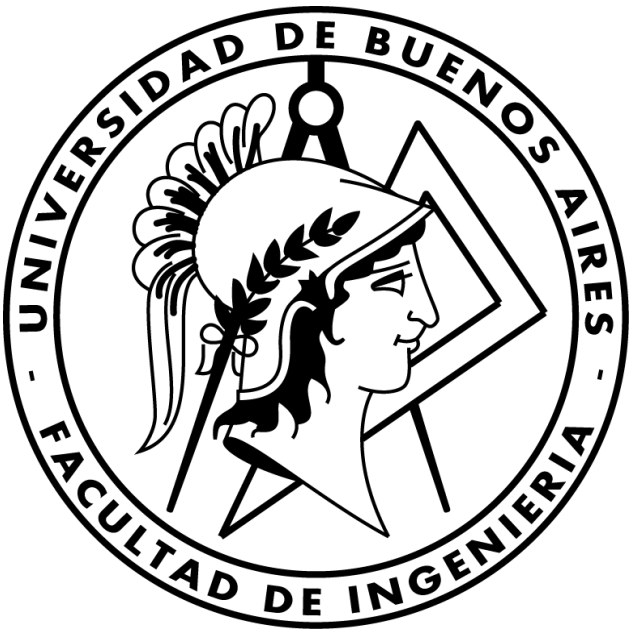


Facultad de Ingeniería, Universidad de Buenos Aires

Aprendizaje Automático

Docentes

Hernan Daniel Merlino

Alumnos

107239 - Ricardo Luis Contreras Villarroel - 95971093

104031 - Josefina Iterman - 39654113

106280 - Nazarena Rueda - 42103193

101043 - Ronchi Santiago Agustín - 40135719

107789 - Gonzalo Ebbes Tenor - 43974474

Fecha de entrega

26/06/2025

Informe: Detección de Ensayos en Inglés Generados por IA

Resumen

Este trabajo práctico desarrolla un sistema de **clasificación automática** capaz de distinguir entre **ensayos largos en inglés** escritos por humanos y aquellos generados por **inteligencia artificial**.

Para ello, se aplican técnicas de **procesamiento de lenguaje natural** y **aprendizaje automático**, utilizando como modelo principal un **clasificador Random Forest** entrenado sobre características **TF-IDF** extraídas de los textos.

El sistema alcanza una **precisión del 99%** en la detección de ensayos académicos en inglés, demostrando su efectividad en contextos educativos y de análisis de contenido.

Introducción

Objetivo del Proyecto

El objetivo principal es desarrollar un sistema capaz de identificar automáticamente si un texto ha sido generado por una IA o escrito por un ser humano. Esta capacidad es cada vez más relevante en el contexto actual donde los modelos de lenguaje generativo están proliferando.

Motivación

- **Detección de contenido generado por IA**: Necesidad de identificar contenido en plataformas educativas, medios de comunicación y redes sociales.
- **Aplicaciones educativas**: Prevenir el uso no autorizado de IA en tareas académicas, especialmente en el contexto de ensayos académicos en inglés.

Alcance del Modelo

El modelo está específicamente diseñado y entrenado para:

- **Idioma**: Ensayos en inglés únicamente
- **Tipo de contenido**: Ensayos académicos y medianamente formales
- **Longitud**: Múltiples oraciones

Limitaciones importantes: El modelo no está optimizado para textos cortos, muy informales o en otros idiomas por fuera del inglés.

Metodología

Dataset

- **Fuente**: Dataset "AI vs Human Text" de Kaggle (https://www.kaggle.com/datasets/shanegerami/ai-vs-human-text)
- **Contenido**: Ensayos escritos en inglés, tanto por humanos como generados por inteligencia artificial (principalmente modelos GPT)
- **Tamaño**: 487,235 ensayos
- **Distribución original**:
    - 37.24% ensayos generados por IA
    - 62.76% ensayos escritos por humanos
- **Muestreo balanceado**: Para evitar que el modelo aprenda un sesgo hacia la clase mayoritaria, se realizó un muestreo aleatorio que seleccionó **10,000 ejemplos de cada clase** (20,000 en total), asegurando una distribución equilibrada y un entrenamiento más justo.
- **Observaciones adicionales:**

- Los textos generados por IA provienen de distintas versiones de Chat GPT y GPT-3.
- Los ensayos humanos provienen de fuentes diversas, incluyendo foros, tareas académicas y publicaciones personales.
- No se aplicó filtrado por tema o estilo, lo que añade diversidad léxica y estilística al conjunto.

Preprocesamiento

1. Limpieza básica

Se aplicaron transformaciones mínimas sobre los textos originales para preservar su estilo natural. En particular, se realizó **eliminación de espacios múltiples**, reemplazandolos por un único espacio, y se eliminaron espacios en blanco al inicio y al final de cada texto.

No se modificaron otros aspectos del texto, como la puntuación, mayúsculas o minúsculas, ni se eliminaron stopwords.

2. División de datos

El dataset balanceado (20,000 ejemplos) se dividió en tres subconjuntos:

- **Entrenamiento:** 80% (16,000 ensayos)
- **Validación:** 10% (2,000 ensayos)
- **Test:** 10% (2,000 ensayos)

Esta división permite evaluar el rendimiento en datos no vistos durante el entrenamiento y ajustar hiperparámetros sin sobreajustar el conjunto de prueba.

3. Estratificación

Se aplicó estratificación en la división para mantener la proporción de clases (50% IA / 50% humano) en cada conjunto, garantizando una evaluación justa y equilibrada.

Extracción de Características

Para transformar los textos en datos numéricos utilizables por el modelo, se utilizó la técnica de **vectorización TF-IDF (Term Frequency - Inverse Document Frequency)**. Esta técnica permite representar cada ensayo como un vector en un

espacio de características, donde cada dimensión corresponde a una palabra o combinación de palabras del vocabulario.

Las configuraciones utilizadas fueron las siguientes:

- **Cantidad máxima de características:** Se limitaron los vectores a un máximo de **50,000 características**, seleccionando las más frecuentes e informativas, para mantener un buen balance entre riqueza representativa y eficiencia computacional.

- **N-gramas de 1 a 2 palabras:** Se consideraron **unigramas (palabras individuales)** y **bigramas (pares consecutivos de palabras)**. Esto permite capturar tanto el uso individual de términos como combinaciones frecuentes que aportan contexto, por ejemplo:

    - Unigrama: "technology"
    - Bigrama: "artificial intelligence"

- **Captura de patrones léxicos y sintácticos:** Esta configuración ayuda a identificar **estilos de escritura**, **estructuras repetitivas**, y otros **patrones lingüísticos** que suelen diferenciar el texto generado por humanos del generado por IA. Por ejemplo, los textos artificiales tienden a repetir ciertas frases estructuralmente similares o usar vocabulario más neutro y formal de forma consistente.

Modelo de Clasificación

- **Algoritmo**: Random Forest Classifier
- **Optimización**: Randomized Search con validación cruzada
- **Parámetros optimizados**:
    - `n_estimators`: 300
    - `max_depth`: None (sin límite)
    - `min_samples_split`: 5
    - `min_samples_leaf`: 1
    - `bootstrap`: False

Fundamentación de la elección del modelo

Se optó por utilizar un **modelo de Random Forest** debido a varias ventajas clave que lo hacen especialmente adecuado para este tipo de tarea de clasificación binaria en procesamiento de lenguaje natural:

- **Robustez frente al sobreajuste:** Al combinar múltiples árboles de decisión entrenados con distintos subconjuntos de datos y características, Random Forest tiende a generalizar mejor que un único modelo, especialmente en contextos con alto volumen de atributos como los vectores TF-IDF.

- **Capacidad para manejar grandes cantidades de características:** La vectorización TF-IDF utilizada en el proyecto generó hasta 50,000 n-gramas como características, lo cual puede dificultar el rendimiento de modelos lineales simples. Random Forest, en cambio, se adapta bien a este tipo de espacios de alta dimensionalidad.

- **Buen rendimiento sin necesidad de extensivo ajuste de hiperparámetros:** A diferencia de modelos más complejos como redes neuronales, Random Forest ofrece muy buen rendimiento inicial sin requerir una infraestructura pesada ni grandes volúmenes de datos para entrenamiento.

En pruebas comparativas preliminares, otros modelos como regresión logística o SVM mostraron buen desempeño, pero Random Forest logró mejores métricas en validación cruzada y fue más consistente en la clasificación de casos difíciles.

Resultados

Los siguientes resultados fueron obtenidos utilizando el **conjunto de test**, que representa una porción del dataset original reservada exclusivamente para evaluación final (10% del total, con distribución balanceada entre clases).

Rendimiento del Modelo

	precision	recall	f1-score	support
0.0	0.98	0.99	0.99	1000
1.0	0.99	0.98	0.99	1000
accuracy			0.99	2000
macro avg	0.99	0.99	0.99	2000
weighted avg	0.99	0.99	0.99	2000

Análisis de Resultados

- **Precisión global**: 99%
- **F1-score**: 0.99 para ambas clases
- **Balance**: Excelente equilibrio entre precisión y recall, lo que indica que el modelo no favorece una clase por sobre la otra
- **Robustez**: Resultados consistentes durante la validación cruzada en entrenamiento (F1 promedio: 0.982), lo que sugiere buena generalización

Casos de Prueba

Ejemplos de Essays Humanos (Correctamente Clasificados):

Essay 1 (fuente):

"I read a post on Askreddit that inquired, "what are some completely legal things that make you a terrible human being." Among the top comments were answers like, standing in the middle of an aisle, not flushing the toilet, or not putting things back where you found them in the grocery store. All these comments had 5 digit upvotes. Are these things somewhat inconsiderate and annoying? Yes. Do they make you a terrible person? No. I feel like these people who view these actions as "terrible" are just looking for any way to feel validated in their perceived moral superiority by barely doing anything at all. They think "I don't stand in people's way, I flush the toilet, I put grocery items back where I found them, I'm a good person and other people are bad". It honestly feels somewhat masturbatory. I think you can do inconsiderate things and even have committed crimes and still be a good person, if you were perfect you wouldn't be human. I can pick apart someone's every action and try to place them neatly into the categories of "good" and "bad" but it probably wouldn't make me any better as a person." → **Predicción: Humano (0.0)** ✓

Essay 2 (fuente):

"I used to rent, so I always used a TV stand. No mess, no stress. But when I finally bought my own place, I figured it was time to do it right and wall-mount the TV. That lasted about two weeks. The height felt off, I couldnt adjust anything without unbolting the whole thing, and dont even get me started on the cable situation. It looked good for about five minutes, then just became a pain. Ended up going back and re-ordering the same fitueyes tv stand I had in my rental. I had left it behind for the next tenant because I thought Id moved on from that phase. Turns out I hadnt. Sometimes the simple option is just... better. Now Im sitting here wondering why wall mounting became the "default" when its such a pain to do right and almost impossible to undo. Feels like we all just accepted it because it looks nice on Instagram. (even though the tv stand can looks nice too?)" → **Predicción: Humano (0.0)** ✓

Essay 3 (fuente):

"I dont know why everyone on Reddit whines and cries about carts not being returned at the supermarket. When I was 15 I had a supermarket job. When I was asked to do carts I was ecstatic, got to go outdoors, got to not deal with annoying customers while bagging. Got to chill with the weird 30 y/o produce guy who hung out under the awning smoking, Why would a kid nowadays want to stay indoors and bag or cashier and deal with annoying people? Just go outside and stretch your legs, maybe sneak over to the other stores in the plaza, whatever. Youre giving these kids something fun to do, whats wrong with that?" → **Predicción: Humano (0.0)** ✓

Ejemplos de Essays Generados por IA (Correctamente Clasificados):

Todos estos ejemplos fueron solicitados a Chat GPT.

Essay 4:

“In today's fast-paced and increasingly interconnected global society, the importance of leveraging technological advancements to optimize learning outcomes cannot be overstated. Artificial intelligence, in particular, provides a wide range of opportunities for enhancing educational efficiency, engagement, and accessibility across diverse populations. By integrating adaptive algorithms, data-driven insights, and scalable platforms, institutions can foster a more personalized and inclusive learning environment that aligns with 21st-century skills and evolving industry demands." → **Predicción: IA (1.0)** ✓

Essay 5:

"Learning a new language can be both exciting and challenging. Many people find that consistent practice and exposure to real conversations help improve their skills more effectively than just studying grammar rules. Using apps, watching movies, and

talking with native speakers are great ways to immerse yourself in the language. Remember, making mistakes is a natural part of the learning process, and patience is key to becoming fluent." → **Predicción: IA (1.0) ✓**

Essay 6:

"Traveling to new places is an amazing way to broaden your horizons and learn about different cultures. It allows you to meet new people, try unique foods, and experience traditions that you might never encounter otherwise. Whether it's a weekend getaway or a longer trip abroad, every journey has the potential to teach you something valuable and create lasting memories" → **Predicción: IA (1.0) ✓**

Essay 7:

"I didn't expect to enjoy cooking as much as I do now. It started as something I had to do just to get by, but over time, trying out new recipes became something I genuinely look forward to. It's relaxing, and there's a small sense of pride when a dish turns out well—especially when someone else enjoys it too." → **Predicción: IA (1.0) ✓**

Implementación del Sistema

El sistema incluye una **interfaz web sencilla e intuitiva** que permite a los usuarios ingresar un texto en inglés y obtener una predicción sobre si fue escrito por un humano o generado por inteligencia artificial. Esta interfaz está diseñada para funcionar en cualquier dispositivo y brinda una experiencia de uso clara y directa.

La comunicación entre el navegador y el modelo se realiza mediante una **API REST desarrollada con FastAPI**, que se encarga de procesar el texto y devolver el resultado.

Funcionalidad General

- El usuario accede a una **página web** con un formulario donde puede pegar o escribir un ensayo en inglés.
- Al hacer clic en el botón de análisis, el texto se envía al servidor mediante una petición HTTP.
- El servidor, utilizando **FastAPI**, recibe el texto, lo analiza con el modelo de machine learning entrenado (Random Forest + TF-IDF) y devuelve el resultado.
- El navegador muestra la predicción en pantalla de forma visual: **verde** si el texto fue escrito por un humano, **rojo** si fue generado por una IA.

Despliegue y Arquitectura

El sistema completo está preparado para ejecutarse en contenedores utilizando **Docker**, lo que permite una instalación rápida y sin configuraciones complejas. Todo se levanta con un único comando, incluyendo tanto el servidor como la interfaz web.

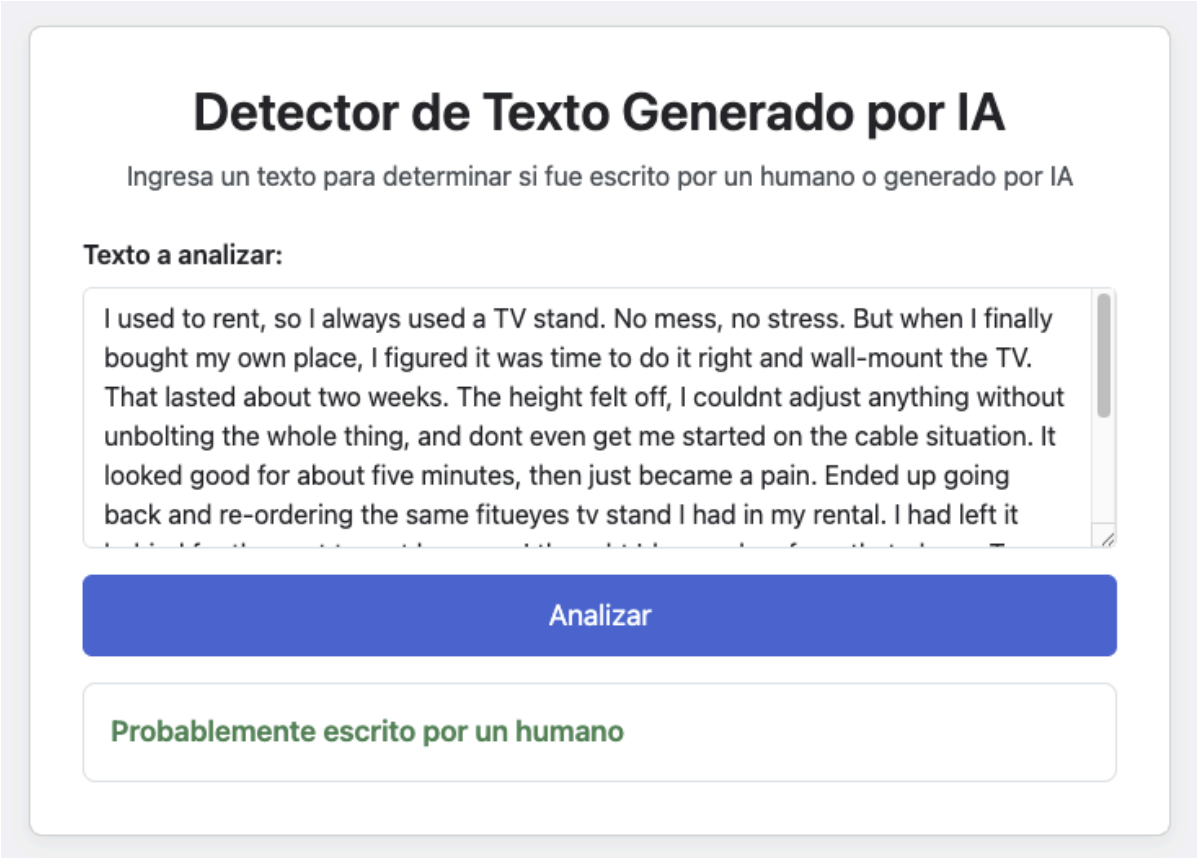
La arquitectura general sigue el patrón clásico **cliente-servidor**:

```
Cliente Web (index.html)
    ↓ HTTP/JSON
Servidor FastAPI (webserver.py)
    ↓ Modelo ML
Random Forest + TF-IDF
```

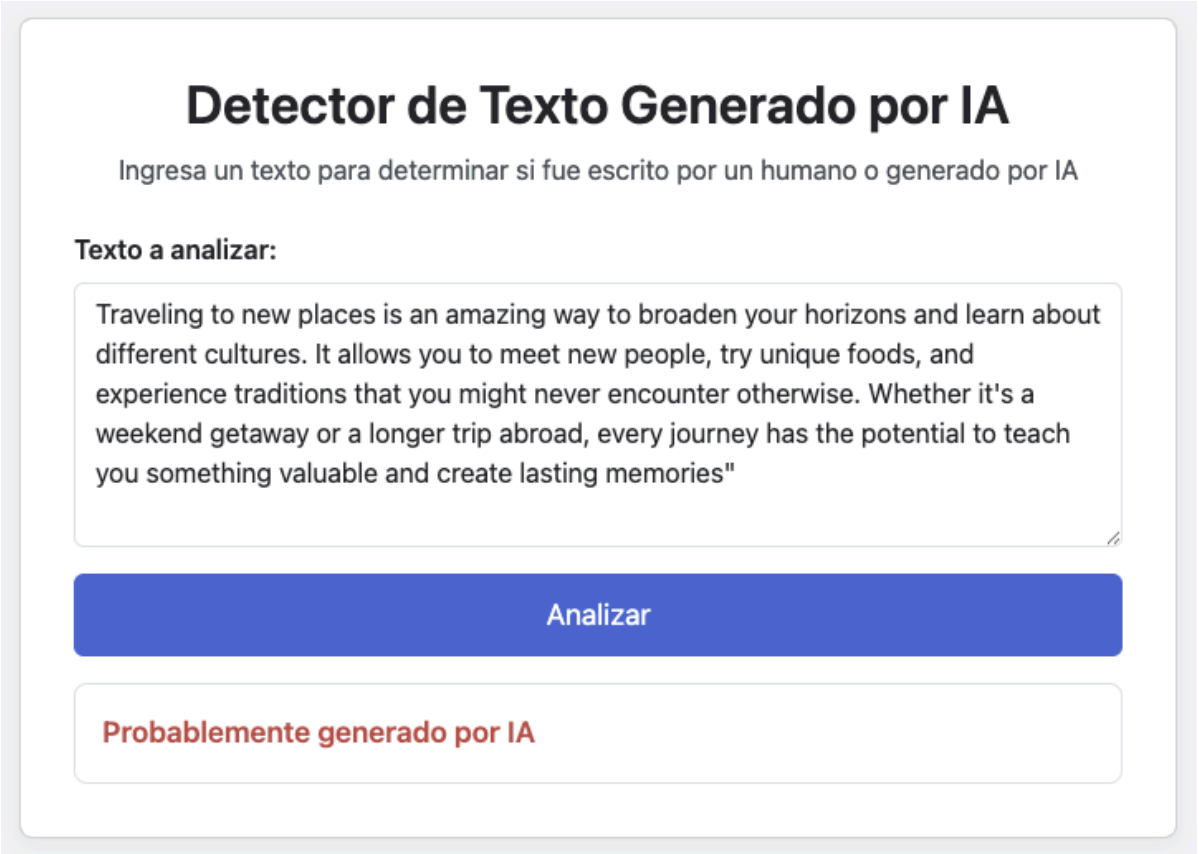
Ejemplos de la Interfaz Web

A continuación se presentan dos ejemplos: uno de un texto correctamente clasificado como generado por inteligencia artificial, y otro como escrito por un ser humano.

- **Ejemplo 1 – Texto escrito por un humano:**



- **Ejemplo 2 – Texto generado por IA:**



Conclusiones

El sistema desarrollado logró cumplir con éxito el objetivo principal del proyecto: distinguir con alta precisión entre ensayos en inglés escritos por humanos y aquellos generados por inteligencia artificial. Gracias al uso de técnicas de procesamiento de lenguaje natural y un clasificador Random Forest optimizado, se alcanzó una precisión del 99% y un excelente equilibrio entre precisión y recall, lo que demuestra la eficacia del enfoque utilizado.

Además, la implementación de una API REST permite que el modelo sea fácilmente integrable en sistemas reales, ampliando su aplicabilidad en contextos educativos y de control de calidad de contenido.

Si bien el sistema presenta limitaciones en cuanto al tipo de texto que puede analizar (ensayos largos, en inglés y de tono formal o académico), representa un paso importante hacia el desarrollo de herramientas automáticas de detección de texto generado por IA, una necesidad creciente en el contexto actual.