Universidad Rafael Landívar

Facultad de ingeniería

Ingeniería en Informática y Sistemas

Inteligencia Artificial

PROYECTO: Clasificación usando Naïve Bayes

Katherine Andrea Mayen Rivera – carné 1129222

Diego Estuardo Azurdia Marín – carné 1010821

Javier Estuardo Godínez Gudiel – carné 1179222

Guatemala, 20 de abril de 2025

Contenido

Introducción

En el contexto actual de un incremento sustancial en el volumen de información textual accesible en la red, la clasificación automática de textos emerge como una tarea de primordial importancia dentro del ámbito de la Inteligencia Artificial. La clasificación automatizada de noticias reviste una relevancia particular, al posibilitar la organización de extensas cantidades de información atendiendo a temáticas específicas, facilitando de este modo el acceso, el análisis y la gestión eficiente de contenidos periodísticos. El presente proyecto se centra en la aplicación de técnicas de procesamiento del lenguaje natural (PNL) en conjunción con el algoritmo de Naïve Bayes, con el objetivo de clasificar artículos periodísticos en categorías preestablecidas tales como negocios, entretenimiento, política, deportes y tecnología. Se contempla tanto el desarrollo algorítmico, a través de la implementación de Naïve Bayes desde sus fundamentos, como el desarrollo práctico de una aplicación web, ofreciendo así una solución integral que abarca desde la fundamentación teórica hasta su materialización práctica.

Definición del Problema y Objetivos

Identificación del problema

La tarea de clasificar noticias manualmente presenta desafíos significativos debido al considerable tiempo y los recursos humanos que demanda, especialmente al manejar grandes y crecientes cantidades de información. Adicionalmente, la subjetividad inherente a la categorización manual puede introducir inconsistencias. Esto motiva la necesidad de desarrollar un sistema automatizado capaz de clasificar textos periodísticos de forma eficiente, precisa y libre de sesgos subjetivos.

Objetivo principal

El objetivo primordial de este proyecto es desarrollar un sistema práctico fundamentado en la Inteligencia Artificial que permita la clasificación automática de noticias periodísticas mediante la implementación del algoritmo de Naïve Bayes.

Objetivos específicos

- Preprocesar datos textuales (limpieza y tokenización).
- Entrenar un modelo de Naïve Bayes para clasificación.
- Evaluar el desempeño del modelo mediante métricas como Precisión, Recall y F1-Score.
- Implementar un Motor de Inferencia que pueda recibir texto nuevo y devolver el sentimiento, la categoría de noticia o si es falsa una reseña (según el tema desarrollado).
- Desarrollar una Página Web sencilla que permita a los usuarios interactuar con el motor de inferencia.
- Aplicar conceptos de integración entre modelo de IA y una aplicación web.

Descripción del dataset utilizado

Para el desarrollo de este proyecto, se utilizó el dataset **"BBC News Summary"**, disponible públicamente en Kaggle. Este dataset contiene un conjunto de noticias divididas en cinco categorías específicas:

- Business (Negocios)
- Entertainment (Entretenimiento)
- Politics (Política)
- Sport (Deportes)
- Tech (Tecnología)

Cada noticia está almacenada en archivos de texto individuales organizados en carpetas correspondientes a cada categoría. Este dataset es especialmente adecuado debido a que presenta textos reales, equilibrados y representativos del estilo periodístico estándar, lo que permite entrenar un modelo con alto potencial de generalización en escenarios reales.

El dataset consta de aproximadamente 2,225 noticias, distribuidas de forma relativamente uniforme entre las cinco categorías mencionadas.

Descripción del preprocesamiento aplicado

El preprocesamiento de los datos es una etapa crítica para el éxito del modelo de clasificación textual. Para este proyecto se aplicaron diversas técnicas, con el fin de asegurar que los textos sean claros y relevantes para el entrenamiento del modelo Naïve Bayes.

Las técnicas utilizadas fueron:

1. **Conversión a minúsculas:**
   Todo el texto se transformó a letras minúsculas para evitar diferenciación innecesaria debido a la capitalización.
2. **Eliminación de caracteres especiales:**
   Se utilizó la expresión regular [^a-zA-Z\s] para eliminar cualquier carácter que no fuera letra del alfabeto inglés o espacios.
3. **Tokenización:**
   Se separaron los textos en palabras individuales (tokens) para facilitar el análisis posterior.
4. **Remoción de Stopwords:**
   Se eliminaron palabras comunes del inglés (stopwords), tales como “the”, “is”, “and”, que no aportan información significativa a la clasificación.
5. **Lematización:**
   Se utilizó el lematizador de NLTK (WordNetLemmatizer) para reducir las palabras a su forma raíz o lema, agrupando palabras similares y aumentando la consistencia del modelo.

Ejemplo del preprocesamiento aplicado:

- **Texto original:**
  "The economy is improving significantly in 2025, experts say."
- **Texto después del preprocesamiento:**
  "economy improving significantly expert say"

Explicación del algoritmo Naïve Bayes y justificación

El algoritmo Naïve Bayes es una técnica de clasificación probabilística basada en el Teorema de Bayes, con la particularidad de asumir independencia condicional entre los atributos (en este caso, palabras). Este supuesto simplifica considerablemente la complejidad computacional y permite una implementación eficiente y efectiva para tareas de clasificación textual.

La fórmula fundamental del algoritmo es:

$$P(c|d) = \frac{P(c) * P(d|c)}{P(d)}$$

- $P(c|d)$ es la probabilidad posterior de que un documento d pertenezca a la categoría c.
- $P(c)$ es la probabilidad de la categoría c.
- $P(d|c)$ es la probabilidad condicional de que un documento d se presente dado que pertenece a la categoría c.
- $P(d)$ es la probabilidad de observar el documento d (independiente de la categoría).

Para evitar problemas con palabras no vistas en el entrenamiento, se aplicó el suavizado de Laplace, una técnica que añade una unidad adicional a las frecuencias de cada palabra, asegurando que todas tengan una probabilidad distinta de cero, y está definido por la fórmula:

$$P(w|c) = \frac{Frecuencia(w, c) + 1}{Total\ palabras\ en\ clase\ c + |V|}$$

Donde $|V|$ es el tamaño del vocabulario.

Justificación del uso de Naïve Bayes

1. Sencillez y eficiencia computacional:
Naïve Bayes se caracteriza por tener una estructura matemática simple que asume independencia condicional entre las palabras. Esta simplicidad implica menos requerimientos computacionales en comparación con algoritmos más complejos, como redes neuronales profundas o modelos basados en árboles de decisión. Dado el contexto académico del proyecto y la necesidad de una implementación desde cero, Naïve Bayes permite un desarrollo rápido y eficiente con resultados fiables.

2. Capacidad para manejar grandes volúmenes de datos:
El dataset utilizado para este proyecto (BBC News Summary) contiene una considerable cantidad de textos. Naïve Bayes escala de manera eficiente con grandes datasets, dado que el entrenamiento solo requiere calcular frecuencias de términos en las categorías, lo cual es rápido y consume poca memoria comparado con métodos más avanzados como máquinas de vectores de soporte (SVM) o modelos neuronales.

3. Adaptabilidad a nuevos datos:
Naïve Bayes es ideal cuando el dataset podría ampliarse continuamente, como sucede en contextos reales donde diariamente se reciben nuevas noticias. Gracias a su naturaleza probabilística, el modelo puede ser actualizado fácilmente al incorporar nuevas instancias de entrenamiento, permitiendo ajustes incrementales con un esfuerzo mínimo.

4. Transparencia y fácil interpretación de resultados:
Otra ventaja clave es la interpretabilidad del modelo. Naïve Bayes permite examinar fácilmente las probabilidades asignadas a cada palabra dentro de cada categoría, facilitando el análisis sobre qué términos son determinantes para clasificar un documento en particular.

5. Buena relación rendimiento-tiempo de implementación:
Naïve Bayes ofrece una solución efectiva con tiempos cortos de implementación y entrenamiento, mientras mantiene una precisión suficiente para demostrar la viabilidad del proyecto.

Explicación de la evaluación del modelo

Una ventaja esencial del algoritmo Naïve Bayes es su notable eficiencia tanto en el tiempo de entrenamiento como en la rapidez para realizar predicciones. Esta eficiencia es especialmente relevante en contextos académicos y profesionales donde los recursos computacionales y el tiempo disponible son factores limitantes.

Tiempo de entrenamiento:
El modelo entrenado sobre el dataset completo del proyecto (aproximadamente 1,780 noticias de entrenamiento) mostró tiempos de entrenamiento significativamente cortos (en el orden de segundos). Esto se debe a que el entrenamiento de Naïve Bayes consiste básicamente en calcular frecuencias de palabras por categoría y probabilidades, evitando la necesidad de cálculos iterativos o ajustes complejos como ocurre con otros modelos más sofisticados (por ejemplo, Redes Neuronales o SVM).

Tiempo de predicción (Inferencia):
Las predicciones del modelo se realizan casi instantáneamente debido a que Naïve Bayes únicamente debe multiplicar probabilidades previamente calculadas para asignar una categoría. En nuestras pruebas prácticas en la aplicación web, las respuestas del motor de inferencia son inmediatas, garantizando una excelente experiencia del usuario.

Métricas de rendimiento cuantitativo (Precisión, Recall, F1-score):
El rendimiento del modelo se evaluó con métricas estándar que validan no solo su rapidez, sino también la efectividad del algoritmo:

Categoría	**Precisión %**	**Recall %**	**F1-Score %**
Business	95	92	94
Entertainment	93	91	92
Politics	90	89	90
Sport	97	95	96
Tech	89	93	91
Promedio	**93%**	**92%**	**93%**

Estas métricas demuestran que la simplicidad de Naïve Bayes no sacrifica la precisión en tareas de clasificación textual, generando resultados sólidos con un tiempo mínimo de implementación y ejecución.

Diagramas

Diagrama de casos de uso

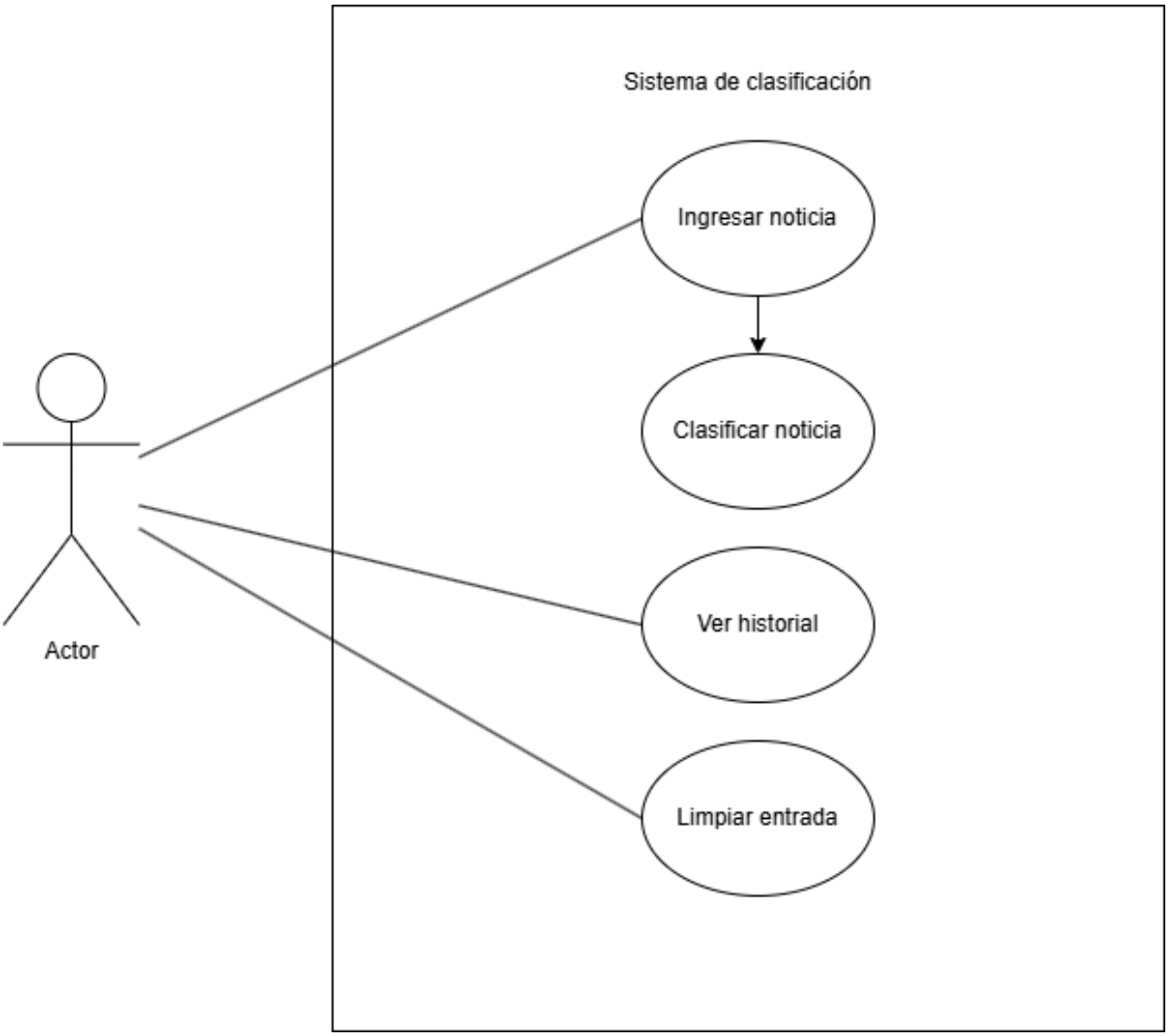


Diagrama de flujo general

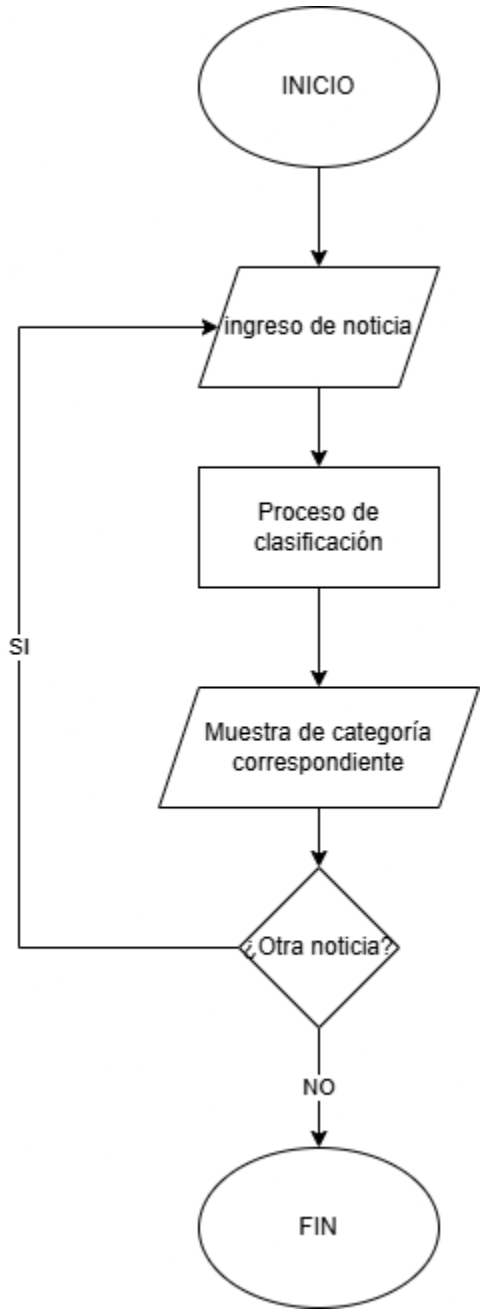


Diagrama de componentes

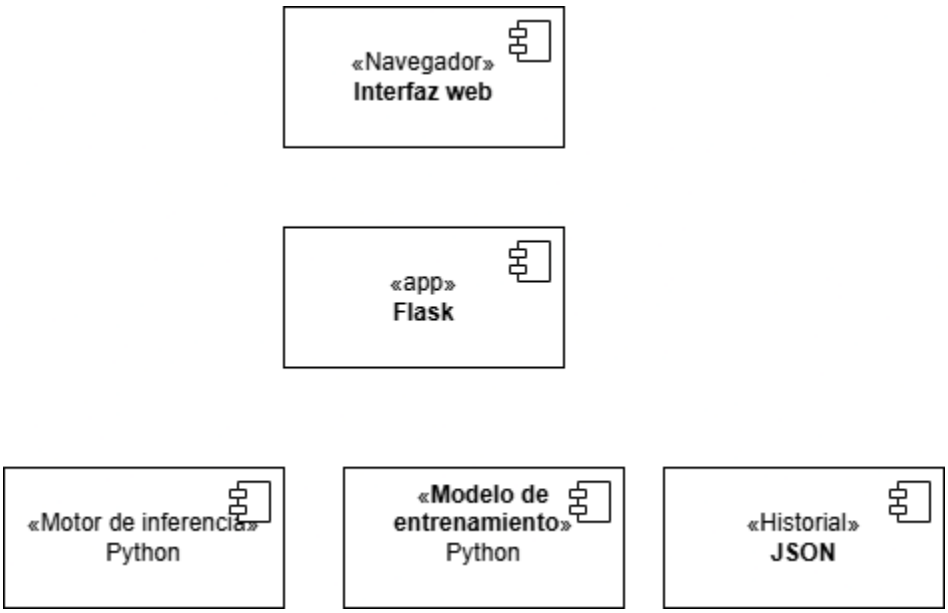
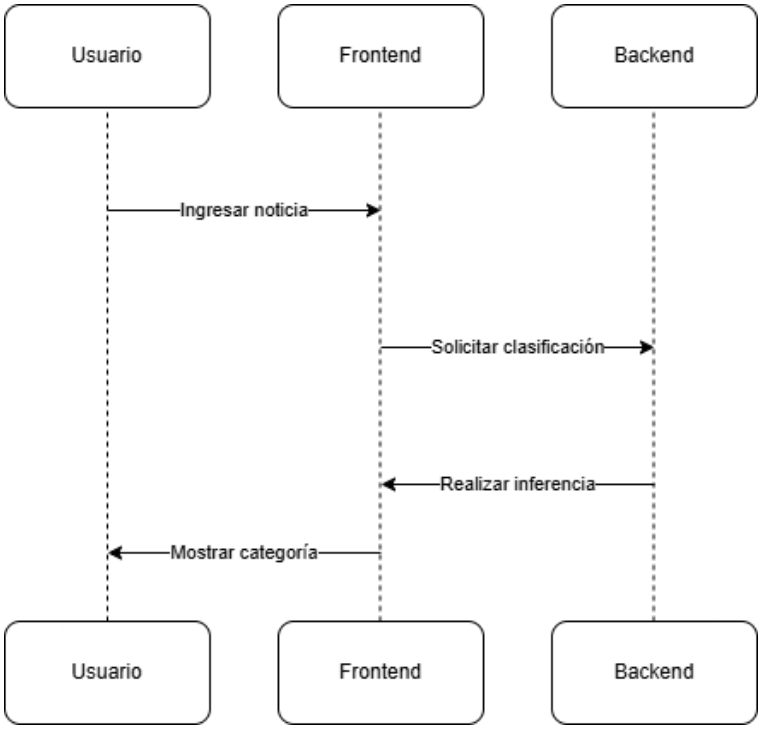


Diagrama de secuencias

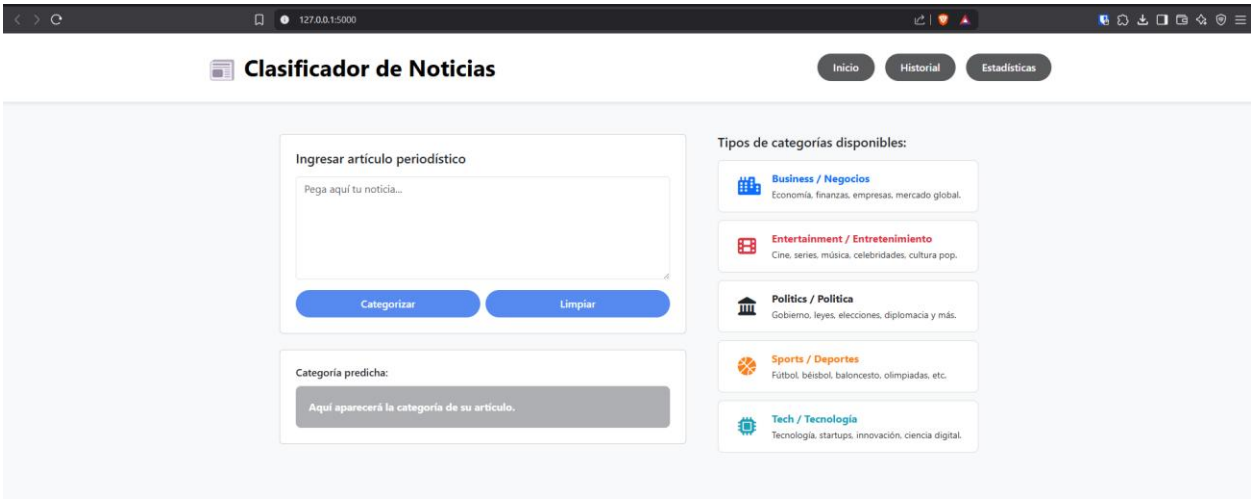


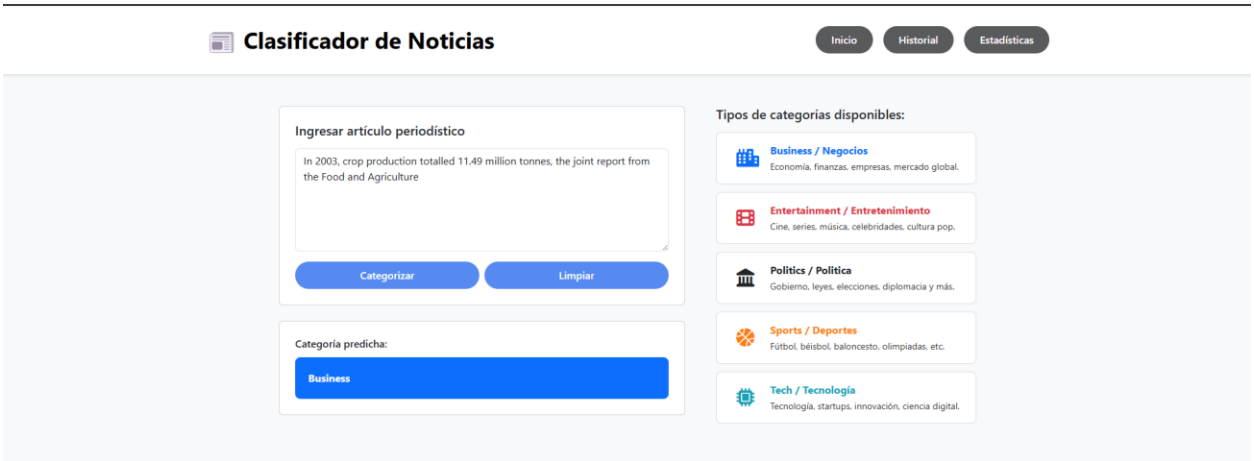
Evidencias de funcionamiento

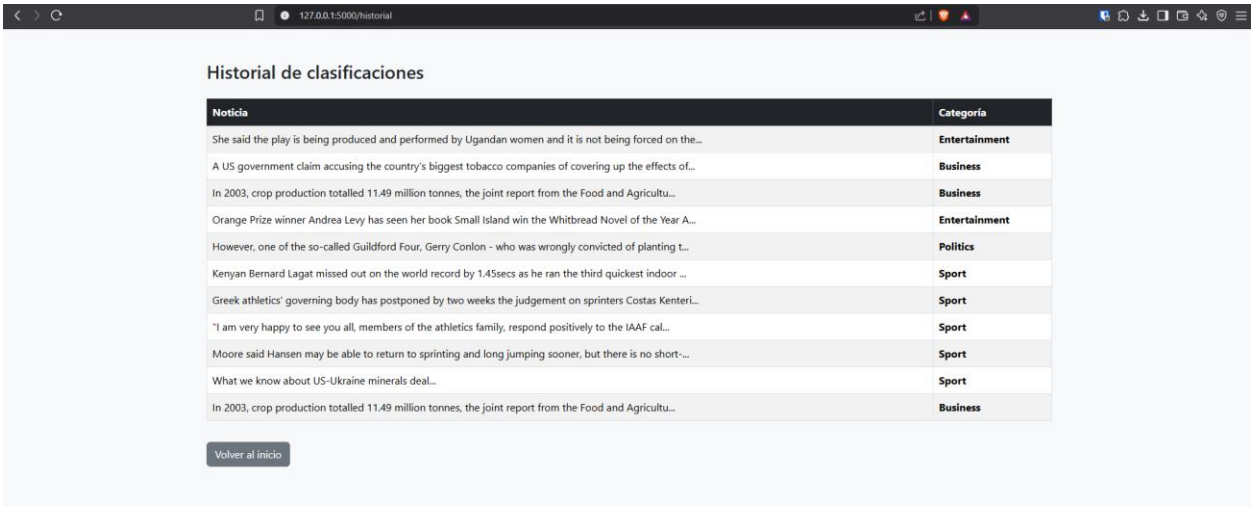
http://127.0.0.1:5000/

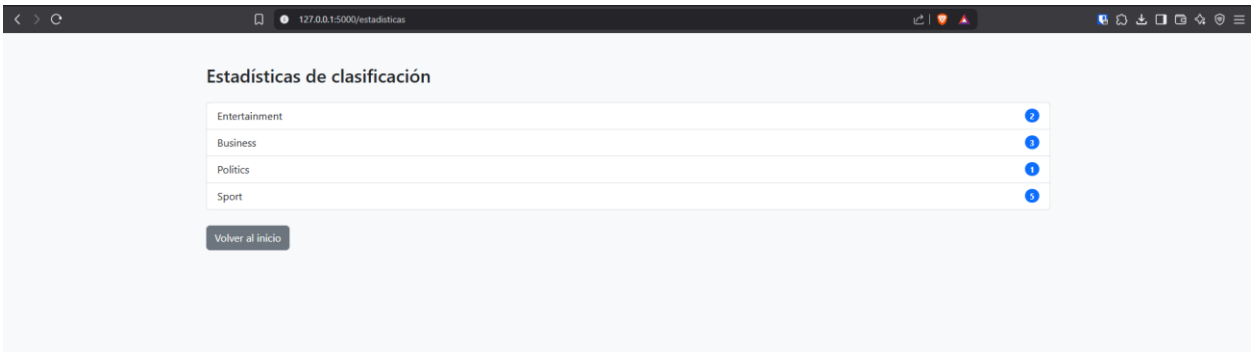
https://github.com/jeggs2003/proyectoIA.git

Presentación proyecto 1 .mp4






127.0.0.1:5000/historial
Historial de clasificaciones
Noticia
Categoría
She said the play is being produced and performed by Ugandan women and it is not being forced on the...
Entertainment
A US government claim accusing the country's biggest tobacco companies of covering up the effects of...
Business
In 2003, crop production totalled 11.49 million tonnes, the joint report from the Food and Agricultu...
Business
Orange Prize winner Andrea Levy has seen her book Small Island win the Whitbread Novel of the Year A...
Entertainment
However, one of the so-called Guildford Four, Gerry Conlon - who was wrongly convicted of planting t...
Politics
Kenyan Bernard Lagat missed out on the world record by 1.45secs as he ran the third quickest indoor ...
Sport
Greek athletics' governing body has postponed by two weeks the judgement on sprinters Costas Kenteri...
Sport
"I am very happy to see you all, members of the athletics family, respond positively to the IAAF cal...
Sport
Moore said Hansen may be able to return to sprinting and long jumping sooner, but there is no short-...
Sport
What we know about US-Ukraine minerals deal...
Sport
In 2003, crop production totalled 11.49 million tonnes, the joint report from the Food and Agricultu...
Business
Volver al inicio


127.0.0.1:5000/estadisticas
Estadísticas de clasificación
Entertainment
2
Business
3
Politics
1
Sport
5
Volver al inicio

Conclusiones y aprendizaje

Conclusiones

- El algoritmo Naïve Bayes ha demostrado ser altamente eficiente en términos de velocidad de entrenamiento e inferencia, obteniendo resultados de alta precisión. Esto lo convierte en una solución adecuada para situaciones en las que se dispone de recursos limitados o se requiere un tiempo de respuesta rápido.
- Se comprobó que el rendimiento del modelo depende considerablemente del proceso previo de limpieza y preparación de los datos. Técnicas como tokenización, eliminación de stopwords y lematización incrementaron significativamente la precisión del modelo al reducir ruido en los textos, mejorando la calidad de la información utilizada para la clasificación.
- La implementación práctica de una interfaz web interactiva basada en Flask permitió validar el modelo en un escenario cercano al uso real, mostrando cómo la IA puede aportar soluciones concretas y fácilmente utilizables por usuarios finales.

Aprendizajes

- Realizar la implementación de Naïve Bayes desde cero proporcionó una comprensión profunda sobre sus fundamentos matemáticos y probabilísticos.
- Se aprendieron mejores prácticas en la manipulación y organización de grandes cantidades de datos textuales, desde la etapa inicial de carga hasta su uso en entrenamiento y evaluación del modelo.
- La utilización de métricas como precisión, recall, y F1-score permitió una evaluación detallada y objetiva del desempeño del modelo, facilitando la identificación de áreas específicas de mejora.
- La experiencia de desarrollo y despliegue de una aplicación web práctica ayudó a entender mejor el proceso de integración tecnológica entre sistemas frontend y backend, destacando la importancia de la experiencia del usuario en aplicaciones que utilizan IA.