Universidad Rafael Landívar

Primer Semestre 2025

Inteligencia Artificial

PROYECTO: Clasificador Naive Bayes

Clasificación de noticias

Diego Azurdia 2528119

César Bocel 1094921

Oswaldo Orellana 1163722

Guatemala 23 de abril de 2025

Índice

Repositorio

diegoazurdia1998/ProyectoIA2025

El video está dentro del repositorio con el nombre de “Video.mp4”

Introducción

El presente proyecto fue desarrollado como parte del curso de **Inteligencia Artificial** de la carrera de **Ingeniería en Sistemas**, y tiene como objetivo aplicar los conceptos de clasificación de texto utilizando el algoritmo **Naïve Bayes**, uno de los modelos probabilísticos supervisados más utilizados por su simplicidad, eficiencia y buen rendimiento en tareas de clasificación textual.

Se trabajó específicamente en la **clasificación automática de noticias**, desarrollando un sistema que permite predecir a qué categoría temática pertenece una noticia ingresada por el usuario. Para ello, se empleó un dataset real proveniente de la **BBC News**, compuesto por aproximadamente 500 artículos por categoría, clasificados en cinco temas principales: `business`, `entertainment`, `politics`, `sport` y `tech`.

Además del desarrollo del modelo de clasificación, el proyecto incluye una **interfaz web funcional**, construida con **HTML, CSS, JavaScript y PHP**, que permite al usuario subir noticias en formato `.txt`, recibir la predicción de la categoría, consultar un historial de clasificaciones previas y visualizar métricas del rendimiento del modelo mediante gráficas.

Esta implementación busca no solo cumplir con los objetivos técnicos del curso, sino también demostrar una **integración práctica entre un modelo de IA y una aplicación web**, fortaleciendo las habilidades de los estudiantes tanto en el ámbito algorítmico como en el desarrollo de soluciones reales orientadas a usuarios finales.

Definición del problema

En la actualidad, la cantidad de información textual disponible en internet crece a un ritmo exponencial, particularmente en el ámbito de los medios de comunicación digitales. Los portales de noticias generan cientos de artículos diariamente, abarcando múltiples temáticas como economía, tecnología, política, deportes y entretenimiento. Ante este volumen creciente de contenido, se vuelve inviable realizar una clasificación manual eficiente.

Por tanto, surge la necesidad de automatizar este proceso mediante herramientas de inteligencia artificial. El problema que se aborda en este proyecto consiste en desarrollar un sistema capaz de **clasificar automáticamente noticias en formato de texto plano (.txt) en una de las siguientes categorías predefinidas**: business, entertainment, politics, sport o tech. Esta clasificación se realiza **basándose en un análisis probabilístico del contenido textual**, identificando patrones léxicos característicos de cada categoría mediante técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático supervisado.

Objetivos Generales

Diseñar e implementar un sistema de inteligencia artificial que permita clasificar noticias escritas en archivos `.txt` en categorías temáticas mediante técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático supervisado.

Objetivos específicos

- Desarrollar un modelo de clasificación basado en **Naive Bayes Multinomial**.
- Utilizar un conjunto de datos etiquetado para entrenar y evaluar el modelo.
- Implementar una interfaz web que permita:
    - Subir archivos `.txt` de noticias.
    - Obtener y visualizar la categoría predicha.
    - Registrar el historial de clasificaciones realizadas.
    - Visualizar métricas de evaluación del modelo (precisión, recall, F1-score).

Descripción del dataset

El dataset utilizado en este proyecto proviene del **BBC News Dataset**, un conjunto de datos ampliamente utilizado en tareas de clasificación de texto. Este dataset incluye noticias reales publicadas por la BBC y organizadas en **cinco categorías principales**:

Categoría	Descripción
business	Noticias relacionadas con economía, empresas, mercados financieros, etc.
entertainment	Noticias sobre música, cine, celebridades, teatro, televisión, etc.
politics	Noticias sobre gobierno, elecciones, política nacional e internacional.
sport	Noticias deportivas: fútbol, tenis, atletismo, etc.
tech	Noticias sobre tecnología, internet, dispositivos, innovación, etc.

Cada categoría contiene aproximadamente **400 a 500 noticias**, distribuidas en archivos .txt individuales, organizados por carpeta. Esto permitió aplicar una estrategia de carga de datos automática para su procesamiento y entrenamiento del modelo.

Este dataset ha sido preprocesado con técnicas como **limpieza de texto**, **tokenización** y **filtrado de palabras raras**, para luego ser dividido en conjuntos de entrenamiento y prueba con un **80/20%** respectivamente.

Descripción preprocesamiento

Los pasos realizados para el preprocesamiento de los datos fueron los siguientes:

Paso 1: Limpieza de texto

Se eliminaron los caracteres especiales tales como signos y hashtags dejando solamente letras y espacios

Se convirtió a minúsculas todo el texto.

Paso 2: Tokenización de las palabras

Se separaron las palabras en función de los espacios.

Paso 3: Filtrado de palabras para el vocabulario

Se construyó un vocabulario con palabras que aparecían al menos tres veces en el conjunto de entrenamiento para reducir el ruido.

Descripción de la implementación Naive Bayes

El algoritmo clasifica noticias en 5 categorías las cuales fueron '*business'*, *'entertainment'*, *'politics'*, *'sport'* y *'tech*', utilizando el enfoque multinomial Naive Bayes, que se ajusta muy bien a las predicciones en base a palabras.

Primero se calcularon las probabilidades a priori en base a la frecuencia de cada clase en los datos de entrenamiento los cuales constituyeron el 80% del dataset utilizado. Luego se realizó un diccionario que almacenó cuantas veces aparecía cada palabra en una categoría, para cada una de las cinco categorías. Finalmente se utilizó el suavizado de LaPlace para evitar probabilidades iguales a cero es decir que una palabra no apareció dentro de una clase.

Red bayesiana

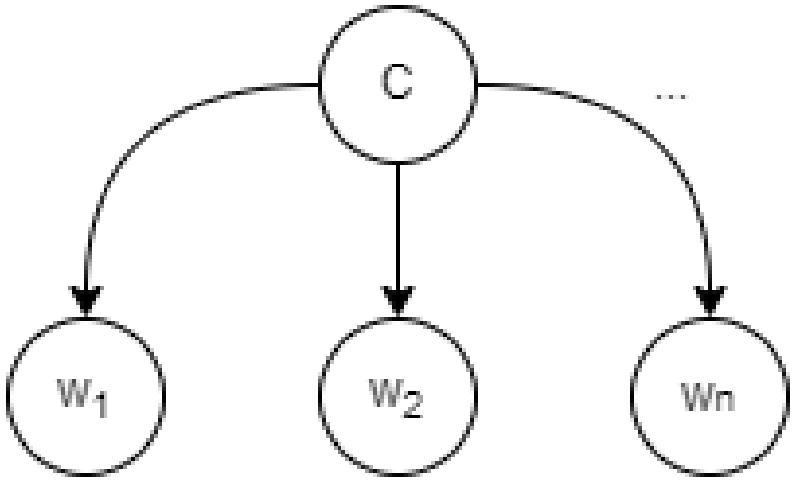
Nodos de la Red

Variables Observadas:

Palabras (w_1, w_2, ..., w_n): Tokens extraídos del texto (ejemplo: "apple", "stock").

Variable Objetivo:

Categoría (C): Clase de la noticia



Explicación de la evaluación del modelo

El modelo fue implementado utilizando una versión personalizada del clasificador Naive Bayes Multinomial, desarrollada paso a paso. Este modelo aprende las probabilidades de que un documento pertenezca a una determinada categoría en función de las palabras que contiene. Además, aplica suavizado de Laplace para evitar divisiones por cero durante el cálculo de probabilidades.

Se trabajó con un conjunto de noticias categorizadas en cinco clases (business, entertainment, politics, sport, y tech), extraídas de una base de datos de la BBC conocida como Pariza.

Previo al entrenamiento, cada noticia es sometida a un proceso de preprocesamiento, el cual incluye:

- Conversión del texto a minúsculas.
- Eliminación de caracteres especiales.
- Tokenización (división del texto en palabras individuales).

Posteriormente, el conjunto de datos se divide en un 80% para entrenamiento y un 20% para prueba, manteniendo el equilibrio de clases mediante estratificación.

Para evaluar el desempeño del modelo, se utilizan las métricas estándar de clasificación:

- Precisión
- Recall (sensibilidad por clase)
- F1-Score
- Soporte (número de ejemplos reales por clase)

Estas métricas se complementan con el uso de una matriz de confusión, la cual permite observar cuántos documentos fueron correctamente clasificados y cuántos fueron confundidos con otras categorías.

Adicionalmente, se almacenan las probabilidades marginales (a priori) de cada clase en el archivo Data.csv, las cuales son reutilizadas durante el proceso de inferencia del modelo.

Diagramas

Arquitetcura de la solución

1. Frontend

Index.html: Encargado del aspecto gráfico del sitio web.

Precision.php: Encargado de el mostrar las gráficas en el sitio web de presicion, recall y f1-score

Subir.php: Encargado de recibir el .txt que ingresará el usuario

Historial.php: Encargado de mostrar y almacenar todas las subidas al sitio web

2. Backend

NaiveBayes.py: Realiza los cálculos en base al teorema de Naive Bayes para obtener el tipo de noticia que viene desde el sitio web.

DownloadDataSet.py: Descarga la base de datos que se usará para las probabilidades.

3. Base de datos

Pariza: Inlcuye toda la base de datos de las noticias de la BBC

Summary BBC: Incluye dichas noticias clasificadas dependiendo su categoría.

Data.csv: Datos probabilísticos que son usados y actualizados según el teorema de Naive Bayes al ingresar nueva información.

Historial.csv: Contiene el historial de texto ingresado.

4. Archivos de soporte

ReadMe.md: Descripción del proyecto y explicación de instalación.

Proyecto 1 IA.pdf: Información de los parámetros a evaluar en el proyecto.

Diagrama de casos de uso

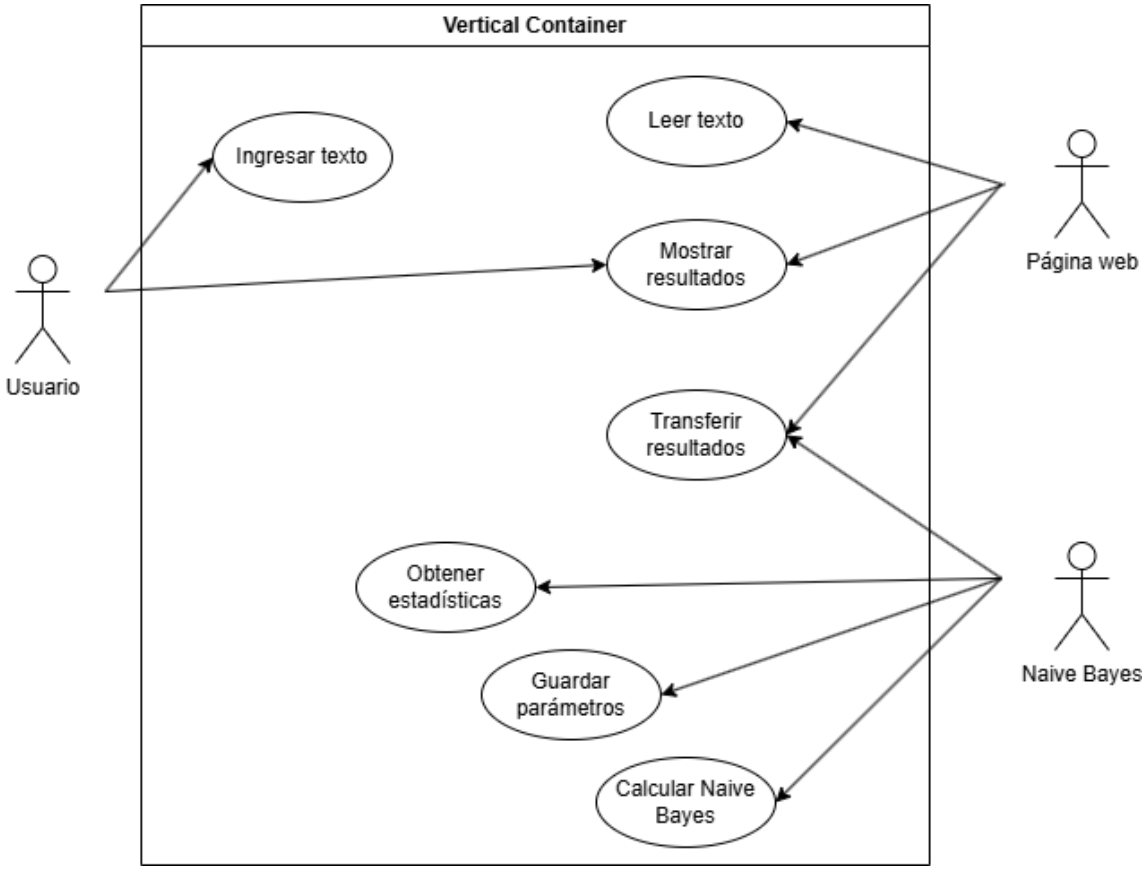


Diagrama de flujo general

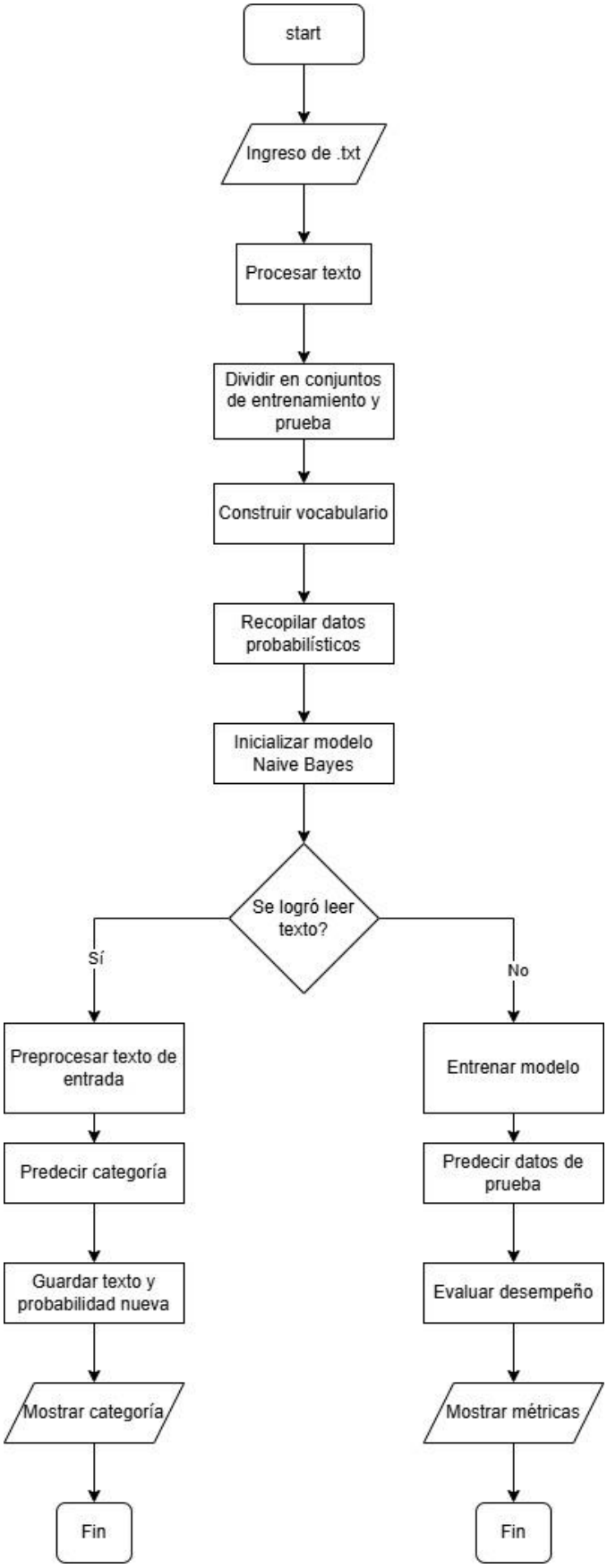


Diagrama de componentes

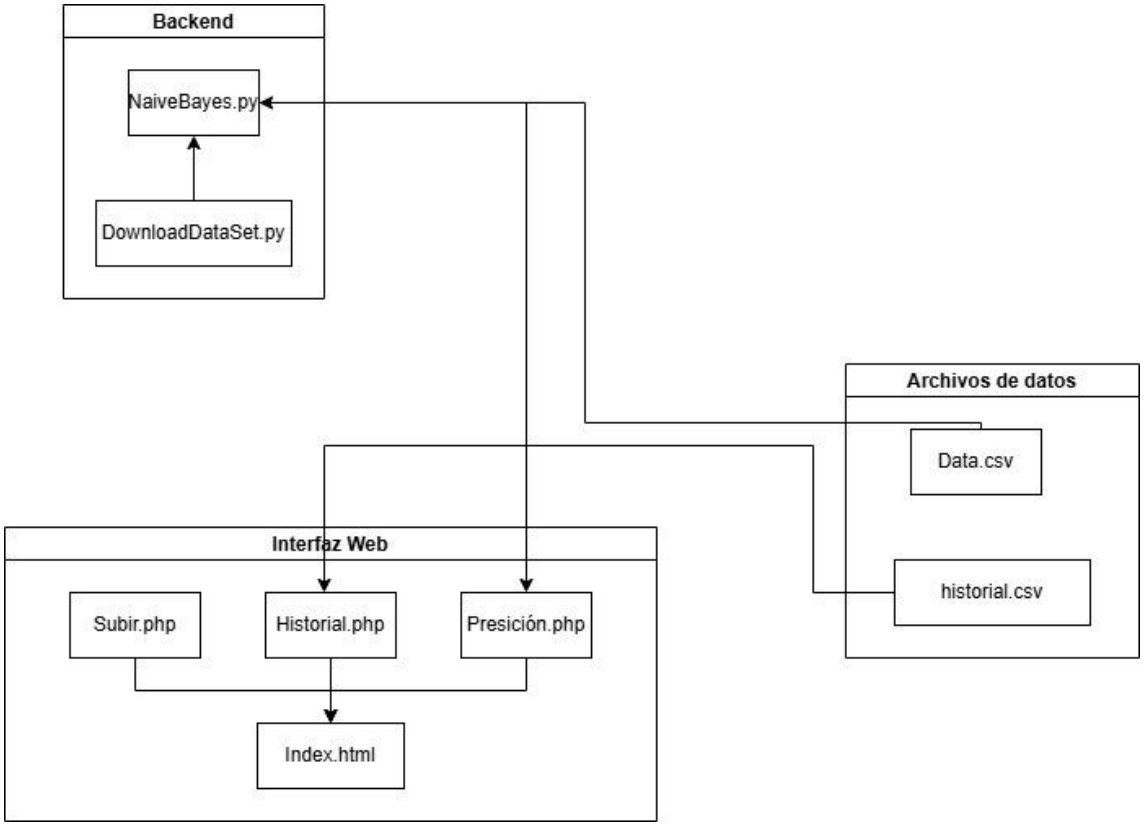
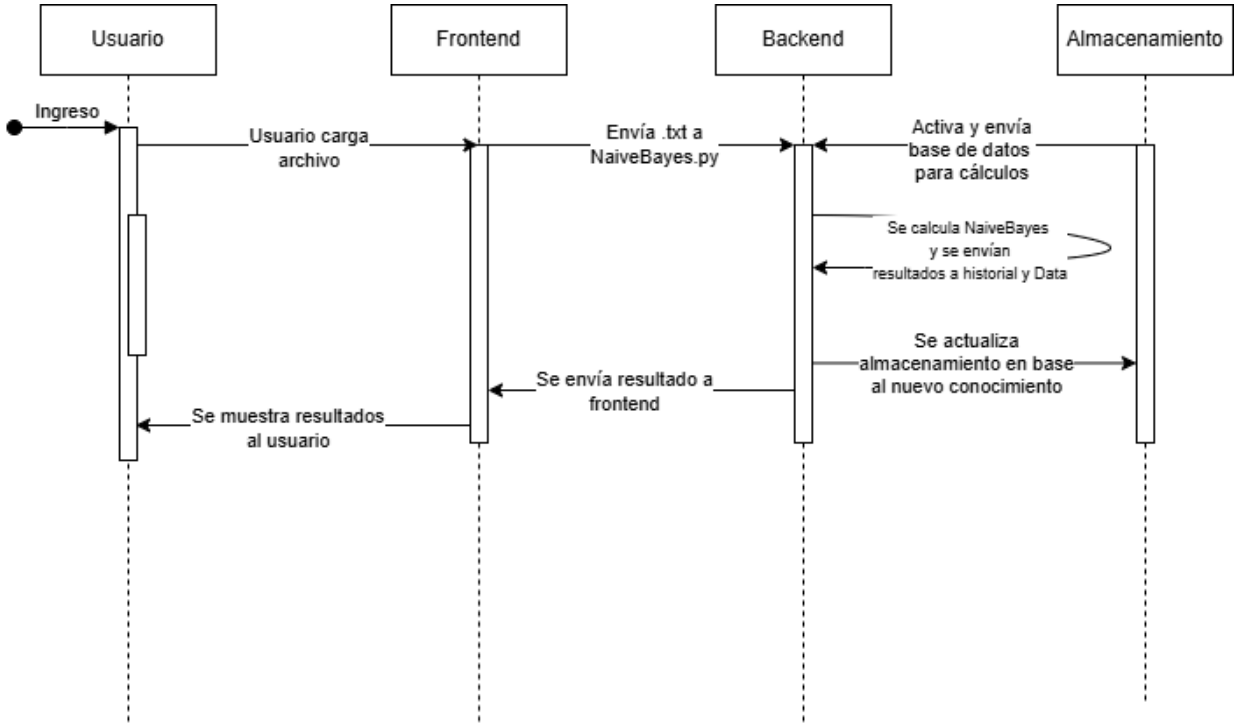
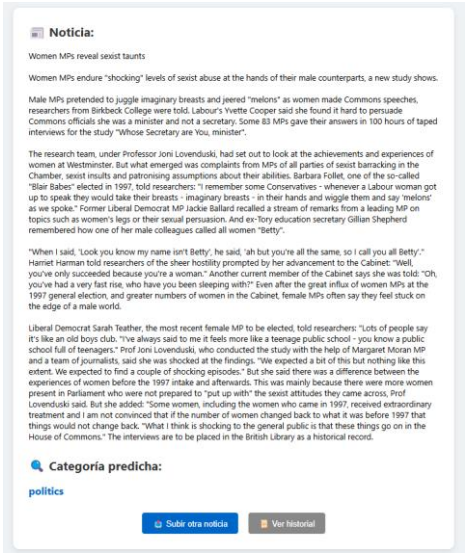


Diagrama de secuencias



Evidencia de funcionamiento



Noticia:

Women MPs reveal sexist taunts

Women MPs endure "shocking" levels of sexist abuse at the hands of their male counterparts, a new study shows.

Male MPs pretended to juggle imaginary breasts and jeered "melons" as women made Commons speeches, researchers from Birkbeck College were told. Labour's Yvette Cooper said she found it hard to persuade Commons officials she was a minister and not a secretary. Some 83 MPs gave their answers in 100 hours of taped interviews for the study "Whose Secretary are You, minister".

The research team, under Professor Joni Lovenduski, had set out to look at the achievements and experiences of women at Westminster. But what emerged was complaints from MPs of all parties of sexist barracking in the Chamber, sexist insults and patronising assumptions about their abilities. Barbara Follet, one of the so-called "Blair Babes" elected in 1997, told researchers: "I remember some Conservatives - whenever a Labour woman got up to speak they would take their breasts - imaginary breasts - in their hands and wiggle them and say 'melons' as we spoke." Former Liberal Democrat MP Jackie Ballard recalled a stream of remarks from a leading MP on topics such as women's legs or their sexual persuasion. And ex-Tory education secretary Gillian Shepherd remembered how one of her male colleagues called all women "Betty".

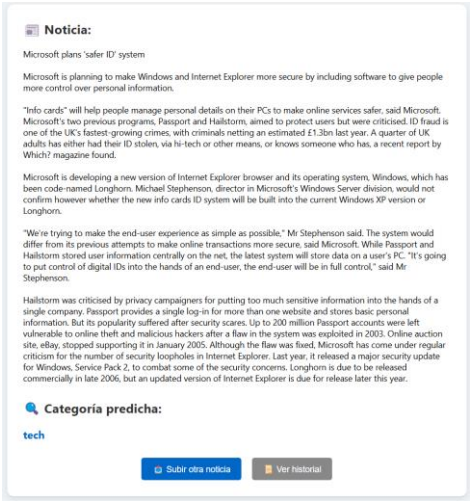
"When I said, 'Look you know my name isn't Betty', he said, 'ah but you're all the same, so I call you all Betty'." Harriet Harman told researchers of the sheer hostility prompted by her advancement to the Cabinet: "Well, you've only succeeded because you're a woman." Another current member of the Cabinet says she was told: "Oh, you've had a very fast rise, who have you been sleeping with?" Even after the great influx of women MPs at the 1997 general election, and greater numbers of women in the Cabinet, female MPs often say they feel stuck on the edge of a male world.

Liberal Democrat Sarah Teather, the most recent female MP to be elected, told researchers: "Lots of people say it's like an old boys club. "I've always said to me it feels more like a teenage public school - you know a public school full of teenagers." Prof Joni Lovenduski, who conducted the study with the help of Margaret Moran MP and a team of journalists, said she was shocked at the findings. "We expected a bit of this but nothing like this extent. We expected to find a couple of shocking episodes." But she said there was a difference between the experiences of women before the 1997 intake and afterwards. This was mainly because there were more women present in Parliament who were not prepared to "put up with" the sexist attitudes they came across, Prof Lovenduski said. But she added: "Some women, including the women who came in 1997, received extraordinary treatment and I am not convinced that if the number of women changed back to what it was before 1997 that things would not change back. "What I think is shocking to the general public is that these things go on in the House of Commons." The interviews are to be placed in the British Library as a historical record.

Categoría predicha:

politics

Subir otra noticia | Ver historial



Noticia:

Microsoft plans 'safer ID' system

Microsoft is planning to make Windows and Internet Explorer more secure by including software to give people more control over personal information.

"Info cards" will help people manage personal details on their PCs to make online services safer, said Microsoft. Microsoft's two previous programs, Passport and Hailstorm, aimed to protect users but were criticised. ID fraud is one of the UK's fastest-growing crimes, with criminals netting an estimated £1.3bn last year. A quarter of UK adults has either had their ID stolen, via hi-tech or other means, or knows someone who has, a recent report by Which? magazine found.

Microsoft is developing a new version of Internet Explorer browser and its operating system, Windows, which has been code-named Longhorn. Michael Stephenson, director in Microsoft's Windows Server division, would not confirm however whether the new info cards ID system will be built into the current Windows XP version or Longhorn.

"We're trying to make the end-user experience as simple as possible," Mr Stephenson said. The system would differ from its previous attempts to make online transactions more secure, said Microsoft. While Passport and Hailstorm stored user information centrally on the net, the latest system will store data on a user's PC. "It's going to put control of digital IDs into the hands of an end-user, the end-user will be in full control," said Mr Stephenson.

Hailstorm was criticised by privacy campaigners for putting too much sensitive information into the hands of a single company. Passport provides a single log-in for more than one website and stores basic personal information. But its popularity suffered after security scares. Up to 200 million Passport accounts were left vulnerable to online theft and malicious hackers after a flaw in the system was exploited in 2003. Online auction site, eBay, stopped supporting it in January 2005. Although the flaw was fixed, Microsoft has come under regular criticism for the number of security loopholes in Internet Explorer. Last year, it released a major security update for Windows, Service Pack 2, to combat some of the security concerns. Longhorn is due to be released commercially in late 2006, but an updated version of Internet Explorer is due for release later this year.

Categoría predicha:

tech

Subir otra noticia | Ver historial

Fecha y hora	Noticia	Categoría predicha
2025-04-25 22:14:08	Microsoft plans 'safer ID' system Microsoft is planning to make Windows and Inte	tech

2025-04-21 18:02:10	Rory McIlroy could go onto win 10 majors now Masters 'shackles are off,' say	sport
2025-04-19 21:22:48	Rory McIlroy could go onto win 10 majors now Masters 'shackles are off,' say	sport
2025-04-19 15:41:43	'Errors' doomed first Dome sale The initial attempt to sell the Millennium Dome	sport
2025-04-19 14:59:05	Japan narrowly escapes recession Japan's economy teetered on the brink of a tech	business
2025-04-19 20:40:08	Microsoft seeking spyware trojan Microsoft is investigating a trojan program tha	tech
2025-04-19 20:38:27	Musical treatment for Capra film The classic film It's A Wonderful Life is to be	entertainment
2025-04-19 14:46:55	Dollar gains on Greenspan speech The dollar has hit its highest level against th	business
2025-04-19 14:59:25	Jobs growth still slow in the US The US created fewer jobs than expected in Janu	business
2025-04-19 21:20:20	Wi-fi web reaches farmers in Peru A network of community computer centres, linke	tech
2025-04-21 17:24:47	Ink helps drive democracy in Asia The Kyrgyz Republic, a small, mountainous stat	tech
2025-04-23 11:12:19	Shannon Sharpe, a Superbowl champion and host of a popular podcast, has been acc	entertainment
2025-04-23 19:45:46	Poles play with GameBoy 'blip-pop' A group of artists in Poland has taken the ca	tech
2025-04-25 21:27:59	Lib Dems' 'bold' election policy Charles Kennedy has told voters his Liberal Dem	politics

Conclusión y aprendizaje

A lo largo de este proyecto logramos entender mucho a mayor profundidad tanto el procedimiento como la aplicación en un contexto real del teorema de Naive Bayes. En donde a través del poder interpretar la forma que la estadística y probabilidad está implícita dentro de la forma que se redactan distintas noticias, se puede determinar de manera automática el tipo de noticia que es la que se está tratando.

Otro aprendizaje que tuvimos fue el como aplicar el recall, precision y F1-score a un contexto real el cual puede ayudar en un contexto real como es el de organizar automáticamente las noticias de un noticiero como la BBC que tiene un gran flujo de información siendo intercambiado en tiempo real.