The Elsagate corpus: Characterising commentary on alarming video content

Panagiotis Soustas and **Matthew Edwards**
University of Bristol
{panagiotis.soustas,matthew.john.edwards}@bristol.ac.uk

Abstract

Identifying disturbing online content being targeted at children is an important content moderation problem. However, previous approaches to this problem have focused on features of the content itself, and neglected potentially helpful insights from the reactions expressed by its online audience. To help remedy this, we present the Elsagate Corpus, a collection of over 22 million comments on more than 18,000 videos that have been associated with disturbing content. We describe the how we collected this corpus and present some insights from our initial explorations, including the surprisingly positive reactions from audiences to this content, challenges in identifying averse comments, and some unusual non-linguistic commenting behaviour of uncertain purpose.

1 Introduction

The topic of Elsagate is one of the most important problems that has emerged recently in online content moderation. The term, which has attracted major media attention (Weston, 2018; Brandom, 2017) and research interest (Balanzategui, 2021; Mai et al., 2022; Tarvin and Stanfill, 2022; Aggarwal and Vishwakarma, 2023; Alqahtani et al., 2023; Choi and Kim, 2024), refers to the widespread distribution of inappropriate and disturbing content aimed at children across multiple online channels, such as video-sharing websites and social networking platforms. These videos frequently incorporate popular children's characters, such as the titular Elsa (from the Disney movie Frozen), but they juxtapose these child-friendly elements with disturbing or harmful themes such as violence, sexual innuendos, and graphic imagery.

YouTube's algorithm often recommends these forms of inappropriate content to children, since at a superficial content level the videos can be similar to otherwise appropriate content (Papadamou et al., 2020). However, the phenomenon is more than a misfiring of content recommendation systems. The makers of this content exploit popular keywords and tags to attract innocent young viewers (Papadamou et al., 2020), thereby potentially causing psychological and emotional harm (Livingstone et al., 2011). There are also more recent references to inappropriate content on YouTube (Tech Transparency Project, 2022; Hern, 2022) which shows that content like this still exists in the platform (Binh et al., 2022).

This paper discusses the creation and initial explorations of a corpus consisting of comments extracted from YouTube videos that have been identified as Elsagate content. While previous work on Elsagate content has focused on its detection as a computer vision problem, YouTube comments provide valuable linguistic insights into forms of user engagement with videos. Our primary interest in this corpus is as a resource that could be used to help automated systems identify future inappropriate content, either on YouTube or in similar online spaces, as we expect the pattern of reactions to Elsagate content to be distinctive when compared to reactions to genuine child-appropriate content. However, this corpus may also provide valuable broader insight into the variable nature of user engagement with disturbing content, and later in this paper we detail several surprising features of our dataset, including unusual non-linguistic commenting behaviour which has not previously been described.

2 Related work

While most research targeting YouTube focuses on either sentiment analysis or hate speech detection, since the rise of the Elsagate phenomenon in 2016, there has been a shift towards detecting disturbing content (Papadamou et al., 2020). Previous attempts at identifying this content have employed image or video data for their analyses. The first

Proceedings of the 1st International Conference on NLP & AI for Cyber Security, pages 147–152
July 29–30, 2024.

attempts at categorisation of videos on YouTube Kids took place before the emergence of the Elsagate phenomenon, using a combination of computer vision with deep learning (Tahir et al., 2012) to categorise videos as benign, explicit or violent.

Ishikawa et al. (2019) were the first to discuss the Elsagate phenomenon as a distinctive online risk, proposing a deep learning detection mechanism derived from the pornography detection literature. Papadamou et al. (2020) presented the first characterisation of disturbing videos targeted at kids by developing a highly accurate deep learning classifier finding that 8.6% of the videos in their dataset were inappropriate but still recommended for toddlers. Yousaf and Nawaz (2022) used a deep learning-based approach to detect inappropriate children's content from YouTube. In later work, the same authors use a BiLSTM network for disturbing video content multiclass classification (Yousaf et al., 2023). Gkolemi et al. (2022) extend the previous video-based approaches to building a detection mechanism for channels creating disturbing content. Most recently, textual content been employed to assist detection mechanisms, with Binh et al. (2022) using subtitle features alongside image data and video metadata to assist in classification. However, no previous approach has considered the reaction expressed by *commenters* as a possible means of detecting or understanding Elsagate material.

3 Corpus description

Our corpus collection was grounded in previous work that had identified specific YouTube channels or videos as disturbing content fitting the description of Elsagate material. Papadamou et al. (2020) provided a list of 33 channels that produce Elsagate content, sourced from a subreddit devoted to tracking this material. After identifying content from the r/ElsaGate using specific keywords they also collected a random sample of the 500 most popular videos uploaded between 18/11/2018 and 2/11/2018 in United States, Great Britain, Russia, India, and Canada.

Binh et al. (2022) separately provided a list of videos from 80 channels that produce age-inappropriate content, as determined by reference to YouTube and FTC guidelines. Their categorization encompassed a wide range of content either visual or linguistic that may be deemed inappropriate, including classic cartoons edited with inappropriate text or visuals, adult gaming content, adult cartoons, toy destruction videos, deceptive channels targeting children and family channels demonstrating child abuse coming from four annotators. As many videos and channels examined in previous research have been removed due to previous reporting, and new content is still being created, we first gathered all still-accessible videos from these sources, and then updated our list using the methodology described by Papadamou et al. (2020), collecting new video IDs reported on the `/r/ElsaGate` subreddit.

In total, our collection covers comments on videos from 53 active channels that have been associated with Elsagate-style content. Out of the 25,861 video IDs identified from these channels, we extracted comment data from 18,324 (71%). The remainder reflects videos identified in previous research that have since been taken down, videos with comment sections disabled, and videos that had no comments. For these 18,324 videos, we used the YouTube API to extract video metadata and all associated comments. To protect user privacy, we anonymised any personally identifiable information. In total, we acquired 22,849,726 comments produced by 7,591,907 unique users.

3.1 Excluded categories

While our comment corpus is large, it contains certain behaviours which require special treatment in processing and analysis. Firstly, our linguistic processing pipeline is currently only capable of dealing with English-language text, and so *non-English* language comments needed to be detected and handled separately. This was accomplished using the `langdetect` Python package. Secondly, we observed a large number of *spam* comments, generated by users who would repeatedly post the same text in an effort to attract attention either to a video or to some other form of online content or product. We identified spamming behaviour by finding exactly duplicated text posted by the same user and we excluded them from our analysis.

Finally, we encountered some unusual comments which did not contain identifiable language. These comments are usually short, and contain a range of unicode symbols usually reserved for niche typographic uses, with no obvious combined meaning. Table 1 provides some example comments of this type selected from our data. While typically such material would be discarded by a natural language

1	Ù?Ø§ Ø§Ù?Ù?Ù? Ù?Ø§ Ø§Ø²Ù?Ø® Ù?Ø°Ø§ Ø§Ù?Ù?Ù?Ø¯ ØØ·Ù? Ø¨Ù?Ù?Ù?
2	Ù?Ø§Ù? ð?¤© ð?¤© ð?¤©
3	ÕµÕ´Õ²Ò½
4	Fwð?¥°ð?¤£ð??
5	Ã°ÂŸÂ~Â‚Ã°ÂŸÂ‘Â!!!...))),

Table 1: Examples of non-linguistic comments on Elsagate videos.

processing pipeline as noise, we highlight its presence within our corpus because the presence of this material has been of interest to Elsagate observers, with some online observers suggesting that the messages are encrypted communications being carried out in public. We do not attempt any cryptanalysis of this material in this paper, but we do filter out slightly less than half a million comments that fit this description. Table 2 provides a full breakdown of the number of comments captured under each excluded category.

Category	Count
Non-Linguistic	434,342
Spam	4,156,675
Non-English	6,461,042

Table 2: Number of comments per excluded category.

3.2 Lexical features of comments

Following all exclusions described in the previous section, a total of 14,777,932 comments from 5,896,553 unique user accounts are included in our main analysis of reactions to Elsagate video content. In what follows, we present an exploratory 'first look' at this content and its features.

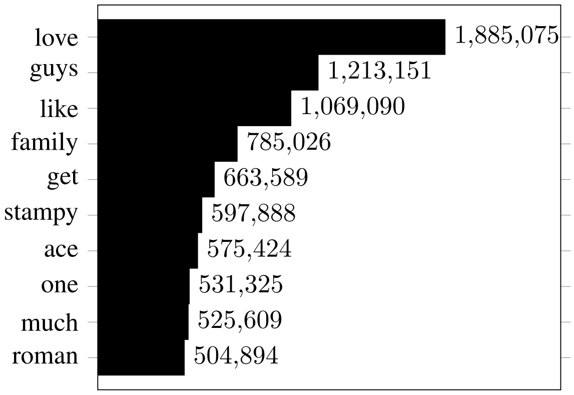

Figure 1: The ten most frequent (non-stopword) English terms within the corpus.

Figure 1 presents the top 10 tokens by frequency within the corpus overall, following stopword removal. An immediate observation is that, despite Elsagate content being characterised by its disturbing or inappropriate nature, positive sentiment is among the most common forms of reaction to these videos, with 'love' being the most frequent term, and 'like' also placing highly. As shown in Table 3, while 'like' is in some cases used in the comparative sense to discuss elements of a video, expressions of positive sentiment and familiarity are commonplace, with commenters showing knowledge of the content creators ('you guys') and their personal background ('your family'). This highlights that, even if Elsagate content may be inappropriate for an age group interacting with it, it does in many cases have an willing audience who enjoy the material and are on good terms with the creators.

collocation			freq.
i	**love**	you	342,989
ˆ	**love**	you	101048
i	**love**	your	96420
you	**guys**	$	159,958
you	**guys**	are	149,765
you	**guys**	so	111,387
i	**like**	the	23,155
would	**like**	to	12,790
looks	**like**	a	12,461
ace	**family**	$	150,003
ace	**family**	i	35,019
your	**family**	$	15,150

Table 3: Most common collocations for common terms (ˆ: start of comment; $: end of comment).

Other common terms visible in Figure 1 relate to particular content or content creators with highly engaged audiences. The presence of these high-volume channels within the corpus highlights an analytic challenge: while certain videos from these creators have been flagged by observers as inappropriate or disturbing content, these labels can be contested, and may not apply to all content from these creators.

Despite the active community focused on Elsagate video identification on YouTube, and our corpus being drawn in large part from materials identified in this way, reference to the phenomenon in these terms was very rare within the comment corpus, with just 60 comments mentioning 'Elsagate' in any form. These occurrences were almost universally warnings or disavowals of content (e.g., *"Known elsagate channel, DO NOT WATCH!"*). many of these commenters were not the natural audience for the video, and appear to have arrived at the content only after having seen it reported in a venue such as the `/r/Elsagate` subreddit. However, as shown in Table 4, comments expressing discomfort in other forms do also appear within the corpus with some regularity, though care must be taken to distinguish tokens from other uses (e.g., the name of 'Weird Al', a popular parodist, appears as a top collocation for 'weird').

	collocation		freq.
so	**messed**	up	1,018
is	**messed**	up	842
really	**messed**	up	340
this	**shit**	is	1,668
the	**shit**	out	1,501
this	**shit**	$	1,460
so	**weird**	$	951
^	**weird**	al	679
is	**weird**	$	638

Table 4: Most common collocations for terms used to express negative reactions (^ : start of comment; $: end of comment).

3.3 Sentiment analysis

Our analysis of sentiment-labelled comments reveals a diverse range of responses from viewers. Utilising the `textBlob` library, we assigned a sentiment tag (*Positive*, *Neutral*, *Negative*) to each comment. The presence of negative comments might be attributed to potentially inappropriate video material, indicating a segment of the audience finds certain content troubling. However, the majority of comments express neutral sentiment, this category accounting for 51.67% of all comments. This suggests a lack of strong emotional polarity among viewers. Furthermore, the widespread nature of positive comments, constituing 39.51% of the total, indicates a largely favourable audience reaction, correlating with the findings from the collocation analysis. Negative comments, comprising only 10.23%, suggests a smaller but still potentially significant portion of the audience expressing dissatisfaction or concern

3.4 Grievance dictionary analysis

To analyse the presence of disturbing content and reactions within our dataset, we employed a dictionary matching technique using the Grievance dictionary (van der Vegt et al., 2021). This resource offers a structured framework for understanding nuances in language. We systematically parsed comments, matching words to predefined categories and scores based on human annotation. The annotation process involved assessing each word on a scale from 0 to 10 denoting how well that word fits in a specific category (van der Vegt et al., 2021).

Category	Count	Score
relationship	11,176,103	4.593
surveillance	4,452,534	5.726
desperation	4,233,406	4.732
loneliness	2,973,491	6.048
murder	2,530,499	5.656
suicide	2,363,681	5.672
violence	1,796,599	6.164
hate	1,437,010	5.949

Table 5: Aggregated grievance dictionary category counts, with mean weighted score.

The results in Table 5 highlight the presence of concerning themes such as hate, violence, suicide and murder within the corpus, raising concern about the nature of content consumption and interaction within online communities.

4 Conclusion

Our large dataset of comments on videos associated with disturbing content contains a variety of behaviours, with a range including highly positive audience engagement, spam, expressions of discomfort with content, and non-linguistic comments that serve no immediately evident purpose. Our analysis to date covers only an initial exploration of this corpus, and we anticipate that it may prove useful to understanding and preventing the spread of disturbing content, both alone and in conjunction with other resources. Of particular interest is the challenge posed by distinguishing content that is directed at children. It is crucial to assess the engagement of various groups including children, adults and threat actors in the comment sections of these videos. Elsagate observers worry about many risks posed by this content, including psychological harm to young children. This language resource sets a foundation for further linguistic studies of reactions to Elsagate content, and provides a first step towards developing language-related technologies that ensure a safer digital space.

Availability

The dataset will be made available for research purposes. Researchers interested in harnessing this linguistic resource for their investigation will be able to access the dataset in Soustas (2024) .

Limitations

While our dataset and analysis contributes some valuable insights into audience reactions on inappropriate video content, it is crucial to acknowledge several limitations. The dataset was collected from a specific online community platform, drawing upon other studies of the same phenomenon. There is an inherent subjectivity involved in determining which content is 'inappropriate', and we did not evaluate the standards of our source community for consistency. Additionally, comments on online platforms are often short and fragmented, making them challenging to analyse comprehensively. This limitation may constrain the depth of insights gleaned from the dataset, as context within comments may be overlooked. The dataset was collected during a specific timeframe, and online discourse surrounding alarming video content may evolve over time. It is worth noting that a significant percentage of the videos of our initial Video ID list had their comment sections closed or were taken down. This aspect adds another layer of complexity to the analysis, as valuable information that could have been derived from these comments is now unavailable. This limitation underscores the dynamic nature of online content and the challenges associated with capturing and analysing user reactions over time. Furthermore, future changes in platform policies could affect the representativeness of this corpus.

Ethics Statement

The data collected for the Elsagate corpus has been obtained following strict ethical guidelines and permission for both data collection and subsequent analysis was obtained from the relevant institutional review board. All data is anonymised and depersonalised to ensure that no personally identifiable information is contained in the dataset. All methodologies, findings and analyses presented in this paper are reported accurately to the best of the authors' knowledge.

Acknowledgements

This research is made possible through the support of the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/S022465/1 for the University of Bristol, EPSRC Centre for Doctoral Training in Trust, Identity, Privacy and Security in Large-scale Infrastructures (TIPS-at-Scale).

References

Sajal Aggarwal and Dinesh Kumar Vishwakarma. 2023. Protecting our Children from the Dark Corners of YouTube: A Cutting-Edge Analysis. In *2023 4th IEEE Global Conference for Advancement in Technology (GCAT)*, pages 1–5. IEEE.

Saeed Ibrahim Alqahtani, Wael MS Yafooz, Abdullah Alsaeedi, Liyakathunisa Syed, and Reyadh Alluhaibi. 2023. Children's safety on YouTube: A systematic review. *Applied Sciences*, 13(6):4044.

Jessica Balanzategui. 2021. Disturbing children's YouTube genres and the algorithmic uncanny. *New Media and Society*, pages 1–22.

Le Binh, Rajat Tandon, Chingis Oinar, Jeffrey Liu, Uma Durairaj, Jiani Guo, Spencer Zahabizadeh, Sanjana Ilango, Jeremy Tang, Fred Morstatter, Simon Woo, and Jelena Mirkovic. 2022. Samba: Identifying inappropriate videos for young children on YouTube. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 88–97. ACM.

Russell Brandom. 2017. Inside elsagate, the conspiracy-fueled war on creepy YouTube kids videos.

Yun Jung Choi and Changsook Kim. 2024. A content analysis of cognitive, emotional, and social development in popular kid's YouTube. *International Journal of Behavioral Development*, page 01650254241239964.

Myrsini Gkolemi, Panagiotis Papadopoulos, Evangelos Markatos, and Nicolas Kourtellis. 2022. YouTubers not madeForKids: Detecting channels sharing inappropriate videos targeting children. In *14th ACM Web Science Conference 2022*, pages 370–381. ACM.

Alex Hern. 2022. YouTube Kids shows videos promoting drug culture and firearms to toddlers.

Akari Ishikawa, Edson Bollis, and Sandra Avila. 2019. Combating the elsagate phenomenon: Deep learning architectures for disturbing cartoons. In *2019 7th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE.

Sonia Livingstone, Leslie Haddon, Anke Görzig, and Kjartan Ólafsson. 2011. Risks and safety on the internet: the perspective of european children: full findings and policy implications from the EU kids online survey of 9-16 year olds and their parents in 25 countries.

Alexandra Mai, Leonard Guelmino, Katharina Pfeffer, Edgar Weippl, and Katharina Krombholz. 2022. Mental models of the internet and its online risks: Children and their parent (s). In *International Conference on Human-Computer Interaction*, pages 42–61. Springer.

Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. 2020. Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 522–533.

Panagiotis Soustas. 2024. Elsagate corpus. Mendeley Data, V1.

Rashid Tahir, Faizan Ahmed, Hammas Saeed, Shiza Ali, and Christo Zaffar, Fareed amd Wilson. 2012. Bringing the kid back into YouTube kids: Detecting inappropriate content on video streaming platforms. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE.

Emily Tarvin and Mel Stanfill. 2022. "YouTube's predator problem": Platform moderation as governance-washing, and user resistance. *Convergence*, 28(3):822–837.

Tech Transparency Project. 2022. Guns, Drugs, and Skin Bleaching: YouTube Kids Still Poses Risks to Children.

Isabelle van der Vegt, Maximilian Mozes, Bennett Kleinberg, and Paul Gill. 2021. The Grievance Dictionary: Understanding threatening language use. *Behavior Research Methods*, pages 1–15.

Phoebe Weston. 2018. YouTube kids app is STILL showing disturbing videos | daily mail online.

Kanwal Yousaf and Tabassam Nawaz. 2022. A deep learning-based approach for inappropriate content detection and classification of YouTube videos. 10:16283–16298.

Kanwal Yousaf, Tabassam Nawaz, and Adnan Habib. 2023. Using two-stream EfficientNet-BiLSTM network for multiclass classification of disturbing YouTube videos.