White Paper - Creating a Repository of Objectionable Online Content: Addressing Undesirable Biases and Ethical Considerations

Thamar Solorio[1], Mahsa Shafaei[1], Christos Smailis[1], Isabelle Augenstein[2], Margaret Mitchell[3], Ingrid Stapf[4], Ioannis Kakadiaris[1]
[1]University of Houston, [2]University of Copenhagen [3]Google [4]University of Tübingen

Executive Summary
This white paper summarizes the authors' structured brainstorming regarding ethical considerations for creating an extensive repository of online content labeled with tags that describe potentially questionable content for young viewers. The workshop focused on four topics: 1) identifying risks for unintended biases in the data and labels, 2) how to reduce risks for unintended biases; 3) identifying ethical considerations of the annotation task, and 4) reducing the risks for the annotators.

A. BACKGROUND

Watching content online, whether it is movies, videos, or other types of content made widely available by the multiple streaming apps (e.g., YouTube, Netflix, Apple TV, Hulu, Amazon Prime, TikTok, Facebook Live), has become a prime form of entertainment for people of all ages. These mass media platforms usually offer harmless content that, in many cases, can be educational, such as the TV program Sesame Street, which has been shown to improve social and reading skills in young children [1]. However, there is also strong evidence showing the many adverse effects of media in young viewers, that range from instigating violent behavior [2]–[4], perpetuating (sex, race, age)-ism mentality, irresponsible sexual or alcohol consumption, to inducing anxiety and fear with possible long-term traumatic effects and content that seems to justify the harming of human dignity. It should be noted that although some of the platforms such as YouTube and Instagram or Facebook filter obvious and extreme forms of sensitive content, there are tons of potentially harmful content available on the internet that intentionally or unintentionally negatively target children and young viewers. Important here is the mobility of devices. Children do not only want to use children's content and find problematic content either on purpose (due to lack of protection - e.g., filters not working on mobile devices) or by accident. Children can easily find problematic content, and it is hard for parents to restrict what children see (or even know about it because they do not see it happen, are not in contact with a child, or simply lack the competence themselves to regulate it).

We define objectionable content as material that can negatively affect young viewers regardless of the severity of the adverse effect. We consider adverse effects as those documented in the fields of psychology and mass media communication. A proposed ontology of objectionable content and a discussion of risks associated is described in [arXvix].

The goal of the intended repository is to support the development of new technology that can reliably identify the presence of objectionable content in a video. This desired technology is meant to support the decision-making processes of parents/guardians, as well as young adults, of whether the content in the video is appropriate for their children and/or themselves. This is different from providing a rating system, such as that from the Motion Pictures Association of America, which is commonly assigned to movies. It is also different from a filtering app. By tagging content with descriptors of objectionable material, users will have the autonomy to decide if the content respects their views/values. It is something they would be comfortable allowing their kids to watch, or watch for themselves, in the case of young adults. The definition

can be culture-specific or adjusted to certain value systems based on general assumptions on what could cause harm/discomfort. That could be important depending on what use/users (e.g., global, national, regional) it is developed for.

B. Risks for unintended biases on labeling objectionable content

A taxonomy of objectionable content should be as comprehensive as possible to accommodate the gamut of different backgrounds, preferences, user values, and risk levels in the general population. To decide what is objectionable to a child, we need to distinguish between content, context, and child [5], [6]. A global perspective on this issue should consider that there are multiple assumptions about "who the child is" and "what a good childhood looks like," "what children should or should not be exposed to." Ignoring extreme cases, there is a considerable grey zone. Children with different capabilities and family cultures, including media habits [6], absorb different amounts of content and react differently to the content. As mentioned earlier, culture, family situation, and religious beliefs are a few of many effective factors when determining what is problematic content.

An additional complicating factor is the "context factor in the content." A good solution addresses the detection of a prominent form of objectionable content, but it needs to be dependent on the context embedded in the narrative. For example, is a child identified with acts of violence or not? What is the type of movie and narrative? How is the potentially problematic content depicted, and how much of the video does it cover? Does violence seem a means to an end in itself - in that it is justifiable no matter what? Is violence embedded in a system of values that are actually negotiated? For instance, the depiction of violence in a war movie, such as "Saving Private Ryan," would not have the same effect on children as a movie that mixes violence with fun (like violent acts by the "Joker" character in "The Dark Knight"), or that portrays violent acts in a humoristic manner.

To alleviate some of the aforementioned challenges, we intend to incorporate fine-grained and contextual tags that also include a level of intensity (None, Moderate, Severe) and a clear definition of what we mean by these levels. In other words, we will not generate a single rating for the entire content; however, we will provide a list of potentially objectionable content present in the video. We plan to provide a set of descriptors [7] to inform the users about the content. We assume that as the parents/guardians of young viewers, these adults can take the information they have about their children. Using the descriptors, we intend to provide. They will have the background knowledge to make informed decisions about content.

C. Reducing and mitigating risks for undesirable biases in data collection

Creating a dataset that is bias-free 100% may not be feasible at this point. It is our experience that all datasets have biases [8]. Our objective is to minimize potential unwanted biases as much as possible by identifying them and devising mitigation strategies (e.g., Datasheets for Datasets [9], Model Cards [10]). The risks of potential biases in the proposed repository can be categorized along the following dimensions: 1) topics covered; 2) languages and dialects covered; 3) geographical bias (for example, YouTube is not available in some countries, so if we only use this platform, we will only consider those countries that have access to this platform); 4) temporal bias (based on when we collect data we are biased only to that duration of the time); 5) gender bias (depending on the platform, most of the content may be posted by a specific gender - e.g., more than 80% of edits by an editor with a declared gender on Wikipedia are made by men [11]); 6) Other sociodemographic dimensions with risks of biases in our dataset: race/ethnicity, age, religion, culture, economic status, ability, and LGBTQ.

To address this issue, in the best-case scenario, we need to collect videos that are diverse over these protected categories for both "creator" of the videos and "content" of the videos. Unfortunately, most of the time, we do not have access to the creators' demographic and personal information. On the other hand, we can capture the personality type of the creator by capturing the type of content they post. As a result, our objective will be to collect diverse content that covers the protected categories that we mentioned and make sure that the result of the model is not biassed towards a specific group.

D. Biases and ethical considerations for the annotation task

Biases are part of being human since they part of the cognitive process. Therefore, we have to plan accordingly to elicit annotations that will support the objectives of this collection with a minimum amount of bias. To do so, we can start by training the annotators on the topic of unconscious bias. Unconscious bias training aligns annotators across different kinds of social categories. In addition to unconscious bias training, an annotation task should recruit a diverse pool of annotators. This includes annotators from different geographic regions and as diverse in their demographics as possible. After the annotation process completes, a few things can be done to reduce the effects of potentially radical input in the annotations collected. There are tools, such as the Implicit Association Test (IAT)[5], that measures attitudes and beliefs that people may be unwilling or unable to report. This test can be a way to evaluate the fairness of annotators, but it also has its own weaknesses, so additional considerations need to be made here.

An alternative approach is to administer a "qualification test" to annotators and evaluate results in comparison with predefined gold standard answers. Here too, there should be a careful consideration of what is considered the gold standard or how the gold standard is generated. A more comprehensive approach could be to have a combination of IAT and Qualification Test, and we can use the results either to choose annotators or give weight to their response. Other considerations involve the process of aggregating annotations into a single gold standard. We have different options available. For example, gold standard tags can be computed as the mean or median value of the annotations. Another way is to have an expert doing final adjudications when annotators disagree on some labels. We can also choose a majority vote and look at the inter-annotator agreement and remove instances from the dataset that do not have a high agreement. These instances can later be revisited to investigate the source of the disagreement. Another approach is to learn how to aggregate annotations [12] by identifying which annotators are reliable and which ones are not. If the problem lies in the annotation guidelines, then we can revise them and initiate another round of annotations with the improved guidelines.

E. Mitigating Risks for the Annotators

One major risk and affording ethical consideration in annotating sensitive content is that annotators themselves may be negatively affected by the content they will see. Annotators from typically marginalized groups may be particularly vulnerable, as they themselves might have experienced the behavior shown in the videos. This presents a dilemma. On the one hand, we would like to have annotators that are sensitive enough about the different types of objectionable content in videos, but on the other hand, we want to avoid exposing subjects to relive a traumatic experience. Below we describe a possible scenario:

Prior to the annotation: potential alert annotators about the content that they will see by using clear descriptions of each type of objectionable content, in addition: 1) we can show a couple of examples before the annotation process starts as part of the consent process; 2) give annotators an outlet to talk about how the task is affecting them; 3) limit the time they can

annotate material and introduce regular breaks. Longer annotation spans might result in increased sensitivity to the material; 4) give annotators support sources, coping mechanisms, and pointers to strategies to develop emotional resilience; and finally, 5) introduce relaxing and funny videos regularly during the annotation to give them a break from watching potentially objectionable content.

E. Discussion

This white paper summarizes significant discussion points of an online brainstorming meeting to identify ethical considerations and potential biases around planned efforts to create an extensive repository of online videos annotated with objectionable content tags. There are several important considerations that emerged:

- Young viewers are not a monolithic group, even when age distinctions are considered. Therefore, it is not advisable to use a viewer guideline that is merely based on audience age. Moreover, an audience rating is also problematic because it can be interpreted as a form of censorship. However, content labels can still be helpful mainly if the intent is to inform parents/guardians and/or viewers about the content included in the video.
- To mitigate biases in the content of the repository, we should aim to collect videos from a diverse user pool. This should also include inspection of the collected data to expose unintended biases in topics where protected groups might be overrepresented as the victims or targets of objectionable content.
- For the annotation process, we must strive for a diverse pool of annotators, and the annotation task must be carefully designed to reduce the risk of emotional trauma to annotators.

Acknowledgment: This work was supported in part by the National Science Foundation under Grant No. IIS-2036368. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] M.-L. Mares and Z. Pan, "Effects of Sesame Street: A meta-analysis of children's learning in 15 countries," *Journal of Applied Developmental Psychology*, vol. 34, no. 3, pp. 140–151, May 2013, doi: 10.1016/j.appdev.2013.01.001.

[2] S. M. Coyne *et al.*, "Parenting and Digital Media," *Pediatrics*, vol. 140, no. Suppl 2, pp. S112–S116, Nov. 2017, doi: 10.1542/peds.2016-1758N.

[3] K. P. Dillon and B. J. Bushman, "Effects of Exposure to Gun Violence in Movies on Children's Interest in Real Guns," *JAMA Pediatr*, vol. 171, no. 11, pp. 1057–1062, 01 2017, doi: 10.1001/jamapediatrics.2017.2229.

[4] E. Scharrer and G. Blackburn, "Images of Injury: Graphic News Visuals' Effects on Attitudes toward the Use of Unmanned Drones," *Mass Communication and Society*, vol. 18, no. 6, pp. 799–820, Nov. 2015, doi: 10.1080/15205436.2015.1045299.

[5] A. R. Lauricella, M. B. Robb, and E. Wartella, *Challenges and Suggestions for Determining Quality in Children's Media*. Routledge Handbooks Online, 2013.

[6] S. Livingstone, G. Mascheroni, and E. Staksrud, "Developing a framework for researching children's online risks and opportunities in Europe," Nov. 2015. http://www.lse.ac.uk/media@lse/research/EUKidsOnline/Home.aspx (accessed Jan. 24, 2021).

[7] "What do the labels mean? | Pegi Public Site." https://pegi.info/what-do-the-labels-mean (accessed Jan. 24, 2021).

[8] "Disembodied Machine Learning: On the Illusion of Objectivity in NLP," *OpenReview*. https://openreview.net/forum?id=fkAxTMzy3fs (accessed Jan. 24, 2021).
[9] M. R. Costa-jussà *et al.*, "MT-Adapted Datasheets for Datasets: Template and Repository," *arXiv:2005.13156 [cs]*, May 2020, Accessed: Jan. 24, 2021. [Online]. Available: http://arxiv.org/abs/2005.13156.
[10] M. Mitchell *et al.*, "Model Cards for Model Reporting," *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, Jan. 2019, doi: 10.1145/3287560.3287596.
[11] "Wikipedia: Who writes Wikipedia?," *Wikipedia*. Jan. 22, 2021, Accessed: Jan. 24, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Wikipedia:Who_writes_Wikipedia%3F&oldid=1001937086.
[12] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy, "Learning Whom to Trust with MACE," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, Jun. 2013, pp. 1120–1130, Accessed: Jan. 24, 2021. [Online]. Available: https://www.aclweb.org/anthology/N13-1132.