

**UNIVERSIDAD EUROPEA DE MADRID**

**ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO**

**MÁSTER EN DATA SCIENCE**

ANTEPROYECTO

DETECCIÓN DE IA GENERATIVA

**DAVID DÍAZ MORIANO**

**CURSO 2024-2025**

**TÍTULO**: DETECCIÓN DE IA GENERATIVA

**AUTOR**: DAVID DÍAZ MORIANO

**TITULACIÓN**: MÁSTER EN DATA SCIENCE

**DIRECTOR DEL PROYECTO**: JAVIER VILLAR GIL

**FECHA**: Abril de 2025

Índice

[**1. JUSTIFICACIÓN 5**](#_z6qmcxch3xy8)

[**2. OBJETIVOS DEL PROYECTO 5**](#_h4xhpkykpdfv)

[2.1. Objetivo General 5](#_wr3dnrxnmh2d)

[2.2. Objetivos Específicos 5](#_77h6dz78t6bh)

[**3. MODALIDADES DEL PROYECTO 5**](#_qki1m76kt88b)

[3.1. Texto 5](#_txr86h1n3920)

[3.1.1. Desarrollo previsto: pipeline para detección de texto generado por IA 6](#_wlh40vijzex)

[3.2. Imágenes 7](#_xeb8cc7dyrp8)

[3.2.1. Desarrollo previsto: pipeline para detección de imágenes generadas por IA 7](#_h657kxk92l7r)

[3.3. Video 8](#_aoptf1ures1v)

[3.3.1. Desarrollo previsto: pipeline para detección de video generado por IA 9](#_pqc186o502r6)

[**4. HERRAMIENTAS Y TECNOLOGÍAS 9**](#_rtbdax2k73je)

[**5. DATASETS TENTATIVOS 10**](#_wo6tviva3ntl)

[**6. PLAN DE PROYECTO 10**](#_36w4hctmpq32)

[**7. RESULTADOS ESPERADOS 11**](#_8hpblsqq1d7d)

[**8. RIESGOS Y ALTERNATIVAS 11**](#_qkm9bgcktymr)

[**9. REFERENCIAS BIBLIOGRÁFICAS 12**](#_7oqsct53m7vf)

# JUSTIFICACIÓN

En la actualidad, la inteligencia artificial generativa ha alcanzado niveles de sofisticación que permiten la creación de contenido digital (texto, imágenes y video) prácticamente indistinguible del contenido producido por seres humanos. Este fenómeno presenta oportunidades, pero también riesgos significativos en ámbitos como la educación, la seguridad, el periodismo y la política. Ante este escenario, la detección precisa de contenido generado por IA se vuelve una necesidad crítica.

# OBJETIVOS DEL PROYECTO

## Objetivo General

Desarrollar un sistema de análisis que permita detectar si un contenido digital ha sido generado por un humano o por una inteligencia artificial, abarcando tres modalidades principales: texto, imagen y video.

## Objetivos Específicos

* Investigar y analizar los patrones característicos del contenido generado por IA en texto, imágenes y video.
* Identificar datasets adecuados para entrenar y validar los modelos de detección en cada modalidad.
* Implementar modelos de machine learning y deep learning para la detección por tipo de contenido.
* Evaluar la eficacia de cada modelo con métricas de precisión, recall y F1-score.
* Explorar técnicas adversariales que busquen engañar al sistema de detección y analizar su resiliencia.
* Evaluar si el contenido generado por IA supera el Test de Turing, y en qué condiciones lo logra.

# MODALIDADES DEL PROYECTO

## Texto

El primer bloque de trabajo se centrará en la detección de texto generado por IA. Este tipo de contenido es el más accesible y cuenta con abundante material para su análisis. Se recopilarán textos generados por modelos como GPT-4, Claude o Gemini, así como textos humanos extraídos de fuentes como Wikipedia, artículos periodísticos o publicaciones de foros.

Se planteará un enfoque comparativo entre técnicas tradicionales de análisis lingüístico (como n-gramas, frecuencia de palabras, puntuación y coherencia léxica) y métodos basados en aprendizaje automático.

Inicialmente se entrenarán modelos clásicos de clasificación como Regresión Logística, Árboles de Decisión o Random Forest, aplicando ingeniería de características a partir de propiedades estadísticas del texto.

Posteriormente, se incorporarán modelos de aprendizaje profundo, especialmente basados en transformers, como BERT, RoBERTa o DetectGPT. Se evaluará su capacidad para distinguir entre texto humano y texto IA sin necesidad de ajuste fino del modelo (zero-shot), y también mediante fine-tuning sobre el dataset recopilado.

La evaluación de resultados se realizará mediante métricas como precisión, recall, F1-score y matriz de confusión. Se buscará identificar patrones comunes en los errores del modelo (falsos positivos y falsos negativos), y se explorarán estrategias para aumentar la robustez del sistema, como la incorporación de ruido o reformulación adversarial del texto.

Además, se analizará si el texto generado por IA es capaz de superar el Test de Turing, ya sea frente a un evaluador humano o mediante criterios formales de indistinguibilidad. Esto permitirá contextualizar los resultados del modelo de detección en términos de percepción humana.

### Desarrollo previsto: pipeline para detección de texto generado por IA

El desarrollo del sistema de detección de texto generado por IA se estructurará en las siguientes etapas:

**Recolección y preparación del dataset**

* Recopilación de textos generados por humanos (Wikipedia, artículos de prensa, foros) y por IA (ChatGPT, Claude, Gemini).
* Limpieza y normalización del texto: eliminación de ruido, estandarización de formato, corrección de codificación.
* Etiquetado binario del dataset (texto humano vs. IA).

**Análisis exploratorio y extracción de características**

* Cálculo de métricas lingüísticas como longitud media de frases, frecuencia léxica, diversidad de vocabulario, puntuación.
* Uso de técnicas clásicas como n-gramas y TF-IDF.
* Visualización de patrones mediante histogramas, boxplots y nubes de palabras.

**Entrenamiento de modelos base**

* Modelos clásicos: Regresión Logística, Random Forest, SVM.
* Evaluación inicial con validación cruzada.

**Modelos avanzados y fine-tuning**

* Aplicación de modelos basados en transformers (BERT, RoBERTa, DetectGPT).
* Fine-tuning con el dataset preparado o uso en modo zero-shot.
* Comparación de resultados frente a modelos clásicos.

**Evaluación del sistema**

* Cálculo de métricas: precisión, recall, F1-score, matriz de confusión.
* Análisis de errores (falsos positivos y falsos negativos).

**Robustez y pruebas adversariales**

* Modificación de textos IA para observar si el sistema puede ser engañado (paráfrasis, errores gramaticales intencionales).
* Estudio de resiliencia del modelo frente a perturbaciones simples.

**Análisis perceptivo (Test de Turing)**

* Revisión manual de muestras para valorar si un humano detectaría la diferencia.
* Comparación de resultados humanos vs. predicciones del modelo.

## Imágenes

El segundo bloque del proyecto estará centrado en la detección de imágenes generadas por IA, especialmente aquellas creadas con modelos como DALL·E, MidJourney o Stable Diffusion. Estos generadores utilizan redes generativas adversariales (GANs) o modelos de difusión para producir imágenes realistas, lo que plantea un reto creciente para su detección automática.

La estrategia comenzará con la recolección de datasets mixtos, incluyendo imágenes reales (por ejemplo, del dataset COCO disponible en Kaggle) y generadas artificialmente mediante diferentes plataformas. Se analizarán patrones característicos de generación IA, como artefactos visuales, inconsistencias en simetría, anatomía o texturas, así como ruido residual propio del proceso de generación.

Para la detección se aplicarán modelos de visión por computadora, con énfasis en redes convolucionales profundas (CNNs). Inicialmente se probarán modelos clásicos como VGG o ResNet, y posteriormente modelos más especializados como XceptionNet o EfficientNet, preentrenados y adaptados mediante fine-tuning.

Se evaluarán técnicas complementarias como análisis de huellas digitales de GANs (GAN fingerprints) y mapas de calor para entender qué zonas de la imagen influyen más en la clasificación.

El sistema será evaluado mediante métricas clásicas de clasificación (precisión, recall, F1-score) y se complementará con un estudio perceptivo: ¿hasta qué punto una imagen generada por IA puede pasar por auténtica a ojos de una persona? Esto se relacionará directamente con el Test de Turing visual.

### Desarrollo previsto: pipeline para detección de imágenes generadas por IA

El desarrollo del sistema de detección de imágenes generadas por IA se estructurará en las siguientes etapas:

**Recolección y preparación del dataset**

* Obtención de imágenes reales del dataset COCO (disponible en Kaggle).
* Generación de imágenes artificiales con Stable Diffusion, DALL·E y MidJourney.
* Redimensionado, normalización y etiquetado binario (real vs. IA).

**Preprocesamiento y extracción de características**

* Conversión de imágenes a tensores.
* Análisis de artefactos visuales, bordes, y ruido.
* Visualización de diferencias mediante mapas de calor.

**Entrenamiento de modelos base**

* Modelos CNN clásicos: VGG, ResNet.
* Validación cruzada inicial para establecer baseline.

**Modelos avanzados y fine-tuning**

* Aplicación de XceptionNet y EfficientNet (vía TensorFlow/Keras).
* Ajuste fino con los datos disponibles para mejorar precisión.

**Evaluación del sistema**

* Métricas: precisión, recall, F1-score, matriz de confusión.
* Comparación de modelos y análisis de errores comunes.

**Robustez y evaluación perceptiva**

* Comprobación manual de imágenes IA que resultan difíciles de detectar.
* Relación con el Test de Turing visual (engañar al ojo humano).

## Video

El tercer bloque se centrará en la detección de contenido audiovisual manipulado por IA, con especial atención a los deepfakes. Estos videos, generados mediante técnicas como el face swapping, el lip-syncing o la síntesis completa del rostro, representan uno de los desafíos más complejos por el nivel de realismo alcanzado.

El trabajo se basará en el análisis de videos reales y manipulados disponibles en datasets como FaceForensics++, DFDC (Deepfake Detection Challenge) o Celeb-DF. Estos pueden encontrarse en plataformas como Kaggle o Papers with Code. La estrategia consistirá en extraer frames individuales de los videos para su análisis cuadro a cuadro, detectando señales de manipulación en la textura de la piel, movimientos oculares, desincronización labial, y artefactos de compresión.

Para ello, se utilizarán redes neuronales convolucionales (CNNs) y modelos preentrenados para clasificación binaria (auténtico/falso), tales como XceptionNet, EfficientNet, o incluso soluciones basadas en arquitectura 3D-CNN si los recursos lo permiten.

Complementariamente, se aplicarán técnicas de detección basadas en landmark facial tracking y análisis de expresividad anómala usando herramientas como Dlib o MediaPipe.

La evaluación se realizará mediante métricas clásicas, pero también considerando un componente perceptivo: ¿es el deepfake lo suficientemente realista como para engañar a un observador humano? Esta dimensión se alineará con el Test de Turing audiovisual.

En caso de limitaciones de tiempo o recursos, se priorizará el análisis en imágenes clave extraídas de los videos, para mantener la cobertura sin sobrecargar el proyecto.

### Desarrollo previsto: pipeline para detección de video generado por IA

El desarrollo del sistema de detección de vídeo generado por IA se estructurará en las siguientes etapas:

**Recolección y segmentación del dataset**

* Descarga de videos desde FaceForensics++, DFDC o Celeb-DF (Kaggle, Google Drive oficial).
* Segmentación en cuadros por segundo (frames) para análisis individual.

**Preprocesamiento y anotación**

* Detección de rostro en los frames (OpenCV, Dlib).
* Etiquetado binario del frame según procedencia (real o deepfake).

**Modelado base con imágenes**

* Uso de CNNs entrenadas sobre los frames extraídos.
* Aplicación de modelos como XceptionNet para detección visual.

**Modelos avanzados (opcional)**

* Exploración de modelos temporales (3D-CNN, LSTM) si el tiempo y recursos lo permiten.

**Evaluación de resultados**

* Precisión, recall, F1-score, matriz de confusión.
* Estudio de falsos positivos y negativos.

**Análisis perceptivo y Test de Turing audiovisual**

* Valoración de realismo de los deepfakes.
* Comparativa entre percepción humana y predicciones del sistema.

# HERRAMIENTAS Y TECNOLOGÍAS

En principio, las tecnologías que se utilizarán para el proyecto, serán las listadas a continuación:

* Lenguaje: Python
* Librerías: Scikit-learn, Transformers (Hugging Face), TensorFlow, PyTorch, OpenCV
* Modelos: BERT, DetectGPT, XceptionNet, EfficientNet
* Procesamiento de audio/video: FFmpeg, Dlib, MediaPipe (si aplica)

# DATASETS TENTATIVOS

* Texto:
  + Wikipedia (texto humano)
  + Textos generados con GPT-4, Claude, Gemini
* Imágenes:
  + COCO dataset (imágenes reales)
  + Imágenes generadas con Stable Diffusion, DALL·E, MidJourney
* Video:
  + FaceForensics++, DFDC, Celeb-DF (videos reales y deepfakes)

# PLAN DE PROYECTO

Esta sería la planificación inicial del proyecto, que dividiremos en 3 fases:

**Fase 1: Documentación y Diseño** (**16 marzo – 13 abril**)

* Definición del problema y objetivos
* Estudio de técnicas y datasets
* Planificación detallada de fases del proyecto

**Fase 2: Desarrollo y Evaluación** (**14 abril – 30 mayo**)

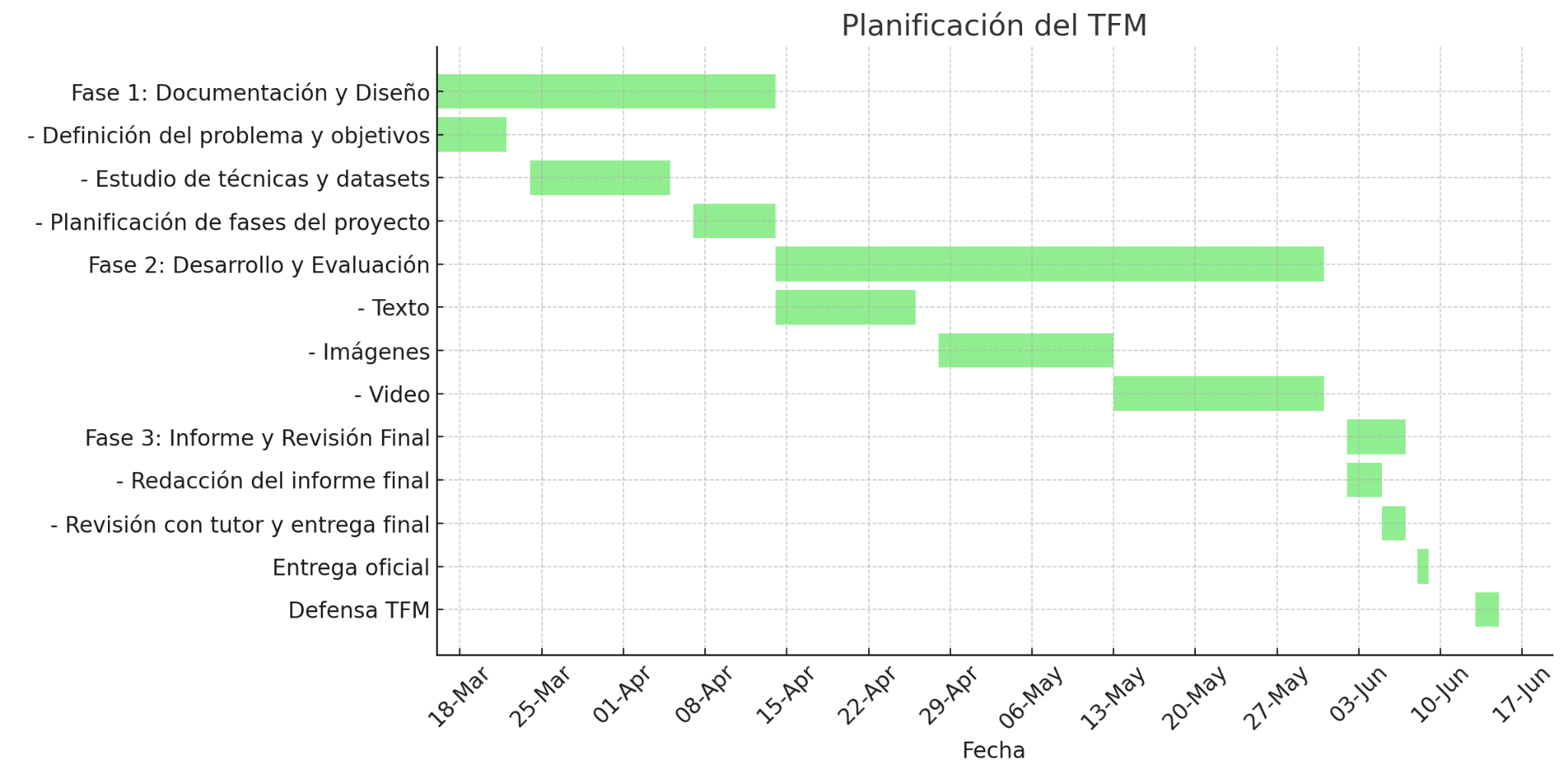
* **Texto**: 14 abril – 25 abril
* **Imágenes**: 28 abril – 12 mayo
* **Vídeo**: 13 mayo – 30 mayo

**Fase 3: Informe y Revisión Final** (**2 junio – 6 junio**)

* Redacción del informe final
* Revisión con el tutor y ajustes para la entrega definitiva

**Entrega definitiva**: 8 junio  
**Defensa del TFM**: 13–14 junio

A continuación se muestra el diagrama de Gantt con la planificación detallada del proyecto, incluyendo las fechas estimadas para cada fase y subfase:



# RESULTADOS ESPERADOS

* Modelos funcionales para cada tipo de contenido con métricas de evaluación.
* Comparativa entre modalidades y análisis de errores.
* Análisis del grado en que los contenidos generados por IA logran superar el Test de Turing.
* Documento técnico y presentación del TFM.

# RIESGOS Y ALTERNATIVAS

* Carga de trabajo elevada por la cobertura multimodal. Se priorizará texto e imagen si el tiempo es limitado.
* Las prácticas externas podrían interferir con el tiempo disponible; los fines de semana actuarán como margen de compensación.

# REFERENCIAS BIBLIOGRÁFICAS

Franganillo, J. (2022). *Contenido generado por inteligencia artificial: oportunidades y amenazas*.<https://franganillo.es/ia.pdf>​[Jorge Franganillo](https://franganillo.es/ia.pdf?utm_source=chatgpt.com)

ISMS Forum. (2024). *Deepfakes: Riesgos, casos reales y desafíos en la era digital*.<https://www.ismsforum.es/ficheros/descargas/deepfake-final1742458135.pdf>​[Foro ISMS](https://www.ismsforum.es/ficheros/descargas/deepfake-final1742458135.pdf?utm_source=chatgpt.com)

WeLiveSecurity. (2024). *Deepfakes en la era de la Inteligencia Artificial*.<https://www.welivesecurity.com/es/seguridad-digital/deepfakes-era-de-la-inteligencia-artificial/>​[ESET Noticias](https://www.welivesecurity.com/es/seguridad-digital/deepfakes-era-de-la-inteligencia-artificial/?utm_source=chatgpt.com)

iProov. (2023). *Deepfakes: Estadísticas y soluciones*.<https://www.iproov.com/es/blog/deepfakes-statistics-solutions-biometric-protection>​[iProov](https://www.iproov.com/es/blog/deepfakes-statistics-solutions-biometric-protection?utm_source=chatgpt.com)

Arte y Estética. (2023). *Arte generado por Inteligencia Artificial y su impacto en artistas*.<https://arteesetica.org/arte-generado-por-inteligencia-artificial-y-su-impacto-en-artistas/>​[Arte es Ética](https://arteesetica.org/arte-generado-por-inteligencia-artificial-y-su-impacto-en-artistas/?utm_source=chatgpt.com)

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.<https://doi.org/10.18653/v1/N19-1423>​[ar5iv+3aclanthology.org+3arXiv+3](https://aclanthology.org/N19-1423/?utm_source=chatgpt.com)

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. *Proceedings of the 40th International Conference on Machine Learning*.<https://dl.acm.org/doi/10.5555/3618408.3619446>​[ACM Digital Library](https://dl.acm.org/doi/10.5555/3618408.3619446?utm_source=chatgpt.com)

Bao, G., Zhao, Y., Teng, Z., Yang, L., & Zhang, Y. (2024). Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. *Proceedings of the International Conference on Learning Representations (ICLR)*.<https://openreview.net/forum?id=Bpcgcr8E8Z>​[OpenReview+1arXiv+1](https://openreview.net/forum?id=Bpcgcr8E8Z&utm_source=chatgpt.com)

Scikit-learn developers. (n.d.). *scikit-learn: Machine Learning in Python*.<https://scikit-learn.org/>​

Hugging Face. (n.d.). *Transformers Documentation*.<https://huggingface.co/docs/transformers>​

Hugging Face. (n.d.). *Scikit-learn Models on Hugging Face Hub*.<https://skops.readthedocs.io/en/stable/hf_hub.html>​[Hugging Face+3skops.readthedocs.io+3scikit-learn Blog+3](https://skops.readthedocs.io/en/stable/hf_hub.html?utm_source=chatgpt.com)