It’s not Sexually Suggestive; It’s Educative | Separating Sex Education from Suggestive Content on TikTok videos

Enfa George **Mihai Surdeanu**
University Of Arizona
{enfageorge,msurdeanu}@arizona.edu

Abstract

We introduce SexTok, a multi-modal dataset composed of TikTok videos labeled as sexually suggestive (from the annotator’s point of view), sex-educational content, or neither. Such a dataset is necessary to address the challenge of distinguishing between sexually suggestive content and virtual sex education videos on TikTok. Children’s exposure to sexually suggestive videos has been shown to have adversarial effects on their development (Collins et al., 2017). Meanwhile, virtual sex education, especially on subjects that are more relevant to the LGBTQIA+ community, is very valuable (Mitchell et al., 2014). The platform’s current system removes/punishes some of both types of videos, even though they serve different purposes. Our dataset contains video URLs, and it is also audio transcribed. To validate its importance, we explore two transformer-based models for classifying the videos. Our preliminary results suggest that the task of distinguishing between these types of videos is learnable but challenging. These experiments suggest that this dataset is meaningful and invites further study on the subject.

1 Introduction

In short-form videos such as in TikTok, accurately identifying sexually suggestive and sex education content amidst a sea of diverse video types poses a significant challenge. In this paper, we delve into this problem, focusing specifically on TikTok, the most downloaded app in 2022, which has a sub stantial user base of early adolescents and young individuals (10-19: 32.5%, 20-29: 29.5%) [1]

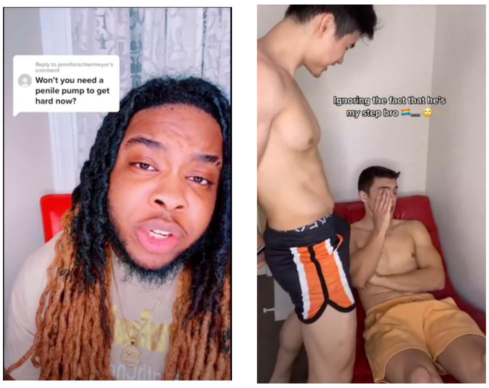

Reply to jenniferschiermeyer's comment
Won’t you need a penile pump to get hard now?
Ignoring the fact that he's my step bro 🏳️‍🌈.... 🙄

Figure 1: Two screenshots from videos in the dataset. On the left, Nyko (@kingnyko2022) addresses a question about his gender transition. The right is from a sexually suggestive video.

(1) **Educative** *(Description)* **:** Video featuring a man discussing a topic while a prominent illustration of a p*n*s with pearly penile papules serves as the background.
(2) **Suggestive** *(Description)* **:** Video shows a man holding a pumpkin over his torso while a woman enthusiastically moves her hand inside, exclaiming, "There is so much in there."
(3) **Educative** *(Transcript)* **:** The average banana in the United States is about 5.5 inches long.That’s the perfect size for baking banana bread most of the time because ...
(4) **Suggestive** *(Transcript)* **:** You are such a good boy. Daddy’s so proud of you.

Table 1: Examples from the dataset, the first two are descriptions, and the latter are video transcripts.

The distinction between suggestive videos and virtual sex education holds crucial significance on multiple fronts. Adolescent sex education in the United States is delivered in a fragmented and often inadequate system, which has long been the subject of intense criticism and is vulnerable to political influence (Fowler et al., 2021).In this context, TikTok presents a novel and promising avenue to conveying comprehensive and accessible sexual health information to adolescents, offering a convenient, private, and inclusive space for learning and discussion (Fowler et al., 2022). At the same time, children’s exposure to sexual media content has been found to influence attitudes and contribute to the formation of adversarial sexual beliefs (Collins et al., 2017).

[1] https://wallaroomedia.com/blog/social-media/tiktok-statistics/

Findings of the Association for Computational Linguistics: ACL 2023, pages 5904–5915
July 9-14, 2023 ©2023 Association for Computational Linguistics

Unfortunately, efforts to moderate explicit content had unintended consequences, as studies have demonstrated the misidentification of non-explicit content due to flawed algorithms and filtering tech niques Peters, 2020. In addition to the above issue, video/video creators (referred to as creators from now on) may also be susceptible to mass reporting. Creators from marginalized communities, partic ularly those within the LGBTQIA+ community, face heightened risks of having their educational content wrongfully flagged or removed [2].

The classification of sexually suggestive and sex education videos presents a complex task, as demonstrated by the examples shown in Table 1. In example 1, we see that p*n*s illustration is not suggestive, while the video with a man holding a pumpkin in example 2 is suggestive. When we look at the transcripts, we see that in example 3, the creator is talking about myths around p*n*s sizes for pleasurable sex, and in example 4, the audio is suggestive. Considering these complexities, accurately categorizing sexually suggestive and sex education videos necessitates a nuanced understanding of contextual cues, subjectivity, evolving language, and robust algorithmic solutions.

The contributions of the paper are as follows:

1. **Introduction of SexTok:** A collection of 1000 TikTok videos labeled as Sexually Suggestive, Sex Education, or Others, along with perceived gender expression and transcription.

2. **Baselines Evaluation:** We evaluate two transformer-based classifiers as baselines for the task of classifying these videos. Our results indicate that accurately distinguishing between these video types is a learnable yet challenging task.

Trigger Warning: Sexual Content and Explicit Language

Please be advised that this research paper and its associated content discuss and analyze sexually suggestive and sex education videos. The examples and discussions within this paper may contain explicit or implicit references to sexual acts, body parts, and related topics. The language used may sometimes be explicit. This material is intended for academic and research purposes and is presented to address challenges in content identification and classification.

2 Related Work

Automatic detection of sexually explicit videos is an area of active study. In a recent survey, Cifuentes et al., 2022 classified the methods into four broad separate strategies. Nudity detection, Analysis of image descriptors (such as Bag of Visual Words), Motion analysis, and other deep learning techniques.

Most works around nudity detection are focused on skin-colored region segmentation to identify nudity. This methodology has been extensively explored in the image domain (Fleck et al., 1996), (Wang et al., 2005) (Platzer et al., 2014), (Garcia et al., 2018),(Lee et al., 2006). (Ganguly et al., 2017)'s work, apart from focusing on the percentage of skin exposure, also gave attention to the body posture of the human in the image and the person's gestures and facial expressions. An alternative strategy is the Bag of Visual Words model, in which the idea is to minimize the existing semantic gap between the low-level visual features and the high-level concepts about pornography. (Deselaers et al., 2008), (Lopes et al., 2009), (Ulges and Stahl, 2011), (Zhang et al., 2013). Approaches based on motion analysis, apart from other features, also capture motion, such as using the periodicity in motion, such as in (Rea et al., 2006). (Zuo et al., 2008) uses a Gaussian mixture model (GMM) to recognize porno-sounds, a contour-based image recognition algorithm to detect pornographic imagery, and are combined for the final decision.

Yet still, sexual activity where the human is mostly clothed or has minimal movement is still challenging. Peters, 2020 studied issues surrounding publicly deployed moderation techniques and called for reconsidering how platforms approach this area, especially due to it's high false positive rates and/or low precision rates for certain types of actions.

3 SexTok Dataset

This section presents the SexTok dataset [3], a collection of 1000 TikTok video links accompanied by three key features: Class Label, Gender Expression, and Audio Transcriptions.

[2] https://mashable.com/article/tiktok-sex-education-content-removal

[3] Data and the experiment codebase will be shared at github.com/enfageorge/SexTok. Videos are shared as links to avoid any potential licensing issues.

3.1 Terminology and Definitions

3.1.1 Class Label

The first feature, Class Label, is a categorical variable with three possible values: *Sexually Suggestive, Sex Education*, and *Others*:

Sexually Suggestive: This category encompasses videos that purposefully intend to elicit a sexual response from viewers. Determining the presence of sexually suggestive content is subjective.

Sex Education: This category encompasses videos aimed at enhancing viewers' knowledge, skills, and attitudes concerning sexual and reproductive health. It covers various topics, including but not limited to sexual orientation, gender, and gender-affirming care.

Others: This category encompasses videos that do not fall within the aforementioned sexually suggestive or sex education categories.

3.1.2 Gender Expression

Gender expression is a form of self-expression that refers to how people may express their gender identity (Summers, 2016). In this paper, we focus solely on the physical visual cues associated with gender expression. We provide five gender expression labels in the dataset: *Feminine, Masculine, Non-conforming, Diverse, and None*.

Feminine and Masculine represent predominantly feminine or masculine expressions, while Non-conforming refers to expressions that deviate from traditional norms. Diverse applies to videos with varying gender expressions among multiple individuals. The None label is for videos without people or only limited visual cues like hands.

The information for the vast majority is not self-reported. When available through the video itself, profile descriptions, or hashtags, we incorporate that information. Otherwise, the annotation is based on the perception of the annotator. *This feature is provided only to serve the purpose of evaluating bias in models built on the dataset.*

3.2 Dataset Construction

Data Collection

The data collection process involved the primary annotator creating a new TikTok account and interacting with the platform in various ways to collect the video links. They carefully watched and hand-selected videos. Two important considerations were taken into account during the dataset construction process: (a) Limit a maximum of five videos per creator in the dataset. (b) Creators appearing in one split of the dataset (train, validation, or test) were excluded from all other splits to ensure independence and prevent data leakage. Detailed information regarding the specific methods used, as well as limitations and ethical considerations, can be found in Appendix A.

Label	Train	Val	Test	Total
Sugg	140	20	40	200
Educative	140	20	40	200
Others	420	60	120	600
Total	700	100	200	1000

Table 2: Video Distribution by Dataset Split and Class Label. Sugg: Suggestive, Edu: Educative.The dataset consists mostly of general videos that do not fall into the categories of sexually suggestive or educative. This reflects a more realistic representation of Tiktok's environment.

Annotator Agreement

A 10% sample of the dataset was independently annotated by a second author to ensure reliability. Cohen's Kappa scores (Cohen, 1960) were used to assess annotator agreement. For Gender Expression, the Kappa score was 0.89, indicating substantial agreement. For Class Label, the Kappa score was 0.93, indicating high agreement. These scores validate the consistency and quality of the dataset's annotations.

Data Processing: Video download and Audio transcription

The videos were downloaded without the TikTok watermark using a TikTok downloader.[4]. The watermark was removed to reduce unnecessary noise in the data.

A smaller sample of videos was first transcribed using OpenAI's whisper (medium) (Radford et al., 2022) and was manually checked for accuracy. The transcriptions were mostly perfect, with a word error rate of 1.79%. After this, all the videos were automatically transcribed using Open AI's Whisper (medium).

3.3 Dataset Properties

In this section, we provide some general statistics about the SexTok dataset. The dataset comprises 1000 TikTok video links with three features: Class

[4] https://github.com/anga83/tiktok-downloader

Label	Fem	Masc	NC	D	None
Sugg	115	84	0	1	0
Edu	85	84	6	8	17
Others	164	170	12	113	141
Total	364	338	18	122	158

Table 3: Video Distribution by Class Label and Gender Expression. Fem: Feminine, Masc: Masculine, NC: Non-conforming, D: Diverse. Sugg: Suggestive, Edu: Educative. The dataset is predominantly feminine in the suggestive category, while in the educative and others categories, both feminine and masculine gender expressions are relatively balanced and dominant.

Label, Gender Expression, and Audio Transcriptions. A breakdown by label and dataset split is given in Table 1. A separate breakdown by Gender Expression and dataset split is given in Table 2.

When the audio was transcribed, a percentage of videos were found not to have any text in the audio transcription, specifically → Suggestive - 15.85%, Educative - 3.97%, Others - 8.4%.

We also observe that suggestive videos tend to be shorter (median duration: 7.86 secs), and have shorter audio transcriptions (median number of words: 14 words), compared to educative videos that are longer (median duration: 50.80 secs) and have longer audio transcriptions (median number of words: 171.5 words). Detailed dataset video length and transcription length are given in Appendix A.)

4 Experimental Setups

In this section, we evaluate the performance of pretrained transformer-based models on the SexTok dataset to assess its significance. The experiments are divided into two subsections: text classification using video transcripts and video classification.

For both transformer-based setups, we utilized models downloaded from Hugging Face Transformers (Wolf et al., 2020), initializing them with three random numbers. Details on hyperparameters are in Appendix C. The reported results are the average of three runs. To assess the performance, we employed four sets of metrics: (1) accuracy, (2) micro precision, recall, and F1 (excluding Others as a negative class from the scores), (3) macro precision, recall, and F1, and (4) overall F1 for each class.

Text Classification using Video Transcript

We fine-tuned `bert-base-multilingual-cased` (Devlin et al., 2018) to perform text classification on the video transcripts. Since we observed that a small percentage of videos do not yield any text in their transcription, we experimented with two setups. One with all video transcriptions and the other with non-empty transcriptions.

Video Classification

We fine-tuned `MCG-NJU/videomae-base`, a VideoMAE base model (Tong et al., 2022) for video classification. The image clips were randomly sampled and preprocessed to align with the default configurations of the model.

5 Results and Error Analysis

The average performance and standard deviation of the models are presented in Tables 4 and 5. Based on these results, we draw the following observations:

- The most accurate model is the text classifier that evaluated videos with a transcription (75%). It demonstrates relatively better performance in identifying educative content but often struggles to differentiate between suggestive content and others, and vice versa. However, it should be noted that this implementation is not realistic in a real-world scenario, as TikTok videos can vary in terms of sound presence and spoken language.
- Both text-based classifiers exhibit higher F1 scores than the video classifier for the Educative and Others classes. But their performance in detecting suggestive content is is comparatively lower than that of the video classifier.
- Notably, neither of the text-based classifiers misclassifies suggestive content as educative, or vice versa, as evident from the confusion matrices in Appendix C.
- The video classifier achieves the highest F1 score for the Suggestive class. However, it frequently confuses Educative and Other videos with each other.

To further understand the hard examples for the model, we manually categorized the errors in both text and video classification experiment setups.

We analysed 54 errors in text classification model. If more than one option was applicable, the video was counted in both: (a) *Audio unrelated to class label (50.00%)*: The audio in these videos

Group	Acc	Micro			Macro		
		P	R	F1	P	R	F1
Majority	0.60	0.00	0.00	0.00	0.20	0.33	0.25
All Text	0.68 ± 0.06	0.76 ± 0.06	0.50 ± 0.06	0.60 ± 0.04	0.71 ± 0.06	0.63 ± 0.03	0.64 ± 0.04
Non-empty Text	0.75 ± 0.02	0.78 ± 0.07	0.54 ± 0.02	0.64 ± 0.02	0.74 ± 0.04	0.65 ± 0.01	0.68 ± 0.00
Video	0.70 ± 0.04	0.61 ± 0.11	0.51 ± 0.07	0.55 ± 0.05	0.68 ± 0.06	0.57± 0.07	0.61 ± 0.01

Table 4: We present the average and standard deviation of results from three different runs of our experiments. We use accuracy, micro-precision, recall, and F1 (with "Others" as a negative class, not included in the scores) and macro-precision, recall, and F1 as metrics. Text-based classification, when transcript is present, has higher overall performance.

Group	Suggestive	Educative	Others
Majority	0.00	0.00	0.60
All Text	0.30 ± 0.14	0.83 ± 0.01	0.80 ± 0.02
Non-empty Text	0.38 ± 0.03	**0.84** ± 0.01	**0.81** ± 0.02
Video	**0.55** ± 0.02	0.63 ± 0.13	0.72 ± 0.15

Table 5: We present the overall F1 of each class label with the average and standard deviation of three random runs. Text-based classification gives a higher F1 for educative content when transcription is present, but suggestive content is detected best in videos where educative content is misclassified higher.

consisted of popular songs or speeches that did not contain any words typically associated with the class label. (b) *Context clues and Euphemism (25.07%)* : These videos relied on context clues or employed euphemistic language (9.26%) or required audio analysis considering the tone and intonation to predict the class label (14.81%). (c) *No or partial transcription (14.81%)*: Approximately 9.26% of the videos had no audio that could be transcribed, while 5.56% had only partial transcriptions available. We analyzed 52 errors in video classification. All educative videos that were classified as others, and vice versa, had the same format that both classes do, i.e., a person looking at the camera speaking. Of the 11 suggestive videos that were not classified correctly, in 63% of videos, some or all of the video frames had fully or mostly clothed people featured in the video. A detailed analysis using Transformers-interpret C (Pierse, 2021) also shows that the text classification shows some signs of overfitting to text.

6 Discussion

The results highlight the complexity of accurately identifying sexually suggestive and educative videos on platforms like TikTok. While the results indicate that text analysis can contribute to detecting educative videos, music clips unrelated to the video topic are commonly used, making reliance on transcription alone insufficient. While existing work in pornographic content detection primarily focuses on visual analysis, our results indicate the need for a multi-modal approach since detecting sexual content requires a more comprehensive understanding encompassing multiple senses, including audio, speech, and text.

Addressing these challenges is crucial for developing effective content moderation systems, ensuring appropriate access to sex education, and creating a safer and more inclusive online environment. It is also crucial to be mindful of potential gender expression bias commonly found in visual datasets (Meister et al., 2022). Moreover, for tasks like this, developing scalable solutions suitable for large-scale systems with millions of users is crucial for effective implementation. Further exploration and investigation of these aspects are left for future research and development.

7 Conclusion

This paper introduces a novel task of identifying sexually suggestive and sex-educative videos and presents SexTok, a multi-modal dataset for this purpose. The dataset includes video links labeled for sexual suggestiveness, sex-educational content, and an other category, along with gender expression and audio transcription. The results highlight the challenging and multi-modal nature of the task and suggest that while the dataset is meaningful and the task is learnable, it remains a challenging problem that deserves future research. This work contributes to promoting online safety and a balanced digital environment.

8 Acknowledgement

This work was partially funded by the LGBTQ+ Grad Student Research Funds by The Institute for LGBTQ Studies at the University Of Arizona. We deeply appreciate the invaluable contributions of Shreya Nupur Shakya throughout this work.

Limitations

We address the limitations of the SexTok dataset and the accompanying experiments here.

SexTok Dataset

- The TikTok account was created and used from a specific geographic location (which will be disclosed in the final version if accepted). This is important to note since the content recommendation of TikTok is influenced by geographic location,[5] among other things; hence a geographic bias may be expected, i.e., certain demographics may be more represented than others, especially in terms of languages used, race, ethnicity, etc.

- The data gathered only represents a small sample of the content available on TikTok and may not represent the entire population of TikTok users or videos.

- Sexual suggestiveness is treated as a discrete class label in the project, whereas in the real world, it has two important properties. 1) The perception of what is sexually suggestive may vary depending on the individual's sexual orientation, worldview, culture, location, and experiences and is highly subjective. 2) Some are more suggestive than others, and we do not account for the variation in the strength of suggestiveness here.

- The dataset is a small snapshot of the TikTok videos from October 2022 to January 2023. Patterns, slang, and other cues may change over time.

- Gender expression has many variations but is referred to as discrete labels here, but in real life, it is not. Additionally, this is as perceived by one annotator and, for the majority, not self-reported by the person in the video. Additional expert annotators may be needed to strengthen the confidence in the label.

- Despite best efforts, it may be possible that the same creator appears more than five times. This is because creators often create multiple accounts to serve as a backup in case TikTok takes down the original account. This is observed to be increasingly common in the sexually-suggestive and sex-ed domains. We show an example in Figure 2

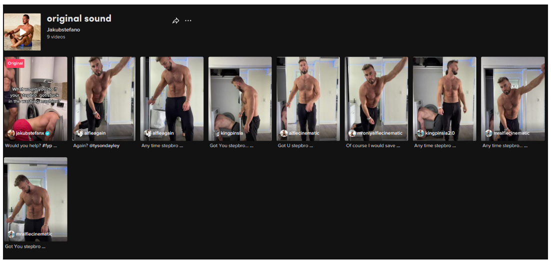


Figure 2: This is a partial screenshot from an audio profile page on Tiktok. Each rectangle is a cover image of a video that uses the same audio. The text on the bottom left of each video is the username of the creator of that video. We can see that the same person has multiple accounts posting the same video.

Other details :

The audio content of the TikTok videos comprises various elements, including background music, spoken dialogue (not necessarily from the video creator), or a combination of both. Notably, TikTok provides voice effects that enable users to modify their voices using predefined options.

Experiments

- The audio transcription of the videos was created automatically using Open AI's Whisper-medium (Radford et al., 2022). Hence this is subject to errors, which may impact the performance of the models.

- For training the models, GPU computing power was used.

Ethics Statement

We address the ethical considerations and consequences of the SexTok dataset and the accompanying experiments here.

- The study's focus is on the technical aspects of the problem. It does not address the broader societal and ethical implications of censorship and of regulating sexually suggestive content on social media platforms. The work only aims to detect sexually suggestive content and sex education content against other video topics but makes no stand on censorship or content regulation of sexually suggestive videos.

[5] https://support.tiktok.com/en/account-and-privacy/account-privacy-settings/location-services-on-tiktok

- Sexual suggestiveness, as well as perceived gender expression, is a subjective matter and is hence susceptible to annotators' bias.

- Gender expression, specifically visual cues only, was annotated and offered only to evaluate bias based on visual cues since such biases are known to exist within large-scale visual datasets (Meister et al., 2022). The authors do not condone the practice of assigning gender identity based on a person's external appearance since gender is an internal sense of identity (Association, 2015). This dataset is not intended to be used for any such practices.

- Due to the nature of the problem, and potential licensing issues, the publicly-collected data is not anonymized.

References

American Psychological Association. 2015. Guidelines for psychological practice with transgender and gender nonconforming people. *American psychologist*, 70(9):832–864.

Jenny Cifuentes, Ana Lucila Sandoval Orozco, and Luis Javier García Villalba. 2022. A survey of artificial intelligence strategies for automatic detection of sexually explicit videos. *Multimedia Tools and Applications*, 81(3):3205–3222.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Rebecca L Collins, Victor C Strasburger, Jane D Brown, Edward Donnerstein, Amanda Lenhart, and L Monique Ward. 2017. Sexual media and childhood well-being and health. *Pediatrics*, 140(Supplement_2):S162–S166.

Thomas Deselaers, Lexi Pimenidis, and Hermann Ney. 2008. Bag-of-visual-words models for adult image classification and filtering. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Margaret M Fleck, David A Forsyth, and Chris Bregler. 1996. Finding naked people. In *Computer Vision—ECCV'96: 4th European Conference on Computer Vision Cambridge, UK, April 15–18, 1996 Proceedings Volume II 4*, pages 593–602. Springer.

Leah R Fowler, Lauren Schoen, and Stephanie R Morain. 2021. Let's tok about sex. *Journal of Adolescent Health*, 69(5):687–688.

Leah R Fowler, Lauren Schön, Hadley Stevens Smith, and Stephanie R Morain. 2022. Sex education on tiktok: a content analysis of themes. *Health promotion practice*, 23(5):739–742.

Debashis Ganguly, Mohammad H Mofrad, and Adriana Kovashka. 2017. Detecting sexually provocative images. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 660–668. IEEE.

Manuel B Garcia, Teodoro F Revano, Beau Gray M Habal, Jennifer O Contreras, and John Benedic R Enriquez. 2018. A pornographic image and video filtering application using optimized nudity recognition and detection algorithm. In *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*, pages 1–5. IEEE.

Hogyun Lee, Seungmin Lee, and Taekyong Nam. 2006. Implementation of high performance objectionable video classification system. In *2006 8th International Conference Advanced Communication Technology*, volume 2, pages 4–pp. IEEE.

Ana PB Lopes, Sandra EF de Avila, Anderson NA Peixoto, Rodrigo S Oliveira, and Arnaldo de A Araújo. 2009. A bag-of-features approach based on hue-sift descriptor for nude detection. In *2009 17th European Signal Processing Conference*, pages 1552–1556. IEEE.

Nicole Meister, Dora Zhao, Angelina Wang, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. 2022. Gender artifacts in visual datasets. *arXiv preprint arXiv:2206.09191*.

Kimberly J Mitchell, Michele L Ybarra, Josephine D Korchmaros, and Joseph G Kosciw. 2014. Accessing sexual health information online: use, motivations and consequences for youth with different sexual orientations. *Health education research*, 29(1):147–157.

Jonathan Peters. 2020. Sexual content and social media moderation. *Washburn LJ*, 59:469.

Charles Pierse. 2021. Transformers Interpret.

Christian Platzer, Martin Stuetz, and Martina Lindorfer. 2014. Skin sheriff: a machine learning solution for detecting explicit images. In *Proceedings of the 2nd international workshop on Security and forensics in communication systems*, pages 45–56.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

N. Rea, G. Lacey, R. Dahyotit, and R. Dahyot. 2006. Multimodal periodicity analysis for illicit content detection in videos. In *The 3rd European Conference on Visual Media Production (CVMP 2006) - Part of the 2nd Multimedia Conference 2006*, pages 106–114.

Randal W Summers. 2016. *Social Psychology: How Other People Influence Our Thoughts and Actions [2 volumes]*. ABC-CLIO.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*.

Adrian Ulges and Armin Stahl. 2011. Automatic detection of child pornography using color visual words. In *2011 IEEE international conference on multimedia and expo*, pages 1–6. IEEE.

Donghui Wang, Miaoliang Zhu, Xin Yuan, and Hui Qian. 2005. Identification and annotation of erotic film based on content analysis. In *Electronic Imaging and Multimedia Technology IV*, volume 5637, pages 88–94. SPIE.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jing Zhang, Lei Sui, Li Zhuo, Zhenwei Li, and Yuncong Yang. 2013. An approach of bag-of-words based on visual attention model for pornographic images recognition in compressed domain. *Neurocomputing*, 110:145–152.

Haiqiang Zuo, Ou Wu, Weiming Hu, and Bo Xu. 2008. Recognition of blue movies by fusion of audio and video. In *2008 IEEE International Conference on Multimedia and Expo*, pages 37–40. IEEE.

A Details of Methods used to collect videos

For sexually suggestive and sex education videos, the annotator interacted with the platform to collect the data in many ways, including search (hashtags, names of people), people reusing the same audio, stitches, duets, the public liked videos of certain profiles pages and the “For you” page. Any video that did not appear to belong to either sexually suggestive or sex education was collected and labeled as Others.

A.1 Sexually Suggestive and Sex ed Videos Videos

- **Search :** Hashtags (including slang usages like #spicyaccountant), Phrases, and Names of popular creators in a domain (discovered through blogs that talk on the subject).

- **Audio Sharing:** TikTok offers multiple people to share and reuse the same audio. So, when a video is found to be, say, sexually suggestive, new creators were discovered by looking into who else used this audio for their video.

- **Stitches and Duets:** A **Duet** allows one creator to post their video side-by-side with a video from another creator on TikTok. A duet contains two videos on a split screen that play at the same time.A **Stitch** is a creation tool on Tiktok that allows a creator to combine another video on TikTok with the one they are creating. Certain videos added in the dataset were discovered as stitches or duets with another creator.

- **Public liked videos:** It is possible to see all videos a certain profile likes by visiting that tab on their profile. By default, this is private but can be set to public. Some profiles share videos of a topic by redirecting visitors to their liked videos. Many videos were found and added to the dataset through this method.

- **"For you" Page:** It’s a recommended feed of videos from creators the user might not follow. The annotator liked and saved videos of sexually suggestive nature, so some similar videos were recommended on the For you Page.

A.2 Other Videos

There are three main strategies for collecting these videos.

- Videos that appeared on the TikTok home page when no user was logged in

- Videos shared with #learnontiktok hashtag

- Videos that reused audio that was also used in a sexually suggestive video.

Each makes up one-third of the total videos collected.

B Detailed stats for transcript length and video length

Parameter	Sugg	Edu	Others	Total
Mean	16.46	231.18	82.18	98.83
Median	14.00	171.50	31.00	33.00
Std	14.33	220.81	126.37	156.08

Table 6: Mean, Median, and Standard Deviation of words present in video transcripts. Words were tokenized using the NLTK package. Sugg stands for Suggestive, and Edu stands for educative. Suggestive videos tend to be significantly shorter than the other classes.

Parameter	Sugg	Edu	Others	Total
Mean	8.96	66.41	39.99	39.06
Median	7.86	50.80	28.30	23.16
Std	3.82	56.92	37.88	42.90

Table 7: Mean, Median, and Standard Deviation of videos in the dataset in seconds. Sugg stands for Suggestive, and Edu stands for educative. Suggestive videos tend to be significantly shorter than the other classes.

C Hyperparameters

Hyperparameters not mentioned below, are default values from Huggingface.

Parameter	Value
Batch size	16
Initial Learning Rate	1e-5
Weight Decay	0.01
Warmup Ratio	0.1
Learning Rate Optimiser	AdamW

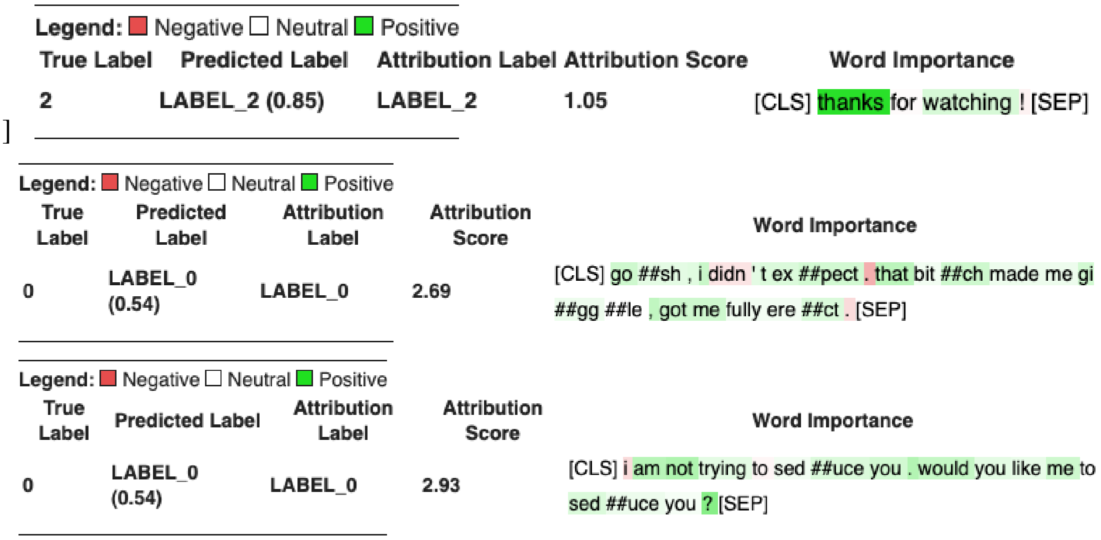
Table 8: Hyperparameters used for the Text Classification Task

Parameter	Value
Batch size	8
Initial Learning Rate	5e-5
Warmup Ratio	0.1
Learning Rate Optimiser	AdamW

Table 9: Hyperparameters used for the Video Classification Task

D Transformer Interpret

Refer to Figure 3 on the next page.



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
2	LABEL_2 (0.85)	LABEL_2	1.05	[CLS] thanks for watching ! [SEP]

]

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	LABEL_0 (0.54)	LABEL_0	2.69	[CLS] go ##sh , i didn ' t ex ##pect . that bit ##ch made me gi ##gg ##le , got me fully ere ##ct . [SEP]

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	LABEL_0 (0.54)	LABEL_0	2.93	[CLS] i am not trying to sed ##uce you . would you like me to sed ##uce you ? [SEP]

Figure 3: Three example transcription and its predictions explanation visualized using Transformers Interpret, a model explainability tool.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Yes. We have. It's not a numbered section and comes right after the conclusion.

☑ A2. Did you discuss any potential risks of your work?
Yes, these are discussed in Ethics on Page 5. Unnumbered section, but immediately follows Limitations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

□ A4. Have you used AI writing assistants when working on this paper?
Not applicable. Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3 introduces the dataset.

□ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 1, page one, footnotes

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 3 introduces and details the dataset. Ethics discussed ethical uses of the dataset.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We state that due to the nature of the task and licensing, it is not possible to anonymize people in the dataset. But the data collected are public information. The dataset contains sexually suggestive content, and this has been repeated throughout the paper.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Such information is not available/not collected.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Tables 1 and 2

C ☑ Did you run computational experiments?

Section 4

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We note that we needed GPU for computation, but the number of hours was not recorded. The paper's focus is on the dataset itself, and details of baselines used were described in detail.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We discuss the experimental setup, including model sources, in Section 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Results are reported in Tables 3 and 4 and are made clear how we reached them. (average of three random runs are reported with the standard deviation. The codebase will be shared if the paper is accepted - for reproducibility and testing.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Sections 3 and 4.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 3

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
The data collected is public.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
May de-anonymise the paper. It will be shared once when the paper is accepted.