Desenvolupament d'un Sistema Automàtic de Classificació d'Edat per a Continguts Audiovisuals

Guillem Cadevall Ferreres

30 de juny de 2025

Resum– Aquest treball de final de grau presenta un sistema automatitzat per classificar continguts audiovisuals segons franges d'edat. El sistema utilitza un pipeline multimodal amb tècniques de visió per computador, processament de llenguatge natural i grans models de llenguatge (LLM). Detecta continguts sensibles com violència, nuesa, drogues i llenguatge ofensiu, i aplica la classificació d'edat basada en la normativa de Malàisia. Inclou transcripció d'àudio, traducció, classificació textual, detecció d'objectes i contingut NSFW, així com una API backend i una interfície frontend intuïtiva. Els resultats mostren una alta fiabilitat en la classificació, especialment per continguts per a tots els públics i per a adults. El projecte ofereix una eina innovadora per adaptar la distribució de continguts a contextos culturals diversos.

Paraules clau– Classificació per edats, Anàlisi multimodal, Moderació de continguts, Grans models de llenguatge (LLM), Visió per computador, Processament del llenguatge natural (PLN), Aprenentatge profund, Adaptació cultural.

Abstract– This project presents an automated system for age-based classification of audiovisual content. It uses a multimodal pipeline combining computer vision, natural language processing, and large language models (LLMs). The system detects sensitive content—such as violence, nudity, drugs, and offensive language—and assigns an age rating based on Malaysian standards. Key features include audio transcription, text translation and classification, object detection, and NSFW filtering. A backend API and user interface were developed for deployment. Evaluation results show high reliability in age classification, particularly for young and adult audiences. This tool supports safer, culturally adaptable digital content distribution.

Keywords– Age Classification, Multimodal Analysis, Content Moderation, Large Language Models (LLMs), Computer Vision, Natural Language Processing (NLP), Deep Learning, Cultural Adaptation.

1 Introducció - Context del treball

El cinema i les sèries de televisió han estat sempre una forma d'art i una font d'entreteniment global que transcendeix fronteres. Amb el creixement de les plataformes de transmissió en línia, la producció i distribució de contingut audiovisual ha arribat a una audiència mundial de manera massiva. No obstant això, la diversitat cultural entre els diferents continents genera una gran varietat de percepcions pel que fa al contingut adequat per a cada grup d'edat. A l'Occident, tot i que existeixen categories per classificar el contingut segons l'edat (com ara les classificacions PG, 13+, R, etc.), aquestes no sempre són estrictes. En canvi, en altres cultures, com les d'Orient, les normatives de classificació de continguts són molt més rigoroses, especialment pel que fa a la violència, el contingut sexual o el llenguatge ofensiu. Aquestes diferències culturals poden dificultar que els usuaris gaudeixin del mateix contingut audiovisual de manera segura i d'acord amb les seves normatives socials i ètiques. [1]

A més, l'exposició a continguts explícits, especialment per part de menors, pot tenir conseqüències negatives en el desenvolupament emocional, social i cognitiu. Diversos estudis han demostrat que el consum prematur de contingut sexual o violent pot augmentar el risc de conductes problemàtiques i afectar la percepció de les relacions socials i de la realitat [2, 3]. Això reforça la necessitat de sistemes

• E-mail de contacte: guillecade22@gmail.com
• Menció realitzada: Computació
• Treball tutoritzat per:
Xim Cerdá Company - CVC
Francesc Tarrés Ruiz - Ugiat Technologies
• Curs 2024/25

que permetin controlar i adaptar el contingut d'acord amb l'edat i el context cultural de l'audiència.

Aquest projecte té com a objectiu el desenvolupament d'un sistema automatitzat de classificació de continguts audiovisuals, com pel·lícules i sèries, capaç d'identificar escenes concretes i assignar-les a diferents categories d'edat. La finalitat d'aquesta eina és proporcionar una solució que permeti adaptar el contingut d'acord amb les preferències culturals de l'audiència. Així, per exemple, cultures més restrictives podrien continuar gaudint de pel·lícules i sèries occidentals, però amb les parts més violentes o sexuals adaptades o eliminades, per complir amb les seves estrictes normatives culturals.

L'eina proposada no només serviria per adaptar pel·lícules i sèries a les necessitats d'audiències diverses, sinó que també podria utilitzar-se per monitoritzar i gestionar el contingut compartit a les xarxes socials. D'aquesta manera, es podria evitar la difusió de continguts inapropiats, contribuint així a un entorn més segur i controlat en línia.

En resum, l'objectiu principal d'aquest Treball Final de Grau és desenvolupar un sistema de classificació multimèdia automatitzat capaç de classificar vídeos, imatges, àudios i textos segons les seves característiques específiques. Aquest sistema utilitzarà tècniques d'aprenentatge automàtic per identificar contingut sensible en funció de diverses categories, com violència, gore, contingut sexual, drogues, tabac, alcohol, llenguatge ofensiu, entre altres. A més, el sistema serà capaç d'assignar una edat adequada a cada tipus de contingut, fent-lo compatible amb diverses normatives culturals. Aquesta eina no només busca millorar l'adaptació de les pel·lícules i sèries a diferents públics, sinó també ajudar a la protecció dels usuaris més vulnerables de l'exposició a continguts inapropiats.

2 ESTAT DE L'ART

En els darrers anys, la classificació automatitzada de continguts multimèdia ha esdevingut un camp d'estudi rellevant dins l'àmbit de la intel·ligència artificial i la visió per computador, especialment degut a l'augment exponencial de contingut audiovisual generat i consumit en plataformes digitals. Diversos enfocaments han estat proposats i implementats per abordar la detecció automàtica de continguts sensibles com la violència, el contingut sexual, el llenguatge ofensiu o l'ús de substàncies, tant en contextos comercials com acadèmics.

Les tècniques d'aprenentatge profund (deep learning) han estat especialment efectives per a l'anàlisi de vídeo i imatge. Xarxes neuronals convolucionals (CNN) s'utilitzen àmpliament per detectar objectes i escenes específiques [4], mentre que les xarxes neuronals recurrents (RNN) i, més recentment, els models basats en Transformers, s'han aplicat amb èxit en la detecció de patrons temporals en seqüències de vídeo [5]. A més, per a la classificació d'àudio, s'han emprat xarxes neuronals capaces d'identificar sons relacionats amb la violència, els insults verbals o altres senyals audibles sensibles.

Pel que fa a les iniciatives comercials, plataformes com YouTube, Instagram o TikTok ja utilitzen sistemes automatitzats per a la detecció de contingut inapropiat, basant-se en models entrenats amb grans volums de dades etiquetades. Tot i això, la majoria d'aquests sistemes són propietaris i no estan dissenyats per adaptar-se a les especificitats culturals de cada mercat.

En l'àmbit acadèmic, s'han publicat diversos estudis que analitzen la viabilitat de sistemes híbrids que combinen visió per computador, processament de llenguatge natural (NLP) i anàlisi d'àudio per aconseguir una classificació multimodal del contingut. Per exemple, projectes com NSFWNet[6] i OpenNSFW[7], desenvolupats per Yahoo, se centren en la detecció de contingut explícit mitjançant CNNs entrenades amb dades etiquetades com a "aptes" o "no aptes".

Relacionat amb el món del cinema, diversos estudis recents han explorat la classificació de contingut mitjançant tècniques d'aprenentatge profund. Per exemple, s'han proposat models basats en xarxes neuronals recurrents (RNN) per predir la classificació MPAA [8] a partir dels diàlegs de les pel·lícules[9]. També s'ha desenvolupat un enfocament multimodal per avaluar tràilers de pel·lícules, utilitzant informació conjunta de vídeo i àudio [10]. En la mateixa línia, s'ha dissenyat un sistema capaç d'identificar automàticament segments explícits dins d'un vídeo i generarne un resum textual [11]. Aquests estudis mostren el potencial dels enfocaments multimodals per a una classificació de continguts explícits més precisa.

També hi ha iniciatives privades en l'àmbit del cloud computing que ofereixen serveis de classificació automàtica de contingut multimèdia, com ara Azure Content Moderator[12] de Microsoft o Google Cloud Vision API[13]. Aquests serveis proporcionen eines per detectar contingut explícit, com ara violència (gore), nuesa o contingut sexual, així com text ofensiu en imatges i vídeos. No obstant això, aquestes solucions presenten limitacions importants pel que fa al nostre objectiu: en general, es basen en classificacions binàries (apte/no apte), sense graduació per edats, i no tenen en compte factors culturals o contextos específics. A més, el seu cost elevat les fa poc accessibles per a projectes que necessiten una personalització avançada o un desplegament a gran escala.

Malgrat els avenços en classificació automatitzada de contingut, actualment no existeix cap eina que ofereixi una classificació per franges d'edat. Les solucions existents es limiten a categories binàries o detecció de contingut explícit, sense considerar els contextos culturals ni proporcionar la flexibilitat necessària per a una personalització avançada segons les necessitats de diferents comunitats i grups d'edat.

3 OBJECTIUS

L'objectiu principal d'aquest Treball Final de Grau és desenvolupar un sistema automatitzat de classificació de continguts audiovisuals capaç d'analitzar de manera integrada múltiples modalitats d'informació. Per assolir aquesta finalitat, s'ha construït un pipeline de processament multimodal que pugui gestionar de forma simultània vídeo, àudio, imatges i text, extraient informació rellevant de ca-

dascuna d'aquestes fonts per generar una classificació precisa i contextualitzada del contingut.

El sistema de classificació desenvolupat ha d'estar específicament dissenyat per al seu desplegament en el mercat malasi, adaptant-se per tant al sistema oficial de classificació de continguts audiovisuals d'aquest país. Aquest marc regulatori estableix cinc categories diferenciades que el sistema ha de ser capaç d'identificar i assignar correctament: **U** (*Umum*), adequat per a totes les edats; **P12** (*Penjaga*), que requereix supervisió parental per a audiències menors de 12 anys; **13**, destinat a audiències de 13 anys en endavant; **16**, per a audiències de 16 anys i superiors; i finalment **18**, restringit a audiències adultes. (vegeu Apèndix A.1 per a detalls de l'empresa, els clients i les directrius de classificació proporcionades.)

Des d'una perspectiva tècnica, un dels principals reptes del projecte és millorar la qualitat de les respostes generades pels grans models de llenguatge (LLM), de manera que la classificació sigui el més precisa, coherent i alineada possible amb les directrius establertes. També és important que els temps de processament es mantinguin dins de límits acceptables per a un ús pràctic, mai comprometent però la qualitat mínima requerida per a una classificació fiable.

Paral·lelament als aspectes de processament, el projecte contempla el desenvolupament d'una arquitectura de desplegament professional mitjançant una API robusta i escalable. Complementàriament, s'ha de desenvolupar un servei web amb una interfície d'usuari intuïtiva que faciliti les proves del sistema de manera més accessible i amb un caràcter més comercial.

Per acabar, a nivell social i comercial, aquest projecte aspira a oferir una solució tècnica que beneficiï tant els productors de contingut com les plataformes de distribució, facilitant l'adaptació cultural del material audiovisual i contribuint a un ecosistema digital més segur i inclusiu.

4 METODOLOGIA

Per al desenvolupament d'aquest projecte, s'ha adoptat una metodologia basada en l'experimentació i l'avaluació iterativa de models i tècniques d'intel·ligència artificial. Aquesta metodologia combina la selecció i ús de models multimodals, la generació i preparació de dades específiques, els processos d'entrenament i optimització de models, donant sempre especial atenció al desplegament local i l'eficiència computacional. En particular, s'ha treballat modificant progressivament l'entorn i els mòduls que interactuen amb el LLM, per tal d'obtenir resultats més precisos i facilitar així un entrenament posterior més senzill i efectiu. Aquesta aproximació ha estat necessària donada la baixa precisió dels resultats inicials obtinguts directament del model original. (vegeu Apèndix A.2, on es detalla la planificació original del projecte mitjançant el diagrama de Gantt, així com les variacions en l'enfocament del treball a causa de diverses circumstàncies.)

4.1 Arquitectura del sistema

Per estructurar de manera clara i escalable el flux de processament del sistema, s'ha optat per utilitzar LangGraph[14] en combinació amb LangChain[15], dues eines especialment dissenyades per construir aplicacions amb grans models de llenguatge (LLM) d'una manera modular i robusta. LangGraph[14] permet definir el comportament del sistema com un graf d'estats on cada node encapsula una funcionalitat concreta, mentre que LangChain[15] aporta una arquitectura basada en components reutilitzables, que facilita la integració de memòria, eines externes i formats estructurats de resposta. Aquesta combinació no només millora la mantenibilitat i claredat del codi, sinó que també proporciona un entorn idoni per gestionar lògiques complexes i condicions dinàmiques en funció de l'estat del sistema.

També cal destacar que un dels grans avantatges de treballar amb grafs d'estats és que permeten detectar fàcilment els processos que es poden paral·lelitzar, fet que permet una optimització significativa del rendiment i reducció del temps d'execució.

Per tal de garantir la fiabilitat i consistència de les sortides generades pels LLM, s'ha fet ús del sistema de parser output de LangChain[15], que transforma les respostes dels models en instàncies validades de classes Pydantic[16], equivalents a diccionaris de Python o objectes JSON. Això assegura que tota la informació segueixi una estructura clara.

La figura 1 mostra l'esquema dels nodes que configuren el graf del sistema. A continuació es presenten els diferents mòduls que formen el sistema desenvolupat.

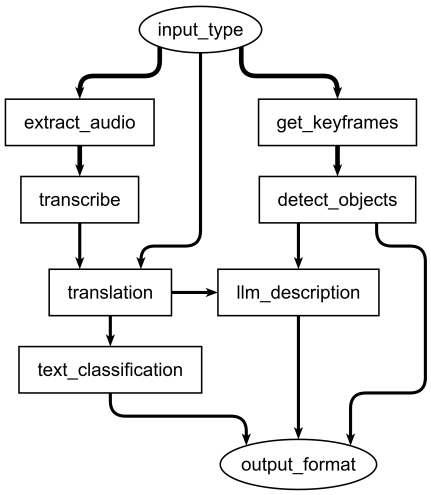

Fig. 1: Graf del sistema

4.2 Mòduls del sistema

4.2.1 Input Type

El primer mòdul del sistema actua com a punt d'entrada del graf i s'encarrega de determinar el tipus d'entrada proporcionada per l'usuari. A partir d'aquesta classificació, cada tipus d'entrada segueix un camí específic dins del sistema, activant únicament els mòduls rellevants per al seu processament.

- **Text**: Es llegeix el fitxer, es detecta l'idioma i es divi-

deix el text en frases per preparar-lo per als següents processos. A continuació, es redirigeix al mòdul de traducció.

- **Àudio**: El fitxer es redirigeix al mòdul de transcripció per convertir-lo a text.
- **Imatge**: L'entrada es processa mitjançant el mòdul de detecció d'objectes.
- **Vídeo**: El sistema l'envia simultàniament als mòduls d'extracció d'àudio i de selecció de fotogrames clau (keyframes).

4.2.2 Extract Audio

Extreu la pista d'àudio dels fitxers de vídeo. Utilitza l'eina FFmpeg[17] per convertir el vídeo d'entrada en un fitxer d'àudio en format WAV, amb una qualitat i format específics (mono, 16 kHz). Aquest fitxer d'àudio resultant es guarda de manera temporal en la mateixa carpeta per a ser processat en etapes posteriors.

4.2.3 Transcribe

Realitza la transcripció automàtica de l'àudio previament extret. Per dur a terme aquesta tasca, s'utilitza el model Faster Whisper[18] en la versió large-v3-turbo amb precisió de 16 bits flotants (bfloat16). La transcripció es realitza configurant el model perquè consideri fins a cinc alternatives simultànies (beam_size=5), millorant així la precisió dels resultats. A més, s'aplica un filtratge de detecció de veu (VAD, Voice Activity Detection[19]) que elimina silencis i sorolls no desitjats, establint com a paràmetre mínim un silenci de 500 mil·lisegons per considerar un segment com a part activa de la veu. Durant aquest procés, el sistema també detecta automàticament l'idioma de l'àudio.

4.2.4 Translation

Tradueix el text transcrit a l'anglès sempre que l'idioma detectat en l'àudio no sigui ja aquesta llengua. Per fer-ho, usa el model preentrenat M2M100 de Facebook[20], en la seva versió "1.2B", un model multilingüe capaç de traduir entre diversos idiomes. La traducció a l'anglès és clau per assegurar la compatibilitat i el rendiment òptim dels models utilitzats en fases posteriors del sistema. En primer lloc, el model de classificació de textos, només està entrenat per treballar amb textos en anglès. També es proporcionen diàlegs al LLM per ajudar-lo a comprendre millor el context del contingut multimèdia, obtenint els millors resultats quan aquests diàlegs estan en anglès, ja que és la llengua principal amb què ha estat entrenat i en la qual processa la informació amb més precisió i eficàcia[21].

4.2.5 Text Classification

Valora qualitativament el contingut textual extret de l'àudio d'un vídeo. Aquesta valoració es fa assignant a cada fragment una etiqueta de classificació que indica el nivell de sensibilitat o intensitat del text. En aquest cas, es passa cada fragment textual (diàleg) a través d'un pipeline de classificació basat en transformers[22]. La tasca de classificació s'ha dut a terme amb cinc categories, etiquetades de menor a major intensitat:

- **VL (Very Low)**: Contingut molt suau i neutral, sense cap element problemàtic ni emocionalment carregat.
- **L (Low)**: Contingut amb alguna expressió lleu, informal o ambigüa, però que no es considera problemàtica.
- **M (Medium)**: Contingut amb una càrrega emocional moderada, potencialment incòmode o amb temes polèmics.
- **H (High)**: Contingut altament explícit, amb llenguatge dur o potencialment ofensiu.
- **E (Extreme)**: Contingut extremament explícit, violent, ofensiu o altament sensible.

Per tal d'adaptar el model a la realitat dels diàlegs audiovisuals, s'ha realitzat un entrenament supervisat mitjançant un corpus recollit via web scraping de la web Genius.com[24], que conté guions de pel·lícules i sèries populars. Així s'obtenen textos reals i representatius d'una gran varietat de gèneres, tons i contextos, des de comèdies infantils fins a thrillers o films per a adults.

Aquestes dades han estat prèviament anotades amb etiquetes personalitzades (VL, L, M, H, E), i preperades per a l'entrenament mitjançant una pipeline que inclou tokenització, codificació de les etiquetes i divisió en conjunts de validació i entrenament.

El model base seleccionat ha estat distilbert-base-uncased[23], una versió lleugera i optimitzada de BERT[25] que redueix significativament el temps d'entrenament i inferència sense sacrificar gaire la precisió. Aquest model és especialment adequat per entorns on l'eficiència és un factor clau.

Durant el desenvolupament del classificador de text també s'ha provar una estratègia alternativa per generar més dades d'entrenament: la generació de dades sintètiques amb un LLM. Tot i que això permet ampliar ràpidament el conjunt d'entrenament amb més exemples, la qualitat de les mostres generades es insuficient. El contingut generat sovint presenta limitacions notables:

- **Alta repetició de patrons lingüístics**: Dificultat per generalitzar el model a causa de la presència reiterada d'estructures similars.
- **Limitacions per generar diàlegs incòmodes**: Els LLM tenen tendència a evitar continguts violents, ofensius o sensibles, cosa que en limita la versatilitat.
- **Manca de naturalitat**: Els diàlegs generats no semblen reals, especialment en contextos informals o col·loquials.

4.2.6 Get Keyframes

Té com a objectiu extreure els frames clau més significatius d'un vídeo, és a dir, aquells moments visuals que representen millor el contingut i que seran utilitzats per models de processament multimodal. Aquest procés, que es detalla a la figura 2, està basat en diverses tècniques avançades de visió per computador i aprenentatge automàtic, integrant detecció de canvis de pla, extracció de característiques i clustering per obtenir una selecció òptima i representativa de frames[26].

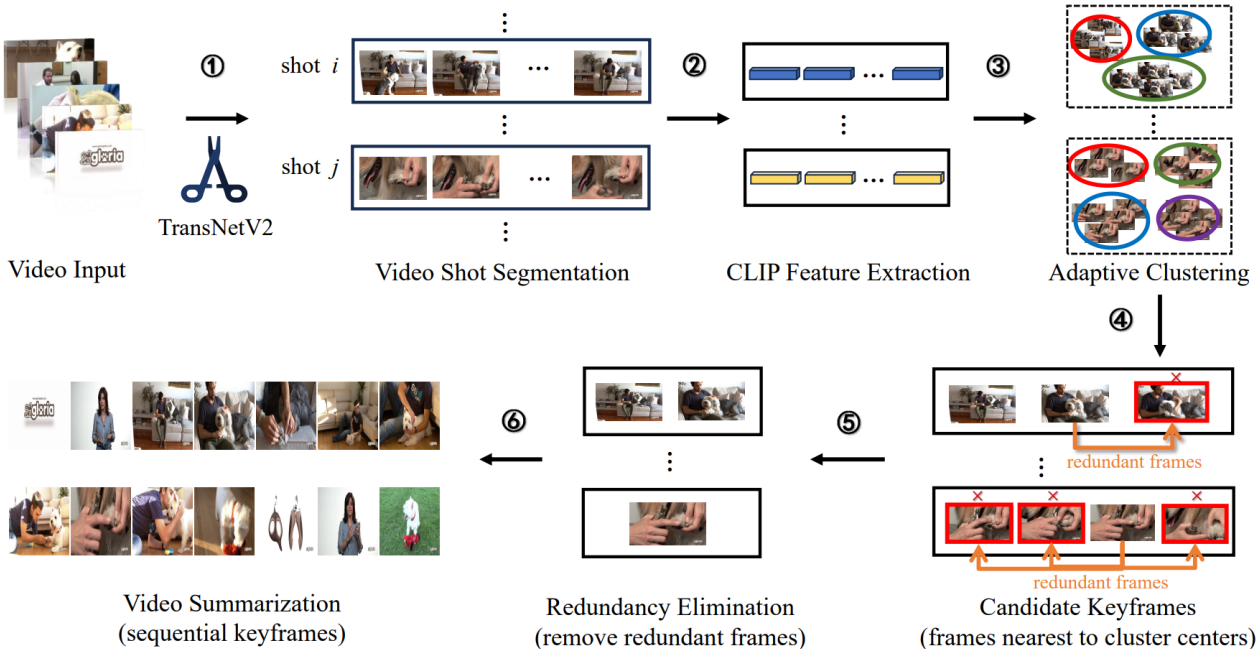

Fig. 2: Passos per extreure els frames clau del vídeo.

El procediment comença amb la preparació de l'entorn, on s'eliminen les dades antigues emmagatzemades a les carpetes de treball i es creen de nou buides. En cas que l'entrada sigui una imatge en lloc d'un vídeo, aquesta es desa directament com a frame clau i es retorna sense necessitat de processar més.

Per al cas de vídeos, el primer pas és dividir-los en fragments basats en els canvis de pla. Aquesta segmentació es fa mitjançant el model TransNetV2[27], especialitzat en la detecció precisa de transicions entre escenes. D'aquesta manera, el vídeo queda dividit en seqüències visuals contínues i homogènies.

Un cop segmentat el vídeo, es captura un frame per segon i s'aplica el model CLIP[28] per extreure vectors de característiques de 768 dimensions. Aquests vectors codifiquen informació visual de cada imatge, facilitant un processament posterior més eficient.

A continuació, dins de cada fragment detectat, s'aplica un algorisme de clustering K-means sobre els vectors de característiques dels frames. Es prova un nombre de centroides entre 2 i 10 i es selecciona el que ofereix millor resultat segons la mètrica silhouette score, que avalua la cohesió i separació dels grups. Per cada cluster, es tria el frame més proper al seu centroide com a representant visual del grup.

Després, es realitza un filtratge per eliminar frames que aportin poca informació útil. Aquesta decisió es basa en mesures d'entropia i variança, que permeten descartar imatges gairebé uniformes o amb poca complexitat, com ara plans en blanc o negre, o transicions difuminades.

Finalment, es controla la redundància entre els frames seleccionats analitzant la similitud entre imatges. Per evitar eliminar frames que, tot i ser visualment semblants, estan separats en el temps i podrien representar contextos o diàlegs diferents, s'incorpora un paràmetre que penalitza la distància temporal, assegurant que aquests frames es mantinguin a la selecció final.

4.2.7 Detect Objects

Identifica objectes específics i continguts sensibles dins dels frames clau extrets prèviament del vídeo. Aquesta detecció és fonamental per contextualitzar les escenes i enriquir la informació que s'utilitzarà en l'anàlisi multimodal posterior. Per aconseguir-ho, es combinen diversos models especialitzats, alguns amb aprenentatge supervisat i finetuning, i altres amb classificació de contingut explícit.

Per a la detecció d'objectes com armes de foc, drogues, tabac, alcohol i violència, es fa servir un model anomenat RFDETRBase[29]. Aquest model és una variant de la família de detectors basats en transformers, inspirada en l'arquitectura DETR (DEtection TRansformer)[30], que integra un model de detecció d'objectes amb la capacitat d'aprendre directament les relacions espacials i semàntiques entre elements visuals en una imatge.

RFDETRBase[29] combina un CNN per a l'extracció inicial de característiques visuals amb un transformador per modelar dependències globals, millorant la precisió i robustesa en la detecció d'objectes.

Per adaptar RFDETRBase[29] a la detecció dels objectes específics que interessen en aquest projecte, s'ha realitzat un procés de finetuning. Aquest procés consisteix a prendre un model preentrenat general (amb pesos inicials entrenats en conjunts de dades amplis com COCO[31]) i continuar l'entrenament amb un conjunt específic de dades etiquetades corresponents a aquestes categories.

Per al finetuning del model s'han utilitzat diversos datasets especialitzats, cadascun format per una recopilació d'imatges etiquetades específicament per a diferents categories: armes de foc (pistoles i rifles), drogues i tabac, alcohol, i situacions violentes. Aquestes imatges s'han obtingut principalment de la plataforma Roboflow Universe[32],

que proporciona conjunts de dades àmpliament anotats i representatius. Mitjançant un entrenament supervisat, s'han ajustat els pesos del model preentrenat, preservant el seu coneixement general mentre s'adapta específicament per detectar amb precisió aquests objectes i escenes d'interès. Durant la fase d'inferència, s'aplica un llindar de confiança del 50%, considerant només les deteccions amb una probabilitat superior a aquest valor per minimitzar falsos positius.

Per detectar contingut sexual explícit i gore, que no és un problema estrictament de detecció d'objectes però sí de classificació d'imatges, s'utilitza un model preentrenat en classificació d'imatges anomenat Falconsai/nsfw_image_detection[33]. Aquest és accessible via la llibreria transformers[22] de Hugging Face[34] mitjançant la pipeline image-classification.

Aquest model retorna per a cada imatge una predicció amb la probabilitat que contingui material NSFW (not safe for work) o material NORMAL. Es defineix un llindar intern, considerant que si la probabilitat de NSFW és superior a 0.3 i major que la de la classe NORMAL, es marca aquesta imatge com a sensible.

4.2.8 LLM Description

Genera una descripció semàntica i una classificació per edats de cadascun dels keyframes d'un vídeo, tenint en compte tant el contingut visual de la imatge com els diàlegs associats al moment temporal corresponent. Aquest procés permet obtenir una avaluació contextualitzada de l'escena, així com una estimació de la classificació per edats.

Per dur a terme aquesta tasca, el sistema utilitza un LLM de Google, concretament el model Gemma-3-27b-it [35], a través de la biblioteca langchain_google_genai [15]. S'han avaluat diversos models de Qwen2.5-VL[36] i Gemma-3[35], amb diferents mides, però després de diverses proves, s'ha determinat que Gemma-3-27b-it[35] ofereix la millor relació entre qualitat dels resultats i mida del model. El funcionament del mòdul s'estructura de la següent manera:

- **Selecció del text:** Segons l'idioma detectat en l'etapa anterior, s'utilitza el text original transcrit o la seva traducció a l'anglès.
- **Iteració sobre fotogrames clau:** Per a cada fotograma clau es busca el conjunt de diàlegs que coincideixen amb el seu segment temporal, basant-se en les marques de temps associades (text_timecodes i key_frames_timecodes). Si no hi ha diàlegs associats, s'indica explícitament que la imatge no conté diàleg.
- **Integració de dades de detecció d'objectes:** A més del contingut visual i textual, el sistema incorpora informació addicional provinent del mòdul de detecció d'objectes. Aquestes dades, que es passen com a flags, informen sobre la presència de certs elements visuals rellevants com ara armes, sang o substàncies com alcohol o drogues.
- **Preparació de la petició:** Es construeix una petició que inclou tant el text com la imatge codificada en base64. El text segueix un prompt estructurat que demana:
  - Una descripció concisa i objectiva del contingut visual i textual de l'escena, destacant elements rellevants per a la classificació per edat (violència, nuditat, drogues, emocions intenses, llenguatge inapropiat, etc.).
  - Una classificació d'edat adequada segons el contingut detectat, utilitzant les següents categories: **U:** Apte per a tots els públics; **P12:** Supervisió suggerida per a menors de 12 anys; **13:** Apte a partir de 13 anys; **16:** Apte a partir de 16 anys; **18:** Apte a partir de 18 anys.
- **Processament de la resposta:** La resposta generada pel model es processa mitjançant un PydanticOutputParser[16], que extreu la descripció i la classificació d'edat.

4.3 Desplegament

Per a l'implementació i desplegament del sistema, s'ha desenvolupat una arquitectura client-servidor completa que permet l'ús pràctic de les capacitats del classificador desenvolupat.

4.3.1 Backend

El backend s'ha implementat utilitzant FastAPI[37]. El sistema fa servir un patró de jobs per gestionar el processament de fitxers de manera eficient, on cada petició d'anàlisi genera un job únic identificat per un UUID, que permet al client comprovar l'estat sense bloquejar l'aplicació. L'arquitectura implementa un sistema de gestió d'estat dels jobs en memòria amb locks per garantir la seguretat en l'accés concurrent (thread-safety), mantenint informació sobre l'estat de cada procés (pending, processing, completed, failed), juntament amb timestamps i resultats.

Els endpoints principals inclouen /classify per iniciar processaments, /job/{job_id} per consultar l'estat d'un procés específic, i /jobs per llistar tots els processos. El processament real s'executa en threads independents per permetre que múltiples usuaris puguin enviar contingut simultàniament sense afectar la responsivitat de l'API.

4.3.2 Frontend

El frontend s'ha desenvolupat utilitzant React[38] amb una arquitectura basada en components reutilitzables. Els components principals inclouen: Upload per gestionar la càrrega de fitxers amb suport drag-and-drop i entrada de text manual, Results per mostrar els resultats amb controls interactius, i diversos components especialitzats per renderitzar diferents tipus de mitjans i presentar transcripcions amb navegació temporal.

La interfície ofereix funcionalitats avançades com la previsualització automàtica del contingut carregat, navegació temporal en vídeos i àudios mitjançant clic a les transcripcions, commutació entre idiomes originals i traduccions, canvi entre modes de transcripció i descripció segons el tipus de contingut, i exportació de resultats en format JSON. El sistema implementa una funció que consulta automàticament l'estat dels jobs cada 2 segons, amb un temps

màxim d'espera de 5 minuts, per evitar el processament de continguts massa llargs.

4.4 Recopilació de dades i entrenament de l'LLM

Inicialment, estava previst que el client proporcionés un conjunt de dades propi, consistent en pel·lícules, sèries o, com a mínim, clips audiovisuals rellevants per a l'entrenament dels models. Aquesta aportació havia de servir com a base principal per a la generació del conjunt de dades. No obstant això, aquesta previsió no s'ha complert i, davant la manca de dades proporcionades pel client, s'ha optat per una alternativa autònoma basada en la recopilació de clips audiovisuals per mitjans propis.

Després d'una exhaustiva cerca, s'ha identificat una plataforma web que ofereix una àmplia col·lecció de clips curts extrets de pel·lícules comercials[39]. Aquesta col·lecció inclou fragments amb una durada compresa entre 5 segons i 2 minuts, provinents de pel·lícules que abracen des de l'any 1915 fins a l'actualitat.

Per tal d'aprofitar aquest recurs, s'ha implementat un sistema de web scraping que permet descarregar automàticament els clips disponibles, juntament amb la informació contextual associada (com ara el títol, any, actors i, si escau, alguna descripció o etiquetes proporcionades pel lloc web).

Un cop generat un volum suficient de dades a partir dels clips audiovisuals obtinguts mitjançant web scraping, s'ha procedit a l'entrenament del model amb l'objectiu de millorar l'eficiència del sistema. Concretament, s'ha dut a terme un procés de fine-tuning al model Gemma-4B-it[35].

L'entrenament s'ha realitzat mitjançant notebooks Jupyter, emprant la infraestructura proporcionada pel projecte Unsloth[40], una eina optimitzada per dur a terme fine-tuning lleuger i eficient de models de grans dimensions.

5 RESULTATS

L'avaluació del sistema desenvolupat s'ha basat principalment en la utilització de conjunts de dades específics i variats, amb l'objectiu de quantificar de manera rigorosa la qualitat i el comportament del model en diferents situacions. Paral·lelament, també s'ha dut a terme una anàlisi dels diferents components que integren el pipeline multimodal.

5.1 Classificador de Text

Les figures 3 i 4 mostren l'evolució de la funció de pèrdua durant el procés d'entrenament del classificador de text al llarg d'uns 5.000 passos d'entrenament.

Ambdues corbes evidencien una convergència adequada del model, amb una disminució pronunciada de la pèrdua durant els primers 1.000 passos, seguida d'una estabilització progressiva. La pèrdua d'avaluació mostra una tendència decreixent des d'aproximadament 0,095 fins a valors propers a 0,030, mentre que la pèrdua d'entrenament descendeix des de valors superiors a 0,9 fins a estabilitzar-se al voltant de 0,020. Aquesta evolució indica que el model

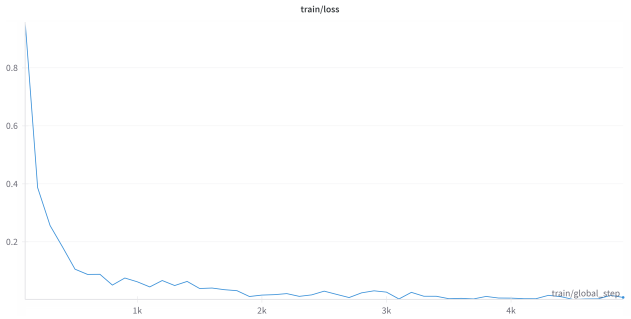

Fig. 3: Gràfica de la pèrdua durant l'entrenament (train loss)

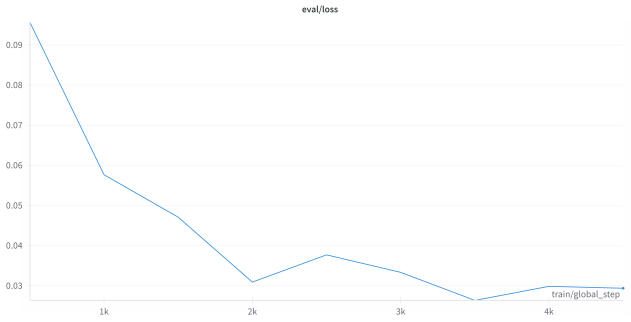

Fig. 4: Gràfica de la pèrdua durant l'avaluació (eval loss)

ha après de manera efectiva sense mostrar signes evidents de sobreajustament.

5.2 Classificador d'Imatges

La figura 5 mostra l'evolució de la funció de pèrdua durant el procés d'entrenament del classificador d'imatges al llarg d'uns 5.000 passos.

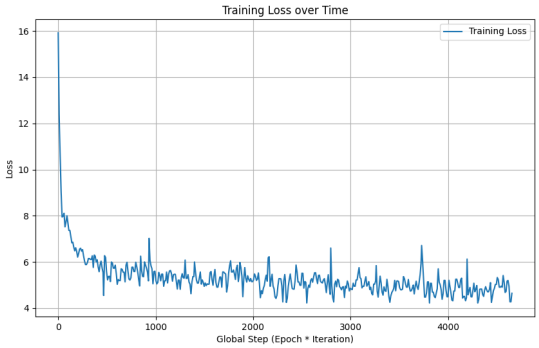

Fig. 5: Gràfica de la pèrdua durant l'entrenament del classificador d'imatges (train loss)

Després d'un descens inicial pronunciat de la pèrdua, de 16,0 a 8,0, la millora es redueix considerablement i la corba s'estabilitza entre 5,5 i 4,5. Tot i això, l'entrenament continua per intentar millorar el rendiment en classes específiques amb pitjors resultats, ja que el valor global de la pèrdua no reflecteix necessàriament aquests ajustos. Per exemple, certes classes com rifle o weapon han obtingut puntuacions altes i s'han classificat correctament de manera consistent, mentre que altres com missile o knife han generat molts falsos positius. Com a conseqüència, aquestes dar-

reres s'han descartat en la versió final del sistema per tal de millorar la fiabilitat global del model i evitar errors crítics en l'etapa de predicció.

5.3 LLM

5.3.1 Entrenament del model

L'entrenament s'ha realitzat utilitzant el model Gemma-3-4B-it[35], ja que era el model de major mida que podien gestionar les targetes gràfiques disponibles. Tot i així, aquest model presenta una capacitat limitada per capturar patrons complexos, i la manca d'un conjunt de dades suficient i de qualitat ha dificultat encara més l'obtenció de resultats satisfactoris.

La quantitat de dades recollides mitjançant tècniques de web scraping ha resultat insuficient per entrenar amb èxit el model. A més, molts dels clips presentaven una qualitat visual molt baixa, fet que en compromet la utilitat per a l'extracció d'informació significativa.

L'únic efecte observable de l'entrenament va ser la capacitat del model de generar sortides amb un format consistent, seguint una estructura similar a una definició de classe, tal com s'havia esperat. Tanmateix, la qualitat de les respostes, en termes de contingut i precisió, es manté molt per sota del que ofereix el model original.

Per aquest motiu, en el producte final i en totes les avaluacions posteriors, s'ha utilitzat el model Gemma-3-27B-it[35] desenvolupat per Google, sense cap fase d'entrenament addicional, ja que aquest model ofereix resultats significativament més robustos i útils per a la tasca plantejada.

5.3.2 Avaluació del classificador

S'han implementat diferents mètodes per valorar de forma quantitativa la qualitat del classificador desenvolupat. Inicialment, es va cercar un dataset que contingués clips de pel·lícules amb edats assignades, però actualment no existeix cap conjunt de dades d'aquestes característiques. Per aquesta raó, s'han recopilat diversos datasets de vídeos i imatges per avaluar el comportament del classificador. La majoria d'aquests datasets són binaris, de manera que el contingut que ells classifiquen com a segur, el nostre classificador hauria d'assignar-lo com a categoria U (tots els públics), i els continguts no segurs haurien de ser assignats a qualsevol altra categoria d'edat superior. A continuació, es detallen els datasets recopilats:

TikHarm Dataset[41]: Es tracta d'una col·lecció curada de vídeos de TikTok dissenyada per entrenar models de classificació de contingut nociu. D'aquest dataset només s'han utilitzat les categories *Adult Content* i *Safe*. No s'ha usat la categoria *Suicide* perquè els vídeos són principalment imatges estàtiques amb textos de caràcter depressiu, cosa que no serveix per testar el classificador, ja que no és contingut de pel·lícules. La categoria *Harmful Content* tampoc s'ha incorporat perquè són vídeos amb accions que els nens no haurien d'imitar, i el classificador no interpreta quines accions s'haurien de poder imitar o no, sinó que es centra en si hi ha contingut explícit.

Violence Dataset[42]: Aquest dataset conté vídeos violents i no violents recopilats de YouTube. Els vídeos violents contenen moltes situacions reals de baralles al carrer en diversos entorns i condicions, mentre que els vídeos no violents mostren diferents accions humanes com esports, menjar, caminar, etc. A més dels vídeos sense violència, n'hi ha molts que mostren persones molt juntes o, per exemple, donant-se una abraçada ràpidament. Això permet verificar si el model confon aquestes accions, que poden ser semblants segons els frames que s'analitzin.

Pornography Dataset[43]: Aquest conjunt inclou imatges pornogràfiques i no pornogràfiques recopilades de diverses fonts. Les imatges pornogràfiques provenen de webs especialitzats, xarxes socials i pel·lícules, i inclouen també casos amb un to sensual però sense arribar a mostrar contingut sexual explícit. Les imatges no pornogràfiques s'han obtingut de plataformes d'imatges de propòsit general.

Smoking Dataset[44]: Aquest conjunt de dades conté imatges repartides equitativament entre persones que fumen i persones que no fumen. Les imatges inclouen gestos similars en ambdues classes (com beure, tossir o parlar per telèfon), cosa que permet avaluar la capacitat del classificador per distingir accions específicament associades al consum de tabac.

L'anàlisi dels resultats presentats a la taula 1 mostra que el classificador desenvolupat presenta mètriques relativament altes i globalment positives en la detecció de contingut segur i no segur. En particular, els resultats obtinguts en la detecció d'objectes específics com el consum de tabac (*Smoking Dataset*) i escenes violentes (*Violence Dataset*) són molt satisfactoris, amb una alta precisió i poca confusió entre categories. Així mateix, la classificació de contingut pornogràfic també mostra un bon rendiment general.

No obstant això, cal destacar algunes limitacions importants. En el cas del conjunt *Pornography*, la classificació

TAULA 1: DISTRIBUCIÓ PERCENTUAL DE CLASSIFICACIONS PER CONTINGUT SEGUR I NO SEGUR PER DATASETS.

Nom	U (%)	P12 (%)	13 (%)	16 (%)	18 (%)	Correcte Safe (%)	Correcte Unsafe (%)
TikHarm - Safe	95.45	4.55	0.00	0.00	0.00	**95.45**	**83.8**
TikHarm - Unsafe	16.20	41.90	2.23	27.38	12.29		
Violence - Safe	91.00	2.00	4.00	2.00	1.00	**91.00**	**92.00**
Violence - Unsafe	8.00	11.00	8.00	67.00	6.00		
Porn - Safe	73.27	17.11	1.97	6.22	1.43	**73.27**	**98.96**
Porn - Unsafe	1.04	4.17	0.62	10.96	83.21		
Smoking - Safe	92.22	7.78	0.00	0.00	0.00	**92.22**	**100.00**
Smoking - Unsafe	0.00	32.22	0.00	66.67	1.11		

de contingut segur és relativament baixa (73,27%), fet que s'atribueix principalment a un elevat nombre de falsos positius, especialment concentrats a la categoria *P12*.Aquesta incidència es deu al fet que el model és molt sensible a la presència de pell o vestimenta suggerent (per exemple, persones amb banyador). Per aquest motiu, tendeix a classificar aquestes imatges com a suggerents o potencialment inapropiades, tot i no tractar-se de contingut pornogràfic estrictament.

D'altra banda, pel que fa als vídeos procedents de xarxes socials (*TikHarm Dataset*), s'observa un menor rendiment en la categoria *Unsafe*, amb una *accuracy* del 83,8%. Aquesta dificultat es relaciona amb la naturalesa dels vídeos, sovint consistents en un únic pla seqüència, que dificulta l'extracció dels frames més representatius i, per tant, afecta negativament la detecció precisa del contingut no segur. Això es tradueix en un nombre més elevat de falsos negatius en aquesta categoria.

6 Conclusions

Aquest Treball Final de Grau ha demostrat la viabilitat tècnica i l'eficàcia del desenvolupament d'un sistema automatitzat multimodal per a la classificació de continguts audiovisuals segons criteris d'edat.

Els resultats obtinguts evidencien que s'ha creat una eina nova i potent que representa un avenç significatiu en el camp de la moderació automàtica de continguts multimèdia. Una de les principals innovacions d'aquest projecte radica en la incorporació d'un LLM per a l'anàlisi contextual d'imatges i escenes audiovisuals. Actualment, no existeix cap eina d'ús públic o de codi obert capaç d'inferir de manera precisa atributs com l'edat a partir de contingut visual i audiovisual. Les solucions existents se centren majoritàriament en classificacions binàries o en processament d'imatges estàtiques, sense aprofitar el potencial dels LLM per extreure metadades riques, context i informació semàntica complexa. La capacitat dels LLM per integrar informació multimodal i generar descripcions contextualitzades supera significativament les limitacions dels sistemes de classificació tradicionals basats únicament en detecció d'objectes o anàlisi d'imatge estàtica.

El sistema desenvolupat té com a funció principal agilitzar substancialment la tasca del classificador manual, actuant com a primera barrera de filtratge automàtic. La robustesa del pipeline multimodal implementat ofereix un alt grau de confiança en les seves decisions, especialment en els casos extrems de classificació. Quan el sistema automàtic determina que un contingut és apte per a tots els públics, aquesta decisió pot ser acceptada amb un nivell de seguretat molt elevat. Per tal que una escena inapropiada sigui erròniament classificada com a apta per a tots els públics, seria necessari que fallessin simultàniament múltiples components del sistema: el classificador de diàlegs, el detector d'objectes, el selector de fotogrames clau i el raonament del gran model de llenguatge. La probabilitat que tots aquests components fallin alhora és extremadament baixa, garantint així la fiabilitat del sistema en aquesta categoria específica.

De manera similar, la sensibilitat configurada en el gran model de llenguatge per identificar contingut restringit a majors de 18 anys ofereix també un alt grau de precisió. Aquesta característica és especialment rellevant considerant que l'eina està enfocada principalment a mercats i cultures amb normatives restrictives pel que fa al contingut audiovisual. En aquests contextos, l'aplicació de criteris estrictes per a la classificació de continguts per a adults resulta coherent amb les expectatives i normatives culturals dels usuaris finals.

No obstant això, l'avaluació dels resultats ha identificat certes limitacions en les categories intermèdies de classificació. Concretament, les categories de P12, 13 i 16 anys presenten una major tendència a la confusió i solapament en les decisions del classificador. Aquesta problemàtica és consistent amb la naturalesa inherentment subjectiva d'aquestes franges d'edat intermèdies, on la frontera entre contingut adequat i inadequat sovint depèn de factors molt més subtils. Futures iteracions del sistema haurien de contemplar una revisió i refinament dels criteris de classificació per a aquestes categories, possiblement mitjançant l'incorporació de dades d'entrenament més específiques i diferenciades per a cada franja d'edat.

En definitiva, aquest projecte ha contribuït a desenvolupar les bases per a una nova generació d'eines de classificació automàtica de continguts que integren diferents modalitats d'anàlisi amb l'ús de models de llenguatge avançats. Malgrat les limitacions observades en certes categories, el sistema desenvolupat suposa un avenç rellevant en la recerca d'alternatives més accessibles i adaptables culturalment als sistemes de moderació de continguts actualment disponibles.

Agraïments

Vull expressar el meu sincer agraïment als meus tutors, el Dr. Xim Cerdà Company (CVC) i el Dr. Francesc Tarrés Ruiz (Ugiat Technologies), per la seva orientació, paciència i suport al llarg de tot aquest treball. El seu coneixement, dedicació i comentaris han estat claus per donar forma al projecte.

També agraeixo a tot l'equip d'Ugiat Technologies la seva col·laboració, especialment pel suport tècnic, els coneixements compartits i l'accés a eines que han permès desenvolupar els experiments amb rigor. El seu compromís amb la innovació ha estat molt inspirador.

Finalment, vull donar les gràcies als companys de la Facultat d'Informàtica de la UAB, així com als meus amics i família, pel seu suport i els bons consells.

Referències

[1] Social Norms in Cinema: A Cross-Cultural Analysis of Shame, Pride and Prejudice. arXiv preprint arXiv:2402.11333, 2024. [En línia]. Disponible clicant **aqui**

[2] Exposure to sexual content and problematic sexual behaviors in children and adolescents: A systematic review and meta-analysis. Pediatrics. Epub June 19, 2023. [En línea]. Disponible clicant **aqui**

[3] Consumption of sexually explicit internet material and its effects on minors' health: latest evidence from the literature. Italian Journal of Pediatrics, Epub February 13, 2019. [En línea]. Disponible clicant **aqui**

[4] ImageNet Classification with Deep Convolutional Neural Networks, 2012. [En línia]. Disponible clicant **aqui**

[5] Attention is All You Need. 2017. [En línia]. Disponible clicant **aqui**

[6] NSFWNet. .NET Standard library per a la classificació d'imatges pornogràfiques mitjançant models preentrenats. [En línia]. Disponible clicant **aqui**

[7] OpenNSFW. Model de classificació "Not Suitable For Work" publicat per Yahoo. [En línia]. Disponible clicant **aqui**

[8] Motion Picture Association film rating system. Wikipedia, the Free Encyclopedia. [En línia]. Disponible clicant **aqui**

[9] Age Suitability Rating: Predicting the MPAA Rating Based on Movie Dialogues. 2020. [En línia]. Disponible clicant **aqui**

[10] , 2021. [En línia]. Disponible clicant **aqui**

[11] Extraction and Summarization of Explicit Video Content using Multi-Modal Deep Learning, 2023. [En línia]. Disponible clicant **aqui**

[12] Azure AI Content Safety. Servei de moderació automatitzada de contingut multimèdia. [En línia]. Disponible clicant **aqui**

[13] Google Cloud Vision API. Servei d'anàlisi d'imatges i vídeo per a moderació de contingut. [En línia]. Disponible clicant **aqui**

[14] LangGraph. Modular State Machine Library for LLM Applications. [En línia]. Disponible clicant **aqui**

[15] LangChain. Framework for Developing Applications Powered by Language Models. [En línia]. Disponible clicant **aqui**

[16] Pydantic. Data Validation and Settings Management Using Python Type Annotations. [En línia]. Disponible clicant **aqui**

[17] FFmpeg. A Complete, Cross-Platform Solution to Record, Convert and Stream Audio and Video. [En línia]. Disponible clicant **aqui**

[18] Faster Whisper. Fast and Accurate Automatic Speech Recognition Using CTranslate2 and Whisper Models. [En línia]. Disponible clicant **aqui**

[19] Voice Activity Detection. Wikipedia, the Free Encyclopedia. [En línia]. Disponible clicant **aqui**

[20] Beyond English-Centric Multilingual Machine Translation. arXiv preprint arXiv:2010.11125, 2020. [En línia]. Disponible clicant **aqui**

[21] Quantifying Multilingual Performance of Large Language Models. arXiv preprint arXiv:2404.11553. [En línia]. Disponible clicant **aqui**

[22] Transformers. State-of-the-Art Natural Language Processing for PyTorch and TensorFlow 2. Hugging Face. [En línia]. Disponible clicant **aqui**

[23] DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. [En línia]. Disponible clicant **aqui**

[24] Genius. Song Lyrics and Musical Knowledge Platform. [En línia]. Disponible clicant **aqui**

[25] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805, 2018. [En línia]. Disponible clicant **aqui**

[26] Large Model based Sequential Keyframe Extraction for Video Summarization. arXiv preprint arXiv:2401.04962. [En línia]. Disponible clicant **aqui**

[27] TransNet V2. Scene Change Detection in Videos Using Deep Neural Networks. [En línia]. Disponible clicant **aqui**

[28] Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020. [En línia]. Disponible clicant **aqui**

[29] RF-DETR: Object Detection vs YOLOv12: A Study of Transformer-based and CNN-based Architectures for Single-Class and Multi-Class Greenfruit Detection in Complex Orchard Environments Under Label Ambiguity. arXiv preprint arXiv:2504.13099, 2025. [En línia]. Disponible clicant **aqui**

[30] DETR: End-to-End Object Detection with Transformers. arXiv preprint arXiv:2005.12872, 2020. [En línia]. Disponible clicant **aqui**

[31] Microsoft COCO: Common Objects in Context. ECCV 2014. [En línia]. Disponible clicant **aqui**

[32] Roboflow Universe. Community for sharing datasets and computer vision projects. [En línia]. Disponible clicant **aqui**

[33] Falconsai/nsfw_image_detection. Model de classificació d'imatges NSFW basat en Vision Transformer. [En línia]. Disponible clicant **aqui**

[34] Hugging Face. Plataforma per a models de llenguatge i visió artificial. [En línia]. Disponible clicant **aqui**

[35] Gemma3 Report. DeepMind. [En línia]. Disponible clicant **aqui**

[36] Qwen2.5 Technical Report. Alibaba Group. [En línia]. Disponible clicant **aqui**

[37] FastAPI. Modern, fast (high-performance) web framework for building APIs with Python. [En línia]. Disponible clicant **aqui**

[38] React – A JavaScript library for building user interfaces. [En línia]. Disponible clicant **aqui**

[39] Clip.Cafe. Plataforma en línia amb una base de dades de més d'1 milió de cites i clips de pel·lícules i sèries. [En línia]. Disponible clicant **aqui**

[40] Unsloth. Plataforma per a fine-tuning de models de llenguatge de codi obert amb optimització de rendiment i ús eficient de memòria. [En línia]. Disponible clicant **aqui**

[41] TikHarm Dataset. Conjunt de dades per a la detecció de contingut d'odi i extremisme. Kaggle. [En línia]. Disponible clicant **aqui**

[42] Real-Life Violence Situations Dataset. Conjunt de dades de situacions de violència en vídeos reals. Kaggle. [En línia]. Disponible clicant **aqui**

[43] Adult Content Dataset. Conjunt de dades per a la detecció de contingut per a adults. Figshare. [En línia]. Disponible clicant **aqui**

[44] Smoking Dataset. Conjunt de dades d'imatges per a la detecció de persones fumant. Kaggle. [En línia]. Disponible clicant **aqui**

APÈNDIX

A.1 Detalls de l'Empresa, Clients i Directrius de Classificació

El present projecte es desenvolupa en col·laboració amb dues entitats principals: Ugiat Technologies i Astro BIZ. Ugiat Technologies (`https://ugiat.com/`) és l'empresa encarregada del disseny i implementació tècnica del sistema automatitzat de classificació de continguts audiovisuals. Per la seva banda, Astro BIZ (`https://biz.astro.com.my/`) és l'empresa client, amb seu a Malàisia, que ha encarregat la implementació del sistema per tal d'assegurar el compliment de la normativa nacional sobre classificació de continguts.

Astro BIZ compta amb un document de suport titulat *Astro Mapping - New Classification Guidelines 2023*, que recull de manera estructurada els criteris, exemples i casos pràctics més rellevants per a la implementació del sistema. A la Figura 6 es mostra una captura de pantalla extreta de la guia oficial per il·lustrar el format i contingut del material de referència proporcionat.

Fig. 6: Captura de pantalla de les categories de consum de tabac segons les directrius d'Astro Mapping 2023.

Aquest document es pot descarregar en línia a través del següent enllaç: Astro Mapping - New Classification Guidelines 2023

A.2 Canvis en l'enfocament del treball i planificació original

La planificació original del projecte preveia un enfocament fortament basat en l'ús de models de llenguatge de gran escala (LLM) per a la classificació automàtica de contingut audiovisual proporcionat per clients. L'estratègia contemplava l'ús del model multimodal QwenVL per generar un primer conjunt de dades a partir de vídeos reals, seguit d'un procés iteratiu de revisió manual, ajust dels models (fine-tuning) i entrenament de models més lleugers que permetessin una detecció precisa amb un cost computacional reduït.

Durant el desenvolupament, aquest enfocament es va veure greument limitat per dues circumstàncies imprevistes:

Limitacions dels LLM: Els models de llenguatge provats (com Qwen, Gemma o LLaMA) ofereixen resultats molt inferiors als esperats. Mostren una sensibilitat desproporcionada i una interpretació inconsistent de les pautes de classificació. Això obliga a substituir l'enfocament inicial de classificació automatitzada per un de més estructurat, basat en classificacions manuals, regles específiques, eines de visió per computador i l'ús de models específics per a tasques concretes (detectors d'insults, NSFW, objectes, etc.).

Retirada de dades per part dels clients: Per motius legals i de privacitat, els clients han renunciat a proporcionarn vídeos, fet que limita severament el conjunt de dades disponible. Això obliga a cercar alternatives públiques i fer ús de tècniques de web scraping per obtenir fragments audiovisuals de qualitat i duració molt variables, cosa que dificulta tant l'entrenament com l'avaluació del sistema.

Davant aquests obstacles, s'ha adoptat un nou enfocament centrat en la millora del context proporcionat als models.

Aquest canvi d'enfocament ha comportat una reestructuració important de la planificació inicial, com es mostra al Diagrama de Gantt original i es detalla al present annex. Diverses fases previstes (com l'entrenament iteratiu de models lleugers sobre dades reals) s'han substituït per processos més manuals i l'exploració de solucions híbrides per compensar la manca de dades i les limitacions dels models generals.

El diagrama de Gantt original del projecte es pot descarregar en línia a través del següent enllaç: Planificació original - Diagrama de Gantt.

A.3 Accés a l'eina de classificació

L'eina web per a la classificació automatitzada de continguts és accessible a través de: `https://nsfw.ugiat.com/`. Si es vol provar l'eina, es recomana fer-ho amb imatges en lloc de vídeos, ja que l'anàlisi de vídeo pot trigar força més temps a completar-se.