White Paper - Creating a Community of Scholars: Automatic labeling of questionable online content

Ioannis Kakadiaris[1], Christos Smailis[1], Mahsa Shafaei[1], Hugo Jair Escalante[2], Elisa Ricci[3], Albert Ali Salah[4], Vitomir Struc[5], Thamar Solorio[1]

[1]University of Houston, [2] INAOE, [3]University of Trento, [4]Utrecht University, [5]University of Ljubljana

Executive Summary

This white paper summarizes the authors' structured brainstorming regarding creating an extensive repository of online content labeled with tags that describe potentially questionable content for young viewers. The workshop focused on three topics:

1. Creating a community of scholars that will contribute to the problem
2. Establishing broad definitions of what constitutes questionable content and of their sources
3. Setting a sound and ethical approach to data collection and annotation.

A. BACKGROUND

The lead research team organized a series of workshops with experts of different fields to gather relevant feedback regarding the three topics mentioned above. The first brainstorming meeting [arXiv] covered psychologists' and media experts' opinions to understand better what types of content can have a significant negative impact on young audiences and the role of pro-social content.

The second brainstorming meeting [arXiv] covered the views of experts in Ethics and AI. This meeting's outcome shed light on the risks of unintended biases in a system that describes media content. Also, we identified potential elements of the research with risks for misuse and identified threats to annotators and possible solutions to prevent trauma to them from the content they viewed.

In this meeting, representatives of the computer vision community brainstormed about the challenges of the problem relevant to computer vision and the identification of annotation settings that can benefit the larger computer vision community.

B. Creating a Community of Scholars

It was suggested that it would be beneficial to build a community of scholars around the repository of questionable online content in order to have a plurality of researchers approaching the problem. One way to accomplish this is yearly competitions/challenges focused on specific challenges/tasks. Each challenge can be structured using predefined datasets, metrics, and tasks that the community will use. To maximize the attention and the potential impact that the resource will get, each task will focus on a specific area of the problem, trying to address specific challenges, as it will become difficult to attract interest from different communities otherwise. We thus want the resource to be targeted to experts working on different modalities (e.g., image, audio, text).

Computer scientists often adopt a very reductionistic approach to solving problems. This happened due to their tendency to narrow their perspective by defining concrete labels. However, social sciences do not work this way. In order to make our repository useful across different fields and especially the humanities (e.g., sociology, psychology), we need to bring specific tools/applications that will help answer specific questions by social scientists. Therefore, a scientific steering committee can be formed that will review various proposals for challenges and will select the ones that are doable and will attract the most interest from the scientific community.

Yearly Competitions: For example, the steering committee can hold yearly competitions based on the repository for each of the fields that its members are part of. This will unfold because people will submit ideas, and the steering committee will decide and then approve them and authorize the collection of content.

Venues: The venues that this type of research would be most appropriate for should be multimedia related conferences such as the ACM International Conference on Multimodal Interaction (ICMI) and targeted workshops of major computer vision conferences such as the Conference on Computer Vision and Pattern Recognition (CVPR), the International Conference on Computer Vision (ICCV) and the European Conference on Computer Vision (ECCV). Also, this type of research might be of interest to the community of the IEEE International Conference on Automatic Face and Gesture Recognition Conference (FG).

C. Broad definitions of what constitutes questionable content and its sources

In this repository, we opted for labeling content as questionable. The word questionable was chosen to describe the content of the repository so as to avoid imposing any specific perspective about the resources that creates a specific judgment, as the goal of this resource is to assist in describing multimedia video content and not impose censoring. We thus avoided using words in the content type ontology that can introduce bias. However, broad terms that describe whether a specific type of content exists or not were preferred as it is harder for them to be biased (e.g., violence)

It is clear that the cultural setting is important for defining the terms. For example, a picture of circumcision may be labeled as gore for western cultures, but it is a very common sight for those that participate in them. Additionally, the presence of nudity in a film can be considered as sensitive material in some cultures, while other cultures are more tolerant of it. Avoiding using culturally dependent terms to describe content types can be a way to limit this problem by making the terms less controversial.

Morality is often culture-dependent, as has been demonstrated by the moral machine experiment from MIT [1]. Thus, we can create a web interface to collect labels for a variety of cultures. Adding an annotation type related to the age that a parent would allow his/her kid to watch specific types of content would be a way to capture more information.

Regarding the sources of video data that will be used in our repository, one option is to use publicly available videos from streaming websites such as YouTube or Reddit and movie publishing organizations. However, since the web is a space preserving the privacy of content, uploaded content has become an increasingly important social and legislative need and concern. To this end, we will opt for asking for permission before collecting data that are publicly available from content owners and ensure that their data can be removed from our resource even after providing initial consent according to GDPR standards and by consulting privacy scholars that can assist us in the design of the data collection process. We propose that data will not be able to be downloaded from our repository.

D. Establishing a sound and ethical approach to data collection and annotation

In order to annotate the data in our repository, there are several factors regarding minimizing the bias of the process. Specifically, we need to collect balanced representative groups of annotators regarding demographics such as gender, race, and age. The collected information could serve to identify potential biases in annotators and to improve the labeling process accordingly iteratively.

Regarding the annotation units, it is proposed that the annotation proceeds by scene and not by the full video. Thus, labels will be applicable on a scene level basis.

To ensure the annotations' quality, a pilot study can be run to measure the inter-annotator agreement in the participants of the annotation process, using a subset of the data. Annotations will label the content on a temporal basis. However, we may also consider spatially localizing questionable content in individual frames of videos. In general, the more diverse the annotation set, the more fields of computer vision will benefit from our resource. To accomplish this goal, using semi-automated methods for annotation will help.

To ensure the well-being of the annotators, since they may be exposed to content that has the potential of mentally traumatizing them, special measures need to be taken so as to minimize the impact of the content on their mental well-being. For example, after being exposed to a small number of questionable videos, a more cheerful video could be shown to them.

In addition, to minimize different bias factors such as the annotators' cultural background, the labels used during the annotation process to describe the data should refer to well-defined broad terms. These terms will explain whether a specific type of content exists or not. Since the labels will be general, there will be sub-labels to make the broad categories semantically rich. For example, the category "aggression" will also have sub-categories that will specify the type of aggression, such as "*Verbal Aggression*" or "*Physical Aggression*."

E. Discussion

In conclusion, the main points raised in this meeting were that providing content descriptors for online video content could be very interesting to various scientific communities in addition to the computer vision community. However, for the resource and the data to be useful, specific steps must be taken in order to minimize several bias factors that can limit its potential impact. Specifically, issues related to bias induced by annotator demographics and the definition of the ontology of labels must be taken. Quality control for the annotation process has to be in place by establishing specific annotation guidelines and inter-annotator measures. Regarding creating a community of scholars, establishing a cross-disciplinary steering committee to promote the repository by defining yearly tasks is desirable.

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 of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] E. Awad, S. Dsouza, A. Shariff, I. Rahwan, J.-F. Bonnefon (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. Proceedings of the National Academy of Sciences. 117(5)