration and interactive transparency on user trust. J Comput Mediat Commun 27(4):zmac010. https://doi.org/10.1093/jcmc/zmac010

Nguyen DQ, Vu T, Nguyen AT (2020) BERTweet: a pre-trained language model for English Tweets. arXiv preprint arXiv:200510200

One AI (2024) Toxicity detection skill. https://oneai.com/skill/toxicity. Retrieved 10 September 2024

Pavlopoulos J, Sorensen J, Laugier L et al (2021) Semeval-2021 task 5: toxic spans detection. In: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021). Association for Computational Linguistics, USA. https://doi.org/10.18653/v1/2021.semeval-1.6

PerspectiveAPI (2023) https://www.perspectiveapi.com/. Retrieved 10 September 2024

Poletto F, Basile V, Sanguinetti M et al (2021) Resources and benchmark corpora for hate speech detection: a systematic review. Lang Resour Eval 55(2):477–523. https://doi.org/10.1007/s10579-020-09502-8

Protect AI (2025) LLM guard—the security toolkit for LLM interactions. https://llm-guard.com/

Raisio H, Valkama K, Peltola E (2014) Disability and deliberative democracy: towards involving the whole human spectrum in public deliberation. Scand J Disabil Res 16(1):77–97. https://doi.org/10.1080/15017419.2013.781957

Read B (2018) Truth, masculinity and the anti-elitist backlash against the university in the age of trump. Teach Higher Educ 23(5):593–605. https://doi.org/10.1080/13562517.2018.1457636

Ribeiro MT, Singh S, Guestrin C (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, KDD'16. Association for Computing Machinery, New York, NY, USA, pp 1135–1144

Risch J, Krestel R (2020) Toxic comment detection in online discussions. In: Agarwal B, Nayak R, Mittal N et al (eds) Deep learning-based approaches for sentiment analysis. Algorithms for intelligent systems. Springer, Singapore

Rojszczak M (2023) The digital services act and the problem of preventive blocking of (clearly) illegal content. Inst Admin J Admin Sci 3(2):44–59

Saha K, Chandrasekharan E, De Choudhury M (2019) Prevalence and psychological effects of hateful speech in online college communities. In: Proceedings of the 10th ACM conference on web science. ACM, pp 255–264

Salminen J, Hopf M, Chowdhury SA et al (2020) Developing an online hate classifier for multiple social media platforms. HCIS 10:1–34

Sarker J, Sultana S, Wilson SR et al (2023) ToxiSpanSE: an explainable toxicity detection in code review comments. In: 2023 ACM/IEEE international symposium on empirical software engineering and measurement (ESEM). IEEE, pp 1–12. https://doi.org/10.1109/esem56168.2023.10304855

Schmidt A, Wiegand M (2017) A survey on hate speech detection using natural language processing. In: Ku LW, Li CT (eds) Proceedings of the fifth international workshop on natural language processing for social media. Association for Computational Linguistics, USA, pp 1–10

Shakil MH, Alam MGR (2022) Hate speech classification implementing NLP and CNN with machine learning algorithm through interpretable explainable AI. In: 2022 IEEE region 10 symposium (TENSYMP). IEEE, pp 1–6

Silva L, Mondal M, Correa D et al (2021) Analyzing the targets of hate in online social media. Proc Int AAAI Conf Web Soc Media 10(1):687–690. https://doi.org/10.1609/icwsm.v10i1.14811

Smirnov I, Oprea C, Strohmaier M (2023) Toxic comments are associated with reduced activity of volunteer editors on wikipedia. PNAS Nexus 2(12):pgad385. https://doi.org/10.1093/pnasnexus/pgad385

Sunstein CR (1995) Democracy and the problem of free speech. Publ Res Q 11:58–72

Suzor NP, West SM, Quodling A et al (2019) What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. Int J Commun 13:18

TrollWall AI (2025) Trollwall AI: AI-powered community management solution. https://trollwall.ai

Vaughan-Johnston TI, Nguyen A, Jacobson JA (2024) A surprising lack of consequences when constraining language. Front Soc Psychol 2:1260974. https://doi.org/10.3389/frsps.2024.1260974

Ventura F, Greco S, Apiletti D et al (2022) Trusting deep learning natural-language models via local and global explanations. Knowl Inf Syst 64(7):1863–1907

Wachs S, Gámez-Guadix M, Wright MF (2022) Online hate speech victimization and depressive symptoms among adolescents: the protective role of resilience. Cyberpsychol Behav Soc Netw 25(7):416–423. https://doi.org/10.1089/cyber.2022.0009

Waseem Z, Davidson T, Warmsley D et al (2017) Understanding abuse: a typology of abusive language detection subtasks. http://arxiv.org/abs/1705.09899

Werthner H, Ghezzi C, Kramer J et al (2024) Introduction to digital humanism: a textbook. Springer, Cham

Yee K, Sebag AS, Redfield O et al (2022) A keyword based approach to understanding the overpenalization of marginalized groups by English marginal abuse models on twitter. https://doi.org/10.48550/arXiv.2210.06351

Young GK (2022) How much is too much: the difficulties of social media content moderation. Inf Commun Technol Law 31(1):1–16

Zait A, Andrei AG (2019) Civic engagement at the crossroads of online and offline spaces: a PLS-SEM assessment. Sci Ann Econ Bus 66(4):559–572

Ziegele M, Jost P, Bormann M et al (2018) Journalistic counter-voices in comment sections: patterns, determinants, and potential consequences of interactive moderation of uncivil user comments. Stud Commun 7(4):525–554. https://doi.org/10.5771/2192-4007-2018-4-525

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.