Mapping the Landscape of Unregulated eXplicit Contents on Reddit

MSVPJ Sathvik[1,*], **Manan Roy Choudhury**[2,*], **Rishita Agarwal**[3], **Sathwik Narkedimilli**[4], **Thao Ha**[2], **Liesel Sharabi**[2], and **Vivek Gupta**[2,†]
[1]University of Birmingham, [2]Arizona State University
[3]Indian Institute of Technology Guwahati [4]National University of Singapore
[*]Equal contribution (co-first authors) [2,†]Corresponding Author
msvpjsathvik@gmail.com, mroycho1@asu.edu, rishita@iitg.ac.in,
sathwik.narkedimilli@ieee.org, thaoha@asu.edu, liesel.sharabi@asu.edu
[2,†]**Correspondence:** vgupt140@asu.edu | **Project Website:** Link

Abstract

The rise of online platforms has facilitated the dissemination of explicit content, posing significant challenges for detection and regulation. Often using coded language to bypass moderation, this content erodes user trust and may be associated with scam-related risks, posing direct financial and personal risks. In this study, we map the landscape of online explicit content posts, focusing on their categorization, linguistic strategies, and temporal and behavioral patterns as they appear within the mainstream platform Reddit. We investigated five distinct content categories including Virtual Services (VS), Physical Services (PS), Exhibitionism (Ex), Couples and Group Interactions (CGI), and Content Creation and Sales (CCS) and performed large-scale experimentation using state-of-the-art large language models (LLMs) such as GPT-4, LLaMA 3.3-70B-Instruct, Gemini 1.5 Flash, Mistral 8×7B, Qwen 2.5 Turbo, and Claude 3.5 Haiku. Our work demonstrates that a nuanced classification of these services requires moving beyond simple keywords, and we establish that expressive signals, such as sentiment, emotion, and tone, are critical for accurate detection. Our analysis reveals distinct behavioral and psychosocial expression patterns for each service category, providing a robust framework for future moderation.

1 Introduction

Modern social media platforms are vast, multipurpose ecosystems that enable unprecedented self-expression and community formation. However, this openness also creates vulnerabilities that facilitate the proliferation of unregulated explicit-service solicitations, posing systemic challenges to online safety (Marche et al., 2023; Raponi et al., 2022). While such solicitations may appear across diverse communities, identifying and studying them requires reliable ground-truth examples. To establish a foundational benchmark, we analyze explicit-service posts within dedicated Reddit communities where solicitation behaviors are observable and consistently labeled (Baumgartner et al., 2020; Gothard, 2021). These posts frequently employ coded language, euphemisms, and suggestive signals designed to bypass moderation, and are sometimes associated with scams and fraudulent schemes that pose financial and personal risks. This controlled setting enables systematic analysis of linguistic, behavioral, and engagement patterns that can inform future detection efforts in broader, heterogeneous social media environments (Trager et al., 2022; Teitelbaum, 2020).

The scope of these solicitations spans multiple modalities, including Virtual Services (VS), such as live video calls; Physical Services (PS), involving in-person transactions; Exhibitionism (Ex); Couples and Group Interactions (CGI); and Content Creation and Sales (CCS). These categories reflect the adaptability of actors promoting illicit services and expose the limitations of moderation strategies that rely solely on surface-level keyword filtering rather than contextual and behavioral indicators.

The risks associated with such content are multilayered. Minors face heightened exposure and vulnerability to grooming and exploitation, while adult users may encounter scams, financial loss, and erosion of trust in digital spaces(Amirkhani et al., 2026; Sanchez and Genelza, 2025; Gupta, 2025). At the governance level, these solicitations strain moderation infrastructure, often necessitating hybrid human–AI approaches to manage scale and contextual ambiguity (Peter and Valkenburg, 2016; Ferguson and Hartley, 2022; Mitchell et al., 2003). Furthermore, the issue intersects with complex legal and regulatory frameworks (Government of NCT of Delhi, 2023). Many jurisdictions criminalize aspects of digital solicitation, while others regulate certain forms of sex work under strict licensing regimes (Government of India, 2000; Ministry of Electronics and Information Technology, Gov-

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 271–292

ernment of India, 2021). This global patchwork underscores the need for scalable, context-aware moderation mechanisms that can adapt across legal, cultural, and linguistic contexts.

In response and also seeing recent efforts to stress-test LLMs through domain-specific benchmarks (Choudhury et al., 2025, 2026), this paper introduces REDDIX-NET, a structured benchmark dataset designed to map and analyze explicit-service solicitations on a mainstream platform. We do not frame this work as a direct detection task in mixed-content environments; rather, we provide a systematic behavioral and linguistic characterization that serves as a precursor for developing detection systems in broader settings. Beyond categorization, we demonstrate that expressive signals, i.e., sentiment, emotion, and tone, play a meaningful role in classification and in understanding engagement dynamics across service types, thereby establishing a more context-aware moderation benchmark.

In this paper, we present the analysis, experimentation, and dataset construction underlying REDDIX-NET. By leveraging multilingual data and evolving online trends, the dataset supports AI-driven categorization of distinct solicitation modalities that frequently operate outside regulated channels and platform policies.

The contributions of this work are as follows:

- We provide the first large-scale analysis of solicited explicit content within service-oriented communities on a mainstream social media platform.
- We introduce REDDIX-NET, a curated benchmark dataset that fills a critical resource gap for moderation and safety research.
- We analyze sentiment and user expression patterns to derive psychosocial engagement indicators that extend beyond surface-level content filtering.

2 Related Work

Sex Trafficking and Escort Advertisement Detection. Prior work on online sexual services primarily targets sex trafficking networks and illicit escort advertisements on dedicated platforms. (Ibanez and Suthers, 2016b,a) used network and content analysis to identify trafficking indicators from structured metadata such as phone numbers and social links. Keskin et al. (2021) analyzed large-scale escort ads to model mobility and service circuits, while Giommoni and Ikwu (2021) extracted trafficking indicators from UK-based advertisements. These approaches focus on overt ads and structured signals rather than conversational content.

Supervised Classification and Ordinal Risk Modeling. Machine learning methods have also been applied to classify illicit businesses and trafficking-related advertisements. Diaz and Panangadan (2020) trained supervised models on review-site data to distinguish legitimate from illicit services, and Wang et al. (2020a,b) proposed Ordinal Regression Neural Networks for predicting trafficking likelihood scores from ad text. These formulations emphasize binary or ordinal risk detection over explicit advertisements, not nuanced solicitation behavior in social media contexts.

NSFW Filtering and Content Moderation. Traditional NSFW detection relies on keyword filtering and image-based recognition (Davidson et al., 2017; Founta et al., 2018; Vidgen et al., 2021; Li et al., 2026; Bao et al., 2025; Jangra et al., 2025; Yang et al., 2025). While effective for overt content, such systems struggle with coded language, euphemisms, and contextual ambiguity, and typically treat detection as surface-level classification rather than modeling expressive linguistic and interaction dynamics.

User Behavior and Engagement in Sexual Content Communities. Studies on online sexual content communities indicate that engagement patterns differ from general social media, particularly in anonymous environments like Reddit. Users often engage in sensitive self-disclosure and receive varied responses such as support and validation (Andalibi et al., 2018), while also seeking sexual advice through intent-driven interactions (PettyJohn et al., 2025). Such disclosures tend to attract higher engagement and repeated participation (Haq et al., 2025). Anonymity facilitates open expression (Shelton, 2015), and community norms shape interaction behavior (Brown, 2018). Overall, interactions combine social, informational, and transactional elements, where comments may include requests, verification, or negotiation, offering insights into user intent beyond the original post.

Our Contribution. In contrast, we examine explicit-service solicitations on a mainstream platform (Reddit), specifically within dedicated service-oriented communities that provide high-

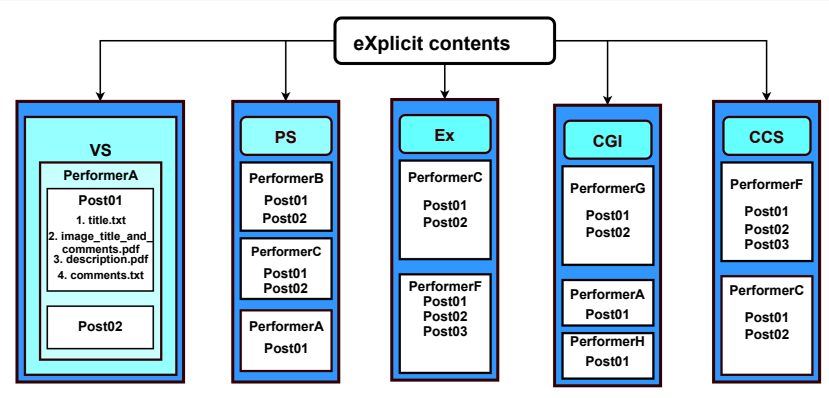


Figure 1: Structure of the proposed dataset, categorizing online sexual services across five distinct categories and their further subdivisions.

confidence ground-truth examples of solicitation behavior. This controlled setting enables systematic analysis of linguistic, behavioral, and engagement patterns that can inform future detection systems in broader environments. Methodologically, we move beyond surface lexical cues by leveraging expressive signals, i.e., sentiment, emotion, and tone, benchmarking state-of-the-art LLMs for context-aware classification. We further analyze comment-level interactions and temporal dynamics, offering a computational social science perspective on solicitation behavior within explicit-service communities on a mainstream platform.

3 REDDIX-NET CONSTRUCTION

This section outlines the methodology and sources used to build REDDIX-NET, providing a comprehensive, data-driven foundation. It highlights the strategic approach taken to construct a robust and insightful dataset.

3.1 Data Collection

The dataset was collected from three large subreddit channels (each with over 50K members) focused on online sexual services with users from all over the world. Although these communities explicitly focus on adult services, they exist within a mainstream social media platform and retain conversational, user-generated characteristics distinct from those of structured advertising marketplaces. This allows us to analyze solicitation behavior in a realistic social interaction context while maintaining high labeling confidence. Our dataset includes posts from 2014 to 2023. Due to privacy policies, their names are withheld. Using the Reddit API with PRAW API, we gathered posts offering services like paid meetups, nude video calls, and couple swaps. More details on data labeling, collection, cleaning, and pre-processing are provided in Appendix 4.

3.2 Data Annotation

For the experiments, three annotators, including the authors, manually reviewed and categorized the posts. The team was selected to mitigate gender and regional bias, comprising two male and one female annotator from three different states in India, bringing diverse linguistic and socio-cultural perspectives across English and several Indian languages. All annotators possessed domain familiarity and contextual sensitivity to handle euphemistic or implicit references to adult services. The annotation process was conducted carefully to ensure reliable and balanced classification.

The posts were divided into the following five categories:

1. **Content Creation and Sales (CCS)**: This category includes services related to the production and sale of adult content, such as videos, photos, or written material, often tailored to specific user requests.

2. **Couples and Group Interactions (CGI)**: Services in this class involve collaborative engagements, typically between two or more individuals, either for personal interaction.

3. **Exhibitionism (Ex)**: This category refers to services where individuals perform live or recorded acts for an audience, often emphasizing the act of showcasing themselves in a sexual or provocative context.

4. **Physical Services (PS)**: Services involving physical interaction, such as in-person meetings, escort services, or any physical contact-based activities, are classified here.

5. **Virtual Services (VS)**: This category includes services provided online, such as video chats, private messages, or virtual performances, which do not involve any physical meetings but are sexual or adult-oriented in nature.

The five service categories were derived through iterative qualitative coding grounded in prior online solicitation and digital sexual commerce studies. Two annotators independently reviewed sampled Reddit posts, consolidated recurring behavioral codes, and refined five solicitation modalities. Pilot annotation validated separability, minimized overlap, and ensured contextual consistency. Figure 1 presents the five-class folder structure reviewed.

3.3 Data annotation guidelines

Posts were annotated into five behavioral categories:

Virtual Services (VS), Physical Services (PS), Exhibitionism (Ex), Couples and Group Interactions (CGI), Content Creation and Sales (CCS).

Annotators reviewed the post title, body, any associated (anonymized) media, and user comments. 1) Guidelines included decision heuristics and examples to disambiguate cases (e.g., distinguishing VS from CCS based on service type and platform mention). 2) Posts were labeled via majority vote from three annotators. Uncategorizable posts were removed. 3) Inter-annotator agreement was measured using Krippendorff's alpha, Cohen's kappa, and Fleiss' kappa. 4) Ethical training and precautions were implemented for the annotator's well-being.

3.4 Statistical Analysis

Table. 1 provides an overview of posts. The dataset comprises 8,146 posts across five categories, with Ex containing the most posts (2,302) and CGI the fewest (1,103), indicating a moderate imbalance in class representation. The overall word count across the dataset is 557,764, with Ex again contributing the most words (164,131) and CGI the fewest (73,751). Despite these variations in volume, the average word length per post remains relatively consistent, ranging between 65 and 75 words across categories. Engagement levels, measured through comment activity, follow a similar pattern: the dataset contains 60,240 comments in total, with Ex receiving the largest share (17,923), although the average number of comments per post remains stable across all categories (7.3–8.0), suggesting uniform interaction behavior. Lexical density, reflected in the average tokens per post, ranges from 63 in CCS to 75 in Ex, indicating only minor variation in text granularity.

3.5 Inter-annotator Agreement Score

To assess annotation consistency, we conducted an Inter-Annotator Agreement (IAA) analysis using Krippendorff's Alpha (K), Cohen's Kappa (C) for pairwise comparisons, and Fleiss' Kappa (F) for group agreement across the full dataset. Pairwise Krippendorff scores were $K(1,2) = 0.6633$, $K(1,3) = 0.7470$, and $K(2,3) = 0.6783$, while the overall agreement among all three annotators was $K(1,2,3) = 0.6963$ (Table. 2). These results indicate substantial inter-annotator reliability and consistent labeling across annotators.

4 Analysis of REDDIX-NET

This section outlines key experiments on REDDIX-NET that use LLMs and PLMs for user classification, sentiment analysis, comment classification, and metadata-temporal analysis. Details follow in subsequent subsections.

4.1 REDDIX-NET User Classification

This experiment aims to identify users offering specific services based on their posts using LLMs. Users often employ coded language and suggestive presentation to evade moderation. While our analysis focuses primarily on textual signals (titles, descriptions, and comments), the dataset retains contextual references to associated media, enabling future multimodal extensions. Users' services are treated as ground truth, and LLMs are prompted to classify posts into predefined service categories.

Posts that remained uncategorized were embedded using Sentence-BERT and clustered via K-means (k = 5, aligned with the five service categories). Cluster centroids were manually reviewed and mapped to the closest category based on dominant semantic patterns. We evaluated multiple state-of-the-art LLMs for this task, including GPT-4 (Wiggers, 2022), LLaMA 3.3-70B-Instruct (Touvron et al., 2023), Gemini 1.5 Flash (Google, 2023), Mistral 8×7B (Jiang et al., 2023), and Claude Haiku.

Metric	VS	PS	Ex	CGI	CCS	Overall
No. of Posts per Category	1588	1501	2302	1103	1547	8146
No. of Words per Category	103844	102917	164131	73751	102490	557764
No. of Comments per Category	11373	10984	17923	7776	11176	60240
Avg No. of Comments per Post per Category	7.5	7.6	8	7.3	7.5	-
Avg No. of Tokens per Post per Category	72	74	75	69	63	-
Avg No. of Posts per Performer per Category	150	162	120	97	175	-
No. of Performers per Category	11	10	19	12	9	-

Table 1: Statistical summary table of the REDDIX-NET dataset, detailing post counts, word counts, comment volumes, and average engagement metrics across the five defined service categories (VS, PS, Ex, CGI, CCS)

Annotator	Krippendorff(K)	Cohen(C)	Fleiss(F)
(1,2)	0.6633	0.554	—
(1,3)	0.7470	0.681	—
(2,3)	0.6783	0.740	—
(1,2,3)	0.6963	—	0.608

Table 2: Inter-Annotator Agreement Scores

4.2 REDDIX-NET Expression Analysis

This analysis examines user responses to posts to understand how engagement patterns reflect emotional and psychosocial reactions rather than direct mental health outcomes.

We performed sentiment analysis on REDDIX-NET using both pre-trained and GoEmotions-fine-tuned (Demszky et al., 2020) BERT-based models (PLMs) and LLMs (Qwen 2.5 Turbo, GPT-4o). The models predict fine-grained emotion labels, which are aggregated into higher-level psychosocial categories (e.g., sadness/anxiety → mental health concern; joy/trust → positive experience; anger/disgust → exploitation indicators). LLMs extract sentiment polarity (positive, neutral, negative, mixed), emotional spectrum, tonal variation (casual, formal, playful, aggressive), confidence scores, and key phrases. A stratified sample of LLM-generated sentiment outputs was manually reviewed to verify consistency with the intended polarity and emotion labels.

GoEmotions fine-tuned BERT models produced probability distributions across emotion categories, later aggregated into broader psychosocial indicators using predefined affective computing mapping rules. Emotional spectrum, confidence scores, keyphrases, and tonal variation were derived through weighted emotion analysis, normalized classifier certainty, transformer-based attention ranking, and stratified manual validation, with all features stored as aggregated statistics from direct Reddit post responses. The fine-tuned BERT model further evaluates emotional dependency, varied emotional states, exploitation signals, user experience, and mental health-related expressions. Using Hugging Face transformers, BERT maps star ratings (e.g., "5 stars" → satisfaction; "1 star" → aggression) to discrete emotions via a custom dual-mapping framework.

4.3 REDDIX-NET Comments Classification

While sentiment analysis captures emotional tone, it fails to represent interaction intent. Therefore, comments were categorized into 19 thematic buckets (Table 6) derived through an iterative qualitative analysis of representative samples. The taxonomy captures recurring intents, including purchase intent, negotiation, verification, harassment, emotional dependency, and authenticity concerns, while remaining moderation-oriented for analysis of explicit service interactions.

Classification Methodology. Comments were classified using GPT-4 with a few-shot prompting setup (Appendix). Each comment was processed individually, and a multi-label scheme was adopted to account for overlapping intents. Comments not confidently assigned were embedded and clustered, then re-evaluated using the same few-shot prompt to map clusters to predefined buckets.

Handling Ambiguity and Validation. Ambiguous comments received all semantically relevant labels. A stratified random sample validated labeling consistency and taxonomy reliability. Nineteen interaction buckets were derived through grounded, qualitative coding by three annotators, who iteratively merged overlapping themes. Validation included manual review and inter-annotator agreement analysis. The final dataset features store normalized proportions of aggregated comment-label distributions.

4.4 Time-based Analysis on REDDIX-NET

We analyzed metadata to examine temporal engagement trends, tracking fluctuations in posts and comments to identify peak and low activity periods. Hour-of-day analysis revealed engagement cycles

in sexual service-related discussions, offering insights into user behavior, participation patterns, and content visibility dynamics.

5 Results and Analysis

In this section, we will discuss all the results that we have obtained from the experiments that we mentioned in the previous section.

5.1 REDDIX-NET User Classification

Table. 3 summarizes the performance of multiple large language models for classifying posts into predefined service categories using precision, F1-score, accuracy, and error-based metrics including MSE, MAE, and JSD. Performance varies considerably across categories, with Virtual Services (VS) and Physical Services (PS) generally achieving higher F1-scores than Exhibitionism (Ex) and Couples and Group Interactions (CGI), indicating stronger textual separability. GPT-4 demonstrates relatively stable performance across categories, whereas Claude 3.5-Haiku attains the highest F1-score in VS, and LLaMA-3.3-70B-Instruct performs competitively in CCS classifications overall.

The accuracy values appear comparatively high because classification was performed on balanced confidence-solicitation samples from dedicated communities rather than in realistic mixed-content environments. Therefore, F1-scores provide a more reliable estimate of category separability under contextual ambiguity. Preliminary ablation trends, influenced by lexical leakage from repeated promotional phrases, were excluded from the final evaluation. For methodological clarity, exploratory clustering analyses should be separated from supervised classification results. Furthermore, reporting macro-averaged precision, recall, and confusion matrices would improve interpretability and better characterize misclassification behavior.

Cat.	Pre (%)	F1 (%)	MSE	MAE	JSD	Acc (%)
GPT-4						
VS	**45.1**	**62.0**	**0.05**	**0.14**	0.44	85.7
PS	39.3	56.1	0.07	0.18	0.56	81.8
Ex	25.0	39.6	0.06	**0.14**	0.49	**85.8**
CGI	28.6	43.2	0.13	0.20	**0.70**	79.5
CCS	33.3	49.6	**0.05**	0.15	0.48	84.8
Llama-3.3-70B-Instruct						
VS	49.0	64.0	0.05	**0.14**	0.51	84.7
PS	**50.1**	62.2	0.05	0.15	0.51	**85.2**
Ex	43.0	58.0	0.11	0.23	**0.71**	77.4
CGI	42.0	54.0	0.11	0.20	0.67	79.9
CCS	49.0	**67.0**	**0.04**	0.15	0.46	85.1
Mistral 8×7B						
VS	42.0	56.0	**0.03**	**0.12**	0.42	**87.9**
PS	**47.0**	**61.0**	0.06	0.18	0.64	82.2
Ex	40.0	55.0	0.10	0.22	**0.70**	78.4
CGI	41.0	56.0	0.08	0.19	0.63	81.1
CCS	44.0	58.0	0.04	0.13	0.43	86.9
Gemini 1.5 Flash						
VS	**48.0**	63.2	0.05	**0.14**	0.47	85.9
PS	47.5	**64.3**	**0.04**	0.13	0.48	**87.2**
Ex	32.5	48.2	0.06	0.17	0.61	82.7
CGI	31.0	46.2	0.11	0.20	**0.71**	80.1
CCS	39.5	56.1	0.05	0.16	0.48	84.3
Claude 3.5-Haiku						
VS	**52.0**	**66.0**	0.06	0.16	0.54	84.1
PS	46.3	56.9	0.05	**0.13**	0.46	**86.6**
Ex	35.8	51.8	0.08	0.19	0.63	81.0
CGI	36.7	52.4	0.13	0.21	**0.68**	79.3
CCS	37.8	52.4	**0.04**	0.15	0.46	85.5

Table 3: Comparative evaluation of different large language models (LLMs) across various service categories (VS, PS, Ex, CGI, CCS). The models are assessed based on multiple performance metrics, including Precision (Pre), F1-score, Mean Squared Error (MSE), Mean Absolute Error (MAE), Jensen-Shannon Divergence (JSD), and Accuracy. The category is abbreviated as Cat.

5.2 REDDIX-NET Expression Analysis

We analyze user expressions using sentiment, emotion, and tone, employing both LLM- and PLM-based approaches.

Sentiment Analysis: As shown in Figure 2, positive sentiment constitutes the largest proportion across all categories (approximately 40–50%), followed by neutral sentiment (around 30%). Negative and mixed sentiments appear less frequently. This distribution reflects the predominance of positive or neutral expressions in the dataset's user comments. Variations across categories are evident; for example, VS and PS exhibit slightly higher proportions of mixed or negative sentiment than CCS. These observations describe sentiment patterns within REDDIX-NET and do not necessarily generalize beyond this context.

Emotion Type Analysis: Table 4 presents the distribution of emotion categories derived from user comments. Emotions such as desire and joy appear frequently across multiple categories, particularly in VS and Ex. PS shows comparatively higher proportions of emotions such as lust and disgust, while CGI and CCS exhibit more balanced distributions, including neutral expressions. These results reflect model-inferred emotional signals and provide an approximate characterization of user responses within the dataset.

Tonality Analysis: Table 5 illustrates the distribution of tonal expressions. The casual tone is most prevalent across all categories, especially in VS and Ex. PS shows relatively higher propor-

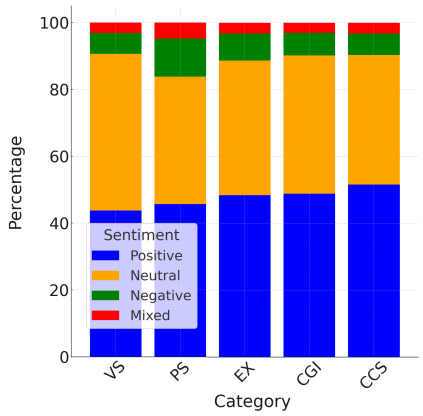


Figure 2: Distribution of sentiment classifications across five different service categories (VS, PS, Ex, CGI, CCS).

Emo(%)	VS	PS	EX	CGI	CCS
Des	31.8	25.3	32.6	27.6	31.3
Joy	31.8	11.4	27.2	22.7	23.7
None	7.8	0.3	11.0	15.1	10.7
Int	9.2	1.5	9.8	10.1	9.2
Sur	5.6	0.1	6.0	6.5	6.9
Ant	4.4	10.1	4.3	5.4	4.0
Lus	2.5	16.4	3.3	2.1	4.6
Ang	1.8	6.3	1.4	3.2	2.9
Exc	1.4	9.5	2.2	2.5	2.1
Ind	1.1	0.6	0.8	1.7	0.9
Pla	0.8	1.4	0.1	1.0	0.8
Dis	0.7	12.6	0.5	0.8	1.4
Inf	0.1	0.1	0.3	1.1	1.0
Frus	1.0	4.4	0.5	0.2	0.5

Table 4: Detailed breakdown percentages of top 14 emotions (Emo) types across five service categories (VS, PS, EX, CGI, CCS). The emotions include Desire (Des), Joy, Interest (Int), Surprise (Sur), Anticipation (Ant), Lust (Lus), Anger (Ang), Excitement (Exc), Indifference (Ind), Playfulness (Pla), Disgust (Dis), Informational (Inf), and Frustration (Frus), along with instances labeled as None.

tions of erotic and playful tones compared to other categories. Formal and neutral tones are consistently present across categories, while aggressive and explicitly sexual tones occur less frequently. These patterns highlight differences in linguistic style across service types.

Cross-Correlation Analysis: Correlation analysis (Figure 6 in Appendix) indicates that relationships between sentiment and tone are generally weak or inconsistent. In contrast, certain emotions align more closely with specific tones (e.g., anger with an aggressive tone, joy with a playful tone). Category-level variations are also observed, with PS showing weaker correlations than other categories. These findings describe associations within the dataset rather than causal relationships.

Tone(%)	VS	PS	EX	CGI	CCS
Cas	62.3	28.6	61.8	61.0	62.8
For	13.8	8.9	14.9	16.3	13.3
Inf	6.9	8.9	6.4	6.3	5.5
Neu	6.2	10.3	7.7	8.1	8.1
Play	5.5	13.6	4.3	3.3	5.2
Agg	1.7	4.2	1.7	2.4	2.0
Ero	1.4	14.5	1.3	1.2	0.9
Flir	1.2	9.1	1.1	0.8	1.5
Se	1.0	1.9	0.9	0.7	0.7

Table 5: Illustration of the distribution of top nine tonal expressions across five service categories. The tones include Casual (Cas), Formal (For), Informal (Inf), Neutral (Neu), Playful (Play), Aggressive (Agg), Erotic (Ero), Flirtatious (Flir), and Sexual (Se).

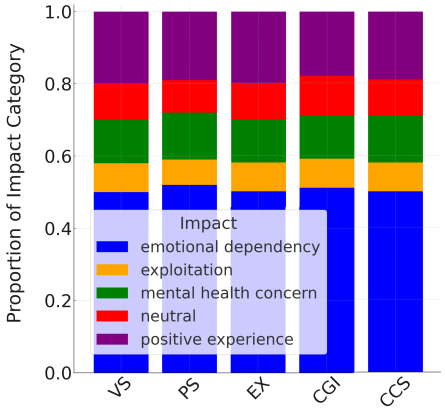


Figure 3: Category-wise impact proportion highlighting the proportion of emotional dependency, exploitation, mental health concerns, neutral perceptions, and positive experiences across different online prostitution categories. This provides insights into the psychosocial expression patterns associated with various engagement types.

5.3 REDDIX-NET Comments Classification

The baskets/categories are defined as: b_{t_1}: Payment or delivery complaints. b_{t_2}: Fantasy and violent demands. b_{t_3}: Legal and ethical concerns. b_{t_4}: Competition or self-promotion. b_{t_5}: Emotional support requests. b_{t_6}: Unclassified comments. b_{t_7}: Price or service negotiations. b_{t_8}: Verification and identity inquiries. b_{t_9}: Specific content requests. $b_{t_{10}}$: External link sharing. $b_{t_{11}}$: Unsolicited requests or harassment. $b_{t_{12}}$: Reviews and recommendations. $b_{t_{13}}$: Service demands (intent to purchase/engage). $b_{t_{14}}$: Skepticism or authenticity questions. $b_{t_{15}}$: Multilingual comments. $b_{t_{16}}$: Positive engagement

Bucket(%)	VS	PS	Ex	CGI	CCS
b_{t_1}	0.4	0.5	0.3	0.7	0.7
b_{t_2}	0.4	0.5	0.9	1.6	1.1
b_{t_3}	0.5	0.9	1.7	1.7	1.8
b_{t_4}	0.7	1.1	0.4	1.9	1.7
b_{t_5}	1.3	1.1	4.5	3.8	3.6
b_{t_6}	1.3	1.8	1.7	0.9	0.7
b_{t_7}	0.5	0.8	0.7	1.9	0.9
b_{t_8}	1.7	2.7	2.8	4.5	6.7
b_{t_9}	12.3	11.9	6.1	3.9	4.5
$b_{t_{10}}$	0.7	3.2	5.0	4.5	4.5
$b_{t_{11}}$	2.7	5.5	4.5	3.2	3.1
$b_{t_{12}}$	5.3	4.6	3.3	6.4	3.6
$b_{t_{13}}$	7.9	3.6	21.2	16.7	20.6
$b_{t_{14}}$	6.2	6.4	7.8	6.4	5.4
$b_{t_{15}}$	1.3	2.0	2.2	9.7	5.4
$b_{t_{16}}$	24.6	27.3	13.9	12.8	14.3
$b_{t_{17}}$	2.6	2.3	3.4	5.2	4.5
$b_{t_{18}}$	19.3	19.3	10.6	7.1	8.1
$b_{t_{19}}$	10.1	4.6	8.9	7.1	8.9

Table 6: This table presents the comments classification of the posts of the different users. Each of the 19 bucket types $b_{t_i} \in \{1, 2, \ldots, 19\}$ captures a distinct user comment or behavior.

(enjoying the post). $b_{t_{17}}$: Self-assertive or confident expressions. $b_{t_{18}}$: Sexual propositions or explicit requests. $b_{t_{19}}$: Ambiguous or multi-response comments.

Table 6 summarizes comment category distributions across service types using a multi-label classification scheme, where percentages are reported independently for each bucket. Positive engagement ($b_{t_{16}}$) is more common in VS and PS, while service demand ($b_{t_{13}}$) is more prominent in Ex, CGI, and CCS. Content requests (b_{t_9}) and explicit propositions ($b_{t_{18}}$) occur more frequently in VS and PS. Lower-frequency categories (e.g., b_{t_1}–b_{t_4}) remain rare across all service types. Overall, the results provide a descriptive overview of interaction patterns and labeled comment behaviors rather than mutually exclusive user intents.

5.4 REDDIX-NET Temporal Analysis

Temporal activity patterns are illustrated in Figures 4 and 5, which show post and comment distributions across hourly intervals. Engagement levels increase from early hours, peak during mid-day to evening periods (approximately 12–19 hours), and decline thereafter. A relatively high comment-to-post ratio is observed during peak periods, indicating increased interaction levels. Since timestamps are aggregated without user location normalization, these patterns represent global activity trends within the dataset rather than region-specific behaviors. Overall, the temporal analysis highlights recurring activity cycles in REDDIX-NET and provides insights into when higher interaction volumes occur within the observed data.

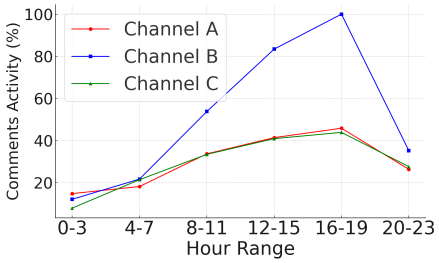


Figure 4: Illustration of comment activity trends over different hourly ranges in a day, highlighting peak engagement times for Channels A, B, and C. The x-axis categorizes days into six time periods, while the y-axis measures the proportion of total posts in each window.

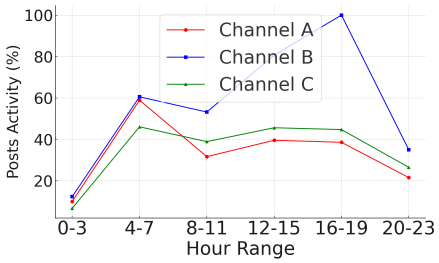


Figure 5: Visualization of temporal posts activity patterns across distinct hourly intervals, emphasizing peak engagement periods for Channels A, B, and C.

6 Conclusion

This study introduces REDDIX-NET, a benchmark dataset for mapping and moderating solicited explicit content on platforms such as Reddit. Using state-of-the-art LLMs, we classify interactions into five behavioral classes through scalable AI-driven moderation. Findings highlight challenges, including evasion tactics and contextual complexity, demonstrating that accurate classification depends on both post content and behavioral patterns derived from sentiment and comment analysis.

Our framework does not make legal judgments, as service legality is jurisdiction-dependent. Instead, it identifies and categorizes unregulated solicitation on mainstream platforms that often violates platform policies and bypasses regulatory oversight. The resulting behavioral classifications, including PS, VS, and CCS, are intended to support moderation and policy enforcement rather than provide definitive legal assessment.

Limitations

This study, while offering a novel dataset and benchmark for online sexual service moderation, is subject to certain limitations. Contextual ambiguity in online discussions makes classification difficult, even for advanced LLMs. While we employed diverse annotation strategies, human bias may still affect labeling. The dataset is Reddit-specific, limiting generalizability to other platforms. Also, the dataset primarily contains explicit-service posts, so model performance on this benchmark should not be interpreted as indicative of real-world moderation accuracy in mixed-content environments where explicit posts represent a small minority. Additionally, evolving evasion tactics pose ongoing challenges for AI moderation, requiring frequent updates.

Future research should focus on several key directions. These include enhancing sentiment analysis with more deterministic methods, integrating AI systems with human-in-the-loop approaches to improve contextual understanding, and further investigating how emotionally charged engagement patterns may relate to psychosocial indicators in online discourse. Our study also offers several important real-life insights, which are discussed in detail in Appendix-6. Crucially, all these analyses done in this paper are confined to data sourced from specific Reddit channels and do not purport to replicate real-world conditions. The findings reflect only the content of this dataset, and interpretations are subject to its inherent limitations and may vary across regional and individual perspectives. Another future work will be to extend this benchmark by incorporating mixed-content datasets with negative examples to evaluate detection performance in real-world moderation scenarios.

Ethics Statement

The ethical dimensions of research concerning online sexual services are multifaceted and deeply sensitive, demanding rigorous safeguards and deliberate ethical oversight. Our study seeks to enhance online safety and inform content moderation strategies while remaining acutely aware of the potential for misuse, exploitation, or harm. To uphold the highest ethical standards, we established a comprehensive framework that emphasizes participant privacy, robust data security, and principled use throughout the research lifecycle.

To promote transparency and responsible engagement, we will release a detailed datasheet alongside the REDDIX-NET dataset. This datasheet outlines the dataset's structure, data collection pipeline, annotation methodology, and usage limitations. We explicitly state that REDDIX-NET is intended strictly for benchmarking and not for model training or any application that could facilitate harm or exploitation.

We prioritized annotator safety and well-being by collaborating with a technical, community-driven organization to recruit and manage our annotation team (One of the authors is a part of this organization). This organization (due to anonymity considerations, we are unable to disclose the organization's name. However, upon request, all relevant details will be shared with authorized personnel) independently oversaw ethical review procedures and secured formal IRB approval for the study. One female annotator was intentionally included in the process to ensure sensitivity toward gender dynamics and to mitigate implicit biases in the annotation of content related to online sexual services. All annotators were clearly informed of the emotionally sensitive nature of the data and were provided with mental health resources and protocols for emotional self-care. Annotators were encouraged to take breaks and access professional support as needed throughout their work.

We implemented a layered, automated anonymization protocol to ensure complete de-identification of users and channels. All usernames, profile links, subreddit identifiers, timestamps, and geolocation metadata were stripped or replaced with generic placeholders such as `[USER]`, `[PROFILE]`, or `[SUBREDDIT]`. Personally identifiable information (PII), including phone numbers, email addresses, and physical locations, was redacted or tokenized as `[INFO REDACTED]`. Images were processed through automated pipelines to blur or crop identifiable regions, removing any visual cues to re-identification. A secondary manual verification step will precede any public release to ensure comprehensive compliance with anonymization requirements.

The dataset was sourced from Reddit, a publicly accessible platform, and all data collection adhered to Reddit's terms of service. We emphasize that despite these precautions, both Reddit and the LLMs used may carry inherent biases. Future users are encouraged to critically evaluate and mitigate such biases in their downstream applications.

REDDIX-NET will be distributed under

a restrictive Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) license, requiring users to accept an ethical usage agreement. Access will be granted only after a valid research use case is submitted and approved via a dedicated form. We release only a curated subset of the dataset, not the full corpus, to minimize risks. Additionally, all dataset access will be logged, ensuring accountability and traceability in case of any ethical breaches or re-identification attempts.

To further reinforce ethical use, we outline a responsible usage checklist:

- **Permitted Uses:** Academic, non-commercial research with recognized IRB or ethical committee approval; development of harm-reduction strategies; bias detection and fairness audits; and responsible studies involving sensitive content.

- **Prohibited Uses:** Attempts to re-identify users or profiles; any activity contributing to sexual exploitation or violating relevant legal frameworks such as the Immoral Traffic (Prevention) Act (1956); or commercial use without explicit written approval.

A clarification is warranted regarding the annotation team: the annotation effort was directly coordinated by one of the authors, a member of the IRB granting organization's research team. All annotators were also members of this organization, and the work was conducted collaboratively with the research team. Although no financial compensation was provided, the annotators volunteered because they were driven by their strong belief in and commitment to the project's objective. The annotators were fully informed of the study's goals and aligned with its ethical standards as outlined in the approved IRB protocol.

By adhering to these principles, we aim to enable responsible research that promotes societal benefit, respects individual dignity, and avoids infringement upon consensual adult expression. This project exemplifies our commitment to ethical innovation and accountable AI deployment in sensitive domains. We acknowledge the use of AI assistants to draft portions of the paper and support related tasks, such as editing and formatting, with all content reviewed and finalized by the authors.

References

Sima Amirkhani, Mahla Fatemeh Alizadeh, Dave Randall, Gunnar Stevens, and Douglas Zytko. 2026. My parents expectations were overwhelming: Online dating romance scams targeting minors in iran through exploitation of parental pressure. *arXiv preprint arXiv:2601.16321*.

Nazanin Andalibi and 1 others. 2018. Social support, reciprocity, and anonymity in responses to sexual abuse disclosures on reddit. In *Proceedings of the ACM on Human-Computer Interaction*.

Han Bao, Qinying Wang, Zhi Chen, Qingming Li, Xuhong Zhang, Changjiang Li, Zonghui Wang, Shouling Ji, and Wenzhi Chen. 2025. Vmoda: An effective framework for adaptive nsfw image moderation. *arXiv preprint arXiv:2505.23386*.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

D.K. Brown. 2018. Reddit's veil of anonymity: Predictors of engagement. *Social Media + Society*.

Manan Roy Choudhury, Adithya Chandramouli, Mannan Anand, and Vivek Gupta. 2026. Better call CLAUSE: A discrepancy benchmark for auditing LLMs legal reasoning capabilities. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5776–5818, Rabat, Morocco. Association for Computational Linguistics.

Manan Roy Choudhury, Anirudh Iyengar Kaniyar Narayana Iyengar, Shikhhar Siingh, Sugeeth Puranam, and Vivek Gupta. 2025. TABARD: A novel benchmark for tabular anomaly analysis, reasoning and detection. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21783–21817, Suzhou, China. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Preprint*, arXiv:1703.04009.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *Preprint*, arXiv:2005.00547.

Maria Diaz and Anand Panangadan. 2020. Natural language-based integration of online review datasets for identification of sex trafficking businesses. In *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 259–264. IEEE.

Christopher J Ferguson and Richard D Hartley. 2022. Pornography and sexual aggression: Can meta-analysis find a link? *Trauma, Violence, & Abuse*, 23(1):278–287.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Luca Giommoni and Ruth Ikwu. 2021. Identifying human trafficking indicators in the uk online sex market. *Trends in Organized Crime*, pages 1–24.

Gemini Team Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. https://arxiv.org/abs/2312.11805.

Kelly Caroline Gothard. 2021. *The incel lexicon: Deciphering the emergent cryptolect of a global misogynistic community*. The University of Vermont and State Agricultural College.

Government of India. 2000. The information technology act, 2000.

Government of NCT of Delhi. 2023. Implementation of immoral traffic (prevention) act.

Vivek Kumar Gupta. 2025. The digital childhood dilemma: Reconciling children's rights, online safety and legal safeguards against exploitation in an era of cyber vulnerabilities. *Journal of Teachers and Teacher Education*, 2(1):01–11.

Ehsan Ul Haq and 1 others. 2025. Exploring self-disclosure norms and engagement dynamics on reddit. *arXiv preprint arXiv:2502.10701*.

Michelle Ibanez and Dan Suthers. 2016a. Detecting covert sex trafficking networks in virtual markets. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 876–879.

Michelle Ibanez and Daniel D Suthers. 2016b. Detecting covert sex trafficking networks in virtual markets. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 876–879. IEEE.

Shalini Jangra, Zaid Almahmoud, Suparna De, Gareth Tyson, Ehsan Ul Haq, and Nishanth Sastry. 2025. Understanding the complexities of responsibly sharing nsfw content online. *arXiv preprint arXiv:2511.15726*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Burcu B Keskin, Gregory J Bott, and Nickolas K Freeman. 2021. Cracking sex trafficking: Data analysis, pattern recognition, and path prediction. *Production and Operations Management*, 30(4):1110–1135.

Xian Li, Yuanning Han, Di Liu, Pengcheng An, and Shuo Niu. 2026. When generative ai is intimate, sexy, and violent: Examining not-safe-for-work (nsfw) chatbots on flowgpt. *arXiv preprint arXiv:2601.14324*.

Claudio Marche, Ilaria Cabiddu, Christian Giovanni Castangia, Luigi Serreli, and Michele Nitti. 2023. Implementation of a multi-approach fake news detector and of a trust management model for news sources. *IEEE Transactions on Services Computing*.

Ministry of Electronics and Information Technology, Government of India. 2021. Information technology (intermediary guidelines and digital media ethics code) rules, 2021.

Kimberly J Mitchell, David Finkelhor, and Janis Wolak. 2003. The exposure of youth to unwanted sexual material on the internet: A national survey of risk, impact, and prevention. *Youth & Society*, 34(3):330–358.

Jochen Peter and Patti M Valkenburg. 2016. Adolescents and pornography: A review of 20 years of research. *The Journal of Sex Research*, 53(4-5):509–531.

Mary E. PettyJohn and 1 others. 2025. Teens seeking information and advice about sexual behaviors on reddit. *Journal of Adolescent Health*.

Simone Raponi, Zeinab Khalifa, Gabriele Oligeri, and Roberto Di Pietro. 2022. Fake news propagation: a review of epidemic models, datasets, and insights. *ACM Transactions on the Web (TWEB)*, 16(3):1–34.

Skye Roisen P Sanchez and Genesis G Genelza. 2025. A systematic literature review on sexual exploitation and abuse of children: Prevalence, risk factors, and societal responses. *Universe International Journal of Interdisciplinary Research*, 5(10):207–219.

M. Shelton. 2015. A qualitative study of disclosure within and beyond reddit.

Ayliana Teitelbaum. 2020. *The Ethics of Reddit and an Artificial Moral Compass*. Ph.D. thesis, New York, NY. Stern College for Women. Yeshiva University.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 21 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, and 1 others. 2022. The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Longshaokan Wang, Eric Laber, Yeng Saanchi, and Sherrie Caltagirone. 2020a. Sex trafficking detection with ordinal regression neural networks. *Preprint*, arXiv:1908.05434.

Longshaokan Wang, Eric Laber, Yeng Saanchi, and Sherrie Caltagirone. 2020b. Sex trafficking detection with ordinal regression neural networks. *arXiv preprint arXiv:1908.05434*.

Kyle Wiggers. 2022. Gpt-2 output detector demo. https://openai-openai-detector.hf.space/. [Online; accessed 2023-01-02].

Fan Yang, Yihao Huang, Jiayi Zhu, Ling Shi, Geguang Pu, Jin Song Dong, and Kailong Wang. 2025. Seeing it before it happens: In-generation nsfw detection for diffusion-based text-to-image models. *arXiv preprint arXiv:2508.03006*.

Appendix-1: Prompts

Classification/Evaluation Prompt

Classification Prompt

Here is the content of a post. It has the following attributes:
1. Title: "title"
2. Image Description: "image_description"
3. Comments: "comments"

Your task is to classify this post into one of the following five categories of services related to online prostitution: 1. Physical Services: Posts offering in-person sexual services or physical prostitution. 2. Virtual Services: Posts offering virtual interactions such as video calls, virtual sex, or promoting platforms like OnlyFans. 3. Exhibitionism: Posts showcasing exhibitionistic behavior, such as public displays or other forms of showcasing oneself. 4. Content Creation and Sales: Posts promoting or selling photos, videos, or other content without direct interaction. 5. Couples and Group Interactions: Posts seeking interactions involving couples, threesomes, or group scenarios. 6. Miscellaneous Fun/Exploration: Posts describing non-specific fun, exploration, or interactions that do not fall into the other categories.
Carefully analyze the title, image description, and comments. Then determine which of the five categories the post best fits. Respond with only the category name. If the information is insufficient, respond with 'Uncategorizable'. Return only the category name or "Uncategorizable" based on your analysis, in the response.

Hyperparameters

Key hyperparameters we used in our experiments include the temperature, which controls the randomness of predictions and is typically set between 0.2 and 0.5 to ensure a more deterministic setting, and the max token length, chosen based on the average post length. Additionally, the models are fine-tuned with a learning rate range of 1e-5 to 1e-3 and a batch size of 16-64, with the number of training epochs determined by the loss function's convergence. During inference, models may use nucleus sampling (top-p) with a probability threshold of 0.9. The evaluation metrics for this task include Distribution Accuracy, Accuracy, F1 Score, Precision, Mean Absolute Error (MAE), and Shannon Entropy of the distribution.

Prompt 1 instructs an AI to classify a post into one of five distinct categories related to online prostitution services based on the post's title, image description, and comments. It delineates clear definitions for each category, ranging from physical and virtual services to exhibitionism, content creation, couples and group interactions, and miscellaneous fun/exploration, ensuring that the classification process is both structured and comprehensive. The prompting strategy emphasizes a careful, context-based analysis of the provided attributes and requires the AI to return only the appropriate category name or "Uncategorizable" if the information is insufficient, thereby promoting precise, deterministic decision-making in the classification process.

Sentiment Analysis Prompt

Sentiment Analysis

You are an expert AI performing sentiment analysis.
Analyze the following text and provide the following insights: 1. Sentiment: Positive, Neutral, or Negative, with a confidence score (0-1). 2. Emotion Classification: Identify the dominant emotion (e.g., joy, anger, sadness, surprise, etc.). 3. Keywords: Extract the main keywords or phrases relevant to the context. 4. Tone: Determine the tone (e.g., formal, casual, playful, persuasive, etc.).

Output Format:

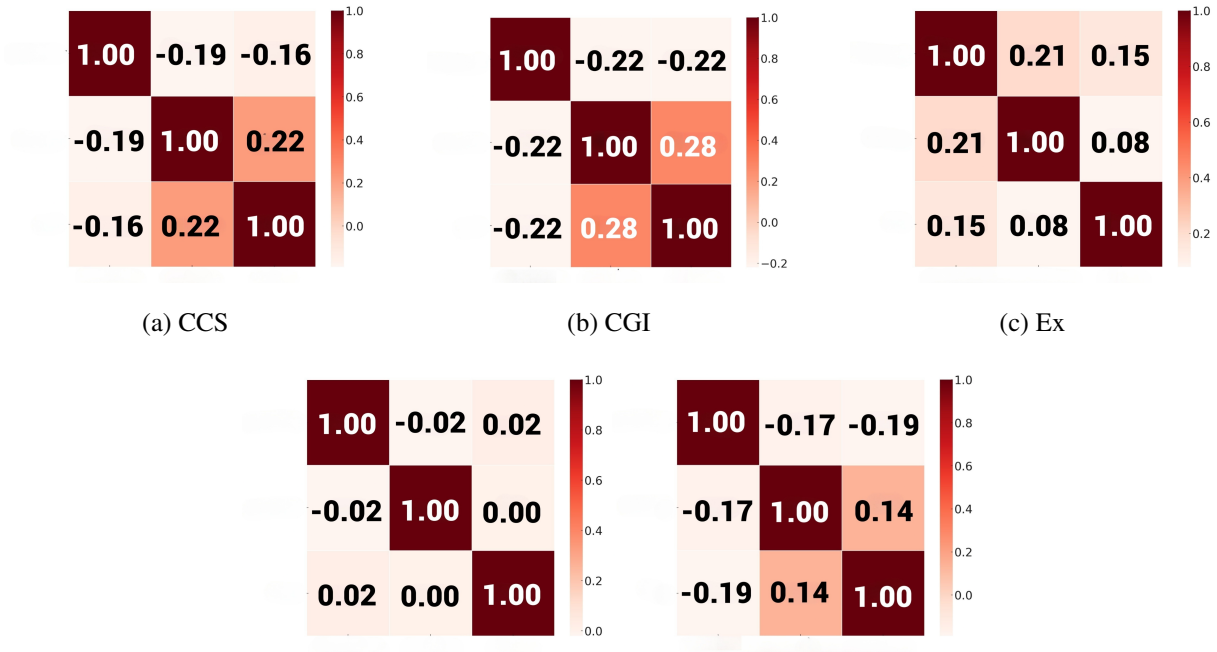
- Sentiment: [label], Confidence: [score]
- Emotion: [emotion]
- Keywords: [keywords]
- Tone: [tone]

The provided sentiment analysis prompt instructs an expert AI to perform a comprehensive evaluation of a given text by extracting multidimensional insights. Specifically, it requires the AI to determine the overall sentiment - positive, neutral, or negative while providing a confidence score, classify the dominant emotion (such as joy,

anger, or sadness), extract key phrases or keywords pertinent to the context, and assess the tone (e.g., formal, casual, or playful) of the text. The prompt also specifies a structured output format, ensuring results are returned consistently and in a standardized manner. This prompting strategy is designed to facilitate a detailed, context-aware analysis that leverages both qualitative and quantitative dimensions, thereby enhancing the interpretability and reliability of the sentiment analysis process.

Comments Classification Prompt

```
You are an expert content analyst. For each
comment provided, classify it into exactly one of
the following 18 classes:

1. Users Who Are Enjoying the Post and Its
Contents (Engagement & Positive Sentiment)
   - "Wow! You look absolutely stunning!"
   - "Absolutely mesmerizing, I'm hooked!"
   - "Stunning visuals, keep up the great work!"

2. Users Who Are Demanding Such Services (Intent
to Purchase/Engage)
   - "How much do you charge for this service?"
   - "Where can I reach you for more details?"
   - "Are you available for a private session?"

3. Users Who Are Requesting Specific Content
(Content Demand Trends)
   - "Can you do a video in a red dress?"
   - "Id love to see more dance moves from you!"
   - "Could you post more outdoor shoots?"

4. Users Who Are Skeptical or Questioning
Authenticity (Trust & Credibility Issues)
   - "Is this actually you or just edited?"
   - "Has anyone actually met her? Looks fake."
   - "Seems too polished—are these authentic?"

5. Users Who Are Providing Reviews &
Recommendations (Word-of-Mouth & Service
Feedback)
   - "Shes super professional and amazing to work
   with!"
   - "Had a great time, shes very professional!"
   - "Overpriced, not worth it."

6. Users Who Are Discussing Legality & Ethics
   - "Isn't this kind of thing banned here?"
   - "Should this even be allowed on this
   platform?"
   - "Im concerned about the legality of this
   content."

7. Users Who Are Competing or Self-Promoting
Services
   - "Check out my profile if you like this!"
   - "I offer exclusive content at a discount!"
   - "I can do this for half the price. DM me!"

8. Users Who Are Negotiating Prices or Services
   - "Can you do this for \$40 instead?"
   - "Is there any discount if I book multiple
   sessions?"
   - "How about a special rate for returning
   customers?"

9. Users Who Are Complaining About Payment or
Delivery Issues
   - "I paid but never got my order!"
   - "She stopped replying after I sent the
   payment!"
   - "This is a scam, don't fall for it!"

10. Users Who Are Making Unsolicited Requests or
Harassment
    - "Send me something for free first!"
    - "I'll find you if you don't reply!"
    - "Do this for me, or else!"

11. Comments that are Multi-Lingual
    - "Teri nazaron mein vo jadu hai."
    - "Include comments from languages which are
    not English."
    - "acha hai"

12. Comments that are Fantasy and Violent Demands
    - "Show me an online act that blends erotic
    fantasy with a violent edge."
    - "I demand you to enact a dark fantasy scene
    with intense aggression."

13. Comments that are Emotional Support
    - "Your posts always brighten my day!"
    - "I appreciate your openness; it helps me
    feel less alone."

14. Comments that are Verification and Identity
Inquiries
    - "Is this really you or just an
    impersonator?"
    - "Can you prove that this is your real
    account?"

15. Comments that are External Link Sharing
    - "Check this link out for exclusive content:
    [external link]"
    - "Visit my page for more: [link]"

16. Comments that are Illicit Propositions /
Explicit Requests
    - "Can you send me private pics? Ill pay
    extra."
    - "Do you do custom videos with nudity?"

17. Comments that are Self-Assertive/Confidence
Expressions
    - "Im the best at what I do, no one compares!"
    - "I always get what I want, and this is no
    different."

18. Comments that are Ambiguous or Multi-Response
Comments
    - "I'm not sure what to think about this..."
    - "Interesting... I wonder whats really going
    on."

For each comment provided, classify it into
exactly one of the above categories and return
the output as a JSON object with each original
comment as a key and its classification as the
value.
```



(a) CCS (b) CGI (c) Ex

(d) PS (e) VS

Figure 6: Correlation heatmaps depicting the interplay between sentiment, emotion, and tone across multiple dataset categories. Darker shades denote stronger correlations, while lighter hues indicate weaker associations.

Appendix-2: Sentimental Analysis

Sentimental Analysis using LLMs

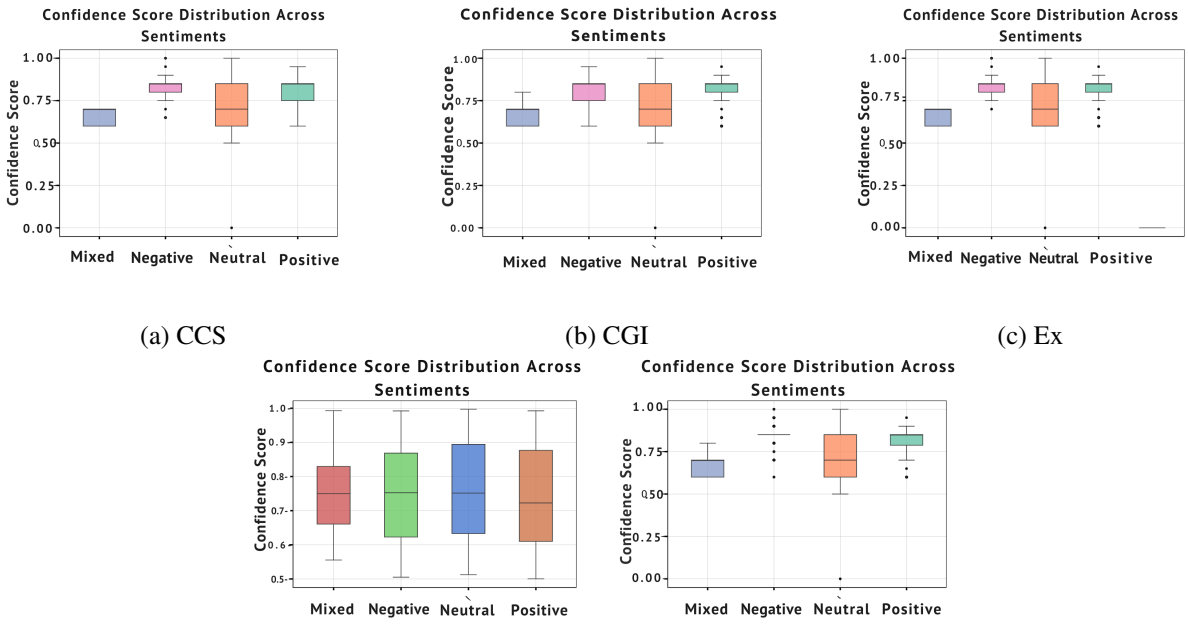
The above prompt positions the AI as an expert content analyst, using a persona-based approach to accurately classify user comments into 18 predefined categories. This structured prompt incorporates few-shot examples for each category, ranging from positive engagement and service inquiries to external link sharing and ambiguous expressions, thereby guiding the AI with concrete instances of desired outputs. By instructing the AI to evaluate each comment and return a JSON object mapping each comment to its classification, the strategy leverages contextual cues and demonstration-based learning to ensure consistent, precise categorization.

Distribution of Tone Types Across Sentiment Types: Based on the two nested pie charts (Figure 7) showing sentiment analysis alongside emotions and tone distribution, here's a comprehensive analysis: The sentiment distribution in both charts shows a predominantly positive and neutral outlook, with 48% positive sentiment the largest segment, followed by 42.9% neutral and 7.21% negative. This indicates that the overall communication style in couples and group interactions tends to maintain a constructive, balanced emotional atmosphere, with very few instances of negative exchanges.

Looking at the emotional aspects in the first chart, Desire (29.4%) and Joy (23.9%) emerge as the dominant emotions, collectively accounting for over half of the emotional expressions. This is followed by a notable segment of "None" (14.8%) and "Interest" (9.03%), suggesting that while interactions are generally emotionally engaged, there are also periods of neutral or emotionally reserved communication. The presence of other emotions, such as Anticipation, Surprise, and Excitement, in smaller proportions indicates a rich diversity of emotional expression, though negative emotions like Anger remain minimal (3.04%).

The tone analysis in the second chart provides interesting insights into the communication style: a Casual tone strongly dominates at 65.7%, followed by a Formal tone at 17.2%. This suggests that most interactions maintain a relaxed, comfortable atmosphere while still preserving some level of formality when needed. The presence of Neutral (7.89%), Informal (5.26%), and Playful (2.28%) tones, with minimal Aggressive tone (1.67%), indicates that the communication environment is generally conducive to open and comfortable interaction while maintaining appropriate boundaries and respect.

Confidence Score Distribution by Sentiment: From Figure. 7, the four sentiment categories (Pos-



(a) CCS (b) CGI (c) Ex

(d) PS (e) VS

Figure 7: The Confidence Score Distribution by Sentiment: Box Plot Comparison of Positive, Neutral, Mixed, and Negative Categories

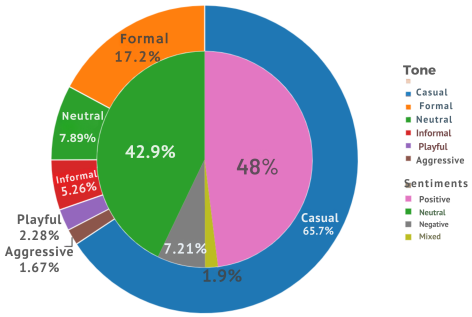
itive, Neutral, Negative, Mixed). Overall, each box plot reveals moderate to high median confidence values, suggesting that the underlying model generally assigns sentiment labels with a notable degree of certainty. However, the presence of outliers and varying interquartile ranges across subplots indicates that classification confidence can fluctuate depending on the specific context or linguistic cues in the data.

A closer inspection of the individual subplots highlights subtle differences in how sentiments are classified. In contrast, other classes (e.g., (a) CCS and (c) Ex) exhibit a broader spread for Neutral or Mixed sentiments, suggesting that the model occasionally encounters more ambiguity when distinguishing between emotionally neutral content and text that blends multiple affective tones. Negative sentiment typically shows a slightly wider distribution, suggesting potential variability in the strength with which negative cues are detected.

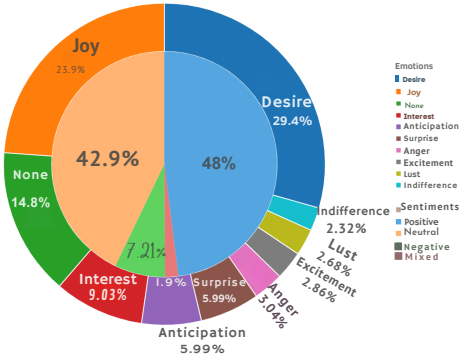
Collectively, these findings underscore a robust, yet context-sensitive classification process. While Positive sentiment often emerges with higher, more consistent confidence scores, Neutral, Mixed, and Negative categories reveal more diverse confidence intervals, reflecting the nuanced nature of human language and emotional expression. The recurring outliers across subplots further emphasize that certain instances may challenge the model's ability to assign a definitive sentiment category. Overall, the distribution of confidence scores across these five classes illustrates a generally reliable classification framework, albeit one that must navigate the inherent complexities of the sentiment-laden text.

Sentimental Analysis using PLM (BERT)

In reviewing Figure 9 describing the sentiment proportions across five categories, a clear trend emerges. Positive sentiments such as "appreciation" and "satisfaction" account for a substantial share across most categories, indicating that user feedback skews favorably. "Neutral" sentiment also appears consistently, though at varying levels, suggesting a notable fraction of content that neither leans strongly positive nor negative. In contrast, "aggression" and "frustration" are relatively lower, suggesting that overtly negative expressions are less common in the overall dataset.



(a) Sentiment and Tones Distribution across sentiment types for different classes in the dataset for CGI Category.



(b) Sentiment and Emotions Distribution across sentiment types for different classes in the dataset for CGI Category.

Figure 8: Distribution of tone and emotions across sentiment types for different classes in the dataset.

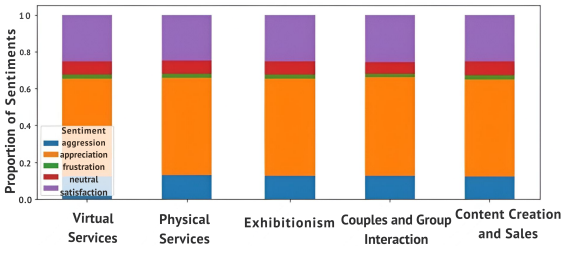


Figure 9: Sentiment Analysis using BERT Model

A closer look reveals subtle differences in sentiment composition among the categories. For example, Physical Services, Content Creations, and Sales exhibit a higher prevalence of "appreciation," indicating more frequent expressions of gratitude or praise. These observations are derived from a BERT-based model that leverages contextual embeddings to classify text with a high degree of nuance. Consequently, the analysis highlights both the generally positive nature of user communications and the importance of contextual factors in shaping sentiment.

Appendix-3: REDDIX-NET Metadata Analysis

Figure 10 provides a clear visual summary of how performers are distributed and overlap across multiple categories, including Miscellaneous Fun, Content Creation and Sales, Physical Services, Virtual Services, Couples and Group Interaction, and Exhibitionism. Each row corresponds to a category, and the black dots indicate shared performers among these categories. The bar chart at the top shows the number of performers participating in each specific combination of categories. Notably, "Exhibitionism" and "Couples and Group Interaction" are the most prevalent, as evidenced by the tallest bars, suggesting their high popularity or frequent reporting. Overall, the figure underscores that while many performers concentrate on a single category, a noteworthy subset engages in multiple overlapping areas, highlighting the importance of cross-category involvement. Looking at the intersection sizes in the top bar chart, it is evident that most performers participate in only one or two categories. However, a distinct group of five performers is active across a broader range of categories, indicating significant overlap in their services and interactions. Additionally, other smaller clusters—such as a group of three performers—reveal that although single-category involvement is common, a considerable minority diversifies their participation.

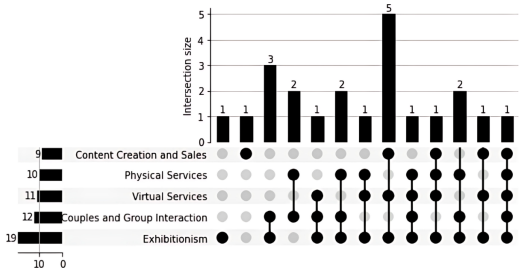


Figure 10: Total No. of performers per category and their overlap

Appendix-4: Labelling Process

The dataset was annotated through a structured labeling process executed by a team of three trained annotators (two male, one female). The

composition of the annotation team was deliberately designed to include a female perspective, a methodological choice intended to mitigate potential gender-based observational biases. This ensures a more balanced and representative interpretation of the illicit services offered across different genders within the corpus.

The annotation workflow proceeded as follows: Each annotator was assigned a subset of user profiles and their associated activities within the designated subreddits. The fundamental unit of analysis was the user's explicit service offering, distilled from their posts, comments, and profile descriptions. Annotators were tasked with performing a qualitative analysis of this content to identify the nature of the advertised service and classify it according to a predefined annotation schema.

This schema was developed iteratively based on a preliminary content analysis of the data. It consists of five mutually exclusive categories that encapsulate the primary types of illicit services observed. The categories are formally defined as:

- **Content Creation and Sales (CCS):** This category encompasses the production and sale of digital media. It includes, but is not limited to, the sale of pre-made or custom photosets, video clips, and subscriptions to platforms like OnlyFans or Fansly, where explicit content is monetized. The key differentiator is the transactional nature of acquiring a static or pre-recorded media product.

- **Couples and Group Interactions (CGI):** This label applies to services that explicitly involve more than one participant, such as live performances by couples, or services marketed towards groups. This category is distinct from individual services and highlights collaborative or multi-person offerings.

- **Exhibition (Ex):** This category is defined by performative acts where the service is a live or public display for an audience. This primarily includes real-time webcam services (camming) or other forms of live, voyeuristic exhibition where the interaction is centered on the act of being watched.

- **Physical Services (PS):** This label is assigned to any service that requires direct, in-person physical contact or meetups. Annotators identified these services through explicit keywords related to location, availability for "in-call" or "out-call" appointments, and other language indicating a non-virtual transaction.

- **Virtual Services (VS):** This category covers interactive, one-on-one services conducted remotely. Examples include synchronous activities such as sexting, live video calls, and "Girlfriend/Boyfriend Experience" (GFE/BFE) simulations that occur in real time but without physical presence. This is distinct from CCS as the service is an interactive experience rather than a media product.

Upon completion of the independent annotation phase, the labels were compiled for a quantitative reliability assessment. To validate the consistency and reproducibility of our schema and the annotators' judgments, we calculated Inter-Annotator Agreement (IAA) scores. This analysis was conducted for all possible annotator pairs (1,2), (1,3), and (2,3), as well as for the complete triad of annotators (1,2,3) to provide a comprehensive measure of concordance across the dataset.

Procedure Involving Data Collection and Construction

1. **Data Collection** The data for this study was collected from three specific subreddits identified as primary hubs for discussions related to illicit services. Data extraction was performed using the Reddit API, facilitated by the PRAW (Python Reddit API Wrapper) library, which enabled the retrieval of both posts and comments from these subreddits.

2. **Data Cleaning** The initial dataset underwent cleaning to remove irrelevant or extraneous content. Posts and comments deemed non-substantive, such as greetings (e.g., "Hi," "Hello!"), were removed to ensure the dataset focused solely on meaningful exchanges related to the research topic.

3. **Data Preprocessing** To protect the anonymity of individuals involved, several preprocessing steps were implemented. Posts containing visible faces were excluded from the dataset, as most posts naturally blurred such identifying features. Additionally, all usernames and Reddit IDs were stripped from the data, retaining only the content of the posts and comments for analysis.

Appendix-5: Ablation studies on feature importance

The evaluation is summarized in the table. 7 provides insights into the relative contribution of different feature sets to classification performance.

Exclusion of Emotion Features: The model maintains near-perfect performance (approximately 0.99 across Accuracy, Precision, Recall, and F1 Score) even when emotion-related features are removed. This indicates that emotion features are complementary rather than strictly required for classification accuracy, suggesting that other expressive signals already capture sufficient discriminative information for the core prediction task.

Use of Individual Feature Sets in Isolation: When the model is trained using only sentiment features, only emotion features, or only tone features, performance drops substantially, with metric values ranging between approximately 0.07 and 0.14. This demonstrates that while each feature set contains a useful signal, none is sufficient on its own to support robust classification. Effective performance, therefore, depends on combining multiple expressive cues.

Exclusion of Comment Features: The removal of comment-related features results in extremely poor performance (approximately 0.07 across metrics), indicating a near-total failure of classification. This sharp decline underscores the central role of comment-derived contextual information in distinguishing service categories, highlighting that engagement context is foundational to the model's predictive capability.

Relative Contribution of Expressive Features: The ablation results indicate that expressive features contribute unequally to model performance. Removing sentiment, tone, or metadata features leads to substantial degradation in classification accuracy, whereas removing emotion features alone does not significantly affect performance in this experimental setup (Accuracy 0.99, F1 0.99). This suggests that sentiment polarity and tonal cues serve as the primary discriminative signals for category prediction, while emotion features play a complementary role, supporting psychosocial interpretation rather than being strictly necessary for classification accuracy.

Feature Set	Accuracy	Precision	Recall	F1 Score
No Sentiment Features	0.03	0.04	0.03	0.03
No Emotion Features	0.99	0.98	0.99	0.99
No Tone Features	0.02	0.03	0.02	0.01
No Comment Features	0.07	0.06	0.09	0.06
No Metadata Features	0.03	0.03	0.02	0.03
Sentiment Features Only	0.14	0.12	0.13	0.11
Emotion Features Only	0.07	0.08	0.07	0.07
Tone Features Only	0.09	0.09	0.08	0.09

Table 7: Table representing ablation study on feature importance.

Overall, these findings indicate that REDDIXNET benefits from a balanced combination of contextual comment features and LLM-derived expressive signals, particularly sentiment and tone. Together, these feature groups provide the strongest predictive contribution, while emotion features enhance interpretability and support downstream psychosocial analysis.

Appendix-6: Psychological and Social Implications Study

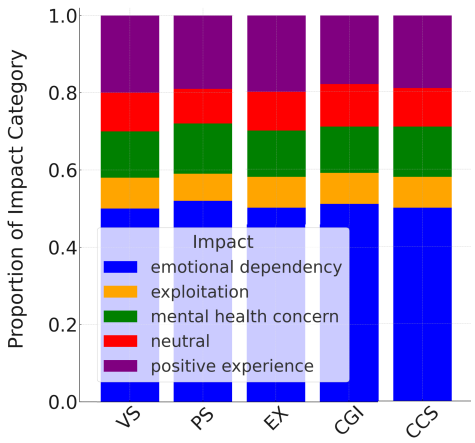


Figure 11: Category-wise impact proportion highlighting the proportion of emotional dependency, exploitation, mental health concerns, neutral perceptions, and positive experiences across different online prostitution categories. This provides insights into the psychosocial expression patterns associated with various engagement types.

For interpretability, emotion predictions were grouped into five impact categories: emotional dependency, exploitation indicators, mental health concerns, neutral perception, and positive experience, which form the basis of Figure 11. The aggregation into higher-level psychosocial impact categories was performed using a few-shot LLM-based classification setup, where emotion predictions and contextual comment text were provided to the LLM to assign one or more impact categories. A strati-

Model	VS				PS				Ex				CGI				CCS			
	TP	FP	TN	FN	TP	FP	TN	FN	TP	FP	TN	FN	TP	FP	TN	FN	TP	FP	TN	FN
LLaMA	0.19	0.01	0.60	0.20	0.19	0.01	0.61	0.19	0.19	0.01	0.55	0.25	0.18	0.02	0.55	0.3	0.2	0.0	0.6	0.2
GPT-4	0.20	0.00	0.56	0.24	0.20	0.00	0.49	0.31	0.20	0.00	0.21	0.59	0.19	0.01	0.31	0.5	0.2	0.0	0.4	0.4
Gemini	0.19	0.01	0.59	0.21	0.20	0.00	0.58	0.22	0.20	0.00	0.39	0.41	0.20	0.00	0.36	0.4	0.2	0.0	0.5	0.3
Claude	0.19	0.01	0.62	0.18	0.18	0.02	0.59	0.21	0.20	0.00	0.45	0.35	0.20	0.00	0.46	0.3	0.2	0.0	0.5	0.3
Mistral	0.19	0.01	0.54	0.26	0.19	0.01	0.59	0.21	0.19	0.01	0.51	0.29	0.19	0.01	0.52	0.3	0.2	0.0	0.6	0.2

Table 8: True positive, False positive, True negative, and False negative for each model across categories.

fied subset of the automatically assigned labels was manually verified by annotators to ensure consistency and reliability. Figure 11 presents the distribution of impact categories across service types (VS, PS, EX, CGI, CCS). The stacked bar chart shows that emotional dependency (blue) makes up the largest proportion across all categories, consistently occupying nearly half of each bar. Exploitation (yellow) and mental health concern (orange) appear in smaller proportions, with exploitation being slightly higher in EX and PS compared to other categories. Neutral perception (red) is relatively minor but present across all categories at comparable levels. Positive experience (purple) consistently accounts for the second-largest segment, reaching close to one-third in some categories, such as CCS and CGI. Overall, the chart highlights that emotional dependency and positive experiences dominate the distribution, while exploitation, mental health concerns, and neutral perceptions remain comparatively lower.

Note: We emphasize that these categories represent inferred psychosocial indicators derived from linguistic expression and should not be interpreted as clinical assessments of users' mental health.

Appendix-7: Additional Observational Insights

This appendix presents additional observational insights derived from expressive and engagement patterns within REDDIX-NET. These observations are based on aggregated linguistic signals and user interaction behavior and should be interpreted as indicators of expressed psychosocial reactions rather than direct measures of psychological or mental health outcomes.

1. Engagement Intensity Across Service Categories. Different service categories exhibit distinct engagement dynamics. Categories such as Exhibitionism (Ex) and Virtual Services (VS) show higher proportions of emotionally expressive and positively engaged comments, suggesting stronger interaction intensity. These patterns reflect differences in how users linguistically respond to different solicitation types.

2. Expressive Signals and Interaction Context. Sentiment, tone, and emotion distributions reveal variation in how users frame their interactions. For example, certain categories demonstrate elevated levels of desire, joy, or transactional tone, while others show more neutral or skeptical expressions. These expressive signals function as indicators of conversational context rather than evidence of underlying psychological states.

3. Psychosocial Indicator Patterns. The aggregated impact categories (emotional dependency, exploitation indicators, mental health-related expressions, neutral perception, and positive experience) reflect patterns in how users articulate reactions within comments. These categories capture observable linguistic markers associated with engagement behavior and do not imply clinical diagnosis or causal psychological effects.

4. Platform-Level Behavioral Rhythms. Temporal engagement patterns highlight cyclical activity in comment and post interactions. These patterns reflect platform-level usage rhythms and collective participation dynamics rather than region-specific or individual behavioral conclusions.

5. Moderation-Oriented Implications. The expressive and engagement patterns identified in this study may assist platform moderators and policy designers in understanding how solicitation posts generate interaction signals. Such insights can inform the development of context-aware moderation tools without implying direct psychological impact assessment.

Overall, these findings provide descriptive and computational insights into how users linguistically engage with explicit-service content within dedicated communities on a mainstream platform. They should be interpreted as platform-level behavioral observations rather than claims about individual-level psychological consequences.

Metric	G-L	G-C	G-M	L-C	L-M	C-M	G-L-C	G-L-M	G-C-M	L-C-M
Pre (%) VS	70.59	70.59	70.59	61.33	61.33	65.66	67.89	72.12	74.23	63.40
F1 (%) VS	15.89	15.89	15.89	22.44	22.44	26.08	19.00	19.38	18.77	23.57
Pre (%) PS	53.11	53.11	53.11	52.59	52.59	52.02	52.88	54.51	54.01	52.13
F1 (%) PS	47.42	47.42	47.42	51.71	51.71	51.87	51.40	47.15	47.51	50.52
Pre (%) Ex	71.10	80.45	63.50	75.88	69.83	67.03	74.63	70.87	83.98	90.69
F1 (%) Ex	40.10	40.10	40.10	17.11	17.11	20.52	28.44	33.59	36.61	17.76
Pre (%) CGI	85.00	85.00	85.00	86.18	86.18	86.86	85.61	86.54	86.01	87.95
F1 (%) CGI	17.19	17.19	17.19	25.17	25.17	23.20	22.14	25.84	23.84	27.68
Pre (%) CCS	58.47	58.47	58.47	60.00	60.00	59.26	61.96	55.64	56.78	59.78
F1 (%) CCS	19.03	19.03	19.03	14.74	14.74	5.05	16.31	20.00	18.48	15.74
MSE	0.464	0.464	0.464	0.486	0.486	0.475	0.471	0.465	0.461	0.482
MAE	0.464	0.464	0.464	0.486	0.486	0.475	0.471	0.465	0.461	0.482
JSD	0.128	0.128	0.128	0.142	0.142	0.144	0.134	0.126	0.125	0.139
Acc (%)	62.43	62.43	62.43	55.88	55.88	56.10	60.19	62.28	63.24	57.07

Table 9: Evaluation Results of aggregation of LLMs (ensemble methods). Here, G→ Gemini 1.5 Flash, L→ LlaMA 3.3-70B-Instruct, M→ Mistral 8×7B, Q→ Qwen 2.5 Turbo, C→ Claude 3.5 Haiku.

Appendix-8: Ensemble Ablation Analysis

An evaluation of dyadic (two-model) and triadic (three-model) ensembles was conducted to measure the impact of model aggregation. The results, shown in Table 9, highlight several key findings regarding the performance of these combined models.

The analysis reveals that triadic combinations consistently outperform dyadic pairs, particularly in overall accuracy and F1 scores. The G-C-M (Gemini-Claude-Mistral) ensemble is the top performer, achieving the highest accuracy (63.24%) and the lowest Jensen-Shannon Divergence (JSD) of 0.125, which indicates the best alignment with the ground truth distribution. This configuration also registered the lowest overall error (MSE/MAE of 0.461). A notable trade-off between precision and recall was observed for the 'Exhibition' (Ex) category, where some ensembles maintained perfect precision but at the cost of a significantly lower F1 score.

In summary, aggregating three diverse models, especially the G-C-M combination, yields more robust and accurate predictions than simpler two-model ensembles. Additionally, we have conducted an ablation study on feature importance, focusing on sentiment, emotion, and tone, which yielded valuable insights detailed in Appendix 5.

Note: It is important to note that the above observations are based on a limited and domain-specific dataset; therefore, the interpretations should be viewed as preliminary hypotheses rather than definitive conclusions.

Appendix-9: Error Analysis

We performed a comprehensive error analysis, summarized in the table. 8 (for our multi-label classification task across six distinct categories, assessing performance based on per-category True Positives, False Positives, False Negatives, and True Negatives. A predominant challenge across all models was the high rate of FNs, largely due to the failure to interpret subtle, coded, and slang-based language. For instance, Virtual Services and Exhibitionism have missed classifications when posts lacked explicit keywords, relying instead on implicit phrases like online sessions" or suggestive imagery. False Positives typically arose from contextual misinterpretations, such as flagging benign terms like meet and greet" as Physical Services or generic pronouns like "we" as Couples and Group Interactions. Significant confusion was also observed due to the inherent ambiguity of the Miscellaneous Fun category and the substantial thematic overlap between Virtual services and Content Creation and Services, making it hard for models to distinguish between live interactions and content sales.

Model-specific behaviors revealed distinct trade-offs between precision and recall. GPT-4 and Llama adopted a more conservative, high-precision approach, minimizing FPs but resulting in high FNs by overlooking nuanced cues, particularly in the Ex, MF, and CGI categories. Conversely, Mistral demonstrated stronger recall for VS and CCS by recognizing industry-specific terms (e.g., OnlyFans"), but this came at the cost of more FPs in the PS category. Gemini proved adept at identifying subtle Ex cues but tended to over-classify CGI and

CCS content. Claude provided a more balanced performance but struggled to resolve ambiguity between borderline PS and MF classifications. In summary, all models were consistently challenged by contextual ambiguity in vague terms like arrangements," the nuances of slang, and pervasive overlaps between service categories.