TURBOCHARGED: AUTOMATING QUALITY ANALYSIS IN TRUST & SAFETY

Alexandra Bianca Tîrnăcop [1*]

[1]*Bucharest University of Economic Studies, Bucharest, Romania, tirnacopalexandra21@stud.ase.ro*

Abstract

Trust and Safety (T&S) is a key framework for online platforms, aiming to protect users from harm such as misinformation, harassment, and exploitation, while also supporting free expression. Although policies, AI tools, and cross-platform collaboration (e.g., GIFCT, StopNCII.org) enhance moderation, significant challenges remain. This study uses a demo dataset of 15 social media posts, reviewed by 9 moderators and checked by a single analyst. Each ticket has been reviewed by three raters to ensure agreement. The model achieved a precision, recall, and F1 score of 70.37%, with an overall accuracy of 64.44%. Automation improves efficiency but requires bias moderation, transparency, and human intervention to address challenging content. However, outsourcing and underinvestment in moderators raise ethical concerns, as human reviewers face psychological risks without adequate support. To address these issues, this paper proposes a decision matrix for use in both machine learning training and moderator and quality analyst training.

Keywords: artificial intelligence; key performance indicators; machine learning

JEL Classification: M11, O22

DOI: 10.24818/CTS/7/2025/2.11

1. Introduction

The purpose of this paper is to provide a clearer understanding of the Trust and Safety domain, followed by a comprehensive literature review and statistical analysis based on the author's research and experience. In the first part, the paper explains the Trust and Safety field. The literature review examines human involvement in this area and how artificial intelligence can enhance raters' abilities in content moderation, as well as the potential risks associated with each actor. In the second part, a methodological analysis is conducted using metrics currently applied in Machine Learning to implemented Artificial Intelligence models. The results are followed by an innovative process structure that supports the paper's main argument: automating quality analysis.

2. Literature Review

2.1 Trust & Safety fundamentals

2.1.1 Definition

Trust & Safety (T&S) refers to the set of policies, guidelines, and operational measures that digital platforms use to manage risks related to online content and user interactions. The aim is to protect users from potential harm, such as misinformation, harassment, fraud, and various forms of exploitation, while also preserving key rights like free expression, access to information, and active digital engagement (DTSP, 2024a). A robust Trust & Safety approach includes proactive design choices, risk evaluation, content moderation, detection and response strategies, quality control, and transparency initiatives. As online threats are constantly evolving, Trust & Safety practices must continually adapt to address emerging challenges (DTSP, 2024a). It is important to note that content moderation is the process of regulating user-posted content on a platform to ensure compliance with pre-established rules (Zeng and Kaye, 2022). Organisations such as the Digital Trust & Safety Partnership (DTSP) help shape industry standards, promote responsible practices, and advocate for greater transparency in the field (DTSP, 2024a). Effective Trust & Safety practices require a proactive approach to managing risks related to online content and user behaviour. Companies can enhance their ability to mitigate potential

* Corresponding author
Authors' ORCID:
Alexandra Bianca Tîrnăcop

harm by developing analytical tools to detect patterns of misuse and by integrating preventive safeguards into digital products. Involving Trust & Safety teams early in the product development process ensures that potential risks are identified and addressed before new features are launched (DTSP, 2024b).

2.1.2 Principles

Content moderation on digital platforms is shaped by cultural, legal, and institutional factors that influence their guidelines. External cultural frameworks include broad concepts such as free speech, exemplified by Enlightenment principles or the U.S. First Amendment, as well as more specific cultural values, such as the French libertine attitude compared to the Northern European Protestant approach to decency. Platform terms and conditions are formed through ongoing contests between different interest groups, each advocating for their values to be recognised. These groups often apply informal pressure and pursue legal challenges to influence platform policies, while platforms themselves also seek to shape and adapt to the broader cultural norms around them (Walker, 2025).

The Santa Clara Principles (SCP) on Transparency and Accountability Around Content Moderation were introduced in 2018 to guide companies in approaching content moderation in a transparent, fair, and rights-respecting manner (BHRRC, 2021). These principles require platforms to be clear about their moderation processes, hold themselves accountable for their decisions, and provide users with a process to appeal moderation actions. The principles also stress that moderation should be proportionate and free from bias. In December 2021, the second edition of these principles was released, expanding their scope to include both Foundational and Operational principles (BHRRC, 2021). This update responds to the increased use of automated moderation tools and incorporates global feedback, advocating greater transparency regarding the content and actions being moderated. The aim is to ensure that content moderation remains just, accountable, and aligned with users' rights as digital environments evolve (BHRRC, 2021). YouTube, GitHub, LinkedIn, Twitter, Apple, Instagram, Facebook, Medium, Reddit, Snap, Tumblr, and Google committed to adhere to the SCP. While some have made progress in increasing visibility into their moderation processes, very few have fully met the standards set by the principles. There remains a significant gap between commitments and actual implementation (SCP, 2021a).

Companies should regularly publish data on the number of posts removed and accounts suspended for violations of their content guidelines. Reports should include detailed breakdowns, such as the total number of flagged posts and accounts, categories of rule violations, content formats (e.g. text, video, images), and sources of flags (e.g. users, governments, automated tools) (SCP, 2021b). Additionally, companies must notify users when their content is removed or their account is suspended, explaining which guidelines were violated, how the content was detected, and the process for appealing the decision. The appeal process should involve a human review, allowing the user to present additional information and providing a clear rationale for the final decision. Over time, independent external reviews may help users seek redress when necessary (SCP, 2021b). Establishing a dedicated team to develop, maintain, and update content and conduct policies ensures consistency and adaptability. Incorporating user feedback into policy drafting enhances relevance, while clear, accessible policy descriptions improve understanding and compliance. Mechanisms for integrating user research and community input strengthen rulemaking, and collaboration with third-party civil society groups and experts helps align policies with industry best practices. Embracing transparency leads to a safer digital environment (DTSP, 2024b).

2.1.3 Focus – Most important areas

a) The fight against terrorism

To identify content shared online that supports or represents dangerous organisations or individuals, the Global Internet Forum to Counter Terrorism (GIFCT) implemented identification through hashes (GIFCT, 2024). A perceptual hash is a unique numerical identifier for digital content such as images, videos, or PDFs, designed to detect similar material without storing the original file. In the GIFCT database, these hashes are catalogued with details about the content type, origin, and context. GIFCT members use them to identify and assess matching content on their platforms, enabling the sharing of

insights on potential terrorist material without exchanging user data. Members can also review and provide feedback on hashes based on GIFCT's taxonomy, fostering collaboration to manage harmful content online (GIFCT, 2024). GIFCT's taxonomy relies on general and behavioural inclusion parameters, such as content linked to entities on the UN Security Council Sanctions List or incidents triggering the Incident Response Framework (IRF). Behavioural criteria focus on material from non-governmental sources, featuring extremist symbols, promoting hate-based ideologies, or encouraging violence. Additional considerations include attacker manifestos, publications with terrorist branding, and URLs flagged by Tech Against Terrorism (TAT). GIFCT's identification through hashes enables platforms to recognise and manage harmful content while maintaining security and adherence to platform policies (GIFCT, 2022).

b) The fight against child sexual exploitation material

INTERPOL employs a global strategy to combat crimes against children, including exploitation, trafficking, forced labour, and abduction. To help locate missing children, it issues Yellow Notices, while experts in human trafficking collaborate with member countries to recover victims (INTERPOL, 2024a). The Internet has intensified the risks, enabling offenders to share illegal content, exploit social platforms, and engage in real-time abuse. The mission of INTERPOL's Crimes Against Children unit is to identify and rescue victims, restrict access to harmful content, and prevent offenders from travelling to exploit children or evade law enforcement, reinforcing the international commitment to safeguarding vulnerable minors (INTERPOL, 2024a). INTERPOL uses advanced software to compare images and videos, allowing investigators to quickly connect victims, abusers, and locations, and to determine whether an image has been previously identified in other countries or shares characteristics with others. This system enables more than 70 countries to collaborate and share information. INTERPOL's Child Sexual Exploitation database, containing over 4.9 million images and videos, has identified more than 42,300 victims (INTERPOL, 2024b). A 2018 joint report with ECPAT International highlighted disturbing trends, such as the extreme abuse of very young children, with 84% of images showing explicit sexual activity. The findings revealed that many victims were prepubescent, including infants and toddlers, with girls being the majority. Most offenders were male, and severe abuse was more common in cases involving boys (INTERPOL, 2024b).

c) The fight against non-consensual intimate imagery (NCII)

Stopping the spread of NCII is a collective effort involving technology companies, advocacy groups, and law enforcement to prevent the distribution of private images or videos shared without consent. As NCII can spread rapidly online, platforms employ a combination of human review, user reporting, and advanced technology such as hash matching to detect and remove such content (StopNCII.org, 2025). Initiatives like StopNCII.org provide tools that enable victims to create a digital fingerprint (hash) of their content, enabling platforms to block and remove matching images without requiring victims to reshare sensitive material. While these measures enhance detection and enforcement, challenges remain, particularly for smaller platforms that may lack the resources for extensive moderation. Education, stronger policies, and collaboration between platforms remain essential in protecting individuals from NCII abuse (DTSP, 2024b).

2.1.4 Importance

Departments responsible for maintaining safe and respectful online spaces play a key role in shaping digital interactions. They develop and implement guidelines that balance open discussion with the need to prevent harmful material. Managing large volumes of user-generated content presents challenges, including interpreting context and addressing emerging risks associated with evolving technologies. These teams establish frameworks that influence how information is shared and regulated. Trust and Safety teams are essential in structuring online conversations and maintaining a well-regulated digital environment (Shulruff, 2024). Trust and Safety departments are often considered cost centres, leading to a lack of investment in their expertise. Content moderation requires continuous recruitment, training, and oversight to manage the increasing volume of content and shifting policies (Eissfeldt and Mukherjee, 2023). To reduce expenses, many platforms turn to outsourcing, resulting in precarious working conditions for moderators who often do not receive the same rights and benefits as permanent staff. Although outsourcing may lower costs, it raises concerns about the well-being of those exposed

to harmful content. Despite these challenges, platforms recognise the importance of investing in trust and safety measures to prevent reputational damage, legal risks, and to ensure a positive experience for users (IHRB, 2025). The rise of AI-generated content adds further complexity, as AI can automate some moderation tasks but also creates new challenges, such as detecting harmful AI-driven content. Ultimately, economic pressures on Trust and Safety departments can affect both the quality of moderation and the health of those responsible for it (multiple aspects discussed in Weigl and Bodo, 2025).

Once terms and conditions are established, they are translated into concrete policy guidelines and training materials for content moderators. These moderators, who are often outsourced or work remotely, apply these rules to assess content, although the norms they are trained on may be culturally distant from their own. These human moderators form part of a global network of content moderation services, frequently operating under exploitative conditions (Siapera, 2021). AI tools assist by filtering and flagging content before it reaches human moderators, using probabilistic models that attempt to interpret cultural norms without fully accounting for local context. Human moderators are essential in refining these AI systems, with the aim of eventually automating the moderation process. While AI aims to enhance efficiency and reduce costs, it still requires human oversight due to the inherent uncertainty in AI decision-making. At the most rigid level of this system, content is filtered through automatic blacklists, creating clear rules for exclusion, although these methods often miss the nuance of individual cases. Ultimately, this multilayered process reduces complex cultural values to binary rules, often at the expense of context and understanding (Horatio Insights, 2025).

2.2 Automation in Trust and Safety (T&S)

AI is transforming content moderation on online platforms, shaping public discourse and influencing societal power dynamics. Social media companies use AI-driven moderation tools to ensure compliance with content policies and speech regulations. Although AI reduces moderation costs, the rise of AI-generated content introduces new challenges, as unreliable or harmful material can spread rapidly. These moderation systems, which integrate algorithms, human oversight and policy frameworks, ultimately affect how information is controlled and disseminated (Ahmed and Khan, 2024). AI-driven automation is essential for maintaining safe digital environments, enabling platforms to identify and address harmful content efficiently. By combining machine intelligence with human analysis, these technologies enhance content moderation and ensure compliance with established guidelines. They are particularly effective at recognising content that aligns with predefined legal standards or matches known harmful material (DTSP, 2024b). Maintaining accuracy is crucial; therefore, ongoing monitoring to prevent errors is imperative. Advanced AI tools also expand moderation capabilities, though careful evaluation is necessary. When applied responsibly, AI supports digital governance, strengthens enforcement and enhances transparency, ultimately contributing to safer online interactions while assisting moderation teams (DTSP, 2024b).

It is important to consider that Artificial Intelligence has been implemented in both areas, with real users and AI models generating content on the platforms. Content moderation is now a complex challenge due to the vast volume of online communication, the difficulty of interpreting context, and the increasing presence of both user-generated and AI-generated harmful material (Reelmind, 2025). The Internet has expanded access to speech, enabling people worldwide to share their perspectives, but this freedom also brings the responsibility of regulation. Digital platforms must navigate a complex landscape of legal requirements, cultural norms, and societal expectations while managing the rapid spread and permanence of content (Ricknel, 2020). Online speech is not merely random conversation but part of broader discussions shaped by various influences. Moderation decisions require careful consideration to avoid both excessive restrictions, which may be seen as suppression, and insufficient oversight, which can allow harmful material to persist. As platforms work to refine their approaches, a more adaptable and context-aware system is essential to ensure a balance between maintaining open discourse and protecting users from harmful content (Habibi, Hoby and Schwartz, 2025).

To implement a trustworthy Artificial Intelligence (AI) model in current organisational practices, it is crucial to address bias, ensure fairness, and maintain responsible governance. Tackling bias begins with carefully examining models for signs of discrimination, evaluating fairness in decision-making processes, and retraining models to correct these issues (Woods, 2022). Data balancing methods are essential for ensuring all groups are represented, such as generating synthetic data, using models to include underrepresented populations, and maintaining a balanced demographic mix in datasets. Companies should also focus on AI governance by equipping data scientists with the knowledge to identify and mitigate bias, taking proactive steps to eliminate biases, and continually improving strategies for building trust. These approaches help make AI systems fairer, more accountable, and more dependable (Woods, 2022).

Generative Artificial Intelligence (GenAI) has the potential to improve quality control in content moderation by providing explanations for enforcement actions, helping to track patterns and identify inconsistencies (DTSP, 2024b). Although promising, this approach is still developing, and AI-generated explanations may sometimes lack accuracy or completeness, necessitating ongoing refinement. GenAI can also support the evaluation of human decision-making by cross-referencing flagged content between human reviewers and AI, prompting further assessment when discrepancies occur. In addition, it can streamline data labelling by converting reviewer insights into structured formats, improving the detection of ambiguous content. To ensure reliability, robust oversight and validation processes must support AI-driven moderation (DTSP, 2024b).

Content moderation on digital platforms increasingly relies on a combination of automated systems and human oversight. Initially, AI-based filters scan uploaded material, identifying and categorising content based on predetermined criteria (Oversight Board, 2025). These systems evolve over time, refining their accuracy by incorporating insights from human moderators. If content is deemed safe, it proceeds without interruption; if flagged, it undergoes further evaluation, with complex cases assigned for manual review. Platforms also use user reports as a secondary moderation method, ensuring a balanced approach. AI tools assist in identifying trends and patterns, handling repetitive tasks, and streamlining moderation. Companies continue to refine these technologies to improve efficiency, reduce operational costs, and alleviate the psychological strain on human reviewers. Recent industry trends indicate a shift towards greater automation, with some companies restructuring their moderation teams to integrate these advancements. Additionally, backend systems are being optimised to manage regulatory requirements more effectively, reducing reliance on manual processing (Vargas Penagos, 2025).

2.3 Quality Analysis Metrics

Content moderation operates through a structured system designed to uphold platform policies. Quality assessment teams monitor moderators' accuracy using real-time evaluation methods such as sampling and pattern matching. Performance is also measured by efficiency metrics like average handling time (TSPA, 2025). Moderation is divided into tiers: the first level addresses clear-cut cases, while the second handles more nuanced or complex content. Some organisations add a third level to address novel or undefined content, with these moderators occasionally contributing to policy refinement. Unlike the first two tiers, which are often managed by external contractors, this advanced tier typically consists of in-house employees due to their specialised expertise and decision-making role (Tremau, 2025).

Key performance metrics are essential in AI-driven content moderation. The commonly used indicators are False Positive Rate (FPR), False Negative Rate (FNR), precision, recall, and the F1 score. These metrics ensure accuracy and effectiveness. FPR measures cases where content is mistakenly flagged as a violation, while FNR tracks actual violations that go undetected. Precision evaluates how accurately the system identifies harmful content, whereas recall determines how well it detects all instances of violations (DTSP, 2024b). A model optimised for precision reduces incorrect flagging but may overlook some violations, while a recall-focused approach ensures more violations are caught but may increase errors. The F1 score balances these factors, helping moderation teams refine their strategies. By

carefully adjusting these metrics based on the risks of excessive flagging or missed violations, AI systems can improve content moderation while minimising unintended consequences (DTSP, 2024b).

True positives (TP) and true negatives (TN) are key components in evaluating the performance of a classification model. A true positive occurs when the model correctly predicts a positive case, while a true negative occurs when it correctly predicts a negative case. These values, together with false positives (FP) and false negatives (FN), are used to calculate important performance metrics. Accuracy, which measures overall correctness, is calculated as:

$$\text{Accuracy: } \frac{TP+TN}{TP+TN+FP+FN} \text{ (Google for Developers, 2025) (1)}$$

However, accuracy can be misleading if one class is much more common than the other. That is why precision and recall provide deeper insights.

Precision indicates the proportion of correct positive predictions among all predicted positives:

$$\text{Precision: } \frac{TP}{TP+FP} \text{ (Google for Developers, 2025) (2)}$$

Recall measures how many actual positives the model correctly identified. Together, these metrics help assess a model's effectiveness beyond overall accuracy (Juba and Le, 2019, pp. 4041–4043).

$$\text{Recall: True Positive Rate (TPR): } TPR = \frac{TP}{TP+FN} \text{ (Google for Developers, 2025) (3)}$$

Accuracy can be misleading if one class is much more common than the other. This is why precision and recall offer deeper insights, as they are metrics used to evaluate the effectiveness of classification models, particularly when dealing with imbalanced datasets where one class appears more frequently than another. Precision measures the proportion of correct positive predictions out of all predicted positives, while recall assesses the model's ability to identify all actual positive cases.

Precision ensures that positive predictions are accurate, while recall ensures that the model identifies as many true positives as possible. Achieving high precision often results in lower recall, and vice versa, making it necessary to strike a balance based on the specific application. If precision exceeds 50%, maintaining constraints on both precision and recall can help stabilise overall accuracy. To improve both metrics in imbalanced datasets, ensuring diverse and representative training data is critical. By addressing class imbalances, models can enhance their predictive reliability and reduce bias in classification tasks (Juba and Le, 2019). On a given dataset, recall shows how many violations a moderator caught, while precision shows how many of the posts taken down were actually violating. These are referred to as overenforcement and underenforcement.

Consistency is a key indicator in content moderation. It demonstrates a system's ability to apply platform guidelines effectively, ensuring harmful content is flagged accurately while avoiding excessive restrictions across different contexts, languages, and evolving user behaviours (Juba and Le, 2019). Its greatest challenge is class imbalance: harmful content, such as hate speech, is much less frequent than normal posts. This causes machine learning models to be biased towards the more common category, which may result in missed harmful content or overzealous censorship (Juba and Le, 2019). Approaches such as resampling or ensemble techniques can address this imbalance, but the most effective way to maintain high accuracy appears to be scaling quality training data, complemented by human oversight, rather than relying solely on algorithmic adjustments. Consistent moderation requires both robust machine learning models and ongoing alignment with platform policies, closely supervised by human analysts, as even minor mistakes can undermine trust and user safety (Juba and Le, 2019).

3. Methodology

Precision and Recall are typically calculated by class (Shweta, Bajpai and Chaturvedi, 2015, p. 22). For the purposes of this paper, the selected class is hate speech. To respect data privacy, this study uses an

artificial demo dataset created by the author, designed to illustrate the measurement system currently employed in content moderation for digital platforms. The dataset comprises 15 tickets (social media posts) evaluated by nine moderators and one analyst. Each ticket is reviewed by three human moderators to ensure reliability and to enable the assessment of inter-rater agreement during the production phase (i.e., while the content is live on the platform). Once consensus among moderators is reached, the ticket is automatically removed or retained on the platform. The analyst acts as the quality assurance (QA) authority, providing the reference or "ground truth" classification during the post-production phase, after the content has been either removed or retained. The task involves binary classification of posts as either “hate speech” or “not hate speech”. To evaluate the performance of the moderation process, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are calculated at the individual level. Breakdown: TP occurs when both the consensus and QA decision label a post as hate speech; TN when both label it as not hate speech; FP when the consensus classifies a post as hate speech but the QA does not; and FN when the consensus decides not hate speech but the QA identifies hate speech (Cyberhaven, 2015). The formulas used are provided in Appendix A.

The detection quality was assessed using precision, recall, F1 score, and accuracy. Precision measured the avoidance of false positives, recall measured the avoidance of false negatives, and the F1 score provided a balanced measure of both. Accuracy, while informative, was interpreted with caution as it can be misleading in cases of class imbalance. At the moderator level, systematic biases were identified by summing true positive and false positive rates per moderator, revealing tendencies to over- or under-flagging. A similar approach to hate speech identification systems is used in the ETHOS dataset (Mollas et al., 2021).

Table 1 visually represents the synthetic dataset used. Individual tickets and moderators are identified by unique numerical values. Ticket ID refers to the numerical value assigned to a ticket (social media post). Rater ID is the numerical value assigned to a moderator. Moderator Decision is the decision each moderator made on the tickets during the production phase. QA Decision represents the correct decision that all moderators should have made for a given ticket, also referred to as ‘ground truth’. The table shows two tickets (TKT-10045782 and TKT-10045813) and six raters (moderators) (MOD1023, MOD4567, MOD2891, MOD5678, MOD2345, MOD8901). Each ticket was assigned to three moderators, and each moderator selected an individual label (hate speech/not hate speech). QA is a unique individual who provides the ground truth.

Table 1. Data Structure Sample

Ticket ID	Rater ID	Moderator_Decision	QA_Decision
TKT-10045782	MOD1023	Hate Speech	Hate Speech
TKT-10045782	MOD4567	Hate Speech	Hate Speech
TKT-10045782	MOD2891	Hate Speech	Hate Speech
TKT-10045813	MOD5678	Not Hate Speech	Not Hate Speech
TKT-10045813	MOD2345	Hate Speech	Not Hate Speech
TKT-10045813	MOD8901	Not Hate Speech	Not Hate Speech

Source: Author (2025)

The synthetic dataset and the experiment mirror the data and process structure of a global content moderation site for digital platforms. The content moderation site chose to remain anonymous. It is important to note that the dataset is a simplified version of real-world datasets. In real-world scenarios, the number of times a ticket may circulate, the variety of decisions, and the number of quality analysts may vary. During the audit phase, raters are assigned numerical values to protect the auditing process from human bias on the QA’s end. Tickets are opened and verified by the expert (QA) using their IDs. Once the audit process is complete, the identity of the raters is revealed to the quality analyst to provide feedback and identify the root cause of the error.

Based on the dataset, the true positive, true negative, false positive, and false negative values were identified according to the ground truth labels established by the quality analyst. The frequency of each variable was then computed. Using these counts, the evaluation metrics – precision, recall, accuracy,

and F1 score – were calculated to assess the moderators' performance. All computations were performed in Microsoft Excel, and the corresponding formulas are provided in Appendix A.

4. Results

Following the computations, the results show a precision of 70.37%, recall of 70.37%, accuracy of 64.44%, and an F1 score of 70.37%. A precision of 70.37% indicates that approximately seven out of ten items flagged as policy-violating content were correctly identified, suggesting a moderate false-positive rate. The recall score of 70.37% demonstrates that the model successfully detected 70% of all actual violations, implying that about 30% of harmful or inappropriate content remains undetected. The identical F1 score reflects a balance between precision and recall, indicating consistent performance across both detection and false-alarm dimensions.

An accuracy of 64.44% is relatively lower than the precision and recall metrics, which may indicate class imbalance – a common feature of content moderation datasets where non-violating content significantly outnumbers violating instances. Therefore, accuracy alone may not provide a complete picture of model performance in this context. Nonetheless, the model demonstrates moderate and balanced detection capability, making it suitable for initial deployment or use as an assistive moderation tool. Further improvements could be achieved through enhanced data representation of under-detected content types, threshold optimisation, or model fine-tuning to reduce both false positives and false negatives.

5. Suggestions

Building an effective auditing system begins with understanding the types of content reviewed and the associated risks. Developing a content taxonomy that categorises violations by legal sensitivity, potential harm, and moderation complexity can have a significant positive impact. Each category is scored and mapped into a prioritisation matrix, which guides how often certain content types should be audited and assesses their automation potential. For example, high-risk and high-volume content such as hate speech or misinformation is sampled daily, while low-risk areas like off-topic comments are reviewed less frequently or flagged only when model confidence is low. Consistency is supported by adaptive QA checklists tailored to each violation type, updated as policies evolve. Regular calibration sessions help ensure alignment across analysts, particularly when interpreting nuanced or region-specific cases.

Operationally, tracking core metrics such as accuracy rates, time spent per review, escalation volumes, and discrepancies between model outputs and human decisions is essential. These data points help identify where guidelines require clarification or where retraining is needed. Automation is crucial for scaling: risk-based sampling, workflow routing, and content pre-triage are all managed algorithmically. However, human analysts must retain control over final decisions, especially for edge cases. Based on the author's experience, this hybrid approach significantly improves quality without compromising throughput. It enables the team to remain focused on what matters: protecting users, reducing policy risk, and maintaining audit integrity at scale. The author reiterates the importance of the audit process: while content moderators are responsible for removing content from platforms, quality analysts audit whether moderators make correct decisions on that content. Through audits, quality analysts also provide feedback and determine which moderators require additional training or guidance.

Table 2 visually represents a suggested prioritisation matrix which, according to the author's findings, could simplify the audit process for quality analysts while increasing efficiency. This is particularly relevant given the growing need for automation as the volume of content requiring audit continues to rise. The table proposes a 'narrow down' approach, beginning with the identification of the content type (Content Category), such as Hate Speech. The main categories are then subdivided into smaller categories (for example, Hate Speech is divided into Racial Slurs and Homophobic Attacks). After narrowing down the content, the next step is to determine the Risk Tier, which assesses how prohibited a type of content is based on current legislation (e.g. the distribution or facilitation of CSAM is illegal) and business interests (e.g. the distribution of pornographic material on certain platforms may harm brand image). This Risk Tier depends on the Risk Drivers. Audit Priority is used to determine how often

a particular type of content should be audited (e.g. daily, weekly) based on its risk tier. Automation Potential is determined by the Content Category (as some policies are more complex than others and not all can be automated), Risk Tier, and Risk Drivers. For example, Nudity/Pornography can be easily identified, has a medium risk tier, and the issue is brand-related rather than legal, so it has high automation potential. In contrast, CSAM distribution has a high-risk tier with legal risk drivers, so its automation potential is low; the author recommends that high-risk tiers involving legal concerns require human input.

Table 2. Suggested Prioritization Matrix

Content Category	Subcategory / Examples	Risk Tier	Risk Drivers	Audit Priority	Automation Potential
Hate Speech	Racial slurs, homophobic attacks	High	Legal (EU DSA), platform bans, user safety	Daily	Medium
Child Exploitation	Grooming, inappropriate comments, age misrepresentation	High	Legal (CSAM), immediate takedown risk	Hourly	Low
Nudity / Pornography	Explicit videos, sexting	Medium	Brand alignment	Weekly	High
Minor Policy Violations	Off-topic posts, profanity, trolling	Low	Low regulatory concern	Monthly	High

Source: Author (2025)

6. Conclusions

The evaluation results indicate that the content moderation model demonstrates moderate and balanced effectiveness, achieving precision, recall, and F1 scores of 70.37%, with an overall accuracy of 64.44%. The system is competent at correctly identifying most policy-violating content while maintaining false positives at a manageable level. Nonetheless, the relatively lower accuracy suggests some misclassification of non-violating posts, likely due to class imbalance or context-sensitive content. To improve performance, it is recommended to enrich the dataset with a greater diversity of low-occurrence violation types, adjust decision thresholds to better balance false positives and false negatives, and integrate contextual analysis. Nevertheless, the findings establish a solid baseline for an automated moderation framework, though further refinement of precision and comprehensiveness is needed.

While automation improves efficiency, it also necessitates bias avoidance, transparency, and human judgement to manage contextual content. The example of StopNCII.org and GIFCT demonstrates how hash-matching and intersite cooperation can combat harmful content, but this is generally not feasible for smaller sites. Positive change can be promoted by optimising moderator training to address definitional ambiguities, re-engineering AI models to minimise false positivesand negatives in class-imbalanced datasets, increasing transparency through periodic audits, and ensuring compliance with guidelines such as the Santa Clara Principles.

Designing a robust quality auditing process in trust and safety involves making practical decisions aligned with real-world risks. By clearly defining content types, understanding the potential harm they may cause, and organising review efforts according to risk, resources can be focused where they are most needed. The prioritisation matrix guides this process, while automation supports scalability without replacing the need for human judgement. This approach helps maintain responsiveness, accuracy, and alignment among teams despite rapid changes in content, provided there is consistent training and regular policy updates.

Limitations

Despite the accuracy and relevance of this paper, the study has been limited primarily by data privacy concerns, which prevented the author from exploring the topic in greater depth and providing a more comprehensive overview of the current situation through more extensive analytics. The very limited sample size is also a minor impediment. The lack of transparency from content moderation sites has been the most significant setback.

Acknowledgement

I would like to express my heartfelt gratitude to Prof. Univ. Dr. Schiopu Andreea Fortuna for her thoughtful guidance and continuous support throughout this research. Her deep expertise and constructive feedback played a crucial role in shaping both the structure and substance of the work. From refining the methodology to clarifying the analysis, her insights were instrumental at every stage.

References

Ahmed, A. and Khan, M.N. (2024). *AI and Content Moderation: Legal and Ethical Approaches to Protecting Free Speech and Privacy* [online] Available at: <https://www.researchgate.net/publication/383661951_AI_and_Content_Moderation_Legal_and_Ethical_Approaches_to_Protecting_Free_Speech_and_Privacy> [Accessed 18 October 2025].

Business and Human Rights Resource Centre (BHRRC) (2021). *Santa Clara Principles present standards for tech platforms to provide transparency and accountability in content moderation*. [online] Available at: <https://www.business-humanrights.org/en/latest-news/the-santa-clara-principles-on-transparency-and-accountability-in-content-moderation/> [Accessed 20 March 2025].

Cyberhaven (2015), *What are False Positives?* [online] Avaliable at: <https://www.cyberhaven.com/infosec-essentials/what-are-false-positives> [Accessed 13th October 2025].

Digital Trust and Satefy Partnership (DTSP) (2024a). *Trust & Safety Best Practices Framework*. [pdf] Digital Trust & Satefy Partnership. Available at: <https://dtspartnership.org/wp-content/uploads/2021/04/DTSP_Best_Practices.pdf> [Accessed 20 March 2025].

Digital Trust and Satefy Partnership (DTSP), (2024b). *Best Practices for AI and Automation in Trust & Safety*. [pdf] Digital Trust & Satefy Partnership. Available at: <https://dtspartnership.org/wp-content/uploads/2024/09/DTSP_Best-Practices-for-AI-Automation-in-Trust-Safety.pdf> [Accessed 21 March 2025]

Eissfeldt, J. and Mukherjee, S. (2023*). Evaluating the Forces Shaping the Trust & Safety Industry* [online] Available at: <https://www.techpolicy.press/evaluating-the-forces-shaping-the-trust-safety-industry/> [Accessed: October 19 2025].

Global Internet Forum to Counter Terrorism (GIFCT) (2024). GIFCT's Hash-Sharing Database. *GIFCT*. [online] Available at: <https://gifct.org/hsdb/> [Accessed 21 March 2025].

Global Internet Forum to Counter Terrorism (GIFCT) (2022). *HSDB Taxonomy - FOR PUBLICATION (Dec 2022)*. [pdf] GIFCT. Available at: <https://gifct.org/wp-content/uploads/2022/12/HSDB-Taxonomy-FOR-PUBLICATION-Dec-2022-1.pdf> [Accessed 21 March 2025].

Google for Developers (2025). Machine Learning Concepts. Classification: *Accuracy, recall, precision, and related metrics*. [online] Available at: <https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall> [Accessed 14 October 2025].

Habibi, M., Hovy, D. and Schwartz, C. (2025). The Content Moderator's Dilemma: Removal of Toxic Content and Distortions to Online Discourse, *Social and Information Networks*. [online] Available at: <https://doi.org/10.48550/arXiv.2412.16114> [Accessed 18 October 2025].

Horatio Insights (2025). *What is Content Moderation? Pros, Cons, and Best Practices*. [online] Available at: <https://www.hirehoratio.com/blog/what-is-content-moderation> [Accessed 18 October 2025]

IHRB (2025). *Content moderation is a new factory floor of exploitation – labour protections must catch up.* [online] Available at: <https://www.ihrb.org/latest/content-moderation-is-a-new-factory-floor-of-exploitation-labour-protections-must-catch-up> [Accessed 18 October 2025]

INTERPOL, (2024a). Crimes against children. *INTERPOL,* [online] Available at: <https://www.interpol.int/en/Crimes/Crimes-against-children> [Accessed 21 March 2025].

INTERPOL, (2024b). Crimes against children. *International Child Sexual Exploitation database*. INTERPOL, [online] Available at: <https://www.interpol.int/en/Crimes/Crimes-against-children/International-Child-Sexual-Exploitation-database> [Accessed 21st March 2025].

Juba, B. and Le, H. S. (2019*).* Precision-Recall versus Accuracy and the Role of Large Data Sets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, pp. 4039–4048. https://doi.org/10.1609/aaai.v33i01.33014039.

Listen Data (2024). *How to Calculate Confusion Matrix in Excel*. [online] Available at: <https://www.listendata.com/2024/06/confusion-matrix-in-excel.html> [Accessed 13 October 2025].

Microsoft (2025). *AND function*. [online] Available at: <https://support.microsoft.com/en-us/office/and-function-5f19b2e8-e1df-4408-897a-ce285a19e9d9> [Accessed 13 October 2025].

Mollas, I., Chrysopoulou, Z., Karlos, S. and Tsoumakas, G. (2021). *Ethos: an online hate speech detection dataset*. [online] Available at: <https://arxiv.org/pdf/2006.08328> [Accessed 19th October 2025].

Oversight Board (2025). *Content Moderation in a New Era for AI and Automation*. [online] Available at: <https://www.oversightboard.com/news/content-moderation-in-a-new-era-for-ai-and-automation/> [Accessed 18 October 2025].

Reelmind (2025). *Ametures Gone Wild: AI Content Moderation Challenges*. [online] Available at: <https://reelmind.ai/blog/ametures-gone-wild-ai-content-moderation-challenges> [Accessed 18 October 2025].

Ricknell, E. (2020). Freedom of Expression and Alternatives for Internet Governance: Prospects and Pitfalls. *Media and Communication,* 8(4), pp. 110-120. https://doi.org/10.17645/mac.v8i4.3299.

Santa Clara Principles (SCP), (2021a). SCP 2.0 Toolkit for Companies. *The Santa Clara Principles on Transparency and Accountability in Content Moderation*. [online] Available at: <https://santaclaraprinciples.org/toolkit-companies/> [Accessed 20 March 2025].

Santa Clara Principles (SCP), (2021b). Santa Clara Principles 2.0 Open Consultation Report. *The Santa Clara Principles on Transparency and Accountability in Content Moderation*. [online] Available at: <https://santaclaraprinciples.org/open-consultation/> [Accessed 21 March 2025].

Shulruff, T. (2024). Trust and Safety work: internal governance of technology risks and harms. *Journal of Integrated Global STEM* 1(2), pp. 95-105. https://doi.org/10.1515/jigs-2024-0003/html [Accessed 19 October 2025].

Shweta, R.C. Bajpai, R.C. and Chaturvedi, H.K. (2015). Evaluation of Inter-Rater Agreement and Inter-Rater Reliability for Observational Data: An Overview of Concepts and Methods, *Journal of the Indian Academy of Applied Psychology*, 41(3), pp. 20-27.

Siapera, E. (2021). AI Content Moderation, Racism and (de)Coloniality, *International Journal of Bullying Prevention*, 4, pp. 55-65, https://doi.org/10.1007/s42380-021-00105-7.

StopNCII.org, (2025). How StopNCII.org Works. *Stop Non-Consensual Intimate Image Abuse*. [online] Available at: <https://stopncii.org/chi-siamo/> [Accessed 22 March 2025].

Tremau (2025). *Content Moderation: Key Practices & Challenges*. [online] Available at: <https://tremau.com/resources/content-moderation-key-practices-challenges/> [Accessed 19 October 2025].

TSPA (2025). *Content Moderation Quality Assurance*. [online] Available at: <https://www.tspa.org/curriculum/ts-fundamentals/content-moderation-and-operations/content-moderation-quality-assurance/>, [Accessed 19 October 2025].

Vargas Penagos, E. (2025). Platforms on the hook? EU and human rights requirements for human involvement in content moderation, *Cambridge Forum on AI: Law and Governance*, 1, e23. https://doi.org/10.1017/cfl.2025.3.

Walker, A.R. (2025). *Legal Defense Fund exits Meta civil rights advisory group over DEI changes*. [online] Available at: <https://www.theguardian.com/technology/2025/apr/11/meta-ldf-dei-policy> [Accessed 19 October 2025].

Weigl, L. and Bodo, B. (2025). Trust and safety in the age of AI - the economics and practice of the platform-based discourse apparatus. *Amsterdam Law School Legal Studies & Institute for Information Law*. 2025-1. http://dx.doi.org/10.2139/ssrn.5116478.

Woods, J. (2022). Bias in AI Program: Showing Businesses How to Reduce Bias and Mitigate Risk. *Vector Insitute*, [online] Available at: <https://vectorinstitute.ai/bias-in-ai-program-showing-businesses-how-to-reduce-bias-and-mitigate-risk/> [Accessed 20 March 2025].

Zeng, J., & Kaye, D. B. V. (2022). From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet*, 14, pp. 79–95. https://doi.org/10.1002/poi3.287.

Appendix A

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$ (Google for Developers, 2025) (1)

Precision: $\frac{TP}{TP+FP}$ (Google for Developers, 2025) (2)

Recall: True Positive Rate (TPR): $TPR = \frac{TP}{TP+FN}$ (Google for Developers, 2025) (3)

F1 Score: $\frac{2\,\mathrm{TP}}{2TP+FP+FN}$ (Google for Developers, 2025) (4)

True Positives TP: =AND(C2="Hate Speech", D2="Hate Speech") (Microsoft, 2025)

False Positives FP: =AND(C2="Hate Speech", D2="Not Hate Speech") (Microsoft, 2025)

False Negatives FN: =AND(C2="Not Hate Speech", D2="Hate Speech") (Microsoft, 2025)

True Negatives TN: =AND(C2="Not Hate Speech", D2="Not Hate Speech") (Microsoft, 2025)

TP COUNT: = COUNTIF (E2:E51,TRUE) – Idem FP, FN, TN (Listen Data, 2024)

Table 3. Glossary

Scenario	Moderator_Decision	Analyst_Decision	Classification
True Positive (TP)	Hate Speech	Hate Speech	Correctly identified as hate speech
False Positive (FP)	Not Hate Speech	Hate Speech	Incorrectly flagged as not hate speech
False Negative (FN)	Hate Speech	Not Hate Speech	Missed hate speech
True Negative (TN)	Not Hate Speech	Not Hate Speech	Correctly ignored