



Procesamiento del lenguaje

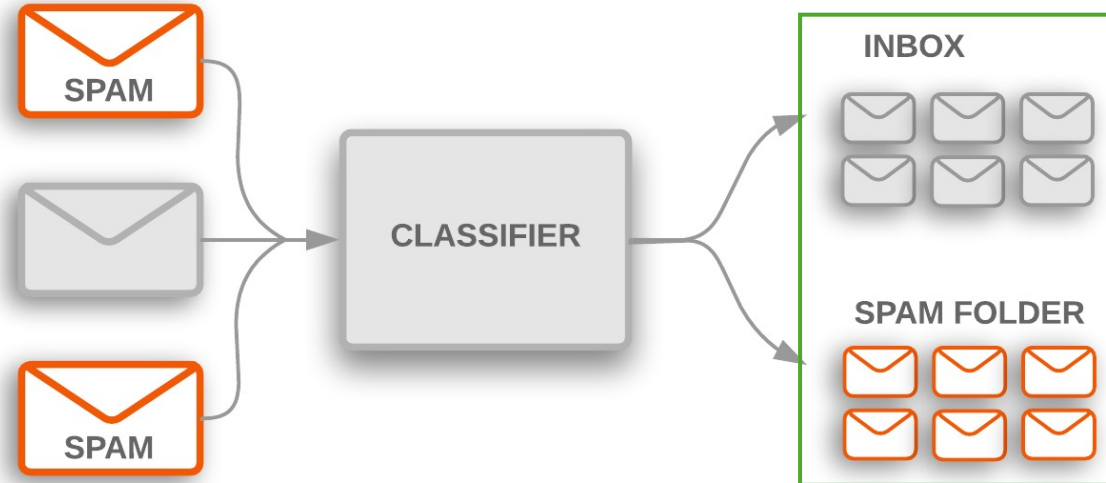


Ph.D. Janneth Chicaiza Espinosa

Clasificación de texto y análisis de sentimientos

Clasificación de textos

OBJETIVO: Construir modelos de aprendizaje automático que se enfoquen en el análisis de sentimientos y la clasificación de textos en categorías específicas.



- La era actual se caracteriza por la explosión de la información, clasificar grandes cantidades de datos de texto manualmente requiere mucho tiempo.
- La clasificación de texto es una tarea fundamental en el NLP e implica asignar categorías o etiquetas predefinidas a documentos de texto.

Aplicaciones de la clasificación de textos

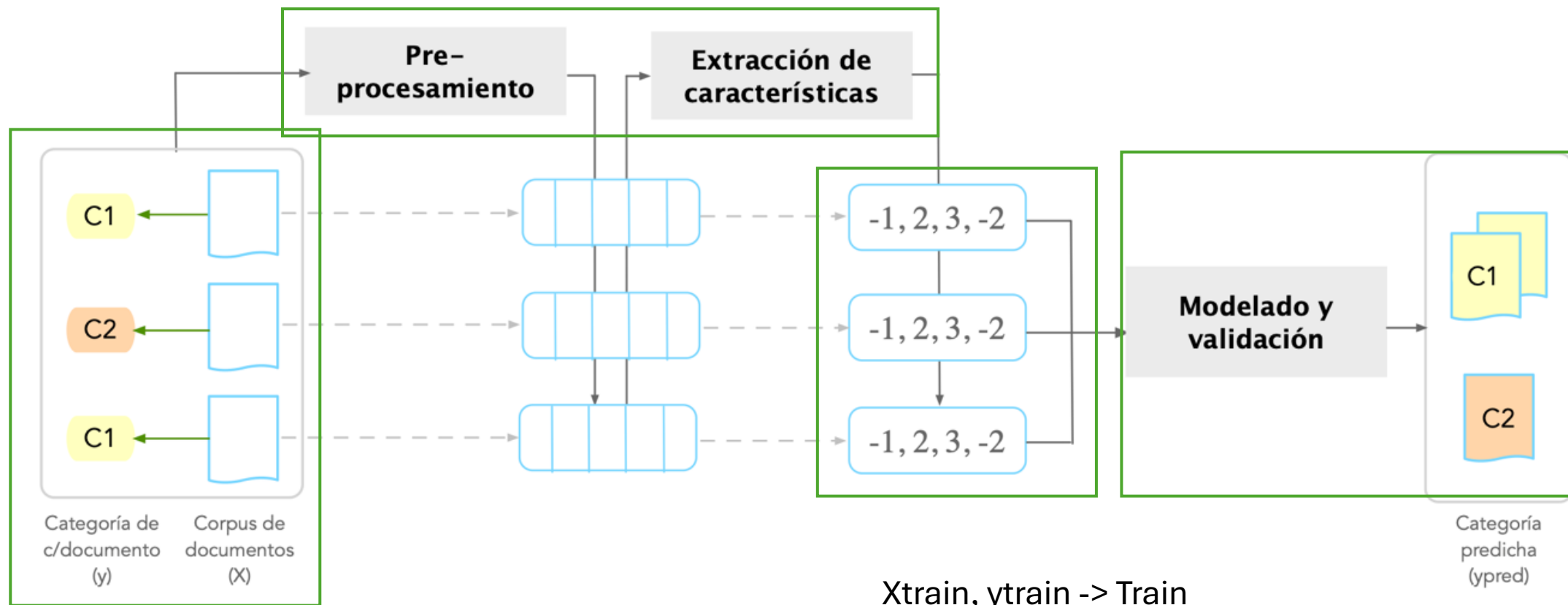
- Clasificar **significa** asignar una (o varias) etiqueta(s) o categoría(s) a un documento según su contenido. La clasificación de texto también puede ser útil en contextos, como:
 - En el ámbito de los catálogos o **bibliotecas de recursos digitales**, la clasificación por áreas temáticas es fundamental para orientar o [permitir a las personas la navegación, el filtrado y la búsqueda de recursos específicos en un área de conocimiento](#).
 - **Análisis de correos electrónicos:** La clasificación de correos puede ser realizada considerando al menos dos criterios, 1) categorizar correos electrónicos no deseados o no solicitados (filtrado de spam), y 2) categorizar correos asignando una prioridad de atención o tiempo de respuesta.
 - **Análisis de sentimientos:** En marketing, la clasificación de sentimientos es la base para realizar investigación de mercado, seguimiento de la reputación de marca en redes sociales y análisis de opiniones de clientes.
 - **Recomendación de contenido:** La clasificación automática del contenido de usuario permite generar recomendaciones personalizadas, como ocurre en plataformas de streaming de música, películas o libros.
 - **Identificación de temas emergentes:** Clasificar automáticamente grandes volúmenes de texto permite identificar temas emergentes o tendencias en redes sociales, noticias en línea, etc.

Tipos de clasificación de textos

- Dependiendo de la cantidad de categorías que se pueden asignar a los textos a clasificar, esta tarea puede ser de tres tipos: clasificaciones binarias, multiclase y multi-etiqueta. La Figura 3.3 ilustra en qué consiste cada tipo



Proceso de clasificación de textos



Xtrain, ytrain -> Train

Xtest, (ytest – ypred) -> test

Clasificación multiclase

Modelo	Descripción	Ventajas	Desventajas
Random Forest Classifier RandomForestClassifier()	Ensemble de árboles de decisión.	Alta precisión, manejo de datos desbalanceados, importancia de características.	Puede ser computacionalmente costoso para grandes conjuntos de datos, difícil de interpretar individualmente cada árbol.
Multinomial Naive Bayes MultinomialNB()	Modelo probabilístico basado en el teorema de Bayes.	Rápido, simple de implementar, adecuado para datos de texto con alta dimensionalidad.	Asume independencia entre las características.
K-Nearest Neighbors Classifier KNeighborsClassifier()	Clasifica un nuevo punto basado en los k vecinos más cercanos.	Simple, no requiere entrenamiento explícito.	Sensible a la elección de k y a la escala de los datos. Puede ser lento para grandes conjuntos de datos.
SVC (Support Vector Classifier) SVC()	Encuentra el hiperplano que mejor separa las clases.	Buen rendimiento en problemas de clasificación binaria y multiclase, capacidad de manejar datos no lineales con kernels.	Puede ser lento para grandes conjuntos de datos, elección de parámetros (kernel, C) puede ser compleja.
Gaussian Naive Bayes GaussianNB()	Similar a MultinomialNB, pero asume una distribución gaussiana para las características numéricas.	Rápido, simple de implementar.	Asume independencia entre las características y una distribución gaussiana.

Evaluación del clasificador

Clasificador binario

```
# Generar predicciones y matriz de confusión:  
y_pred = svm_model.predict(X=X_test)  
  
print(metrics.confusion_matrix(y_test, y_pred))  
  
# Ver métricas por clase:  
print(metrics.classification_report(y_test, y_pred))
```

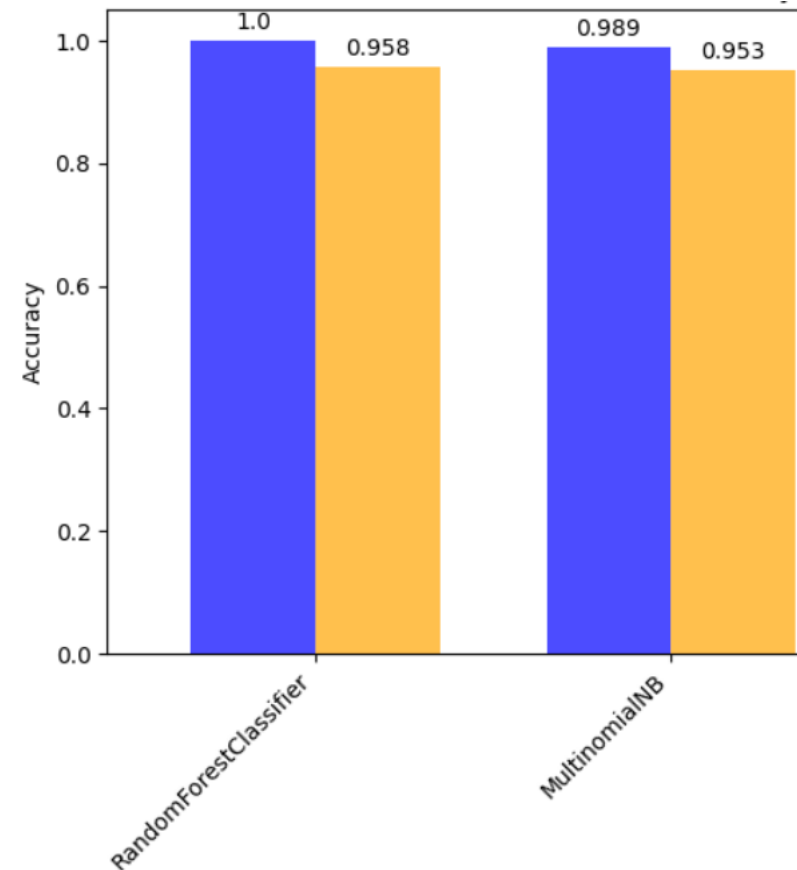
Tiempo de ejecución: 20.66 segundos

[[961 4]

[24 126]]

	precision	recall	f1-score	support
ham	0.98	1.00	0.99	965
spam	0.97	0.84	0.90	150
accuracy			0.97	1115
macro avg	0.97	0.92	0.94	1115
weighted avg	0.97	0.97	0.97	1115

Clasificador multiclase (accuracy)

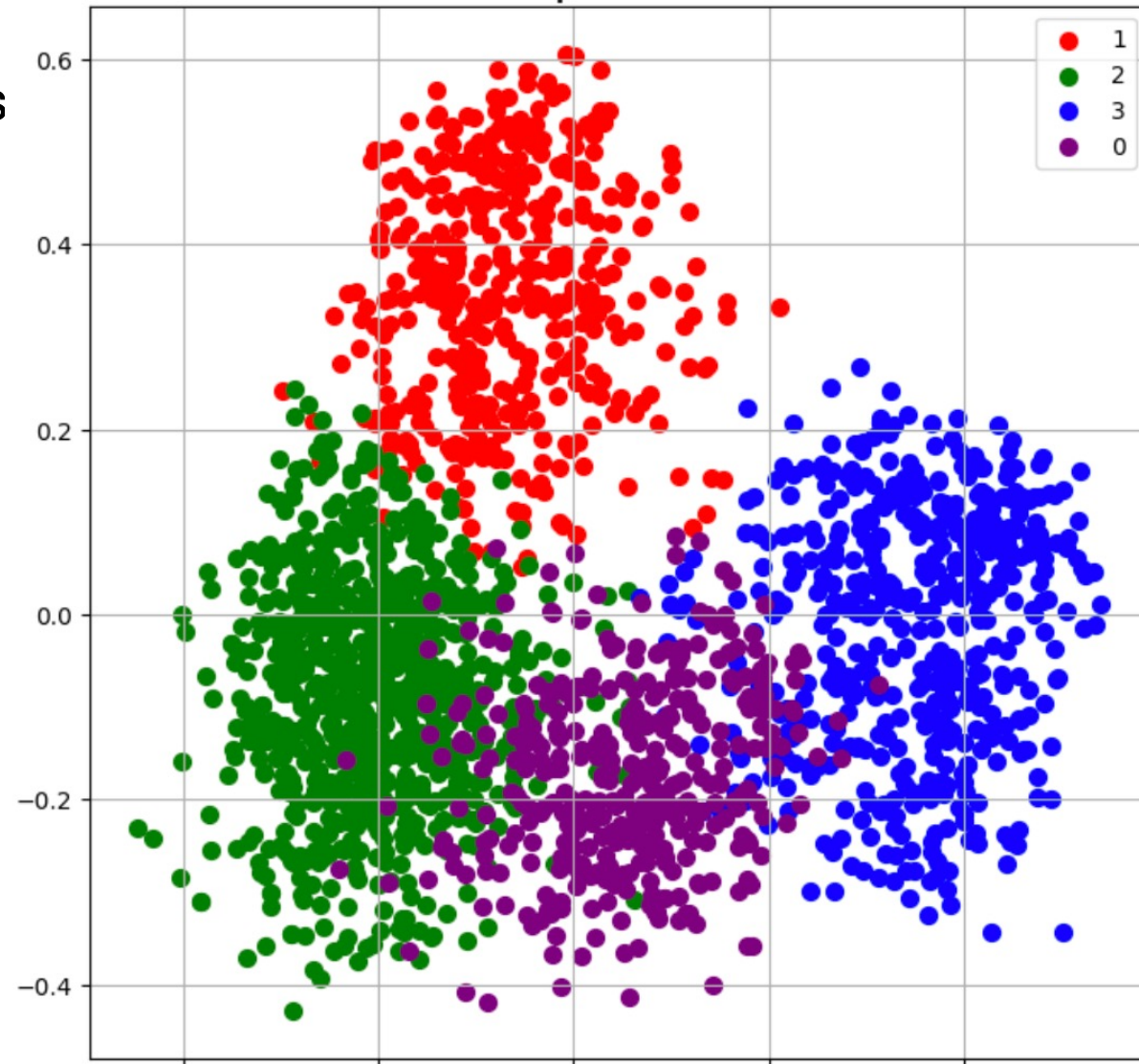


Accuracy:
proporción de predicciones correctas que el modelo realiza sobre un conjunto de datos

Luego de la creación de un modelo de clasificación debemos evaluarlo, desde el punto de vista de su rendimiento, utilizando para ello un conjunto de datos de prueba, diferente del utilizado durante el entrenamiento.

Técnicas no supervisadas

- **¿Cuándo usar?** Cuando no conocemos la etiqueta o categoría a la que pertenece cada documento o es complicado generar datasets etiquetados.
- **Clustering de texto:** consiste en agrupar documentos de texto similares en función de su contenido.
 - Algoritmos de agrupamiento: **K-means**, DBSCAN, **HDBSCAN**, y otros.
- **Topic Modeling:** trata de descubrir temas latentes presentes en una colección de documentos de texto.
 - Latent Dirichlet Allocation (**LDA**).



Técnicas de clustering útiles en texto

Característica	K-means	HDBSCAN
Definición	Agrupar datos en un número predefinido (k) de clusters.	Algoritmo basado en densidad: crea una jerarquía de clusters, identificando clusters de diferentes densidades.
Número de clusters	Debe ser especificado por el usuario.	Determinado automáticamente basado en la densidad de los datos.
Forma de los clusters	Asume clusters de forma convexa y aproximadamente iguales.	Puede identificar clusters de cualquier forma y tamaño, incluyendo clusters no convexos y de densidad variable.
Valores atípicos	Tiende a asignar valores atípicos a clusters cercanos, lo que puede distorsionar los resultados.	Identifica y etiqueta los valores atípicos como ruido (-1).
Complejidad computacional	Relativamente eficiente para conjuntos de datos de tamaño mediano.	Puede ser más costoso computacionalmente, especialmente para conjuntos grandes y densos.
Parámetros clave	Número de clusters (k).	Tamaño mínimo de clusters (min_cluster_size), número mínimo de muestras para formar un cluster (min_samples).

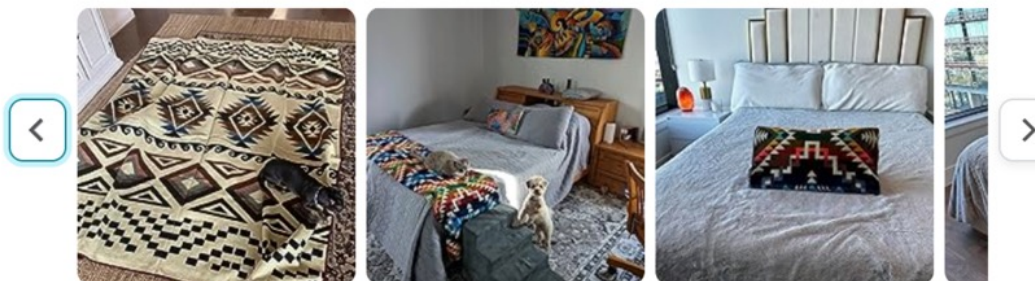
Análisis de sentimientos

Rating



Opiniones con imágenes

[Ver todas las fotos](#)



Opiniones destacadas de los Estados Unidos

Producto



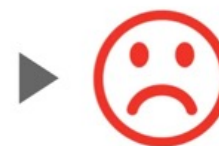
Opiniones

Increíble calidad y supersuave. Es muy delgada pero muy cálida.

Pedí esta manta para mi esposo, que estaba pasando por algunos tratamientos médicos.

Es una manta muy, muy fina... la descripción dice «gruesa». **Publicidad engañosa.**

Sentimiento




How will your business succeed if you don't know how customers feel about your brand?

- ✓ Consiste en extraer el sentimiento (opinión, percepción, actitud), con la cual la gente se expresa a través de comentarios o publicaciones, como los reviews.

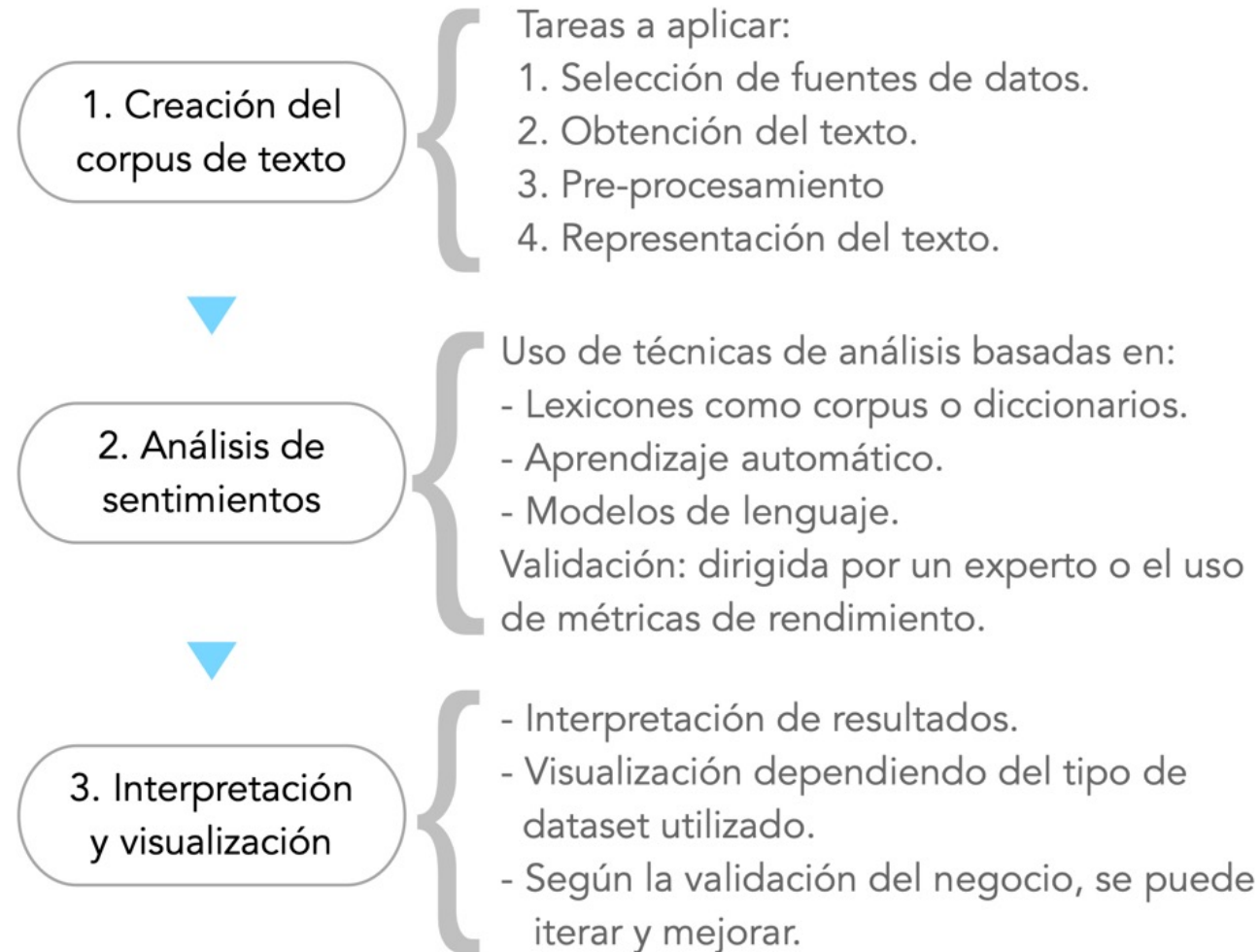
Niveles de análisis

Documento	Se determina la polaridad de un documento de texto. Un documento puede representar a un libro, un capítulo, una noticia, el resumen de un paper, o la reseña completa de un producto. Es el nivel más abstracto de análisis.
Oración	Se realiza el análisis de sentimientos a nivel de oración. En este caso, es necesario identificar a las diferentes unidades gramaticales de un documento. Es útil cuando un documento tiene una variedad y de sentimientos asociados.
Frase	La polaridad se determina a nivel de frase o fragmento de texto. Cada frase puede contener múltiples aspectos o aspectos únicos. Por ejemplo, el servicio del restaurante fue <i>excelente</i> (+), pero, la comida fue <i>decepcionante</i> (-).
Aspecto	El análisis se enfoca en los aspectos específicos mencionados en un texto. Por ejemplo, el servicio (<i>aspecto 1, +</i>) del restaurante fue excelente, pero, la comida (<i>aspecto 2, -</i>) fue decepcionante.



Nivel de detalle

Proceso de análisis de sentimientos

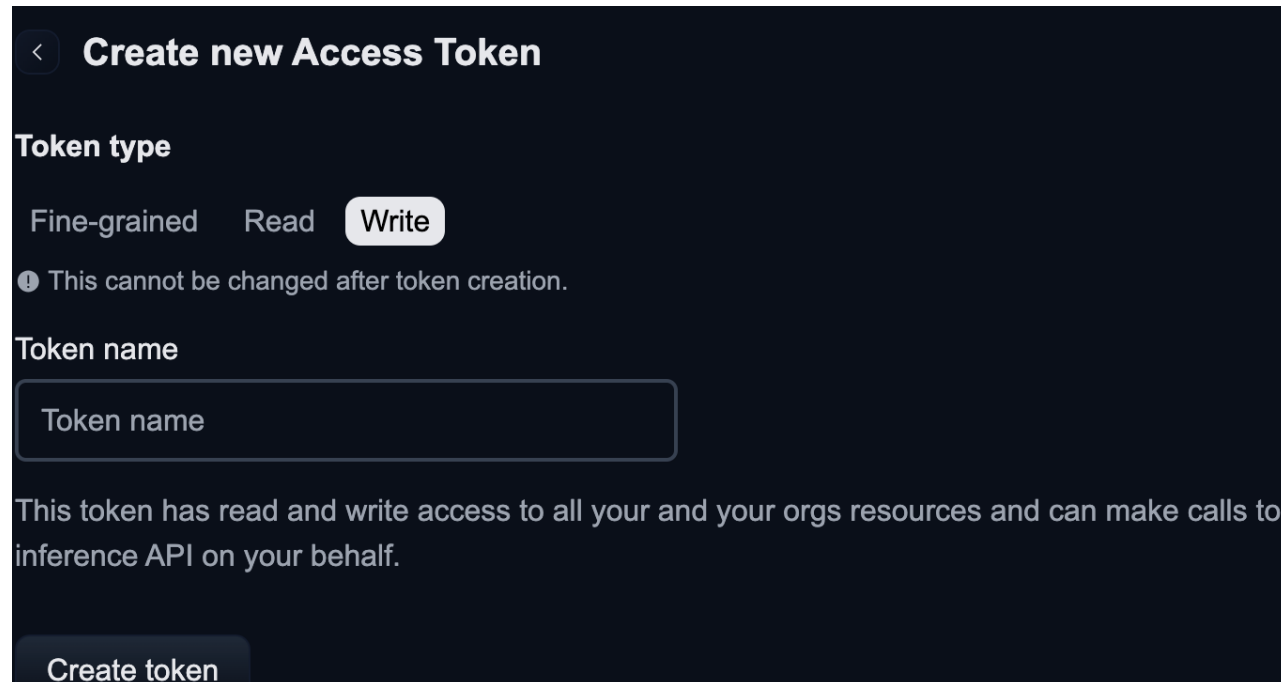


Requisitos para la UNIDAD 4 (1)

- **Gemini (Google)**





- Solicitar la API Key en: <https://aistudio.google.com/>

- Crear cuenta en HuggingFace: <https://huggingface.co/>
Crear un Access Token de escritura.



The screenshot shows the 'Create new Access Token' page on Hugging Face. At the top, there is a back arrow and the title 'Create new Access Token'. Below this, the 'Token type' section has three buttons: 'Fine-grained', 'Read', and 'Write'. The 'Write' button is selected and highlighted. A warning icon and text state: 'This cannot be changed after token creation.' The 'Token name' section features a text input field with the placeholder 'Token name'. Below the input field, a descriptive text reads: 'This token has read and write access to all your and your orgs resources and can make calls to inference API on your behalf.' At the bottom, there is a 'Create token' button.

Requisitos para la UNIDAD 4 (2)

Gated Repos Status			
View the gated repositories that you have requested access to.			
<input type="text" value="Filter by repo name"/>			
Repo Name	Type	Date	Request Status
Meta's Llama 3.1 models & evals ⓘ	 models group	Oct 26	ACCEPTED
Meta's Llama 3 models ⓘ	 models group	Aug 21	ACCEPTED
mistralai/Mistral-7B-v0.1	 model	Aug 6	ACCEPTED
Google's Gemma models family ⓘ	 models group	Feb 23	ACCEPTED

- **Solicitar autorización para los repositorios:**
 - Llama 3.1: <https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>
 - Llama 3: <https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>
 - Gemma: <https://huggingface.co/google/gemma-2b-it>

Gracias



uhemisferios



uhe.oficial



uhe_oficial



Universidad Hemisferios

UHE

uhemisferios.edu.ec