RADMAD: A Resource-Aware Decentralized Moderation and Deployment Framework for Sarcastic Toxicity

Anonymous ACL submission

Abstract

Sarcasm is frequently used to mask malicious intent, driving online toxicity and cyberbullying. While recent emoji-centric models effectively detect sarcasm, deploying them via traditional cloud-based moderation introduces latency and relies on reactive, post-hoc policies. To address these limitations, this paper proposes the Resource-Aware Decentralized Moderation and Deployment (RADMAD) framework for sarcastic toxicity. RADMAD shifts active moderation to the edge via localized inference, enabling low-latency interventions while transmitting only anonymized data for continuous cloud learning. To justify this emoji-centric approach, a statistical validation of 678,823 posts is conducted, establishing emojis as a critical modality present in more than 20% of online posts. A pre-existing EmoCentricSarcBERT model is compressed for edge deployment using techniques like knowledge distillation and post-training quantization. One of the resulting models, QuantizedEmoCentricSarcBERT with reduced RAM usage and physical storage (by 75%) is utilized to demonstrate real-world viability, a stochastic deployment simulation was conducted on a longitudinal dataset of 50 users. This stochastic simulation of users' responses to RADMAD's proactive nudging demonstrates the framework's ability to successfully isolate persistent toxic actors while protecting the compliant majority. Moreover, the quantized model maintains a Cohen's Kappa (κ) of 0.8378 and an R^2 alignment of 0.762737 with the 32-bit cloud baseline.

1 Introduction

Toxic online discourse remains a perpetual challenge for digital platforms, with sarcasm serving as a sophisticated mechanism for cyberbullying and trolling. Sarcasm detection is recognized as a challenging task in sentiment analysis due to the inherent ambiguity between literal and intended meaning (Joshi et al., 2017). By using ironic or satirical utterances designed to mock or insult, aggressors can mask malicious intent behind seemingly benign text, allowing them to evade traditional keyword-based filters while still inflicting psychological harm on victims (Dhumpati et al., 2025; Pamecha et al., 2024; Al-Harigy et al., 2022; Van Hee et al., 2018). This form of sarcastic cyberbullying has become a significant concern in digital spaces, necessitating advanced detection frameworks capable of identifying hostility within complex linguistic contexts.

Despite the efficacy of sarcasm detection models, traditional moderation typically relies on centralized, post-hoc cloud pipelines that identify and remove content only after it has been published. This reactive approach fails to prevent initial exposure to harm and raises significant privacy concerns. Proactive interventions can mitigate these issues by triggering inference during the message creation phase. For instance, research conducted on platforms such as X (formerly Twitter) has demonstrated that real-time evaluation and interventions can result in 34% of participants revising or abandoning their drafts (Katsaros et al., 2022). However, many existing proactive systems focus primarily on explicit toxicity and require cloud-based inference (Islam et al., 2026; Murthy and Siddesh, 2024; Maity et al., 2022), causing latency bottlenecks that make them non-viable for edge devices such as mobile phones.

Furthermore, identifying sarcasm is inherently difficult due to the paucity of context in text-only environments. Consequently, researchers have explored various modalities, including audio and visual cues, to capture non-verbal intent (Chen et al., 2024; Liang et al., 2022; Tomar et al., 2023; Singh and Sharma, 2023; Dutta and Bhattacharyya, 2022). One highly effective modality is emojis. In digital communication, emojis act as crucial paralinguistic cues, effectively serving as digital facial expressions that signal sarcastic intent (Kumar et al.,

2023; Grover, 2022). While emojis are often polysemic, their specific usage, frequency, and placement within a message provide decisive markers for detection (Grover and Banati, 2026, 2024; Chauhan et al., 2022; Jain et al., 2022). Integrating emojis for sarcasm detection helps identify toxic content on platforms where users frequently rely on visual symbols to imply irony or hostility.

To address these gaps, this paper proposes the Resource-Aware Decentralized Moderation and Deployment (RADMAD) framework for sarcastic toxicity. RADMAD shifts active moderation from the cloud to the edge, enabling low-latency, localized evaluation and real-time interventions, while transmitting only anonymized data to the cloud for continuous learning. While RADMAD primarily employs an emoji-centric approach for lightweight edge inference, simulation results demonstrate that the pipeline also remains a highly viable deployment strategy for emoji-free text.

The primary contributions of this work include:

- **Validation of emojis as a viable modality:** A statistical analysis of 678,823 posts confirms that emojis are present in $\geq 20\%$ of online posts, establishing them as a reliable modality for text analysis.

- **Edge models:** The creation of resource-aware edge models by employing compression strategies, such as Knowledge Distillation and Post-Training Quantization, on a pre-existing EmoCentricSarcBERT baseline. The resulting 8-bit model (QuantizedEmoCentricSarcBERT) will be made publicly available to the research community via HuggingFace.

- **RADMAD framework:** A decentralized, edge-centric moderation pipeline that integrates on-device behavioural intervention with a cloud-learning loop. It enables low-latency, privacy-preserving sarcasm detection and proactive *nudging* to deter sarcastic toxicity in online discourse.

- **Stochastic deployment simulation of RADMAD framework:** Demonstration of the efficacy of the RADMAD framework through a stochastic simulation of an experimental dynamic scoring system. Longitudinal simulations on the social media posts of 50 real-world users incorporating a Sarcasm Score (S), Nudge Penalty (λ), and Recovery Score (ζ) demonstrate the framework's ability to isolate or rehabilitate users based on their response to nudges over a time frame.

- **Assessment of base and edge model inference agreement:** A statistical agreement analysis of EmoCentricSarcBERT and QuantizedEmoCentricSarcBERT showing that the 8-bit compression maintains a Cohen's Kappa of 0.8723 with cloud baselines while significantly reducing hardware overhead.

2 Related Work

Recent literature addressing online toxicity and its mitigation spans three distinct domains: sarcasm detection as an indicator of implicit toxicity, resource-efficient inferencing for edge devices, and localized proactive moderation techniques. Some of the notable work across these domains are discussed as follows.

2.1 Modalities for Sarcasm Detection

Accurately detecting sarcasm is essential for maintaining digital discourse integrity, as it is frequently leveraged to disseminate satirical content, fake news, and facilitate implicit cyberbullying (Pamecha et al., 2024; Singh and Sharma, 2023; Moores and Mago, 2022). Traditional approaches typically rely on heavy multimodal feature fusion or ensemble methods to capture non-verbal signals and disambiguate sentiment (Tomar et al., 2023; Dutta and Bhattacharyya, 2022; Liang et al., 2022; Badlani et al., 2019; Pan et al., 2020; Castro et al., 2019). The identification of implicit abuse, misogyny, and hate speech heavily benefits from sarcasm and irony-aware models (Cabrera et al., 2025; Frenda et al., 2023; Muaad et al., 2022). Various studies have demonstrated that integrating sarcasm detection significantly improves cyberbullying classification (Philipo et al., 2026; Guo and Gauch, 2024; Maity et al., 2022; Chia et al., 2021). Recent sarcasm detection models frequently employ multimodal approaches such as complex image-text pairs and social interaction graphs along with architectures like transformers, fusion architectures, and graph based networks among others (Grover and Banati, 2026; Khan et al., 2025; Liu et al., 2024; Fang et al., 2024a,b; Liang et al., 2024; Tian et al., 2023; Potamias et al., 2020). However, these architectures demand significant computational overhead and rely on modalities like video, audio, and images that are expensive for low-latency, text-first

social media settings. In text-first environments, emojis serve as paralinguistic features that can be processed concurrently with text. Consequently, they represent a computationally efficient alternative to aforementioned resource-intensive modalities. Research demonstrates that augmenting text with emojis substantially improves sarcasm and cyberbullying detection in the absence of other multimodal cues (Ibrahim et al., 2025; Kusaila and Gerhart, 2025; Kumar et al., 2023; Mohan et al., 2023; Grover, 2022; Liang et al., 2022; Danesi, 2021; Filik et al., 2016). For instance, emoji-focused models like EmoCentricSarcBERT significantly outperform standard text-only baselines (Grover and Banati, 2024, 2026). Despite these robust detection capabilities, existing emoji-aware models like EmoCentricSarcBERT have a high parameter count and massive footprint making them unsuitable for deployment on resource-constrained edge devices.

2.2 Resource-Efficient Sarcasm Detection Models

Deploying robust transformer models to resource-limited devices requires significant architectural compression to reduce parameter count, memory footprint, and inference latency. While techniques like pruning and weight-sharing reduce size, Knowledge Distillation (KD) has emerged as a primary method for preserving generalization ability. KD trains a compact student model to replicate the softened output probabilities of a larger teacher via temperature scaling (Hinton et al., 2015). This approach produces highly efficient architectures; for example, DistilBERT utilizes pre-trained distillation to retain 97% of BERT's baseline performance with only 40% of the parameters (Sanh et al., 2019). Similarly, TinyBERT employs a two-stage distillation process for massive inference speedups (Jiao et al., 2020), while MobileBERT incorporates inverted-bottleneck structures to achieve significant on-device acceleration (Sun et al., 2020). However, KD requires computationally expensive layer-wise feature matching and extensive retraining. Alternatively, Post-Training Quantization (PTQ) achieves hardware-level compression by mapping continuous 32-bit floating-point weights directly into discrete 8-bit integer bins. This bypasses retraining overhead and substantially accelerates inference latency, as mobile processors compute integer matrices highly efficiently. Although forcing continuous weights into rigid integer slots introduces quantization errors, PTQ offers an optimal trade-off between semantic resolution and deployment speed. This paper bridges these domains by empirically benchmarking both task-specific KD and PTQ to distil the semantic capacity of an emoji-centric sarcasm detection model into edge format.

2.3 Proactive Inference Systems

Reactive moderation relies on massive, centralized cloud pipelines to process and remove harmful content post-dissemination. Conversely, proactive inference systems integrate models and moderation mechanisms directly during message composition to tackle toxic content prior to publication. Studies indicate that accuracy and privacy nudges can effectively restrict unintended disclosures and misinformation without rigid blocking mechanisms (Pennycook and Rand, 2022; Kitkowska et al., 2020; Wang et al., 2014, 2013). In practice, real-time evaluation pipelines, such as the Reconsidering Tweets initiative, triggered review prompts that achieved a 34% revision or abandonment rate for flagged replies (Katsaros et al., 2022). Similarly, Google's Perspective API deployed proactive cloud-based warnings utilizing multilingual character-level transformers (Lees et al., 2022). Recent work has also explored cloud-based psychological nudging mechanisms that intercept cyberbullying attempts by translating toxic words into positive multilingual equivalents and presenting simulated affirmative social reactions to induce long-term behavioral shifts (Hirano and Tanaka, 2024). However, many of these architectures fundamentally rely on continuous cloud connectivity, demanding the transmission of unedited text drafts to central servers, thereby introducing latency bottlenecks and privacy vulnerabilities. Existing edge-based alternatives, such as the ReThink application (Prabhu, 2019), operate locally to preserve privacy but often lack the representational capacity to process nuanced, implicit toxicity such as sarcasm. The RADMAD framework presented in this work addresses this gap by synthesizing on-device, emoji-centric text inference with a configurable penalty and recovery mechanism entirely at the edge.

3 Methodology and The RADMAD Framework

The methodology for this work is structured across three foundational phases. First, the prevalence of emojis is statistically validated to justify their

use as an effective modality for text analytics in online discourse. Second, Knowledge Distillation (KD) and Post-Training Quantization (PTQ) are employed to compress an existing model (EmoCentricSarcBERT) (Grover and Banati, 2026) to meet edge-hardware constraints. Finally, the RAD-MAD framework is detailed, proposing a nudge-based proactive self-moderation approach utilizing the compressed edge-based sarcasm detection model.

3.1 Emojis: A Viable Modality for Text Analysis

To empirically establish the prevalence of emojis in modern online discourse, a sampling frame of 678,823 social media posts spanning 2022 to 2025 across various platforms (e.g., X, Facebook, Reddit, YouTube, Instagram) was aggregated. A series of binomial hypothesis tests evaluated the true proportion (p) of posts containing emojis within a population of $678,823$ records, testing against baseline thresholds $p_0 \in \{0.15, 0.16, \ldots, 0.23\}$ ($H_0 : p \leq p_0$ vs. $H_a : p > p_0$). With parameters set for a one-tailed proportion test ($Z = 1.64$, $\alpha = 0.05$, $E = 0.02$, $1 - \beta = 0.90$), the analysis was conducted on a sample size of $3,775$ posts for three independent iterations for each baseline proportion p_0. It was conclusively determined with $\alpha = 0.05$ that more than 20% of social media posts contain emojis. Complete historical data, p-values, and test statistics for these hypotheses are provided in *Appendix A, Table 7*.

3.2 Edge Optimization via Model Compression

The efficacy of leveraging emojis for sarcasm and cyberbullying detection has been firmly established in recent literature. Building upon this modality, a base model EmoCentricSarcBERT, which features an emoji-augmented vocabulary and is fine-tuned on the SarcOji dataset (Grover and Banati, 2022). While EmoCentricSarcBERT has proven effective on unseen datasets such as SarcOjiTest1 and SarcOjiTest2 (Grover and Banati, 2024, 2026), its substantial size necessitates significant architectural compression before it can be deployed in mobile environments. For this, two distinct optimization strategies were utilized: Knowledge Distillation (KD) and Post Training Quantization (PTQ). For KD, a compact student model is trained to mimic the feature representations of the teacher model. A composite loss function was utilized, combining the Cross-Entropy loss between the students predicted logits and true labels with the Kullback-Leibler (KL) divergence between the temperature-scaled soft probabilities of the teacher and student. The total loss $\mathcal{L}$ is computed as:

$$\mathcal{L} = \alpha \mathcal{L}_{CE}(S_x, y) + (1-\alpha) T^2 \mathcal{L}_{KL}\left(\sigma\left(\frac{S_x}{T}\right), \sigma\left(\frac{T_x}{T}\right)\right) \tag{1}$$

where σ represents the softmax function. Extensive grid searches across temperature (T) and balance (α) hyperparameters were conducted to distill EmoCentricSarcBERT into DistilBERT, TinyBERT, and MobileBERT architectures (detailed in Appendix B). While KD physically reduces architectural capacity, it still relies on high-precision mathematics. Alternatively, PTQ serves as a direct hardware-level optimization. The architecture is exported into a static Open Neural Network Exchange (ONNX) format, and the high-precision 32-bit floating-point weights (W_{FP32}) are mathematically mapped directly into a discrete 8-bit integer space ($q_{min} = -128$ to $q_{max} = 127$). This is achieved by computing scale factors (S) and zero-point offsets (Z):

$$S = \frac{\max(W_{FP32}) - \min(W_{FP32})}{q_{max} - q_{min}} \tag{2}$$

By calculating specific scale factors and zero-point offsets for each layer:

$$Z = \text{Round}\left(q_{min} - \frac{\min(W_{FP32})}{S}\right) \tag{3}$$

$$W_{INT8} = \text{Clamp}\left(\text{Round}\left(\frac{W_{FP32}}{S}\right) + Z, q_{min}, q_{max}\right) \tag{4}$$

PTQ yields the QuantizedEmoCentricSarcBERT model, dramatically reducing storage size and inference latency without the extensive compute overhead of retraining. Crucially, among the compressed variants, the quantized model exhibited minimal performance degradation and most closely mirrored the predictive capabilities of the full-precision baseline. A comprehensive performance comparison on SarcOjiTest1 and SarcOjiTest2 between the quantized model, the base EmoCentricSarcBERT, and the knowledge-distilled models is detailed in *Appendix B, Tables 8, 9, 10*.

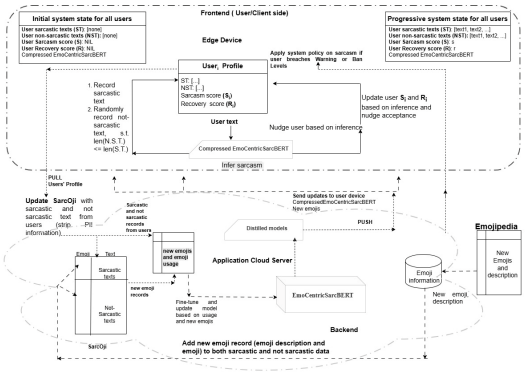


Figure 1: The execution pipeline of the RADMAD framework.

3.3 The RADMAD Framework

With the edge-viable model established, the RADMAD framework operates as a continuous feedback loop between the user's localized edge device and a centralized cloud platform (Figure 1). The cloud platform maintains the base EmoCentricSarcBERT model, the central SarcOji dataset, and system policy threshold. Simultaneously, the user's edge device is initialized with the QuantizedEmoCentricSarcBERT inference engine, a localized text repository, and a baseline Sarcasm score (S) that indicates the user's behaviour within the system. During real-time message composition, the edge model evaluates drafted text locally to classify sarcastic intent. This inference occurs entirely on-device, ensuring unedited and potentially sensitive drafts are never transmitted over the network. If a message is flagged as sarcastic, the system delivers a localized UI nudge to prompt a non-sarcastic rephrase. To ensure the foundational model accurately adapts to evolving language and data drift, localized texts are periodically stripped of all Personally Identifiable Information (PII) and transmitted to the cloud. The cloud server leverages this anonymized data to fine-tune the base model, subsequently pushing an updated, quantized inference engine back to the edge nodes. The edge-based Sarcasm Score (S) tracks a users cumulative toxicity load. This score increments by a fixed Nudge Penalty (λ) when nudges are ignored and decays by a Recovery Score (ζ) through non-sarcastic interactions. As S fluctuates, RADMAD enforces a policy-specific escalation mechanism when S reaches a system threshold θ. For example, reaching a warning threshold θ_w flags an account as "Risky," while breaching the absolute threshold (θ) triggers a ban. RADMAD offers a flexible architecture where θ, θ_w, ζ, λ, δ and associated policies can be adapted to platform-specific needs and the evolving nature of online emoji discourse. This ensures the framework remains relevant as communication practices and emoji usage shift over time. The complete execution pipeline of the RADMAD framework is detailed in *Appendix C*. The experimental results indicate that the proposed framework performs robustly across both emoji-free and emoji-integrated texts.

4 Experimental Setup and Simulation

To empirically evaluate the viability of the RADMAD framework and demonstrate its readiness for real-world deployment, an operational thresholding policy was tested using a longitudinal simulation. While the core architecture is policy-agnostic and capable of supporting varied moderation guidelines, this evaluation instantiates an experimental proportional-recovery policy to model user behavior. Executed on a CPU architecture to mirror edge-device constraints, the evaluation utilized the social media posts of 50 real-world users (Grover and Banati, 2023). To mirror live system dynamics, these historical posts were processed chronologically over a two-year timeline. Each user is assigned a stochastic nudge acceptance probability (p_i), determining user i's likelihood of complying with system nudges to rephrase sarcastic content. The Sarcasm Score (S) quantifies a users cumulative toxicity load. A score of $s = 0.0$ represents a baseline state, signifying that the user is in full compliance with the system's moderation policies. If a user ignores a system nudge and publishes flagged text, their score is incremented by a fixed Nudge Penalty (λ). To prevent premature banning of users who occasionally use sarcasm, the experimental framework integrates a Recovery Score (ζ). This acts as a mathematical forgiveness mechanism that dynamically balances against penalty points based on the platform's allowed sarcasm rate (p), calculated as:

$$\zeta = \frac{\lambda \cdot p}{1 - p} \tag{5}$$

Conversely, a persistently non-compliant user exhibiting an elevated frequency of sarcasm (p_s) accumulates penalties faster than the system's recovery rate. This expected net increase in their penalty score per interaction is defined as the Troll Drift (Δ):

$$\Delta = \lambda p_s - \zeta(1 - p_s) \tag{6}$$

Using this established expected drift, a Ban Threshold (θ) is defined by setting the total number of interactions (γ) the system tolerates from a non-compliant user before triggering administrative action. To facilitate preemptive intervention, a corresponding Warning Threshold (θ_w) is established using a fractional coefficient (α) to flag at-risk accounts prior to restriction:

$$\theta = \gamma \cdot \Delta, \quad \theta_w = \alpha \cdot \theta \tag{7}$$

Table 1 defines the hyperparameter configuration for this work. It is important to note that these specific thresholds and recovery rates are experimental parameters employed for this simulation. Real-world deployment would require extensive, platform-specific research to optimize these policies based on unique community guidelines.

Parameter	Symbol	Value
Nudge Penalty	λ	1.0
Baseline Sarcasm Rate	p	0.35
Recovery Factor	ζ	0.538
Worst-case Sarcasm Rate	p_s	0.85
Expected Troll Drift	Δ	0.769
Target Ban Horizon	γ	300
Ban Threshold	θ	230.77
Warning Threshold	θ_w	115.38

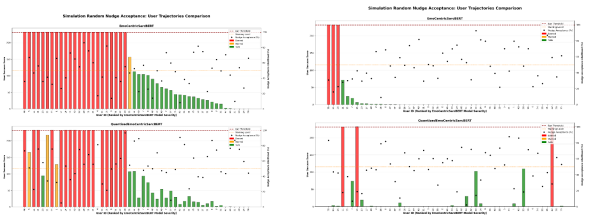
Table 1: Hyperparameter configurations for the RADMAD simulation.

5 Results and Discussion

This section details the empirical evaluation of the RADMAD framework across both hardware constraints and behavioral outcomes.

5.1 Architectural Efficiency

To validate the viability of decentralized deployment, the predictive performance and hardware requirements of the 32-bit EmoCentricSarcBERT baseline were compared against the 8-bit QuantizedEmoCentricSarcBERT model. The high-precision teacher established the upper bound for NLP performance, while the quantized edge model successfully reduced its physical storage footprint by approximately 75% (from 417.21 MB to 106.39 MB) with only a marginal drop in prediction performance. Crucially for mobile deployment, the Post-Training Quantization process drastically reduced active RAM utilization from 478.60 MB down to a highly constrained 32.49 MB, while nearly halving the CPU inference time. This confirms that the compressed architecture provides the real-time processing speed and memory efficiency necessary to trigger localized inference pipelines natively, without bottlenecking host device resources or requiring external compute calls. Table 2 reports the performance across both the Zero-compliance and Stochastic compliance scenarios, evaluating the complete set of user records alongside the specific subset containing emojis."



(a) Zero compliance scenario (b) Stochastic compliance scenario

Figure 2: Comparison of the Zero-compliance and Stochastic-compliance scenarios. Simulating nudge acceptance significantly reduces the number of banned users (Figure (b)), empirically demonstrating RADMAD's effectiveness as a moderation tool.

5.2 Simulation Outcomes

A longitudinal simulation of 50 users evaluated the system under two behavioral conditions: Zero compliance (worst-case) and Stochastic (Probabilistic) compliance. Under Zero Compliance, where users strictly ignored nudges, both models exhibited near-perfect alignment. Because users rejected all nudges, they absorbed continuous penalties and rapidly breached the ban threshold. Notably, the base model consistently calculated slightly higher sarcasm scores, indicating that weight compression causes the edge model to occasionally drop sensitivity to edge-case sarcasm.

As observed from Figure 2, in the Stochastic scenario, users with a high nudge-acceptance likelihood maintained a baseline equlibrium state, as the mathematical recovery factor (ζ) actively offset their occasional penalties. This validates that proactive nudging, combined with a recovery loop, systematically isolates genuinely toxic actors while rehabilitating compliant users.

While metrics such as MAE, R^2, and KL Divergence confirm the 8-bit model accurately preserves the teacher's continuous probability distribution, high Cohen's Kappa scores demonstrate

Scenario	Text Filters	Evaluated Posts	CPU Inference Time (s)		Active RAM Used (MB)	
			Base	Quantized	Base	Quantized
Zero compliance	No filter (all records)	34269	7656.67	4321.33	478.60	32.49
	Records with emojis	27731	6836.55	2935.14	459.09	23.44
Stochastic compliance	No filter (all records)	66659	13349.61	6143.16	502.48	38.79
	Records with emojis	27731	6232.98	2860.89	477.11	24.36

Table 2: Architectural Efficiency and Hardware Performance. (Inference metrics are based on the continuous evaluation of 34,269 raw posts strictly on CPU).

Scenario	Text Filters	Evaluated Posts	Users Banned		Disagreed Posts
			Base	Quant.	
Zero comp.	No filter	34269	23	18	2740
	With emojis	27731	8	8	1703
Stochastic comp.	No filter	66659	2	1	5014
	With emojis	27731	0	2	1703

Table 3: Users' evaluation statistics for EmoCentricSarcBERT and QuantizedEmoCentricSarcBERT.

robust inter-rater reliability for discrete labels (Table 4). Statistically, the quantized model maintained a highly robust Cohens Kappa (κ) of 0.8378 and an R^2 of 0.7627 with the cloud baseline on unfiltered text. This dual-level assessment proves extreme integer compression retains high fidelity in both model logic and behavioral interventions. The complete user trajectory scatter plots (for all posts and filtered set with emoji posts) are provided in *Appendix D, figures 3, 4, 5, and 6.*

5.3 Model Disagreement Analysis

A qualitative analysis of conflicting predictions (Table 5) reveals that performance degradation occurs primarily in brief, informal, or emoji-dense texts, introducing two primary failure modes.

Shorter, emoji-heavy texts: For extremely short inputs consisting solely of user tags and emojis, the edge model frequently assigns false positives (1). Without sufficient lexical context, the smaller model over-interprets standalone emojis as having complex, non-literal meanings rather than genuine expressions.

Loss of context in slang and brevity: When processing brief texts, slang, or external links, the edge model often predicts false negatives (0). The representational capacity lost during quantization impairs the model's ability to resolve semantic ambiguity, indicating the edge model relies heavily on explicit lexical cues compared to the baseline's higher tolerance for nuanced internet discourse.

5.4 Deployment Implications and Challenges

Executing inference locally carries significant privacy advantages: potentially sensitive, unedited drafts are never transmitted to centralized servers. Cloud communication is *strictly reserved for anonymized data to mitigate linguistic drift*. Crucially, RADMAD transitions moderation from a reactive, post-hoc pipeline to an *a priori* proactive system. Introducing constructive algorithmic friction during message composition intercepts implicit toxicity before dissemination.

However, the framework introduces complex deployment challenges. While the thresholding system mathematically isolates toxicity, determining permissible sarcasm rates is a highly subjective policy choice. Overly strict configurations risk generating unsustainable friction, potentially suppressing legitimate satirical commentary. Therefore, deploying RADMAD requires dynamic threshold calibration tailored to the specific cultural and social interaction goals of the host platform.

6 Conclusion

This paper introduces the Resource-Aware Decentralized Moderation and Deployment (RADMAD) framework to address the latency, privacy, and compute limitations of traditional cloud-based moderation. By validating emojis as highly prevalent paralinguistic anchors, this study established an empirical foundation for computationally lightweight classification. The research demonstrated that Post-Training Quantization successfully distills a complex transformer (EmoCentricSarcBERT) into an 8-bit edge-viable format. This compression reduces physical storage by 75% and drastically minimizes RAM utilization while maintaining a highly robust Cohens Kappa of 0.8723 against the 32-bit baseline. Ultimately, by shifting the initial inference workload to the mobile edge, RADMAD provides social media platforms with a resource-efficient, privacy-preserving paradigm to proactively deter

Scenario	Text Filters	MAE	RMSE	R^2	Mean KL Div.	Cohen's Kappa
Zero compliance	No filter	0.079549	0.227841	0.771260	0.225961	0.8399
	Records with emojis	0.060179	0.200960	0.818470	0.184972	0.8723
Stochastic compliance	No filter	0.073741	0.219741	0.762737	0.209329	0.8378
	Records with emojis	0.060179	0.200960	0.818470	0.184972	0.8723

Table 4: Statistical alignment of EmoCentricSarcBERT and QuantizedEmoCentricSarcBERT.

Text Example	Base Model	Compressed Model
"@needytommo Yeh babyyyyyy"	0	1
"@_sxphss_ ... lets get it into the millionsssss"	0	1
#americanidol"	0	1
"@DANIELLEJTROT yeeesss 💚"	0	1
"@sweetnsaltymama 🛒 🤗 💛"	0	1
"@tobi just applied 🤞"	0	1
"sad [URL]"	1	0
"I missed my bro too lol [URL]"	1	0
"Oh, brilliant. Another flat tire. My day is just going perfectly."	1	0
"Wow, I absolutely love it when my Wi-Fi cuts out in the middle of a meeting."	1	0
"@13minuteslate 🖤"	1	0
"Aww Damn! I was headed there to watch our game tonight and have a drink! Welp. 🧐"	1	0
"Lets all take a moment to appreciate this beautiful dish from Microsoft Dining. 🤩"	1	0

Table 5: Examples of prediction disagreements between EmoCentricSarcBERT and QuantEmoCentricSarcBERT.

sarcastic toxicity.

Limitations

While the RADMAD framework demonstrates significant architectural potential for edge-centric natural language inference, several limitations regarding modality scope, data drift, privacy protocols, and threshold calibration must be acknowledged.

- **Reliance on emojis:** To strictly adhere to low computational overhead and on-device memory constraints, this architecture intentionally isolates text and lightweight emojis, avoiding heavy multimodal data such as images or audio. Consequently, the edge framework cannot process sarcasm derived from multimodal incongruity (e.g., benign text paired with an ironic image). Future work must investigate combining this lightweight pipeline with broader, server-side multimodal architectures to create a comprehensive moderation tier without sacrificing edge efficiency.

- **Linguistic data drift:** Digital conversational streams frequently utilize rapidly evolving regional slang and code-mixed languages. The current foundational SarcOji dataset requires continuous expansion to include code-mixed text and culturally specific paralinguistic symbols to prevent semantic degradation. This ensures equitable classification across diverse edge environments.

- **Privacy guarantees during updates:** Although the RADMAD pipeline strips localized data of explicit Personally Identifiable Information (PII) prior to cloud transmission for model refinement, centralized dataset aggregation still presents theoretical vulnerabilities. The current architecture lacks cryptographic privacy guarantees. Future iterations must explore privacy-preserving update protocols, such as federated learning or differential privacy, to secure user data payloads while allowing the foundational transformer to adapt to linguistic drift.

- **Algorithmic threshold calibration:** The stochastic deployment simulation evaluates longitudinal viability using static, experimentally defined parameters (ζ, Δ, θ, λ, p). In live production environments, rigid parameters risk generating unacceptable levels of false-positive system friction. For proactive inference to operate seamlessly, the framework requires live A/B deployment testing to dynamically calibrate thresholding mechanisms based on true user mitigation yields, ensuring ambiguous edge cases are processed with appropriate algorithmic tolerance.

7 Ethical Considerations

This research focuses on the proactive mitigation of sarcastic toxicity and cyberbullying in online discourse. In developing and evaluating the RADMAD framework, the following ethical principles were observed to ensure responsible data handling and to align with the ACL Ethics Policy.

Data Privacy and Anonymization The empirical evaluations and longitudinal simulations conducted in this study utilized text and emoji data sourced exclusively from Twitter posts (up to the year 2021, complying with Terms of Service for Twitter in the year 2021). Specifically, this dataset comprised publicly available posts (tweets) from high-profile public figures, including celebrities and verified brand accounts. Because these accounts operate in a highly public capacity, they carry a significantly diminished expectation of privacy compared to private individuals. This study involved strictly passive observational evaluation; no direct interactions with the users or accounts were conducted.

Despite the public nature of the data, rigorous anonymization was applied to protect user identities. Prior to analysis and model simulation, unique User IDs (UIDs) and recognizable account handles were systematically stripped from the corpus.

Intended Use and Privacy by Design The core objective of the RADMAD framework is to encourage healthier digital communities by introducing constructive algorithmic friction, not to facilitate surveillance or punitive censorship. Traditional cloud-based moderation requires the transmission of users' unedited, potentially sensitive text drafts to centralized servers. In contrast, RADMAD was explicitly engineered with privacy-by-design principles. By shifting the predictive inference of the QuantizedEmoCentricSarcBERT model directly to the user's edge device, the framework ensures that unedited drafts never leave the user's mobile environment. For the proposed asynchronous model updates to the centralized cloud platform, it is designed to anonymize data and store these records without users' personally identifiable information, mitigating the surveillance risks inherent in modern content moderation architectures.

Potential Misuse and Bias Automated moderation systems, including those based on sarcasm and toxicity detection, carry an inherent risk of misuse. If deployed with overly aggressive algorithmic thresholds, such systems could inadvertently suppress legitimate satirical commentary or disproportionately affect communities that frequently utilize sarcasm as a coping mechanism or in-group dialect. The RADMAD framework addresses this by remaining strictly advisory (via UI nudging) during the draft phase, allowing users to retain agency over their final published discourse. Future deployments of this framework must ensure that the underlying datasets are continuously updated to reflect diverse, multi-demographic emoji usage to prevent representational biases. At present, this work focuses on the English language. In future, this approach can be applied to diverse set of languages, including low-resource languages and code-mixed discourse.

8 Declaration

The author(s) utilized Google Gemini to refine the manuscript's writing and to check mathematical formula and code for errors. The author(s) reviewed and edited all AI-assisted text and code and take full responsibility for the content and accuracy of this paper.

References

Lulwah M Al-Harigy, Hana A Al-Nuaim, Naghmeh Moradpoor, and Zhiyuan Tan. 2022. Building towards automated cyberbullying detection: A comparative analysis. *Computational Intelligence and Neuroscience*, 2022(1):4794227.

Rohan Badlani, Nishit Asnani, and Manan Rai. 2019. Disambiguating sentiment: An ensemble of humour, sarcasm, and hate speech features for sentiment classification. *W-NUT*, 2019:337–345.

Angelly Cabrera, Linus Lei, and Antonio Ortega. 2025. Transfer learning via lexical relatedness: A sarcasm and hate speech case study. *arXiv preprint arXiv:2508.16555*.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4619–4629.

Dushyant Singh Chauhan, Gopendra Vikram Singh, Aseem Arora, Asif Ekbal, and Pushpak Bhattacharyya. 2022. An emoji-aware multitask framework for multimodal sarcasm detection. *Knowledge-Based Systems*, 257:109924.

Wangqun Chen, Fuqiang Lin, Guowei Li, and Bo Liu. 2024. A survey of automatic sarcasm detection: Fundamental theories, formulation, datasets, detection methods, and opportunities. *Neurocomputing*, 578:127428.

Zheng Lin Chia, Michal Ptaszynski, Fumito Masui, Gniewosz Leliwa, and Michal Wroczynski. 2021. Machine learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection. *Information Processing & Management*, 58(4):102600.

Marcel Danesi. 2021. The law and emojis: Emoji forensics. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, 34(4):1117–1139.

Raghu Dhumpati, Archana Sasi, Shaik Johny Basha, Suryanarayana Vadhri, P Saroja, Swathi Alluri, and Ramesh Vatambeti. 2025. Enhancing sarcasm detection in sentiment analysis for cyberspace safety using advanced deep learning techniques. *Scientific Reports*, 15(1):22681.

Poulami Dutta and Chandan Kumar Bhattacharyya. 2022. Multi-modal sarcasm detection in social networks: A comparative review. In *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 207–214. IEEE.

Hong Fang, Dahao Liang, and Weiyu Xiang. 2024a. Multi-modal sarcasm detection based on multi-channel enhanced fusion model. *Neurocomputing*, 578:127440.

Hong Fang, Dahao Liang, and Weiyu Xiang. 2024b. Single-stage extensive semantic fusion for multi-modal sarcasm detection. *Array*, 22:100344.

Ruth Filik, Alexandra urcan, Dominic Thompson, Nicole Harvey, Harriet Davies, and Amelia Turner. 2016. Sarcasm and emoticons: Comprehension and emotional impact. *Quarterly Journal of Experimental Psychology*, 69(11):2130–2146.

Simona Frenda, Viviana Patti, and Paolo Rosso. 2023. When sarcasm hurts: Irony-aware models for abusive language detection. In *International conference of the cross-language evaluation forum for european languages*, pages 34–47. Springer.

V. Grover and H. Banati. 2024. An attention approach to emoji focused sarcasm detection. *Heliyon*, 10(17):e36214.

V Grover and H Banati. 2026. An emoji centric approach to sarcasm detection in online discourse. *Scientific Reports*, 16(1):3891.

Vandita Grover. 2022. Exploiting emojis in sentiment analysis: A survey. *Journal of The Institution of Engineers (India): Series B*, 103(1):259–272.

Vandita Grover and Hema Banati. 2022. Understanding the sarcastic nature of emojis with sarcoji. In *Proceedings of the Fifth International Workshop on Emoji Understanding and Applications in Social Media*, pages 29–39.

Vandita Grover and Hema Banati. 2023. Emorile: a personalised emoji prediction scheme based on user profiling. *International Journal of Business Intelligence and Data Mining*, 22(4):470–485.

Xiaoyu Guo and Susan Gauch. 2024. Using sarcasm to improve cyberbullying detection. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying@ LREC-COLING-2024*, pages 52–59.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Taichi Hirano and Fumihide Tanaka. 2024. Development of sns applications to control cyberbullying with nudge using multilingual translation. In *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3974–3979. IEEE.

Shaikh Ambreen Mohd Ibrahim, Manoj M Deshpande, and Vijaykumar N Pawar. 2025. Automated sarcasm detection in english tweets using ccnn and ellstm with text and emoji embeddings. In *2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, pages 1–8. IEEE.

Md Shariful Islam, Md Jalal Uddin Chowdhury, Md Misbah Uddin Tareq, Tufayel Ahmed, Md Aktarujjaman, Md Shahab Uddin, Samah Alshathri, Ahmad Taher Azar, and Walid El-Shafai. 2026. Sarcasmsense: A novel multitask learning framework for vlog-based political sarcasm and irony detection. *IEEE Access*.

Deepak Kumar Jain, Akshi Kumar, and Saurabh Raj Sangwan. 2022. Tana: The amalgam neural architecture for sarcasm detection in indian indigenous language combining lstm and svm with word-emoji embeddings. *Pattern Recognition Letters*, 160:11–18.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4163–4174.

A. Joshi, P. Bhattacharyya, and M. J. Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

M. Katsaros, K. Yang, and L. Fratamico. 2022. Reconsidering tweets: Intervening during tweet creation decreases offensive content. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 16, pages 443–453.

Shumaila Khan, Iqbal Qasim, Wahab Khan, Khursheed Aurangzeb, Javed Ali Khan, and Muhammad Shahid Anwar. 2025. A novel transformer attention-based approach for sarcasm detection. *Expert Systems*, 42(1):e13686.

Agnieszka Kitkowska, Yefim Shulman, Leonardo A Martucci, and Erik Wästlund. 2020. Psychological effects and their role in online privacy interactions: A review. *IEEE Access*, 8:21236–21260.

Akshi Kumar, Saurabh Raj Sangwan, Adarsh Kumar Singh, and Gandharv Wadhwa. 2023. Hybrid deep learning model for sarcasm detection in indian indigenous language using word-emoji embeddings. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(5):1–20.

Daniel Kusaila and Natalie Gerhart. 2025. Sarcastic emojis: why dont you understand? *Global knowledge, memory and communication*, 74(5-6):1572–1589.

Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3197–3207.

Bin Liang, Lin Gui, Yulan He, Erik Cambria, and Ruifeng Xu. 2024. Fusion and discrimination: A multimodal graph contrastive learning framework for multimodal sarcasm detection. *IEEE Transactions on Affective Computing*, 15(4):1874–1888.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1767–1777.

Hao Liu, Runguo Wei, Geng Tu, Jiali Lin, Cheng Liu, and Dazhi Jiang. 2024. Sarcasm driven by sentiment: A sentiment-aware hierarchical fusion network for multimodal sarcasm detection. *Information Fusion*, 108:102353.

Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 1739–1749.

Anuraj Mohan, Abhilash M Nair, Bhadra Jayakumar, and Sanjay Muraleedharan. 2023. Sarcasm detection using bidirectional encoder representations from transformers and graph convolutional networks. *Procedia Computer Science*, 218:93–102.

B. Moores and V. Mago. 2022. A survey on automated sarcasm detection on twitter. *arXiv preprint arXiv:2202.02516*.

Abdullah Y Muaad, Hanumanthappa Jayappa Davanagere, JV Bibal Benifa, Amerah Alabrah, Mufeed Ahmed Naji Saif, D Pushpa, Mugahed A Al-Antari, and Taha M Alfakih. 2022. Artificial intelligence-based approach for misogyny and sarcasm detection from arabic texts. *Computational Intelligence and Neuroscience*, 2022(1):7937667.

Jamuna S Murthy and GM Siddesh. 2024. A smart video analytical framework for sarcasm detection using novel adaptive fusion network and sarcasnet-99 model. *The Visual Computer*, 40(11):8085–8097.

Pinky Pamecha, Chaitya Shah, Divyam Jain, Kashish Gandhi, Kiran Bhowmick, and Meera Narvekar. 2024. Cyberbullying or just sarcasm? unmasking coordinated networks on reddit. *arXiv preprint arXiv:2410.20170*.

H. Pan, Z. Lin, P. Fu, Y. Qi, and W. Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of EMNLP 2020*, pages 1383–1392.

Gordon Pennycook and David G Rand. 2022. Nudging social media toward accuracy. *The Annals of the American Academy of Political and Social Science*, 700(1):152–164.

Adamu Philipo, Jianguo Ding, Doreen Sarwatt, Jumanne Mohamed, Afidhu Yusufu, Mahmoud Daneshmand, and Huansheng Ning. 2026. Sentiment-enhanced cyberbullying detection models on social media platforms. *ACM Transactions on the Web*, 20(1):1–26.

Rolandos Alexandros Potamias, Georgios Siolas, and Andreas-Georgios Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32(23):17309–17320.

Trisha N Prabhu. 2019. Detecting messages with offensive content. US Patent 10,250,538.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

B. Singh and D. K. Sharma. 2023. A survey of sarcasm detection techniques in natural language processing. In *6th International Conference on Information Systems and Computer Networks (ISCON)*, pages 1–6. IEEE.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2158–2170.

Yuan Tian, Nan Xu, Ruike Zhang, and Wenji Mao. 2023. Dynamic routing transformer network for multimodal sarcasm detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2468–2480.

Mohit Tomar, Abhisek Tiwari, Tulika Saha, and Sriparna Saha. 2023. Your tone speaks louder than your face! modality order infused multi-modal sarcasm detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3926–3933.

C. Van Hee and 1 others. 2018. Automatic detection of cyberbullying in social media text. *PLOS ONE*, 13(10):e0203794.

Yang Wang, Pedro Giovanni Leon, Alessandro Acquisti, Lorrie Faith Cranor, Alain Forget, and Norman Sadeh. 2014. A field trial of privacy nudges for facebook. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2367–2376.

Yang Wang, Pedro Giovanni Leon, Kevin Scott, Xiaoxuan Chen, Alessandro Acquisti, and Lorrie Faith Cranor. 2013. Privacy nudges for social media: an exploratory facebook study. In *Proceedings of the 22nd international conference on world wide web*, pages 763–770.

A Statistical Validation of Emoji Signals

To evaluate the frequency of emoji usage in online discourse, data samples from platforms including X, Facebook, Instagram, YouTube, and Reddit were aggregated. Due to recent restrictions on proprietary data access, existing publicly available datasets on Kaggle spanning the years 2022 to 2025 were utilized. Datasets were selected strictly if they met four criteria: they measure user engagement, capture a global demographic, contain timestamps within the specified timeframe, and cover diverse topics such as news, sports, and COVID-19.

Dataset	Platform	Records	With Emojis
(Karunarathna, 2023)	YouTube	11,109	1,443
Global News Eng. (Karunarathna, 2024)	Facebook	2,734	437
#ChatGPT 1000 Daily (Domingo, 2023)	X	239,839	34,692
(Karunarathna, 2024)	X	257	187
(Karunarathna, 2024)	Facebook	1,996	1,926
(Karunarathna, 2023)	YouTube	50,183	10,713
(Karunarathna, 2024)	YouTube	100,000	20,981
Social Media SA (Parmar, 2024)	X, FB, Insta	352	11
(2024)	Instagram	170,335	56,925
Reddit top posts (Savi, 2025)	Reddit	100,997	11,916

Table 6: Collection of social media records for testing emoji usage.

The hypothesis tests were conducted for p_0 in 0.15, 0.16, 0.17, 0.18, 0.19, 0.20, 0.21, 0.22, and 0.23. To evaluate the prevalence of emojis, a conservative sample size of n = 3,889 was established (calculated at $p_0 = 0.23, = 0.05, 1 - \beta = 0.90, andE = 0.02$). This uniform sample size was applied to all subsequent tests. To ensure the robustness of the findings against sampling variance, the hypothesis test was repeated three times for each baseline proportion (p_0). For every individual test iteration, a new, independent random sample of n = 3,889 records was drawn directly from the aggregated corpus described in Table 6.

Based on the results presented in Table 7 it can be concluded that more than 20% of online posts contain emojis.

B Knowledge Distillation Configurations

Knowledge distillation was employed to compress the fine-tuned EmoCentricSarcBERT model into DistilBERT, TinyBERT, and MobileBERT architectures using an 80:20 split of the SarcOji dataset. A grid search optimized the softmax temperature ($T \in \{3, 4, 5, 6, 8\}$) and the loss balancing factor ($\alpha \in \{0.5, 0.6, 0.7, 0.8\}$). The classification performance of the resulting compressed models was then evaluated on SarcOjiTest1 (T1) and SarcOjiTest2 (T2).

C The RADMAD Framework Pipeline

The following execution pipeline is proposed for the RADMAD framework.

Initial state configuration: The system is initialized across two distinct environments. The user device initializes with a baseline Sarcasm score ($S = 0$), a Recovery score ($\zeta = 0$), an empty local cache, the top five most frequently utilized emojis, and the QuantizedEmoCentricSarcBERT model. The platform server initializes the SarcOji dataset, a configurable threshold (θ), and the base EmoCentricSarcBERT model.

Server-side maintenance: As new emojis are introduced or develop new conversational contexts, the SarcOji dataset is updated. Localized user texts (flagged as sarcastic, alongside non-sarcastic texts) are periodically anonymized (stripped of PII) and transmitted to the cloud to fine-tune the base model and mitigate data drift.

On-device execution: During message composition, the edge model evaluates the drafted text locally. If flagged, the system triggers a UI intervention. If the user ignores the nudge, S is incremented, utilizing the nudge penalty λ. If the user accepts the nudge and publishes non-sarcastic content, S decays by ζ. As S breaches warning θ_w or ban θ thresholds, platform-specific escalation mechanisms (warnings or bans) are enforced.

D Simulation Results

Zero compliance scenario: The most prominent outcome of this scenario is that many users, regardless of which model evaluated them, breached the absolute Ban Threshold. Because the users never accepted nudges, they absorbed a continuous stream of penalties without any recovery. The base EmoCentricSarcBERT consistently calculated a higher final sarcasm score than the 8-bit QuantizedEmoCentricSarcBERT.

Stochastic compliance scenario: Stochastic compliance significantly reduced Final Sarcasm Scores compared to the Zero Compliance baseline. Users with high Nudge Acceptance Likelihood maintained an equilibrium state with a sarcasm score of 0.0 because the recovery factor (ζ) successfully mitigated penalty accumulation.

Both the scenarios evaluate users' post across two year timeline, one on the unfiltered records and one on the records that contained emojis. The user trajectories are reported in Figures 3, 4, 5, and 6.

p0	Test Iteration	Sample $\hat{p}$	Z-statistic	P-value	Decision
0.15	1	0.220879	10.65514	8.25E-27	Reject H0
0.15	2	0.219337	10.44946	7.37E-26	Reject H0
0.15	3	0.216251	10.03561	5.32E-24	Reject H0
0.16	1	0.223451	9.499054	1.06E-21	Reject H0
0.16	2	0.214965	8.344088	3.59E-17	Reject H0
0.16	3	0.227822	10.08401	3.25E-24	Reject H0
0.17	1	0.223965	8.072362	3.45E-16	Reject H0
0.17	2	0.219851	7.506522	3.04E-14	Reject H0
0.17	3	0.229108	8.770949	8.86E-19	Reject H0
0.18	1	0.217537	5.673828	6.98E-09	Reject H0
0.18	2	0.223708	6.540721	3.06E-11	Reject H0
0.18	3	0.230908	7.533450	2.47E-14	Reject H0
0.19	1	0.220108	4.531747	2.92E-06	Reject H0
0.19	2	0.224736	5.189707	1.05E-07	Reject H0
0.19	3	0.217794	4.199361	1.34E-05	Reject H0
0.20	1	0.219337	2.914147	0.001783311	Reject H0
0.20	2	0.216251	2.461674	0.006914512	Reject H0
0.20	3	0.220879	3.138753	0.000848343	Reject H0
0.21	1	0.214965	0.753763	0.225495830	Fail to Reject H0
0.21	2	0.227822	2.649844	0.004026445	Reject H0
0.21	3	0.215222	0.792454	0.214048087	Fail to Reject H0
0.22	1	0.219851	-0.022460	0.508958384	Fail to Reject H0
0.22	2	0.229108	1.351490	0.088269248	Fail to Reject H0
0.22	3	0.223451	0.516604	0.302716161	Fail to Reject H0
0.23	1	0.223708	-0.941590	0.826798660	Fail to Reject H0
0.23	2	0.230908	0.134322	0.446573960	Fail to Reject H0
0.23	3	0.215480	-2.202390	0.986181010	Fail to Reject H0

Table 7: Hypothesis test results for proportion of emojis on sample frame of social media posts

Temp	Alpha	Val_MCC	Val_F1	T1_MCC	T2_MCC
2	0.2	0.7877	0.8698	0.2627	0.2242
2	0.5	0.7875	0.8703	0.2537	0.2734
2	0.8	0.7707	0.8586	0.2527	0.2505
3	0.8	0.7840	0.8657	0.2499	0.2684
4	0.8	0.7877	0.8698	0.2693	0.2208

Table 8: Grid search results for Knowledge Distillation (DistilBERT).

Temp	Alpha	Val_MCC	Val_F1	T1_MCC	T2_MCC
3	0.5	0.734674	0.832744	0.228394	0.234751
3	0.6	0.738857	0.842331	0.231642	0.233535
3	0.7	0.757283	0.848947	0.208325	0.190654
3	0.8	0.748389	0.842438	0.222234	0.172758
4	0.5	0.747453	0.845919	0.231265	0.252984
4	0.6	0.743368	0.841685	0.215392	0.192943
4	0.7	0.746570	0.842717	0.204757	0.207050
4	0.8	0.755154	0.852122	0.245219	0.224182
4	0.8	0.755938	0.848458	0.207931	0.160839
5	0.6	0.742059	0.843561	0.234232	0.172582
5	0.7	0.742553	0.842198	0.226970	0.205707
5	0.8	0.746186	0.846452	0.215092	0.234190
6	0.5	0.752529	0.848218	0.252316	0.240606
6	0.8	0.743327	0.843573	0.241876	0.218307
8	0.5	0.744000	0.843941	0.255063	0.200836
8	0.8	0.740511	0.844369	0.246524	0.243014

Table 9: Grid search results for Knowledge Distillation (TinyBERT)

Temp	Alpha	Val_MCC	Val_F1	T1_MCC	T2_MCC
4	0.8	0.755154	0.852122	0.245219	0.224182
6	0.5	0.752529	0.848218	0.252316	0.240606
4	0.5	0.747453	0.845919	0.231265	0.252984
8	0.5	0.744	0.843941	0.255063	0.200836
6	0.8	0.743327	0.843573	0.241876	0.218307
8	0.8	0.740511	0.844369	0.246524	0.243014

Table 10: Grid search results for knowledge distillation process (MobileBERT)

Figure 3: Zero compliance scenario for unfiltered posts for 50 users.

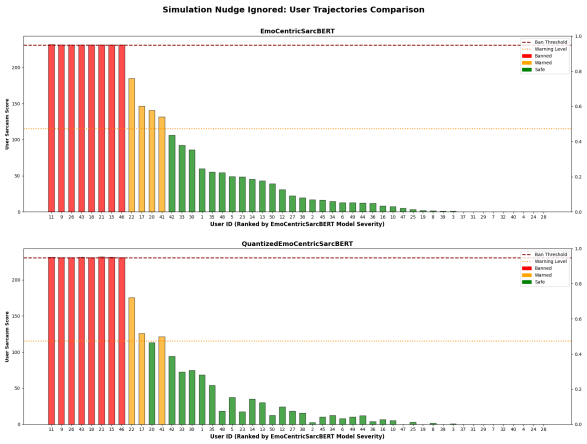


Figure 4: Zero compliance scenario for posts strictly with emojis for 50 users.

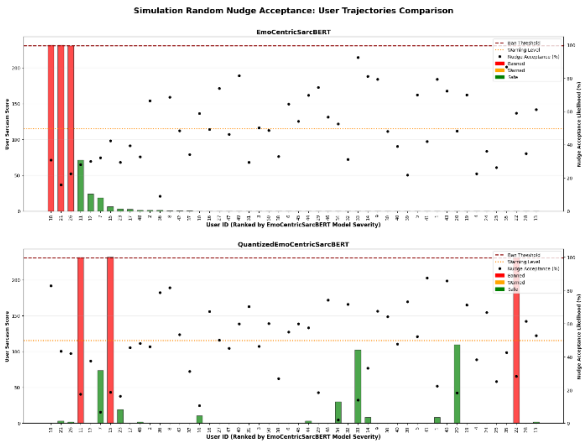


Figure 5: Dynamic compliance scenario for unfiltered posts for 50 users.

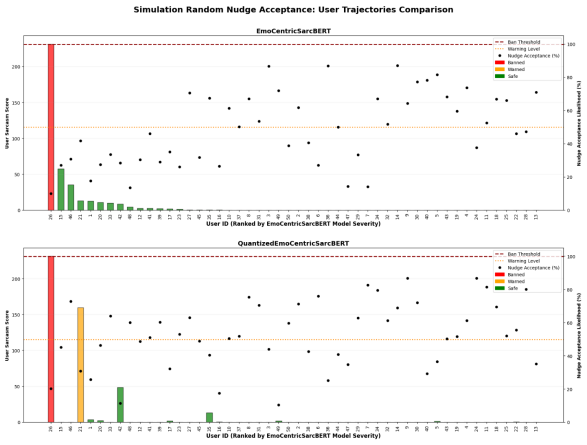


Figure 6: Stochastic compliance scenario for posts strictly with emojis for 50 users.

Model	Size (MB)	SarcOji Test 1			Infer time	SarcOji Test 2			Infer time
		F1	MCC	RoC_AuC	CPU (s)	F1	MCC	RoC_AuC	CPU (s)
EmoCentricSarcBERT	417.21	65.50	0.33	71.30	988.6999	53.55	0.35	74.18	581.1884
DistilledEmoCentricSarcBERT	267.00	56.77	0.25	65.48	413.6220	53.73	0.36	69.58	
QuantizedEmoCentricSarcBERT	106.39	63.01	0.30	70.97	425.0536	53.39	0.35	74.24	215.9556
MobileEmoCentricSarcBERT	99.66	59.31	0.25	68.34	450.2944	45.03	0.24	65.88	232.9474
TinyEmoCentricBERT	71.49	60.87	0.25	67.82	81.7695	45.57	0.25	68.84	50.1441

Table 11: Compressed models comparison against the EmoCentricSarcBERT baseline.