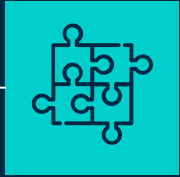


# CLASIFICACIÓN DE NOTICIAS POR CATEGORÍA

Daniel Pazmiño Ortega  
Juan Felipe Mazo

# Tabla de contenidos



01

CONTEXTO



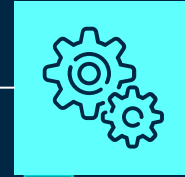
02

MOTIVACIÓN  
Y METODOLOGIA



03

RESULTADOS



04

CONCLUSIONES

# CONTEXTO

- Una de las aplicaciones más importante del ML es la clasificación
- La clasificación se basa en asignar una categoría a un conjunto de datos basándose en ciertas características



# CONTEXTO

## TF-IDF

- Es una medida numérica que expresa cuán relevante es una palabra para un documento en una colección
- Relaciona la frecuencia de término (TF) con la frecuencia de ocurrencia del término en la colección de documentos (IDF) [1].

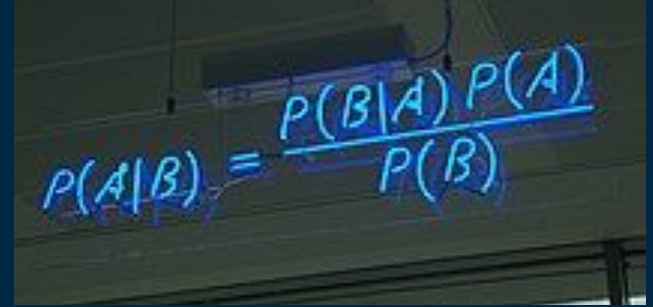
$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$



# CONTEXTO

## Naive bayes

- Es un clasificador probabilístico simple que se basa en el teorema de Bayes.
- Probabilidad condicional.
- Supone una independencia en los atributos (ingenuo) [2]


$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Formula teorema de Bayes



# Dataset utilizado [3]

- Contiene 2225 noticias de BBC
- Está separado en dos archivos (entrenamiento y testeo)
- El archivo de prueba contiene el Id del artículo y el texto
- El archivo de entrenamiento también contiene su categoría



## CONTEXTO

1582	howard truanted to play snooker conservative leader michael howard has admitted he used to play tr...	politics
651	wales silent on grand slam talk rhys williams says wales are still not thinking of winning the grand...	sport
1797	french honour for director parker british film director sir alan parker has been made	entertainment

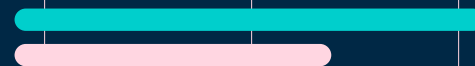
Imagen Dataset de entrenamiento

# MOTIVACIÓN

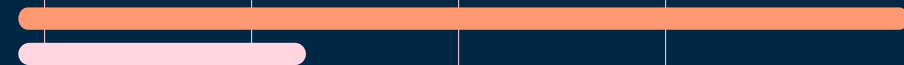
ENFRENTAR LA SOBRECARGA DE INFORMACIÓN



AHORRAR TIEMPO Y RECURSOS



GENERAR VALOR



FACILITAR ACCESO A LA INFORMACIÓN



# Metodología

- Separar por palabras
- Remover Stopwords
- Vectorización de las palabras
- Aplicar idf
- Pasar categoría a números

Preprocesamiento  
de datos

Naive  
Bayes

A partir del TF-IDF y la  
categoría

General y por categoría

Precisión

Boton de  
carga

Se muestra la  
predicción



# RESULTADOS

## PRESICIÓN

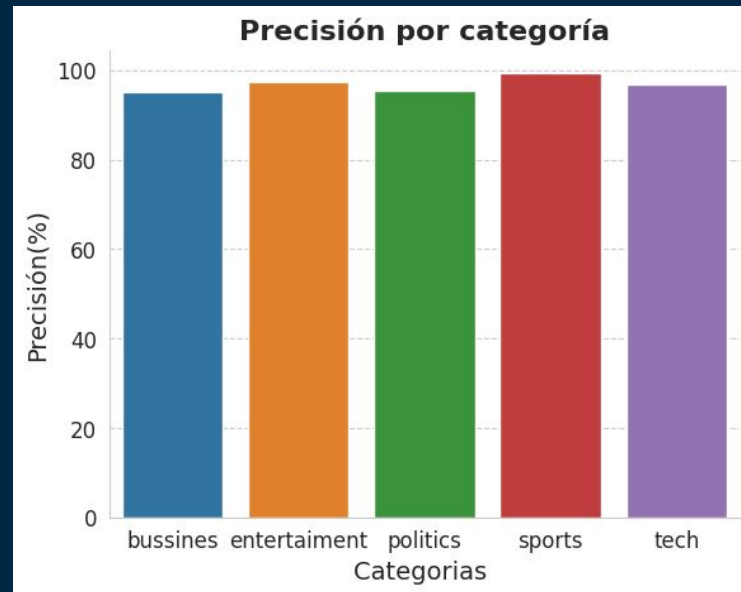
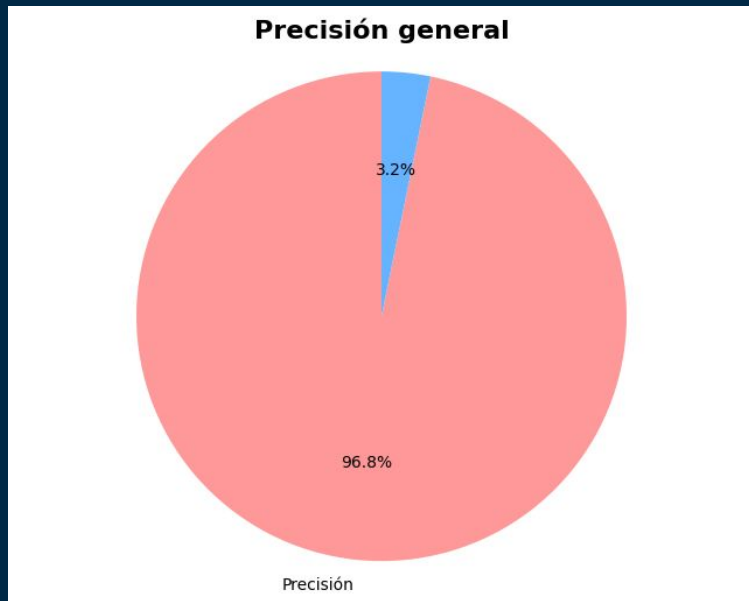


Diagrama de torta - precisión general

Diagrama de barras - precisión por categoría

# RESULTADOS

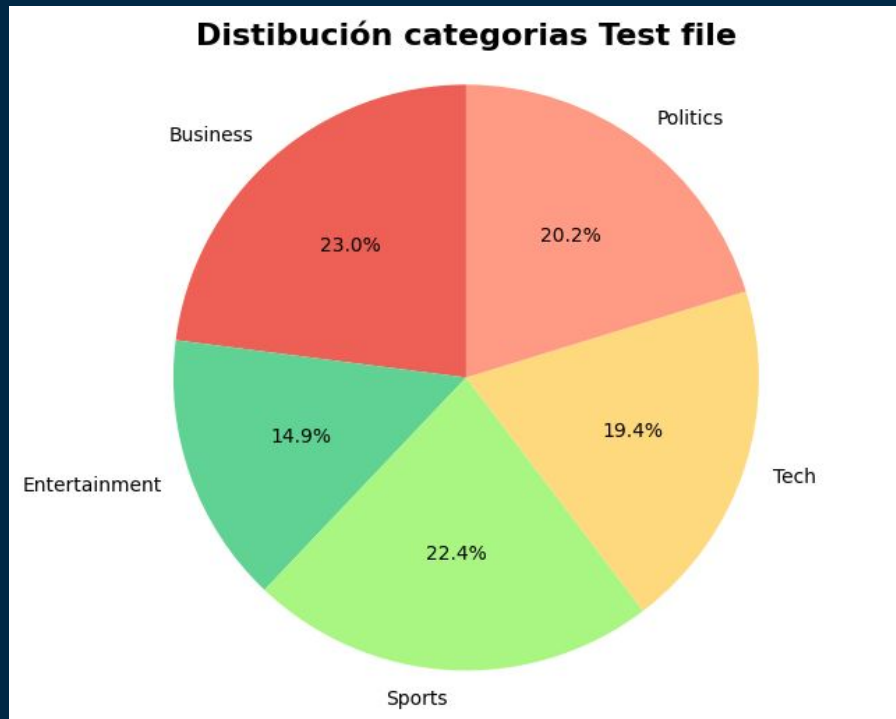
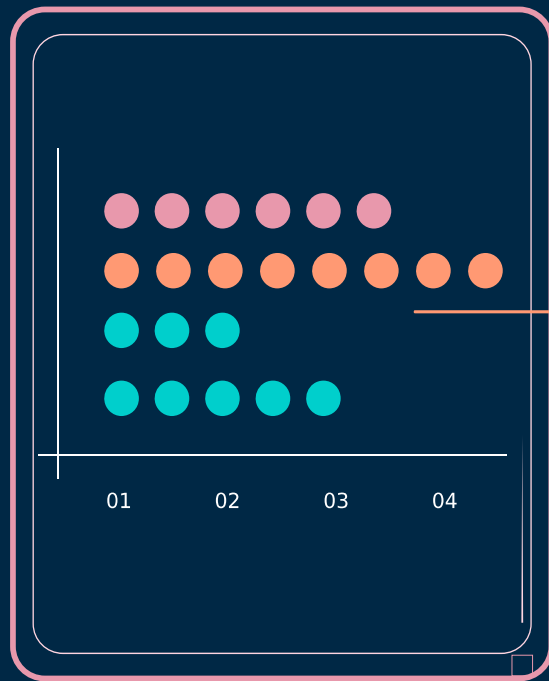


Diagrama de torta - distribución de noticias por categoría  
archivo de pruebas

# CONCLUSIONES

- El modelo alcanzó una precisión aceptable dadas las condiciones y al tratamiento simple que se le aplicó.
- La precisión de la categoría "sports" del 99% se puede atribuir a un lenguaje más diferenciable, caso contrario de "bussines" y "politics" que puede presentar ambigüedad.
- La clasificación se basa en asignar una categoría a un conjunto de datos basándose en ciertas características.
- A partir del conteo que se realiza a cada categoría se podría llegar a observaciones valiosas para el contexto donde se realice la predicción.



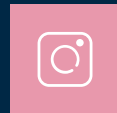
# REFERENCIAS

- [1] K. Sparck Jones, “A statistical interpretation of term specificity and its application in retrieval,” *J. Doc.*, vol. 28, no. 1, pp. 11–21, 1972.
- [2] R. Mosquera, O. D. Castrillón, and L. Parra, “Máquinas de Soporte Vectorial, Clasificador Naïve Bayes y Algoritmos Genéticos para la Predicción de Riesgos Psicosociales en Docentes de Colegios Públicos Colombianos,” *CIT Inform. Tecnol.*, vol. 29, no. 6, pp. 153–162, 2018.
- [3] “BBC news classification,” *Kaggle.com*. [Online]. Available: <https://www.kaggle.com/competitions/learn-ai-bbc/overview>. [Accessed: 17-May-2023].

Do you have any questions?

daniele.pazminoo@autonoma.edu.co  
juanf.mazos@autonoma.edu.co

# THANKS



CREDITS: This presentation template was created by [Slidesgo](#),  
including icons by [Flaticon](#), and infographics & images by [Freepik](#)  
Please keep this slide for attribution