

Dự án tin học trẻ

Ứng dụng nhận diện video không phù hợp với trẻ em trên Youtube



April 25, 2024

THPT: Trần khai nguyên

**T**

**PHIẾU THUYẾT MINH SẢN PHẨM SÁNG TẠO  
--------------------**

**ĐĂNG KÝ THÔNG TIN SẢN PHẨM DỰ THI**

|  |  |  |  |
| --- | --- | --- | --- |
| **1** | **Đăng ký Bảng dự thi** | **Bảng D1** | **Bảng D2** |
| **☐** | **☒** |
| **2** | **Sản phẩm dự thi** | **Phần mềm** | **SP tích hợp phần cứng** |
| **☒** | **☐** |
| **3** | **Tên SPST dự thi** | Ứng dụng nhận diện video không phù hợp với trẻ em trên Youtube | |
| **4** | **Ngôn ngữ lập trình hoặc nền tảng** | Python 3.9 , Html, Css, Javascript | |
| **5** | **Cấu hình cài đặt** | Flask==3.0.3  huggingface\_hub==0.23.4  transformers==4.41.2  youtube\_transcript\_api==0.6.2  tf-keras  torch  flask-cors | |

**THÔNG TIN TÁC GIẢ (NHÓM TÁC GIẢ)**

|  |  |  |
| --- | --- | --- |
| **Số lượng thí sinh trong đội thi** | **1 người ☐** | **2 người ☒** |
| **Thí sinh thứ nhất (đội trưởng)** | | |
| **Họ và tên:** | Vương Chánh Hưng | |
| **Ngày/tháng/năm sinh:** | 12/6/2007 | |
| **Lớp, trường:** | 11A15, THPT Trần Khai Nguyên | |
| **Quận, huyện:** | 5 | |
| **Điện thoại:** | 0965645280 | |
| **Email:** | [vuongchanhhung@gmail.com](mailto:vuongchanhhung@gmail.com) | |
| **Thí sinh thứ hai** | | |
| **Họ và tên:** | Mai Lộc Phúc Thịnh | |
| **Ngày/tháng/năm sinh:** | 19/4/2007 | |
| **Lớp, trường:** | 11A13, THPT Trần Khai Nguyên | |
| **Quận, huyện:** | 5 | |
| **Điện thoại:** | 0347720419 | |
| **Email:** | [mailocphucthinh194@gmail.com](mailto:mailocphucthinh194@gmail.com) | |
| **Giáo viên hoặc chuyên gia hướng dẫn** | | |
| **Họ và tên:** | Trần Lê Hùng Phi | |
| **Đơn vị công tác:** | THPT Trần Khai Nguyên | |
| **Chức vụ:** | Tổ trưởng chuyên môn | |
| **Điện thoại:** | 0937125977 | |
| **Email:** | [Tranlehungphi@gmail.com](mailto:Tranlehungphi@gmail.com) | |

**GIỚI THIỆU VỀ SẢN PHẨM**

1. **Ý tưởng của sản phẩm**

Nhóm muốn tạo ra một tiện ích trên nền tảng Google Extension giúp nhận diện những video không phù hợp với trẻ em thông qua, những đoạn hội thoại, cũng như độc thoại trong video.

**2. Giới thiệu tổng quan**

Nhóm có ý định sử dụng công cụ Youtube\_transcript\_api của Python để lấy ra được những đoạn hội thoại, sau đó sử dụng một Pretrained-AI model s-nlp/roberta\_toxicity\_classifier một model được train lại (fine-tune) từ model RoBERTa model, từ nhiều nguồn dữ liệu khác nhau, để lọc ra những đoạn thoại mang tính chất thù ghét, phản cảm… Nhóm sẽ gọi model thông qua api pipeline từ HuggingFace với mục đích tăng tốc độ, và tiết kiệm dữ liệu cho api của nhóm khi host lên cloud. Sau đó trả ra kết quả dưới dạng Json bằng thư viện Flask. Cuối cùng nhóm xây dựng giao diện trước khi đưa lên Chrome web store.

**MÔ TẢ SẢN PHẨM**

**1. Chức năng chính của sản phẩm**

***1.1. Mô tả rõ các tính năng của sản phẩm***

Sản phẩm có khả năng nhận diện các video không phù hợp với trẻ em, chủ yếu là những Video có các đoạn thoại tục tĩu, phản cảm bằng tiếng Anh. Từ đó đưa ra cảnh báo đối với người sử dụng. Đặc biệt là các bậc phụ huynh có con nhỏ ở nhà. Ngoài ra tiện ích còn có khả năng trả về các thông tin chi tiết như số lượng những đoạn thoại có khả năng ảnh hưởng xấu đến trẻ em và chi tiết các đoạn thoại đó, để giúp phụ huynh có cái nhìn tốt hơn.

***1.2. Mô tả nền tảng phát triển của sản phẩm***

- Ngôn ngữ lập trình: Python, Html, Css, Javascript

- Nền tảng để phát triển các ứng dụng: Visual Studio Code, Google colab

- Mô tả hoạt động của phần mềm

Ứng dụng sẽ được xây dựng và đặt trên nền tảng Google extension sau khi người dùng nhấp vào video. Nhóm sử dụng API của ngôn ngữ lập trình Javascript và hàm lấy URL để trả về máy chủ URL của video đó được xây dựng bằng Flask sau khi nhận được URL từ API. Nhóm sẽ sử dụng phương pháp xử lý chuỗi trong Python để lấy video-id. Từ đây tại máy chủ nhóm sẽ gửi video-id về youtube\_transcript\_api để chuyển đổi video thành dạng văn bản. Sau khi nhận được văn bản được trích xuất từ video nhóm sẽ sử dụng model <https://huggingface.co/s-nlp/roberta_toxicity_classifier> có được từ Hugging Face để phân tích. Sau khi phân tích xong mô hình sẽ cho kết quả là NSFW (non-safe for work) hoặc SFW (safe for work) tương ứng với có chứa nội dung độc hại và không. Sau đó gửi về Google Extension kết quả đã được dịch lại của nhóm lần lượt là “Video phù hợp với trẻ em” và “Video được khuyến cáo không phù hợp với trẻ em” kèm theo chi tiết cụ thể. Và cảnh báo cho người dùng.

***1.3. Kết luận:***

- Sản phẩm đã giải quyết được vấn đề gì, kết quả của sản phẩm trong thực tế;

* Sản phẩm có khả năng cao sẽ bảo vệ trẻ em trên môi trường Internet khỏi những nội dung độc hại gây sai lệch nhận thức. Góp phần tạo nên môi trường Internet an toàn lành mạnh để trẻ em có thể tự do khám phá, học hỏi và giải trí.
* Giúp trẻ em nhận thức và tránh xa những video có lời thoại không phù hợp đối với trẻ em.
* Là một đóng góp nhỏ, cho toàn bộ cộng đồng, và có thể được sử dụng hoặc tham khảo như một module nhỏ cho toàn bộ chiến dịch chống lại video độc hại cho trẻ em.

- Những ưu, nhược điểm của sản phẩm.

* Ưu điểm của sản phẩm bao so với các sản phẩm hiện có là tính linh hoạt, độc lập cao không bị phụ thuộc vào nhiều yếu tố điển hình là phụ để của video. Chỉ cần có được video thì có thể thực hiện đánh giá sơ bộ, sàng lọc.
* Nhược điểm của sản phẩm là đôi khi có video độc hại không phải vì những đoạn đối thoại hoặc độc thoại mà là do mang hình ảnh máu me, ghê rợn mang tính bạo lực cao.
* Đôi khi Model AI có thể bị nhầm lẫn giữa những câu thoại chỉ mang tính chất mỉa mai đơn thuần thành một câu nói mang tính thù ghét. Nên kết quả trả về có thể không được chính xác hoàn toàn.
* Hiện tại kết quả được quyết định dựa trên tỉ lệ các đoạn thoại độc hại trên tổng video. Tỉ lệ của nhóm đưa ra hiện tại chỉ mang tính cá nhân, sau khi được lên nền tảng Google Extension, có được nhiều dữ liệu từ phản hồi của người dung thì kết quả sẽ chính xác hơn.

**2. Đánh giá sản phẩm**

***2.1. Tiềm năng ứng dụng***

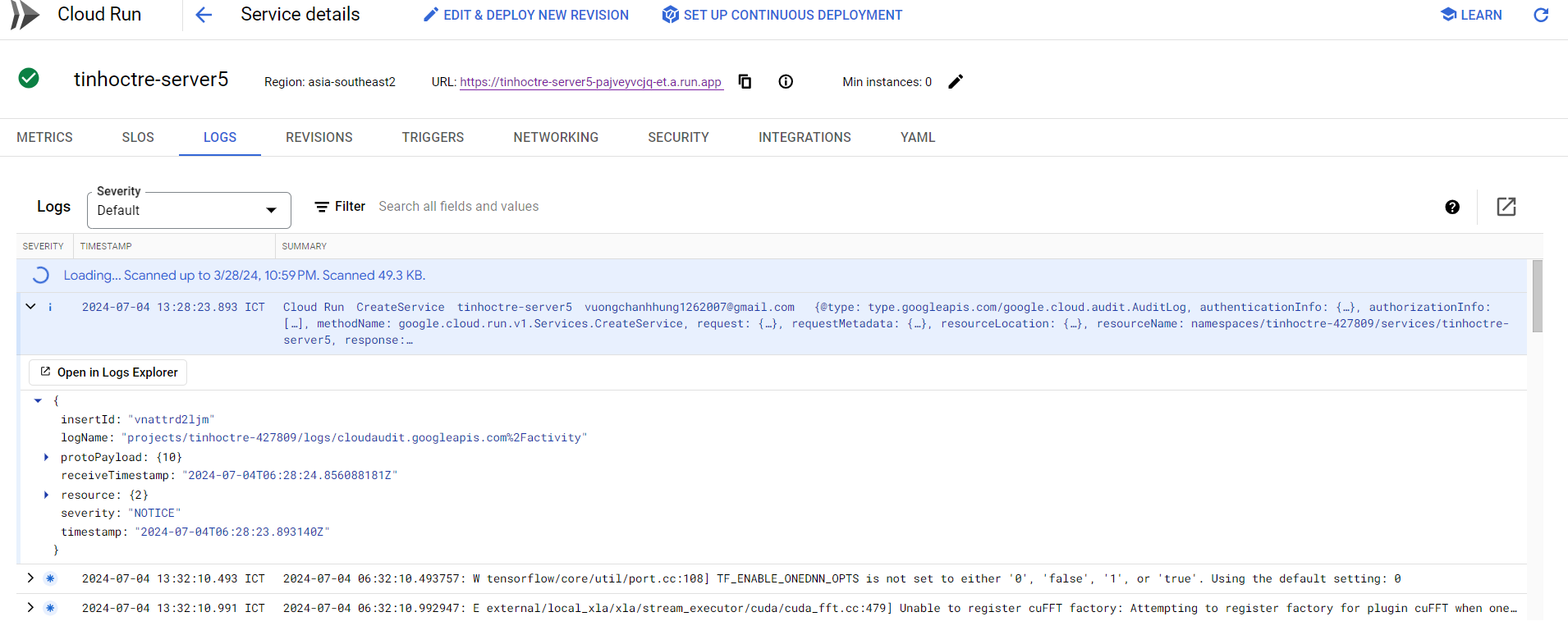
Khả năng ứng dụng của sản phẩm rất có tiềm năng trong việc ứng dụng vào thực tế. Và có thể được các bậc phụ huynh tin dùng. Và cũng như là một công cụ bổ sung hỗ trợ bên cạnh Youtube kids nhằm giảm khả năng trẻ em xem phải những video ảnh hưởng đến sự phát triển, tâm lý của trẻ. Ứng dụng có khả năng ứng dụng trên quy mô toàn cầu bởi vì nền tảng chính để phát triển ban đầu của nhóm là Google Extension với ưu điểm tiện lợi và nhanh chóng. Với độ phủ và phổ biến cao trên phạm vi toàn cầu của trình duyệt Chrome. Điều đó làm cho sản phẩm có tiềm năng phát huy và dễ dàng tiếp cận với nhiều người dùng trên toàn thế giới. Mặc dù hiện tại ngôn ngữ hỗ trợ chính chỉ là tiếng Anh do sự hạn chế trong việc training các Model NLP với các loại ngôn ngữ có độ phức tạp cao. Nhưng trong tương lai nhóm sẽ có thể thực hiện được việc đa dạng hóa ngôn ngữ cho ứng dụng nhờ vào sự nghiên cứu hỗ trợ của cộng đồng AI trên toàn thế giới như Hugging Face và Tensorflow Hub.

Đường dẫn đến mã nguồn của sản phẩm https://github.com/hung1321/inapproriate-video-detector\_

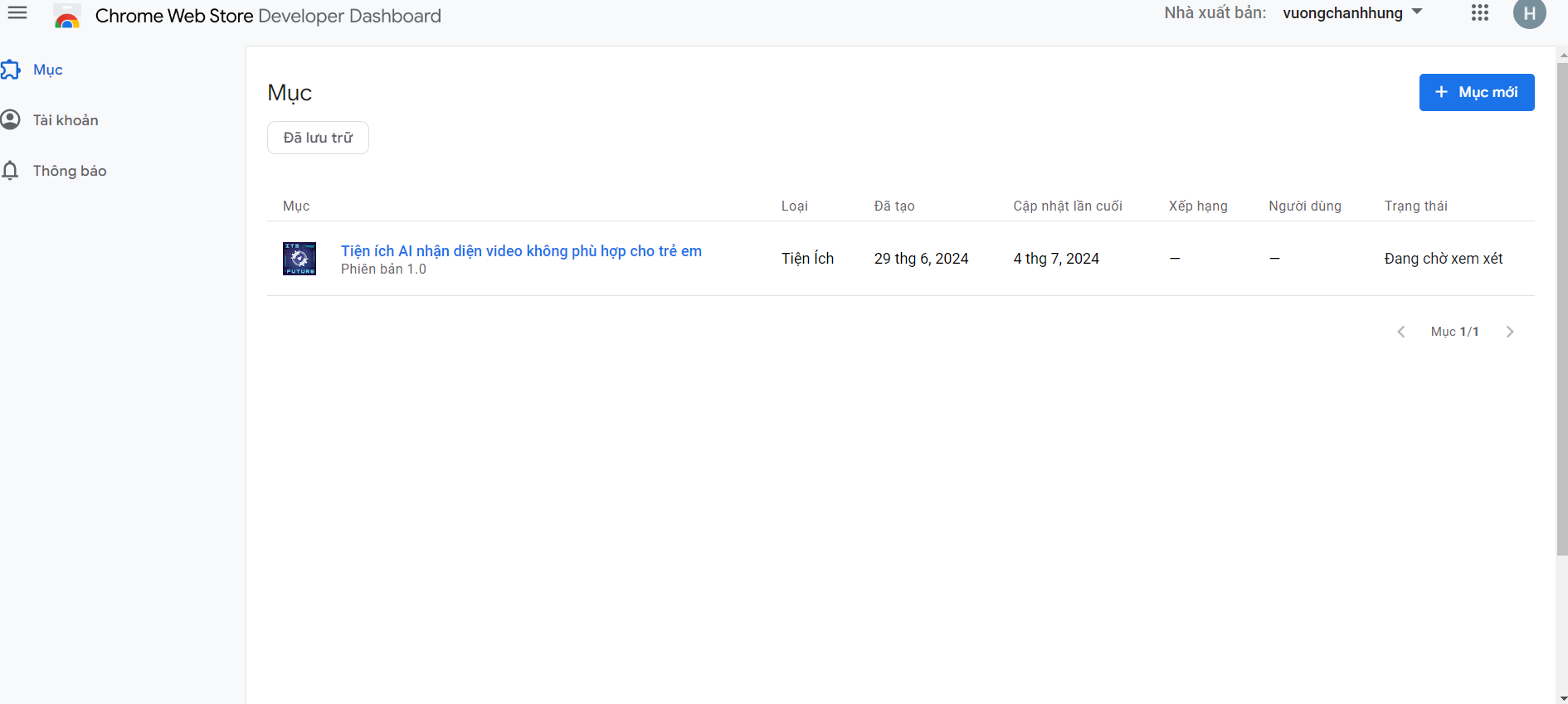
***2.2. Hiệu quả đem lại khi ứng dụng sản phẩm***

Nhóm tin rằng công cụ này là việc vô cùng cần thiết để đảm bảo một môi trường mạng an toàn trong sạch. Bảo vệ trẻ em khỏi những nội dung độc hại trên Internet. Ngoài ra ứng dụng cũng có tiềm năng được thương mại hóa khi đã đáp ứng được tất cả điều kiện của một phần mềm thương mại. Đem lại hiệu quả kinh tế và tài chính bởi vì ứng dụng mang lại nhiều lợi ích thực tế.

1. **Yêu cầu đối với cơ sở hạ tầng cần thiết để triển khai ứng dụng sản phẩm**
   1. Một hosting, vps hoặc là dịch vụ đám mây (nhóm quyết định xử dụng Google cloud Artifact registry)



* 1. Ngoài ra nhóm cũng cài đặt hệ thống giao diện lên Google Extension. Hiện tại sản phẩm của nhóm đã được gửi lên hệ thống để xét duyệt.



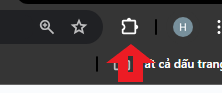
**4. Sản phẩm được phát triển ước tính trong khoảng thời gian:**

Số tháng:1 *(Từ 3 đến 4)*

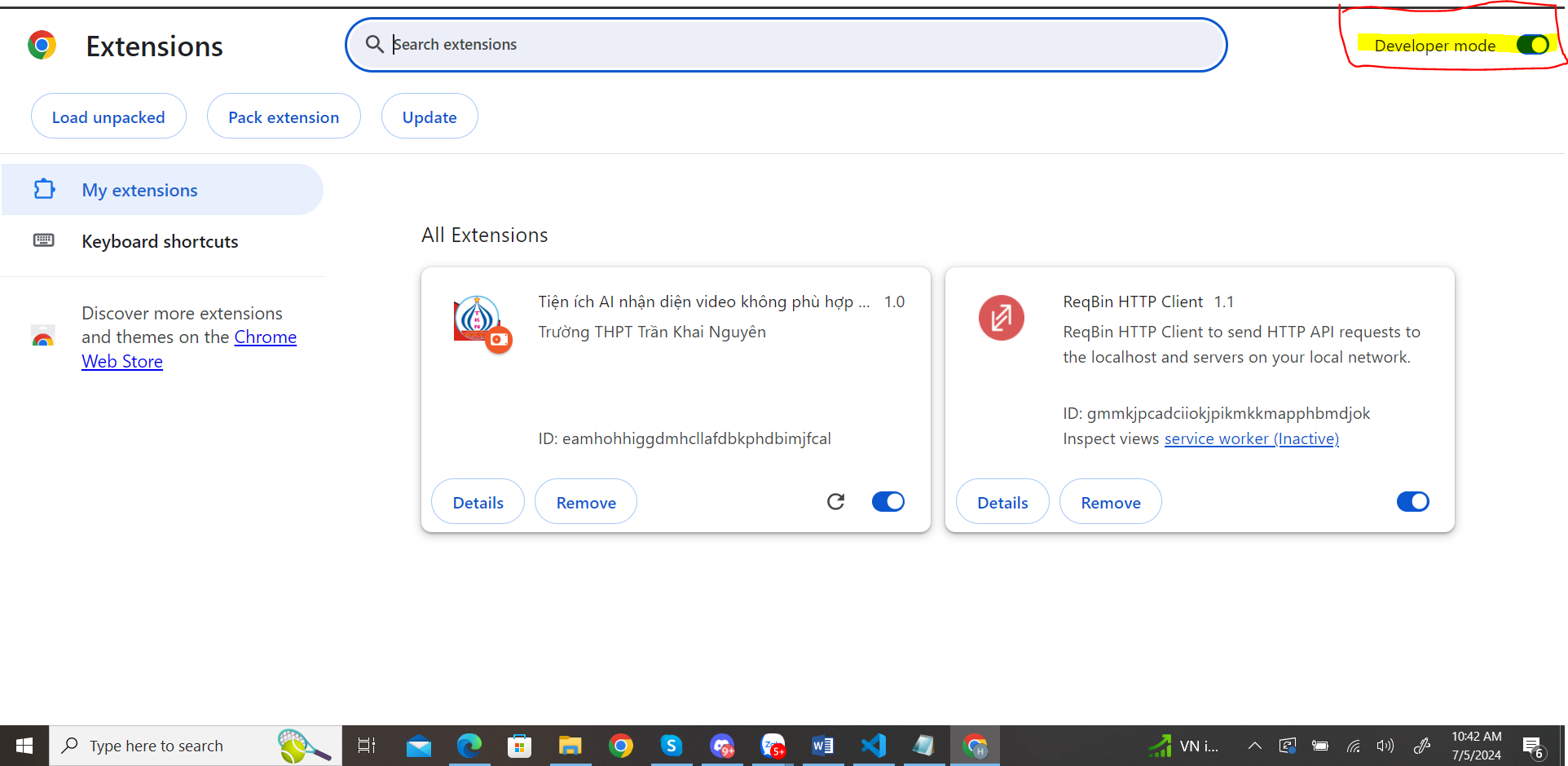
**5. Hướng dẫn sử dụng sản phẩm**

**Hiện tại sản phẩm vẫn đang được kiểm duyệt từ phía Google trước khi được đưa lên Google Extension web store nên hiện tại sản phẩm chỉ được sử dụng local trên máy.**

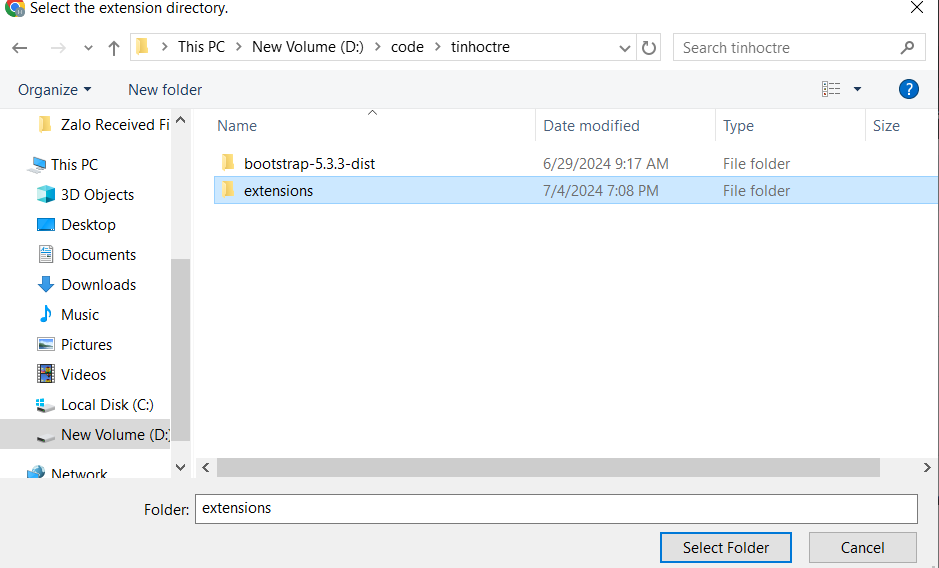
**Để có thể sử dụng Local** trước hết phải tải được Chrome. Sau đó nhấp vào biểu tượng ở góc trên bên phải cửa sổ.



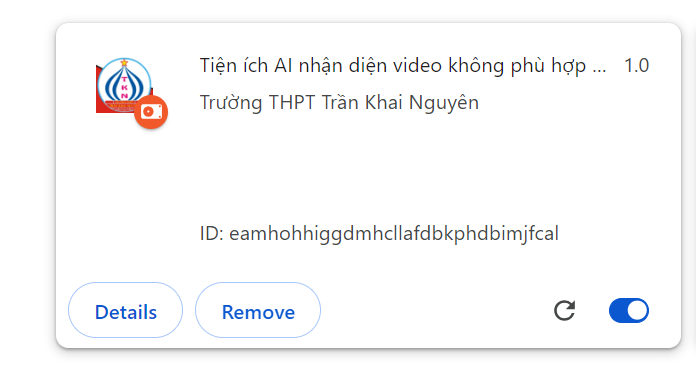
Sau đó mở chế độ **Developer mode**



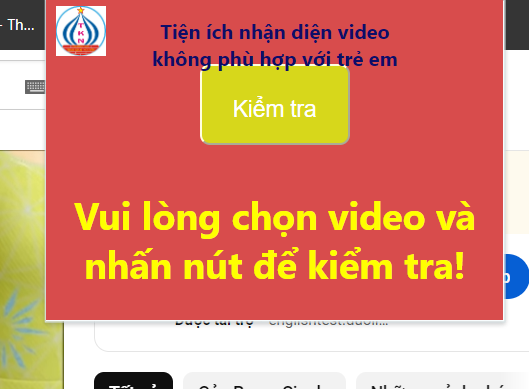
Sau đó bấm vào nút **Load unpacked**



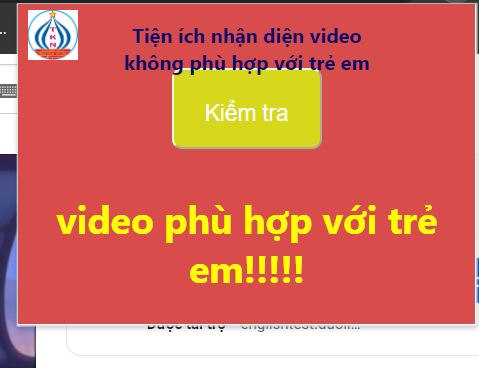
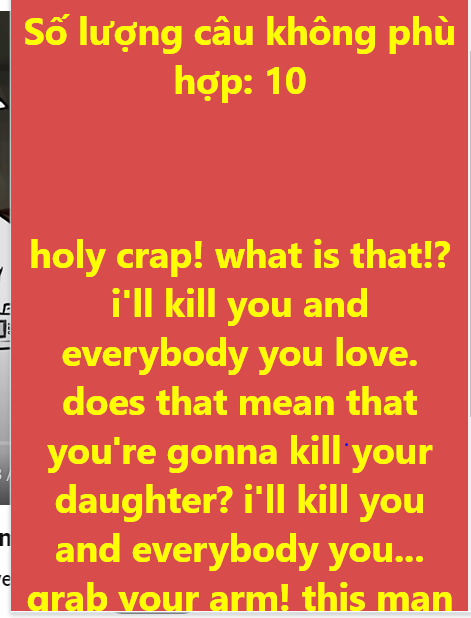
Chọn Folder **extensions** được đính kèm theo.



Bật tiện ích lên nhấp vào một video trên Youtube, sau đó nhấn vào video trên Youtube để kiểm tra cụ thể video đó và chờ kết quả sẽ được trả về sau khoảng 1 phút (nhóm sẽ cố gắng cải thiện thời gian phản hồi trong tương lai).



Các kết quả trả về sẽ bao gồm

* 
* 
  + Trong trường hợp kết quả là không phù hợp thì hệ thống sẽ trả về số lượng đoạn thoại bị nghi là không phù hợp với trẻ em, kèm với những đoạn thoại ấy.
  + 
* 

Đối với trường hợp này nguyên nhân là do nhóm chưa hoàn thành việc hỗ trợ ngôn ngữ tiếng Việt cho tiện ích, tạm thời chỉ hỗ trợ tiếng Anh.

**6. Tự đánh giá về những mặt còn tồn tại chưa giải quyết được của sản phẩm để khắc phục.**

Sản phẩm hiện tại vẫn chưa giải quyết được một số vấn đề triệt để trong một số trường hợp như video không có âm thanh nhưng vẫn mang hình ảnh bạo lực, không phù hợp với trẻ em. Hoặc là với những đứa trẻ thông minh, nhanh nhẹn trong việc sử dụng các thiết bị điện tử thì có thể tắt ứng dụng một cách dễ dàng nên vẫn chưa đảm bảo sự an toàn tuyệt đối. Hiện tại nhóm chỉ mới nghiên cứu và phát triển trên nền tảng Google Extension nên chỉ áp dụng được trên máy tính, chưa ứng dụng được trên các nền tảng khác điển hình là di động. Và một vấn đề cốt lõi nữa là những kẻ xấu với ý đồ cố tình đăng tải nội dung độc hại lên nền tảng Youtube vẫn có thể sử dụng nhiều chiêu trò để qua mặt thuật toán do hiện tại nhóm vẫn chỉ có một mô hình AI được train lại từ một mô hình có sẵn với dữ liệu (dataset) hạn chế nên vẫn rất dễ dàng bị qua mặt bởi nhiều chiêu trò.

**KẾT LUẬN**

**1. Hướng phát triển của sản phẩm trong tương lai**

Trong tương lai nhóm sẽ tiếp tục củng cố model Ai hiện có bằng cách thu thập thêm nhiều dữ liệu (dataset) từ nhiều nguồn khác nhau để cải thiện độ chính xác của sản phẩm. Tiếp tục nghiên cứu phát triển trên đa nền tảng như di động. Tìm hiểu và tiếp tục ứng dụng các model để bổ sung thêm nhiều tiêu chí cho mục đích nhận diện. Như đã đề cập do nhiều vấn đề chủ yếu là do thiếu kinh nghiệm trong lĩnh vực AI nhóm chỉ mới có được một tiêu chí đó là các lời đối thoại mang nội dung tiêu cực như: Xúc phạm, lăng mạ, sỉ nhục, đe dọa,,, Nhóm có ý định bổ sung thêm nhiều tiêu chí khác nữa giống như: Nội dung liên quan đến chính trị không phù hợp với trẻ em, các từ ngữ chỉ phù hợp cho một số độ tuổi nhất định và các từ ngữ không mang tính tiêu cực nhưng tục tĩu, mất văn hóa…

**2. Nguyện vọng trong tương lai**

Trong tương lai khi đã có đủ nguồn lực tài chính và kinh nghiệm. Nhóm mong muốn tự xây dựng một cấu trúc AI hoàn chỉnh chuyên về việc xử lý ngôn ngữ chủ yếu là việc nhận dạng nội dung độc hại thông qua các lớp riêng, để đảm bảo sự bảo vệ tối đa. Cũng như là phát triển cấu trúc thị giác máy tính (Computer-vision) để nhận diện được những video độc hại qua các hình ảnh không phù hợp. Cũng như là mở rộng sản phẩm trên nhiều nền tảng khác nhau và cũng như là các hệ điều hành khác nhau. Và cuối cùng nhóm muốn phát triển hệ thống mang tính bao quát trên internet không chỉ riêng Youtube mà còn nhiều nền tảng xã hội khác như: Facebook, TikTok, Instagram, Telegram… và cũng như là Internet nói chung. Nhóm muốn đóng góp để tạo nên một môi trường an toàn lành mạnh để từ đó các thế hệ trẻ được tiếp xúc với công nghệ từ sớm có thể tự do học tập, khám phá mà không phải lo sợ về sự nguy hiểm của Internet.

**TÀI LIỆU THAM KHẢO**

1. Mã nguồn
   1. <https://github.com/AryenSinghal/YouTube-Video-Summariser> (Tham khảo cách sử dụng nền tảng Google Extension)
   2. <https://huggingface.co/>
   3. <https://huggingface.co/datasets>
   4. <https://colab.research.google.com/github/huggingface/notebooks/blob/main/transformers_doc/en/tensorflow/sequence_classification.ipynb#scrollTo=aobRTqXZLVWu>
   5. https://huggingface.co/s-nlp/roberta\_toxicity\_classifier
2. Tài liệu tham khảo
   1. <https://huggingface.co/docs/transformers/tasks/sequence_classification>
   2. <https://developer.chrome.com/docs/extensions/get-started/tutorial/hello-world?hl=vi>
   3. <https://www.freecodecamp.org/news/javascript-post-request-how-to-send-an-http-post-request-in-js/>
   4. https://www.researchgate.net/publication/363133906\_Samba\_Identifying\_Inappropriate\_Videos\_for\_Young\_Children\_on\_YouTube
3. Model AI được sử dụng để fine-tune
   1. <https://huggingface.co/michellejieli/NSFW_text_classifier?not-for-all-audiences=true>
   2. https://huggingface.co/s-nlp/roberta\_toxicity\_classifier

*……, ngày 25 tháng 4 năm 2024*

**Xác nhận của nhà trường Chữ ký của tác giả/nhóm tác giả**