Differences Between Pretrained Models and Models Trained on Children's YouTube Content

Weston Van Erp

Abstract

The increasing prevalence of digital media, especially YouTube and TikTok, has shifted young children's content consumption from traditional television to online platforms, raising concerns about content quality and its impact on cognitive development. Unlike regulated television programming, content on these platforms is often saturated, commercially-driven, and less educational. This paper proposes an NLP-based method to examine language patterns in children's online media through token prediction. By training a custom model on a corpus of creator-generated transcripts and comparing it to a pre-trained model, we aim to identify key linguistic and thematic trends. Our findings offer insights into the potential developmental effects of current digital content for young audiences.

1 Description/Motivation of the Problem

Technology and access to it has rapidly evolved during the past two decades, and some of the most tangible innovations have been the advent and widespread infiltration of iPads and YouTube. 43 percent of Generation Alpha children have a tablet before the age of 6 (Basis) and more children in the 8-10 consume YouTube Shorts and TikTok than traditional television (AECF). 4 out of the 7 largest YouTube channels (measured by number of subscribers) are made-for-kids channels, primarily targeting children under 10 years old. If the scope were expanded to consider a primary audience of pre-teens, it would also encompass the largest channel on the platform, Mr Beast (Socialblade). The creative process that content creators on these internet platforms follow is less regimented and regulated than that of television programs, and this presents a problem when considering content intended for children. Some parents and psychologists worry about the highly saturated, non educational, greed-inducing content being consumed by today's children and its impact on brain development (CNBC). The novelty of this industry is part of what makes it so concerning, as its effects will not be properly understood for years to decades, and the vast number of content creators making made-for-kids content makes it hard to adequately assess. Natural Language Processing is another field that has seen a lot of development in recent years and may offer some insight into this situation. Language models specialize in consuming large amounts of text and completing next-token-prediction tasks. If a model could be trained on these videos and its token prediction results were compared to that of a "normal" pre-trained model for some given instances of interest, more specific discourse about the subject could be generated without requiring experts to think critically about an impossibly large collection of content.

2 Proposed Solution

This project looks at transcripts from 7,000 videos uploaded by the most popular made-for-kids YouTube channels (Socialblade). The corpus consists of transcripts from channel videos in the following quantities: 2,148 Ryan's World, 942 T-Rex Ranch, 1,169 Toys and Colors, 1,211 Kids Diana Show, 710 Vlad and Niki, 820 Like Nastya. Only English transcripts were analyzed for this project. This project assumes that focusing on the analysis of top channels provides a realistic representation of a hypothetical child's media diet, which appears reasonable when considering the top-heaviness of the industry. This project eschewed large channels such as Cocomelon that primarily produce nursery rhymes, as the focus of this project is ordinary language, not lyrics. More than 250,000 lines of captions were compiled and used to fine-tune the pre-trained BERT-based-uncased model. Then, it completed a series of fill-in-the-blank tasks alongside a pre-trained BERT-based-uncased model that had not consumed the lines of captions. Each model's

top 5 answers for each prompt and their probabilities were listed. The fill-in-the-blank tasks were chosen to target perceived/expected differences in performance, and/or perceived areas of importance. Using a pre-trained BERT model as a starting point as opposed to a blank slate model ensures that it is not completely devoid of influence other than the videos, which is a more realistic analog to the speech a hypothetical actual child that would be watching the videos.

3 Similar Papers

Papers dealing with this specific issue were unable to be found in the ACL Anthology, but there are two papers that looked into analyzing YouTube closed captioning. One by Rachael Tatman in 2017 looked into gender and dialect biases in automatic closed captioning, but this project will strictly be looking at non-automatically generated captioning for data. There is another paper by Edison Marrese-Taylor, Jorge Balazs, and Yutaka Matsuo in 2017 titled "Mining fine-grained opinions on closed captions of YouTube videos with an attention-RNN". This paper may serve as a jumping off point for general caption collection and analysis. Both papers are in the references section.

4 Description of Process

The process of going from a handful of YouTube channels to next token prediction task results is a four-step process. First, one would navigate to a given channel, click on the "Videos" tab, and then run a Python script. This script would scroll through all videos, adding each one to the user's "Watch Later" playlist. Then, the second script would click through a playlist (in this case, you would use it on the Watch Later playlist), open the video description, click on "Show Transcript", highlight the entire video's transcript, and then paste it into the end of a text document. This step was the most time consuming, as each line of a video's transcript needed to be "manually" highlighted, no keyboard shortcuts could be used. Once the text file had been created, it would be uploaded to the Python notebook on Google Colab, and used to fine tune a pretrained BERT model. Then, it and a non-fine-tuned BERT model were given the next-token-prediction tasks, and the results were displayed.

5 Description/Analysis of Results

The results of the next token prediction will be included, but some analysis of the results could be as follows. The fine-tuned model (the one that consumed made-for-kids content) had many more innocent, optimistic, and playful associations such as "fun", "toys", "bravery", and "colors" while the non fine-tuned model (the one that did not consume the made-for-kids content) hasd more mature associations such as "beer", "cigarette", "think", and "listen". The children's model also focused on more immediate, cause-and-effect actions such as "stop", "help", and "focus" as opposed to "think", "ask", and "write" as with the adult model. Emotions are presented in a more straightforward way with the children's model, using words such as "pretend", "relax", and "fun", while the adult model uses words such as "stay", "expect", and "understand". There are also some knowledge gaps in the children's model and odd associations, such as "The biggest animal in the ocean is the octopus" and "To solve a problem, the first thing you should do is eat". There is a general concern of oversimplicity with the children's model, with a lack of problem-solving and emotional reasoning. Also, heavy reliance on entertainment and play might overshadow educational educational/reflective lessons.

6 Analysis of Limitations

As it currently stands, the text file must be uploaded (which could take approximately 2 minutes) and the model must be trained (which on the T4 GPUs can take approximately 2.5 hours) to achieve results. An attempt was made to create a zip archive of the fine-tuned model, but the excruciatingly long download time means that connection interruptions stifled attempts to do so (it is a full sized BERT model, after all). The predicament is that with currently available computing resources, the notebook needs to be run on the T4 GPUs offered by Google Colab, but models cannot be downloaded from the notebook nor can one afford runtime disconnections. Further limitations include corpus size. While 250,000 lines of transcript is significant, it could be larger and could potentially better represent what it is attempting to.

7 Potential Follow Up Work

Potential follow up work could include many of the things discussed in the Analysis of Limitations

section. There is a seemingly endless supply of made-for-kids YouTube content and going further down the list of the largest channels in that category would only enhance the validity of this project. 1 million lines would be a nice round milestone to shoot for next. Especially if considering quadruple the corpus size, the problem of training time would need to be solved. The current training time on T4 GPUs is approximately 2.5 hours, so if the corpus size were to increase drastically, so too would the training time. More processing power would need to be secured to solve this issue. Also, finding some way to properly store the fine tuned model would be ideal.

8 References

Reschke, Megan. "Generation Alpha: Online Habits and Media Preferences by the Numbers." Basis Technologies, 21 Aug. 2024, basis.com/blog/generation-alpha-online-habits-and-media-preferences-by-the-numbers.

The Annie E. Casey Foundation. "The Impact of Social Media and Technology on Gen Alpha." The Annie E. Casey Foundation, 22 Oct. 2024, www.aecf.org/blog/impact-of-social-media-on-gen-alpha.

Josephinebila. "YouTube's Dark Side Could Be Affecting Your Child's Mental Health." CNBC, CNBC, 13 Feb. 2018, www.cnbc.com/2018/02/13/youtube-is-causing-stress-and-sexualization-in-young-children.html.

Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Edison Marrese-Taylor, Jorge Balazs, and Yutaka Matsuo. 2017. Mining fine-grained opinions on closed captions of YouTube videos with an attention-RNN. In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 102–111, Copenhagen, Denmark. Association for Computational Linguistics.

Top 100 Youtubers Made-for-Kids Channels - Socialblade Youtube Stats | YouTube Statistics, socialblade.com/youtube/top/category/made-for-kids. Accessed 20 Dec. 2024.