Identifying Harmful Video Content With Movie Trailers and Machine Learning

Originally published at https://www.unite.ai/identifying-harmful-video-content-with-movie-trailers-and-machine-learning/

A research paper from the Swedish Media Council outlines a possible new approach to the automatic identification of 'harmful content', by considering audio and video content separately, and using human-annotated data as a guiding index for material that may disturb viewers.

Entitled *Is this Harmful? Learning to Predict Harmfulness Ratings from Video*, the paper illustrates the need for machine learning systems to take account of the entire context of a scene, and illustrates the many ways that innocuous content (such as humorous or satirical content) could be misinterpreted as harmful in a less sophisticated and multimodal approach to video analysis – not least because a film's musical soundtrack is often used in unexpected ways, either to unsettle or reassure the viewer, and as a counterpoint rather than a complement to the visual component.

A Dataset Of Potentially Harmful Videos

The researchers note that useful developments in this sector have been impeded by copyright protection of motion pictures, which makes the creation of generalized open source datasets problematic. They also observe that to date, similar experiments have suffered from a sparsity of labels for full-length movies, which has led to prior work oversimplifying the contributing data, or keying in on only one aspect of the data, such as dominant colors or dialogue analysis.

To address this, the researchers have compiled a video dataset of 4000 video clips, trailers cut down into chunks of around ten seconds in length, which were then labeled by professional film classifiers that oversee the application of ratings for new movies in Sweden, many with professional qualifications in child

psychology.

Under the Swedish system of film classification, ‘harmful’ content is defined based on its possible propensity to produce feelings of anxiety, fear, and other negative effects in children. The researchers note that since this ratings system involves as much intuition and instinct as science, the parameters for the definition of ‘harmful content’ are difficult to quantize and instill into an automated system.

Defining Harm

The paper further observes that earlier machine learning and algorithmic systems addressing this challenge have used specific facet detection as a criteria, including the visual detection of blood and flames, the sound of bursting, and the frequency of shot length, among other restricted definitions of harmful content, and that a multi-domain approach seems likely to offer a better methodology for automatic rating of harmful content.

The Swedish researchers trained an 8×8 50-layer neural network model on the Kinetics-400 human movement benchmark dataset, and created an architecture designed to fuse video and audio predictions.

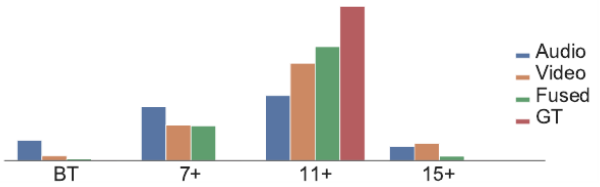
In effect, the use of trailers solves three problems for the creation of a dataset of this nature: it obviates copyright issues; the increased turbulence and higher shot frequency of trailers (as compared to the originating movies), allows for a greater frequency of annotation; and it ensures that the low incidence of violent or disturbing content in an entire movie does not unbalance the dataset and accidentally class it as suitable for children.

Results

Once the model was trained, the Swedish researchers tested the system against video-clips.

In this trailer for *The Deep* (2012), the two models used for testing the system (randomly sampled labels vs. probabilistic labels) successfully classified the movie as suitable for viewers aged 11 and over.

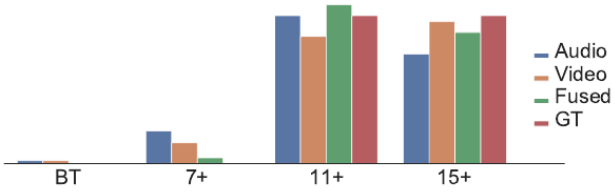
7c07587540f8



Source: https://arxiv.org/pdf/2106.08323.pdf

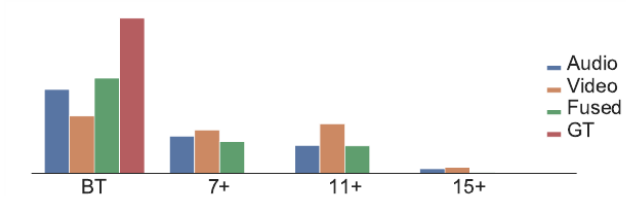
For a scene from *Discarnate* (2018) where a monstrous antagonist is introduced, the dual framework again correctly estimated the target age range as 11+/15+.

c373f1f5efe7



However, a clip from the trailer for *A Second Chance* (2014) presented greater difficulty, since the model was unable to agree with the human annotations for the scene, which had classified it as ‘BT’ (universally acceptable). In effect, the algorithm has detected potential for harm which the human evaluators have not ascribed to it.

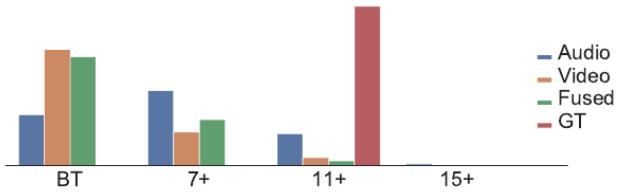
522ca5c192c1



Though the researchers attest a high accuracy score for the system, some failures did occur, such as this clip from *City State* (2011), which features a detained naked man threatened with a rifle.

6d1718bc7b91

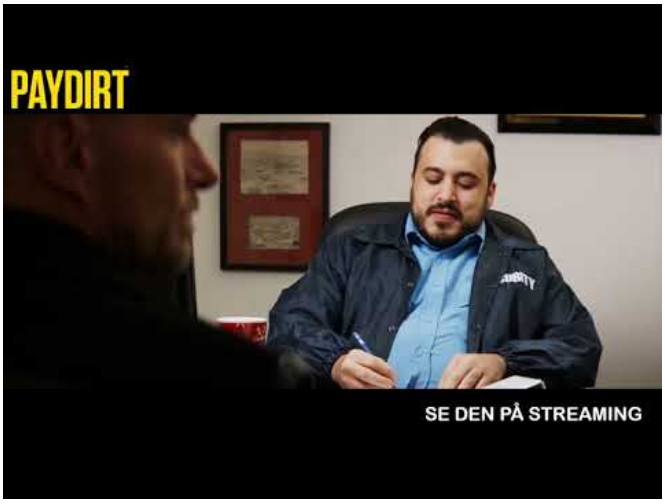
In this case, the system has assigned an 11+ rating to the clip, in contrast to the human annotations.

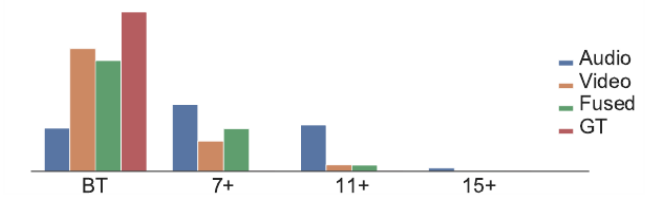


Dissonance Of Intent And Harmfulness

The paper notes that in evaluating a clip from the trailer for *Paydirt* (2020), the system correctly assigns a 'universal' rating to the clip based on the visual and linguistic aspects (though characters are discussing firearms, the intent is comedic), but is confused by the dissonantly threatening music used, which may have a satirical context.

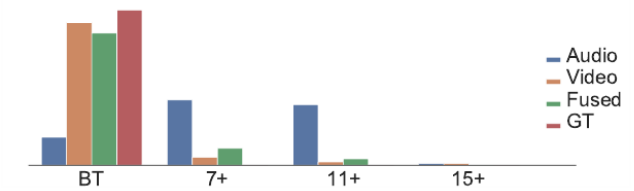
16b8d1ce0fa1





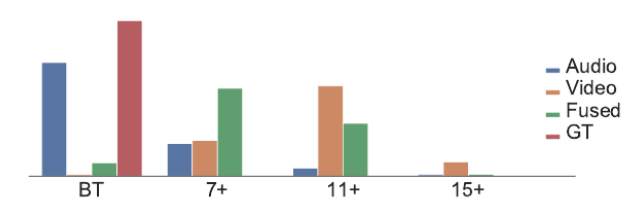
Likewise in a trailer for the film *For Sama* (2019), the threatening style of the musical content is not matched by the visual content, and once again, the system experiences difficulty in disentangling the two components to make a uniform judgement that covers both the audio and video content of the clip.

7e2bc2a6f2b3



Finally, the system correctly navigates audio/video dissonance in a trailer clip for *Virgin Mountain* (2015), which contains some threatening visual cues (i.e. a broken window) that are undermined by the music. Thus the framework correctly guesses that the clip is rated 'universal' (BT).

80dcdfbb9f9e



The researchers concede that a system of this nature is exclusively focused on children, with the results unlikely to generalize well to other types of viewer. They also suggest that codifying 'harmful' content in this linear way could potentially lead to algorithmic rating systems that are less unpredictable, but note the potential for unwanted repression of ideas in the development of such approaches:

> *'Assessing whether content is harmful is a delicate issue. There exists an important balancing act between freedom of information and protecting sensitive groups. We believe that this work takes a step in the right direction, by being as transparent as possible about the criteria being used to assess the harmfulness. Furthermore, we believe separating harmfulness from appropriateness is an important step towards making classification of harmful content more objective.*
>
> *'…Detecting harmful content is also of interest to online platforms such as YouTube. On such platforms, the balancing act between information freedom and protection becomes even more important and is further complicated by the proprietary nature of the algorithms responsible.'*

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai