AI-Based Offensive Speech Detection in Social Media

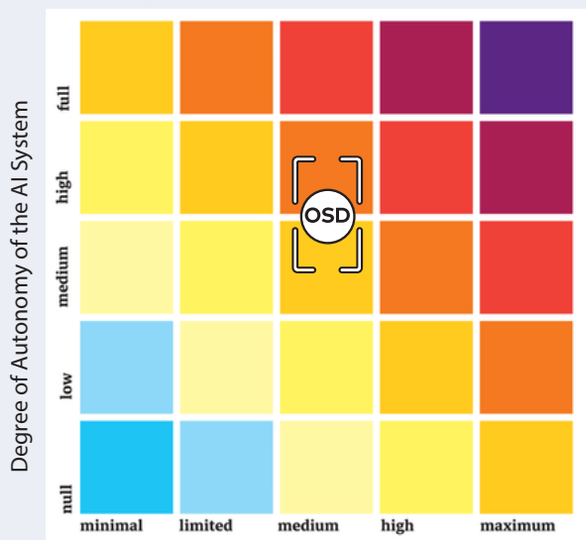
Executive Summary

Offensive speech and in extreme cases hate speech, are a major challenge in digital communication, particularly on online platforms on which content is shared.[2] Hate speech differs from offensive speech in that it targets very specific individuals or groups in order to discriminate against and demean them on the basis of specific characteristics such as origin, religion, gender, sexual orientation, and others. Offensive and hate speech are expressed through texts, images, symbols or gestures. Its widespread occurrence and the tension between the violation of human rights and restrictions on freedom of speech render the issue highly challenging from an ethical and social perspective. Due to the large volume of problematic content and the legal obligation in Europe to take action against hate and incitement, online platforms are increasingly using AI systems for the automated detection and removal of offensive and hate speech.[1,5,10,20]

From a technical perspective, a clear and fair assessment of language by AI systems is challenging, as the evaluation of offensive statements must take place within the respective context and is often challenging even for humans. Furthermore, AI systems struggle to recognise irony and sarcasm. Consequently, there is a risk of insufficient protection against problematic and illegal statements. Conversely, this can also lead to the unjustified filtering and removal of content, which, from the users' perspective, may curtail the right to freedom of speech.[10]

In the ALAIT Risk Radar, the use of AI systems to detect and remove offensive comments and hate speech is classified as posing a high to medium overall risk (dark orange and light orange). With regard to the level of autonomy of the AI system, the AI systems used generally have a medium to high level of autonomy (Levels 3 and 4), as human review only takes place in problematic cases, such as when users raise concerns or when findings are unclear. The risk in the context of use is classified as 'medium' (level 3), as whilst there is generally no immediate danger to life or limb, AI systems that operate incorrectly or inaccurately can have a significant impact on people's fundamental rights, such as protection against discrimination, personal safety, physical integrity and freedom of expression. In Austria, citizens can take action against digital hate speech that insults or harms them, or is illegal for other reasons, and file a charge. The European Digital Services Act (DSA) also obliges platform providers to actively take action against hate speech and to protect marginalised groups. For example, it must be ensured that content moderators monitor systems' functioning and output. Content moderators need to be continuously trained – both in the functioning and limitations of AI systems as well as in relation to social sensitivities. In January 2025, the DSA was expanded by the "Code of Conduct on Countering Illegal Hate Speech Online", which goes be-

ALAIT Risk Radar for AI-based detection of offensive speech



AI Applications in the Field of Offensive Speech Detection

The ALAIT Risk Radar is a scientifically developed risk analysis tool for artificial intelligence (AI) that classifies AI applications in a context-specific manner, taking into account their degree of technical autonomy, thereby making the risk landscape visible to users at a glance. The principle is as follows: the higher the operational risk arising from the application context and the greater the degree of autonomy of the AI system in terms of decision-making, the higher the risk associated with its use. An extended bracket indicates a range within the risk classification. A low degree of autonomy in an AI system does not mean that one can simply sit back and relax. It requires a strong role for the people using it. (For details on the tiered model, see p. 8ff).

**ALAIT is a research project initiated by the Federal Ministry for Innovation, Mobility and Infrastructure (BMIMI) to develop a novel form of technology assessment for AI technologies using scientific methods.*

yond voluntary commitments to establish concrete operational guidelines for platform operators.

Introduction

Offensive and hate speech pose particular challenges on online platforms where users share content such as texts, images or videos. Unlike offensive speech, hate speech is directed towards members of a particular group. Studies[1] show that women, Muslims, Jews and people of colour are particularly affected by hate speech online. Offensive and hate speech can lead to psychological distress and social exclusion, and in extreme cases, physical harm.[2] They restrict freedom of expression by creating a climate of fear and through unfair filtering or censorship on social media. This can reinforce social polarisation and normalise discriminatory attitudes, making them appear increasingly acceptable. In some social groups, this can foster radicalisation processes because hostile narratives are perceived as legitimate and thus reinforced. Hate speech can manifest in language, images, videos, caricatures, emojis and symbols. Generative AI technologies increase the potential to create and disseminate offensive and hateful content on social media.

Generative AI technologies are increasing the scope for the creation and dissemination of offensive speech and hate speech on social media, for example in the form of so-called 'deepfakes': media content such as images, videos or audio files that appear realistic and are perceived as genuine, but have in fact been created or edited using AI and depict real or fictional people in situations or making statements that are fabricated.[1,3]

As both freedom of expression and user safety must be balanced, the effective detection and prevention of offensive comments and hate speech is essential. Automated content moderation, which scans large volumes of user-generated content for harmful and illegal material and removes it where necessary, is indispensable in practice. The AI systems used are intended to benefit three groups:

1. Operators of online platforms, who can thereby significantly increase the volume of content moderation.
2. Content moderators, who can be protected from psychologically challenging tasks by automation.
3. Users, who can ideally experience social media without offensive, illegal or violent content

However, these three groups are counterbalanced by three other groups who are disadvantaged by the automation described:

1. The data labellers, who often have to label emotionally distressing data – which serves as training data for these AI systems – under precarious conditions in the Global South.

2. Marginalised groups whose content is often wrongly classified as hate speech due to biases in AI-powered content moderation, and is therefore more frequently censored.

3. Democracy as a whole, along with its democratically legitimised institutions. This is because the AI systems used by the operators of platforms and forums, which are usually not made transparent and are therefore not subject to democratic control, decide which content is classified as "good" or "bad" and what may and may not be said

Technology Description

AI-based detection of offensive comments relies on advances in machine learning (ML), particularly in the field of natural language processing (NLP), which deals with textual content. At its core, this technology involves training algorithms using large datasets containing labelled content (***supervised learning***), so that they learn to automatically classify new content as 'acceptable' or 'in breach' of a specific policy (e.g. offensive language vs. normal language).[6] Whilst such systems can obviously flag hateful words, they struggle to interpret text correctly – for example, in terms of context – and also have difficulty with spelling mistakes, sarcasm and irony, as well as seemingly harmless code words that carry specifically hostile meanings. Generally speaking, they perform better in English than in other languages due to the large volume of English training data. There are various techniques designed to help mitigate these challenges. Among others, the following two techniques are used:

1. Named Entity Recognition (NER) and Entity Linking (EL) are used to identify identity groups in texts and determine exactly which words refer to the target groups, linking them to an external knowledge database.[26,27] These technologies are designed to help identify the subject matter of the post and who/what is being attacked.
2. Sentiment and emotion analyses are used to distinguish real hate speech from other forms of offensive language.[28,29]

Advances in performance are achieved in particular through the use of transformer models (e.g. BERT, RoBERTa, XLM, GPT), which can recognise contextual relationships between words, e.g. in sentences, and are therefore suitable for language analysis.[7,8] However, these advances mainly apply to the English language and are less applicable to other languages such as German or Italian. In addition, these models also contain biases that often disadvantage particularly marginalised groups even further.

In practice, primarily language-based systems are used. Since offensive statements go beyond purely textual content, there is a need for multimodal systems that can also analyse images or videos and are needed for computer vision technologies.[9] However, the development of such multimodal systems poses major challenges in terms of fairness, transparency and contextual understanding.

Opportunities

The use of AI to detect offensive comments on online platforms has the potential to benefit users, content moderators and social media providers alike:[6,8,10]

Opportunities for Users:

- **Enhanced user experience on social media:** The integration of AI into content moderation – even independently of the detection of offensive comments – can help to ensure a personalised experience and reduce the volume of spam, disinformation and content ranging from disturbing to illegal. Combined with the option, enshrined in the Digital Services Act (DSA), to choose the criteria by which content is moderated, this can improve the overall experience and (ideally) reduce the risk of users being exposed to hateful and harmful content.[3,12]

Opportunities for Content Moderators:

- **Lower psychological stress and workload:** AI tools can obscure sensitive images and censor offensive language in real time. Consequently, the volume of material reaching content moderators decreases, shielding them from the most disturbing content. This alleviates the psychological strain of the job.[3,13]

Opportunities for Online Platform Operators:

- **Speed and scope:** According to its own statements, over 500 million posts are generated on platform X (formerly Twitter) every day.[11] fficient moderation of this content can only be conducted through the integration of AI systems that can automatically detect and remove the majority of obviously illegal and offensive content within minutes of its publication.[10]
- **Consistency and accuracy:** When properly trained and configured, AI systems can apply a high degree of consistency to all analysed content. Research shows that transformer models such as RoBERTa, and specialised variants such as HateBERT (a BERT model trained on hateful or offensive speech), can perform well in detecting hate speech.[12]

Challenges and Risks

Despite the opportunities mentioned above, there are a number of risks and challenges that online platform providers must address in order to ensure a trustworthy and reliable integration of AI for detecting offensive speech:

1. **Many false positives and false negatives as well as *AI Bias***: Although AI systems are designed to recognise and delete offensive comments, they often make mistakes. In particular, AI systems struggle to recognise irony and sarcasm.[8,15] In multimodal systems, the problem lies in the fact that images do not have a clear semantic context like text, but rely on complex visual cues, symbolism and cultural understanding that the systems cannot recognise. Errors in a classification system can be categorised as "false positives" (when a harmless post is mistakenly identified as hate speech) or "false negatives" (when actual hate speech is not recognised by the AI system). Such errors often occur more prominently among marginalised or vulnerable groups, leading to further discrimination against these groups. Studies have shown that automated hate speech detectors often mislabel content written in African American English (AAE) or minority dialects as "toxic".[8,17] This can lead to the reinforcement of prejudices and even censorship of these groups. Conversely, an accumulation of false negative results can cause some groups to remain exposed to a high level of hate and insults, which can lead to psychological distress and withdrawal from online forums.

2. **Restriction of freedom of speech:** Article 19 of the International Covenant on Civil and Political Rights (ICCPR) enshrines the protection of freedom of speech, while Article 20(2) prohibits any incitement to national, racial or religious hatred that leads to discrimination, hostility or violence.[1] The filtering and deletion of content therefore directly conflicts with the right to freedom of speech. AI moderation tools can be misused by governments or social media providers to significantly restrict fundamental individual freedoms. On social media, AI systems have been shown to restrict even harmless content – due to false positive classification – and thus limit the freedom of speech of their users.[10]

3. **Lack of contextual understanding:** One of the biggest challenges in automatically detecting offensive language and hate speech is that AI lacks contextual understanding.[14] Expressions such as sarcasm, metaphors, irony and coded language are generally difficult for AI systems to recognise.[15] Coded language involves using code words for ethnic insults that are clearly understood within a particular community, but not by outsiders. Hateful comments can thus bypass filtering.[16] The performance of an AI model is always based on test data containing a certain proportion of sarcasm, irony, or coded language. In practice, however, many users know how to exploit sarcasm and coded statements to bypass AI-supported moderation. Consequently, AI systems for detecting offensive speech can only partially ensure that hate speech is identified.

4. **Lack of *transparency:*** Decisions made by AI systems are generally opaque and difficult to understand or verify, due to factors including their architecture, the enormous volumes of data and the complexity of the systems themselves, as well as the secrecy surrounding proprietary data and software. This makes the verification process labour-intensive and time-consuming, or simply impossible (***black box problem***).[18] This contradicts users' rights under Article 17 of the

Digital Services Act (DSA). The General Data Protection Regulation (GDPR) also requires an explanation for the deletion of content.[6,19]

5. **Relocation of psychological burden:** A significant issue is the relocation of psychologically demanding work to the Global South. Studies show that so-called data workers in countries with fewer regulations and labour rights often undertake tasks such as data annotation, content moderation and labelling, often under precarious conditions and in informal employment relationships. In doing so, they frequently have to label disturbing images and texts to generate training data for AI systems to automatically detect hate speech. They are poorly paid and lack adequate protection from psychological harm.[32,33,34]

6. **General uncertainty regarding offensive statements:** Most social media providers define their own terms of use and definitions of offensive statements. For example, Meta has set up a department dedicated to dealing with this issue.[4,10] This shifts the decision of what is considered acceptable language/speech to private companies that are not subject to democratic processes. Proprietary, non-transparent AI models thus set the gold standard for what is and is not permitted. However, offensive statements are perceived differently by different individuals, so it is not possible to provide an accurate and complete list of hate speech, as it can be combined and invented. Legislators should therefore use democratic means to address the conflict of values between freedom of expression and the protection of users from hate speech and violence. While a clear definition of hate speech could be used to train AI systems, it would need to be contextualised when applied to specific conditions. Open-source initiatives could provide a democratic alternative by making labelled data from affected communities available for training open-source AI models to detect hate speech, allowing users to choose their own AI model.

Recommendations for Practical Use

To minimise the above-mentioned risks associated with the practical use of AI-powered offensive speech detection, a combination of regulatory measures, technical safeguards and proven moderation procedures is required.

1. **Establish human supervision:** One of the most effective principles for mitigating risk is to ensure that AI does not make final decisions about content that affects users' fundamental rights, such as the right to freedom of speech. AI systems should therefore be developed and deployed in accordance with at least the "human-in-the-loop" principle, which combines AI with human expertise, for example through the randomised testing of AI decisions by moderators. As part of this strategy, moderators should also be trained to recognise the limitations of AI tools as well as legal standards regarding hate speech, and not to blindly trust the AI's warning messages. Additionally, involving moderators or advisors from marginalised and frequently hate speech-affected groups can facilitate an understanding of the context.[1] Combined with ***explainable AI (XAI)***, this approach increases transparency and efficiency for moderators.[7,10,23,26]

2. **Taking cultural and linguistic contexts into account:** At present, content moderation pays little attention to cultural and linguistic contexts, although this is a prerequisite for identifying problematic content. AI systems should be developed and applied to a greater extent for cultural regions outside the Anglo-American linguistic sphere.

3. **Ensuring improved data quality and continuously learning datasets:** The language on online platforms is evolving rapidly, with new slang terms and nuanced usages constantly emerging. Continuous learning frameworks, in which models gradually incorporate newly labelled data (including context), ensure that the recognition algorithms remain up to date. To

avoid ***bias***, companies should also ensure that their training datasets are representative, carefully annotated and include more examples of non-hateful content from minority languages and dialects. This helps the model learn to recognise when a racial slur is being used aggressively. Significant progress has been made in context recognition using NLP models such as Hate-BERT, providing a solid foundation for further AI system development. This has led to improvements in overall accuracy and reductions in systematic biases against protected groups.[8,21]

4. **Use of open-source models:** The use of balanced and transparent open-source models not only improves overall accuracy and reduces systematic bias against protected groups, but also contributes to strengthening user trust.[8,2]

5. **Publishing regular audits and making effective use of the results:** In accordance with Articles 34 and 37 of the Digital Services Act (DSA), platforms are obliged to identify, analyse and evaluate systemic risks such as the dissemination of illegal content, negative impacts on fundamental rights, and potential harm to civil society discourse, electoral processes, public safety, gender-based violence, public health and minors.[30] In addition, platforms are required to disclose the datasets used for training, as well as the AI models, so that their performance can be independently tested. These audits must be published to enhance accountability and must lead to concrete measures and adjustments of the models.[20]

6. **Development of multimodal analysis to identify deepfakes:** In the context of hate speech, detecting ***deepfakes*** requires the development of new technologies that incorporate image analysis alongside text analysis. Organisations such as the 'Coalition for Content Provenance and Authenticity' (C2PA) and practices such as "Open Source Intelligence" (OSINT) are already taking the first steps towards this. These organisations combat the spread of misleading information on the internet by establishing technical standards for certifying the source and history of media content.[24,25]

7. **Establishing complaint procedures for users:** According to the Digital Services Act (DSA), platform providers must provide users with the opportunity to request a review if they disagree with the company's decision to remove or retain controversial content and have exhausted all legal remedies.[31] IIn Germany, there is also a legal framework for this. Since 2017, the Network Enforcement Act (NetzDG) has required social networks to respond appropriately to complaints from users.[10]

Important Terms

Black box systems: Refers to all systems whose internal decision-making processes are not transparent or traceable to humans. This means that only inputs and outputs can be observed, without understanding exactly how the processing in between takes place.

Data protection in AI: The full range of practices and concerns relating to the ethical collection, storage and use of personal data by artificial intelligence systems.

Deepfakes: Realistic-looking media files (photos, audio files or videos) that have been altered, created or falsified using AI techniques.

Explainable AI (XAI): Refers to methods and techniques that make it possible to better understand and account for the decisions and outcomes of AI systems.

Generative AI: Artificial intelligence that can generate new content such as text, images, music or code, based on patterns learnt from training data.

European Union Law on Artificial Intelligence (EU AI-Act): A European regulation on artificial intelligence (AI) – the first comprehensive regulation on AI issued by a major regulatory authority. It focuses in particular on high-risk AI systems.

AI autonomy: The ability of an AI system to achieve a set of goals under a range of uncertainties in its environment, independently and without external intervention.

AI accuracy: Refers to an AI system's ability to make proper predictions or decisions. It is an important measure of its performance and crucial for determining its effectiveness and reliability.

AI bias: Bias refers to the systematic unequal treatment of certain objects, individuals or groups compared to others. Treatment refers to any kind of action, including perception, observation, representation, prediction or decision-making.

Transparency: This means that the functionality, decision-making processes and areas of application of an AI system are comprehensible, explainable and openly accessible – to developers, users and other stakeholders.

Supervised learning: A subfield of machine learning in which the training data is labelled in advance and the system learns the pattern between the content and the label. The task of such an AI system is to find a relationship that maps each input in the training set (the data) to an output (the label).

Explanation of the ALAIT Risk Radar's tiered model

The ALAIT AI Risk Radar illustrates the relationship between application risk and the autonomy of an AI system. The risk levels are based on the EU AI Act, in particular Article 6 and Annex III, which deal with high-risk areas of AI application. Lower levels of system autonomy and application risks are represented by cooler colours (blue), whilst higher levels of system autonomy and application risks are represented by warmer colours (red).

The change of color conveys the increased risk of AI applications. This color scale can be used to identify the overall risk: violet and dark red – very high, red and dark orange – high, light orange and yellow tones – medium, blue tones – low overall risk. Ideally, high application risks and system autonomy should be avoided or only be used after very careful consideration.

Level of Autonomy of the AI System

Level 1: No Autonomy

AI is a passive tool; it is people who make all the decisions and take action.

Example: Diagnostic systems that display raw medical data or analyse the data (without providing recommendations!).

Recommended use cases: Scenarios involving high risks or in which ethical decisions are of crucial importance (e.g. medical diagnostics, the justice system).

Level 2: Low Degree of Autonomy (Human-in-the-Loop)

The AI provides recommendations or options, but users remain responsible for selecting and approving actions.

Example: AI suggests optimal routes for logistics, or recommendation systems in e-commerce.

Recommended use cases: Tasks of medium complexity involving moderate risks (e.g. supply chain optimisation).

Level 3: Medium Degree of Autonomy (Human-on-the-Loop)

The AI carries out certain tasks autonomously, with humans intervening in exceptional cases.

Example: AI-supported manufacturing processes where the system controls the machines, but users intervene in the event of anomalies.

Recommended use cases: Scenarios in which continuous human involvement is not required, but critical risks necessitate human oversight (e.g. industrial automation, monitoring of financial transactions).

Level 4: High Degree of Autonomy (Human in Control)

The AI system operates largely autonomously, but allows users to override it themselves in order to avoid undesirable results.

Example: Autonomous vehicles.

Recommended use cases: Low- to medium-risk environments (e.g. logistics, basic traffic management).

Level 5: Full Autonomy with Minimal Supervision

The AI system operates independently and requires minimal or no human intervention.
Human involvement is limited to long-term audits.

Example: Autonomous agricultural machinery, AI for electricity distribution, underground railways, airport rail links.

Recommended use cases: Environments with low safety or ethical risks and high reliability of the AI system (e.g. repetitive tasks in controlled environments).

Degrees of Application Risk

Level 1: Minimal Risk

The AI system has no impact on users or decision-making.

Examples: Filters, NPCs, recommendation algorithms without serious consequences (DeepL, other translation systems).

Criteria: No direct impact on the high-risk areas covered by the ***EU AI-Acts***.

Stufe 2: Limited Risk

AI systems that interact with users but do not make high-risk decisions. The risk increases when there is a lack of transparency regarding the involvement of AI.

Examples: Chatbots and AI-generated content without disclosure, simple automation tasks.

Criteria: Areas not included in the 'high-risk' list under the EU AI Act.

Level 3: Medium Risk

AI systems do not have any particular impact on individuals, but they do have an effect at a collective or societal level.

Examples: Generative AI such as ChatGPT and other systems that can indirectly influence the environment in which they are used.

Criteria: AI systems that are available for public use and have the potential to influence and, in the long term, change existing practices.

Level 4: High Risk

Any algorithm used in what the EU AI Act defines as 'high-risk areas' or which has a direct impact on individuals.

Examples: Medicine, biometrics, critical infrastructure, education and vocational training, employment, access to public sector services, law enforcement, migration.

Criteria: Classification as a 'high-risk sector' under the EU AI Act is only applicable if the rules on transparency and data quality are complied with.

Level 5: Extreme Risk

Any algorithm used in what the EU AI Act defines as 'high-risk areas'.

Examples: Medicine, biometrics, critical infrastructure, education and vocational training, employment, access to public sector services, law enforcement, migration.

Criteria: Classification as a 'high-risk area' if the rules on transparency and data quality are NOT complied with.

Sources

1 European Union Agency for Fundamental Rights (Ed.). (2023). Online content moderation: Current challenges in detecting hate speech. Publications Office. https://doi.org/10.2811/332335

2 Nations, U. (2025, June 27). What is hate speech? United Nations; United Nations. https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

3 Horan, S. (2025, March 12). Deepfake Dangers: How Content Moderation Teams Can Fight AI-Generated Misinformation. Zevo Health. https://www.zevohealth.com/blog/deepfake-dangers-how-content-moderation-teams-can-combat-ai-generated-misinformation/ Accessed 22 July 2025

4 Meta. (2020, July 1). Sharing Our Actions on Stopping Hate. Meta for Business. https://www.facebook.com/business/news/sharing-actions-on-stopping-hate Accessed 22 July 2025

5 Information provided by the IT companies about measures taken to counter hate speech, 2022. (n.d.). Retrieved 6 June 2025, from https://commission.europa.eu/system/files/2022-12/Information%20provided%20by%20the%20IT%20companies%20about%20measures%20taken%20to%20counter%20hate%20speech%20%E2%80%93%202022.pdf

6 Cortiz, D., & Zubiaga, A. (2020). Ethical and technical challenges of AI in tackling hate speech. The International Review of Information Ethics, 29. https://doi.org/10.29173/irie416

7 Mehta, H., & Passi, K. (2022). Social Media Hate Speech Detection Using Explainable Artificial Intelligence (XAI). Algorithms, 15(8), 291. https://doi.org/10.3390/a15080291

8 Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. PeerJ Computer Science, 7, e598. https://doi.org/10.7717/peerj-cs.598

9 Barragán-Montero, A., Javaid, U., Valdés, G., Nguyen, D., Desbordes, P., Macq, B., Willems, S., Vandewinckele, L., Holmström, M., Löfman, F., Michiels, S., Souris, K., Sterpin, E., & Lee, J. A. (2021). Artificial intelligence and machine learning for medical imaging: A technology review. Physica Medica, 83, 242–256. https://doi.org/10.1016/j.ejmp.2021.04.016

10 Dietrich, F. (2024). AI-based removal of hate speech from digital social networks: Chances and risks for freedom of expression. AI and Ethics. https://doi.org/10.1007/s43681-024-00610-7

11 Shepherd, J. (2025, June 5). 21 Essential Twitter (X) Statistics You Need to Know in 2025. The Social Shepherd. https://thesocialshepherd.com/blog/twitter-statistics

12 Abusaqer, M., Saquer, J., & Shatnawi, H. (2025). Efficient Hate Speech Detection: Evaluating 38 Models from Traditional Methods to Transformers. Proceedings of the 2025 ACM Southeast Conference, 203–214. https://doi.org/10.1145/3696673.3723061

13 Teo, D. M. (2024, November 10). Protecting Content Moderators' Wellbeing. Zevo Health. https://www.zevohealth.com/blog/moderating-harm-maintaining-health-protecting-the-wellbeing-of-content-moderators/ Accessed 14 June 2025

14 Shen, T., Jin, R., Huang, Y., Liu, C., Dong, W., Guo, Z., Wu, X., Liu, Y., & Xiong, D. (2023). Large Language Model Alignment: A Survey (No. arXiv:2309.15025). arXiv. https://doi.org/10.48550/arXiv.2309.15025

15 Potamias, R. A., Siolas, G., & Stafylopatis, A.-G. (2020). A transformer-based approach to irony and sarcasm detection. Neural Computing and Applications, 32(23), 17309–17320. https://doi.org/10.1007/s00521-020-05102-3

16 Taylor, J., Peignon, M., & Chen, Y.-S. (2017). Surfacing contextual hate speech words within social media (No. arXiv:1711.10093). arXiv. https://doi.org/10.48550/arXiv.1711.10093

17 Coldewey, D. (2019, August 14). Racial bias observed in hate speech detection algorithm from Google. TechCrunch. https://techcrunch.com/2019/08/14/racial-bias-observed-in-hate-speech-detection-algorithm-from-google/

18 Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, M., Gallivanone, F., Cozzi, A., D'Amico, N. C., & Sardanelli, F. (2021). AI applications to medical images: From machine learning to deep learning. Physica Medica, 83, 9–24. https://doi.org/10.1016/j.ejmp.2021.02.006

19 Art. 1 DSGVO – Gegenstand und Ziele. (2016). Datenschutz-Grundverordnung (DSGVO). https://dsgvo-gesetz.de/art-1-dsgvo/

20 European Union. (2025, May 27). How the Digital Services Act enhances transparency online | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/policies/dsa-brings-transparency

21 Rawat, A., Kumar, S., & Samant, S. S. (2024). Hate speech detection in social media: Techniques, recent trends, and future challenges. WIREs Computational Statistics, 16(2), e1648. https://doi.org/10.1002/wics.1648

22 Anjum, & Katarya, R. (2024). Hate speech, toxicity detection in online social media: A recent survey of state of the art and opportunities. International Journal of Information Security, 23(1), 577–608. https://doi.org/10.1007/s10207-023-00755-2

23 Calabrese, A., Neves, L., Shah, N., Bos, M., Ross, B., Lapata, M., & Barbieri, F. (2024). Explainability and Hate Speech: Structured Explanations Make Social Media Moderators Faster. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp.398–408). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-short.38

24 Coalition for Content Provenance and Authenticity. (2025). Overview—C2PA. https://c2pa.org/

25 Sharma, A., Maier, F., & Breeden, J. (2025, April 17). Open Source Intelligence: Die besten OSINT Tools. Computerwoche. https://www.computerwoche.de/article/2795282/wie-viel-wissen-hacker-ueber-sie.html Accessed 19 June 2025

26 Carvallo, A., Quiroga, T., Aspillaga, C., & Mendoza, M. (2024). Unveiling Social Media Comments with a Novel Named Entity Recognition System for Identity Groups. arXiv preprint arXiv:2405.13011.

27 Lin, J. (2022). Leveraging world knowledge in implicit hate speech detection. arXiv preprint arXiv:2212.14100.

28 Mnassri, K., Rajapaksha, P., Farahbakhsh, R., & Crespi, N. (2023, May). Hate speech and offensive language detection using an emotion-aware shared encoder. In ICC 2023-IEEE International Conference on Communications (pp.2852–2857). IEEE.

29 Martins, R., Gomes, M., Almeida, J. J., Novais, P., & Henriques, P. (2018, October). Hate speech classification in social media using emotional analysis. In 2018 7th Brazilian Conference on Intelligent Systems (BRACIS) (pp.61–66). IEEE.

30 Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act) (Text with EEA Relevance), 277 OJ L (2022). http://data.europa.eu/eli/reg/2022/2065/oj/eng

31 Meta Oversight Board Bylaws: (2024). https://www.oversightboard.com/wp-content/uploads/2024/03/Oversight-Board-Bylaws.pdf Accessed 22 July 2025

32 Data Workers' Inquiry. (n.d.). Data Workers' Inquiry. Retrieved 18 September 2025, from https://data-workers.org/

33 Qhala. (2025, June 10). Data Workers in AI: A New Frontier of Labour Exploitation in the Global South. Medium. https://qhalahq.medium.com/data-workers-in-ai-a-new-frontier-of-labour-exploitation-in-the-global-south-362e22eae01b

34 Tasin, J. J. and F. (2024, July 22). How the global South may pay the cost of AI development. OMFIF. https://www.omfif.org/2024/07/how-the-global-south-may-pay-the-cost-of-ai-development/

About the ALAIT Project

The Austrian Lab for AI Trust (ALAIT) is a research and development project funded by the Austrian Federal Ministry for Innovation, Mobility and Infrastructure (BMIMI) aimed at building trust through knowledge in the field of artificial intelligence (AI). The ALAIT project aims to empower interested parties and key societal groups to use AI technologies responsibly and to establish ethical and high-quality standards for the use of AI.

The project is led by **winnovation** (Gertraud Leimüller and Lena Müller-Kress) and is being carried out in collaboration with **leiwand.ai** (Rania Wazir and Silvia Wasserbacher-Schwarzer), **TU Wien** (Sabine Köszegi and Ilya Faynleyb) and A**ustria Presse Agentur – APA** (Verena Krawarik and Sophia Marecek).

The ALAIT dossiers are available on the project website: https://science.apa.at/project/alait/

The contents of this dossier reflect the current state of the art and have been carefully compiled in accordance with scientific criteria. However, they do not constitute legally binding information or advice.

Imprint

Media owner and publisher:
winnovation consulting gmbh
Linke Wienzeile 124/5
1060 Vienna

Published in 2026, funded by the Austrian Federal Ministry for Innovation, Mobility and Infratructure.

Bundesministerium Innovation, Mobilität und Infrastruktur

winnovation

leiwand.ai

TU WIEN

APA

Acknowledgements:

We would like to thank the following experts for their support with earlier versions of this dossier: Florian Schmidt (APA), Ingrid Brodnig, Mira Reisinger (leiwand.ai)