Detecting Child Objectification on Social Media: Challenges in Language Modeling

Miriam Schirmer[1], Angelina Voggenreiter[2], Jürgen Pfeffer[2], Emőke-Ágnes Horvát[1]
[1]Northwestern University
[2]Technical University of Munich
miriam.schirmer@northwestern.edu, angelina.voggenreiter@tum.de
juergen.pfeffer@tum.de, a-horvat@northwestern.edu

Abstract

Online objectification of children can harm their self-image and influence how others perceive them. Objectifying comments may start with a focus on appearance but also include language that treats children as passive, decorative, or lacking agency. On TikTok, algorithm-driven visibility amplifies this focus on looks. Drawing on objectification theory, we introduce a Child Objectification Language Typology to automatically classify objectifying comments. Our dataset consists of 562,508 comments from 9,090 videos across 482 TikTok accounts. We compare language models of different complexity, including an n-gram-based model, RoBERTa, GPT-4, LlaMA, and Mistral. On our training dataset of 6,000 manually labeled comments, we found that RoBERTa performed best overall in detecting appearance- and objectification-related language. 10.35% of comments contained appearance-related language, while 2.90% included objectifying language. Videos with school-aged girls received more appearance-related comments compared to boys in that age group, while videos with toddlers show a slight increase in objectification-related comments compared to other age groups. Neither gender alone nor engagement metrics showed significant effects. The findings raise concerns about children's digital exposure, emphasizing the need for stricter policies to protect minors.

1 Introduction

Recent investigations have raised serious concerns about children's presence on social media. A New York Times report uncovered a troubling trend on Instagram: a "marketplace of girl influencers," often managed by parents, that draws the attention of individuals with exploitative intentions (Valentino-DeVries and Keller, 2024). These findings underscore the broader risks associated with children's online presence, including on platforms like TikTok, where short-form videos encourage engagement from vast audiences. As of 2024, TikTok is estimated to have around 900 million users (Statista, 2024), making it a major platform for self-expression, creative content, and social engagement. However, this visibility also exposes children to harmful language, including targeted harassment, inappropriate comments, and objectification. Objectification, in this context, refers to language that reduces a child to their physical attributes rather than recognizing them as individuals with agency. This includes excessive focus on appearance, comparisons to adult beauty standards, and possessiveness (Glick and Fiske, 2018).

Discussions about children on social media often focus on access, i.e., whether they should be allowed to participate and consume content (Martínez Allué and Martín Cárdaba, 2024). A bigger challenge, however, is when children themselves become the content. Most often, this content is shared by parents. 90% of parents in the United States who regularly use social media have shared content about their children online (Amon et al., 2022). While some of these videos portray children in everyday contexts, others—intentionally or unintentionally—place a strong emphasis on their physical appearance (Figure 1).

Existing studies on online harm, such as hate speech detection and cyberbullying, have developed robust models for identifying harmful lan-

Figure 1: Anonymized examples of typical child video content on TikTok.

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 396–412
August 1, 2025 ©2025 Association for Computational Linguistics

guage (Basile et al., 2019; Fortuna and Nunes, 2018; Zampieri et al., 2019). However, objectification presents a distinct challenge: it often manifests in subtle, seemingly positive, or ambiguous ways that evade traditional detection methods. Objectifying comments may not contain overt insults or hate speech but instead fixate on a child's physical traits, compare them to adults, or sexualize their appearance under the guise of admiration (Bernard et al., 2018; Glick and Fiske, 2018). Prior work on complex social phenomena in NLP shows that theory-driven approaches improve both reliability and interpretability (Davis, 2018; Kovács et al., 2021; Breazu et al., 2025; Hovy and Yang, 2021), highlighting the need for a theory-grounded typology that captures how objectification is expressed.

This work is centered around the research questions of how objectifying language manifests in TikTok comments on children and what challenges state-of-the-art natural language processing (NLP) models face in detecting it:

(1) We introduce the **Child Objectification Language Typology**. We implement and test this typology on a dataset of 562,508 comments from 9,090 unique videos across 482 TikTok accounts. [1]

(2) We evaluate language models of different complexity, identifying their strengths and limitations in capturing implicit objectification in comments.

(3) We find that 10.35% of comments contain appearance-related language, and 2.90% include objectifying language. Videos featuring school-aged girls receive more appearance-related comments than those with boys.

2 Context and Measurement of Child Objectification

Although inappropriate content involving minors exists across many platforms, TikTok has faced particular criticism for facilitating the sexual exploitation of children and adolescents (Polito et al., 2022). Young users are often drawn to imitate trending content, which can include sensual or provocative dances or appearances in swimwear or underwear (Suárez-Álvarez et al., 2023). In addition, studies found that minors frequently receive sexually explicit comments and requests (Silva, 2019; Soriano-Ayala et al., 2023). Engagement metrics, such as likes, play a critical role in amplifying certain types of online discourse, including objectification. Content emphasizing physical appearance tends to receive higher engagement (Frederick et al., 2022; Fardouly and Vartanian, 2016). Additionally, girls and young women receive more appearance-based comments than boys on social media (Döring and Mohseni, 2019; Kim, 2021).

Most insights into the online exploitation of children have come from investigative journalism by major news outlets (Valentino-DeVries and Keller, 2024; Silva, 2019; Levine, 2022; Barry et al., 2021), while scientific studies on the topic remain limited. Although broader analyses of children on TikTok are beginning to emerge (Stephenson et al., 2024), academic research has so far focused mainly on qualitative reports and individual case studies (Khan and Bhattacharjee, 2022; Soriano-Ayala et al., 2023). One exception is a recent study suggesting that up to one fifth of videos on the platform may feature children, though it does not analyze the nature of this content in depth (Steel et al., 2025). Large-scale quantitative research focused specifically on children is needed for a more systematic investigation in this area.

2.1 *Sharenting* on Social Media

"Sharenting", a term used to describe parents sharing information about their children on social media, has become increasingly common in today's digital society (Cataldo et al., 2022; Verswijvel et al., 2019). Sharenting includes a wide range of activities, from posting photos and videos to sharing personal stories and milestones, often introducing children to the online world from an early age. While sharenting can help families stay connected and celebrate meaningful moments, it also raises important concerns about privacy, consent, and the long-term impact of creating a digital footprint for children (Stephenson et al., 2024; Walrave et al., 2022). The full extent of sharenting remains unclear and differs across countries and platforms. In a survey of 493 United-States-based parents who regularly use social media, nearly 90% reported sharing content about their children online (Amon et al., 2022). This practice can infringe on children's right to privacy, especially since many parents do not seek their child's consent before posting (Kopecky et al., 2020; Ní Bhroin et al., 2022; Van den Abeele et al., 2024). The shared content often includes sensitive information: for instance, in a sample of Facebook posts from 168 parents, 90.5% mentioned their child's first name,

[1] All code is publicly available at https://github.com/MiriamSchirmer/child-objectification.

83.9% included birthdates, and 32.7% shared personal documents or videos (Brosch, 2016).

2.2 Measuring Objectifying Language

2.2.1 Objectification Theory

Theories of objectification have been widely explored in psychology, feminist studies, and media research, describing how individuals are reduced to their physical attributes rather than being recognized as full persons. One central framework in this area is *Objectification Theory* (Fredrickson and Roberts, 1997). This theory conceptualizes objectification as the process by which individuals are perceived and treated primarily as bodies to be evaluated based on appearance. A key component of this framework is self-objectification, which occurs when individuals internalize an external observer's perspective, leading them to engage in body surveillance and appearance-based self-evaluation. This phenomenon has been linked to psychological consequences such as increased body shame, anxiety, and reduced cognitive performance, particularly in social media contexts where visual presentation is central (Moradi and Huang, 2008). In a related line of research, Glick et al. (1997) introduced *Ambivalent Sexism Theory*. This framework distinguishes between hostile sexism, characterized by overtly negative and demeaning attitudes, and benevolent sexism, which manifests as seemingly positive but ultimately restrictive perceptions that reinforce traditional gender roles. Similarly, *The Fragmented Body Theory* (Bernard et al., 2012, 2018) highlights how media representations frequently depict individuals as isolated body parts rather than whole persons, reinforcing objectifying narratives and shaping the way people perceive and describe others.

2.2.2 Natural Language Processing for Objectification Detection

NLP models have been widely used to detect harmful language online, including hate speech, cyberbullying, and toxicity (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). However, while hate speech is often characterized by hostility, threats, or dehumanization based on identity (Waseem et al., 2017; Vidgen and Derczynski, 2020), objectification can occur without overt negativity. Instead–similar to misogynistic language (Nozza et al., 2019; Samghabadi et al., 2020)–it can take the form of seemingly positive or neutral language that nonetheless reduces individuals to their appearance or sexualizes them (Glick and Fiske, 2018).

Measuring objectifying language requires a more refined approach that accounts for implicit linguistic cues and context. Prior studies have explored dictionary-based methods and supervised machine learning models to identify objectification (Farrell et al., 2019; Sik et al., 2023). While lexicon-based approaches provide a starting point, they struggle with context sensitivity and often lack the ability to distinguish between benign and problematic uses of appearance-related language. Supervised models, trained on manually labeled data, can improve accuracy but are constrained by the quality and representativeness of their training datasets. Advances in the development of LLMs have improved performance in capturing implicit language features by learning contextual patterns (Abdurahman et al., 2024; Ding et al., 2024). Still, language models of all sizes face challenges in distinguishing admiration from objectification, particularly when there is no obvious negativity (ElSherief et al., 2021; Li et al., 2024).

2.2.3 Child Objectification Language Typology for Social Media

Applying traditional objectification theory to social media comments directed at children is complex, as these theories were primarily developed to analyze the objectification of adult women. Regarding children, objectification rarely involves obvious sexualization; instead, it tends to appear in repeated focus on physical appearance, exaggerated admiration, and comparisons that reinforce external valuation. Concepts like infantilization, central to adult objectification, are less applicable here since childlike traits are inherent. To address these gaps, we propose a **Child Objectification Language Typology** (Table 1). This typology encompasses both explicit and subtle forms of objectification in social media discourse by distinguishing general appearance-related remarks and specifically objectifying language. It categorizes different forms of objectifying language on social media, incorporating concepts from objectification theory adaptable to children. Sexualizing or age-inappropriate language applies adult beauty norms to children, implying maturity beyond their actual age or sexualizing physical traits. Comparative and competitive appearance commentary ranks children's attractiveness, reinforcing social hierarchies and treating them as objects of comparison. Diminutive framing exaggerates cuteness, portraying children as fragile, doll-like, or dependent rather than recognizing them as developing individuals. Possessive comments include language that implies ownership, entitlement, or undue familiarity. Appearance-based language includes comments that describe a child's physical features, clothing, or attractiveness. While these comments are not necessarily objectifying, they shift the focus to physical features and are thus included as a baseline.

3 Scope and Research Questions

This study addresses how objectification manifests in social media discourse, evaluating the effectiveness of different language models in its detection. To systematically analyze this child objectification, we build on these research questions:

RQ1: What linguistic context characterizes how children are discussed in TikTok comments, and how do these patterns relate to objectification?

RQ2: How can we implement a typology to classify and analyze objectifying language in TikTok comments, and how well do different language models perform in this task?

RQ3: How are demographic factors (e.g., gender, age) and metadata (e.g., likes, downloads) related to the prevalence of objectifying language?

4 Methods

4.1 Video Collection and Annotation

As TikTok's user guidelines prohibit individuals under 13 from holding their own accounts (TikTok, 2024), this study does not examine accounts run directly by minors. Instead, we focused on accounts that feature children under 13 but are managed by adults—usually their parents. To create our dataset, we began by identifying TikTok accounts with the highest follower counts. To ensure a substantial and representative sample of approximately 500 TikTok accounts, we initially screened 25,000 of the most-followed accounts. We identified 825 accounts that regularly featured children and collected the first 100 videos from each. Both the videos and their comments were collected using the Ensemble Data TikTok API between November and December 2024.

A research team consisting of two postdoctoral researchers and two research assistants then reviewed the collected videos ($n = 82{,}500$) to identify those featuring a child under the age of 13. We relied on visual cues (e.g., toddlers), and when uncertain, we used any age information provided in TikTok videos or profile descriptions. Given the high volume of videos, we manually labeled a subset of 12,000 videos and trained a neural network classifier to detect children based on image frames. Our model achieved an accuracy of 93% (see Appendix, Section A.3 for details). We manually reviewed the videos again during a final annotation stage and could thus correct misclassifications. To ensure meaningful representation, our final dataset includes only accounts with at least 20 videos featuring children. In the final annotation round, we annotated the child's perceived gender, categorizing each instance as female, male, or having children of multiple genders present (mixed). When there were no visual cues indicating gender, we labeled the gender as other.[2] Age groups were similarly coded as infant, toddler, or school-aged.

4.2 Comment Collection and Classification

For each video, we collected up to the first 300 comments, along with associated video metadata. To ensure consistency and reliability in the analysis, we included only English-language comments based on TikTok's metadata. Focusing on English as the most widely used language on the platform helped reduce linguistic variation and improved the generalizability of our findings. The final dataset includes 562,508 comments from 9,090 unique videos across 482 TikTok accounts (Table 2). These accounts are distributed across 21 countries (see Appendix, Section A.4 for details).

Topic Modeling. Addressing RQ1 on the linguistic context of how children are discussed on TikTok, we first examined whether appearance-related language was prevalent enough in our dataset to justify applying our theoretically motivated typology to TikTok comments. To investigate this, we conducted topic modeling using BERTopic (Grootendorst, 2022) to identify common topics. The extracted topics were visualized using UMAP (Uniform Manifold Approximation and Projection).

Implementing the Objectification Typology. To implement our Child Objectification Language Typology (RQ2), we developed two distinct classifiers for both general appearance-related and specifically objectifying comments according to the typology (see Appendix, Figure A.1). For our training data,

[2] We acknowledge that gender identity is diverse. Our use of visual cues or descriptions is based on conventional perceptions and does not intend to exclude or invalidate non-binary, genderqueer, or other gender identities.

Category	Definition	Examples
Objectification-Based Language		
Sexualizing or Age-Inappropriate Comments	Implies maturity beyond the child’s age or applies adult attractiveness standards.	*“She’s gonna be a heartbreaker."* *“She doesn’t look 12 at all!"*
Comparative Comments	Ranks children’s attractiveness, reinforces beauty standards, or compares them to adults.	*“She’s the prettiest one in this group."* *“Such a strong little man."*
Diminutive Framing	Uses baby-like descriptors that emphasize smallness or cuteness.	*“Aww, such a tiny baby doll!"* *“She’s just a little angel!"*
Possessiveness	Implies ownership, entitlement, or personal attachment to the child.	*“She’s mine!"* *“Our little angel."*
General Appearance-Based Language	Describes physical features, clothing, or attractiveness. These comments may not be objectifying per se, but can amplify objectification through an emphasis on physical attributes.	*“I like her hair!"* *“Where did you buy his shoes?"*

Table 1: Categories of child objectification in social media comments.

Group	Videos	Comments
Child Gender		
Female	4,289 (47.18%)	259,705 (46.17%)
Male	3,346 (36.81%)	204,392 (36.34%)
Other	280 (3.08%)	12,677 (2.25%)
Mixed	1,145 (12.60%)	84,920 (15.10%)
Child Age Group		
Infant	1,766 (19.43%)	94,056 (16.72%)
Toddler	3,000 (33.00%)	171,560 (30.50%)
School-aged	2,995 (32.95%)	207,777 (36.94%)
Mixed	1,300 (14.30%)	88,289 (15.70%)
Total	**9,090** (100.00%)	**562,508** (100.00%)

Table 2: Overview of the TikTok Children Dataset. Absolute counts and percentages (in brackets) are provided to show the distribution of videos and their corresponding comments by child gender and age group. Mixed refers to videos featuring children of different perceived genders or age groups being featured in one video.

two postdoctoral researchers and one research assistant labeled a sample of 6,000 comments according to our typology. We calculated Fleiss’ Kappa for inter-annotator agreement (Fleiss, 1971), yielding moderate agreement with scores of $\kappa = .74$ for general appearance-based comments and $\kappa = .63$ for the objectification category. After the initial round of agreement checks, any ambiguous cases were discussed within the research team, and final labels were assigned collaboratively. Given the inherently subjective nature of interpreting objectification in social media comments, some disagreement was expected. While not perfect, this level of agreement is comparable to tasks like implicit hate speech detection (Li et al., 2024; Matter et al., 2024).

Model Selection. We implemented five classification models, evaluating their suitability for detecting objectification-related language in TikTok comments. Hyperparameter details for each model are included in the Appendix, Section A.5. During training, we used a 5-fold cross validation and a classification threshold of 0.5 for prediction. We evaluated each model using F1-score (binary), precision, recall, and AU-ROC.

N-Gram Neural Network Model. We implemented a fully connected neural network (NN) trained on n-gram representations, where n = [1, 2, 3]. In this model, the input text is transformed into term frequency representations of n-grams, which are then passed through dense feed-forward layers with non-linear activations. This model balances interpretability with expressive power, capturing both lexical patterns and basic phrase structures while remaining computationally efficient.

RoBERTa (finetuned). To incorporate deep contextual representations, we used RoBERTa (Liu et al., 2019). We fine-tuned RoBERTa on our dataset, optimizing key hyperparameters including learning rate and batch size. RoBERTa’s strength lies in its ability to capture subtle linguistic patterns and implicit biases, making it useful for detecting indirect forms of objectification in online discourse.

Large Language Models (GPT-4, LLaMA 2, and Mistral 7B). To evaluate the capabilities of large-scale generative models, we included GPT-4 (OpenAI et al., 2024) via OpenAI's API and open-source models LLaMA 2 (Touvron et al., 2023) and Mistral 7B (Jiang et al., 2023) in both zero-shot and few-shot settings. These models process classification as a text generation task, where we reframe objectification detection as a structured prompt-based task. Each prompt included instructions to either classify general appearance-related comments or objectifying comments targeted towards children. For the objectification classification task, we included the categories of our typology in both zero- and few-shot prompts (see Appendix, Section A.6).

4.3 Sociodemographic Characteristics and User Engagement Indicators

To analyze the relationship between gender, age, engagement metrics, and the prevalence of appearance-related and objectification-related comments (RQ3), we employed Ordinary Least Squares (OLS) regression. We estimated separate models for appearance-related and objectification-related comments, using the predicted probabilities of each comment (aggregated at the video level) as the dependent variables. Independent variables included gender, age group, and their interaction, along with the number of downloads and total comments to account for engagement.

5 Results

Answering RQ1 on language characteristics associated with comments on TikTok videos featuring children, we found that it is primarily shaped by expressions of affection and cuteness, familial references, aesthetic discussions, and engagement with TikTok trends. Overall, this reflects language that focuses on emotional reactions, social aspects, and platform-driven interactions. Figure 2 shows the semantic similarity among the 25 most frequent topics, illustrating how different themes cluster together based on shared linguistic patterns in their BERT-based embeddings. These embeddings were projected into two dimensions using UMAP for visualization (see Appendix, Section A.2 for details). Topics related to affection and cuteness, such as those containing words like "cute" and "baby," are closely grouped, reflecting their frequent co-occurrence in sentimental expressions. Family-related topics, including mentions of "sis," "sister," "momma," and "dad," form distinct clusters, highlighting the strong presence of familial framing in comments. Aesthetic discussions are also prominent, with words like "hair," and "shoes" indicating attention to fashion. The influence of TikTok culture is evident in words such as "tiktok(s)".

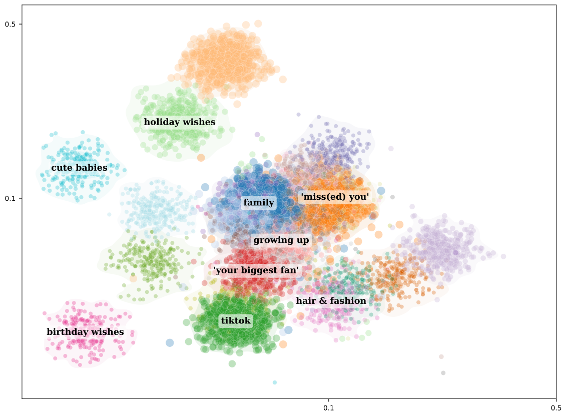

Figure 2: Top 25 topic clusters based on BERT embeddings of comments, reduced to two dimensions using UMAP. Each point represents a comment, positioned according to its semantic similarity to others. Point size reflects the topic frequency.

5.1 Objectification Detection

Table 3 presents the performance of our classification models in detecting general appearance-related and objectification-related language in TikTok comments (RQ2). For detecting general appearance-related language, RoBERTa achieved the highest F1-score (0.74), along with the top precision (0.62) and AU-ROC (0.98) values. Mistral (few-shot) reached the highest recall (0.97), but with very low precision (0.09), indicating a high rate of false positives and thus limited reliability. Among the generative LLMs, GPT-4 (few-shot) performed best with an F1-score of 0.51, slightly ahead of the open-source models. The NGram Neural Network performed comparably to GPT-4 (zero-shot), achieving an F1-score of 0.46, suggesting that lexical patterns still play a notable role in identifying appearance-related language. For objectification-related language, overall model performance was lower. Again, RoBERTa achieved the best F1-score (0.51) and led across most metrics. Among generative LLMs, LLaMA 3-7B (few-shot) showed relatively stronger results, reaching an F1-score of 0.25, surpassing GPT-4 and Mistral in this task. Overall, RoBERTa was the top-performing model across both classification tasks. Among open-source mod-

Classifier	General Appearance				Objectification			
LM	F1 (bin.)	Prec.	Recall	AU-ROC	F1 (bin.)	Prec.	Recall	AU-ROC
NGramNeuralNetwork	0.46	0.59	0.38	0.72	0.42	0.46	0.39	0.67
RoBERTa (finetuned)	**0.74**	**0.62**	0.91	**0.98**	**0.51**	**0.63**	0.44	**0.97**
OpenAI GPT-4 (zero-shot)	0.48	0.50	0.47	0.72	0.18	0.21	0.15	0.57
OpenAI GPT-4 (few-shot)	0.51	0.59	0.44	0.71	0.08	0.05	0.27	0.54
LLaMA 3-7B (zero-shot)	0.42	0.35	0.54	0.73	0.18	0.11	0.40	0.65
LLaMA 3-7B (few-shot)	0.45	0.40	0.52	0.73	0.25	0.17	0.47	0.69
Mistral-7B (zero-shot)	0.16	0.09	0.95	0.62	0.08	0.04	0.54	0.56
Mistral-7B (few-shot)	0.16	0.09	**0.97**	0.62	0.08	0.04	**0.69**	0.58

Table 3: Classification performance of the language models used in this work. We report Binary F1-Scores, Precision, Recall, and AU-ROC. Bold values indicate the best-performing model for each metric.

els, LLaMA 3-7B (few-shot) offered the strongest alternative for both appearance and objectification detection. The NGram Neural Network performed competitively with GPT-4, nearly matching its performance on general appearance and outperforming it in objectification detection. With few-shot settings, most models showed modest gains over their zero-shot counterparts. Only in one case, objectification classification with GPT-4, the few-shot variant led to worse results. This might be explained by the provided examples limiting the models' flexibility or introducing biases. For Mistral, F1 scores were identical for both settings.

Error Analysis. A closer look at misclassified comments revealed that context seems to be essential in determining whether a phrase is objectifying in certain cases. One recurring example that all language models misclassified is the phrase "my little one", which appeared in sentences with varying interpretations. When used by a parent to refer to their own child (e.g., *"My little one also loves these sweets!"*), the phrase does not refer to the child in the video and is thus not objectifying. However, when the unrelated child in the video is referenced (e.g., *"She's so precious, she's my little one"*), it can imply a sense of symbolic possession, fitting within the possessiveness category of objectification. Another example that posed challenges were comments that specifically contained sexualizing language but were not directed at the child. The comment *"Mommy looks so sexy in that dress"*, for example, was classified as positive by all language models, but not by the human annotators. This comment does not qualify as child objectification because the sexualizing language is directed at an adult (the mother), not the child depicted in the video. Although such comments may still be inappropriate in the context of children's content, they do not target the child and therefore fall outside the scope of objectification as defined in our study. This highlights the importance of target awareness in classification; models must not only detect harmful or sexualizing language but also correctly identify who it is directed at.

Objectification Prevalence. We used our finetuned RoBERTa model as the best-performing model to classify the full dataset of 562,508 comments. Through this approach, we found a prevalence of 58,266 comments that were generally related to physical appearance (10.35%) and 16,351 comments that had an objectifying nature (2.90%). On the video level (N = 9,090), each video received an average of 6.41 appearance-related comments and 1.80 objectifying comments.

Out of these comments, we looked at the top 10 most frequent words associated with appearance and objectification (Figure 3) to better understand the language used in classified comments. In general appearance-related comments, *"beautiful"* was the most frequently occurring word, followed by *"pretty," "precious," and "lovely"*. Fashion-related words, such as *"shirt" and "dress"* were also among the top words. The expression *"(s)he looks"* is usually followed by a descriptive adjective like *"cute"*. These terms primarily describe physical appearance and clothing, often in a positive or admiring way, and suggests a strong emphasis on aesthetics. Objectification-related comments, in contrast, contained words such as *"doll", "my little," "mouth," "fit," and "makeup"*. These terms reflect a shift in focus from general appearance to specific body parts or implicit evaluations of attractiveness. *"(s)he looks like"* was usually followed by a description of a famous person (e.g., *"she looks like mini Rihanna"*). Words like *"model"* and *"fit"* were usually used in a context that described

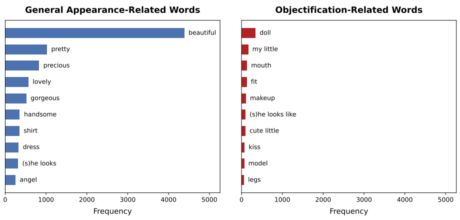

Figure 3: Frequency of appearance- (left) and objectification-related (right) words found in comments.

the child as a future model or complimented the fit of a garment, which suggests comparisons to adult beauty standards. The prevalence of *"kiss"* and *"legs"* indicates a more concerning focus on physical features. The presence of possessive phrases such as *"my little"* also aligns with objectification patterns, as it implies a sense of ownership or control.

5.2 Sociodemographic Characteristics and User Engagement Indicators

The regression models examined the relationship between gender, age, engagement metrics (downloads and total comments), and the predicted likelihood of appearance-related and objectification-related comments. We used RoBERTa-based predicted probabilities as dependent variables to retain classifier uncertainty and avoid distortions from binary thresholding. Results indicate that videos featuring school-aged children received significantly fewer appearance-related comments, especially when the child was male ($\beta = -0.029$, $p = .001$). For objectification-related comments, videos with toddlers showed a slight but significant increase ($\beta = 0.004$, $p = .046$) compared to other age groups, while videos with school-aged boys received fewer such comments ($\beta = -0.012$, $p < .001$). Gender alone and engagement metrics (downloads and comments) did not significantly predict either type of comment (see Table 4).

6 Summary and Discussion

Overall, 10.35% of comments are related to physical appearance and 2.90% of comments are objectifying based on our typology. These comments are embedded in language that centers around beauty, lifestyle, and expressions of cuteness and admiration. RoBERTa consistently achieved the highest F1-scores across both tasks. In contrast, generative LLMs struggled to match its performance. For objectification classification, traditional fine-tuned models appear to outperform large-scale generative approaches, possibly because the task relies less on broad world knowledge and more on recognizing subtle, context-dependent language patterns. School-aged girls received significantly more appearance-related comments than boys of the same age, reflecting gendered patterns observed in earlier studies on social media discourse (Sidani, 2023). This supports broader findings that girls and women are frequently judged based on their appearance (Zurbriggen et al., 2007). Objectification-related comments appeared slightly more often in videos featuring toddlers than other age groups.

The patterns in TikTok comments raise serious concerns about children's online safety. Language that objectifies children might reinforce harmful norms and subject children's appearances to scrutiny, impacting digital well-being and privacy. Exposure to such comments can increase vulnerability among young users, highlighting the need for stronger protections (Gerrard and Thornham, 2020; Gongane et al., 2022). Platforms must enhance content moderation and improve AI detection of inappropriate language. Raising awareness about sharenting could also help reduce unintended exposure and exploitation (Polito et al., 2022; Stephenson et al., 2024).

7 Limitations

Performance of Language Models in Classification Tasks. Across all models, detecting objectification remained challenging with low to moderate F1-scores. However, the F1 scores obtained are consistent with similar setups regarding, for example, implicit hate speech (ElSherief et al., 2021; Li et al., 2024), misogyny (Park and Lee, 2017; Zeinert et al., 2021), and trauma detection (Schirmer et al., 2023, 2024a). The n-gram-based approach performed competitively to LLMs, indicating that simpler linguistic feature-based methods may have value in identifying objectification. Few-shot prompting led to modest performance improvements, which is consistent with prior research on online harm (Agarwal et al., 2023; Nozza, 2021; Pan et al., 2024). However, GPT-4 performed slightly worse in one few-shot setting, and Mistral showed no difference between the two. Improvement through few-shot learning might thus depend on model architecture and task details (Plaza-del Arco et al., 2023). Finally, class imbalance likely impacted performance, with objectifying comments underrepresented in the data and models

Variable	Appearance-Related			Objectification-Related		
	Coef (β)	95% CI	p	Coef (β)	95% CI	p
Intercept	0.133	[0.120, 0.145]	<.001	0.027	[0.023, 0.030]	<.001
Gender (Male)	-0.005	[-0.021, 0.012]	.557	0.002	[-0.003, 0.007]	.479
Age (School)	-0.053	[-0.066, -0.039]	<.001	-0.002	[-0.006, 0.002]	.383
Age (Toddler)	-0.025	[-0.039, -0.011]	.001	0.004	[0.000, 0.009]	.046
Male × School	-0.029	[-0.047, -0.011]	.001	-0.012	[-0.017, -0.006]	<.001
Male × Toddler	-0.006	[-0.025, 0.013]	.543	-0.004	[-0.011, 0.002]	.163
Downloads	-7.51e-08	[-5.96e-07, 4.45e-07]	.777	1.42e-08	[-9.98e-08, 1.28e-07]	.807
Comments	3.78e-07	[-1.16e-06, 1.91e-06]	.630	-1.50e-07	[-3.46e-07, 4.67e-08]	.135
R^2		0.043			0.012	
Adj. R^2		0.042			0.011	
Observations		7,025			7,025	

Table 4: OLS regression results predicting appearance- and objectification-related comments with interaction terms. CI = Confidence Interval.

biased toward the majority class (Buda et al., 2018). Still, representing objectifying comments in their real-world proportion is important for improving model robustness.

Annotation and Data. The subjectivity of annotating objectification-related comments led to only moderate agreement. Given these challenges, it is unreasonable to expect perfect accuracy from language models (Li et al., 2024). However, their ability to detect nuanced patterns at scale may allow them to recognize implicit objectification more consistently than rule-based approaches (Gligorić et al., 2024; Matter et al., 2024; Wang et al., 2024). The dataset was collected from accounts with high follower counts. Therefore, these videos are likely to exhibit higher overall engagement. The observed prevalence of such comments may not be representative of less visible or lower-engagement content on the platform.

Gender and Metadata Differences. Engagement metrics (i.e., the number of downloads and likes) showed no significant association with appearance-related or objectifying comments, suggesting that while engagement may contribute to the visibility of content (Kopecky et al., 2020; Schirmer et al., 2024b), it is not a primary driver of objectifying language. With overall low model fits and small effect sizes for all predictors, the results must be interpreted with caution. They likely capture only a limited part of a more complex interplay involving platform norms, audience composition, and broader social context.

Context and Real-World Implications. To make these findings more generalizable, future research should explore cross-platform comparisons (Horvát and Hargittai, 2021; Matassi and Boczkowski, 2021). Given the psychological and social implications of these findings, further work is needed to assess the real-world impact of such comments on children's self-perception and digital well-being (Garmendia et al., 2022; Ouvrein and Verswijvel, 2019). Additionally, expanding the dataset with richer context, such as including longer comment threads, structured vignettes, or multimodal analysis, could improve model sensitivity to implicit forms of objectification (Chistova and Smirnov, 2022; Muti et al., 2022; Rehman et al., 2025; Schirmer et al., 2025).

Ethics Statement

This study was approved by the ethics committee at the Technical University of Munich. No identifying information, such as account names or individual-level metadata, is shared. Only anonymized comments were included in the paper. While we were prepared to report any material classified as child pornography under German law, we did not come across such content. All annotators were briefed in advance and trained for the task. They had the option to pause or withdraw from the annotation process at any time.

Acknowledgments

Funded by the Bavarian Research Institute for Digital Transformation (bidt), an institute of the Bavarian Academy of Sciences and Humanities, and the Bavarian State Ministry of Science and Art (project "Youth Safety in an AI-Driven Online World").

References

Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2024. Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, 3(7):245.

Vibhor Agarwal, Yu Chen, and Nishanth Sastry. 2023. Haterephrase: Zero-and few-shot reduction of hate intensity in online posts using large language models. *arXiv preprint arXiv:2310.13985*.

Mary Jean Amon, Nika Kartvelishvili, Bennett I. Bertenthal, Kurt Hugenberg, and Apu Kapadia. 2022. Sharenting and children's privacy in the united states: Parenting style, practices, and perspectives on sharing young children's photos on social media. *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, pages 1–30.

Ben Barry, Georgia Wells, James West, Jennifer Stern, and Jeffrey French. 2021. How tiktok serves up sex and drug videos to minors. Wall Street Journal.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.

Philippe Bernard, Sarah J Gervais, Jill Allen, Sophie Campomizzi, and Olivier Klein. 2012. Integrating sexual objectification with object versus person recognition: The sexualized-body-inversion hypothesis. *Psychological Science*, 23(5):469–471.

Philippe Bernard, Sarah J Gervais, and Olivier Klein. 2018. Objectifying objectification: When and why people are cognitively reduced to their parts akin to objects. *European Review of Social Psychology*, 29(1):82–121.

Petre Breazu, Miriam Schirmer, Songbo Hu, and Napoleon Katsos. 2025. Large language models and the challenge of analyzing discriminatory discourse: human-ai synergy in researching hate speech on social media. *Journal of Multicultural Discourses*, pages 1–19.

Anna Brosch. 2016. When the child is born into the internet: Sharenting as a growing trend among parents on facebook. *The New Educational Review*, 43:225–235.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.

Ilaria Cataldo, An An Lieu, Alessandro Carollo, Marc H. Bornstein, Giulio Gabrieli, Albert Lee, and Gianluca Esposito. 2022. From the cradle to the web: The growth of "sharenting"—a scientometric perspective. *Human Behavior and Emerging Technologies*, 2022:1–12.

Elena Chistova and Ivan Smirnov. 2022. Discourse-aware text classification for argument mining. *Computational Linguistics and Intellectual Technologies*, pages 93–105.

Andrew R. Chow. 2025. What happened when India banned TikTok? *Time Magazine*.

Stefanie E Davis. 2018. Objectification, sexualization, and misrepresentation: Social media and the college experience. *Social Media + Society*, 4(3):2056305118786727.

Leyuan Ding, Praboda Rajapaksha, Aung Kaung Myat, Reza Farahbakhsh, and Noel Crespi. 2024. Can hallucination reduction in llms improve online sexism detection? *Intelligent Systems Conference*, pages 625–638.

Nicola Döring and M Rohangis Mohseni. 2019. Male dominance and sexism on youtube: results of three content analyses. *Feminist Media Studies*, 19(4):512–524.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363.

Jasmine Fardouly and Lenny R Vartanian. 2016. Social media and body image concerns: Current research and future directions. *Current Opinion in Psychology*, 9:1–5.

Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring misogyny across the manosphere in reddit. *Proceedings of the 10th ACM Conference on Web Science*, pages 87–96.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

David A. Frederick, Eva Pila, Vanessa L. Malcarne, Emilio J. Compte, Jason M. Nagata, Cassidy R. Best, Catherine P. Cook-Cottone, Tiffany A. Brown, Lexie Convertino, Canice E. Crerand, Michael C. Parent, Jamie-Lee Pennesi, Marisol Perez, Rachel F. Rodgers, Lauren M. Schaefer, J. Kevin Thompson, Tracy L. Tylka, and Stuart B. Murray. 2022. Demographic predictors of objectification theory and tripartite influence model constructs: The US Body Project I. *Body Image*, 40:182–199.

Barbara L Fredrickson and Tomi-Ann Roberts. 1997. Objectification theory: Toward understanding women's lived experiences and mental health risks. *Psychology of Women Quarterly*, 21(2):173–206.

Maialen Garmendia, Gemma Martínez, and Carmelo Garitaonandia. 2022. Sharenting, parental mediation and privacy among spanish children. *European Journal of Communication*, 37(2):145–160.

Ysabel Gerrard and Helen Thornham. 2020. Content moderation: Social media's sexist assemblages. *New Media & Society*, 22(7):1266–1286.

Peter Glick, Jeffrey Diebold, Barbara Bailey-Werner, and Lin Zhu. 1997. The two faces of adam: Ambivalent sexism and polarized attitudes toward women. *Personality and Social Psychology Bulletin*, 23(12):1323–1334.

Peter Glick and Susan T Fiske. 2018. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. In Susan T Fiske, editor, *Social Cognition*, pages 116–160. Routledge.

Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel J Candès, and Dan Jurafsky. 2024. Can unconfident llm annotations be used for confident conclusions? *arXiv preprint arXiv:2408.15204*.

Vaishali U Gongane, Mousami V Munot, and Alwin D Anuse. 2022. Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*, 12(1):129.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Emőke-Ágnes Horvát and Eszter Hargittai. 2021. Birds of a feather flock together online: Digital inequality in social media repertoires. *Social Media+ Society*, 7(4):20563051211052897.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 588–602.

Xing Jiang, Hugo Touvron, Ludovic Denoyer, Hervé Jégou, and Guillaume Lample. 2023. Mistral 7b: A small model that thinks big. *arXiv preprint arXiv:2310.06825*.

Md Masudul Islam Khan and Himel Bhattacharjee. 2022. A new avenue of crime in bangladesh: TikTok as a weapon of violence against women. *1st International Conference of Social Sciences on Bangladesh*, 50.

Hye Min Kim. 2021. What do others' reactions to body posting on instagram tell us? the effects of social media comments on viewers' body image perception. *New Media & Society*, 23(12):3448–3465.

Kamil Kopecky, Rene Szotkowski, Inmaculada Aznar-Díaz, and José-María Romero-Rodríguez. 2020. The phenomenon of sharenting and its risks in the online environment: Experiences from czech republic and spain. *Children and Youth Services Review*, 110:104812.

György Kovács, Pedro Alonso, and Rajkumar Saini. 2021. Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources. *SN Computer Science*, 2(2):95.

Alexandra S. Levine. 2022. These tiktok accounts are hiding child sexual abuse material in plain sight. Forbes.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2):1–36.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Miriam Martínez Allué and Miguel Ángel Martín Cárdaba. 2024. "Kidfluencers" Children influencers on YouTube and TikTok and their impact on the child audience. *VISUAL REVIEW. International Visual Culture Review Revista Internacional De Cultura Visual*, 16(5):261–270.

Mora Matassi and Pablo Boczkowski. 2021. An agenda for comparative social media studies: The value of understanding practices from cross-national, cross-media, and cross-platform perspectives. *International Journal of Communication*, 15:22.

Daniel Matter, Miriam Schirmer, Nir Grinberg, and Jürgen Pfeffer. 2024. Investigating the increase of violent speech in incel communities with human-guided gpt-4 prompt iteration. *Frontiers in Social Psychology*, 2:1383152.

Bonnie Moradi and Yu-Ping Huang. 2008. Objectification theory and psychology of women: A decade of advances and future directions. *Psychology of Women Quarterly*, 32(4):377–398.

Arianna Muti, Katerina Korre, and Alberto Barrón-Cedeño. 2022. Unibo at semeval-2022 task 5: A multimodal bi-transformer approach to the binary and fine-grained identification of misogyny in memes. *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 663–672.

Niamh Ní Bhroin, Thuy Dinh, Kira Thiel, Claudia Lampert, Elisabeth Staksrud, and Kjartan Ólafsson. 2022. The privacy paradox by proxy: Considering predictors of sharenting. *Media and Communication*, 10(1):371–383.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech 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 2: Short Papers)*, pages 907–914.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 149–155.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and et al. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Gaëlle Ouvrein and Karen Verswijvel. 2019. Sharenting: Parental adoration or public humiliation? a focus group study on adolescents' experiences with sharenting against the background of their own impression management. *Children and Youth Services Review*, 99:319–327.

Ronghao Pan, José Antonio García-Díaz, and Rafael Valencia-García. 2024. Comparing fine-tuning, zero and few-shot strategies with large language models in hate speech detection in english. *CMES-Computer Modeling in Engineering & Sciences*, 140(3).

Hyanghee Park and Joonhwan Lee. 2017. Do private and sexual pictures receive more likes on instagram? *International Conference on Research and Innovation in Information Systems (ICRIIS)*, pages 1–6.

Flor Miriam Plaza-del Arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. *The 7th Workshop on Online Abuse and Harms (WOHA), Association for Computational Linguistics*.

Vinícius Polito, George Valença, Maria Wanick Sarinho, Fernando Lins, and Rodrigo Pereira dos Santos. 2022. On the compliance of platforms with children's privacy and protection requirements: An analysis of tiktok. *International Conference on Software Business*, pages 85–100.

Mohammad Zia Ur Rehman, Sufyaan Zahoor, Areeb Manzoor, Musharaf Maqbool, and Nagendra Kumar. 2025. A context-aware attention and graph neural network-based multimodal framework for misogyny detection. *Information Processing & Management*, 62(1):103895.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas Pykl, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and misogyny detection using bert: A multi-task approach. *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131.

Miriam Schirmer, Tobias Leemann, Gjergji Kasneci, Jürgen Pfeffer, and David Jurgens. 2024a. The language of trauma: Modeling traumatic event descriptions across domains with explainable ai. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13224–13242.

Miriam Schirmer, Isaac Misael Olguín Nolasco, Edoardo Mosca, Shanshan Xu, and Jürgen Pfeffer. 2023. Uncovering trauma in genocide tribunals: An nlp approach using the genocide transcript corpus. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 257–266.

Miriam Schirmer, Jürgen Pfeffer, and Sven Hilbert. 2025. Talking about torture: A novel approach to the mixed methods analysis of genocide-related witness statements in the khmer rouge tribunal. *Journal of Mixed Methods Research*, 19(1):83–102.

Miriam Schirmer, Angelina Voggenreiter, and Jürgen Pfeffer. 2024b. More skin, more likes! measuring child exposure and user engagement on tiktok. *arXiv preprint arXiv:2408.05622*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics*, pages 1–10.

Karen Sidani. 2023. The hypersexualization of young girls and the infantilization of adult women. *American Journal of Humanities and Social Sciences Research*, 7(1):193–197.

Domonkos Sik, Renáta Németh, and Eszter Katona. 2023. Topic modelling online depression forums: beyond narratives of self-objectification and self-blaming. *Journal of Mental Health*, 32(2):386–395.

Mariana Silva. 2019. Video app TikTok fails to remove online predators. BBC.

Encarnación Soriano-Ayala, María Bonillo Díaz, and Verónica C. Cala. 2023. TikTok and child hypersexualization: Analysis of videos and narratives of minors. *American Journal of Sexuality Education*, 18(2):210–230.

Statista. 2024. Number of TikTok users worldwide from 2020 to 2025.

Statista. 2025. Countries with the largest TikTok audience as of February 2025.

Benjamin Steel, Miriam Schirmer, Derek Ruths, and Juergen Pfeffer. 2025. Just another hour on tiktok: Reverse-engineering unique identifiers to obtain a complete slice of tiktok. *arXiv preprint arXiv:2504.13279*.

Sophie Stephenson, Christopher Nathaniel Page, Miranda Wei, Apu Kapadia, and Franziska Roesner. 2024. Sharenting on tiktok: Exploring parental sharing behaviors and the discourse around children's online privacy. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Rebeca Suárez-Álvarez, Antonio García-Jiménez, and María L. Urbina Montana. 2023. Sexualising characteristics of adolescent on tiktok: Comparative study great britain–spain. *Convergence*, 29(5):1262–1282.

TikTok. 2024. Community guidelines. TikTok.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Sharan Batra, Prajjwal Bhargava, Shruti Bhosale, and et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jennifer Valentino-DeVries and Michael H. Keller. 2024. A marketplace of girl influencers managed by moms and stalked by men. The New York Times.

Elisabeth Van den Abeele, Ini Vanwesenbeeck, and Liselot Hudders. 2024. Child's privacy versus mother's fame: unravelling the biased decision-making process of momfluencers to portray their children online. *Information, communication & society*, 27(2):297–313.

Karen Verswijvel, Michel Walrave, Kris Hardies, and Wannes Heirman. 2019. Sharenting, is it a good or a bad thing? Understanding how adolescents think and feel about sharenting on social network sites. *Children and Youth Services Review*, 104:104401.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE*, 15(12):e0243300.

Michel Walrave, Karen Verswijvel, Gaëlle Ouvrein, Luna Staes, Lara Hallam, and Kris Hardies. 2022. The limits of sharenting: Exploring parents' and adolescents' sharenting boundaries through the lens of communication privacy management theory. *Frontiers in Education*, page 803393.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *Proceedings of the First Workshop on Abusive Language Online, Association for Computer Linguistics*, pages 78—-84.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1415–1420.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*, pages 3181–3197.

Eileen L Zurbriggen, Rebecca L Collins, Sharon Lamb, Tomi-Ann Roberts, Deborah L Tolman, and L Monique Ward. 2007. APA task force on the sexualization of girls. *American Psychological Association*.

A Appendix

A.1 Overview of Classification Pipeline

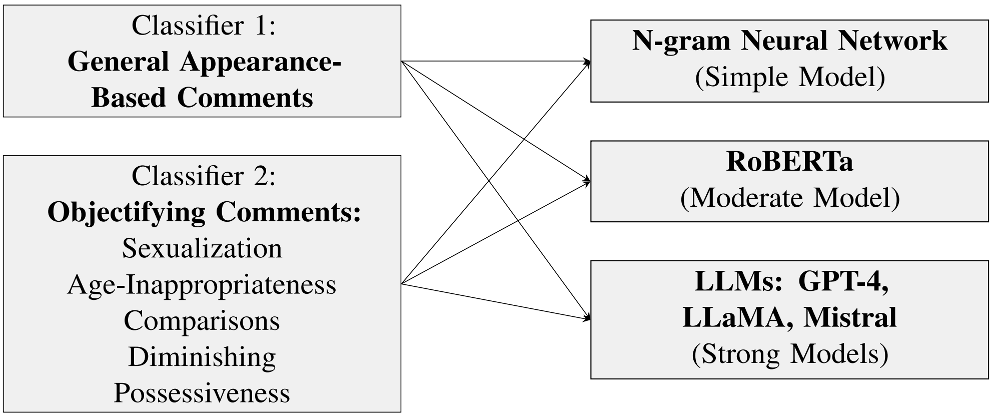

Figure 4: Overview of Classification Pipeline for Appearance-Based and Objectifying Comments

A.2 Topic Modeling Details

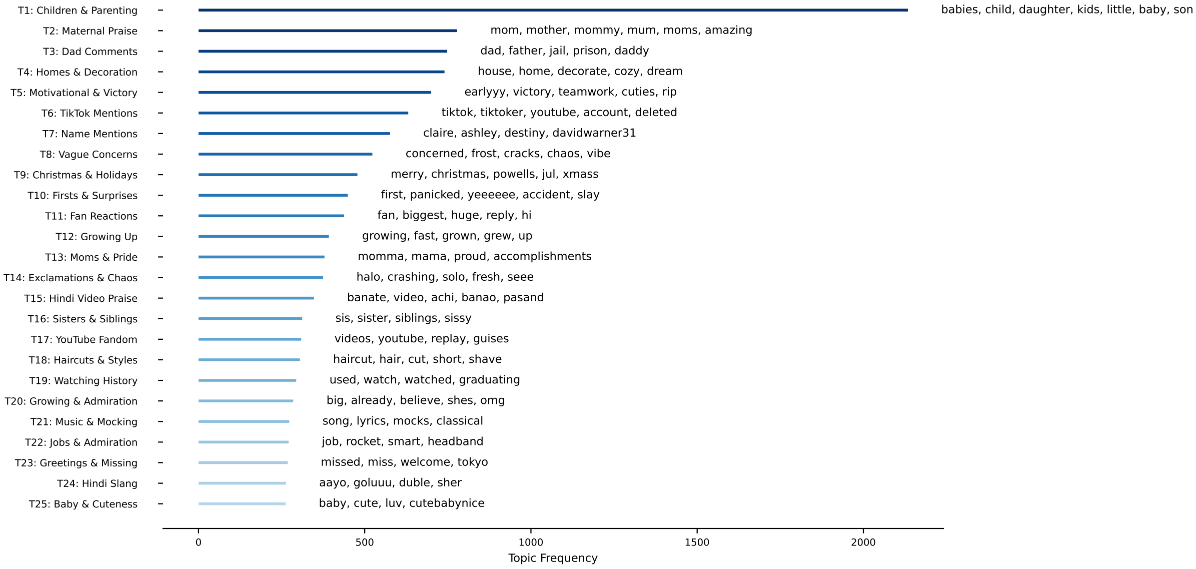

Figure 5: Top 25 meaningful topics with their corresponding representative words and frequency.

A.3 Video Classifier Details

After having collected the first 100 videos from accounts that generally featured children in some of the videos, we had to decide if each single video actually contained a child. Due to the high number of videos, we trained a neural network classifier to do this. Our training data set consisted of approximately 12,000 manually labeled videos that were split into a train and test set (80:20 ratio). We employed a Multi-Layer Perceptron (MLP) classifier trained on video embeddings extracted using the CLIP model (`openai/clip-vit-base-patch32`). The classifier consisted of an input layer matching the embedding size, a hidden layer with ReLU activation, a dropout layer to prevent overfitting, and an output layer for binary classification. The final class label was determined using a softmax activation function. The detailed model architecture and hyperparameters are provided in Table 5. For feature extraction, videos were processed using CLIP. Each video was represented by a set of 12 evenly spaced frames, with each frame passed through CLIP to obtain a 512-dimensional embedding. The mean embedding across all frames was computed to obtain a single feature vector per video, which served as input to the classifier. The classifier was trained using supervised learning with a 5-fold cross-validation strategy. The Adam optimizer was used with weight decay to improve generalization, and learning rate scheduling was applied to adjust training dynamics. A batch size of 512 was used, and training was conducted for 20 epochs on a GPU-enabled environment.

Parameter	**Value**
Model Architecture	MLP
Input Size	512 (CLIP embedding)
Hidden Size	256
Dropout Rate	0.5
Output Size	2 (Binary classification)
Activation Function	ReLU
Optimizer	Adam
Learning Rate	0.001
Weight Decay	1e-4
Loss Function	Cross-Entropy Loss
Batch Size	512
Epochs	20
Learning Rate Scheduler	ReduceLROnPlateau (patience = 3)
Cross-Validation	5-Fold
Validation Split	20%
Frames per Video	12
Feature Extractor	CLIP (‘openai/clip-vit-base-patch32‘)
Frame Aggregation	Mean embedding

Table 5: Hyperparameter configuration of the MLP classifier.

A.4 Geographical Distribution of Accounts

The top five countries represented in the dataset are the United States, which account for a third of all accounts (33.7%), followed by India with nearly nine percent (8.61%),[3] Indonesia with almost eight percent (7.69%), Brazil with seven percent (7.14%), and Mexico with five percent (5.31%). This distribution roughly aligns with TikTok's global user base, with the largest TikTok user bases being in Indonesia, the United States, and Brazil (Statista, 2025).

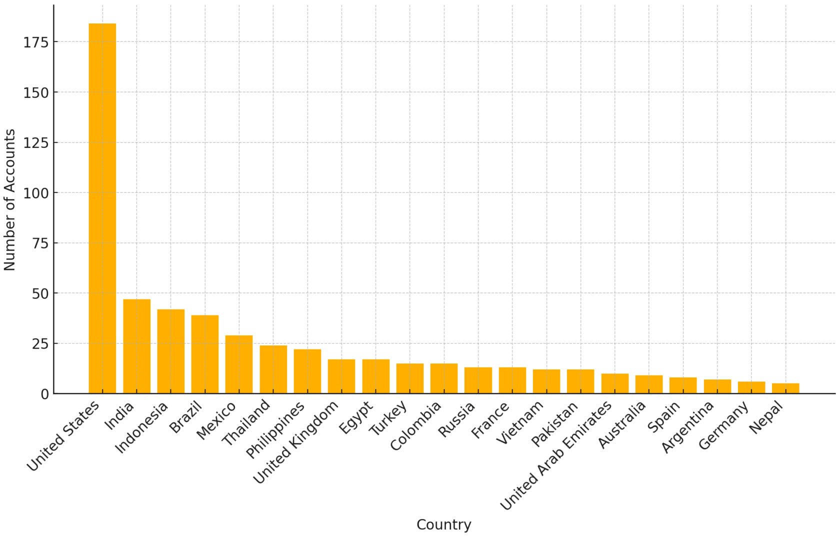

Figure 6: Country distribution of accounts included in our dataset.

[3] India banned TikTok in 2020. Before the ban, it had nearly 200 million users, making it one of the platform's largest markets (Chow, 2025). The high share of accounts from India in this dataset may be due to accounts operating through VPNs and videos that had been uploaded before the ban.

A.5 Hyperparameter Details for Language Models

Model	General Appearance	Objectification
NGramNN	n_gram_range: [1, 2, 3] hidden_dim: 128 dropout_rate: 0.4 lr: 0.001 batch_size: 32 epochs: 10	n_gram_range: [1, 2, 3] hidden_dim: 128 dropout_rate: 0.3 lr: 0.001 batch_size: 32 epochs: 10
RoBERTa (finetuned)	pretrained: roberta-base total layers: 12 trainable layers: 3 dropout_rate: 0.3 lr: 2.00e-05 epochs: 3	pretrained: roberta-base total layers: 12 trainable layers: 3 dropout_rate: 0.3 lr: 2.00e-05 epochs: 3
OpenAI GPT-4	target_model: gpt-4-turbo temperature: 0.1	target_model: gpt-4-turbo temperature: 0.1
Mistral 7B	temperature: 0.1 max_tokens: 512	temperature: 0.1 max_tokens: 512
LLaMA 3	temperature: 0.1 max_tokens: 512	temperature: 0.1 max_tokens: 512

Table 6: Hyperparameters for each model and task.

A.6 Prompts Used For LLM Comment Classification

General Appearance Comments (Zero-Shot)

```
Identify all comments that describe the physical appearance of the child in the video, including
those that consist only of emojis. The comments should specifically refer to the child shown and
include observations about their appearance. Only include comments that are positive or neutral in
tone. Ignore any comments that do not reference the child's physical appearance or visual traits.
Include comments that can be used to describe appearance. Take note that these are TikTok comments,
so be aware of TikTok-specific language.
- 1 = Describes physical appearance or body parts or outfit-related things.
- 0 = Does not describe physical appearance (e.g., behavior, voice, actions).
Comment: "comment"
Classification (ONLY return 0 or 1):
```

General Appearance Comments (Few-Shot)

```
Identify all comments that describe the physical appearance of the child in the video, including
those that consist only of emojis. The comments should specifically refer to the child shown and
include observations about their appearance. Only include comments that are positive or neutral in
tone. Ignore any comments that do not reference the child's physical appearance or visual traits.
Include comments that can be used to describe appearance. Take note that these are TikTok comments,
so be aware of TikTok-specific language.
- 1 = Describes physical appearance or body parts or outfit-related things.
- 0 = Does not describe physical appearance (e.g., behavior, voice, actions).
Examples:
- "She looks so cute" → 1
- "I keep wondering who he looks like. He looks like Harry Potter." → 1
- "The way she says ball is so cute!" → 0
Comment: "comment"
Classification (ONLY return 0 or 1):
```

Objectifying Comments (Zero-Shot)

```
Identify comments that describe the physical appearance of the child in the video, including those
consisting only of emojis. Focus on language that may be objectifying, overly personal, possessive,
sexualizing, diminishing, or otherwise inappropriate in tone.  Take note that these are TikTok
comments, so be aware of TikTok-specific language. Exclude comments that are unrelated to the child.
- 1 = Contains objectifying or inappropriate language.
- 0 = Does not contain objectifying or inappropriate language.
Comment: ”comment”
Classification (ONLY return 0 or 1):
```

Objectifying Comments (Few-Shot)

```
Identify comments that describe the physical appearance of the child in the video, including those
consisting only of emojis. Focus on language that may be objectifying, overly personal, possessive,
sexualizing, diminishing, or otherwise inappropriate in tone.  Take note that these are TikTok
comments, so be aware of TikTok-specific language. Exclude comments that are unrelated to the child.
- 1 = Contains objectifying or inappropriate language.
- 0 = Does not contain objectifying or inappropriate language.
Examples:
- ”What a cute little princess” → 1
- ”I Love You SO SO Much, Can I Babysit?” → 1
- ”She looks adorable today!” → 0
Comment: ”comment”
Classification (ONLY return 0 or 1):
```