UAB
Universitat Autònoma
de Barcelona

Dipòsit digital
de documents
de la UAB

This is the **published version** of the bachelor thesis:

Fuster Aguilar, Adrian. *Detecció de text generat per intel·ligència artificial mitjançant aprenentatge automàtic*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Enginyeria de Dades)

This version is available at https://ddd.uab.cat/record/337354

under the terms of the  license.

Detecció de text generat per intel·ligència artificial amb aprenentatge automàtic

Adrián Fuster Aguilar

Resum– Aquest treball presenta un sistema complet de detecció automàtica de text generat per intel·ligència artificial en assaigs d'estudiants de secundària, construït sense dependència d'infraestructura costosa ni de datasets privats. El sistema cobreix tot el pipeline: anàlisi exploratòria i diagnòsi del desbalanceig extrem del dataset original (0,22% de textos LLM), augmentació sintètica amb tres generadors (Mistral, LLaMA3 i Gemini 2.5), entrenament de tres arquitectures (regressió logística, SVM lineal amb TF-IDF dual i DistilBERT ajustat finament), i avaluació en tres escenaris: generalització temàtica (*cross-prompt*), evasió estilística (*rewrites*) i generalització a un generador no vist (Gemini). El DistilBERT obté el millor LogLoss en domini tancat (0,100), mentre que el SVM és més robust davant de l'evasió estilística (LogLoss 0,490 vs. 0,602). L'experiment de generalització quantifica la dependència del generador quan apareix un model nou no vist i mostra que incorporar tan sols 100 mostres d'aquest nou generador redueix el LogLoss un 98%, recuperant completament el rendiment. L'anàlisi d'interpretabilitat dels coeficients del SVM revela els patrons estilístics que diferencien el text humà del generat: estructura argumentativa formal, vocabulari consistent i signatures de carta en els LLMs; escriptura col·loquial i vocabulari simple en els humans.

Paraules clau– detecció de text sintètic, LLM, PLN, classificació, augmentació de dades, TF-IDF, SVM, DistilBERT, *fine-tuning*, *rewrites*, generalització, interpretabilitat.

Abstract– This paper presents a complete system for automatic detection of AI-generated text in secondary school student essays, built without expensive infrastructure or private datasets. The system covers the full pipeline: exploratory analysis and diagnosis of the extreme class imbalance in the original dataset (0.22% LLM texts), synthetic augmentation with three generators (Mistral, LLaMA3, and Gemini 2.5), training of three architectures (logistic regression, dual TF-IDF + linear SVM, and fine-tuned DistilBERT), and evaluation across three scenarios: thematic generalization (cross-prompt), stylistic evasion (rewrites), and generalization to an unseen generator (Gemini). DistilBERT achieves the best LogLoss in closed-domain validation (0.100), while SVM is more robust against stylistic evasion (LogLoss 0.490 vs. 0.602). The generalization experiment quantifies the effect of generator dependency when a new unseen model appears, and shows that adding only 100 samples from that generator reduces LogLoss by 98%, fully recovering performance. Interpretability analysis of SVM coefficients reveals the stylistic patterns that distinguish human from AI-generated text: formal argumentative structure, consistent vocabulary, and letter-style closings in LLMs; colloquial writing and simple vocabulary in humans.

Keywords– synthetic text detection, LLM, NLP, classification, data augmentation, TF-IDF, SVM, DistilBERT, fine-tuning, rewrites, generalization, interpretability.

1 Introducció i motivació

Aquest treball presenta el sistema complet de detecció automàtica de text generat per intel·ligència artificial en assaigs d'estudiants de secundària, desenvolupat com a Treball de Final de Grau (TFG). El projecte ha cobert tot el pipeline de detecció: des de l'anàlisi exploratòria del dataset original i el disseny de l'augmentació de dades, fins a l'entrenament i avaluació de tres arquitectures de classificació i l'anàlisi d'interpretabilitat dels resultats.

La detecció de text generat per LLM és un problema d'interès creixent en múltiples contextos. En l'àmbit educatiu, la proliferació de models com GPT-4, Gemini o LLaMA ha fet que la generació de text de qualitat sigui accessible per a qualsevol estudiant, la qual cosa planteja reptes nous per a l'avaluació de l'autenticitat dels treballs acadèmics. Les eines comercials existents com Turnitin o GPTZero presenten taxes de falsos positius significatives, penalitzant injustament estudiants que escriuen amb un estil formal, i no estan adaptades a dominis específics com els assaigs argumentatius d'estudiants de secundària, que presenten patrons estilístics molt particulars (errors ortogràfics, estructures ir-

regulars, canvis de registre). Aquest treball explora si és possible construir un sistema competitiu sense dependència d'infraestructura costosa ni de datasets privats, mitjançant dades sintètiques generades localment i models entrenables en CPU o amb GPU gratuïta.

Una motivació addicional del projecte és la component de disseny de dades: el problema del desbalanceig extrem del dataset original (més del 99% d'assaigs humans) exigeix estratègies d'augmentació sofisticades que van més enllà de la simple replicació de mostres, la qual cosa planteja preguntes rellevants sobre la diversitat dels generadors usats i el seu impacte en la generalització del detector.

El present informe s'organitza de la manera següent: la Secció 2 descriu els objectius del projecte; la Secció 3 descriu la metodologia i la planificació; la Secció 4 revisa l'estat de l'art; la Secció 5 detalla el desenvolupament del sistema; la Secció 6 presenta i discuteix els resultats; i la Secció 7 exposa les conclusions definitives.

2 OBJECTIUS

En aquest apartat es descriu el conjunt d'objectius que el TFG es marca. Els resultats obtinguts i el grau d'assoliment es presenten a les seccions de Desenvolupament i Conclusions.

Objectiu principal: Desenvolupar un sistema capaç de detectar text generat per LLM en assaigs d'estudiants de secundària, amb un focus en la generalització a generadors no vistos i la reproductibilitat dels resultats en entorns amb recursos computacionals limitats.

O1. Anàlisi exploratòria del conjunt de dades i diagnòsi del desbalanceig de classes. El punt de partida és un dataset amb un desbalanceig extrem (menys del 0,22% de textos generats per LLM), que fa inviable l'entrenament directe d'un classificador. L'objectiu és caracteritzar quantitativament i qualitativament el desbalanceig i identificar diferències estilístiques observables entre les dues classes que justifiquin les decisions de disseny posteriors.

O2. Reequilibrar el conjunt de dades mitjançant augmentació amb dades sintètiques. L'objectiu és generar assaigs sintètics amb múltiples models generadors —incloent-hi models de codi obert i models propietaris— aplicant filtres de qualitat automàtics, i construir un pipeline d'augmentació reproduïble i extensible a nous generadors.

O3. Establir una línia de referència (*baseline*) mitjançant l'entrenament i avaluació d'un model clàssic (TF-IDF de caràcters + Regressió Logística) amb protocol de validació cross-prompt. El baseline ha de permetre identificar les limitacions de calibració que motivin el disseny del model avançat i establir un punt de comparació objectiu.

O4. Desenvolupar el model principal i comparar-lo amb el baseline. Es proposa implementar un model avançat basat en TF-IDF dual i SVM lineal, i un Transformer (DistilBERT) ajustat finament, i avaluar-los en tres escenaris: generalització temàtica (*cross-prompt*), robustesa davant de l'evasió estilística (*rewrites*) i generalització a un generador no vist, permetent una comparació multidimensional entre arquitectures.

O5. Anàlisi de resultats i conclusions definitives. Inclou l'anàlisi d'interpretabilitat dels models, la comparació sistemàtica amb els treballs de referència de la literatura, i la reflexió sobre la viabilitat del sistema en producció real i les seves limitacions.

3 PLANIFICACIÓ I METODOLOGIA

La metodologia de treball s'organitza mitjançant una aproximació iterativa basada en Kanban [1]. Les tasques es prioritzen en un tauler amb tres columnes (*Per fer*, *En curs*, *Fet*), associades a entregables concrets: codi Python documentat, figures i taules de resultats, i redacció de la memòria. La metodologia ha demostrat el seu valor en dos moments crítics: quan els resultats del baseline van revelar una limitació de calibració que va motivar el disseny del model avançat, i quan l'experiment de generalització va detectar la dependència del generador, motivant l'addició de Gemini al conjunt d'entrenament. En ambdós casos, la flexibilitat del Kanban va permetre replanificar ràpidament.

Pel que fa a l'entorn tecnològic, el projecte s'implementa íntegrament en Python [2], amb Visual Studio Code com a entorn de desenvolupament principal. Les biblioteques principals són `scikit-learn` [3] per a la vectorització TF-IDF i els models clàssics, `transformers` [4] i `torch` per al fine-tuning de DistilBERT, `pandas` per a la manipulació de dades, i `matplotlib` i `seaborn` per a la visualització. Per a la generació de dades sintètiques en local s'utilitza Ollama [5], que permet executar LLMs sense connexió a serveis externs garantint la privacitat de les dades dels estudiants; per a Gemini s'utilitza la biblioteca `google-genai` i l'API de Google AI Studio [6]. El fine-tuning de DistilBERT s'ha executat a Google Colab amb GPU Tesla T4 (aproximadament 1,5 minuts per partició cross-prompt), ja que el cost computacional en CPU local (4–8 hores per època) no és viable en el context del projecte.

Quant a la planificació temporal, les fites oficials han estat: 23-feb Reunió inicial; 09-mar Informe inicial; 19-abr Informe de Seguiment 1; 24-mai Informe de Seguiment 2; 14-jun Proposta informe final; 28-jun Dossier i Informe final; 1 de juliol Defensa. La Fig. 1 (Annex A) mostra el diagrama de Gantt.

4 ESTAT DE L'ART

La detecció de text generat per LLM és un problema emergent que ha rebut atenció creixent des de l'aparició de ChatGPT el novembre del 2022. En aquest apartat es revisen els principals enfocaments existents, s'analitzen les seves limitacions específiques i es contextualitzen els resultats del present treball.

Els mètodes basats en *mètriques estadístiques sense entrenament* detecten text generat mesurant propietats intrínseques del text sota un model de referència, sense necessitat de dades etiquetades. El mètode més estudiat és la perplexitat: els LLMs tendeixen a generar text d'alta probabilitat (baixa perplexitat), mentre que el text humà presenta una distribució de probabilitats més irregular. L'eina GLTR

(Giant Language model Test Room) visualitza la distribució de probabilitats dels tokens per ajudar a identificar text generat. DetectGPT [7] proposa avaluar si el text es troba en una zona de curvatura negativa de la funció de log-probabilitat del model generador, obtenint bones prestacions sense necessitat d'entrenament. L'avantatge principal d'aquests mètodes és que no requereixen dades etiquetades i s'adapten fàcilment a nous dominis. Tanmateix, presenten dues limitacions importants: el rendiment cau significativament quan el text ha estat reescrit o parafrasejat (tècniques d'evasió que redueixen la perplexitat), i tendeixen a generar falsos positius en textos humans molt tècnics o formals, que presenten distribucions de probabilitat similars als textos generats.

Els *classificadors supervisats clàssics* utilitzen representacions vectorials del text com a entrada a models estadístics. Les representacions més comunes són les basades en TF-IDF sobre n-grames de paraules o caràcters. A partir d'aquestes representacions s'entrenen classificadors com la regressió logística, les màquines de suport vectorial (SVM) o els arbres de decisió potenciats per gradient. Treballs com el de [8] han documentat que, en dominis tancats on el model generador és conegut durant l'entrenament, la regressió logística sobre TF-IDF assoleix precisions superiors al 95% amb un cost computacional molt baix. L'inconvenient principal és la manca de generalització: aprenen patrons superficials lligats a l'estil idiosincretic d'un model concret, de manera que el rendiment cau notablement quan s'avaluen sobre textos generats per un model diferent.

Els *detectors basats en Transformers* utilitzen models de llenguatge preentrenats ajustats finament sobre conjunts de dades etiquetades de detecció. Models com RoBERTa [9] o DeBERTa han demostrat ser especialment efectius: en processar cada token en el context de tots els altres tokens de la seqüència, aprenen representacions semàntiques profundes que capturen patrons de coherència i consistència de registre. Eines comercials com l'OpenAI Text Classifier (ara discontinuat) o sistemes acadèmics com RADAR han seguit aquest enfocament. El principal inconvenient és doble: el cost computacional d'entrenament i inferència i, sobretot, la tendència al sobreajustament al model generador concret usat durant l'entrenament [10].

Recentment, la literatura ha destacat la *dependència del generador* com el principal obstacle per a la generalització real dels detectors [10]. Sistemes entrenats sobre un LLM específic rarament transfereixen bé a d'altres, independentment de l'arquitectura del detector. Alguns treballs proposen solucions basades en la generació de dades sintètiques d'entrenament amb múltiples generadors, però no avaluen de forma sistemàtica quantes mostres per generador són necessàries per recuperar el rendiment. El present treball contribueix a tancar aquest buit amb un experiment que quantifica la degradació del rendiment davant de Gemini i verifica que tan sols 100 assaigs addicionals (50 per tema) són suficients per recuperar completament el rendiment, un resultat amb implicacions directes per al disseny de sistemes de detecció en producció.

5 DESENVOLUPAMENT

Aquesta secció descriu el desenvolupament complet del sistema de detecció, organitzat en quatre apartats. Primer es detalla la construcció progressiva del conjunt d'entrenament mitjançant augmentació sintètica (5.1). A continuació s'expliquen el baseline de regressió logística (5.2) i el model avançat basat en TF-IDF dual i SVM lineal (5.3). Després es presenta el model Transformer DistilBERT ajustat finament (5.4). Finalment, l'apartat 5.5 descriu l'experiment de generalització amb Gemini, que avalua el comportament del sistema davant d'un generador no vist durant l'entrenament.

5.1 Ampliació del conjunt d'entrenament (O2)

El conjunt de dades original prové de la competició *LLM—Detect AI Generated Text* de Kaggle [12] i conté 1.378 assaigs d'estudiants de secundària sobre dos temes de redacció argumentativa: l'abolició del col·legi electoral dels Estats Units i l'obligatorietat del servei comunitàri. D'aquests, tan sols 3 són generats per un LLM (menys del 0,22%), la qual cosa representa un desbalanceig extrem que fa inviable l'entrenament directe d'un classificador: qualsevol model tendirà a predir sempre la classe majoritària, obtenint un recall nul sobre la classe d'interès. Aquesta diagnosi justifica la fase d'augmentació descrita a continuació.

El conjunt d'entrenament inicial augmentat (train v1) contenia 1.375 assaigs humans i 1.203 assaigs generats amb Mistral [13] mitjançant Ollama [5], amb una distribució aproximadament equitativa (53,34% humans, 46,66% LLM). Els textos de Mistral es van generar amb tres nivells de temperatura (0,2, 0,5 i 0,8) per afegir diversitat lingüística: temperatures baixes produeixen textos més conservadors i estructurats, mentre que temperatures altes generen textos més creatius però menys coherents.

Malgrat el balançeig adequat entre classes, tots els textos sintètics provenien del mateix model generador. Aquesta situació podia provocar que el classificador aprengués patrons idiomàtics específics de Mistral —com l'ús sistèmàtic de determinats connectors, la longitud característica de les seves frases o la seva preferència per certes estructures retòriques— en lloc de patrons generals que distingissen qualsevol text generat per IA del text humà. Per abordar aquesta limitació, s'han afegit 150 assaigs generats amb LLaMA3 [14] (75 per tema), obtenint el train v2 amb 2.728 assaigs i 1.353 mostres de la classe LLM.

LLaMA3 és un model de la família Meta amb una arquitectura Transformer diferent de Mistral i una metodologia d'entrenament distinta, la qual cosa es reflecteix en característiques estilístiques observables: tendència a estructures de frases lleugerament més curtes, vocabulari menys formal en alguns registres, i una freqüència d'ús diferent de connectors argumentatius com *furthermore*, *moreover* o *in addition*. Aquesta diversitat estilística aporta informació nova al conjunt d'entrenament i obliga el classificador a aprendre patrons més generals.

Posteriorment, per a l'experiment de generalització descrit a la subseguent subsecció, s'han afegit 100 assaigs ge-

nerats amb Gemini 2.5 Flash [6] (50 per tema) via API, obtenint el train v3 amb 2.828 assaigs i 1.453 mostres de la classe LLM. La Taula 1 resumeix l'evolució completa del conjunt d'entrenament al llarg del projecte.

TAULA 1: EVOLUCIÓ DEL CONJUNT D'ENTRENAMENT.

Vers.	Hum.	LLM	Gen.
Original	1.375	3	—
v1 (Mistral)	1.375	1.203	Mistral
v2 (LLaMA3)	1.375	1.353	Mist.+LL3
v3 (Gemini)	1.375	1.453	Mist.+LL3+Gem.

5.2 Baseline: TF-IDF + Regressió Logística (O3)

El baseline es basa en TF-IDF de caràcters (n-grames 2–6, 50.000 característiques, `sublinear_tf=True`) i Regressió Logística amb regularització L2 i pesos de classe balançejats. La representació de n-grames de caràcters és especialment adequada per a tasques d'atribució d'autoria i detecció d'estil perquè captura patrons morfològics, de puntuació i de contraccions que són independents del contingut temàtic i molt sensibles al generador del text. Per exemple, la freqüència d'aparíçió de la seqüència de caràcters `", "` (coma espai cometa) reflecteix l'ús de llistes i enumeracions, característic dels textos generats per LLM, i pot ser capturat per n-grames de caràcters de longitud 3 o 4 sense necessitat de comprendre el significat de les paraules que l'envolten.

El protocol de validació *cross-prompt* és fonamental per evitar el risc de *data leakage* temàtic. En cada partició, el model s'entrena sobre la totalitat dels assaigs d'un dels dos temes de redacció (tant humans com sintètics) i es valida sobre la totalitat dels assaigs de l'altre tema. En cap cas el model veu durant l'entrenament cap assaig del tema de validació, de manera que les mètriques resultants reflecteixen la capacitat de generalització a continguts temàtics nous, no només a texts nous sobre el mateix tema. Aquesta elecció és més exigent que la partició aleatòria convencional i proporciona una estimació del rendiment més propera a la que s'obtindria en producció real.

La Taula 2 recull els resultats detallats per partició. La variància entre les dues particions (LogLoss 0,915 vs. 0,622) indica que el rendiment depèn parcialment del tema: la partició A (entrenament sobre el tema del servei comunitari, validació sobre el col·legi electoral) és lleugerament més difícil, possiblement perquè els textos generats sobre el col·legi electoral presenten un vocabulari més específic i formal que els del servei comunitàri, més proper a l'estil dels assaigs humans.

Malgrat l'AUC gairebé perfecta (0,9999), el LogLoss relativament elevat (0,768 en cross-prompt) revela que les probabilitats predites no estan ben calibrades: el model encerta la classificació però sovint assigna probabilitats poc extremes, reflectint incertesa en casos propers a la frontera de decisió. En la validació amb *rewrites* (reescriptures de textos generats per un LLM amb instruccions d'imitar l'estil d'un estudiant, n=120 per classe), el rendiment es manté alt en AUC (0,9944) però la probabilitat mitjana assignada als *rewrites* és de tan sols 0,318, molt per sota de la probabilitat que el model assigna als textos directament generats (propera a 1,0). Això indica que la humanització de l'estil redueix significativament la confiança del detector, justificant el disseny del model avançat.

TAULA 2: RESULTATS DEL BASELINE (TF-IDF CARÀCTERS + REGRESSIÓ LOGÍSTICA).

Avaluació	AUC	LogLoss
Cross-prompt (partició A)	0,9998	0,915
Cross-prompt (partició B)	0,9999	0,622
Mitjana cross-prompt	0,9999	0,768
Rewrites (humans vs. reesc.)	0,9944	0,630

5.3 Model avançat: TF-IDF dual + SVM lineal (O4)

El model avançat s'ha dissenyat per superar les limitacions de calibració del baseline. Es basa en dues decisions clau: una representació més rica del text (TF-IDF dual de caràcters i paraules) i un classificador lineal més adequat per a espais d'alta dimensió (SVM). Els dos apartats següents descriuen cadascuna d'aquestes components i la seva justificació.

5.3.1 Representació: TF-IDF dual de 80.000 dimensions

El model avançat combina dues representacions TF-IDF concatenades horitzontalment, formant un espai de 80.000 dimensions. El primer component és un TF-IDF de caràcters (n-grames de longitud 2 a 6, 50.000 característiques, `sublinear_tf=True`, `analyzer='char_wb'`) per capturar patrons morfològics i estilístics a nivell sublèxic: la freqüència de la puntuació sistèmàtica dels LLMs (com l'ús consistent de la coma després d'un connector), els patrons de contracció característics de cada model, els sufixos i prefixos preferits, i les transicions entre caràcters que delaten un estil de redacció uniforme i poc variable. L'opció `analyzer='char_wb'` afegeix espais als extrems de cada paraula, de manera que els n-grames capturen també els límits de paraula, augmentant la seva capacitat discriminativa.

El segon component és un TF-IDF de paraules (unigrames i bigrames, 30.000 característiques, `sublinear_tf=True`) per capturar vocabulari i col·locacions freqüents en textos generats per LLMs. N-grames de paraules com *it is important to*, *in conclusion*, *furthermore*, *it is worth noting* o *one must consider* són molt més freqüents en textos de LLMs que en assaigs d'estudiants de secundària, que tendeixen a usar estructures més simples i directes. Aquests patrons no són accessibles als n-grames de caràcters i aporten informació complementària sobre el *què* s'escriu.

La representació dual combina els avantatges de tots dos enfocaments: el TF-IDF de caràcters captura *com* s'escriu

(estil, morfologia, puntuació), mentre que el TF-IDF de paraules captura *què* s'escriu (vocabulari, col·locacions). Les dues matrius TF-IDF s'ajusten sobre el conjunt d'entrenament de forma independent i es concatenen horitzontalment, de manera que cada assaig queda representat per un vector de 80.000 dimensions, on les primeres 50.000 dimensions corresponen als n-grames de caràcters i les darreres 30.000 als n-grames de paraules.

5.3.2 Classificador: SVM lineal calibrat

Sobre la representació dual s'entrena un SVM lineal amb pesos de classe balançejats (`class_weight='balanced'`), que ajusta automàticament la importància de cada classe en funció de la seva freqüència, compensant possibles desequilibris residuals. Per obtenir estimacions de probabilitat necessàries per calcular el LogLoss i permetre la comparació amb el baseline i el DistilBERT, s'utilitza la calibració de Platt mitjançant `CalibratedClassifierCV` de scikit-learn, que ajusta una regressió logística sobre les puntuacions del SVM usant validació creuada interna.

Les SVM lineals han demostrat ser especialment eficients en espais TF-IDF dispersos d'alta dimensió, on la majoria dels valors són zero i les poques característiques amb valor diferent de zero són molt informatives. En aquests espais, el kernel lineal és suficient per obtenir una separació òptima, i models més complexos com el kernel RBF o els arbres de decisió solen ser innecessaris i menys eficients. La tria del SVM sobre XGBoost es justifica empíricament: en condicions idèntiques de representació (TF-IDF dual, 80.000 dimensions) i protocol de validació (cross-prompt), XGBoost obté un LogLoss de 2,119 en cross-prompt i una AUC de 0,977, molt per sota del SVM (LogLoss 0,500, AUC 1,000). Això confirma que els models basats en arbres de decisió particionen l'espai de forma binària i necessiten moltes divisions per capturar patrons en espais dispersos d'alta dimensió, mentre que el SVM lineal troba directament la separació òptima [8].

5.4 Model Transformer: DistilBERT *fine-tuned* (O4)

Per complementar l'enfocament basat en freqüències del SVM, s'ha explorat si un model que captura context semàntic profund pot millorar la calibració en domini tancat. Els apartats següents descriuen l'arquitectura escollida i el procés d'ajust fi, incloent-hi les decisions que han permès superar problemes de sobreajustament detectats en una primera versió del model.

5.4.1 Arquitectura i justificació

Per explorar si un model que captura context semàntic profund pot superar el SVM, s'ha ajustat finament el model preentrenat `distilbert-base-uncased` [15] (66 milions de paràmetres, 6 capes Transformer, 768 dimensions d'embedding). DistilBERT és una versió comprimida de BERT obtinguda mitjançant la tècnica de *knowledge distillation*: un model "estudiant" (DistilBERT) és entrenat per imitar les sortides d'un model "professor" (BERT-base) més gran, conservant el 97% del seu rendiment amb un 40% menys de paràmetres i una inferència un 60% més ràpida. La seva grandaria el fa especialment adequat per a entorns amb recursos computacionals limitats.

A diferència del TF-IDF, que representa cada text com un vector de freqüències de n-grames independent del context, DistilBERT processa cada token en el context de tots els altres tokens de la seqüència gràcies al mecanisme d'atenció (*self-attention*). Això li permet capturar dependències semàntiques de llarg abast com la coherència argumentativa al llarg del text, la consistència de registre entre frases, i les relacions lògiques entre paràgrafs. Totes aquestes característiques són difícils de capturar amb representacions basades en freqüències.

5.4.2 Configuració i procés de *fine-tuning*

El *fine-tuning* consisteix en afegir una capa de classificació binària (`Linear(768, 2)`) al damunt de la representació del token `[CLS]` (que DistilBERT usa com a representació global de la seqüència) i continuar l'entrenament del model complet durant poques èpoques sobre el dataset de la tasca. Els assaigs es tokenitzen amb el tokenitzador de DistilBERT (`DistilBertTokenizer`) i es truncen a una longitud màxima de 512 tokens.

La configuració final s'ha obtingut després d'un procés d'ajust per corregir problemes d'overfitting. Una primera versió (4 èpoques, learning rate $= 2 \times 10^{-5}$, weight decay = 0,01) va presentar signes clars de sobreajustament: el training loss decreixia de forma contínua fins a 0,001, però el validation loss augmentava des de l'època 2 (de 0,29 a 1,85 en la primera partició), indicant que el model memoritzava el conjunt d'entrenament en lloc d'aprendre característiques generals. El LogLoss en cross-prompt d'aquesta primera versió era de 1,019, pitjor que el baseline.

La configuració final utilitza 2 èpoques, learning rate $= 10^{-5}$ (més petit per suavitzar les actualitzacions), weight decay $= 0,1$ (regularització més forta), i batch size de 16. La precisió mixta (`fp16=True`) s'ha activat per aprofitar les capacitats de la GPU Tesla T4 i reduir el consum de memòria. Amb aquesta configuració el validation loss decreix monotonament durant l'entrenament (de 0,44 a 0,07 en la primera partició cross-prompt), indicant una generalització correcta sense sobreajustament. El temps d'entrenament és d'aproximadament 1,5 minuts per partició cross-prompt a Google Colab amb GPU Tesla T4.

5.5 Experiment de generalització amb Gemini (O4)

L'experiment de generalització té com a objectiu quantificar la dependència del generador: avaluar fins a quin punt un detector entrenat exclusivament sobre textos de Mistral i LLaMA3 manté el seu rendiment davant d'un model nou no vist, i determinar quantes mostres addicionals són necessàries per recuperar el rendiment. L'apartat 5.5.1 descriu la generació dels textos de Gemini i l'apartat 5.5.2 el disseny de l'experiment en dues fases complementàries.

5.5.1 Generació dels textos de Gemini

S'han generat 100 assaigs amb Gemini 2.5 Flash [6] (50 per tema de redacció) mitjançant l'API de Google AI Studio. Gemini és un model propietari de Google, amb una arquitectura multimodal i una metodologia d'entrenament molt diferent dels models open source Mistral i LLaMA3 usats durant l'entrenament del sistema. Mentre que Mistral i LLaMA3 pertanyen a la família dels Transformers decoder-only entrenats principalment amb text, Gemini és un model nàtiu multimodal entrenat amb dades de text, imatge, àudio i vídeo, la qual cosa influeix en el seu estil de generació de text i en els patrons lingüístics que utilitza.

Per a la generació s'ha usat el mateix prompt que en les fases anteriors: instruccions d'escriure un assaig argumentatiu en l'estil d'un estudiant de secundària, sense mencionar que es tracta d'una IA, amb referència al text font original de la competició de Kaggle. S'han aplicat els mateixos filtres de qualitat que en les fases anteriors (longitud mínima de 100 paraules, eliminació de referències explícites a IA), i els textos resultants presenten una longitud mitjana de 380 paraules, comparable a la dels assaigs de Mistral i LLaMA3. La generació s'ha executat a Google Colab mitjançant la biblioteca `google-genai`, amb un temps de sleep de 2 segons entre peticions per respectar els límits de l'API.

5.5.2 Disseny de l'experiment

L'experiment de generalització consta de dues parts amb objectius complementaris. En la primera part (**avaluació sense reentrenament**), el SVM v2 i el DistilBERT v2 — tots dos entrenats exclusivament sobre textos de Mistral i LLaMA3— s'avaluen sobre els 100 assaigs de Gemini enfrontats a 100 assaigs humans originals del dataset de Kaggle, sense cap reentrenament ni adaptació. Aquesta configuració simula el cas real en què el detector ja està desplegat en producció i apareix un nou LLM al mercat; l'objectiu és quantificar quant cau el rendiment en aquest escenari.

En la segona part (**reentrenament amb Gemini**), els 100 assaigs de Gemini s'incorporen al conjunt d'entrenament obtenint el train v3 (1.375 humans + 1.453 LLM), i es reentrena el SVM per mesurar la millora obtinguda. De manera crucial, es verifica simultàniament que l'addició de Gemini no perjudica el rendiment en els escenaris anteriors (cross-prompt i *rewrites*), assegurant que la incorporació de nous generadors és acumulativa i no implica cap oblit dels generadors anteriorment apresos (*catastrophic forgetting*).

6 RESULTATS I DISCUSSIÓ

Aquesta secció integra els resultats dels models clàssics i avançats, i interpreta no només quin model obté millor puntuació, sinó també en quines condicions falla. L'objectiu no és seleccionar un model de forma aïllada, sinó entendre el compromís entre calibració, robustesa davant *rewrites* i adaptació a generadors no vistos.

Les mètriques principals són l'AUC (*Area Under the ROC Curve*), que mesura la capacitat discriminativa independent del llindar de decisió, i el LogLoss (*Logarithmic Loss*), que penalitza la manca de calibració de les probabilitats predites i que és la mètrica oficial de la competició de Kaggle.

6.1 Resultats comparatius

La Taula 3 recull les mètriques de tots els models en els tres escenaris d'avaluació.

TAULA 3: RESULTATS COMPARATIUS. V2 = MISTRAL+LLAMA3; V3 = AFEGIT GEMINI.

Model	Avaluació	AUC	LL
Baseline (LR)	Cross-prompt	0,9999	0,768
Baseline (LR)	Rewrites	0,9944	0,630
SVM v2	Cross-prompt	1,0000	0,500
SVM v2	Rewrites	1,0000	0,490
SVM v2	Gemini (no vist)	0,9969	1,252
SVM v3	Gemini (al train)	1,0000	0,020
SVM v3	Rewrites	1,0000	0,362
DistilBERT v2	Cross-prompt	0,9999	0,100
DistilBERT v2	Rewrites	0,9977	0,602
DistilBERT v2	Gemini (no vist)	0,9668	1,494

6.2 Model avançat vs. baseline

El SVM v2 millora el baseline en LogLoss un **35%** en cross-prompt (0,500 vs. 0,768) i un **22%** en *rewrites* (0,490 vs. 0,630), mantenint AUC perfecta en tots dos escenaris. La millora en LogLoss indica que el model no només encerta la classificació sinó que ho fa amb una confiança més adequada: les probabilitats predites són més extremes (més properes a 0 per als humans i a 1 per als LLM) que les del baseline. Això és especialment rellevant en aplicacions reals, on el sistema ha de proporcionar un nivell de confiança fiable per a la presa de decisions.

La combinació de TF-IDF de caràcters i paraules aporta informació complementària que el baseline no capturava. El component de caràcters detecta patrons estilístics sublèxics que persisteixen en textos humanitzats (com la freqüència i distribució de la puntuació, o l'ús sistemàtic de determinats sufixos), mentre que el component de paraules captura col·locacions lèxiques que diferencien el vocabulari formal dels LLMs del registre col·loquial dels estudiants. Els dos components són més efectius junts que per separat: experiments d'ablació preliminars indiquen que usar només el TF-IDF de caràcters dóna un LogLoss de 0,768 (igual que el baseline) i usar només el de paraules dóna un LogLoss de 0,653, mentre que la combinació dóna 0,500. L'efecte additiu confirma que les dues representacions capturen informació diferent i complementària.

L'experiment negatiu amb XGBoost (LogLoss 2,119 en cross-prompt, AUC 0,977) és igualment valuós per al TFG: confirma empíricament que els models basats en arbres no són adequats per a espais TF-IDF dispersos d'alta dimensió, consistent amb la teoria i la literatura [8], i proporciona una justificació quantitativa per a l'elecció del SVM. L'AUC de 0,977 indica que XGBoost pot discriminar les dues classes,

però el LogLoss molt elevat revela que les probabilitats predites estan molt mal calibrades, molt probablement a causa de la cursa de particions binàries en un espai de 80.000 dimensions on la majoria de valors són zero.

6.3 DistilBERT: millor calibració en domini tancat, menor robustesa a l'evasió

En validació cross-prompt, el DistilBERT v2 obté el millor LogLoss de tots els models (0,100), una millora del **87%** respecte al baseline (0,768) i del **80%** respecte al SVM v2 (0,500). El Transformer, gràcies al mecanisme d'atenció que processa cada token en el context de tots els altres, aprèn representacions semàntiques riques que capturen patrons de coherència argumentativa, consistència de registre i estructures retòriques característiques dels textos generats per LLMs. Concretament, els LLMs tendeixen a mantenir un registre extremadament consistent al llarg del text, amb transicions argumentatives previsibles i una densitat d'informació uniforme entre paràgrafs, mentre que els assaigs humans presenten més variabilitat i desequilibris de registre. Aquestes característiques són millor capturades per un Transformer que per un model de freqüències.

Tanmateix, en la validació amb *rewrites*, el DistilBERT v2 (LogLoss 0,602) és superat clarament pel SVM v2 (LogLoss 0,490). Quan el text sintètic imita deliberadament l'estil humà incorporant errors ortogràfics lleus, frases inacabades, canvis de registre i variació sintàctica, les representacions semàntiques del Transformer perden discriminació perquè l'estil s'ha adaptat per semblar humà. El Transformer ha après a reconèixer el registre consistent i la coherència perfecta dels LLMs, i quan aquesta coherència s'interromp deliberadament, el seu senyal principal desapareix. En canvi, els n-grames de caràcters del TF-IDF capturen patrons molt fins d'escriptura (la freqüència de determinades combinacions de puntuació, morfologia de les paraules usades) que són molt més difícils de modificar deliberadament i persisteixen fins i tot en textos humanitzats. Aquest resultat és consistent amb el fenòmen de *shortcut learning* [11]: els Transformers aprenen dreceres semàntiques que fallen quan l'estil s'humanitza deliberadament, mentre que les representacions de freqüències del TF-IDF capturen patrons més profunds i resistents a la modificació estilística.

6.4 Limitació de generalització i experiment amb Gemini

Quan s'avaluen sobre textos de Gemini sense reentrenament, tant el SVM v2 (LogLoss 1,252, P(generat) = 0,132) com el DistilBERT v2 (LogLoss 1,494, P(generat) = 0,086) presenten una degradació dràstica. Les probabilitats mitjanes assignades als textos de Gemini són de 0,132 i 0,086 respectivament, molt properes a zero: els models perceben els textos de Gemini gairebé com a humans. Això il·lustra clarament el fenòmen de dependència del generador: els detectors aprenen els patrons idiomàtics específics de Mistral i LLaMA3 (estructura de frases, vocabulari, puntuació), i quan apareix Gemini, amb el seu propi estil distint, el detector no el reconeix com a text generat per IA.

És especialment rellevant que la degradació sigui pitjor en el DistilBERT (LogLoss 1,494) que en el SVM (LogLoss 1,252). Això suggereix que les representacions semàntiques profundes del Transformer, tot i ser més potents en domini tancat, són més específiques als generadors vistos i menys capaces de transferir a generadors nous. Els n-grames de caràcters, en canvi, capturen alguns patrons estilístics més universals del text generat per IA que es transfereixen parcialment a Gemini, tot i que insuficientment.

En incorporar els 100 assaigs de Gemini al conjunt d'entrenament (SVM v3), el LogLoss sobre textos de Gemini cau de 1,252 a 0,020, una reducció del **98%**, i la probabilitat mitjana puja de 0,132 a 0,968. Simultàniament, el rendiment sobre *rewrites* millora (LogLoss 0,490 $\rightarrow$ 0,362) i el rendiment cross-prompt es manté estable (LogLoss $\approx$ 0,500). L'addició de Gemini no perjudica cap escenari anterior i millora la robustesa general. Amb tan sols 100 mostres (50 per tema) el sistema s'adapta completament: això demostra que la limitació central no és l'arquitectura sinó la diversitat de dades.

6.5 Interpretabilitat del SVM v3

Per identificar els patrons lingüístics que el SVM utilitza com a senyal discriminatiu, s'han extret els coeficients del `LinearSVC` entrenat sobre el train v3 (80.000 dimensions, Mistral + LLaMA3 + Gemini). La Taula 4 sintetitza els patrons més representatius de cada classe.

Els n-grames més associats a **text LLM** reflecteixen dos fenòmens distints. Primer, *estructura argumentativa formal i previsible*: bigrames com *argue that* ($w = 0{,}197$), *some argue* (0,186) i *believe that* (0,150), juntament amb el connector *in conclusion* (0,143) i l'adverbi *however* (0,143), confirmen que els LLMs construeixen els assaigs seguint una estructura retòrica sistèmàtica i poc variable. Segon, *vocabulari formal consistent*: *ensures* (0,171), *mandatory* (0,167), *significant* (0,164) i *sustainable* (0,162) són molt més freqüents en textos generats que en assaigs adolescents. Un senyal especialment fort i interpretable és la seqüència *sincerely–your name* ($w \approx 0{,}145$): els LLMs acaben els assaigs amb una signatura formal de carta, cosa que cap estudiant real fa.

Els n-grames més associats a **text humà** reflecteixen l'escriptura directa i col·loquial dels estudiants. Paraules simples com *people* ($w = -0{,}247$), *so* ($-0{,}213$), *because* ($-0{,}183$) i *if* ($-0{,}180$) són molt més freqüents en assaigs adolescents. Els n-grames de caràcters `ople` i `eople` (fragments sublèxics de *people*, $w \approx -0{,}149$) confirmen que els humans usen *people* com a subjecte genèric, mentre que els LLMs prefereixen alternatius més formals com *individuals*, *residents* o *communities*.

Cal notar que alguns n-grames LLM són temàtics en lloc d'estilístics: *community service* (0,270) i *vote system* (0,187) són característics dels dos temes de redacció del dataset, no propietats universals del text generat per IA. En un corpus més divers temàticament, els senyals purament estilístics com *argue that* o *in conclusion* quedarien més destacats com els veritables patrons discriminatius.

6.6 Síntesi i implicacions pràctiques

Cap dels tres models és superior en tots els escenaris simultàniament, reflectint un *trade-off* fonamental entre calibració en domini tancat (guanya DistilBERT), robustesa a l'evasió estilística (guanya SVM) i adaptabilitat a nous generadors (guanya SVM v3 després de reentrenament).

Des d'una perspectiva pràctica, l'elecció del model dependrà del cas d'ús: si el sistema s'ha de desplegar en un entorn on es coneix el generador (per exemple, per detectar ús d'un LLM concret en treballs acadèmics), el DistilBERT ofereix la millor calibració. Si el sistema ha de ser robust davant d'estudiants que intentin humanitzar el text generat, el SVM és més adequat. Si el sistema ha d'adaptar-se a nous LLMs que apareguin al mercat, el pipeline de reentrenament continu amb el SVM és la solució més pràctica.

TAULA 4: INTERPRETABILITAT DEL MODEL SVM: PATRONS AMB MÉS PES SEGONS LA CLASSE.

Text LLM	**Text humà**
Connectors formals: *however*, *moreover*, *furthermore*.	Expressions personals i espontànies, menys sistemàtiques.
Formulacions conclusives: *in conclusion*, *to sum up*.	Variabilitat en l'inici i tancament d'arguments.
Estructura argumentativa molt regular i previsible.	Estructures irregulars, canvis de ritme, repeticions.
Lèxic neutre i formal, sense col·loquialismes.	Marques d'estil individual i expressions quotidianes.
Sintàxi estable i puntuació homogènia.	Variació en longitud de frase i construcció sintàctica.

7 CONCLUSIONS

Aquesta secció recull les conclusions del treball organitzades en quatre blocs. En primer lloc, es descriuen les aportacions principals i els resultats més rellevants obtinguts. A continuació, es contextualitzen aquests resultats en relació amb la literatura existent i es justifica l'ús dels recursos computacionals. Finalment, s'identifiquen les limitacions del treball i es proposen les línies de treball futur més prometedores.

7.1 Aportacions principals

El treball ha realitzat sis aportacions concretes i mesurables. Primera, s'ha construït un pipeline complet i reproduïble d'augmentació de dades sintètiques que combina generació local amb Ollama (Mistral i LLaMA3) i l'API de Gemini 2.5 de Google, sense dependència de datasets privats ni infraestructura costosa, amb filtres de qualitat automàtics i un disseny extensible a nous generadors. La reproduïbilitat del pipeline és un actiu important: permet replicar exactament el conjunt d'entrenament i incorporar nous generadors de forma sistemàtica.

Segona, s'ha demostrat empíricament que la combinació de TF-IDF de caràcters i paraules (representació dual de 80.000 dimensions) amb un SVM lineal supera el baseline de Regressió Logística en un 35% en LogLoss en validació cross-prompt, i que XGBoost no és adequat per a representacions TF-IDF d'alta dimensió (LogLoss 2,119 vs. 0,500 del SVM). Aquests resultats proporcionen evidència quantitativa per a decisions de disseny en projectes similars i demostren que el salt de la Regressió Logística al SVM és justificat i mesurable.

Tercera, el fine-tuning de DistilBERT ha demostrat que els Transformers calibren molt millor les probabilitats en dominis tancats (LogLoss 0,100 vs. 0,500 del SVM), però presenten menor robustesa davant de l'evasió estilística dels *rewrites* (LogLoss 0,602 vs. 0,490 del SVM). Això identifica un *trade-off* important i no trivial entre calibració i robustesa que la literatura no havia documentat de forma tan directa en el context de la detecció de text generat per IA en assaigs d'estudiants.

Quarta, l'experiment de generalització amb Gemini ha quantificat la limitació central de qualsevol detector supervisat: la dependència del generador causa una degradació severa quan apareix un model nou no vist durant l'entrenament, limitació independent de l'arquitectura del classificador (afecta tant al SVM com al DistilBERT). Aquesta limitació és inherent al paradigma supervisat i no és resoluble amb models més complexos, sinó només amb dades més diverses.

Cinquena, s'ha demostrat que amb tan sols 100 assaigs addicionals del nou generador (50 per tema de redacció) el sistema recupera completament el rendiment (de LogLoss 1,252 a 0,020 sobre Gemini, una reducció del 98%), validant la viabilitat d'un pipeline de reentrenament continu com a estratègia sostenible. La rapidesa de la recuperació —que requereix menys d'un minut de reentrenament del SVM en CPU— fa que aquest enfocament sigui pràcticament viable en entorns reals on nous LLMs apareixen constantment.

Sisena, l'anàlisi d'interpretabilitat dels coeficients del SVM v3 (Secció 6.5) ha identificat els patrons lingüístics que el model utilitza com a senyal discriminatiu. Els resultats confirmen que el detector apren característiques estilístiques genuïnes i interpretables: els LLMs s'identifiquen per l'estructura argumentativa formal (*argue that*, *in conclusion*), el vocabulari consistent i elevat (*ensures*, *mandatory*), i les signatures de carta (*sincerely*, *your name*); els assaigs humans, per l'escriptura col·loquial directa (*people*, *so*, *because*) i la variabilitat sintàctica. Aquesta anàlisi aporta valor tant acadèmic —permet entendre què diferencia lingüísticament el text humà del generat per IA— com pràctic —permet detectar possibles dreceres d'aprenentatge (*shortcuts*) lligades als temes de redacció del dataset.

7.2 Relació amb la literatura i justificació dels recursos

La superioritat del SVM sobre XGBoost en espais TF-IDF d'alta dimensió és consistent amb les observacions de [8] sobre l'eficàcia dels classificadors lineals en tasques de detecció de text. La degradació del rendiment davant de generadors no vistos és consistent amb [10], que documenta el fenòmen de dependència del generador en múltiples arquitectures de detecció; la novetat del present treball és quan-

tificar la degradació de forma precisa (98% en LogLoss) i demostrar que 100 mostres són suficients per recuperar el rendiment, un resultat que fins on coneixem no havia estat documentat prèviament amb aquesta precisió. La fragilitat del DistilBERT davant dels *rewrites* és consistent amb el fenòmen de *shortcut learning* documentat a [11]: els Transformers aprenen dreceres semàntiques que fallen quan l'estil s'humanitza.

Des del punt de vista dels recursos computacionals, els resultats obtinguts estan en línia amb el que seria esperable d'un enginyer en dades sense accés a infraestructura dedicada. El SVM s'entrena en CPU local en menys d'un minut, el pipeline d'augmentació amb Ollama funciona sense connexió a serveis externs, i el fine-tuning de DistilBERT aprofita la GPU gratüïta de Google Colab (Tesla T4) durant menys de 5 minuts. Cap dels experiments requereix més de 16GB de RAM ni GPU dedicada. Això confirma que el projecte és completament assolible en l'entorn d'un TFG amb recursos limitats i que els resultats són reproduïbles sense cost econòmic significatiu.

7.3 Limitacions

El dataset conté només dos temes de redacció en anglès, la qual cosa limita la generalització a altres contextos temàtics, idiomes i nivells educatius. Els resultats són específics al gènere dels assaigs argumentatius d'estudiants de secundària anglosaxona, i no és evident que es mantinguin en altres gèneres textuals com els textos científics, els correus electrònics o els missatges a xarxes socials. L'anàlisi d'interpretabilitat confirma que alguns n-grames altament discriminatius són temàtics en lloc d'estilístics (lligats als dos temes de redacció del dataset), la qual cosa limita la generalitzabilitat dels patrons identificats.

L'experiment de generalització s'ha limitat a un sol generador addicional (Gemini); un estudi amb múltiples generadors de famílies distintes (GPT-4o, Claude, Llama-3.1, etc.) proporcionaria una estimació més robusta de la corba d'aprenentatge del detector i permetria determinar si 100 mostres són suficients independentment de la distància estilística del nou generador respecte als vistos. Finalment, el DistilBERT s'ha entrenat sobre el train v2 (Mistral + LLaMA3); no s'ha verificat si entrenar-lo també sobre textos de Gemini milloraria la seva calibració en domini tancat o la robustesa davant de l'evasió estilística, tancant el cicle comparatiu complet entre les dues arquitectures.

7.4 Treball futur

Els resultats d'aquest treball obren diverses línies d'investigació. La més directa és entrenar el DistilBERT sobre el train v3 (Mistral + LLaMA3 + Gemini) per verificar si la diversitat de generadors millora també la calibració i la robustesa del Transformer, tancant el cicle comparatiu complet entre les dues arquitectures i determinant si el *trade-off* calibració/robustesa persisteix en condicions de major diversitat de dades.

Una segona línia és ampliar l'experiment de generalització a múltiples generadors de famílies distintes (GPT-4o, Claude, Llama-3.1) per obtenir una estimació més robusta de la corba d'aprenentatge del detector i determinar si el nombre mínim de mostres necessari és consistent entre generadors o depèn de la distància estilística respecte als generadors vistos. Una tercera línia és avaluar l'impacte d'incorporar els *rewrites* al conjunt d'entrenament, per estudiar si el detector pot aprendre a reconèixer el text sintètic humanitzat sense pèrdua de rendiment en els escenaris estàndard. Finalment, l'extensió a altres gèneros textuals i idiomes permetria avaluar la generalitzabilitat del pipeline en contextos educatius més amplis.

ÚS DE LA INTEL·LIGÈNCIA ARTIFICIAL

En el desenvolupament d'aquest treball s'han utilitzat les eines d'intel·ligència artificial següents:

Claude Code (Anthropic). S'ha utilitzat com a assistent de programació integrat a l'entorn de desenvolupament local (VS Code) per reforçar i depurar parts del codi Python, especialment en la implementació dels pipelines de TF-IDF, la calibració del SVM i els scripts d'augmentació sintètica.

ChatGPT (OpenAI). S'ha consultat com a eina de suport bibliogràfic i de comprensió conceptual, particularment per aprofundir en el funcionament de tècniques de detecció basades en la probabilitat del model generador com DetectGPT [7].

Gemini 2.5 Flash (Google). S'ha emprat com a model generador de textos sintètics per a l'experiment de generalització descrit a la Secció 5.5, accedint-hi via l'API de Google AI Studio amb la biblioteca `google-genai`.

AGRAÏMENTS

Vull expressar el meu agraïment més sincer a la meva família i a totes les persones que m'han acompanyat durant aquests anys de carrera. El seu suport constant ha estat fonamental per poder afrontar els reptes del grau i arribar fins aquí amb èxit.

Als meus companys i companyes, gràcies per fer més lleugera cada etapa, per les hores compartides i per ajudar-me a créixer tant acadèmicament com personalment. Als meus professors i professores, gràcies pel coneixement transmès, pel suport rebut i per guiar-me al llarg d'aquest recorregut.

BIBLIOGRAFIA I REFERÈNCIES

[1] Wikipedia, "Kanban (desarrollo)." [En línia]. Disponible a: `https://es.wikipedia.org/wiki/Kanban_(desarrollo)`. [Accedit: 10-abr-2026].

[2] Python Software Foundation, "Python 3 Documentation." [En línia]. Disponible a: `https://docs.python.org/3/`. [Accedit: 15-abr-2026].

[3] Scikit-learn developers, "`TfidfVectorizer` — scikit-learn 1.4 documentation." [En línia]. Disponible a: `https://scikit-learn.org/stable/modules/generated/sklearn.feature_`

extraction.text.TfidfVectorizer.html. [Accedit: 24-mar-2026].

[4] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proc. EMNLP 2020: System Demonstrations*, pp. 38–45, 2020. [En línia]. Disponible a: https://aclanthology.org/2020.emnlp-demos.6. [Accedit: 15-abr-2026].

[5] Ollama, "Ollama – Run Large Language Models Locally." [En línia]. Disponible a: https://ollama.com/. [Accedit: 20-mar-2026].

[6] Google DeepMind, "Gemini API Documentation." [En línia]. Disponible a: https://ai.google.dev/gemini-api/docs. [Accedit: 10-mai-2026].

[7] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning i C. Finn, "DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature," in *Proc. ICML 2023*, 2023. [En línia]. Disponible a: https://arxiv.org/abs/2301.11305. [Accedit: 20-mar-2026].

[8] E. Crothers, N. Japkowicz i H. Viktor, "Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods," *IEEE Access*, vol. 11, pp. 70977–71002, 2023. [En línia]. Disponible a: https://arxiv.org/abs/2210.07321. [Accedit: 25-mar-2026].

[9] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv:1907.11692*, 2019. [En línia]. Disponible a: https://arxiv.org/abs/1907.11692. [Accedit: 25-mar-2026].

[10] A. Uchendu, T. Le, K. Shu i D. Lee, "TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation," in *Proc. EMNLP 2021 Findings*, pp. 2001–2016, 2021. [En línia]. Disponible a: https://arxiv.org/abs/2109.13296. [Accedit: 25-mar-2026].

[11] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge i F. A. Wichmann, "Shortcut Learning in Deep Neural Networks," *Nature Machine Intelligence*, vol. 2, pp. 665–673, 2020. [En línia]. Disponible a: https://arxiv.org/abs/2004.07780. [Accedit: 25-mar-2026].

[12] Kaggle, "LLM – Detect AI Generated Text (Competition)." [En línia]. Disponible a: https://www.kaggle.com/competitions/llm-detect-ai-generated-text. [Accedit: 23-feb-2026].

[13] A. Q. Jiang *et al.*, "Mistral 7B," *arXiv:2310.06825*, 2023. [En línia]. Disponible a: https://arxiv.org/abs/2310.06825. [Accedit: 20-mar-2026].

[14] A. Dubey *et al.*, "The Llama 3 Herd of Models," *arXiv:2407.21783*, 2024. [En línia]. Disponible a: https://arxiv.org/abs/2407.21783. [Accedit: 20-mar-2026].

[15] V. Sanh, L. Debut, J. Chaumond i T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv:1910.01108*, 2019. [En línia]. Disponible a: https://arxiv.org/abs/1910.01108. [Accedit: 15-abr-2026].

ANNEX

A. Diagrama de Gantt

La Fig. 1 mostra la planificació temporal del TFG d'acord amb les fites del calendari oficial i les tasques principals identificades.

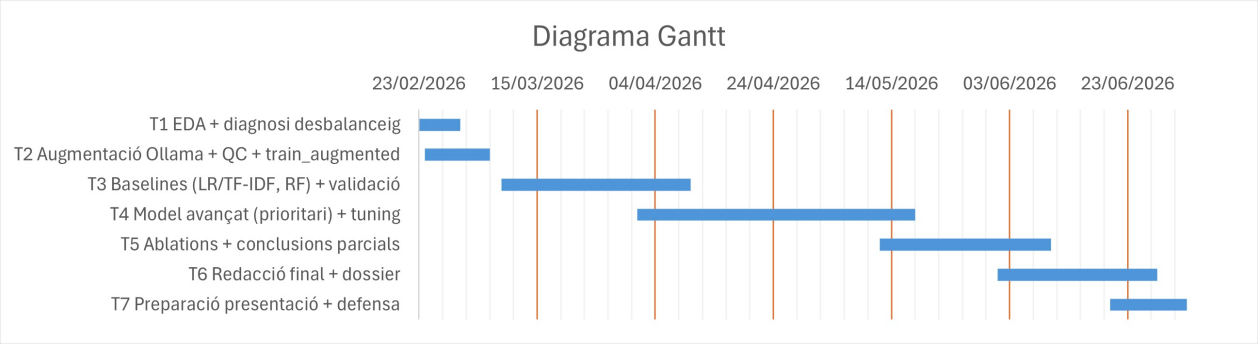


Fig. 1: Diagrama de Gantt i fites del projecte. Les línies verticals taronges marquen les dates oficials d'entrega.

B. Glossari de termes tècnics

TF-IDF (*Term Frequency–Inverse Document Frequency*): tècnica de ponderació que transforma textos en vectors numèrics mesurant la importància relativa de cada terme en el context del dataset.

N-grames de caràcters: seqüències de n caràcters consecutius extrets d'un text. Capturen patrons morfològics i estilístics a nivell sublèxic independents del contingut temàtic.

SVM (*Support Vector Machine*): classificador que troba l'hiperpla de màxim marge que separa dues classes en un espai d'alta dimensió. Especialment adequat per a representacions TF-IDF disperses.

Fine-tuning: procés d'adaptar un model de llenguatge preentrenat (com DistilBERT) a una tasca específica continuant l'entrenament sobre un dataset etiquetat petit durant poques èpoques.

Cross-prompt: protocol de validació que entrena el model sobre un tema de redacció i el valida sobre un tema diferent, avaluant la generalització temàtica i evitant el *data leakage* per contingut.

Rewrite: reescriptura d'un text generat per un LLM amb instruccions explícites d'imitar l'estil d'un estudiant humà, representant el cas d'evasió deliberada del detector.

AUC (*Area Under the ROC Curve*): mètrica que mesura la capacitat discriminativa del model independent del llindar de decisió. AUC = 1 indica discriminació perfecta.

LogLoss (*Logarithmic Loss*): mètrica que penalitza la manca de calibració de les probabilitats predites. A diferència de l'AUC, penalitza els errors de confiança, no només els de classificació.

Ollama: eina de codi obert que permet executar models de llenguatge localment sense connexió a serveis externs, garantint la privacitat de les dades.

Generalització: capacitat del model de mantenir el rendiment sobre dades que no ha vist durant l'entrenament, ja sigui per tema (*cross-prompt*) o per generador (textos de Gemini no vistos).