Keep It PG or Let It Go?

Exploring the Presence of (In)appropriateness in YouTube Videos for Young Children and Opportunities for Safeguarding

J.J.P. de Water

TU Delft

Keep It PG or Let It Go?

Exploring the Presence of (In)appropriateness in YouTube Videos for Young Children and Opportunities for Safeguarding

Thesis report

by

J.J.P. de Water

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended publicly on 30-04-2025

Thesis committee:

Chair:	Dr. Sole Pera
Supervisor:	Dr. Sole Pera & Robin Ungruh
External examiner:	Dr. Cynthia Liem
Place:	Faculty of Electrical Engineering, Mathematics, Computer Science, Delft
Project Duration:	04-09-2024 - 30-04-2025
Student number:	4661109

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

Faculty of EEMCS · Delft University of Technology

Abstract

As online video platforms like YouTube and YouTube Kids continue to shape young children's daily media use, concerns about their exposure to inappropriate content persist. While these platforms implement various safeguards to protect young audiences, inappropriate videos continue to surface in search results and next-video recommendations, sometimes even at the top of the list. This study explores how metadata-derived features can be used to identify and address such content. Drawing on a ground-truth dataset of YouTube videos labeled for toddler appropriateness, we conduct a detailed feature analysis to uncover patterns linked to (in)appropriateness and train a classifier capable of predicting video appropriateness for young children. Building on these insights, we develop and evaluate score-based reranking strategies designed to reduce exposure to inappropriate videos while promoting age-appropriate content. Our findings show that metadata-informed reranking significantly improves the prioritization of suitable content, raising HitRate@1 of suitable videos from 14% to as high as 62%, but also reveal critical trade-offs: misclassified inappropriate videos, particularly when predicted with high confidence, may still appear in top-ranked positions. As such, detection and reranking methods like ours represent a first step that warrants further steps for safer recommendation environments for young children. This research provides a practical framework for improving recommendation outcomes and contributes to the broader conversation on designing safer, more transparent, and child-centered media systems.

Preface

This thesis marks the final step in my Master's journey at Delft University of Technology. Throughout this research, I explored the nuanced and at times unsettling landscape of YouTube recommendations for young children, driven by the realization that even platforms designed to be safe can fall short in complex, subtle ways.

I am incredibly grateful to my supervisors, Dr. Sole Pera and Robin Ungruh, for their continued guidance, critical feedback, and inspiring conversations. Their expertise not only sharpened this research but made it infinitely more enjoyable. I would also like to thank my external examiner, Dr. Cynthia Liem, for her thoughtful interest in this project.

Special thanks go out to my girlfriend, friends, family, and peers, who provided both technical insight and moral support, especially during moments when "letting it go" was not an option. To those who reviewed early drafts, challenged my assumptions, or reminded me to step away from my screen: thank you.

I also want to express my heartfelt gratitude to my parents, whose unwavering support has carried me not just through this thesis, but through all my years of studying. You have always encouraged me to follow my interests and pursue my goals, no matter how ambitious. Thank you for believing in me, and for standing beside me every step of the way.

Finally, a personal thanks to every snowman, reindeer, and curious toddler that helped shape this thesis in spirit, tone, and structure.

Delft, April 2025
J.J.P. de Water

Contents

List of Figures

List of Tables

1

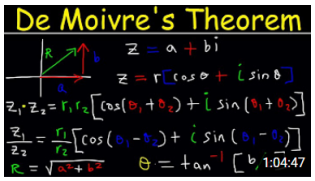
For the First Time in Forever: Introduction

Online video platforms, such as YouTube and YouTube Kids, have become a central part of young children's daily entertainment. In 2015, Google introduced YouTube Kids for young children aged 0-12 [1]. YouTube Kids is specifically designed for a young audience and aims to provide safe and age-appropriate content. By 2021, over 35 million users across more than 80 countries viewed YouTube Kids videos weekly [2]. The YouTube Kids mobile app reached 99.4 million downloads in that same year [3]. This trend continued to grow, with the app reaching 103.21 million downloads globally in 2022 and 131.35 million in 2023. Reflecting this growth, recent findings from Qustodio's[1] 2023-2024 Annual Data Report[2], which provides insights into children's online app habits on mobile and desktop devices, highlights how streaming video content has become a dominant activity for young children worldwide [4]. The data from over 400,000 families with children aged 4-18 reveals that children spent, on average, 57 minutes per day streaming video content. Notably, 63% of the children watched YouTube with an average of 70 minutes per day, while 6% of them spent an astounding average of 96 minutes per day watching YouTube Kids. This widespread and frequent engagement highlights the significant presence of YouTube and YouTube Kids in young children's daily lives.

YouTube and YouTube Kids have a large library of content that is **appropriate** for young users, i.e., videos that are specifically made for them and could have a positive impact on their development [5] or those that are not directly interesting to them but cannot be categorized as problematic. Examples of such appropriate videos can be found in Figure 1.1. Well-planned and developmentally appropriate videos can support social-emotional, language, and academic development in young children [5]. These videos often use child-directed speech, slower pacing, and clear learning goals to engage young viewers in ways that align with their cognitive and emotional needs. However, on both platforms **inappropriate** videos surface frequently [6]. Such videos typically provide content that is not age-appropriate and can oftentimes be described as harmful.

Sesame Street: Sing the Alphabet Song! | Sesame Street Alphabet

(a) Sesame Street video



Complex Numbers In Polar - De Moivre's Theorem

(b) De Moivre's Theorem video

Figure 1.1: Appropriate videos for young children

Exposure to inappropriate videos can negatively impact children's well-being and development in various ways. Repeated exposure to violent content may desensitize children and increase the risk of

[1] Qustodio is one of the global leaders in online safety and digital well-being for families.

[2] https://www.qustodio.com/en/born-connected-rise-of-the-ai-generation/online-video-qustodio-annual-data-report-2023/

aggressive behavior [7], while early exposure to sexual content has been linked to problematic sexual behaviors and psychological distress, including shame, confusion, and anxiety [8, 9]. Exposure to horror or frightening content can provoke prolonged fear reactions, such as nightmares, sleep disturbances, and avoidance behaviors, potentially disrupting emotional stability and daily routines [10]. Given these risks, it is crucial to carefully manage and moderate the content accessible to children on online video platforms.

To address this ongoing issue, YouTube has recently implemented various moderation and filtering techniques. These include machine learning methods to detect inappropriate content [11], a strike system to apply up to three strikes to a channel that breaks community guidelines [12, 13], which explicitly prohibit harmful, dangerous, or disturbing content, and the introduction of age groupings, as shown in Figure 1.2, which aims to tailor recommendations to a child's developmental stage by segmenting content for preschool, younger, and older children [14]. Most recently, they introduced the flag *made for kids*, which allows creators to declare whether their content is appropriate for children or not [15, 16, 17]. The introduction of the *made for kids* flag followed YouTube's settlement with the Federal Trade Commission (FTC) after the platform was found in violation of the Children's Online Privacy Protection Act (COPPA) [18, 19]. Despite these efforts, recent research shows the issue persists [20, 21, 22, 23].

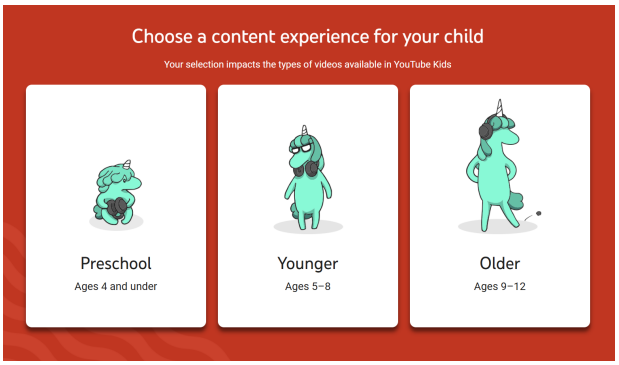
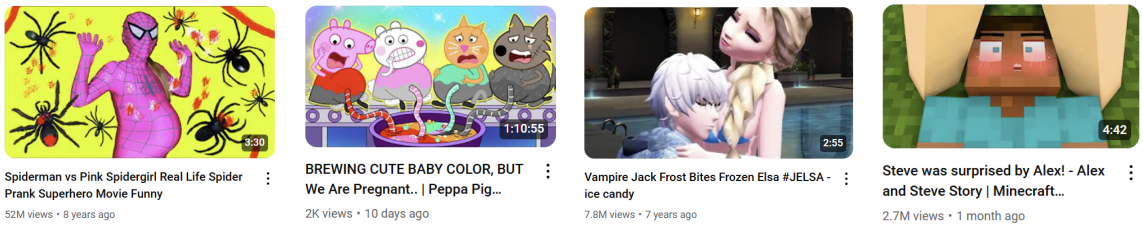


Figure 1.2: Age group content settings as found on `https://www.youtubekids.com/`

YouTube and YouTube Kids rely on recommender algorithms and machine learning models to filter and rank videos and personalize next-video suggestions [24]. Similar to YouTube's decision-making algorithms, these strategies rely heavily on video metadata [25, 26]. Due to the reliance on these algorithms, the platforms need to rely on safeguarding mechanisms or manual tagging by users. However, safeguards within these systems often fail to account for nuanced differences in disturbing content, leading to the exposure of children to videos that are inappropriate to their age group or do not pertain to their interests and skills [24, 27, 28, 21, 29, 30, 22]. These videos either evade detection by the platform's filtering algorithms or are incorrectly categorized due to the system's lack of detailed content analysis [20].

Disturbingly, some of these inappropriate videos are intentionally designed to evade platform safeguards. Often referred to in the context of the *Elsagate* phenomenon[3], these videos use popular cartoon characters, e.g., Elsa or Spiderman, to attract young viewers, but contain inappropriate themes, e.g. violence, horror, or sexual connotation, all with the intent to appear child-friendly while circumventing detection of them being unsuitable for young audiences [6, 31]. The examples in Figure 1.3 illustrate several characteristics used to attract young viewers while avoiding detection by platform safeguards. The videos feature disturbing themes disguised with familiar, child-friendly characters, i.e., Spiderman, Peppa Pig, Elsa, and Steve, creating a deceptive visual appeal. The seemingly playful thumbnails use vibrant colors to hide graphic or unsettling content, making them appear innocent at first glance, but inappropriate upon closer inspection (cf. Figure 1.3a and Figure 1.3b). Despite these inappropriate elements, many of these videos have remained on the platform for years, gaining millions of views, demonstrating the alarming persistence and popularity of such content (cf. Figures 1.3a and 1.3c). Additionally, recent uploads have shown rapid growth in viewership, with new videos quickly reaching large audiences despite their explicit nature (cf. Figure 1.3b and Figure 1.3d).

[3]The *Elsagate* phenomenon refers to a wave of videos that emerged on YouTube and YouTube Kids, featuring popular children's characters, such as Elsa from *Frozen* or Spiderman, in inappropriate scenarios [6].



(a) Spiderman video **(b)** Peppa Pig video **(c)** Frozen video **(d)** Minecraft video

Figure 1.3: Examples of inappropriate videos intentionally created to evade safeguards

Other videos may not be intentionally created to target young children and evade safeguards, but they are still unsuitable for young audiences. These videos may be educational or entertaining for their intended, older audiences, i.e., school-aged children (aged 6-11 years), adolescents (aged 12-17 years), or adults (aged 18+ years), but they are not fitting for younger viewers due to their complexity or themes, or simply due to being irrelevant to their interests [24]. The examples in Figure 1.4, available on YouTube Kids with the *Preschool* (i.e., ages 0-4) content setting on, visualize those aspects that might deem videos unfitting for younger age groups. Videos might simply be too complex to understand or irrelevant for young audiences (cf. Figure 1.4b and Figure 1.4c), or they might explore aspects that, while educational, are too advanced or potentially unsettling (cf. Figure 1.4a and Figure 1.4d). Such unfitting content can be impressionable to children since its impact can vary, from being confusing to inducing fear or anxiety.

(a) Inappropriate theme: reproduction **(b)** Irrelevant complexity: complex numbers **(c)** Irrelevant interest: stock splits **(d)** Inappropriate theme: death

Figure 1.4: Examples of videos available on YouTube Kids

While YouTube and similar platforms have implemented automated safeguards to protect young children from harmful or developmentally inappropriate content, these systems often fall short. Inappropriate videos continue to surface in young children's recommendations, either by evading detection or by being unintentionally recommended due to metadata signals that do not reliably align with child-appropriateness. The limitations of these safeguards stem from their general-purpose design, which is not tailored to the specific developmental needs and vulnerabilities of young children. This highlights a broader need for child-centered improvements in both content classification and recommendation strategies, approaches that not only block harmful content more effectively but also promote content that aligns with the cognitive, emotional, and social needs of young audiences.

Addressing this need requires more than simply improving filtering systems; it demands a deeper understanding of how different types of inappropriateness manifest within children's content. Despite the growing concern about inappropriate content, YouTube's existing safeguards struggle to differentiate between videos that can impact children positively as well as negatively in nuanced ways and with varying degrees of severity. The inability to make these distinctions results in current safeguards either failing to detect inappropriate videos or mistakenly filtering out content that is suitable.

Optimally, we argue that video recommendations for children should exclusively contain content that is *appropriate*; this includes videos with age-appropriate content created for children and videos with content not directly interesting to young viewers but posing no harm. However, beyond simply ensuring appropriateness, recommender systems should prioritize *truly* age-appropriate content. Well-planned and

developmentally appropriate videos can support social-emotional, language, and academic development in young children [5], making it essential that high-quality, engaging, and beneficial content is ranked higher over less relevant alternatives.

Achieving this requires a precise classification system that captures a nuanced understanding of appropriateness, one that moves beyond a binary distinction of appropriate versus inappropriate to account for the varying ways content may align with young children's needs. Without a clear differentiation between videos that deceptively mimic child-friendly content, videos that contain mature themes unsuitable for children but were never intended for them in the first place, and videos that, though not harmful, may still disrupt a child's viewing experience by failing to align with their developmental stage, recommender systems cannot effectively reduce exposure to harmful content or ensure that beneficial content remains accessible and well-ranked. Ensuring these classifications are accurate is crucial not only for protecting young children from harmful content but also for improving the quality of video recommendations tailored to their needs. This leads us to our first Research Question (RQ):

Research Question 1

What key features derived from video metadata characterize inappropriate content in young children's online video recommendation?

To address this question, we conduct a feature analysis on features derived from video metadata, including textual metadata (e.g., titles, descriptions, and tags) and engagement metrics (e.g., view count and video likes/dislikes), to identify patterns that distinguish appropriate from inappropriate content. These Metadata-derived features are particularly valuable in this context because they are readily available, do not require processing of raw video or audio, and can be scaled to large datasets. Moreover, metadata offers structured and interpretable signals, making it well-suited for the distribution and correlation analyses applied in this study to examine feature relationships with content labels. The features in this study are selected based on their accessibility, their ease-of-use, their distributions, and their observed correlations with content appropriateness for young children.

Beyond their practicality, metadata-derived features offer a rich source of implicit information. Textual metadata often reflects how videos are marketed or framed for children, including the use of keywords like "for kids" or popular character names, tactics frequently observed in deceptive or inappropriate content targeting young audiences [28, 24]. Engagement metrics, such as view counts and like-to-dislike ratios, provide behavioral signals that can highlight anomalous content [32, 33]. Moreover, metadata is also often one of the few real-time accessible inputs for content moderation and remains highly interpretable, making it especially useful for moderation workflows that require transparency and human oversight [25]. Together, these properties make metadata an effective and deployable foundation for identifying patterns that correlate with inappropriateness in ways that are both explainable and deployable. Based on these motivations, we extract the most relevant features for classification and develop a classifier model that predicts video appropriateness for young children.

While classification is essential for understanding inappropriateness, it does not directly address how content is ranked and recommended. Even if inappropriate videos are identified, they may still appear in recommendations unless the recommender strategy explicitly incorporates this information. To mitigate young children's exposure to inappropriate content, recommender systems must integrate these insights into their ranking strategy. This leads us to our second Research Question:

Research Question 2

To what degree can features deemed relevant for predicting inappropriateness contribute to the mitigation of young children's exposure to inappropriate content by recommender systems?

To address this question, we design and evaluate a set of score-based reranking strategies that incorporate insights from our feature analysis and classifier. These include classifier-based, categoryId-based, and viewCount-based reranking approaches, as well as two fusion methods that combine their signals (i.e., CombMNZ and CombSUM). Our goal is to assess how these strategies can reduce exposure to

inappropriate content while also promoting videos that are truly age-appropriate without compromising the overall quality of recommendations for young children.

Through this research, we present a structured approach to improving content safety in video platforms serving young audiences by distinguishing between appropriate and inappropriate videos based on metadata-derived features. We show how video-level classification can identify nuanced signals of inappropriateness using interpretable, scalable inputs. Building on this, we integrate classification outputs and metadata signals into score-based reranking strategies that reshape recommendation lists to promote truly age-appropriate content and demote inappropriate content. Together, these methods offer a layered framework that reflects the complexity of online content moderation and provides a practical foundation for safer, developmentally aligned video recommendations for young children.

These findings are intended to encourage reflection among online video platforms on the limitations of current content moderation and recommendation mechanisms. By highlighting the challenges of existing safeguards, our work highlights the need for more adaptive systems that can better differentiate between various types of inappropriateness. Platforms that rely on automated filtering and recommendation should critically assess how their systems handle nuanced cases of inappropriateness and consider integrating more sophisticated classification and recommender techniques to improve content safety.

To ensure transparency and reproducibility, the full codebase used for both experiments is made publicly available at: `https://github.com/JoeydeW/KeepItPGorLetItGo`.

The remainder of this manuscript is organized as follows: an overview of related work is presented in Chapter 2. The feature analysis and classifier, which address RQ1 by identifying metadata-derived signals of inappropriateness, are described in Chapter 3. The reranking strategies, which address RQ2 by applying these insights to mitigate exposure to inappropriate content, are presented in Chapter 4. We address the ethical implications of our research in Chapter 5. Ending with the discussion in Chapter 6 and conclusion in Chapter 7.

2

Into the Unknown: Literature Background

From the comforting glow of morning cartoons to endless autoplay, the digital media landscape young children navigate today is vast, algorithmically curated, and not without its dangers. In this chapter, we situate our study within prior research on children's engagement with video platforms, particularly YouTube and YouTube Kids. We explore the dynamics of how inappropriate content infiltrates these platforms, the risks posed by inappropriate content, and the technological strategies developed to detect and mitigate these risks.

In Section 2.1, we examine the rise of YouTube and YouTube Kids as pivotal platforms in shaping the media consumption of young children, focusing on the dual aspects of their potential to foster learning and the risks posed by exposure to inappropriate content. To better understand these risks, we examine the potential impacts of such exposures, detailing how they might affect the cognitive, emotional, or behavioral development of young viewers. In Section 2.2, we lay out the structural challenges and vulnerabilities of YouTube's content moderation pipeline and recommender algorithm, with a focus on why inappropriate content can still be surfaced despite existing safeguards.

In Section 2.3, we provide an overview of metadata-informed approaches to detect video inappropriateness. Furthermore, we examine classification-based efforts to distinguish harmful from benign content. In Section 2.4, we explore the role of recommender systems in shaping children's viewing experiences and existing proposals for ranking or filtering content aimed at children. Finally, in Section 2.5, we provide an overview of regulatory frameworks and ethical concerns, including platform responsibility, algorithmic transparency, and the tension between engagement incentives and child safety.

2.1. YouTube Videos for Young Children

2.1.1. The Rise of YouTube and YouTube Kids in Early Childhood Media Use

Over the past decade, YouTube has become one of the most influential platforms in shaping children's digital experiences [34, 4]. With its vast, almost always-online video library, YouTube offers young viewers content that spans entertainment, education, music, and animation. According to Qustodio's annual digital report [4], YouTube has consistently ranked as one of the most used platforms among children, with usage continuing to increase year after year across age groups and countries alike.

Particularly in recent years, the significance of YouTube as a daily fixture in children's lives was further underscored during the COVID-19 pandemic. Hussain et al. [35] reveal that YouTube Kids became an instrumental part of maintaining children's routines, social interactions, and educational engagements during school closures. The study suggests that parents turned to the platform not merely for entertainment but as a vital educational resource, highlighting the platform's role as a cornerstone in the daily lives of children and stressing the importance of reliable content delivery systems.

In response to growing parental demand for safer and more child-appropriate video experiences, YouTube introduced YouTube Kids in 2015 [36, 1]. This specialized version of YouTube was developed to offer a safer, family-oriented environment, equipped with enhanced features like larger images, voice search, and robust parental controls, including screen-time management tools. Ben-Yair, YouTube Kids Director of Product Management, described it as "the first Google product built from the ground up with little ones in mind" [1]. The launch followed broader efforts by Google to tailor services for children under 13 in accordance with child privacy regulations and usability concerns [36].

Despite these intentions, the quality and child-appropriateness of content aimed at young children on YouTube, both on the main platform and within YouTube Kids, vary considerably. As a platform, YouTube offers significant potential for supporting early learning and development. Henderson et al. [5] argue that well-designed, age-appropriate video content can aid children's language acquisition, academic readiness, and social-emotional development. However, this potential is not always realized in practice. A systematic review by Alqahtani et al. [20] highlights that a substantial portion of content aimed at children fails to align with developmental needs, frequently lacking in educational quality and sometimes featuring overstimulating or inappropriate elements. Similarly, Aggarwal and Vishwakarma [21] note that while some videos are genuinely supportive of children's developmental needs, others may pose risks that are not immediately evident, such as subtle negative messaging or emotionally manipulative thumbnails. These findings point to ongoing variation in content quality and may suggest a disconnect between the platform's objectives and the actual viewing experiences for children.

2.1.2. The Emergence of Inappropriate and Harmful Content

While YouTube Kids was designed to safeguard young audiences, the platform has repeatedly struggled to filter out harmful content [21, 22, 23]. Over recent years, an alarming trend has emerged where videos that appear harmless at first glance contain material inappropriately targeting young viewers [6, 31, 24]. This deceptive content often bypasses both algorithmic recommendations and manual filtering systems, sometimes making its way into the supposedly secure environment of YouTube Kids.

A defining example of this phenomenon is Elsagate, a term used to describe a wave of videos featuring popular characters among children, like Elsa from Frozen, Spiderman, or Peppa Pig, in unsettling or inappropriate scenarios [6, 31]. These videos are often structured to mimic child-appropriate content in style, title, and thumbnail but contain disturbing elements such as violence, horror, or sexual content. Papadamou et al. [24] conducted a systematic analysis of such content and were among the first to compile a manually annotated dataset of almost 5,000 videos, distinguishing between different types of appropriateness for toddlers (i.e., aged 1-5). Their study demonstrated that a non-negligible proportion of harmful videos evade YouTube's systems, with 1.1% of 233,337 Elsagate-related videos and 0.5% of 154,957 other child-related videos classified as inappropriate for young children. They further simulated toddler browsing behavior through live simulation, mimicking a toddler randomly clicking on YouTube's suggested videos, revealing a 3.5% chance of encountering an inappropriate video within ten hops if she starts from a video that appears among the top ten results of a toddler-appropriate keyword search (e.g., 'Elsa' or 'Spiderman').

Rather than relying on overt signals of inappropriateness, many videos use vibrant thumbnails, emotionally charged titles, and familiar characters to mask their actual themes. Balanzategui [37] explores how these videos often exhibit what she terms the algorithmic uncanny: a set of formal and thematic qualities shaped more by algorithmic trends and search engine optimization than by human pedagogical intent. These aesthetic markers, such as lifeless animation, surreal scenarios, and repetitive structures, are not just bizarre but potentially disorienting, especially when served repeatedly to young children through recommendation loops.

On a broader scale, the issue extends beyond individual videos to entire channels. Gkolemi et al. [28] analyzed YouTube channels known for distributing misleading or harmful content, revealing that many such channels adopt child-friendly aesthetics to camouflage their intentions. Their findings indicate that this problem is not just a product of isolated uploads but often part of a larger network of systematic exploitation.

It is also worth noting that inappropriateness is not always confined to the video content itself. Alshamrani [38] examined the comment sections of children's videos and found a substantial presence of inappropriate language, harassment, and sexual innuendo. This suggests that even videos labeled as appropriate may still be embedded in harmful viewing environments if their surrounding interactions are not properly moderated, a systemic issue that underscores the complexity of defining and detecting appropriateness.

Finally, the severity of these findings is not limited to academia. Investigative reports by news outlets, such as The Verge [6] and Boston25News [7], exposed how violent and inappropriate videos featuring popular children's characters had been repeatedly recommended through autoplay or search to very young viewers. These exposures have spurred significant public and regulatory backlash, pushing for greater accountability and more stringent content moderation practices on the platform.

2.1.3. The Impact of Inappropriate Content on Young Children

Exposure to inappropriate content can impact children's development and emotional well-being in various ways [7, 39, 40, 41, 8, 9, 10]. For instance, repeated consumption of inappropriate content may desensitize children to violence, increasing the likelihood that they might reenact aggressive behaviors observed in these videos [7]. This tendency towards imitation aligns with Bandura's social cognitive theory, which suggests that children learn behaviors through observing adults and peers [42]. Bandura [42, 43] found that children who observe aggressive behavior initiate more aggressive behavior during play than those who observe non-aggressive behavior.

Young children are particularly likely to imitate behavior when they perceive the observed model as similar to themselves, such as those similar in age or gender. This imitation mechanism is so powerful that children may even imitate actions of fantasy characters, e.g., cartoon and superhero characters [39]. For young viewers, seeing familiar and trusted characters engaging in violent or inappropriate acts can be particularly distressing, often leading to frustration and anxiety, especially if they feel an affinity toward the character [40]. Furthermore, repeated exposure to violent and aggressive video content can more deeply affect children's behavior, cognition, and emotions compared to occasional exposure [39]. Such frequent viewing of violent material, especially during evening hours, has also been associated with increased sleep problems, including difficulty falling asleep, frequent nighttime awakenings, nightmares, and daytime tiredness, all of which negatively affect children's daily functioning and overall well-being [41]. Moreover, according to Gerbner's cultivation theory, children's social perceptions can gradually shift toward the versions of reality portrayed in video content, causing them to adopt the exaggerated or distorted perspectives frequently presented in inappropriate videos [44].

Beyond violence, exposure to inappropriate sexual content at an early age has been linked to significant developmental consequences. Children exposed to sexual content are nearly twice as likely to exhibit problematic sexual behaviors, with even stronger effects noted when the content is explicitly violent [8]. These problematic sexual behaviors may include sexual aggression, coercion, and inappropriate sexual acts towards peers, deeply affecting children's socio-emotional and behavioral development. Additionally, children accidentally exposed to sexually explicit content can experience psychological ramifications, such as feelings of shame, guilt, anxiety, and confusion [9]. These emotional impacts can hinder a child's healthy development, causing social withdrawal, diminishing self-esteem, and potentially long-term disruptions in interpersonal relationships and sexual development.

Similarly, exposure to horror or frightening content can have enduring negative effects on young viewers. Children exposed to horror media can experience prolonged anxiety, recurring nightmares, and persistent fears, some lasting months or even years after exposure [10]. Common fright reactions include difficulty sleeping, obsessive thinking about frightening content, and avoidance behaviors, e.g., reluctance to sleep alone or fear of dark spaces. These persistent reactions can significantly interfere with normal daily routines and emotional stability, possibly leading to chronic stress and anxiety [10].

Collectively, these negative effects emphasize the importance of carefully managing and moderating the content accessible to children on online video platforms. Whether stemming from violent themes, sexual innuendo, or unsettling imagery, the risks are not merely theoretical; they manifest in tangible developmental, emotional, and behavioral consequences [7, 40, 8, 9, 10]. While the responsibility to shield children from such content often falls to parents and guardians, it is equally a systemic issue tied to how platforms like YouTube organize and recommend content.

2.2. Platform Challenges and Systematic Risks

While platforms like YouTube and YouTube Kids have implemented various safeguards to filter and moderate content aimed at children, inappropriate videos continue to appear in recommendations [27, 24]. This raises a broader question: What safeguards has YouTube implemented to protect young audiences, and where do they fall short? In this section, we explore the mechanisms YouTube and YouTube Kids have implemented to filter content and the vulnerabilities that persist within their algorithmic infrastructure.

2.2.1. Platform Safeguards

To address the ongoing issue of inappropriate content surfacing on its platforms, YouTube has implemented a variety of safeguards aimed at improving content moderation and curation, particularly for younger audiences. These include machine learning models designed to detect policy-violating content [11],

a three-strike enforcement system for repeat violations of community guidelines prohibiting harmful or inappropriate material [12, 13], and the introduction of age groupings within YouTube Kids to better tailor recommendations to children's developmental stages [14]. The platform has also introduced the *made for kids* flag, allowing content creators to declare whether their videos are intended for children, in response to regulatory pressure and to improve filtering accuracy [15, 16, 17].

2.2.2. Exploitation of the Algorithm

While safeguards aim to reduce harm, YouTube's recommendation algorithm itself has become a key mechanism through which inappropriate content gains visibility. Designed to maximize user engagement and watch time, the algorithm frequently promotes videos that accumulate high interaction metrics, such as views, likes, and comments, regardless of their content type [26]. However, as multiple studies have shown, these systems rarely evaluate the developmental appropriateness of the content they promote, particularly in child-facing environments [27, 21]. This creates an opening for content creators seeking to exploit the system.

Aggarwal and Vishwakarma [21] highlight how malicious actors exploit the limitations of YouTube's moderation systems by embedding inappropriate content in otherwise benign-looking videos. These videos may use familiar characters, visually engaging thumbnails, and emotionally charged titles to attract clicks while subtly inserting harmful elements. Such tactics are particularly effective at avoiding detection by both human moderators and automated detection systems, especially when the inappropriate segments are brief or coded in ambiguous imagery. The authors underscore that such content continues to appear on YouTube, despite platform safeguards, and emphasize the need for more effective content filtering mechanisms to protect young viewers.

Further compounding the issue, Tahir et al. [27] observe that even within the confines of YouTube Kids, a platform introduced with the promise of offering a safer environment for children and ensuring compliance with laws like the Children's Online Privacy Protection Act (COPPA) [18], videos containing inappropriate material continue to surface in recommendations. Their findings suggest that current moderation mechanisms, often reliant on surface-level signals such as keywords, thumbnails, and video metadata, are insufficient to detect harmful videos designed to exploit the system. They argue that these features can be easily manipulated by uploaders seeking to trigger algorithmic promotion and that moderation systems must instead examine the audiovisual content itself to effectively flag inappropriate material. Their study further underscores the risks posed by high volumes of content, which outpace the capacity of manual review and exacerbate the platform's vulnerability to abuse.

Further, Kaushal et al.'s [45] work illustrates a systemic issue: the proximity of unsafe content to safe videos. Through network analysis, they demonstrate that videos with unsafe content frequently exist within close proximity to safe content, forming closely knit clusters within the larger network of YouTube content, increasing the likelihood that children navigating from benign videos might stumble upon harmful ones. The study brings to light the insidious nature of content promotion on YouTube, where the algorithm's inability to discern the true nature of content can inadvertently lead to harmful exposure.

These collective findings underscore the inherent vulnerabilities within YouTube's algorithmic framework. When coupled with safeguards that are incomplete or susceptible to manipulation, the platform's recommendation system not only fails to adequately filter out inappropriate content but can also actively contribute to its proliferation. Understanding the structural weaknesses of the platform and the manipulative tactics employed by some content creators is crucial for devising effective countermeasures.

2.3. Feature Analysis & Classification

In this section, we explore how inappropriate content can be identified through metadata-derived features and automated classification techniques. Specifically, we examine prior work that analyzes engagement patterns, metadata, and content-based features to detect inappropriate videos for young children. Moreover, we provide an overview of classification systems developed to flag such content at scale, ranging from metadata-based classifiers to deep learning models that combine visual, linguistic, and behavioral signals.

2.3.1. Feature Analysis in Inappropriate Video Detection

Identifying inappropriate content on YouTube, particularly content targeting young children, requires the ability to detect subtle signals embedded in video metadata, textual content, and channel-level behavior.

Several studies have explored how these features can be analyzed and operationalized to detect harmful or miscategorized content.

Hoiles et al. [26] emphasize the predictive power of metadata features such as view count, like ratio, and comment activity, noting that these engagement metrics, commonly utilized by YouTube's recommendation algorithms, can effectively model video popularity and viewer engagement. These findings suggest that engagement-related metadata, while originally designed to capture attention, could also serve as useful input for detecting anomalies or suspicious patterns when paired with other signals. However, engagement signals can be distorted. Shah [32] and Kuchhal and Li [33] reveal how coordinated view fraud, i.e., fake accounts and bot-generated views, can artificially inflate a video's popularity, making it appear trustworthy or appealing when it is not. These findings underscore the need for a more robust analysis of engagement data, particularly when attempting to detect misleading or inappropriate content targeting children.

While engagement and view-based signals provide useful indicators, complementary cues can be found in textual and audiovisual content. Yousaf et al. [22] focus on video-based cues, employing a two-stream EfficientNet-BiLSTM model that analyzes both static RGB frames and motion-based optical flow representations. Their model excels at capturing spatiotemporal patterns in children's cartoon videos and achieves strong performance in multiclass classification of inappropriate content types, such as violence or nudity. While their approach does not incorporate metadata, it highlights the importance of visual structure and motion in content assessment.

In contrast, Binh et al. [29] developed SAMBA, a fusion model that leverages video metadata and subtitles to improve appropriateness classification. Their recurrent fusion approach combines embedded representations of tags, titles, thumbnails, and subtitles, achieving a significant performance boost over metadata-only classifiers. Their work demonstrates that subtitle content can reveal inappropriate themes otherwise masked by misleading metadata, offering a powerful text-based complement to visual or statistical features.

In addition to engagement and statistical features, visual and linguistic cues can significantly influence children's interaction with content. Pinney et al. [46, 47] and Milton et al. [48] demonstrate that simple phrasing, emotionally charged language, and familiar thumbnails can sway children's viewing preferences, making them more susceptible to videos designed to attract clicks rather than provide developmentally appropriate material.

Building on these insights into feature representation, it becomes essential to validate whether observed feature patterns meaningfully align with video appropriateness labels. Correlation analysis can serve as a valuable method to confirm that selected features align consistently with the different types of appropriateness. Akoglu [49] provides a practical guide for interpreting different types of correlation coefficients, which has become a foundational reference in data mining and applied machine learning. Prematunga [50] further expands on best practices in statistical correlation, offering recommendations for the selection of appropriate correlation metrics depending on the scale and distribution of the data.

2.3.2. Inappropriate Video Classification

In addition to feature exploration, several studies have focused on developing classification systems to detect inappropriate YouTube videos aimed at children. These efforts span traditional machine learning models, deep learning architectures, and hybrid approaches, often evaluated on datasets curated specifically for this task.

Papadamou et al. [24] laid foundational groundwork by assembling a dataset of nearly 5,000 YouTube videos and categorizing them into appropriate and inappropriate classes, which serves as a critical resource for our classification task. To detect inappropriate content targeting toddlers, the authors developed a deep learning classifier that processes multiple input modalities, including titles, tags, thumbnails, and metadata features, such as video statistics and stylistic cues. The model architecture combines LSTM networks for textual features, a pre-trained CNN for thumbnail analysis, and dense layers for metadata, merging them into a unified classification output. Notably, their classifier served not only to distinguish harmful content but also to assess the likelihood of inappropriate videos appearing through YouTube's recommendation system, demonstrating how such content can still surface during typical toddler browsing behavior.

To address the challenges of identifying inappropriate content at scale, several researchers have proposed systems that leverage deep learning models and real-time feature analysis. Reddy et al. [51]

introduced a kid-friendly access model based on neural network architectures, which demonstrated a higher ability to distinguish inappropriate videos compared to traditional filters. Their model combines metadata cues with learned representations, offering a more dynamic and scalable moderation approach. However, they also emphasize that no safeguard system is foolproof and recommend pairing automated detection with human oversight. Their findings highlight the difficulty of designing robust protective mechanisms in an environment where malicious creators continuously adapt to evade detection.

Advancements in machine learning models have introduced sophisticated approaches to content classification. Aggarwal and Vishwakarma [21] introduced a hybrid model that integrates EfficientNet with a BiLSTM network to capture both spatial and sequential data. Their model effectively detects ambiguous or borderline content, i.e., videos that combine appropriate and inappropriate elements, illustrating the importance of using multimodal signals in the classification process.

Building on these ideas, Faheem Nikhat and Sait [23] proposed a more advanced deep learning architecture that combines unsupervised clustering and a double-branch recurrent neural network (PDBRNN) to detect inappropriate content. Their system achieved high accuracy by learning statistical and affective traits from large-scale datasets, such as disproportionate engagement levels or emotional cues in video descriptions, which tend to be associated with inappropriate videos targeting young children.

Ishikawa et al. [30] similarly developed a deep learning model to detect Elsagate-style cartoons by analyzing visual and temporal signals. While primarily focused on classification, their study highlights the potential for such models to be directly integrated into recommendation pipelines, proactively identifying and filtering inappropriate content before it reaches young audiences.

On the implementation side, many of these classification approaches are grounded in traditional algorithms such as Random Forests [52] and Support Vector Machines (SVMs) [53], which remain reliable baselines for comparative evaluation. Tools like Scikit-learn [54] continue to provide accessible and standardized implementations of these models. These models are typically evaluated using stratified k-fold cross-validation, a technique shown to improve reliability and consistency in performance estimation, especially for imbalanced datasets [55].

2.4. Recommender Systems & Safeguards

Online video platforms like YouTube heavily rely on recommendation algorithms to keep users engaged [25]. Although YouTube Kids was introduced as a safer, more controlled version of the main platform, the underlying algorithms of these recommendation systems largely remain opaque, driven by engagement signals, metadata, and user interactions that often prioritize popular content without adequately considering its appropriateness for young viewers [25, 26].

Studies such as those by Hoiles et al. [26] highlight how engagement metrics like views, likes, and comments can disproportionately amplify content that, while engaging, may not be appropriate for all audiences. This is particularly problematic for young viewers, who, due to their developmental stage, may lack the capacity to critically evaluate or effectively navigate the platform. Research by Duarte Torres [56] and Gwizdka and Bilal [57] illustrates that children tend to click on content that appears at the top of search results or recommendation lists, rarely scrolling beyond the first few suggestions. This behavior makes them especially susceptible to the biases embedded in recommendation and ranking mechanisms.

Efforts to improve recommender systems for child audiences have yielded promising alternatives like the SAMBA model proposed by Binh et al. [29]. SAMBA incorporates a fusion-based classifier that leverages both video metadata and subtitle content to detect inappropriate content. Once flagged, these classifications feed into an attention-based recommendation model tailored for children. Their end-to-end system combines filtering and reranking, ensuring that inappropriate videos are deprioritized while age-appropriate videos are surfaced.

To evaluate the effectiveness of our own reranking strategies, we draw upon a range of established ranking metrics from recommender systems research. Based on the framework proposed by Tamm et al. [58], we adapt Mean Reciprocal Rank (MRR), Mean Average Precision (MAP), HitRate@k, and Normalized Discounted Cumulative Gain (NDCG) to assess how effectively our reranking strategies prioritize desirable content while suppressing inappropriate material. These metrics are crucial for providing a robust and interpretable framework for performance comparison across different ranking strategies.

To complement these standard metrics, we also draw inspiration from the REC-ST metric introduced by Ungruh et al. [59, 60], which was originally designed to assess the prevalence and positioning of stereotypical content in recommendation results. Although developed in a different context, the structure of this metric offers a useful template for quantifying the ranking severity of harmful recommendations. In our study, we adapt this concept to define REC-INAP, a new metric that penalizes the appearance of inappropriate videos, especially when they appear at the top of recommendation lists. This adaptation enables us to systematically evaluate not only how well desirable content is promoted but also how effectively inappropriate content is suppressed.

2.5. Regulatory & Ethical Concerns

Beyond technical detection and system design, the regulation of child-appropriate content on platforms like YouTube raises serious ethical and legal challenges. As Miroshnichenko [25] argues, platforms like YouTube face growing pressure to increase transparency and accountability, especially regarding algorithmic recommendations and data collection practices. These concerns are amplified in the context of child audiences, where the General Data Protection Regulation (GDPR) and similar legislation impose stricter requirements around consent, profiling, and the right to explanation. The FairTube campaign highlights these tensions, advocating for workers' rights and ethical AI governance, while also underscoring the challenges of holding platform algorithms accountable in the absence of regulatory oversight.

The imperative to protect young users from surveillance and exploitation was starkly illustrated by a landmark legal action against YouTube and its parent company, Google. The companies were fined a record $170 million by the U.S. Federal Trade Commission (FTC) for violations of the Children's Online Privacy Protection Act (COPPA) [18]. The complaint alleged that YouTube collected personal information from children without parental consent by using cookies for behavioral advertising. In response, YouTube pledged to limit data collection on child-directed content and rolled out several platform changes, including disabling personalized ads and comments on videos marked for children [17]. However, these measures were met with criticism by content creators and digital rights advocates, who argued that the burden of compliance was being shifted from the platform to individual uploaders [61].

The FTC further emphasized that channel owners themselves are now legally responsible for marking their content as directed at children and may face penalties if they mislabel content to avoid algorithmic or monetization penalties [19]. This shift raises ethical concerns about responsibility and platform governance. While YouTube outsources moderation and labeling duties to individual creators, the algorithmic system that drives recommendation and monetization continues to operate with limited transparency [25]. As Brandom [6] points out in their journalistic investigation into the Elsagate scandal, the platform's structural incentive system rewards engagement above appropriateness, often amplifying harmful or misleading content.

These instances may point to a broader need to align technical safeguards with ethical responsibilities and legal mandates. While the FTC's enforcement actions against COPPA violations help establish important legal benchmarks, the dynamic and often opaque nature of recommendation algorithms continues to pose challenges to traditional accountability frameworks. We argue that ensuring the safety and well-being of child users on platforms like YouTube may require not only adjustments to algorithmic processes and content filters, but also the development of more transparent, enforceable, and ethically grounded governance structures.

Ultimately, ensuring a safe digital space for children requires coordinated efforts between platforms, regulators, content creators, and researchers, each playing a role in addressing the persistent gaps between algorithmic practices and child protection standards.

3

Lost in the Woods: Analyzing YouTube Video Features and Appropriateness Classification

In order to safeguard young children from unfitting content online and actively promote content that aligns with their developmental needs, we must first understand how video metadata can help distinguish between different types of appropriateness. To distinguish between appropriate and inappropriate video content for young children, we present in this chapter a detailed analysis of YouTube video metadata and its potential for predicting inappropriateness. First, we describe the methodological setup, including the dataset used for feature exploration and classifier training, the analysis and preparation of features for classification, and the classifier development process of different supervised learning algorithms. Then, we analyze a broad set of metadata-derived features, investigating their distributions across the data and their statistical relationships with video appropriateness. This exploratory analysis informs the selection of a final set of features that are meaningful for classification. Developing a classifier enables us to transform these features into actionable predictions of content appropriateness, allowing us to distinguish between different types of content suitability for young children. Finally, we evaluate the accuracy of the best-performing model and assess its ability to support effective recommendation strategies. This analysis lays the foundation for Chapter 4, where we incorporate these insights into reranking approaches designed to reduce the likelihood of inappropriate content being recommended to young children.

3.1. Setup

This section outlines the methodological foundation of our analysis. It introduces the dataset used to explore appropriateness in children's video content and describes the preparation and selection of metadata features for classification. We then present the supervised learning setup, detailing how these features are used to train and evaluate models capable of distinguishing between different types of appropriateness. Together, these components establish the groundwork for the feature analysis and classifier evaluation presented in Section 3.2.

3.1.1. Data

This section describes the dataset used for analyzing and classifying YouTube videos into different types of appropriateness for young children.

YouTube Dataset

In order to obtain a realistic exploration scenario that accurately reflects the type of content recommended to young children on YouTube and YouTube Kids, we use a ground-truth dataset that includes real-world video data along with expert-annotated classifications. The dataset, obtained from [24], consists of 4,797 video entries, each containing video metadata requested from YouTube's API and labeled according to its appropriateness for toddlers (i.e., children aged 1-5 years). In the remainder of this work, when we refer to young children, we specifically mean toddlers.

Unfortunately, YouTube does not provide an API for retrieving videos from YouTube Kids [24]. As

a result, the dataset is based on videos collected from YouTube's main platform. Nevertheless, this remains highly relevant, as many young children consume content on YouTube instead of, or in addition to, YouTube Kids [4]. Thus, the metadata and annotations in this dataset provide a necessary foundation for exploring how video features correlate with the defined classifications, serving as the basis for training our classifier.

The videos in the dataset are categorized into four distinct classifications based on established definitions of appropriateness. Videos intentionally targeting children but containing harmful or misleading content, with *Elsagate* being a prominent example introduced in Chapter 1, are classified as ***disturbing***. These videos, exemplified in Figure 1.3a and Figure 1.3c, deceptively incorporate child-friendly characters or themes but include inappropriate elements, such as violence, horror, or sexual content. Conversely, videos containing mature themes, strong language, or explicit content unfitting for children under 17 are classified as ***restricted*** as they are intended for an older audience and include elements that are inappropriate for young children. These videos, though not intended for young audiences, may still surface in recommendations and often contain inappropriate language, online gambling, drug use, alcohol, or graphic nudity.

In contrast, videos that, while appropriate for general audiences, do not cater to the cognitive, emotional, or entertainment needs of young children, are defined as ***irrelevant***. These may include complex educational topics; see, for instance, the example of complex numbers in Figure 1.4b. Although not explicitly harmful, irrelevant videos may be confusing or misaligned with the developmental stage and interests of young viewers. Finally, videos explicitly created for young children, containing age-appropriate and beneficial content, are classified as ***suitable***. These videos offer educational or engaging content directly aligned with the developmental stage of young viewers. An example of such a video can be found in Figure 1.1.

These four labels, i.e., *suitable*, *irrelevant*, *restricted*, and *disturbing*, are summarized in Table 3.1, which also groups them under the broader categories of *appropriate* and *inappropriate*. This typology not only helps to distinguish varying degrees of inappropriateness but also forms the conceptual backbone of this study's feature analysis and classification.

Table 3.1: Appropriateness Types [24]

Appropriate		**Inappropriate**	
Suitable	**Irrelevant**	**Restricted**	**Disturbing**
Age-appropriate and relevant to young children's development	Safe but not relevant to young children's development or interests	Mature themes not aimed at young children	Harmful content disguised as child-friendly

Understanding and distinguishing between these labels allows us to move beyond a simple appropriate-versus-inappropriate binary. It reflects the more nuanced reality of video content on YouTube and YouTube Kids. By incorporating these distinctions, we aim to better understand how inappropriateness manifests across different types of content. These classifications serve as the foundation for the feature analysis and classification described in the remainder of this chapter.

Metadata

To support the development of a classifier that predicts video appropriateness, we aim to analyze metadata features that may carry informative signals about video content. Attributes such as title, description, related videos, tags, thumbnail, comment count, dislike count, like count, and view count enable a comprehensive analysis of video characteristics and their relationship with the ground-truth labels. The goal is to identify meaningful distinctions in metadata that can be used to train a classifier capable of predicting video appropriateness for young children.

The dataset contains comprehensive metadata for each of the 4,797 video entries, capturing diverse attributes useful for analyzing video content and context. Each video entry consists of structured metadata obtained directly from YouTube's API, encompassing details about the video's technical characteristics, user interaction metrics, and thematic content indicators. Specifically, the dataset includes the following metadata attributes:

- *Caption*: A boolean indicating whether the video includes captions.
- *Definition*: The video's quality level, either standard definition ('sd') or high definition ('hd').
- *Dimension*: The video's dimensional format, i.e., '2d'.
- *Duration*: The duration of the video in ISO 8601 format [62].
- *Licensed content*: A boolean indicating whether the content in a video is licensed or not.
- *Etag*: The video's e-tag.
- *Id*: A unique identifier for the video, assigned by YouTube.
- *Elsagate related seed*: A boolean indicating whether the video is related to an Elsagate-related video.
- *Kind*: The kind of video. Only one kind can be found in this dataset, i.e., '$youtube\#video$'.
- *Related videos*: A list of YouTube video IDs recommended by the platform as related to the given video.
- *Category id*: Identifier corresponding to one of YouTube's defined video categories, e.g., "Gaming," "Music," or "Education."
- *Channel id*: The unique identifier of the channel that uploaded the video.
- *Channel title*: Name of the channel that uploaded the video.
- *Default audio language*: The default language of the video's audio.
- *Description*: The description of the video, provided by the uploader.
- *Live broadcast content*: The status indicating whether the video content was originally broadcast live ('live') or not ('none').
- *Tags*: Keywords associated with the video to aid in recommendation.
- *Thumbnail*: A visual preview image displayed for the video.
- *Title*: The title of the video, provided by the uploader.
- *Comment count*: The total number of comments viewers have left on the video.
- *Dislike count*: The number of dislikes the video has received.
- *Favorite count*: The number of times users have marked the video as one of their favorites.
- *Like count*: The number of likes the video has received.
- *View count*: The total amount of views accumulated by the video.
- *Embeddable*: A boolean indicating whether the video is embeddable.
- *License*: The license under which a video is shared, i.e., 'creativeCommon' or 'youtube'.
- *Privacy status*: The video's privacy status, i.e., 'public' or 'unlisted'.
- *Public stats viewable*: A boolean indicating whether the video's statistics are publicly accessible.
- *Upload status*: The video's upload processing status, i.e., 'uploaded' or 'processed'.

The dataset is structured such that each video entry is linked to multiple related videos, as recommended by YouTube. These related video lists not only reflect how content is positioned by YouTube's recommender systems but also enable a practical testbed for evaluating ranking strategies. In Chapter 4, we utilize these related video lists to assess how effectively we can improve the presentation of recommendations. Finally, to ensure the data's suitability for meaningful analysis, preprocessing steps were applied to clean and organize the data.

Splitting the Data

To systematically evaluate both the classifier and the reranking strategies, we split the dataset into four distinct subsets. Each subset serves a different purpose within our experiments and is constructed to support a rigorous and representative evaluation of our methods.

We define two subsets for training and evaluating the performance of our classifier and two subsets for assessing the reranking strategies. First, we construct the *Unseen Reranking Set*, consisting of 355 video entries. This subset consists exclusively of videos with at least two related videos also present in the

dataset, a property necessary for our experiment in Chapter 4. As such, it is excluded from this experiment and will be utilized further in the respective chapter.

Second, we extract the *Related Videos Set*, which contains 909 videos. These are all the videos that appear in the related video lists of entries in the dataset. To avoid data leakage and unintended label influence, this subset is excluded from classifier training. This ensures that the classifier does not learn patterns directly from related videos, as these are important for our exploration in the second experiment; we exclude these from this part of our study

Third, we define the *Unseen Classifier Set*, consisting of 459 entries. This set is sampled from the remaining data and remains untouched during training and validation. It serves as a final unbiased assessment of the classifier's performance. Importantly, this subset is not stratified by class, resulting in a label distribution that differs from the training data. This design simulates a real-world scenario in which a classifier encounters new videos with unpredictable class proportions, reflecting the messier, imbalanced conditions found in actual recommendation environments.

The remaining 3,656 videos form the *Classifier Train-Test Set*, used to train and validate the classifier through stratified K-fold cross-validation. We split this subset into eight folds using stratified sampling, where each fold consists of 3,199 training entries and 457 test entries. Stratification ensures that each fold retains the proportional distribution of labels found in the full training data. This allows for robust training and validation across varied subsets, reducing evaluation bias and providing a reliable foundation for model comparison.

By organizing the data in this way, we ensure that both parts of our exploration are evaluated on data not seen during training. This structure supports controlled experimentation while also exposing models to realistic data conditions.

3.1.2. Features: Selection & Analysis

This section discusses the features selected from the dataset for training the classifier, with the broader goal of predicting a video's appropriateness for young children. To achieve this, we examine a range of metadata attributes, including textual, visual, and engagement-related features, to identify a set of features whose characteristics could meaningfully distinguish between the defined types of video appropriateness. For each feature, we describe the rationale for inclusion, the preprocessing or engineering steps applied to derive meaningful values, and the categorization of features into interpretable types. To examine their potential for supporting classification, we outline an exploratory analysis approach that consists of visual inspections of feature distributions across labels and the calculation of statistical descriptors, such as mean, variance, and standard deviation. In addition, we describe the use of correlation analysis to assess the strength of association between individual features and appropriateness types. Together, these steps establish the analytical foundation for selecting a meaningful subset of features to be used in classifier development.

Features Chosen for Analysis

The features chosen for this analysis are derived from the metadata in the dataset. Each feature is chosen based on its potential to capture meaningful distinctions between different types of video appropriateness, as well as its accessibility, interpretability, and scalability. Wherever possible, the selection is informed by prior work on children's media consumption and content appropriateness, though some features are included in a more exploratory capacity to test under-examined or novel signals. The following features are selected for analysis, considering their relevance to understanding the appropriateness of videos for young children:

Derived Features

- *Title*: The video title provides an important first impression for viewers and is often crafted to attract attention. Due to its prominent role in how videos aimed at young children are surfaced and recommended [24], we consider the title a rich source of implicit signals regarding the video's intent and appropriateness. To extract these signals, we derive three features from each title using

established language analysis techniques. First, we analyze the emotions conveyed using NRCLex[1], an emotion detection approach previously validated in child-related contexts to understand children's preferences and engagement with content [48]. Second, we analyze the use of hard or complex words using a phonemic decoding model created by Pinney et al. [46, 47]. Third, we analyze abrupt changes in topic using the `topicChangeDetector_v1` model[2]. From now on we will refer to these features as $emotions_{title}$, $hard\_words_{title}$, and $topic\_change_{title}$ respectively. These elements can indicate whether a video is intended for children or contains content misaligned with their cognitive and emotional development. For example, a title like "Learn Colors and Shapes with Elsa" followed by "Spiderman Attacked by Zombies" represents a sudden change in tone and theme that can be perceived as a topic shift by human viewers. Such sharp contrasts can mislead children into viewing content that initially appears appropriate but transitions into inappropriate themes. The topic change model aims to capture similar shifts in thematic coherence, which may signal potentially inappropriate or misleading content. All three features are encoded numerically, capturing emotion scores, the proportion of complex words, and the degree of topic shift, respectively.

- *Description*: The description offers additional context about the video. Similar to the title, by analyzing emotions ($emotions_{description}$), hard words ($hard\_words_{description}$, and topic changes ($topic\_change_{description}$) in descriptions, we aim to uncover patterns that correlate with video appropriateness or inappropriateness. Descriptions might contain clues about a video's intended audience or other hidden intentions.
- *Thumbnail*: Thumbnails serve as a visual gateway to videos and are designed to capture attention. To derive meaningful features from this visual modality, we employed the `Salesforce/blip-image-captioning-base` model[3], a vision-language transformer designed for automated image captioning. This model generates a descriptive sentence summarizing the visual content of thumbnails, allowing us to analyze its textual properties. Once the generated captions were obtained, we applied the same textual feature extraction methods used for video titles and descriptions. First, we identify the emotional tone conveyed in the generated captions using the `NRCLex` library ($emotions_{thumbnail}$). Second, we compute the proportion of difficult words using the phonemic decoder model ($hard\_words_{thumbnail}$). High linguistic complexity in thumbnail captions may signal content that is not cognitively aligned with a young child's developmental stage. By transforming visual content into interpretable textual features, thumbnail analysis complements our other modalities and introduces an additional channel for detecting potentially inappropriate or deceptive content.
- *Tags*: Tags associated with a video can provide insights into its thematic focus. Tags that reference inappropriate themes may signal videos unfitting for children. Specifically, tags containing terms associated with the Elsagate phenomenon, such as "Elsa" or "Spiderman", could be indicative of intentionally inappropriate videos targeting young children. This assumption is grounded in previous research by Papadamou et al. [24], which demonstrated through analysis of frequent tags that these terms are strongly associated with inappropriate videos. To quantify the thematic relevance and potential appropriateness of tags, we employ TF-IDF analysis. For each video, we compute TF-IDF scores for associated tags, selecting the top 5 tags with the highest scores to capture their relative importance. We then convert these tags into numerical scores using their vector norms derived from SpaCy's language model ("en_core_web_md"), resulting in a consistent numeric representation that can be used for analysis.

Platform Metadata

- *Category id*: Each video belongs to one of YouTube's categories, e.g., "Gaming" or "Music". Certain categories may have a higher likelihood of containing inappropriate or irrelevant content for young children. By examining category distributions, we aim to identify potential trends linked to video appropriateness.
- *Licensed content*: This feature shows whether the content in a video is licensed or not. Videos with licensed content might adhere to stricter copyright and quality standards, which could influence

[1] NRCLex measures emotional affect from a body of text. The affect dictionary contains approximately 27,000 words and is based on the National Research Council Canada (NRC) affect lexicon. See documentation: `https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`

[2] Model is available at `https://huggingface.co/raicrits/topicChangeDetector_v1`

[3] Model is available at `https://huggingface.co/Salesforce/blip-image-captioning-base`

inappropriateness.

- *Default audio language*: The default language of a video's audio could provide information about its intended audience.
- *License*: The license under which a video is shared can offer additional context about its intended use and compliance with certain standards. Certain licenses may correlate with higher or lower quality and appropriateness levels.

Engagement Metadata

- *View count*: Engagement-based features such as view count play a critical role in how YouTube's recommendation algorithm surfaces content to users. Metrics like view count are frequently used to guide ranking decisions on YouTube, meaning that highly viewed videos are more likely to be promoted further [26]. While this amplification effect can help surface popular content, it may also unintentionally elevate videos that are inappropriate if they manage to attract attention early on. One thing to note is that this feature can be manipulated by using fake or automated accounts to artificially inflate view counts, making videos appear more popular and credible than they truly are [32, 33].
- *Comment count*: The number of comments on a video can reflect audience engagement. Videos with unusually low comment counts compared to their view counts may signal irregular patterns.
- *Like count*: Positive feedback in the form of likes can indicate user satisfaction and alignment with viewer expectations. However, analyzing this metric together with other features may reveal discrepancies that hint at inappropriateness.
- *Dislike count*: Similar to like count, dislikes can indicate user dissatisfaction. High dislike counts relative to like counts may suggest content that viewers find inappropriate, controversial, or unsatisfactory.
- *Favorite count*: This feature tracks how many users have marked a video as one of their favorites. A high favorite count could indicate content that is strongly liked and enjoyed by specific audiences.

An overview of the full feature set, including short descriptions, can be found in Appendix A.

To enhance the classifier's ability to distinguish between video categories, we perform feature engineering on selected textual and visual data. In particular, we derive new features from thumbnail images by generating captions and analyzing their emotional tone and linguistic complexity. These engineered thumbnail features mirror previously applied textual features, such as emotion scores, readability measures, and the presence of complex words, used for analyzing titles and descriptions, and extend this analysis to the visual modality. This step complements our earlier feature set by introducing a comparable representation for thumbnail content, enabling a more unified and expressive input space for classification. Overall, the full set of features was selected based on their theoretical relevance, practical potential, and ease of use to differentiate between video categories. The subsequent sections describe the exploratory and statistical techniques used to analyze the selected features and evaluate their utility for classification.

Exploratory Data Analysis
To gain insights into the selected features and their distributions across video appropriateness labels, we conduct an exploratory data analysis (EDA). This step is crucial in understanding the statistical properties of each feature and identifying patterns that could inform classification performance.

Descriptive Statistics To identify patterns that may help distinguish between video appropriateness labels, we examine the statistical properties of the selected features. We perform this analysis on the *Classifier Train-Test Set*, the dataset used throughout model development and evaluation. The goal of this analysis is to explore whether certain features show meaningful differences across the four video categories and whether they offer potential discriminative signals for classification.

For each feature, we compute key descriptive statistics to capture their characteristics and variability. Specifically, we calculate the mean (μ) to represent the average value of each feature across all videos. To quantify the variability and dispersion of feature values, we compute both variance (σ^2) and standard deviation (σ), which help quantify how much feature values deviate from the mean across different classes. These statistical insights are further supported by visualizations of the feature distributions across categories, aiding in the identification of potentially distinguishing patterns.

Correlation Analysis To complement the distribution and statistical relationship analysis conducted in the previous section, we now turn to statistical correlation analysis to validate whether the observed patterns hold under quantitative scrutiny. This step allows us to confirm or refute initial impressions derived from the distribution plots and more rigorously assess the predictive potential of each feature in relation to the target classification labels.

We employ a combination of statistical methods tailored to both categorical and numerical feature types, following best practices outlined in prior correlation methodology literature [49, 50]. For categorical features, we calculate Cramér's V to quantify the strength of association between each feature and the classification labels. Cramér's V ranges from 0 (i.e., no association) to 1 (i.e., perfect association) and provides a normalized, interpretable measure of dependency.

For continuous or ordinal features, we apply Spearman's Rank Correlation (ρ) to measure the degree of monotonic relationship between a feature and the labels. Higher absolute values of Spearman's ρ suggest stronger associations. In each case, we also compute corresponding p-values to assess the statistical significance of the observed correlations.

To determine overall feature importance, we prioritize categorical features based on higher Cramér's V values and numerical features based on higher absolute Spearman correlation coefficients. However, correlation strength alone does not imply reliability. Therefore, only features with statistically significant p-values (≤ 0.05) are considered for selection. This two-step approach ensures that selected features are both meaningfully associated with the labels and unlikely to exhibit those associations due to random chance.

3.1.3. Classification

To train a classifier that can predict the appropriateness of YouTube videos for young children, we first transform the selected metadata features into a format fitting for model input.

Data Vectorization The selected features are vectorized into a numerical format fitting for training the classifier. This transformation results in feature vectors containing 75 values per video, enabling consistent input for all classification models.

The goal of the classifier is to categorize videos into one of four categories: *suitable*, *disturbing*, *restricted*, or *irrelevant*, aligning with our broader understanding of video inappropriateness.

To identify the most effective classification model, we train and evaluate several supervised learning algorithms. These include a Random Forest model [52], which is an ensemble of decision trees that improves predictive performance by averaging the outcomes of multiple decorrelated trees. Random Forests are well-suited for handling heterogeneous data and capturing non-linear decision boundaries. We also evaluate multiple Support Vector Machine (SVM) variants [53], which attempt to find the optimal hyperplane that separates classes in the feature space and are highly effective in high-dimensional settings.

Additionally, we implement a single-layer PyTorch multinomial logistic regression model [4], serving as a simple linear baseline that maps input features directly to output class probabilities through a softmax function. A second version of logistic regression is implemented using Scikit-learn's built-in library [54], allowing for rapid prototyping and integration with `GridSearchCV` for efficient hyperparameter tuning. Finally, we develop a three-layer PyTorch feed-forward neural network, capable of modeling complex, non-linear interactions between input features. This deeper architecture tests whether added capacity improves performance over the simpler models. Each model was implemented using standardized procedures, including consistent data splits and uniform training configurations where applicable.

To ensure robust and unbiased evaluation, we apply stratified 8-fold cross-validation [55]. This method maintains balanced label distributions across all folds, reducing the risk of performance inflation due to imbalanced splits. Each model is trained and tested independently on each fold, and the final evaluation metrics are averaged across all folds to produce stable performance estimates. When relevant, we apply **hyperparameter tuning** using `GridSearchCV` to identify the best achievable performance for each classifier. Specifically, we perform grid search on both the Random Forest and SVM models. This step ensures that each model has the opportunity to reach its optimal configuration. For example, for SVM

[4] For documentation, see `https://pytorch.org/docs/stable/generated/torch.nn.Linear.html`

models, we explore combinations with and without **Principal Component Analysis** (PCA), and for the Random Forest mode, we optimize parameters such as maximum tree depth, split criteria, and number of estimators. We report four standard metrics for model evaluation: classification accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F1-score. These metrics are particularly important in a multi-class setting with class imbalance, as they treat all classes equally in aggregate scores.

3.2. Results

This section presents the empirical results from our analysis of video metadata and its predictive power for assessing content appropriateness. We begin by exploring the behavior of the selected features across the four classification labels using visual and statistical methods. These insights inform our understanding of how appropriateness manifests in the data and guide the final feature selection. In the second part of this section, we evaluate the performance of various classifier models trained on the selected features, comparing their ability to distinguish between appropriate and inappropriate content for young children.

3.2.1. Exploratory Data Analysis

To uncover patterns of appropriateness within the video metadata, we conduct an exploratory analysis of the selected features. This involves visualizing their distributions, inspecting descriptive statistics such as mean and variance, and examining how these characteristics vary across the four types of video appropriateness. By identifying trends and irregularities, we gain early insights into which features may hold discriminative power for classification. The results from this analysis also serve as a qualitative complement to the statistical correlation analysis presented later in this section.

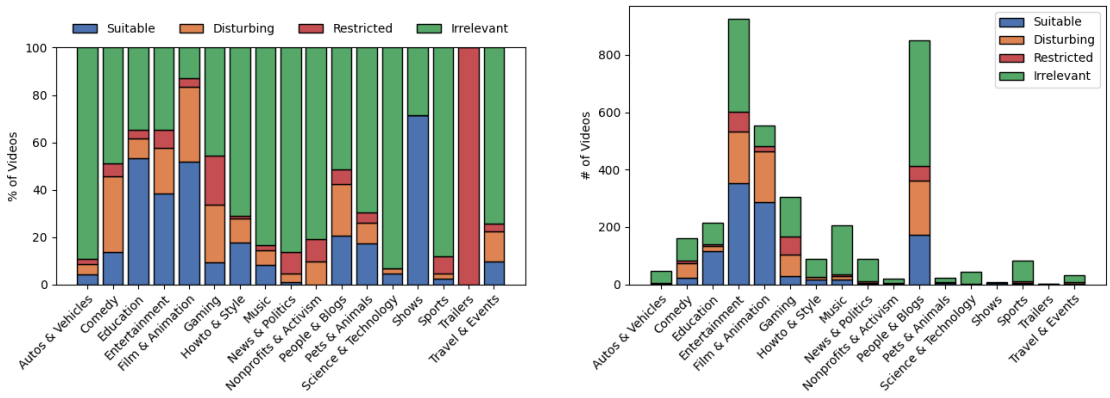
Distributions & Statistical Relationships

This section presents a detailed visual and statistical analysis of the selected features, highlighting their differences across the four appropriateness types. For each feature, we examine its distribution, report summary statistics (e.g., mean, variance, and standard deviation), and reflect on its potential to distinguish between video types. These visualizations serve as an initial lens into the dataset, offering early evidence of discriminative patterns before turning to more formal statistical methods.

To begin uncovering patterns of inappropriateness, we examine the distribution of videos across YouTube's predefined content categories, as visualized in Figure 3.1. While the *categoryId* feature alone does not directly indicate inappropriateness, it often provides strong contextual cues when considered alongside other features. From Figure 3.1b, we observe that the category *Entertainment* dominates in volume, containing 924 videos, followed by *People & Blogs*. These categories appear to serve as a kind of catch-all for a broad range of content types, and perhaps because of this breadth, they also house a notable share of inappropriate videos. As shown in Figure 3.1a, both disturbing and restricted content make up a significant portion of these categories. This observation echoes findings from Papadamou et al. [24], who noted that inappropriate content often masquerades as entertainment and is embedded in seemingly benign categories.

Interestingly, the *Comedy* and *Gaming* categories also show substantial proportions of restricted and disturbing videos. These categories are commonly associated with older audiences, which might explain their tendency to contain content that is mismatched with the developmental needs of young children. In contrast, categories such as *Education* and *Howto & Style* contain a higher proportion of suitable videos, although they are not completely free from inappropriate examples.

A noteworthy pattern emerges in the *Trailers* category, which, despite its extremely small volume, consists exclusively of restricted videos. This suggests a categorical mismatch with the needs of young children, highlighting how certain content categories may be entirely unfitting for inclusion in child-directed recommendation feeds. However, due to the small sample size of this category, caution is warranted in interpreting this pattern too strongly, as limited data may exaggerate or distort underlying trends. Ultimately, while *categoryId* is a broad feature, its distribution reveals meaningful differences across labels. When interpreted in conjunction with other metadata, such as engagement or language cues, it can serve as a useful signal for identifying content categories where inappropriate videos are more likely to appear.



(a) Distribution of *categoryId* by percentage of videos

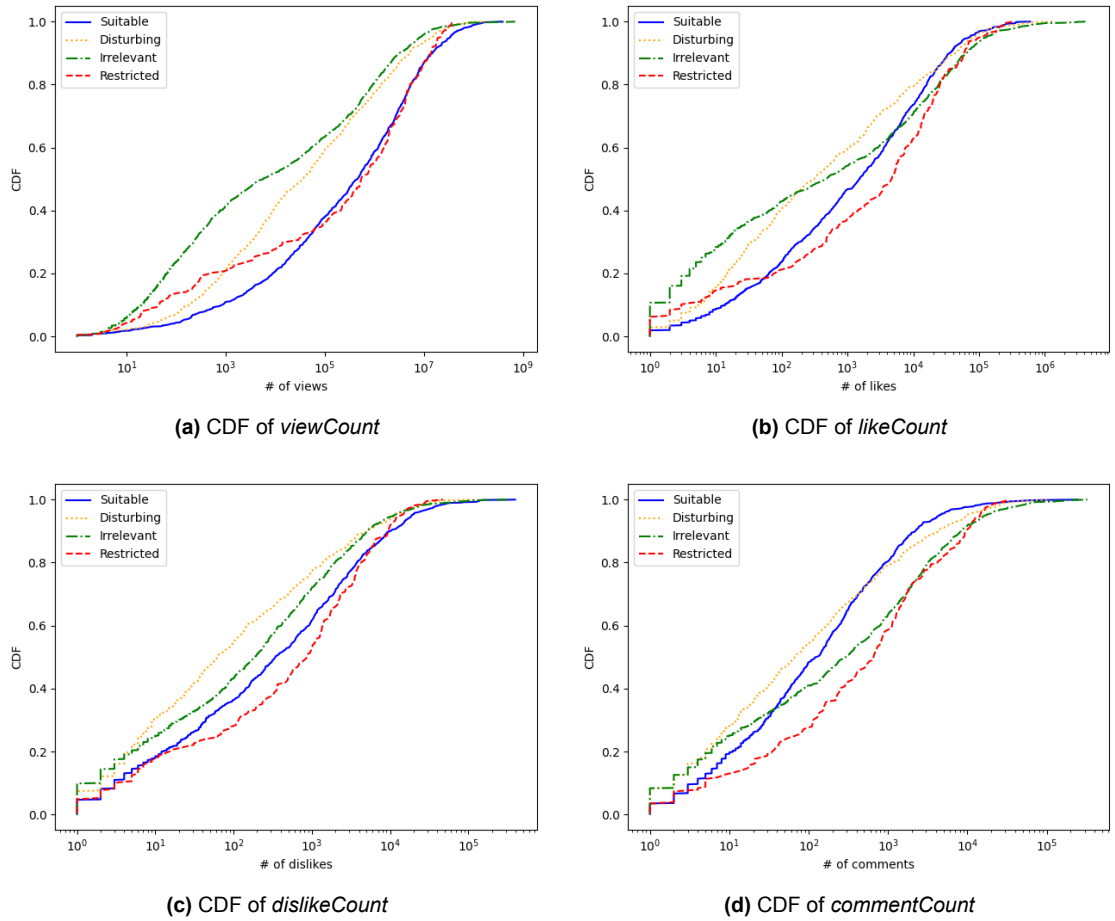
(b) Distribution of *categoryId* by total number of videos

Figure 3.1: Distribution of *categoryId*

To understand how user interaction patterns relate to video appropriateness, we examine the cumulative distributions of several engagement metrics, including view count, like count, dislike count, and comment count, as shown in Figure 3.2. These metrics can signal not just a video's popularity, but also how viewers respond to it, and whether certain types of videos receive disproportionate exposure. Prior research has shown that engagement signals, such as views, likes, and comments, play a significant role in shaping video visibility and recommendation outcomes on YouTube [26]. In Figure 3.2a, we observe that inappropriate videos, particularly disturbing ones, tend to accumulate more views than suitable videos. This disparity suggests that such content may benefit from increased visibility, potentially driven by the recommendation algorithm or through attention-grabbing thumbnails. The average number of views in the dataset is $\mu = 4{,}046{,}783$, with a large variance of $\sigma^2 = 4.98 \times 10^{14}$ and a standard deviation of $\sigma = 22{,}321{,}571$, indicating substantial differences in reach across videos. A similar pattern can be seen in Figure 3.2b, which shows that disturbing videos tend to receive likes more rapidly than suitable ones, especially in the lower range of the distribution. This may reflect early viewer engagement before the video's true nature becomes apparent. The mean number of likes is $\mu = 22,947$, with a variance of $\sigma^2 = 1.91 \times 10^{10}$ and standard deviation $\sigma = 138{,}089$.

Figure 3.2c shows that disturbing and restricted videos also receive a higher number of dislikes, especially in the early part of the distribution. This suggests that viewers often react negatively to such content, although often after it has already reached a substantial audience. The average number of dislikes is $\mu = 2{,}456$, with a variance of $\sigma^2 = 1.61 \times 10^8$ and a standard deviation of $\sigma = 12{,}698$. In Figure 3.2d, a similar pattern emerges. Disturbing videos tend to accumulate more comments than suitable ones, which may reflect confusion, criticism, or engagement through controversy. The mean number of comments is $\mu = 2{,}084$, with a variance of $\sigma^2 = 1.31 \times 10^8$ and a standard deviation of $\sigma = 11{,}452$. Lastly, the *favoriteCount* feature was excluded from further analysis, as it showed no variation across the dataset. All values were zero, resulting in a mean, variance, and standard deviation of zero. This is due to YouTube's official deprecation of the public favorites feature. As stated in the YouTube Data API documentation, the *favoriteCount* property has been deprecated since August 28, 2015, and its value is now always set to zero[5]. Although the field remains in the API for backward compatibility, it no longer reflects any real user interaction. Together, these engagement metrics reveal that inappropriate content is not only present in the dataset but also widely interacted with. While part of this interaction may reflect genuine interest, it also raises concerns about visibility and exposure, especially if algorithms prioritize engagement volume over content suitability.

[5]See the official YouTube Data API documentation: `https://developers.google.com/youtube/v3/docs/videos`, under `statistics.favoriteCount`.



(a) CDF of *viewCount*

(b) CDF of *likeCount*

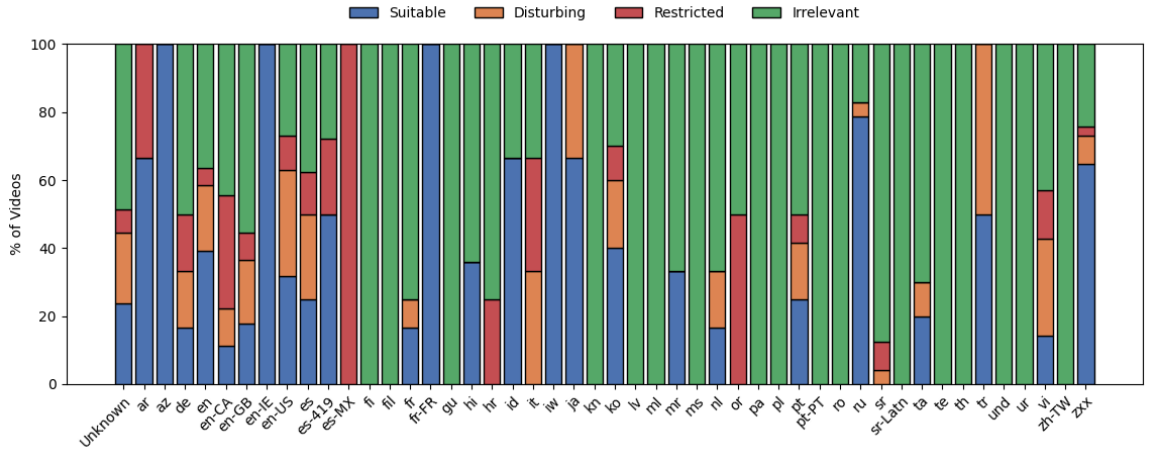
(c) CDF of *dislikeCount*

(d) CDF of *commentCount*

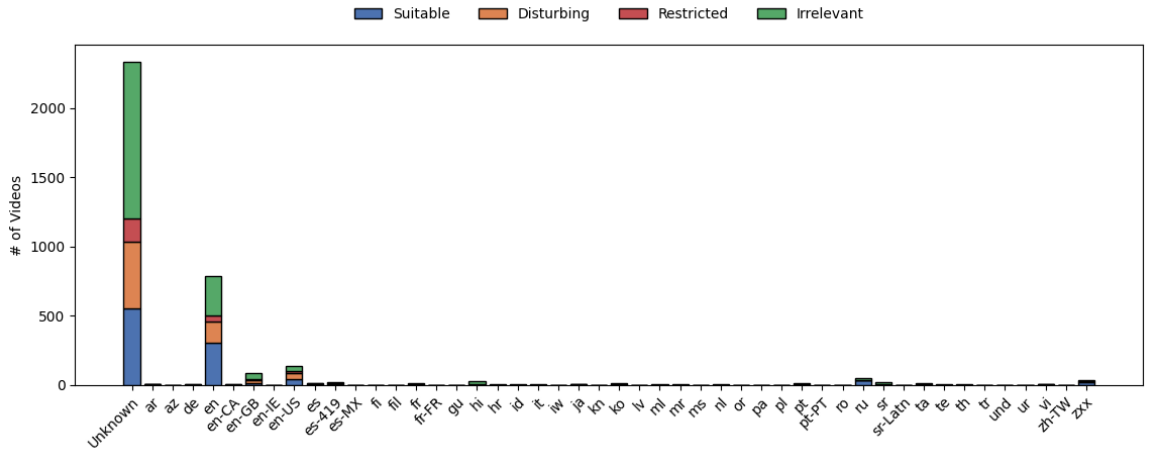
Figure 3.2: CDFs of engagement features

We analyze the distribution of videos across default audio languages in Figure 3.3a and complement this with the total number of videos per language in Figure 3.3b. The most common entry by far is ”Unknown,” which appears in 2,339 videos. We assume this likely reflects videos without spoken dialogue, such as music clips or animations, or cases where the audio language was not correctly detected by YouTube’s systems or manually assigned by the uploader. The ”Unknown” group contains a mix of labels, making it an important but ambiguous indicator when considered in isolation. In Figure 3.3b, we see that most remaining videos are associated with a small number of language codes, including ”ar” (Arabic), ”en” (English), and regional variants such as ”en-US” and ”en-GB.” These languages contribute substantially to the dataset and are more suitable for interpretation.

By contrast, some languages in Figure 3.3a appear to have unusually high proportions of disturbing or restricted videos, such as ”es-MX” (Spanish-Mexico) or ”tr” (Turkish), but these categories are represented by only a small number of videos. Without sufficient sample sizes, these proportions may be misleading. Based solely on these figures, the *defaultAudioLanguage* feature does not appear to offer a clear or consistent separation between labels. However, while the visual presentation of this feature does not appear strongly indicative of appropriateness label distinctions, its value may become more evident when examined through statistical analysis later in this chapter.



(a) Distribution of *defaultAudioLanguage* by percentage of videos

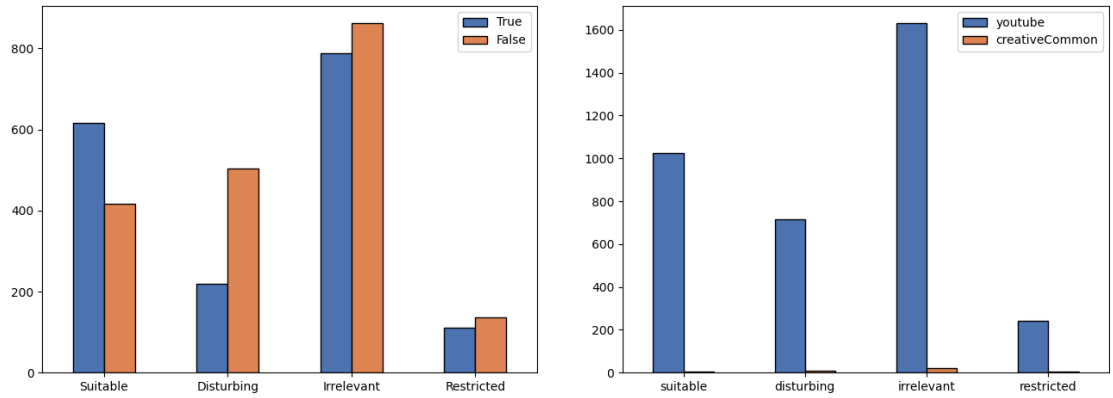


(b) Distribution of *defaultAudioLanguage* by total number of videos

Figure 3.3: Distribution of *defaultAudioLanguage*

Figure 3.4a and Figure 3.4b illustrate the distributions of the *licensedContent* and *license* features across classification labels. The *licensedContent* feature, shown in Figure 3.4a, represents whether a video contains licensed material. While the distribution is relatively balanced across all categories, disturbing and restricted videos appear more frequently in the ”False” group, suggesting that inappropriate content may be more common among unlicensed or unofficial uploads. The overall mean value is $\mu = 0.475$, with a variance of $\sigma^2 = 0.249$ and a standard deviation of $\sigma = 0.499$, indicating that the feature is nearly evenly split between True and False across the dataset.

The *license* feature, visualized in Figure 3.4b, captures whether a video is published under YouTube’s standard license or a Creative Commons license. Almost all videos fall under the YouTube license, with Creative Commons accounting for only a very small subset. The mean value is $\mu = 0.011$, with a variance of $\sigma^2 = 0.011$ and a standard deviation of $\sigma = 0.104$. Although a few disturbing and irrelevant videos are published under the Creative Commons license, the absolute numbers are extremely low, making it difficult to draw strong conclusions from this feature in isolation. Overall, both licensing-related features may hint at content oversight or legitimacy, but neither shows a clear or strong separation between categories based on these visual distributions. Their utility may lie in reinforcing other features rather than serving as primary signals on their own.



(a) Distribution of *licensedContent*

(b) Distribution of *license*

Figure 3.4: Distributions of *licensedContent* and *license*

Figure 3.5 presents the distribution of the *tagScores* feature across the four video categories. These scores reflect the semantic weight and relevance of tags assigned to each video, derived from TF-IDF values and encoded using SpaCy's language model (see Section 3.1.2). Across all categories, the distributions show a clear bimodal shape, indicating that videos tend to have either low or high-scoring tags, with fewer falling in between.

Among the more notable patterns is the large cluster of near-zero tag scores in irrelevant videos, suggesting that these videos either do not rely heavily on tags or use ones that are not distinctive or widely used across the dataset. In contrast, both suitable and disturbing videos exhibit a higher prominent density peak between tag scores of approximately 6.0 and 7.5 compared to irrelevant and restricted videos, suggesting a shared tendency to use high-weight, thematically loaded tags. Restricted videos also exhibit a similar dual-peak structure, but their density is more evenly distributed across the lower and higher ranges.

The overall average tag score is $\mu = 4.06$, with a variance of $\sigma^2 = 9.54$ and a standard deviation of $\sigma = 3.09$. Based on this visual presentation, the *tagScores* feature appears to hold some potential for helping differentiate between appropriateness labels, particularly in distinguishing irrelevant videos from the rest. However, whether this feature actually correlates meaningfully with video label classifications will be examined more rigorously later in this chapter through statistical analysis.

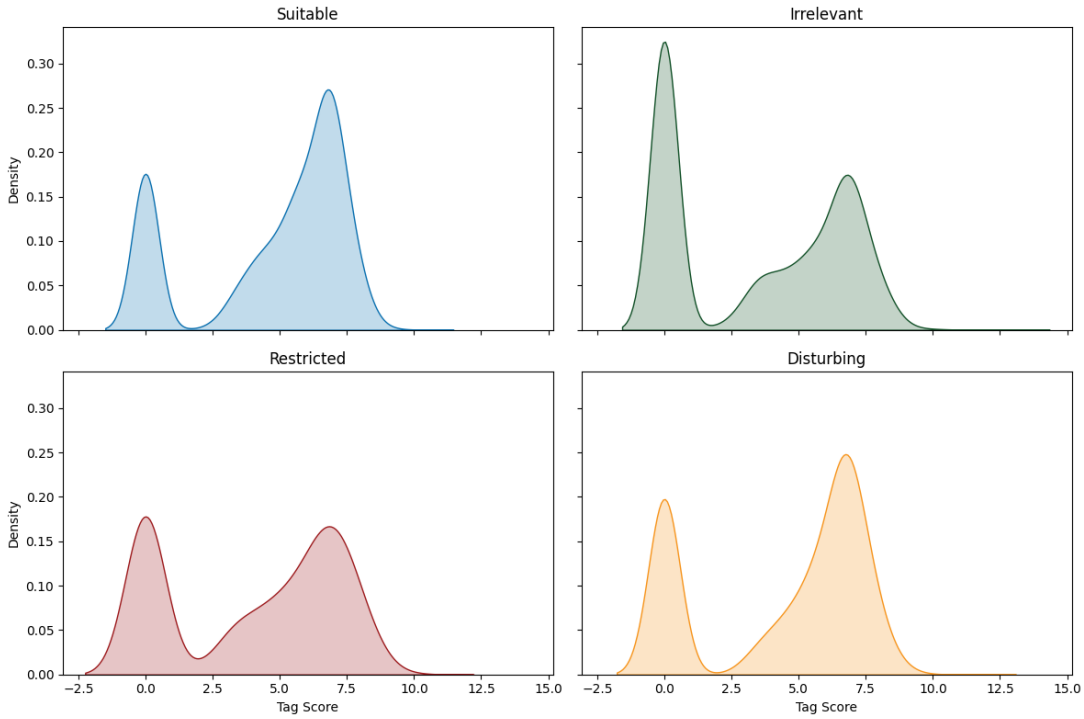


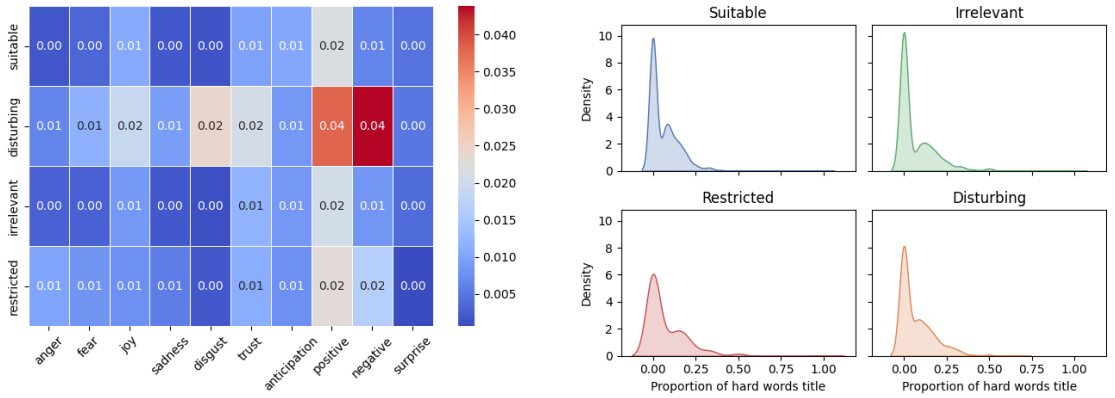
Figure 3.5: Distribution of *tagScores*

Figure 3.6 presents the distribution of emotional and linguistic complexity features extracted from video titles. The heatmap in Figure 3.6a shows the mean values for ten emotion categories, as detected by NRCLex (see Section 3.1.2). While most scores are relatively low, disturbing videos stand out for their elevated levels of both `positive` and `negative` sentiment. This dual emotional tone may reflect attempts to capture attention through emotionally charged language, even when it is contradictory. Disturbing videos also score slightly higher on emotions such as `disgust`, `fear`, and `sadness`, which may hint at darker or inappropriate themes being conveyed subtly in titles.

In contrast, suitable and irrelevant videos generally show lower emotional intensities across all categories. Restricted videos display a more balanced emotional profile, but still with somewhat elevated negative sentiment. Statistically, the most frequently expressed emotions in titles across all videos are `positive` ($\mu = 0.024$, $\sigma^2 = 0.014$, $\sigma = 0.119$) and `negative` ($\mu = 0.015$, $\sigma^2 = 0.009$, $\sigma = 0.093$), with the rest being comparatively lower.

Figure 3.6b illustrates the distribution of $hard\_words_{title}$, representing the proportion of complex words per title. Restricted and disturbing videos tend to have slightly higher concentrations of complex vocabulary, which may either obscure meaning or signal that the video is not intended for young audiences. Suitable and irrelevant videos cluster more closely around lower complexity scores. The average hard word proportion across all titles is $\mu = 0.067$, with a variance of $\sigma^2 = 0.010$ and a standard deviation of $\sigma = 0.102$.

Taken together, these visual patterns suggest that both emotion and linguistic complexity of titles are not arbitrary but may reflect underlying intent or target audience.



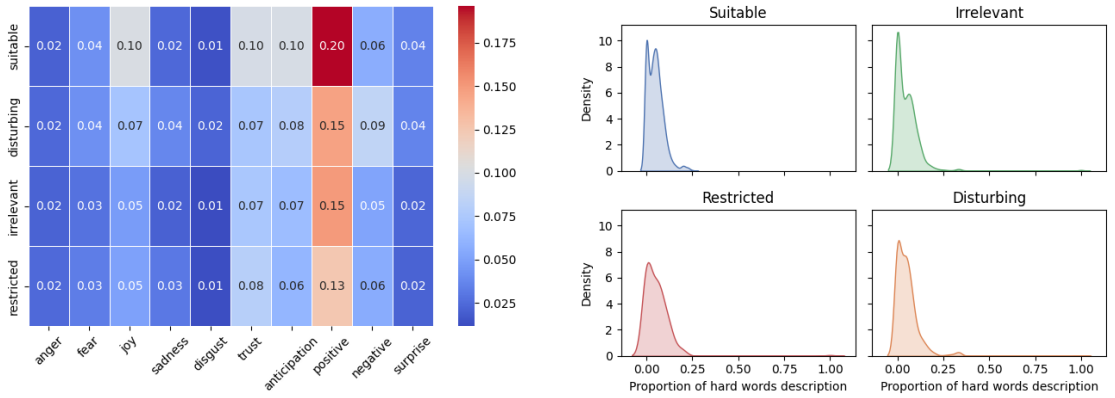
(a) Mean heatmap of $emotions_{title}$

(b) Distribution of $hard\_words_{title}$

Figure 3.6: Distributions & relationships of title-based features

Figure 3.7 illustrates the distribution of emotional tone and linguistic complexity in video descriptions. The heatmap in Figure 3.7a shows that suitable videos consistently express higher levels of `joy`, `trust`, `anticipation`, and overall `positive` sentiment compared to other categories. In contrast, disturbing videos show elevated levels of `sadness`, `disgust`, and `negative`, alongside atypically high `joy` and `positive` scores, echoing patterns observed in the titles. This emotional ambivalence may reflect emotionally manipulative or misleading framing. Irrelevant and restricted videos display lower intensity across most emotional categories, with restricted videos showing a slightly flatter profile overall. The most prevalent emotional scores across all videos are `positive` ($\mu = 0.160$, $\sigma^2 = 0.045$, $\sigma = 0.213$), `trust` ($\mu = 0.081$, $\sigma^2 = 0.018$, $\sigma = 0.135$), and `anticipation` ($\mu = 0.078$, $\sigma^2 = 0.020$, $\sigma = 0.140$), indicating that descriptions tend to convey optimism, reliability, and forward-looking intent, especially in suitable content.

Figure 3.7b presents the proportion of difficult words in each video description. Inappropriate content tends to use slightly more complex vocabulary, possibly to obscure its intent or circumvent keyword filtering. Suitable and irrelevant videos cluster more tightly around lower complexity scores. On average, the proportion of hard words is relatively low across all descriptions ($\mu = 0.049$, $\sigma^2 = 0.004$, $\sigma = 0.066$), suggesting that most descriptions are linguistically simple but still allow for nuanced variation between categories. These visual patterns reinforce earlier trends seen in titles and further suggest that emotionally charged or complex descriptions may offer predictive signals.



(a) Mean heatmap of $emotions_{description}$

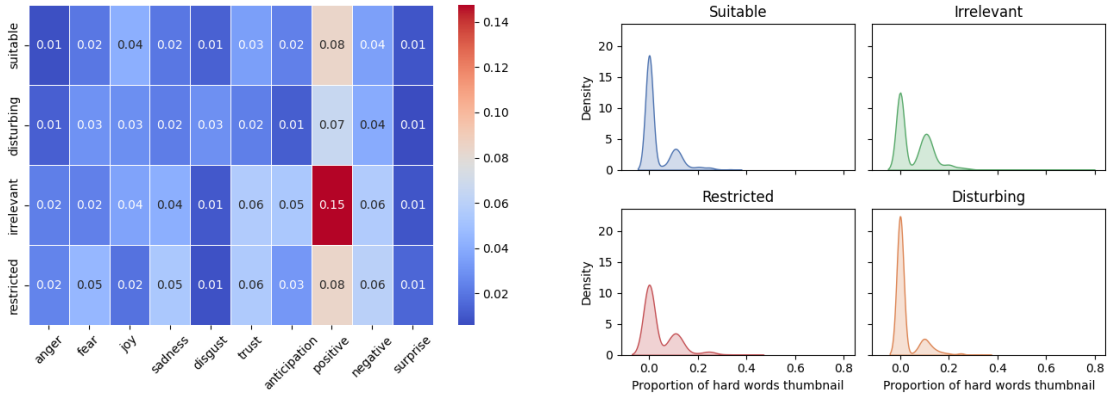
(b) Distribution of $hard\_words_{description}$

Figure 3.7: Distributions & relationships of description-based features

Figure 3.8 presents emotion-based and readability-based features extracted from the thumbnails via image captioning. Similar to the title and description features, these measures aim to detect subtle affective

or linguistic signals embedded in the visual content. In Figure 3.8a, we observe that suitable and irrelevant videos tend to elicit higher values for *joy*, *trust*, and *positive* emotions compared to disturbing and restricted videos. For instance, suitable thumbnails show the highest average value for *positive* ($\mu = 0.109$), followed by irrelevant ones. In contrast, disturbing and restricted thumbnails exhibit lower values on these affective traits while showing slightly elevated levels of *negative* affect, such as *sadness* and *fear*. The overall average *positive* score is $\mu = 0.109$, with a variance of $\sigma^2 = 0.0621$ and a standard deviation of $\sigma = 0.249$, while the *negative* score averages $\mu = 0.0472$ with a variance of $\sigma^2 = 0.0220$ and $\sigma = 0.148$.

In Figure 3.8b, the distribution of $hard\_words_{thumbnail}$ shows consistent patterns across all labels but with slightly elevated complexity for restricted videos. These results suggest that even image captions derived from thumbnails, when transformed into text, may contain subtle cues that reflect the appropriateness of content. The average proportion of difficult words is $\mu = 0.0444$, with a variance of $\sigma^2 = 0.0047$ and a standard deviation of $\sigma = 0.0684$. While the visual patterns do not sharply separate all categories, they suggest that thumbnails may still encode subtle signals that reflect the tone or complexity of the content. Whether these signals align meaningfully with video label classifications will be examined more systematically in the correlation analysis later in this chapter.



(a) Mean heatmap of $emotions_{thumbnail}$

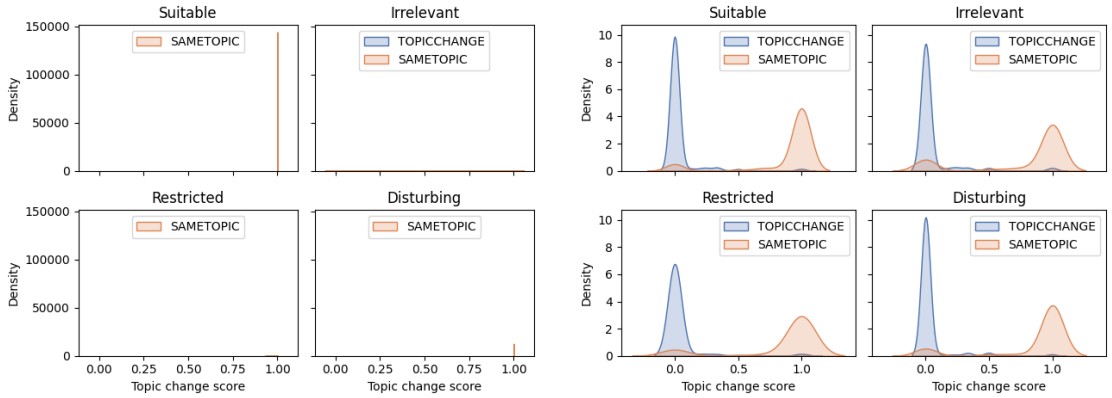
(b) Distribution of $hard\_words_{thumbnail}$

Figure 3.8: Distributions & relationships of thumbnail-based features

Figure 3.9 presents the distribution of topic change scores derived from video titles and descriptions. These scores quantify abrupt shifts in thematic coherence within a video's own text, capturing whether the content changes tone or subject matter in ways that may be confusing or misleading for children. The `TOPICCHANGE` value reflects texts identified as containing such internal topic shifts, while `SAMETOPIC` corresponds to texts deemed thematically consistent.

In Figure 3.9a, the `SAMETOPIC` distributions are sharply concentrated near 1.0 across all categories, with a mean of $\mu = 0.996$ and a standard deviation of $\sigma = 0.058$. This reflects strong internal thematic coherence in most video titles. In contrast, `TOPICCHANGE` scores appear only in irrelevant videos and are centered near 0.0, with a mean of $\mu = 0.003$ and a standard deviation of $\sigma = 0.056$. However, since no disturbing, restricted, or suitable videos in the dataset are labeled with `TOPICCHANGE`, this feature lacks the balanced representation necessary to support classification across categories meaningfully. Its skewed presence suggests that while it may highlight thematic inconsistencies in irrelevant titles, it is not currently reliable or informative for broader appropriateness distinctions based solely on title text.

The pattern in descriptions, shown in Figure 3.9b, follows a similar structure but is less polarized. `SAMETOPIC` scores still skew toward the upper end of the scale, though with greater variability ($\mu = 0.837$, $\sigma = 0.348$), reflecting the naturally broader and more varied nature of video descriptions. In contrast, `TOPICCHANGE` scores average $\mu = 0.041$ with a standard deviation of $\sigma = 0.156$, capturing meaningful deviations in thematic coherence. Unlike the title-based variant, this feature is well-represented across all four labels, making it a more viable candidate for classification. These results suggest that descriptions, with their richer textual content, can provide more consistent and balanced signals for detecting inappropriate shifts in topic.



(a) Distributions of $topic\_change_{title}$

(b) Distributions of $topic\_change_{description}$

Figure 3.9: Distributions & relationships of thumbnail-based features

In summary, the above figures and statistics demonstrate that various features, including emotional tone, linguistic complexity, content category, and engagement metrics, exhibit visually distinguishable patterns across video labels. While many features show only subtle differences in isolation, their combined representations reveal promising signals that could support automated classification.

Feature Selection Results

Based on the results of the correlation analysis, we identify 12 features that exhibit the strongest potential for distinguishing between video categories. These features, listed in Table 3.2, are selected using two thresholds: a correlation score of at least 0.1 and a statistically significant p-value of ≤ 0.05. The correlation score reflects the strength of association between each feature and the classification labels, measured by Cramér's V for categorical features and by absolute Spearman correlation coefficients for numerical features. The two thresholds ensure that only features with both practical relevance and reliable statistical support are included. The selected features span a diverse range of feature types. They include platform-level attributes such as *categoryId*, *licensedContent*, and *defaultAudioLanguage*; engagement metrics including *viewCount*, *likeCount*, and *dislikeCount*; and several affective features derived from emotion analysis of titles and descriptions. The inclusion of both engagement-based and emotional signals highlights the multifaceted nature of video inappropriateness and underscores the importance of combining multiple modalities in classification.

Table 3.2: Features selected based on correlation score threshold and statistical significance

Feature	Correlation Score	p-value
categoryId	0.302939	1.510112e-179
viewCount	0.214811	2.002030e-39
licensedContent	0.200900	8.855065e-32
defaultAudioLanguage	0.180428	1.038449e-20
descriptionEmotions.joy	0.142543	4.712159e-18
dislikeCount	0.137026	8.669567e-17
titleEmotions.disgust	0.136714	1.018876e-16
titleEmotions.negative	0.133481	5.297923e-16
likeCount	0.107725	6.561151e-11
descriptionEmotions.positive	0.107141	8.318086e-11
descriptionEmotions.trust	0.106909	9.133928e-11
descriptionEmotions.anticipation	0.106356	1.141812e-10

This structured feature selection approach not only reduces noise and mitigates overfitting but also ensures that the classifier is guided by features grounded in both theory and data. The final set of 12 features serves as the foundation for training the classifier model. Notably, several of these features align with patterns identified earlier in the distribution analysis, such as the strong differentiation potential of *categoryId*, *licensedContent*, and emotion scores in titles and descriptions, providing further validation of their relevance. Meanwhile, some features that appeared less discriminative in visualizations, like *defaultAudioLanguage*, nonetheless showed significant statistical associations with the classification labels, highlighting the importance of combining qualitative and quantitative perspectives in feature selection. A complete overview of all features considered during the correlation analysis, including their respective correlation scores and p-values, is provided in Appendix B.

3.2.2. Classification Results

Classifier Model Selection

Table 3.3 presents the average performance of each classifier across the eight folds. The best-performing model across all configurations is a Random Forest classifier optimized through hyperparameter tuning using `GridSearchCV`. This tuned version of the model achieved the highest overall performance in terms of accuracy (64.06%), precision (56.98%), and recall (47.91%), while maintaining a comparable macro-averaged F1-score (48.31%) to the untuned variant (48.45%). This discrepancy can be attributed to how macro-F1 is calculated by averaging the F1-scores of all classes equally. In multi-class settings with imbalanced label distributions, even minor reductions in F1-score for underrepresented classes, such as the *restricted* category, can disproportionately impact the macro average.

Despite this small drop, the tuned model outperformed others in nearly all metrics and offered more balanced overall predictions, making it the most robust choice for downstream integration. For the remainder of this manuscript, we refer to this optimized model as the ”Random Forest with GridSearch”. Given its superior performance, this model was selected as the final classifier and re-trained on the full training dataset using the best hyperparameter configuration. Its predictions form the foundation of the classifier-based reranking approach described in the next chapter.

Table 3.3: Classifier performance across 8-fold cross-validation

Classifier Model	Accuracy	Precision	Recall	F1 Score
PyTorch Neural Network	45.13%	11.29%	24.98%	15.55%
Random Forest	62.20%	52.75%	47.87%	**48.45%**
Random Forest w/ GridSearch	**64.06%**	**56.98%**	**47.91%**	48.31%
SVM	49.70%	44.25%	45.60%	43.43%
SVM w/ PCA	50.22%	44.38%	45.73%	43.75%
SVM w/ GridSearch	51.20%	43.95%	43.75%	43.16%
SVM w/ PCA & GridSearch	51.18%	43.40%	42.30%	41.98%
PyTorch Logistic Regression	48.39%	39.41%	31.09%	26.47%
Scikit-learn Logistic Regression	58.78%	42.40%	41.08%	39.73%

Classifier Results

The final results of the classifier are presented in Table 3.4, which displays the model’s performance on the Unseen Classifier Set. This includes per-class precision, recall, and F1-score, as well as macro- and weighted averages and overall accuracy.

This final grid search yielded the following best-performing hyperparameter configuration: **Bootstrap** = True, **Class Weight** = None, **Criterion** = ”entropy”, **Max depth** = None, **Max features** = 0.5, **Min samples leaf** = 3, **Min samples split** = 10, and **Number of estimators** = 100. Using this configuration, the classifier was retrained on the full training set and evaluated on the Unseen Classifier Set to provide an unbiased estimate of real-world performance.

The Unseen Classifier Set was held out entirely from all model training and validation procedures, offering a realistic approximation of how the classifier would perform when encountering new data. This

evaluation setup ensures that reported performance metrics reflect not just internal consistency but also generalization capability beyond the training distribution. As shown in Table 3.4, the results vary substantially across classes, underscoring both the strengths and limitations of the final model when applied under unconstrained, real-world conditions.

Table 3.4: Random Forest classifier results on the Unseen Classifier Set

	Precision	Recall	F1-score
suitable	42.86%	44.26%	43.55%
disturbing	39.33%	36.84%	38.04%
irrelevant	55.74%	67.00%	60.85%
restricted	0.00%	0.00%	0.00%
accuracy			49.02%
macro avg.	34.48%	37.02%	35.61%
weighted avg.	44.18%	49.02%	46.36%

While the final model exhibited promising results during cross-validation, its generalization to unseen data proved more challenging. The classifier achieved an overall accuracy of 49.02%, with macro-averaged precision, recall, and F1-score of 34.48%, 37.02%, and 35.61%, respectively. These metrics indicate a moderate ability to distinguish between video categories under real-world conditions.

Performance across classes varied significantly. The model was most successful in identifying *irrelevant* videos, achieving a recall of 67.00% and an F1-score of 60.85%. Followed by *Suitable* and *disturbing* were the next best-performing categories, with F1-scores of 43.55% and 38.04%, respectively. However, the classifier failed entirely to identify *restricted* content, scoring 0.00% on all evaluation metrics for this category. This shortcoming is likely due to a combination of factors, including the low representation of *restricted* videos in the training data and the greater heterogeneity of this class, which makes it more difficult to model using the selected features. This limitation is noteworthy as it raises the risk that restricted videos may be misclassified into more appropriate categories and inadvertently prioritized in recommendations. It highlights the importance of implementing additional safeguards or strategies when handling underrepresented but high-risk content categories.

Despite these limitations, the results show that the classifier can capture meaningful distinctions between categories. Its strong performance on the irrelevant class provides a valuable signal for refining recommendations aimed at young children. These findings reinforce the value of using the classifier as an initial filtering mechanism within a larger content moderation pipeline. In the following chapter, we explore how the classifier's predictions can be incorporated into a reranking strategy designed to minimize the exposure of young children to inappropriate content while promoting suitable recommendations.

3.3. Discussion

The results of this chapter shed light on the complex and nuanced nature of classifying video appropriateness for young children. The stark contrast between cross-validation performance and evaluation on the Unseen Classifier Set highlights just how challenging it is to build a classifier that generalizes well across different datasets. While the tuned Random Forest classifier performed strongly during stratified 8-fold cross-validation, achieving an average accuracy of 64.06%, precision of 56.98%, recall of 47.91%, and F1-score of 48.31%, its performance on the Unseen Classifier Set dropped, yielding an overall accuracy of 49.02%, precision of 34.48%, recall of 37.02%, and F1-score of 35.61%.

This decline highlights the challenges of generalization in real-world conditions. The Unseen Classifier Set was randomly sampled and not stratified by label, resulting in a distribution that differs from the training and validation data. It simulates the unpredictable nature of content encountered in real-world scenarios, where class imbalance, label noise, and the subtlety of content signals complicate reliable classification.

Accurately assessing the appropriateness of videos for young children is an inherently difficult problem with no simple solution. Video metadata and textual signals often contain only indirect or weak indicators of

appropriateness, and the nature of inappropriate content, particularly disturbing videos, is often intentionally deceptive. Disturbing videos are intentionally crafted to closely resemble suitable videos, making it especially hard to detect using automated methods. These types of videos are often designed to appear child-friendly at first glance while actually embedding harmful or misleading content. As a result, even sophisticated models with rich feature sets can struggle to distinguish between categories reliably.

This also helps explain why platforms like YouTube and YouTube Kids face ongoing challenges in filtering out inappropriate content. Although these platforms employ large-scale classifiers and human moderators, the very nature of the problem (i.e., ambiguous signals, malicious content, and uneven label distributions) makes it difficult to build a universally reliable system. Our findings underscore the need for layered strategies that go beyond single-model classification to improve recommendation quality.

Despite these difficulties, the feature analysis and classifier modeling conducted in this chapter provide valuable insights. Several features, including *categoryId*, *viewCount*, *licensedContent*, *defaultAudioLanguage*, and the emotional tone in titles and descriptions, demonstrated meaningful correlations with video labels. These insights, along with the trained classifier itself, will serve as the foundation for the next stage of this study: developing score-based reranking approaches. Rather than relying solely on the classifier as a filtering mechanism, we will integrate the classifier's predictions and feature-based signals into different recommendation frameworks. The objective is to assess the extent to which practical improvements can be achieved through learned signals and classification, steering recommendations toward content that aligns more closely with young children's cognitive and emotional needs and away from content that is clearly harmful.

In summary, while classification alone does not offer a perfect solution, we demonstrate in this chapter its utility as part of a layered strategy. The learned patterns and selected features provide an informed basis for steering recommendation decisions, allowing for a more nuanced response to the problem of inappropriate content on children's platforms. This integration is the focus of the next chapter, where the classifier's outputs are combined with a score-based reranking strategy to promote safer and more appropriate recommendations.

4

Do you want to build a Recommender Strategy?: YouTube Video Reranking

Recommender systems play a central role in shaping the viewing experience on platforms like YouTube and YouTube Kids. For young children, whose media consumption is often guided by suggestions from such systems, the quality of recommendations is not merely a matter of relevance, it is a matter of safety, development, and well-being. To improve the safety and relevance of video recommendations for young children, we introduce and evaluate in this chapter several score-based reranking strategies designed to minimize exposure to inappropriate content while promoting suitable content for children aged 1–5 years. We leverage the insights from Chapter 3 to develop a recommendation approach that aligns with the cognitive and emotional needs of young children by prioritizing suitable videos and minimizing the visibility of disturbing or restricted content. First, we introduce the setting of the reranking task and define what constitutes an optimal recommendation for children. Next, we address the limitations of the related video data, specifically the presence of unknown items, and how we account for this in our evaluation. We then present four reranking strategies: one based on classifier predictions, two leveraging metadata features identified as most predictive during earlier analysis, and two that fuse the outputs of these approaches. Finally, we evaluate the effectiveness of these strategies using multiple ranking metrics, comparing their ability to surface appropriate content and suppress inappropriate recommendations.

4.1. Setup

Online video platforms typically present video recommendations as ranked lists. For children, these rankings carry additional weight, as research has shown that young users tend to interact most with top-ranked items and rarely scroll far beyond the initial few results [56, 57]. This makes the ranking position of videos crucial: harmful content placed high in the list can be just one click away.

Figure 4.1 illustrates an example of such a ranked list, highlighting the importance of positioning suitable content prominently to maximize visibility and engagement while minimizing potential harm.

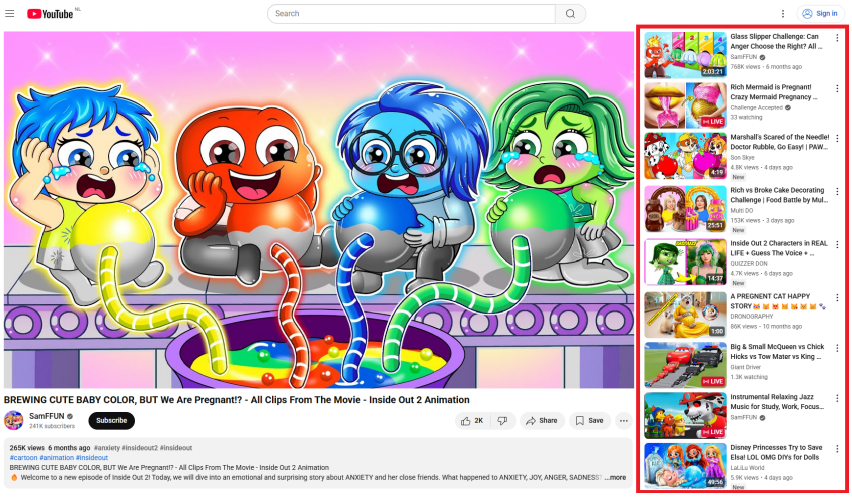


Figure 4.1: Example of a ranked recommendation list

We argue that, optimally, video recommendations for children should exclusively be *appropriate*; this includes videos that are **suitable** since they are made for children and only provide age-appropriate content, and **irrelevant** videos, which, according to the definition by Papadamou et al. [24], are not specifically made for children but do not pose direct harm. Although irrelevant videos may not cater to a toddler's cognitive, emotional, or entertainment needs, their presence is less problematic than content with disturbing or restricted elements.

In contrast, **restricted** videos often include mature themes, such as violence, sexual references, or coarse language, and are clearly inappropriate for young children. Most concerning, however, are **disturbing** videos, those that deliberately mimic child-appropriate content while embedding frightening, abusive, or deceptive material. These videos are intentionally designed to circumvent safeguards and mislead young audiences, making them especially harmful. Therefore, an ideal ranking system should place disturbing content at the very bottom of the list, below all other types of appropriateness, in order to minimize the likelihood of accidental exposure.

In this context, an optimal recommendation list should not only avoid suggesting inappropriate videos but also rank suitable videos as highly as possible. Prioritizing suitable content maximizes the likelihood that children engage with age-appropriate content that aligns with their developmental needs. Our goal in this chapter is to demonstrate how various reranking strategies can help approximate this ideal by reranking recommendation lists to elevate safer, more beneficial videos while demoting those that are harmful or less fitting.

4.1.1. Data

To simulate a realistic recommendation scenario, we use real-world data from the *Unseen Reranking Set*, which we constructed in Section 3.1.1. Each video in this set is accompanied by a list of related videos, collected directly from YouTube's API. These related video lists reflect recommendations shown to users and serve as the input for our reranking strategies. Each item in these lists is represented by a YouTube video identifier (video id), and in ideal circumstances, each id would be matched with a corresponding entry in our main dataset, containing the video's metadata and a label for its appropriateness.

These related videos form the basis for our evaluation: we aim to rerank them in a way that approximates our optimal ranking. However, not all listed videos are known; some appear in the recommendation lists but are not present in the dataset.

4.1.2. Unknown Related Videos

A key challenge in working with YouTube's related videos is the presence of unknown entries, which are video identifiers listed as related to others but not present in the main dataset. We define ***unknown*** videos as those for which we have neither metadata nor ground-truth labels, meaning we cannot extract features

or generate classification predictions. Consequently, we have no basis for assessing their appropriateness or calculating reranking scores.

As shown in Figure 4.2, most lists contain either 10 or 30 entries, reflecting YouTube's common recommendation lengths. Figure 4.3 illustrates the impact of these unknown entries across the related video lists. Figure 4.3a reveals that many of these lists include a substantial number of unknowns, often eight or more, and some lists contain up to 28 entries for which we have no information. Figure 4.3b shows the effective list lengths when unknown entries are removed. Most lists shrink to only two or three known videos, with very few extending beyond that.

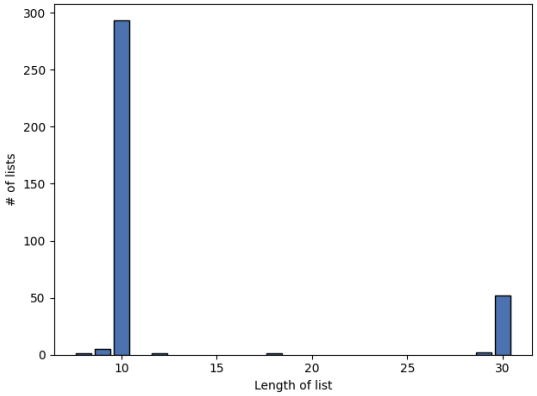
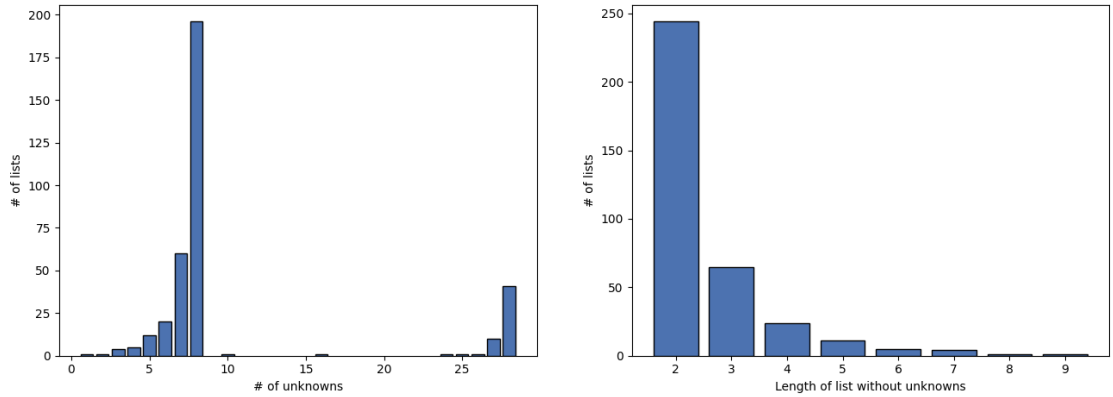


Figure 4.2: List length of related video lists



(a) Number of unknowns by number of lists

(b) List length without unknowns

Figure 4.3: Unknowns in related video lists

Given this situation, we choose not to discard unknown entries from the ranking lists. Removing them would significantly reduce the size of the recommendation lists and, more importantly, it would distort the original structure of YouTube's rankings. Since higher-ranked positions on YouTube typically receive more clicks, artificially promoting lower-ranked items by removing unknowns could introduce unintended biases in the reranking evaluation.

Instead, we take a different approach. Because we cannot determine whether unknown videos are appropriate or inappropriate, we leave them in their original positions but focus our reranking strategy on the known items. Our objective is to ensure that videos labeled as suitable or irrelevant (i.e., appropriate) are promoted to higher ranks than unknowns, where possible, ensuring that items that we predict to be appropriate appear above videos where their impact on children is unknown, or potentially even inappropriate. Conversely, known videos labeled as inappropriate, such as disturbing or restricted, should

be placed below the unknowns, as we assume that they are inappropriate for children; thus, these should appear as low as possible. This way, we prioritize content which we predict to be appropriate over content for which we lack information, and we demote content we assume to be harmful. This approach acknowledges the uncertainty of unknowns while still offering a principled way to structure the recommendation list based on what is known about video appropriateness.

4.1.3. Strategies for Reranking

This study explores four distinct reranking approaches that assign relevance scores to candidate videos based on different types of information. The first approach relies on predicted labels from our trained classifier and the confidence of these predictions. While this method leverages the power of supervised learning, it is inherently sensitive to misclassifications and errors in confidence. The second and third strategies are feature-based, using the video's category ID and view count, respectively, both of which were identified as highly informative during feature correlation analysis. These approaches are less dependent on the classifier's accuracy and provide a more interpretable and stable scoring mechanism based on known label distributions. Finally, a fusion strategy aggregates the signals from the individual reranking approaches using two standard rank aggregation techniques: CombSUM and CombMNZ.

These reranking strategies offer different ways of prioritizing videos that are likely to be fitting for young children while pushing potentially harmful content further down the list. In the following sections, we provide a detailed explanation of the logic and computation behind each approach.

Incorporating Classifier-Based Scores

The first reranking strategy directly leverages the output of the trained classifier introduced in Chapter 3. For each video in the recommendation list that is not *unknown*, the classifier predicts a label, i.e., *suitable*, *irrelevant*, *disturbing*, or *restricted*, and assigns a confidence score to that prediction. To rerank the recommendations, we convert these predictions into numerical scores that reflect how desirable the video is for recommendation to young children.

To approximate an ideal ranking, we assign the highest score to videos predicted as *suitable* since these are explicitly targeted toward young children and provide safe, age-appropriate content. Videos labeled *irrelevant* are given a moderately positive score, as they are not harmful but also not tailored to children's needs. On the other end of the spectrum, we assign strongly negative scores to videos classified as *disturbing*, which are considered the most harmful and deceptive in nature. *Restricted* videos receive a moderate negative score to reflect their clear inappropriateness, even if they may not be directly targeted at children. This way, all videos assumed to be appropriate receive a positive score, while all items predicted to be inappropriate receive a negative score, ensuring that positive videos appear above inappropriate ones. Videos that are *unknown*, for which the classifier cannot make a prediction, receive a score of 0, ensuring that no unknown, and potentially harmful, video appears above a video that is predicted to be appropriate and that items that are predicted to be inappropriate definitely a placed at the bottom of the list. So, the scoring will be as follows:

- **Suitable:** +2.0 (High score)
- **Irrelevant:** +1.0 (Moderate score)
- **Unknown:** 0.0
- **Restricted:** -1.0 (Moderate negative score)
- **Disturbing:** -2.0 (High negative score)

For all *non-unknown* videos, the final score used for reranking is calculated by multiplying the assigned label score with the classifier's confidence in its prediction. This allows the reranking strategy to account not only for the predicted type but also for how certain the classifier is about its decision. In doing so, we avoid treating all predictions equally; high-confidence predictions have more impact on the ranking, while low-confidence ones contribute less. For example, a video predicted as *irrelevant* with very high certainty may receive a slightly higher score than a video predicted as *suitable* with only marginal confidence. This scoring reflects the practical uncertainty in classification and helps temper over-reliance on potentially unreliable predictions. The computation is defined as follows:

$$rank_{final} = score_{label} * confidence_{prediction} \tag{4.1}$$

where $score_{label}$ is the numeric value associated with the predicted label, and $confidence_{prediction}$ reflects the model's probability estimate for that label. The label score is retrieved using a simple lookup function:

$$score_{label} = LOOKUP(label_{prediction}) \tag{4.2}$$

This method provides a simple but effective way to align the recommendation ranking with the classifier's predictions of appropriateness

CategoryId-Based Approach
The second reranking strategy builds on one of the most statistically informative features identified in the correlation analysis in Chapter 3: *categoryId*. As shown earlier, this platform-assigned category was not only strongly correlated with video appropriateness but also exhibited relatively clear label-specific distributions. Given its interpretability and consistent relevance across labels, *categoryId* offers a promising signal for guiding reranking decisions.

This approach assigns ranking scores based on the expected appropriateness of a video given its *categoryId*. Instead of using a classifier prediction, we derive the likelihood that a video with a given category falls under each of the four appropriateness labels, based on distributions observed in the training data. Each label is again matched to the same scoring scheme as used in the classifier-based strategy. By combining these label-specific scores with the conditional probability of a label given a *categoryId*, we obtain a reranking score that reflects how likely a video from that category is to be fitting for children.

To compute this, we first estimate the probability of each label occurring within each *categoryId* using the data from the *Classifier Train-Test Set*. This is calculated with the following formula:

$$P(label|id_{category}) = \frac{\text{\#videos with } \textit{categoryId} \text{ and label}}{\text{\#total videos with } \textit{categoryId}} \tag{4.3}$$

From this distribution, we select the label with the highest probability and apply the score associated with that label. This score is then scaled by the probability itself to reflect the confidence in that outcome:

$$rank_{final} = max_{label} P(label|id_{category}) * probability_{id_{category}} \tag{4.4}$$

This process ensures that videos from categories with a strong association to suitable or irrelevant labels are ranked higher, while those from categories frequently associated with restricted or disturbing content are pushed lower. Unlike the classifier-based approach, this method does not rely on a trained model to predict appropriateness from video metadata. Instead, it leverages known distributional patterns across video categories to inform reranking in a simple way. Because *categoryId* is a platform-defined attribute already available for every video, this strategy is lightweight to implement and easily scalable across large video sets. Furthermore, it can act as a reliable fallback strategy when metadata is incomplete or a classifier is unavailable or uncertain, providing robust support in cold-start or low-confidence scenarios.

ViewCount-Based Approach
The third strategy incorporates another feature-based scoring approach, this time leveraging the second highest feature identified as most important during the feature analysis: *viewCount*. Similar to the *categoryId*-based approach, this method estimates the likelihood of each label occurring at a given view count and maps that label to a predefined reranking score.

Using the same scoring scale as before, i.e., assigning higher scores to appropriate labels and negative scores to inappropriate ones, we compute the conditional probability of a label given a specific view count:

$$P(label|count_{views}) = \frac{\text{\#videos with } \textit{viewCount} \text{ and label}}{\text{\#total videos with } \textit{viewCount}} \tag{4.5}$$

We then identify the label with the highest probability and apply its associated score, scaled by the probability itself:

$$rank_{final} = max_{label} P(label|count_{views}) * probability_{count_{views}} \tag{4.6}$$

Unlike *categoryId*, which has a fixed and limited set of values, *viewCount* is a continuous and highly variable numerical feature. As a result, not every encountered value has a corresponding entry in the scoring table. To address this, we find the closest matching view count from the training data and use its label distribution as a proxy. This approximation allows the strategy to generalize to new or unseen view counts during reranking while preserving consistency in the scoring logic.

This approach offers a straightforward yet flexible mechanism for capturing general trends in how viewership relates to content appropriateness.

Fusion-Based Approaches

The final reranking strategies presented in this study aggregate the outputs of the previously introduced methods, namely the classifier-based, categoryId-based, and viewCount-based strategies, into unified rankings. These Fusion-based approaches aim to balance the strengths of individual strategies and mitigate their respective weaknesses. By combining the classifier's metadata-driven predictions with the broader statistical signals provided by the *categoryId* and *viewCount* distributions, the fusion methods seek to enhance overall ranking performance. While this integration is intended to produce more reliable recommendations, especially in the presence of uncertainty or feature sparsity, the actual impact depends on the interactions between the underlying signals.

To perform this aggregation, we implement two separate and widely used data fusion techniques: `CombMNZ` [63] and `CombSUM` [64]. Instead of combining the methods into a single fusion model, each of these strategies represents a distinct way of combining input from multiple reranking signals to produce an improved recommendation order.

Before fusion, scores from each individual strategy are normalized using min-max normalization to bring them to a common scale:

$$V^c = \frac{S^v - V^c_{min}}{V^c_{max} - V^c_{min}} \tag{4.7}$$

where $S^v$is the raw score from strategy c for video v, and V^c_{min}, V^c_{max} are the minimum and maximum scores generated by that strategy across the recommendation lists.

The `CombMNZ` strategy adds the normalized scores from all contributing strategies and multiplies this sum by the number of strategies that gave the video a non-zero score:

$$rank_{final} = \textit{CombSUM} \times |V^c > 0| \tag{4.8}$$

The `CombSUM` strategy, on the other hand, simply sums the normalized scores from each strategy without adjusting for the number of contributing sources:

$$rank_{final} = \sum_{c}^{N} V^c \tag{4.9}$$

Together, these two fusion-based reranking strategies are designed to serve as a robust ensemble alternative that avoids over-reliance on any single feature or model. By aggregating scoring outputs across the individual reranking models, the fusion strategies aim to produce more stable rankings that better reflect our objective of prioritizing appropriate content for young children.

4.1.4. Evaluation Metrics

The effectiveness of the reranking approaches is assessed using several evaluation metrics, each chosen to reflect a different aspect of appropriateness-aware recommendation quality. Together, they offer a comprehensive assessment of how effectively each strategy promotes content fitting for young children

while minimizing the presence of inappropriate videos. While the length of the recommendation lists in our dataset varies across entries, all selected metrics are either inherently normalized by list length or designed to account for it explicitly, allowing for consistent comparison across rankings of different sizes.

To enable consistent comparisons across different categories of content, we use a single variable, $app-type$, to represent the appropriateness label of interest in each metric. This variable can take on one of six values: *suitable*, *irrelevant*, *restricted*, *disturbing*, *appropriate* (suitable or irrelevant), or *inappropriate* (restricted or disturbing). Using this unified notation allows us to apply the same metric formulations to assess both positive effects, such as prioritization of suitable content, and negative effects, such as undesired exposure to disturbing videos. In the remainder of this section, we describe how each evaluation metric is adapted to incorporate $app-type$ and what aspects of ranking quality it is designed to capture:

MRR$_{app-type}$ (Mean Reciprocal Rank Appropriateness Type) Adapted from the MRR metric [58], which typically measures the ranking quality of search results performance by assessing the position of the first relevant item in a list. Our adapted metric captures the average rank position of the first video of a given type across all ranked lists. For instance, $MRR_{suitable}$ indicates the average position across all lists at which the first suitable video appears. Similarly, $MRR_{disturbing}$ reflects the average rank of the first disturbing video, where lower values indicate more successful suppression of harmful content.

$$MMR_{app-type} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{rank_q^{app-type}} \tag{4.10}$$

where $rank_q^{app-type}$ is the rank of the first occurrence of type $app-type$ in list q.

MAP$_{app-type}$ (Mean Average Precision Appropriateness Type) Based on the MAP metric [58], which typically measures the ranking quality of search results by assessing how well relevant items are distributed across a ranked list. The metric rewards rankings in which relevant items appear earlier and are more consistently placed throughout the list. Our adapted metric calculates the average precision for items of a specific type (denoted as $app-type$), e.g., appropriate, inappropriate, suitable, irrelevant, restricted, or disturbing. It reflects how well videos of that type are ranked throughout all lists. We use this to separately assess how well each reranking strategy promotes appropriate or suppresses inappropriate content.

$$MAP_{app-type} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{|R_q|} \sum_{k=1}^{N} P(k) \cdot type_{q,k}^{app-type} \tag{4.11}$$

where Q is the set of all ranked lists, R_q is the number of items of a given type in list q, N is the top-N recommendations, i.e., the length of list q, $P(k)$ is the precision at position k, defined as the proportion of videos of type $app-type$ among the top k videos in the list, and $type_{q,k}^{app-type}$ is 1 if the label of the video at rank k in list q equals $app-type$, 0 otherwise.

HitRate@1$_{app-type}$ (Hit Rate at Rank 1 Appropriateness Type) Derived from the $HitRate@k$ metric [58], which typically evaluates whether an item of interest appears among the top k positions in a ranked list. Our adapted metric measures whether a video of a specific appropriateness type appears at the top, i.e., rank 1, of the recommendation list. For example, $HitRate@1_{suitable} = 1$ if the first recommended video is suitable, and 0 otherwise. It is particularly useful for assessing how often the reranking strategy succeeds in surfacing appropriate content immediately or avoids surfacing inappropriate content first. This is especially relevant in child-oriented platforms where autoplay functionality often plays the top-related video by default, and young children tend to select one of the first two recommendations presented on screen [56, 57].

$$HitRate@1_{app-type} = \frac{1}{|Q|} \sum_{q \in Q} type_{q,1}^{app-type} \tag{4.12}$$

where Q is the set of all lists, $type_{q,1}^{app-type}$ is 1 if the top-ranked item, i.e., position 1, in list q equals $app-type$, 0 otherwise.

NDCG$_{app-type}$ (Normalized Discounted Cumulative Gain per Appropriateness Type) Derived from the $NDCG$ metric [58], which typically measures ranking quality by evaluating how highly relevant items are placed in a list, with a stronger emphasis on items appearing near the top. Our adapted metric calculates how well the videos of a specific appropriateness type are prioritized within each recommendation list. This adaptation enables us to assess both positive and negative outcomes. For example, a high NDCG$_{suitable}$ reflects the effective promotion of appropriate content, while a low NDCG$_{disturbing}$ indicates the successful suppression of harmful videos.

$$DCG_{app-type} = \sum_{i=1}^{N} \frac{type_i^{app-type}}{log_2(i+1)} \tag{4.13}$$

$$NDCG_{app-type} = \frac{DCG_{app-type}}{IDCG_{app-type}} \tag{4.14}$$

where $type_i^{app-type}$ is 1 if the label of the video at rank i equals $app-type$, else 0. $IDCG$ denotes the maximum possible DCG, i.e., the ideal ranking where all videos of type $app-type$ are placed at the top of the list, ordered such that all videos that equal $app-type$ appear as early in the ranking as possible. A higher $NDCG$ indicates better prioritization of videos of a given type.

REC-INAP (Recommendations Inappropriateness Indicator) This metric is based on the REC-ST metric [59, 60], which accounts for the number of results among the top-N recommendations that convey stereotypes. Our metric, REC-INAP, captures the number of items among the top-N recommendations that convey inappropriate content while considering their ranking position. This will help us assess the severity of inappropriate recommendations after reranking.

$$\textit{REC-INAP@}N = \frac{1}{|Q|} \sum_{q \in Q} \frac{\sum_{i=1}^{N} (type_{q,i}^{inappropriate} \cdot (N-i+1))}{N \cdot (N+1)/2} \tag{4.15}$$

where Q is the set of all ranked recommendation lists, $type_{q,i}^{inappropriate}$ is an indicator function that equals 1 if the video at rank i in list q is labeled as inappropriate and 0 otherwise, and $(N-i+1)$ is a weight that emphasizes items ranked closer to the top of the list. The denominator $\frac{N \cdot (N+1)}{2}$ normalizes the weighted sum, ensuring the final score lies in the range $[0, 1]$. This weighted formulation ensures that inappropriate content appearing higher in the list is penalized more severely.

4.2. Results

To evaluate the effectiveness of the reranking strategies, we apply each strategy to every related video list, i.e., the original recommendation list created by YouTube, in the *Unseen Reranking Set* and compute their performance using a suite of appropriateness-aware evaluation metrics. These include NDCG$_{app-type}$, MAP$_{app-type}$, MRR$_{app-type}$, HitRate@1$_{app-type}$, and REC-INAP. For each metric, we report the mean values across all recommendation lists to ensure consistency in evaluation. Together, these metrics cover both the overall ranking appropriateness for children and the exposure-related risks for specific video types.

4.2.1. Baseline Performance

We use the original ranking order of YouTube's related videos as our baseline. As shown in Tables 4.1, this baseline performs modestly across all metrics designed to measure the promotion of appropriate content and suppression of inappropriate material. For instance, NDCG$_{appropriate}$ is only 0.40, indicating that suitable and irrelevant videos are not strongly prioritized near the top of YouTube's related video lists. Even more concerning are the low scores for NDCG$_{suitable}$ (0.34) and HitRate@1$_{suitable}$ (0.14), which indicate that suitable content rarely appears at the very top of the list and is inconsistently emphasized across the ranked results.

Furthermore, inappropriate content is not entirely excluded from top positions. While baseline scores for inappropriate categories remain relatively low, they are far from negligible. Most notably,

HitRate@1$_{disturbing}$ is 0.01, indicating that in 1% of the related video lists, a disturbing video appears as the very first recommendation. Given that young children frequently select one of the first two recommended videos [56, 57], this translates into a serious real-world exposure risk, even if the frequency seems low in absolute terms.

In short, the baseline rankings from YouTube offer limited protection against inappropriate exposure and lack consistent prioritization of the most fitting content. These results emphasize the need for more refined reranking strategies that better align with child-appropriateness goals.

4.2.2. Effectiveness of Individual Reranking Strategies

The reranking strategies implemented in this study produce substantial improvements across multiple appropriateness-aware evaluation metrics. Notably, all approaches outperform the original YouTube ranking across metrics that measure the promotion of fitting content. For instance, $nDCG_{appropriate}$ increases from 0.40 to as high as 0.73, while $MAP_{appropriate}$ rises from 0.26 to 0.72. Similarly, $MRR_{appropriate}$ and $HitRate@1_{appropriate}$ more than double, indicating that fitting content appears significantly earlier in the reranked lists and more frequently in the top position.

However, these gains come with trade-offs. Metrics reflecting the ranking quality of inappropriate content, such as NDCG$_{inappropriate}$, MAP$_{inappropriate}$, and HitRate@1$_{inappropriate}$, also increase. This is undesirable, as it suggests that some inappropriate videos are ranked higher than before. For example, MAP$_{inappropriate}$ increases from 0.14 to 0.33 in the CategoryId-based approach, and HitRate@1$_{inappropriate}$ rises from 0.09 to as high as 0.31. One of the likely reasons is the residual inaccuracy of the classifier, which may misclassify restricted or disturbing videos as less harmful types, thereby inadvertently boosting their position. Furthermore, the presence of unknown items in recommendation lists complicates accurate ranking, as these items remain in fixed positions. As a result, misclassified inappropriate videos can end up relatively higher in the final ranked list simply due to limited flexibility in reordering.

While these overall increases are modest in scale and do not outweigh the substantial gains in promoting suitable content, it is important to acknowledge the trade-off. For example, although MAP$_{inappropriate}$ increases from 0.14 to as high as 0.33, MAP$_{suitable}$ increases more substantially, from 0.22 to 0.62. Likewise, the improvement in HitRate@1$_{suitable}$ from 0.14 to 0.62 exceeds the corresponding rise in HitRate@1$_{inappropriate}$, which increases from 0.09 to 0.31. These figures suggest that, overall, the reranking strategies are more effective at promoting suitable content than they are at elevating inappropriate content.

However, the rise in HitRate@1$_{inappropriate}$ is particularly concerning. Young children tend to click on the top result in recommendation lists, often selecting one of the first two items presented to them on screen [56, 57]. As such, the presence of even a single inappropriate video at rank 1 poses a serious exposure risk. This finding reinforces the importance of not only improving average performance but also minimizing high-impact ranking failures at the top of the list.

Despite these challenges, the overall improvement in prioritizing suitable videos, while still partially suppressing inappropriate ones, demonstrates the value of the reranking approaches. Below, we analyze each strategy in more detail.

Classifier-Based Strategy

The classifier-based reranking approach shows strong improvements across all appropriate-targeted metrics. NDCG$_{appropriate}$ rises to 0.71, and both MRR$_{appropriate}$ and HitRate@1$_{appropriate}$ reach 0.71 and 0.69, respectively. This highlights that the classifier's label predictions, when converted into reranking scores, are effective in promoting appropriate videos earlier in the list. However, this method also leads to increases in inappropriate content metrics. MAP$_{inappropriate}$ rises to 0.27, and HitRate@1$_{inappropriate}$ increases to 0.27. These increases are likely due to misclassifications from the underlying classifier, especially when confidence scores are not strong enough to push clearly inappropriate videos further down.

CategoryId-Based Strategy

The CategoryId strategy achieves the highest NDCG$_{appropriate}$ (0.73) and MAP$_{appropriate}$ (0.72) among all individual strategies. It also performs best on most label-specific metrics, particularly for NDCG$_{suitable}$

(0.62) and MAP$_{suitable}$ (0.61). This highlights the utility of the *categoryId* feature, which seems to correlate relatively well with video appropriateness. Nonetheless, this method also results in the highest MAP$_{inappropriate}$ (0.33) and HitRate@1$_{inappropriate}$ (0.31). This is likely due to underrepresented categories in the dataset, such as the *Trailers* category, which has a very small volume consisting only of restricted videos, or because certain categories, although generally appropriate, contain a small subset of inappropriate videos that are not distinguishable at the category level alone.

ViewCount-Based Strategy
The ViewCount-based strategy shows the lowest overall improvement among the reranking approaches. While it increases nDCG$_{appropriate}$ to 0.66 and MAP$_{appropriate}$ to 0.59, its effectiveness lags behind the classifier- and categoryId-based methods. This suggests that while the $viewCount$ feature was statistically important during feature analysis, it may not be as precise a signal when used in isolation for reranking.

In terms of suppressing inappropriate content, this strategy performs comparably to the classifier-based method. MAP$_{inappropriate}$ remains relatively low at 0.26, and HitRate@1$_{inappropriate}$ is kept at 0.27.

Fusion Strategies: CombMNZ & CombSUM
The fusion strategies outperform all individual methods in terms of promoting appropriate content. Both nDCG$_{appropriate}$ and MAP$_{appropriate}$ reach 0.73 and 0.72, respectively, and MRR$_{appropriate}$ peaks at 0.73. Additionally, HitRate@1$_{suitable}$ achieves the highest values: 0.62 for both fusion methods. This demonstrates the effectiveness of combining signals from multiple strategies, balancing the strengths of each individual approach.

While MAP$_{inappropriate}$ and HitRate@1$_{inappropriate}$ also increase moderately (up to 0.31 and 0.29, respectively), the overall trade-off remains favorable. However, the increase in HitRate@1$_{inappropriate}$ is particularly concerning. Although the fusion-based reranking significantly boosts the presence of appropriate content at the top, even small increases in the likelihood of inappropriate videos being ranked first must be carefully considered in safety-critical recommendation settings for children.

Overall, the results demonstrate that reranking can meaningfully improve the prioritization of appropriate content in child-directed recommendations, with fusion-based strategies achieving the most balanced performance across metrics. At the same time, the observed increases in exposure to inappropriate videos, particularly in the top-ranked positions, highlight important limitations and trade-offs. The following discussion reflects on these results, considering both their practical implications and the challenges that remain for designing safer, appropriateness-aware recommendation systems.

Table 4.1: Ranking strategies results

Metric	YouTube	Classifier-based	Categoryid-based	Viewcount-based	CombMNZ	CombSUM
MRR$_{appropriate}$	0.31	0.71	0.72	0.69	**0.73**	**0.73**
MRR$_{inappropriate}$	**0.16**	0.30	0.33	0.30	0.32	0.32
MAP$_{appropriate}$	0.26	0.68	**0.72**	0.59	**0.72**	**0.72**
MAP$_{inappropriate}$	**0.14**	0.27	0.33	0.26	0.31	0.31
HitRate@1$_{appropriate}$	0.17	0.69	0.69	0.68	0.71	**0.72**
HitRate@1$_{inappropriate}$	**0.09**	0.27	0.31	0.27	0.29	0.28
NDCG$_{appropriate}$	0.40	0.71	**0.73**	0.66	**0.73**	**0.73**
NDCG$_{inappropriate}$	**0.21**	0.30	0.34	0.29	0.33	0.33
REC-INAP	**0.09**	0.11	0.14	0.11	0.13	0.13
MRR$_{suitable}$	0.26	0.61	0.61	0.60	**0.63**	**0.63**
MRR$_{irrelevant}$	0.05	0.11	**0.13**	0.11	0.12	0.12
MRR$_{restricted}$	**0.12**	0.20	0.21	0.20	0.21	0.21
MRR$_{disturbing}$	**0.05**	0.11	0.13	0.11	0.12	0.12
MAP$_{suitable}$	0.22	0.58	0.61	0.50	**0.62**	**0.62**
MAP$_{irrelevant}$	0.04	0.11	**0.12**	0.11	**0.12**	**0.12**
MAP$_{restricted}$	**0.10**	0.18	0.21	0.17	0.20	0.20
MAP$_{disturbing}$	**0.04**	0.10	0.13	0.10	0.12	0.11
HitRate@1$_{suitable}$	0.14	0.59	0.58	0.58	**0.62**	**0.62**
HitRate@1$_{irrelevant}$	0.03	0.09	**0.11**	0.09	0.09	0.10
HitRate@1$_{restricted}$	**0.08**	0.19	0.20	0.19	0.21	0.20
HitRate@1$_{disturbing}$	**0.01**	0.08	0.11	0.08	0.08	0.08
NDCG$_{suitable}$	0.34	0.61	0.62	0.56	**0.63**	**0.63**
NDCG$_{irrelevant}$	0.07	0.12	**0.13**	0.12	**0.13**	**0.13**
NDCG$_{restricted}$	**0.14**	0.20	0.21	0.19	0.21	0.21
NDCG$_{disturbing}$	**0.07**	0.12	0.14	0.12	0.13	0.13

4.3. Discussion

In this chapter, we demonstrated how score-based reranking strategies can be leveraged to improve the appropriateness of video recommendations for young children. Building upon the insights from our feature analysis and classifier predictions, we designed three reranking strategies, i.e., classifier-based, categoryId-based, and viewCount-based, that aim to elevate appropriate content, especially suitable content, while demoting inappropriate videos. Additionally, we applied two fusion-based methods, namely CombMNZ and CombSUM, to combine signals from multiple strategies, thereby reinforcing the overall effectiveness of our strategies.

All reranking approaches show improvements over the original YouTube ranking on multiple key metrics related to the prioritization of appropriate content. This is evidenced by marked improvements in NDCG$_{appropriate}$, MAP$_{appropriate}$, MRR$_{appropriate}$, and HitRate@1$_{appropriate}$. Suitable videos appear earlier in the lists and more frequently in the top-ranked position, making them more accessible to young viewers. Fusion-based strategies, in particular, offer the most consistent gains across metrics, reinforcing the value of combining different signals to steer recommendations toward safer content.

However, these gains come with notable trade-offs. Metrics that capture the ranking of inappropriate content, such as MAP$_{inappropriate}$, MRR$_{inappropriate}$, and HitRate@1$_{inappropriate}$, also increase across all strategies. This is highly undesirable, especially when examining HitRate@1$inappropriate$, which rises from 0.09 in the original YouTube ranking to as high as 0.31 after reranking. This increase is particularly noteworthy given the critical importance of the top-ranked item in shaping young children's viewing behavior. Young children often click on one of the first two recommended items [56, 57]. Therefore, even modest increases in inappropriate content at rank 1 may affect exposure risk.

Several factors likely contribute to this undesired side effect. First, some inappropriate videos may be misclassified by the underlying classifier and, thus, mistakenly promoted in reranking. This issue becomes more pronounced when such misclassifications are accompanied by high confidence scores, causing these

videos to be weighted more heavily in the reranking process. Second, the presence of unknown items, videos without metadata or ground-truth labels, limits the flexibility of reranking, as these items cannot be moved. As a result, misclassified videos, especially those incorrectly predicted as suitable with high confidence, may be disproportionately promoted, pushing them to the top of the ranked list and increasing potential exposure risk. Third, while certain features, such as *viewCount*, are statistically important, they are less informative when used in isolation, potentially leading to imprecise label probability estimates. Finally, limited data volume for certain features can result in skewed or overestimated label distributions, particularly in strategies that rely on aggregated statistics, such as *categoryId* or *viewCount* mappings.

Together, these factors help explain why the reranking strategies, while clearly improving the prioritization of appropriate content, especially suitable videos, consistently perform worse than YouTube's ranking in suppressing inappropriate content. This dual outcome highlights the inherent trade-offs and complexity involved in optimizing for safety in recommender systems: optimizing for one objective, such as suitability, can inadvertently compromise another, such as minimizing exposure to harmful videos. A challenge that requires careful balancing of competing goals and reliable signals.

These findings raise important questions about system design, classifier reliability, and safety guarantees, topics that will be addressed in Chapter 6.

5

Ethical Considerations

The experiments described in chapters 3 and 4 were conducted in a thoughtful, transparent, and reproducible manner. All methods are fully described, and each step can be reproduced following the details in the manuscript. No data was fabricated or falsified, and we report both positive and negative results, including limitations and trade-offs observed during our evaluation. We have presented work, data, and ideas of others with the appropriate medium of presentation, i.e., citations. All cited work can be found in the reference list. Finally, there is no conflict of interest to declare.

5.1. Data Management

In this research, a robust Data Management Plan (DMP) was meticulously developed and approved according to TU Delft's stringent requirements, ensuring ethical compliance and integrity throughout this study. This research utilizes a ground-truth dataset originally compiled by Papadamou et al. [24], which includes metadata of publicly available YouTube videos. This dataset was chosen due to its relevance to our research goals and its curated annotations that distinguish between appropriate and inappropriate content for toddlers. The dataset does not include personalized user data, containing no identifiable information about individual viewers or content creators. Although certain elements of the metadata, such as thumbnail images, could in principle be traced back to specific content creators, this research does not involve any effort to identify or profile individuals or channels. Importantly, while the dataset includes videos that may have been recommended to children, it is not based on data collected from actual children. All data was secured on TU Delft-approved servers, with regular backups and restricted access. Detailed documentation is maintained to ensure the transparency of data origins, modifications, and methodological applications, facilitating reproducibility and adherence to ethical research norms.

5.2. Ethics

Ethical integrity guided each phase of this research. As the dataset consists entirely of public, non-personal metadata, privacy concerns typically associated with personal data collection were avoided. We refrained from engaging with or analyzing any data that would allow for reidentification of individuals, and we did not interact with human subjects in any form. Our analysis was focused solely on video-level metadata and algorithmic patterns.

Transparency is a central value of this work. Our methodology, feature selection process, classifier performance, and reranking outcomes are thoroughly documented. We deliberately report both successes and limitations, such as the classifier's generalization challenges and the unintended promotion of misclassified inappropriate videos, to foster a responsible and accurate representation of our findings.

In evaluating algorithmic predictions and recommendation strategies, we took care to reflect critically on bias and fairness. Special attention was given to understanding how model decisions might differentially affect exposure to inappropriate content. The classifier and reranking strategies were evaluated not only on aggregate performance but also in terms of their broader implications for safety and content prioritization.

5.3. Compliance and Ethical Responsibility

This research rigorously follows all prescribed guidelines by TU Delft, especially regarding ethical conduct and data management. Consultations with data management support staff have been documented, integrating their recommendations into the research design to enhance ethical compliance.

Although our study does not involve direct personal data processing or collection, it aligns with broader ethical standards and regulatory frameworks that govern digital content and child protection. These practices underscore our commitment to conducting research that is both ethically sound and socially responsible.

5.4. Reflective Ethical Statement

Engaging in this study has broadened our understanding of the ethical complexities involved in analyzing digital content aimed at children. By analyzing algorithmic content curation through the lens of appropriateness, we gained a deeper appreciation for the societal impact of recommender systems and content moderation strategies. Our goal, to reduce young children's exposure to inappropriate content while promoting developmentally aligned media, was pursued through technical contributions that can inform future system designs.

At the same time, we recognize that interventions in algorithmic ranking and content classification are not without risk. Misclassifications, if deployed in real-world systems without safeguards, may inadvertently suppress beneficial content or promote harmful material. These risks are particularly concerning in child-facing environments, where even minor failures can have significant consequences. As such, we have taken care to report both the strengths and limitations of our approach transparently, acknowledging the trade-offs involved in classification accuracy, ranking precision, and uncertainty handling.

This research contributes not only to the academic discourse but also to ongoing conversations about platform responsibility, transparency, and the ethical design of child-facing technologies. By reflecting on both the potential benefits and the risks of our work, we emphasize the need for continued ethical vigilance in the development of tools that shape children's digital experiences.

6

Reindeer... Safeguards are Better than People: Discussion

In this chapter, we reflect on the findings presented in Chapter 3 and Chapter 4, situating them within the broader discourse on child safety, algorithmic decision-making, and content moderation on online video platforms. Here, we not only summarize key results but interpret their implications, limitations, and opportunities for action.

The classification and reranking experiments presented throughout this manuscript offer practical insights into how feature-based strategies can support safer content delivery to young children. Throughout this discussion, we reflect critically on challenges uncovered by our experiments and their implications for the future of recommender design, regulation, and parental agency. We begin by revisiting the two central research questions that guided this work, summarizing how the empirical findings address them and where open questions remain. We then reflect on the broader implications, both technical and societal. We continue with an overview of the study's limitations and conclude with concrete directions for future work.

6.1. Answers to the Research Questions

This section revisits the two central research questions posed in Chapter 1. The first question addresses the identification of appropriate and inappropriate content for young children based on YouTube video metadata, while the second explores the role of these insights in shaping safer recommender systems. Below, we discuss how the results inform each question in turn and reflect on the broader significance of the findings.

6.1.1. RQ1: What key features derived from video metadata characterize inappropriate content in young children's online video recommendation?

Our feature analysis and classification results shed light on the complex relationship between video metadata and content appropriateness for young children. While many metadata features were found to carry weak or indirect signals individually, several demonstrated meaningful correlations with the labels in the ground-truth dataset. These included structural features such as *categoryId*, *licensedContent*, and *defaultAudioLanguage*, as well as engagement features like *viewCount* and the emotional tone of textual fields (e.g., *descriptionEmotions.joy* and *titleEmotions.disgust*).

Through correlation analysis, we validated that these features were not only statistically associated with the classification labels but also offered interpretability in terms of their effect. For example, disturbing and restricted videos often exhibited different language patterns and licensing characteristics compared to suitable content. Some features, like *categoryId*, revealed certain genre-based affinities, with specific YouTube content categories disproportionately linked to inappropriate content.

However, our findings also demonstrate that metadata-based feature signals alone are rarely sufficient for clear-cut classification. Many of the distinguishing patterns we observed, such as emotion distributions or language complexity in video descriptions, show overlap across labels, making it difficult to establish firm boundaries between appropriateness types based on a single signal. This is particularly true for

disturbing videos, which are handcrafted to mimic the surface characteristics of child-appropriate content [24, 6, 37]. In such cases, engagement metrics or benign-sounding descriptions may mask the underlying harm, making automated detection based solely on metadata a formidable challenge.

Despite these findings, the selected features form a strong foundation for classifier development. Even with relatively weak individual signals, their combination allowed for a reasonably accurate classification model, especially when applied to data distributions similar to those seen during training. These results support the use of metadata-derived features as a valuable input space for detecting video-level inappropriateness while also reinforcing the importance of combining signals and validating them across multiple methods.

6.1.2. RQ2: To what degree can features deemed relevant for predicting inappropriateness contribute to the mitigation of young children's exposure to inappropriate content by recommender systems?

The second stage of this study translates the insights gained from feature analysis and our classifier into action by implementing a series of reranking strategies. These approaches aim to reduce young children's exposure to inappropriate videos by promoting content that is either predicted to be suitable or aligned with safer metadata-derived feature signals.

To understand the real-world impact of our strategies, we must consider how children interact with video platforms in practice. When using YouTube or YouTube Kids, young children are typically presented with ranked lists of recommendations. These lists are often influenced by engagement signals like view count, which the platform uses to prioritize content that appears popular or engaging to others [26]. Prior work shows that children tend to click on one of the first two results in these lists [56, 57]. If those top-ranked videos include content that is inappropriate, the child is exposed before any reporting or manual moderation mechanisms have time to intervene. This makes it especially important that content that aligns best with the development of young children, which supports early learning, language development, and emotional regulation [5], appears at the top of the recommendation list, while inappropriate content, such as horror-themed videos that may provoke prolonged anxiety, recurring nightmares, or avoidance behaviors [10], does not appear in the top-ranked results.

Our evaluation of YouTube's original ranking highlights how it fails to prioritize appropriate content. Suitable videos appear inconsistently in top positions, with HitRate@1$_{suitable}$ at only 0.14, limiting exposure to videos that are age-appropriate and relevant to children's development. Meanwhile, inappropriate content is not absent from top-ranked positions: HitRate@1$_{inappropriate}$ remains relatively low at 0.09, but not zero, an important caveat, as even a single instance of exposure to disturbing content may be harmful.

Our reranking strategies, i.e., classifier-based, categoryId-based, viewCount-based, and two fusion methods, demonstrate clear improvements in the surfacing of suitable content. Suitable videos not only appear more frequently but also tend to be ranked higher, as shown by consistent improvements in NDCG$_{suitable}$, MRR$_{suitable}$, and MAP$_{suitable}$ across all strategies. In particular, the probability of a suitable video appearing in the top-ranked position increased substantially, as shown by HitRate@1$_{suitable}$, which jumps from 0.14 in the original YouTube ranking to as high as 0.62 after reranking.

However, these improvements come with an important trade-off. Inappropriate videos, particularly those misclassified with high confidence, are sometimes promoted more aggressively due to the structure of the reranking algorithms and the presence of unknown videos, which cannot be ranked confidently due to missing data. This is especially visible in the rise of HitRate@1$_{inappropriate}$; a concerning outcome, given how influential the first item in a ranked list can be for young viewers. Exposure to inappropriate content can impact young children's development and emotional well-being in various ways [7, 9, 10], making the presence of such content in top-ranked positions a significant concern. Although YouTube's original ranking performs better on this metric (i.e., presents less inappropriate items in top positions), it still allows inappropriate content to appear in top-ranked positions, despite the platform's extensive moderation systems and content restrictions. In this respect, both YouTube's systems and our reranking strategies fail to consistently keep harmful content out of reach, highlighting one of the most pressing and unresolved challenges in recommender systems for young children. Ensuring that inappropriate content, particularly disturbing content, never appears in the top-ranked position should therefore remain a key focus of future safeguarding efforts. While our strategies made significant progress in promoting suitable videos, more sophisticated approaches will be needed to suppress inappropriate ones with equal reliability.

Taken together, these findings show that metadata-informed classifiers can contribute meaningfully to safer recommendation outcomes, but only when applied within a layered and uncertainty-aware framework. Based on our observations, particularly the trade-off between improved suitability ranking and the unintended promotion of inappropriate videos, we propose a two-layer approach for real-world deployment. The first layer would use the classifier as a hard filter to remove videos confidently identified as inappropriate, ensuring they do not appear anywhere in the recommendation list. The second layer would integrate classifier predictions as soft signals into the ranking process, helping prioritize videos with strong suitability signals at the top while demoting those with uncertain predictions, which might otherwise be promoted due to high engagement metrics. Together, these two layers address the risk posed by high-confidence misclassifications and the inability to recover once an inappropriate video is top-ranked. Although such a system would still be limited by the accuracy of classification and our understanding of inappropriateness, it would substantially reduce the likelihood of inappropriate videos being top-ranked and increase the visibility of suitable content. This layered design reflects a pragmatic balance between safety and flexibility and offers a realistic direction for improving recommender systems that may reach children.

6.2. Implications

The findings of this study not only contribute to the technical understanding of detecting and mitigating inappropriate content for young children but also surface broader implications for recommender system design, platform governance, and child protection. Based on both prior work and the findings of our study, we explore the inherent challenges of the problem, reflect on the limits of current safeguarding mechanisms, and consider the role of classifiers, recommender systems, and caregivers in shaping safer media environments for young children.

6.2.1. Inherent Difficulty of the Problem

This study underscores that distinguishing appropriate from inappropriate videos for young children is an inherently complex task, one that resists simple, automated solutions. As discussed in Section 3.3, the variability in classifier performance across different evaluation settings reveals the sensitivity of models to training distributions, label noise, and subtle content shifts. The sharp performance drop between cross-validation and evaluation on the Unseen Classifier Set illustrates how easily predictive signals can falter when confronted with new or imbalanced data.

One reason for this instability is the nature of inappropriate content itself. Disturbing videos are often intentionally designed to imitate child-friendly aesthetics, embedding inappropriate themes beneath familiar thumbnails or seemingly innocuous metadata [24, 6]. This deceptive design makes such content difficult to identify using surface-level cues alone. Even when strong correlations exist between features and labels, the real-world generalizability of these signals is limited by the deceptive nature of the content.

These challenges highlight why platforms like YouTube and YouTube Kids continue to struggle with inappropriate video detection despite deploying large-scale moderation systems. Research has repeatedly found that harmful content still slips through filters and appears in recommendation loops, often due to algorithmic vulnerabilities and metadata manipulation [21, 20]. The task is not only technically hard but also fundamentally shaped by a malicious dynamic in which bad actors continuously adapt to avoid detection. As our study shows, reliable detection depends not only on sophisticated algorithms but also on an understanding of the broader context in which these systems operate.

6.2.2. Reflections on Platform Safeguards

The results of this study also invite critical reflection on the state of current platform safeguards. While YouTube and YouTube Kids have publicly disclosed technical measures such as automated filters [11], a three-strike enforcement system for repeat violations of community guidelines prohibiting inappropriate material [12, 13], and age-based access tools [14], these mechanisms remain largely reactive rather than preventive. Many additional safeguards rely on user reports and manual review, creating a lag between exposure and intervention, by which point the harm may already be done.

Moreover, these safeguards place a substantial burden on parents, caregivers, and even content creators. Parents are expected to supervise or vet content in environments that lack transparency, while creators must label their videos according to evolving platform rules, which have been widely described as confusing and inconsistently enforced by creators and journalists alike [61], under the threat of penalties.

The platform, meanwhile, retains control over recommendation algorithms and monetization, which remain largely opaque [25]. As shown in prior literature and confirmed by our own findings, inappropriate videos can still surface in recommendation loops, especially when uploaders manipulate metadata to circumvent YouTube's systems. This illustrates that the current safeguarding model, which is distributed, reactive, and algorithmically fragile, is insufficient in addressing the systemic nature of the problem.

A deeper issue lies in the platform's underlying incentive structure. YouTube, like many commercial platforms, is built on an engagement-driven business model, where revenue is generated by maximizing watch time, impressions, and viewer interaction. This logic carries over into YouTube Kids, even if indirectly, as channels targeting young children are still monetized based on views and engagement. As stated in Google's official monetization guidelines, content creators earn revenue through mechanisms that reward high view counts, audience retention, and interaction [65]. Hanif [66], YouTube's Vice President of Creator Products, similarly highlights that creators are incentivized to boost watch time and engagement metrics in order to improve monetization outcomes. This dynamic creates strong incentives for creators to maximize exposure by any means available. As long as content visibility and profitability are tied to popularity metrics, many content creators may be incentivized to do whatever they can to capture attention, whether by flooding thumbnails with popular children's characters, exploiting emotionally charged titles, or faking engagement through artificial means. Shah [32] and Kuchhal and Li [33] document how coordinated view fraud, including fake accounts and bot-generated views, can be used to inflate a video's popularity artificially, making it appear more trustworthy or appealing than it truly is.

Our study contributes to this conversation by demonstrating how metadata-informed classifiers and reranking strategies can support more proactive filtering. However, these tools are not silver bullets. As long as algorithmic incentives prioritize engagement over suitability, the risk of exposure remains built into the system.

These dynamics point to a deeper structural imbalance in the platform's priorities. Even a near-perfect classifier or reranking strategy cannot fully mitigate the risk if the surrounding system continues to reward visibility over child safety. As long as some degree of inaccuracy remains and the dominant algorithmic objective remains engagement maximization, the risk of exposure to inappropriate content is not a bug but a feature, embedded in the design of the system itself.

True safeguarding, then, requires more than improved classifiers or better filters. It demands a shift in platform priorities toward developmental appropriateness and content integrity. A safer system will ultimately require rethinking not only how content is detected and ranked but also how it is rewarded. One long-term vision would involve rethinking the platform's monetization logic itself: What if videos were rewarded not for maximizing attention but for aligning well with child development guidelines? Instead of simply measuring clicks or watch time, reward structures could consider factors like age appropriateness, thematic coherence, emotional tone, and verified creator credibility. Such a shift would not only discourage exploitative content strategies but also incentivize the creation of higher-quality, developmentally supportive media for children.

6.2.3. Classifier as One Layer of Safeguarding

Our findings reinforce the value of treating classification not as a definitive gatekeeper but as one layer in a broader safeguarding strategy. While classifiers can help detect and filter inappropriate content, we argue that they should be complemented by other protective mechanisms, such as a recommender system, human moderation, platform-level content restrictions, and transparent reporting tools, to address the full complexity of child safety on algorithmic platforms.

As demonstrated in Chapter 4, the classifier's predictions have been found to be reasonably effective when integrated into a flexible, score-based ranking system. In addition to using classification to block inappropriate content outright, it can help shift the visibility of videos in a more nuanced way, promoting content with high confidence in suitability while demoting content with lower certainty.

This layered approach is especially important given the limitations of classification models, which can misclassify inappropriate content, particularly when such errors occur with high confidence, leading to those videos being inadvertently promoted in ranking. For example, a disturbing video with a misleading title, description, and thumbnail might be misclassified as suitable and end up ranked first. In such cases, reranking allows for the application of additional constraints or risk-aware weighting schemes, such as

adjusting scores based on high-risk categories, to help moderate the impact of these errors. While not completely eliminating the risk of exposure, this strategy can reduce the likelihood that disturbing content appears in high-ranking positions, which is particularly important for young viewers who disproportionately click on top-ranked items.

6.2.4. Recommender System Improvement

A key implication of this work is that recommender systems can be improved even without perfect classification. The reranking strategies developed in this study demonstrate that leveraging imperfect but informative signals can meaningfully reshape ranked lists in ways that increase the prominence of suitable content. In particular, the fusion-based approaches (i.e., CombSUM and CombMNZ) show that combining multiple scoring signals can mitigate weaknesses in individual features or classifiers.

That said, the trade-offs exposed in the evaluation, especially the unintended promotion of misclassified inappropriate videos, serve as a caution. Optimizing for one objective (e.g., suitability) can compromise another (e.g., exposure risk), especially when system constraints like unknown items limit flexibility. Designing safer recommendation systems, therefore, requires both technical nuance and a willingness to engage with trade-offs directly. Mitigation, not elimination, is a more realistic and scalable framing for safeguarding within recommender pipelines.

6.2.5. Implications for Parents and Guardians

Finally, our findings have important implications for parents and guardians navigating video platforms with or on behalf of their children. While tools like YouTube Kids offer a more curated experience, our evaluation shows that, even if these platforms deploy safeguards and moderation mechanisms, inappropriate videos can still surface through recommendations, sometimes even at the top of the list. This is particularly concerning given that exposure to inappropriate content can impact young children's development and emotional well-being in various ways [7, 9, 10]. Given this potential for harm, the presence of even a small number of inappropriate videos in top-ranked positions challenges the assumption that these platforms are safe by default. It forces a reconsideration of the trust parents place in automated systems to protect their children without constant supervision.

This raises difficult but necessary questions: Would parents still feel confident allowing their child to independently use YouTube or YouTube Kids if they knew that disturbing content might appear at the top of search results or recommendation lists? Would their trust in the platforms shift if they understood how easily malicious content can exploit surface-level features to bypass safeguards?

These questions do not have simple answers. Many parents rely on platforms like YouTube Kids as convenient, on-demand entertainment and learning tools, especially in moments where active supervision is not feasible. In these contexts, trust in the platform plays a central role. However, our findings show that even carefully monitored systems like YouTube Kids (i.e., those with curated environments, restricted content types, and simplified interfaces) can fail and that children may still encounter inappropriate material, particularly if misclassifications or engagement-driven promotion bring such content into top-ranked positions. As such, parents and guardians deserve greater transparency about what risks remain and what steps they can take to mitigate them.

At the same time, it is unrealistic to expect parents to monitor every video their child watches. The sheer volume and rapid change of online video content make individual oversight impractical, especially in households where digital media plays a daily role in education, entertainment, or childcare. Children often navigate apps independently and may consume dozens of videos in a single session, watching for hours on end [4]. In these environments, it becomes increasingly difficult for families to shoulder the responsibility of content safety on their own.

Platforms must take greater responsibility for safeguarding, but in the meantime, there is an opportunity to better support parents with smarter tools. YouTube has implemented several such features already, such as content classification for uploaders through the *made for kids* flag [15, 16, 17]. However, other platforms, such as TikTok, Facebook, and Instagram, have introduced additional transparency tools that help users understand why specific content is shown to them. TikTok's "Why this video" feature explains how recommendations in the "For You" feed are influenced by user interactions, followed accounts, regional popularity, and more [67]. Meta has also published detailed breakdowns of how AI influences content ranking across Facebook and Instagram, emphasizing efforts to make recommendation logic more

transparent [68]. Similar explanations could be implemented for child-facing recommendations to help parents and children understand why a particular video is recommended.

Additionally, parental controls could evolve to reflect not only content types but also developmental alignment, enabling guardians to tailor the recommendation environment to their child's specific age and emotional and cognitive stage, similar to YouTube Kids's existing age group profiles. However, these groupings could be expanded or refined to allow for more granular control or to incorporate feedback from developmental psychology, ensuring that content aligns not just with age but with cognitive, emotional, and thematic suitability.

Ultimately, while technical models like those presented in this study can assist in identifying and promoting more suitable content, a safer viewing experience will require collaboration between families, platform designers, and researchers. Giving parents more visibility and more influence over what their children are exposed to is not just a convenience; it is a critical part of responsible system design [69].

6.3. Limitations

While this study offers valuable insights into the detection and mitigation of inappropriate content in video recommendations for young children, it is important to acknowledge the limitations that may have influenced our findings. These limitations span the dataset, the feature design and modeling process, the classification pipeline, and the evaluation of reranking strategies. Recognizing these constraints allows for a more accurate interpretation of results and paves the way for meaningful future work.

6.3.1. Dataset Limitations

The dataset used in this study provided a diverse and structured set of videos with carefully curated labels across four content categories: suitable, disturbing, restricted, and irrelevant. However, the dataset remains relatively small in size compared to the scale of content on platforms like YouTube. This restricts the representativeness of some content types and limits the robustness of certain statistical findings.

Moreover, the dataset is specifically annotated for toddlers (i.e., children aged 1–5), and does not include age-graded appropriateness labels for other developmental stages. As a result, the analysis and findings are not directly transferable to older children, whose content needs and sensitivities may differ substantially.

Additionally, related video lists were inconsistently populated across videos. While some videos contained full lists of related items, others contained very few or none. This uneven distribution created practical constraints for the reranking strategies, which relied on the ability to adjust rankings based on related video metadata. In cases with limited or empty related video lists, the effectiveness of reranking was inherently capped.

The dataset was also unbalanced across labels, with some categories, particularly restricted videos, being less represented. This imbalance could have contributed to classifier performance variation and challenges in generalization. Moreover, the presence of unknown items (i.e., videos for which no metadata or ground-truth label was available) further complicated evaluation.

The presence of unknown entries (i.e., videos lacking metadata or ground-truth labels) in related video lists presented a unique challenge. Since these videos lacked metadata or labels, they could not be reranked or properly assessed within our evaluation metrics. As a result, unknown items acted as static anchors in ranked lists, sometimes disrupting the relative ordering of known items. While the reranking strategies focused exclusively on known content, the influence of unknowns, particularly when appearing in top-ranked positions, may have obscured the full impact of reranking improvements.

6.3.2. Feature Limitations

This study deliberately focused on metadata-derived features, which offer interpretability and scalability but come with trade-offs in depth and expressiveness.

One of the more experimental features in this study was the topic change feature. Our goal was to explore whether abrupt shifts in a video's textual description might correlate with inappropriate or incoherent content. Sudden changes in tone or theme, such as those that start with child-friendly phrases and abruptly introduce violent or unrelated concepts, can be perceived as topic shifts and may signal

potentially inappropriate material. In this context, a high degree of topical inconsistency may reflect a lack of thematic coherence or an attempt to mislead viewers, both of which could be associated with inappropriate content. However, high-quality topic change detection models for short-form content are rare. As a proof-of-concept, we used an existing model trained on Italian news segments, which introduces both linguistic and domain mismatch. While the model did detect some meaningful variation, specifically in longer descriptions, the signal was generally weak and noisy. We include this feature as an exploratory step, but caution that it should be treated as a preliminary indicator rather than a reliable metric.

Thumbnail analysis was included as a way to extract additional content cues from visual metadata. However, unlike textual fields, thumbnail images are less standardized and more context-dependent. In this study, we applied the same emotion and language models used on titles and descriptions to captions generated from thumbnails. While this provided some surface-level comparability, there is currently little prior research validating this approach for appropriateness detection. Consequently, while thumbnail captions contributed marginally to overall signal strength, their interpretation should be approached with caution, especially given the noisiness of image-to-text model output and the lack of semantic context surrounding text extracted from static thumbnail images.

6.3.3. Classifier Limitations

The classification model used in this study demonstrated strong performance under stratified cross-validation but dropped significantly on the Unseen Classifier Set. This highlights a key limitation in the model's ability to generalize to unfamiliar data distributions. Some of this drop can be attributed to the non-stratified sampling of the test set, but it also reflects the broader challenge of training effective classifiers on small, imbalanced datasets.

Further, the Random Forest model, while interpretable and performant, lacks the expressive capacity of more complex deep learning models. Due to computational, data, and time constraints, we opted for traditional machine learning approaches. While this was a reasonable trade-off, it limited our ability to capture more abstract content patterns that may require more expressive models or could be better learned from multimodal inputs.

6.3.4. Reranking Strategy Limitations

Our reranking strategies demonstrate clear gains in promoting suitable content but are limited in suppressing inappropriate content. A major contributing factor is classifier error propagation: videos misclassified as suitable are sometimes promoted to high ranks, particularly when coupled with high confidence scores.

Moreover, the static nature of unknown items also weakened the ability of the reranking strategies to re-optimize the full ranked list. Because these videos could not be properly reranked, they acted as anchors in the ranking list, limiting the overall flexibility of the system and occasionally resulting in distorted outcomes.

Finally, some reranking strategies, such as those based on *categoryId* or *viewCount* mappings, relied on feature-specific distributions that were not always statistically robust. Limited data volume or skewed category representation occasionally led to distorted ranking behavior. While fusion-based strategies mitigated this to some extent, the overall effectiveness of reranking was ultimately bound by the reliability of the signals and the structure of the underlying data.

6.4. Future Work

This study opens several promising avenues for future research and system development aimed at protecting young children from inappropriate content on video platforms. While our feature-informed classifier and reranking strategies provide a meaningful foundation, both the technical approach and the surrounding ecosystem present opportunities for enhancement.

One clear direction is the integration of multimodal features, especially those derived from visual and audio content. Inappropriate content often manifests through visual cues (e.g., disturbing animation, unexpected violence) or audio tracks (e.g., suggestive language, tonal shifts) that cannot be detected through metadata alone. This manuscript focused exclusively on metadata and textual signals due to their accessibility, interpretability, and scalability. However, prior work has shown that deep learning models trained on raw video frames or audio features can effectively detect harmful, suggestive, or otherwise

inappropriate content in children's media [24, 28, 27]. Integrating these features could significantly improve classification robustness, particularly in identifying disturbing videos that are handcrafted to bypass safeguards.

Real-time or near-real-time recommendation testing is another valuable direction. While this work evaluated reranking strategies on static recommendation lists, future studies could simulate dynamic interaction scenarios, such as child-like navigation through a recommendation graph over several hops, to assess how children's exposure might evolve across multiple viewing steps. For example, a system could mimic a child selecting one of the top-ranked videos, then recursively following recommendations over time, while applying reranking updates after each step. Such longitudinal simulations would offer insights into how well reranking strategies hold up over time and how quickly inappropriate content may re-enter a user's recommendation stream under different conditions.

Improving annotation strategies is also critical. The ground-truth dataset used in this work employed discrete categorical labels, but in practice, content appropriateness is rarely binary. Future work could experiment with multi-label annotations or probabilistic scoring systems to reflect the uncertainty and spectrum of inappropriateness. Further, expanding the label taxonomy to account for age-specific sensitivities (e.g., distinguishing between content suitable for toddlers versus older children) could allow for more granular control in recommendation systems.

Another opportunity lies in refining the signal sources used for classification and reranking. Some features explored in this study, such as topic change or thumbnail captioning, require more sophisticated modeling and better underlying tools. Future iterations might replace or enhance these signals using newer, task-specific models for coherence detection, multimodal emotion recognition, or context-aware thumbnail analysis.

In addition to technical and regulatory pathways, future work should also consider the economic incentives and business trade-offs that platforms like YouTube face. Shifting recommendation priorities from engagement maximization to developmental appropriateness may come at the cost of reduced watch time, lower ad impressions, and, therefore, diminished short-term revenue, particularly for child-focused channels that currently thrive on attention-grabbing but borderline inappropriate content. For platforms whose business model is tightly coupled to user retention and advertising, this creates a structural tension between commercial performance and child safety. However, ignoring this balance may carry long-term risks, including multi-million dollar fines, such as YouTube's $170 million COPPA fine, reputational damage, public backlash, or increased regulatory scrutiny. Future research could explore hybrid models that reconcile these goals, such as monetization structures that reward verified suitability or algorithmic tuning that optimizes for engagement within safety bounds, helping platforms future-proof their systems against both ethical and legal challenges.

Finally, we believe that sustainable solutions to the problem of inappropriate content in children's recommendations will require broader stakeholder collaboration. Platforms, researchers, regulators, and parents each have a role to play in shaping digital environments that prioritize developmental appropriateness over engagement metrics. While this study offers a technical foundation through feature-informed classification and reranking, long-term impact will depend on how these tools are adopted, extended, and supported at the platform level. Tools such as more customizable filtering layers, transparency overlays for recommendation logic, and user-friendly reporting systems could empower both parents and designers to make safer content pathways visible and navigable. At the same time, aligning platform incentives through monetization frameworks that reward suitability rather than attention is essential to disincentivize exploitative content strategies. Future work should continue to explore how these technological, social, and economic levers can be combined to design child-centered systems that safeguard young users by default rather than by exception.

7

All Is Not Yet Found: Conclusion

As platforms like YouTube and YouTube Kids continue to play a central role in young children's daily media use, there is a growing need to understand not only what types of content children are exposed to, but how these videos are filtered, ranked, and surfaced through automated recommendation systems. This research set out to examine the nuanced nature of video appropriateness in YouTube recommendations for young children and to explore how these distinctions could be leveraged to improve content safety.

To address this challenge, we followed a two-part approach in this study. First, we conducted an exploratory feature analysis to identify metadata-derived signals, such as engagement metrics, emotional tone in textual fields, and categorical labels, that characterize different types of video appropriateness for young children. These insights guided the development of a classifier capable of distinguishing between ***suitable***, ***irrelevant***, ***restricted***, and ***disturbing*** videos. Second, we designed and evaluated several reranking strategies, based on classifier predictions, metadata distributions, and hybrid fusion methods, to explore how these distinctions in appropriateness could be used to improve recommendation outcomes. This approach allowed us to examine not only how inappropriate content could be filtered more effectively, but also how age-appropriate content could be promoted more reliably within ranked recommendation lists for children.

Our results demonstrate that metadata-based classification offers a scalable and interpretable option for detecting inappropriate content, but generalization remains difficult. Our classifier's performance dropped when applied to unseen data, reflecting the challenge of modeling appropriateness across diverse video types and metadata profiles. Nonetheless, the classifier's predictions proved useful when integrated into score-based reranking strategies, which consistently improved the positioning of suitable content. HitRate@1 of suitable videos increases from 14% to as high as 62%, substantially raising the likelihood that young children encounter appropriate content first, especially given their tendency to click on one of the top two results [56, 57]. However, these gains came with trade-offs. Unknown items (i.e., videos lacking metadata and ground-truth labels) could not be reranked, limiting overall control and flexibility. Moreover, inappropriate videos that were misclassified with high confidence were sometimes promoted to top positions, with HitRate@1 for inappropriate videos increasing from 9% under YouTube's original ranking to as high as 31%. Given that exposure to inappropriate content can impact young children's development and emotional well-being in various ways [7, 9, 10], this is a concerning outcome. As such, detection and reranking methods like ours represent a first step that warrants further steps for safer recommendation environments for young children.

While this study centers on classification and reranking, our findings point to broader design and policy implications. As long as content visibility is tied to engagement-based signals, such as view count or likes, inappropriate videos may continue to surface. We argue that addressing this requires more than technical fixes; it calls for structural changes in platform incentives and greater support for families. This includes not just better filters, but also more customizable controls tailored to children's developmental needs. Platforms like YouTube must take steps to increase transparency around how recommendation and moderation systems work, helping rebuild trust in child-facing media systems.

Looking ahead, several opportunities remain for advancing child-centered content moderation. Future work could integrate multimodal features, such as visual, audio, and subtitle-based signals, to improve the detection of inappropriate content that cannot be captured by metadata alone. In addition, testing reranking

strategies in longitudinal or real-time settings would provide a clearer view of how recommendation exposure evolves over time. Expanding the label taxonomy to reflect age granularity or degrees of developmental alignment could also support more tailored moderation, while refining monetization models may help incentivize the production and surfacing of high-quality, truly age-appropriate content. These directions offer promising paths for strengthening both the precision and fairness of recommendation systems aimed at young audiences.

This manuscript provides a foundation for such efforts. We demonstrate that metadata-informed models, while imperfect, can contribute meaningfully to content safety. We also show that reranking strategies, even simple ones, can significantly improve the promotion of suitable content, though challenges remain in reliably suppressing inappropriate material. And we argue that a nuanced understanding of inappropriateness, embedded within layered systems and informed by real-world complexity, offers a realistic and impactful way forward. While the challenges of content moderation for young children are far from resolved, this work marks a meaningful step toward safer, more transparent, and better tailored recommender systems for children.

All is not yet found, but the path toward a safer digital landscape for children becomes clearer with every step.

References

[1] Shimrit Ben-Yair. “Introducing the newest member of our family, the YouTube Kids app—Available on Google Play and the App Store”. In: *YouTube Official Blog, February* 23 (2015).

[2] Neal Mohan. *Investing to empower the YouTube experience for the next generation of video*. 2021. URL: https://blog.youtube/inside-youtube/neal-innovation-series/.

[3] Laura Ceci. *Global downloads of YouTube Kids App 2023*. 2024. URL: https://www.statista.com/statistics/1251942/global-youtube-kids-app-downloads/.

[4] Qustodio. 2024. URL: https://www.qustodio.com/en/born-connected-rise-of-the-ai-generation/.

[5] Dahlia Henderson et al. “Youtube for young children: what are infants and toddlers watching on the most popular video-sharing app?” In: *Frontiers in Developmental Psychology* 2 (2024), p. 1335922.

[6] Russell Brandom. *Inside Elsagate, the conspiracy-fueled war on creepy YouTube kids videos*. 2017. URL: https://www.theverge.com/2017/12/8/16751206/elsagate-youtube-kids-creepy-conspiracy-theory.

[7] Jacqui Heinrich. *Deceptive online videos with popular kids characters hide sinister scenes*. 2018. URL: https://www.boston25news.com/news/parents-disturbed-by-videos-featuring-violence-on-kids-app-1/718408215/.

[8] Camille Mori et al. “Exposure to sexual content and problematic sexual behaviors in children and adolescents: A systematic review and meta-analysis”. In: *Child abuse & neglect* 143 (2023), p. 106255.

[9] Sana Ali et al. “Apprehending parental perceptions and responses to accidental online pornography exposure among children in Pakistan: a qualitative investigation”. In: *Human Arenas* (2024), pp. 1–24.

[10] Kristen Harrison et al. “Tales from the screen: Enduring fright reactions to scary media”. In: *Media Psychology* 1.2 (1999), pp. 97–116.

[11] The YouTube Team. *More information, faster removals, more people - an update on what we’re doing to enforce YouTube’s community guidelines*. 2018. URL: https://blog.youtube/news-and-events/more-information-faster-removals-more/.

[12] The YouTube Team. *Introducing our new strikes system*. 2019. URL: https://blog.youtube/news-and-events/introducing-our-new-strikes-system/.

[13] The YouTube Team. *Making our strikes system clear and consistent*. 2019. URL: https://blog.youtube/news-and-events/making-our-strikes-system-clear-and/.

[14] Sarah. *YouTube Kids App is now available on the web at www.youtubekids.com - youtube community*. 2019. URL: https://support.google.com/youtube/thread/12980033?hl=en.

[15] Susan Wojcicki. *An update on kids and data protection on YouTube*. 2019. URL: https://blog.youtube/news-and-events/an-update-on-kids/.

[16] Google. *Determining if your content is ”made for kids”*. 2019. URL: https://support.google.com/youtube/answer/9528076.

[17] The YouTube Team. *Our comment on Coppa*. 2019. URL: https://blog.youtube/news-and-events/our-comment-on-coppa/.

[18] Federal Trade Commission. *Google and YouTube Will Pay Record $170 Million for Alleged Violations of Children’s Privacy Law*. 2019. URL: https://www.ftc.gov/news-events/news/press-releases/2019/09/google-youtube-will-pay-record-170-million-alleged-violations-childrens-privacy-law.

[19] Kristin Cohen. *YouTube channel owners: Is your content directed to children?* 2019. URL: https://www.ftc.gov/business-guidance/blog/2019/11/youtube-channel-owners-your-content-directed-children.

[20] Saeed Ibrahim Alqahtani et al. “Children’s Safety on YouTube: A Systematic Review”. In: *Applied Sciences* 13.6 (2023). DOI: 10.3390/app13064044. URL: https://www.mdpi.com/2076-3417/13/6/4044.

[21] Sajal Aggarwal et al. “Protecting our Children from the Dark Corners of YouTube: A Cutting-Edge Analysis”. In: *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)*. IEEE. 2023, pp. 1–5.

[22] Kanwal Yousaf et al. “Using two-stream EfficientNet-BiLSTM network for multiclass classification of disturbing YouTube videos”. In: *Multimedia Tools and Applications* 83.12 (2024), pp. 36519–36546.

[23] H Faheem Nikhat et al. “Inappropriate YouTube content detection and classification by using proposed novel auto-determined k-means clustering and PDBRNN architecture”. In: *Journal of Intelligent & Fuzzy Systems* Preprint (2024), pp. 1–13.

[24] Kostantinos Papadamou et al. “Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children”. In: *Proceedings of the international AAAI conference on web and social media*. Vol. 14. 2020, pp. 522–533.

[25] Iuliia Miroshnichenko. “FairTube Revisited: The GDPR Violations of Content Creators’ Personal Data Rights by Algorithmic Decision-Making on YouTube”. PhD thesis. Tallinn University of Technology, 2021.

[26] William Hoiles et al. “Engagement and popularity dynamics of youtube videos and sensitivity to meta-data”. In: *IEEE Transactions on Knowledge and Data Engineering* 29.7 (2017), pp. 1426–1437.

[27] Rashid Tahir et al. “Bringing the kid back into YouTube kids: detecting inappropriate content on video streaming platforms”. In: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM ’19. Vancouver, British Columbia, Canada: Association for Computing Machinery, 2020, pp. 464–469. DOI: 10.1145/3341161.3342913. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3341161.3342913.

[28] Myrsini Gkolemi et al. “YouTubers Not madeForKids: Detecting Channels Sharing Inappropriate Videos Targeting Children”. In: *Proceedings of the 14th ACM Web Science Conference 2022*. WebSci ’22. Barcelona, Spain: Association for Computing Machinery, 2022, pp. 370–381. DOI: 10.1145/3501247.3531556. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3501247.3531556.

[29] Le Binh et al. “Samba: Identifying Inappropriate Videos for Young Children on YouTube”. In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. CIKM ’22. Atlanta, GA, USA: Association for Computing Machinery, 2022, pp. 88–97. DOI: 10.1145/3511808.3557442. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3511808.3557442.

[30] Akari Ishikawa et al. “Combating the elsagate phenomenon: Deep learning architectures for disturbing cartoons”. In: *2019 7th International Workshop on Biometrics and Forensics (IWBF)*. IEEE. 2019, pp. 1–6.

[31] Reddit. 2017. URL: https://www.reddit.com/r/ElsaGate/comments/6o6baf/what_is_elsagate/.

[32] Neil Shah. “Flock: Combating astroturfing on livestreaming platforms”. In: *Proceedings of the 26th International Conference on World Wide Web*. 2017, pp. 1083–1091.

[33] Dhruv Kuchhal et al. “A view into YouTube view fraud”. In: *Proceedings of the ACM Web Conference 2022*. 2022, pp. 555–563.

[34] Ofcom. 2024. URL: https://www.ofcom.org.uk/media-use-and-attitudes/media-habits-children/children-and-parents-media-use-and-attitudes-report-2024/.

[35] Shahid Hussain et al. “Role Of Youtube Kids Channel In Socialization Of Children During Covid-19.” In: *Journal of Positive School Psychology* 6.9 (2022).

[36] Marco Della Cava. *Google to revamp its products with 12-and-younger focus*. 2014. URL: https://eu.usatoday.com/story/tech/2014/12/03/google-products-revamped-for-under-13-crowd/19803447/.

[37] Jessica Balanzategui. "'Disturbing'children's YouTube genres and the algorithmic uncanny". In: *new media & society* 25.12 (2023), pp. 3521–3542.

[38] Sultan Alshamrani. "Detecting and Measuring the Exposure of Children and Adolescents to Inappropriate Comments in YouTube". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. CIKM '20. Virtual Event, Ireland: Association for Computing Machinery, 2020, pp. 3213–3216. DOI: 10.1145/3340531.3418511. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3340531.3418511.

[39] Brad J. Bushman et al. "Short-term and Long-term Effects of Violent Media on Aggression in Children and Adults". In: *Archives of Pediatrics & Adolescent Medicine* 160.4 (Apr. 2006), pp. 348–352. DOI: 10.1001/archpedi.160.4.348. eprint: https://jamanetwork.com/journals/jamapediatrics/articlepdf/204790/poa50133_348_352.pdf. URL: https://doi.org/10.1001/archpedi.160.4.348.

[40] Sapna Maheshwari. *YouTube kids, criticized for content, introduces new parental controls*. 2018. URL: https://www.nytimes.com/2018/04/25/business/media/youtube-kids-parental-controls.html.

[41] Michelle M Garrison et al. "Media use and child sleep: the impact of content, timing, and environment". In: *Pediatrics* 128.1 (2011), pp. 29–35.

[42] Albert Bandura et al. "Transmission of aggression through imitation of aggressive models." In: *The Journal of Abnormal and Social Psychology* 63.3 (1961), p. 575.

[43] Jacqueline Allan. *An analysis of Albert Bandura's aggression: A social learning analysis*. Macat Library, 2017.

[44] George Gerbner et al. "Growing up with television: Cultivation processes". In: *Media effects*. Routledge, 2002, pp. 53–78.

[45] Rishabh Kaushal et al. "KidsTube: Detection, characterization and analysis of child unsafe content & promoters on YouTube". In: *2016 14th Annual Conference on Privacy, Security and Trust (PST)*. IEEE. 2016, pp. 157–164.

[46] Christine Pinney et al. "Incorporating Word-level Phonemic Decoding into Readability Assessment". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 8998–9009. URL: https://aclanthology.org/2024.lrec-main.788/.

[47] Christine Pinney et al. "How Readability Cues Affect Children's Navigation of Search Engine Result Pages". In: *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*. IDC '24. Delft, Netherlands: Association for Computing Machinery, 2024, pp. 62–69. DOI: 10.1145/3628516.3655818. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3628516.3655818.

[48] Ashlee Milton et al. ""Don't Judge a Book by its Cover": Exploring Book Traits Children Favor". In: *Proceedings of the 14th ACM Conference on Recommender Systems*. RecSys '20. Virtual Event, Brazil: Association for Computing Machinery, 2020, pp. 669–674. DOI: 10.1145/3383313.3418490. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3383313.3418490.

[49] Haldun Akoglu. "User's guide to correlation coefficients". In: *Turkish journal of emergency medicine* 18.3 (2018), pp. 91–93.

[50] Roshani K Prematunga. "Correlational analysis". In: *Australian Critical Care* 25.3 (2012), pp. 195–199.

[51] Sanjana Reddy et al. "Development of kid-friendly youtube access model using deep learning". In: *Data Science and Security: Proceedings of IDSCS 2020*. Springer. 2021, pp. 243–250.

[52] Leo Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32.

[53] Corinna Cortes et al. “Support-vector networks”. In: *Machine learning* 20 (1995), pp. 273–297.

[54] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[55] Ron Kohavi et al. “A study of cross-validation and bootstrap for accuracy estimation and model selection”. In: *Ijcai*. Vol. 14. 2. Montreal, Canada. 1995, pp. 1137–1145.

[56] Sergio Duarte Torres. “Information Retrieval for Children: Search Behavior and Solutions”. English. The PhD thesis was awarded with a cum laude degree. PhD Thesis - Research UT, graduation UT. Netherlands: University of Twente, Feb. 2014. DOI: 10.3990/1.9789036536189.

[57] Jacek Gwizdka et al. “Analysis of Children’s Queries and Click Behavior on Ranked Results and Their Thought Processes in Google Search”. In: *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. CHIIR ’17. Oslo, Norway: Association for Computing Machinery, 2017, pp. 377–380. DOI: 10.1145/3020165.3022157. URL: https://doi.org/10.1145/3020165.3022157.

[58] Yan-Martin Tamm et al. “Quality metrics in recommender systems: Do we calculate metrics consistently?” In: *Proceedings of the 15th ACM conference on recommender systems*. 2021, pp. 708–713.

[59] Robin Ungruh et al. “Mirror, Mirror: Exploring Stereotype Presence Among Top-N Recommendations That May Reach Children”. In: *ACM Trans. Recomm. Syst.* (Mar. 2025). Just Accepted. DOI: 10.1145/3721987. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3721987.

[60] Eslam Hussein et al. “Measuring misinformation in video search platforms: An audit study on YouTube”. In: *Proceedings of the ACM on human-computer interaction* 4.CSCW1 (2020), pp. 1–27.

[61] Makena Kelly et al. *YouTube’s New Kids’ content system has creators Scrambling*. 2019. URL: https://www.theverge.com/2019/11/13/20963459/youtube-google-coppa-ftc-fine-settlement-youtubers-new-rules.

[62] International Organization for Standardization. *ISO 8601-1:2019*. Tech. rep. International Organization for Standardization, 2019. URL: https://www.iso.org/standard/70907.html.

[63] Mohammad Othman Nassar et al. “fCombMNZ: an improved data fusion algorithm”. In: *2009 International Conference on Information Management and Engineering*. IEEE. 2009, pp. 461–464.

[64] Yizheng Huang et al. “Multiple Linear Combination Approaches for Information Search in Ranking”. In: *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE. 2022, pp. 749–749.

[65] Google. *How to earn money on YouTube*. 2013. URL: https://support.google.com/youtube/answer/72857.

[66] Amjad Hanif. *10 ways to monetize on YouTube*. 2025. URL: https://blog.youtube/creator-and-artist-stories/10-ways-to-monetize-on-youtube/.

[67] TikTok. *Learn why a video is recommended for you*. 2022. URL: https://newsroom.tiktok.com/en-us/learn-why-a-video-is-recommended-for-you.

[68] Nick Clegg. *Bringing even more transparency to how we protect our platform*. 2023. URL: https://about.fb.com/news/2023/06/how-ai-ranks-content-on-facebook-and-instagram/.

[69] Jenny Radesky et al. “From moral panic to systemic change: Making child-centered design the default”. In: *International Journal of Child-Computer Interaction* 31 (2022), p. 100351. DOI: https://doi.org/10.1016/j.ijcci.2021.100351. URL: https://www.sciencedirect.com/science/article/pii/S2212868921000623.

A

Features Overview

Table A.1: Overview of metadata-derived features

Feature	Description
$emotions_{title}$	Emotion scores from video titles using NRCLex
$hard\_words_{title}$	Proportion of complex words in titles using a phonemic decoding model by [46, 47]
$topic\_change_{title}$	Detected topic shifts in titles using the `topicChangeDetector_v1` model
$emotions_{description}$	Emotion scores from video descriptions using NRCLex
$hard\_words_{description}$	Proportion of complex words in descriptions using a phonemic decoding model by [46, 47]
$topic\_change_{description}$	Detected topic shifts in descriptions using the `topicChangeDetector_v1` model
$emotions_{thumbnail}$	Emotion scores from generated thumbnail captions using the `Salesforce/blip-image-captioning-base` model and NRCLex
$hard\_words_{thumbnail}$	Proportion of complex words in thumbnail captions using the `Salesforce/blip-image-captioning-base` model and a phonemic decoding model by [46, 47]
tagScores	TF-IDF weighted tag vector norms using SpaCy's language model ("en_core_web_md")
categoryId	YouTube-assigned video category
licensedContent	Indicates whether video contains licensed content
defaultAudioLanguage	Primary audio language of the video
license	License type associated with the video
viewCount	Number of times the video has been viewed
commentCount	Total comment count on the video
likeCount	Number of likes the video has received
dislikeCount	Number of dislikes the video has received
favoriteCount	Number of times video was marked as favorite

B

Correlation Analysis Features

Table B.1: Correlation Analysis Results

Feature	Correlation Score	p-value
categoryId	0.302939	1.510112e-179
viewCount	0.214811	2.002030e-39
licensedContent	0.200900	8.855065e-32
defaultAudioLanguage	0.180428	1.038449e-20
descriptionEmotions.joy	0.142543	4.712159e-18
dislikeCount	0.137026	8.669567e-17
titleEmotions.disgust	0.136714	1.018876e-16
titleEmotions.negative	0.133481	5.297923e-16
likeCount	0.107725	6.561151e-11
descriptionEmotions.positive	0.107141	8.318086e-11
descriptionEmotions.trust	0.106909	9.133928e-11
descriptionEmotions.anticipation	0.106356	1.141812e-10
titleEmotions.sadness	0.081995	6.888635e-07
titleEmotions.fear	0.070209	2.147535e-05
commentCount	0.062355	1.615241e-04
tagScores	0.061911	1.798579e-04
titleEmotions.positive	0.053000	1.346958e-03
descriptionEmotions.surprise	0.050444	2.280955e-03
titleEmotions.anger	0.048388	3.428417e-03
thumbnailEmotions.disgust	0.047532	4.044518e-03
descriptionEmotions.fear	0.047325	4.208408e-03
descriptionHardWords	0.043382	8.705722e-03
titleEmotions.trust	0.041636	1.181177e-02
thumbnailEmotions.negative	0.040811	1.359348e-02
titleEmotions.joy	0.034370	3.770014e-02
thumbnailEmotions.anger	0.034072	3.939613e-02
descriptionTopicChange.SAMETOPIC	0.031951	5.339433e-02
thumbnailEmotions.sadness	0.030841	6.224117e-02
license	0.027719	4.220226e-01
thumbnailEmotions.fear	0.024847	1.330733e-01
thumbnailEmotions.surprise	0.023916	1.482442e-01
titleTopicChange.SAMETOPIC	0.020135	2.235390e-01

Feature	Correlation Score	p-value
titleTopicChange.TOPICCHANGE	0.016168	3.284228e-01
descriptionEmotions.negative	0.015659	3.438796e-01
titleEmotions.surprise	0.014016	3.968718e-01
titleEmotions.anticipation	0.013808	4.039056e-01
thumbnailEmotions.joy	0.013331	4.203540e-01
thumbnailEmotions.positive	0.013134	4.272621e-01
descriptionEmotions.anger	0.012268	4.583651e-01
descriptionTopicChange.TOPICCHANGE	0.009510	5.653996e-01
descriptionEmotions.sadness	0.008992	5.867599e-01
thumbnailEmotions.trust	0.007340	6.572966e-01
thumbnailEmotions.anticipation	0.006505	6.941884e-01
titleHardWords	0.004829	7.703600e-01
descriptionEmotions.disgust	0.002952	8.583925e-01
thumbnailHardWords	0.002249	8.918579e-01