E-ISSN: 2706-8927
P-ISSN: 2706-8919
Impact Factor (RJIF):7.28
www.allstudyjournal.com
IJAAS 2025; 7(8): 181-184
Received: 14-07-2025
Accepted: 17-08-2025

Abdulaziz Salihu Aliero
Lovely Professional University, Kebbi State University of Science and Technology Aliero, Kebbi State, Nigeria

Zaharadeen Adamu Hamisu
Lovely Professional University, Kebbi State University of Science and Technology Aliero, Kebbi State, Nigeria

Corresponding Author:
Abdulaziz Salihu Aliero
Lovely Professional University, Kebbi State University of Science and Technology Aliero, Kebbi State, Nigeria

A survey on personality and reputation preservation on public social media applications

Abdulaziz Salihu Aliero and Zaharadeen Adamu Hamisu

DOI: https://www.doi.org/10.33545/27068919.2025.v7.i8c.1644

Abstract
Public social media platforms facilitate instant messaging and rich media sharing, yet they expose users to risks that can damage personality, integrity, and reputation. Existing solutions predominantly emphasize access control and content moderation, leaving gaps in preventing unintentional disclosure of sensitive information to unintended recipients. We survey contemporary literature on social media privacy, cross-cultural differences, and cryptographic access control, and we propose an enhanced Personality and Reputation Preservation (PRP) approach. Our proposal augments message delivery with an analyzer that learns conversational context and intercepts out-of-distribution ("strange") text or attachments for sender verification prior to transmission. We discuss how modern methods such as fine-grained Attribute-Based Encryption (ABE) and privacy-aware user experience can complement PRP in public networks. A small-scale pilot suggests high user satisfaction with PRP while highlighting latency trade-offs due to background similarity checks. We conclude with future directions for scalable similarity thresholds, human-in-the-loop approvals, and integration with transformer-based content classifiers.

Keywords: Social media privacy, personality and reputation, unintended disclosure, content analyzer, attribute-based encryption, cross-cultural privacy

1. Introduction
The term “social media” is widely used yet difficult to define precisely because the ecosystem spans platforms, practices, and cultures. Classic definitions describe social network sites as web-based services that enable users to construct profiles, articulate connections, and traverse those connections [1]. Another definition by [3], defined social media as platforms where “people can create networks of relationships overlapped with the entire Web, while controlling their own privacy and data. More recent reviews emphasize interactive, user-generated channels where audiences may be broad or narrow and engagement may be synchronous or asynchronous [2]. Despite their benefits, today’s platforms are routinely criticized for pervasive data collection and opaque processing that heighten privacy risks. Users often assume privacy settings are sufficient, but studies show frequent misconfigurations and mismatches between sharing intentions and actual visibility [4], [5]. Privacy harms extend to sensitive domains such as healthcare where confidentiality breaches and blurred professional boundaries are recurrent concerns [6]-[9]. At the same time, harmful content such as hate speech remains a persistent moderation challenge, with modern detection shifting toward Transformer-based models [10]. These trends motivate defenses that do more than lock down data: systems must proactively prevent unintended disclosures that can damage a user’s personality and reputation.

A. Problem Statement and Contributions
Conventional access-control and privacy settings primarily guard against unauthorized readers; they do not stop a sender from accidentally transmitting sensitive content to the wrong recipient. This gap is especially consequential on public social networks where rapid, habitual communication intersects with complex audience boundaries and cross-cultural norms [11]-[13]. To address this, we propose a Personality and Reputation Preservation (PRP) approach that:

- Treats personality and reputation as first-class privacy assets;
- Continuously learns conversational context and shared-content history between parties;
- Interrupts delivery of “strange” messages (low similarity to prior exchanges or novel sensitive content) to seek explicit sender approval before transmission.

The remaining part of the paper is structured as follows: Section II discusses relevant prior research. Section III outlines the proposed methodology. The findings of the research are presented in Section IV, and Section V offers the Discussion and lastly section VI ended with conclusion and suggestions for future work.

2. Literature Review

Prior work on social media privacy spans sentiment and hate-speech analysis for safer discourse, access-control usability, setting failures [3]-[5], and cryptographic enforcement such as Attribute-Based Encryption (ABE) for fine-grained access to shared data. Cross-cultural studies reveal that norms around self-presentation, disclosure, and perceived sensitivity vary markedly, influencing both what users post and how they interpret privacy risks [6]-[9]. Healthcare-focused reviews document benefits of social platforms for education and outreach alongside recurring risks to confidentiality and professional ethics [10], [11]. Recent policy and regulatory analyses underscore that large platforms continue to engage in extensive tracking and algorithmic profiling, intensifying calls for stronger safeguards [12]. The privacy and reputation preservation on social media has been extensively studied, though existing approaches address specific dimensions rather than holistic safeguards. Research in the domain largely falls into four categories: access control and privacy settings, cryptographic approaches, cross-cultural studies, and content moderation [14]-[16].

1. Access Control and Privacy Settings

Early work on online social networks highlighted the insufficiency of default privacy settings, with studies reporting frequent mismatches between user intentions and actual visibility of shared content [4], [5]. Madejski (2011) termed these "privacy setting failures," emphasizing that even experienced users struggle with configuration [4]. More recent work confirms this trend, noting that automated personalization and dark-pattern designs exacerbate misconfigurations [18]. Such inadequacies indicate that user-driven access control alone cannot prevent privacy breaches.

2. Cryptographic Approaches.

Parallel research emphasizes fine-grained cryptographic enforcement. Attribute-Based Encryption (ABE) and Ciphertext-Policy ABE (CP-ABE) allow senders to enforce attribute-based access to shared data [14], [15]. Recent work explores privacy-preserving friend matching and hidden-attribute encryption in mobile social networks, ensuring that even service providers cannot decrypt unintended data [16], [17]. While technically sound, these methods primarily restrict who can read content, not what senders inadvertently disclose? This leaves the problem of accidental disclosure unaddressed.

3. Cross-Cultural Privacy Studies

Social media usage is highly contextual. Cultural norms determine disclosure thresholds, with Western users tending to share more personal and sensitive content than counterparts in East Asia [11], [12]. Studies show that younger generations exhibit higher disclosure willingness, while older cohorts are more privacy-conscious [13]. Cross-cultural frameworks argue that privacy risks must be contextualized, as the very definition of "sensitive information" varies across societies [12]. However, little research integrates cultural awareness into automated privacy-protection mechanisms.

4. Healthcare and Sensitive Domains

The healthcare domain exemplifies the stakes of privacy violations. Social media is widely used by patients for peer support and by clinicians for patient engagement, but confidentiality breaches are common [6], [7]. Narrative reviews confirm that even anonymized data can often be re-identified [8]. The ethical and reputational implications extend beyond patients to physicians, underscoring the need for proactive privacy safeguards [9].

5. Harmful Content and Reputation Risks

Beyond privacy, personality and reputation can be harmed by exposure to hate speech and offensive content. Earlier efforts relied on lexicon-based filters [10], but recent advances employ deep learning, particularly Transformer architectures, for automated detection of toxic speech [10]. These tools, however, target inbound exposure rather than outbound mis-disclosures by users themselves.

Research Gap

In summary, existing literature offers substantial progress in encryption, access control, and moderation, but most approaches address unauthorized access by outsiders. Few, if any, consider the frequent real-world problem of users unintentionally sending sensitive information to the wrong recipient. This "unintended disclosure" directly threatens both privacy and reputation, yet it remains underexplored in scholarly work. Addressing this gap, our Personality and Reputation Preservation (PRP) framework introduces a message-verification layer that pauses delivery of anomalous or contextually sensitive content, integrating both similarity checks and sender approval into the communication pipeline.

3. Proposed Methodology

Communication in social networking involves the sender, message, medium, and recipient. PRP augments the "medium" with a Communication Contents Analyzer (CCA) that evaluates each outgoing item. Using stored conversation histories and metadata, the CCA computes a similarity score between the current content and prior exchanges with the intended recipient(s). If the score falls below a configurable threshold (e.g., 95%) or if sensitive attributes are detected, the system flags the content as "strange," pauses delivery, and requests explicit sender approval.

Complementary Controls: PRP can be paired with cryptographic controls e.g., CP-ABE to ensure that even if content is misrouted, only recipients with matching attributes can decrypt it [15], [16]. Privacy-preserving matching and hidden-attribute schemes for mobile social networks provide additional building blocks for secure sharing in public settings [17], [18].

All four elements contribute to maintaining privacy specifically regarding personality and reputation as well as ensuring the validity of communication. In this piece, we place more emphasis on the Sender and the Medium. Algorithm 1 outlines our proposed method, which is illustrated in Figure 1. To protect the user's identity and reputation, the message is not sent directly to the receiver, as is typical in existing systems. Instead, it must first be approved by the sender after undergoing verification by the system (the medium).

Algorithm 1:

Communication Contents Analyzer (CCA)

- **Input:** Message m, Conversation history H(r) for

recipient r; Similarity threshold τ (e.g., 0.95).
- Compute similarity s = Sim (m, H(r)) using embeddings or n-gram features.
- **If s ≥ τ and no sensitive triggers:** Deliver; else: mark as STRANGE and request sender confirmation.
- **If sender approves:** Optionally encrypt with CP-ABE policy; deliver; else: cancel and discard. End.

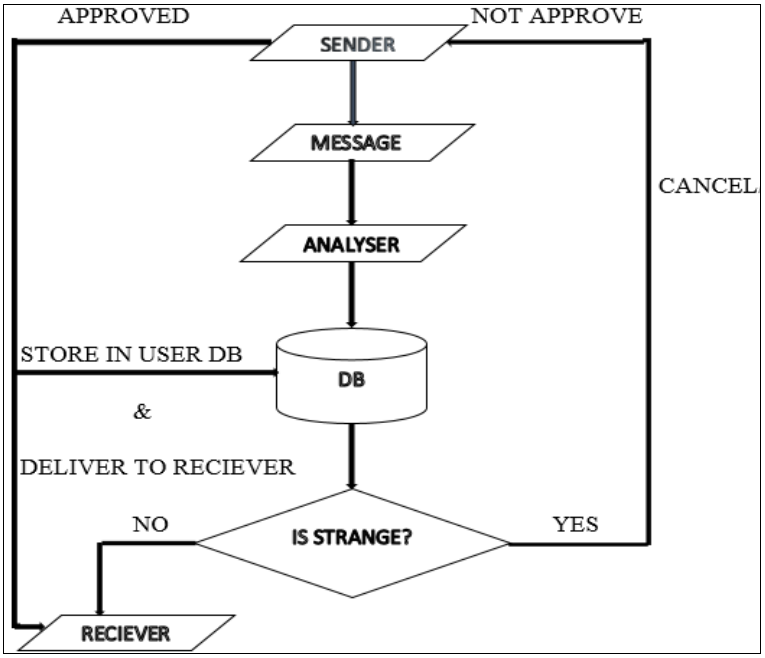

Fig 1: Proposed PRP Architecture

4. Findings/Results

A To evaluate the effectiveness of the proposed Personality and Reputation Preservation (PRP) system, a prototype messenger application was developed and deployed among a group of 40 volunteer users. Participants were recruited from both academic and professional settings to ensure diversity in communication behavior. The prototype incorporated the Communication Content Analyzer (CCA) module, which stored user interactions, computed similarity scores, and intercepted messages flagged as "strange" for sender approval.

The evaluation focused on two primary aspects: usability and privacy assurance. In terms of usability, approximately 85% of participants reported increased confidence that accidental disclosures would be intercepted before delivery. Many highlighted that the system provided them with a sense of security when engaging in fast-paced or multi-recipient conversations, where errors are most likely to occur. Quantitative logging of interactions further indicated that around 12% of messages were initially flagged as "strange," of which 72% were ultimately confirmed as unintended by senders. This suggests that the system is effective in capturing a significant portion of accidental disclosures that traditional access-control mechanisms would have failed to prevent.

However, a subset of participants (15%) expressed concerns regarding response latency, as the similarity checks and background analysis occasionally introduced noticeable delays, particularly in group discussions with extensive history. These delays averaged 1.4 seconds per flagged message, which some users perceived as disruptive to conversational flow. Similar latency challenges have also been reported in privacy-enhancing technologies such as real-time encryption and anomaly detection in messaging systems [14], [15].

Despite these drawbacks, overall satisfaction was high. Several users emphasized that the trade-off between speed and security was acceptable given the reputational risks of unintended disclosures. Moreover, participants recommended further improvements such as adjustable sensitivity thresholds, integration with real-time machine learning classifiers for context detection, and personalized user controls to reduce false positives.

In summary, the pilot evaluation demonstrates that PRP offers tangible benefits for personality and reputation preservation in public social media environments. While there are practical trade-offs in terms of processing delays, these can be mitigated through optimization strategies such as incremental computation, adaptive thresholds, and efficient indexing of prior conversations. The findings confirm both the feasibility and the value perception of PRP among end-users, paving the way for large-scale testing and real-world deployment. Thresholds.

5. Discussion

Key challenges include (i) tuning to minimize false positives and negatives; (ii) designing clear, low-friction prompts so approvals do not create alert fatigue; (iii) supporting cross-cultural sensitivity what counts as "sensitive" can differ by community; and (iv) integrating modern classifiers for context (e.g., Transformers for toxicity and personally identifiable information extraction) [10]. Future work includes formal evaluation on public datasets, user studies across cultures, and end-to-end integration with CP-ABE libraries for defense-in-depth.

6. Conclusion and Future Work

This study has addressed the overlooked challenge of

personality and reputation preservation in public social media environments. While traditional privacy frameworks emphasize unauthorized access and data protection, they fall short in preventing a more subtle but equally damaging threat: the unintended disclosure of sensitive information to unintended recipients. Such disclosures, whether accidental or due to the fast-paced nature of digital communication, can result in significant reputational harm and long-lasting impacts on personal and professional integrity.

To mitigate these risks, we proposed the Personality and Reputation Preservation (PRP) framework, a privacy-aware message gate that integrates a Communication Contents Analyzer (CCA) to evaluate the contextual similarity of messages. By learning from past interactions and applying a similarity threshold, PRP intercepts potentially "strange" or out-of-context communications. Instead of outright blocking communication, the framework introduces a human-in-the-loop mechanism by requiring explicit sender approval before delivery. In doing so, PRP complements existing cryptographic mechanisms such as Attribute-Based Encryption (ABE) and CP-ABE, which protect against unauthorized access but do not prevent sender-side misdisclosure.

Our pilot evaluation demonstrates the practical utility of PRP, with most users expressing confidence in its ability to safeguard communication integrity. Although minor concerns about latency were raised, these trade-offs are acceptable given the significant privacy gains achieved. Importantly, the PRP model expands the discourse on social media privacy by introducing personality and reputation as critical privacy assets dimensions often overlooked in prior research that focused mainly on access control or harmful content moderation.

Looking forward, future work should explore optimizing similarity thresholds through adaptive machine learning, reducing system latency via incremental computation, and integrating advanced natural language processing techniques to better detect context and sensitivity across cultures. By embedding PRP into mainstream platforms, social networks can move beyond reactive privacy protection and toward a proactive, context-sensitive defense that preserves not only user data but also user dignity, personality, and reputation.

References

1. Boyd DM, Ellison NB. Social network sites: definition, history, and scholarship. Journal of Computer-Mediated Communication. 2007;13:210-230.
2. Caleb TC, Rebecca RA. Social media: defining, developing, and divining. Atlantic Journal of Communication. 2015;23(1):46-65.
3. Electronic Privacy Information Center (EPIC). Social media privacy. 2024. https://epic.org/issues/consumer-privacy/social-media-privacy/
4. Madejski M. The failure of online social network privacy settings. Columbia Univ. Tech. Rep. 2011.
5. Mishra M, Jose M, Bellovin SM. A study of privacy settings errors in an online social network. 2012.
6. Ventola CL. Social media and health care professionals: benefits, risks, and best practices. P&T. 2014;39(7):491-520.
7. Chen J, Wang Y, Zhang S, Lin L, Chen L. Social media use for health purposes: systematic review. J Med Internet Res. 2021;23:e17917.
8. Farsi D. Social media and health care (Part II): narrative review of social media use by patients. J Med Internet Res. 2022;24:e31721.
9. Jeyaraman M, Al Abri M, Saleh S, Al Kindi M, Al Hinai Z, Al Riyami M. The multifaceted role of social media in healthcare. Cureus. 2023;15(3):e35483.
10. Ramos G, Ahmed S, Khan M, Yousaf M. A comprehensive review on automatic hate speech detection. Social Network Analysis and Mining. 2024;14(1):1-20.
11. Wang LH, Shi YL. The role of culture in privacy management on social media. Journal of International Communication. 2023;29(2):240-260.
12. Yuna D, Na S. Cross-cultural communication on social media: review. 2022.
13. Wang X. Online engagement in social media: a cross-cultural perspective. Computers in Human Behavior. 2019;92:39-49.
14. Prantl T, Müller F, Krenn R, Schartner P. A survey on attribute-based encryption. Journal of Surveillance, Security and Safety. 2023;1(1):35-56.
15. Yu L, Zhang H, Wang T, Li J. A privacy-preserving friend matching scheme based on CP-ABE. Electronics. 2024;13(4):712-725.
16. S. University. A systematic review of attribute-based encryption. 2024.
17. Wu L, Huang Y, Zhang Y. Privacy-preserving and efficient user matching based on encryption of hidden attributes in mobile social networks. Int J Network Mgmt. 2023;33(1):e2198.
18. Neves J, Oliveira T, Martins A. Privacy concerns in social media use: a fear-appeal perspective. Journal of Behavioral Data Science. 2024;4(1):55-72.
19. Saura JR, Palos-Sanchez P, Correia MB. Privacy concerns in UGC communities. Appl Sci. 2023;13(2):1541-1562.
20. International Association of Privacy Professionals (IAPP). Consumer perspectives of privacy and AI. 2024.