

Clasificación de textos

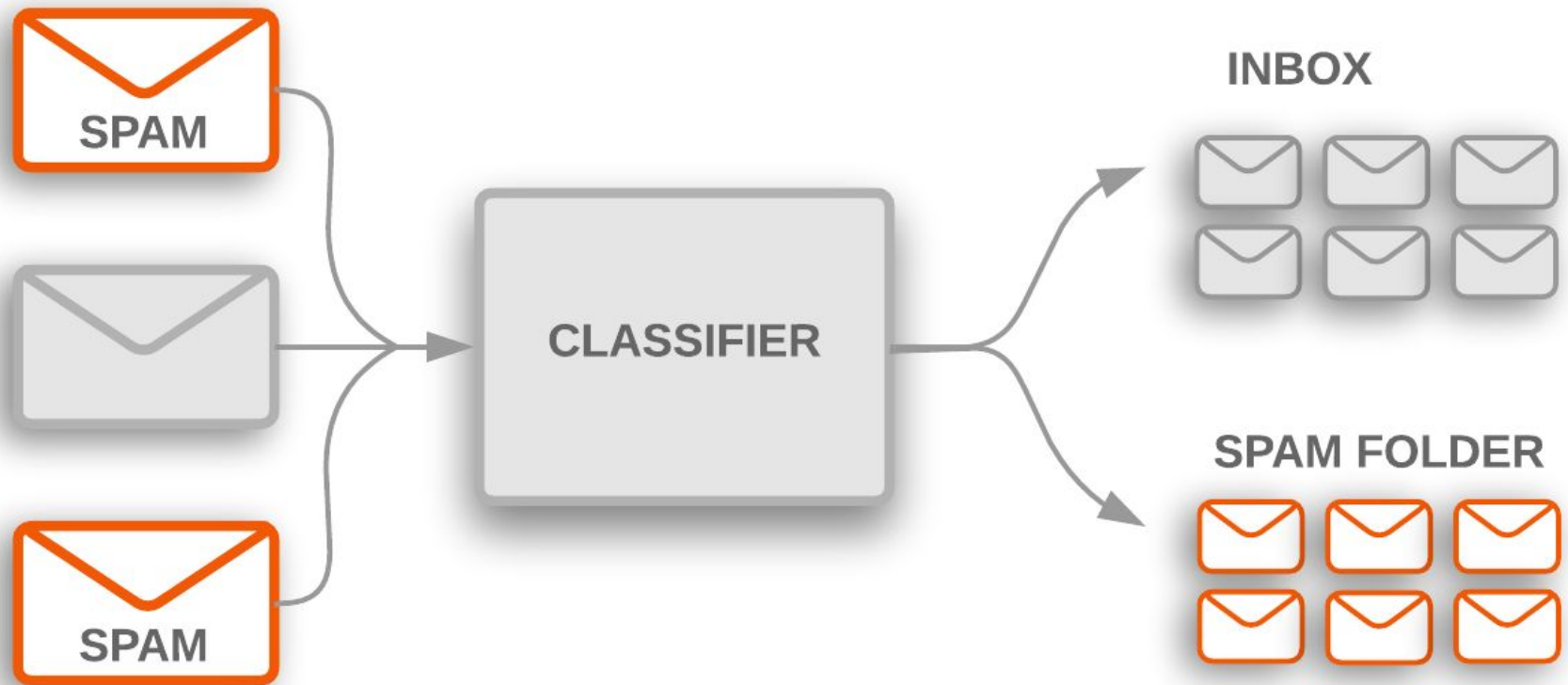
Tecnologías Emergentes en la Sociedad de
la Información

Qué es clasificación de textos?

- Consiste en asignar a un documento una o varias categorías (clases).

Ejemplos de aplicaciones

Identificar si un email es spam o no



Identificar idioma de un texto



Clasificación de noticias



Sentiment Analysis



Sentiment Analysis

amazon.es

prime

Enviar a Isabel Leganes 28914

Todos los departamentos ▾

Volver a comprarAmazon.es de IsabelOfertasPrime Now y La Plaza de DIAChollosCheques regaloVenderAtención al ClienteHogar y cocinaInformación


Amazon.esLos más vendidosChollosOfertasProductos ReacondicionadosLista de deseosCheques regaloAmazon PrimeApps de AmazonVender en AmazonTrabajar en Amazon

Rainbow Socks - Hombre Mujer Calcetines Deporte Colores de Algodón -... > Opiniones de clientes

Opiniones de clientes

★★★★☆ 4,3 de 5 ▾

620 valoraciones de clientes



Rainbow Socks - Hombre Mujer Calcetines Deporte Colores de Algodón - Pares - - Talla UE

por Rainbow Socks

Precio: 16,99 € - 24,99 € + Envíos gratis con Amazon Prime

5 estrellas

61%

4 estrellas

22%

3 estrellas

9%

2 estrellas

2%

1 estrella

6%

Escribir una opinión

☆☆☆☆☆ Calcetines para verano

11 de junio de 2018

Tamaño: 39/41 | Color: 12 X Multicolor | **Compra verificada**

Al verlos lo primero que pensé son demasiado finos y me van a irritar los pies, pero la verdad he quedado muy contenta con ellos, no me irritan y se lavan muy bien. yo los uso para diario (camino bastante) para hacer deporte no se que tal serán.

Actualización después de 6 meses

Al principio estaba muy contenta, pero después de varios lavados se han deteriorado muy rápidamente.

A 2 personas les ha parecido esto útil

Útil

Comentar

Informar de un abuso



Anna Nardi Fontanals

☆☆☆☆☆ Perfectos para caminar

27 de octubre de 2018

Tamaño: 39/41 | Color: 12 X Multicolor | **Compra verificada**

Me gustan los colores, la cantidad de calcetines y el precio, pero sobretodo son comodoss, se ajustan a mi pie genial y se mantienen fijos en los pies. Camino mucho durante el día y con éstos voy muy cómoda, absorbe bien el sudor y NO me hace rozaduras ni nada. Lo único que son un poco finos y no sé que tal me irán ahora en invierno.

A una persona le ha parecido esto útil

Útil

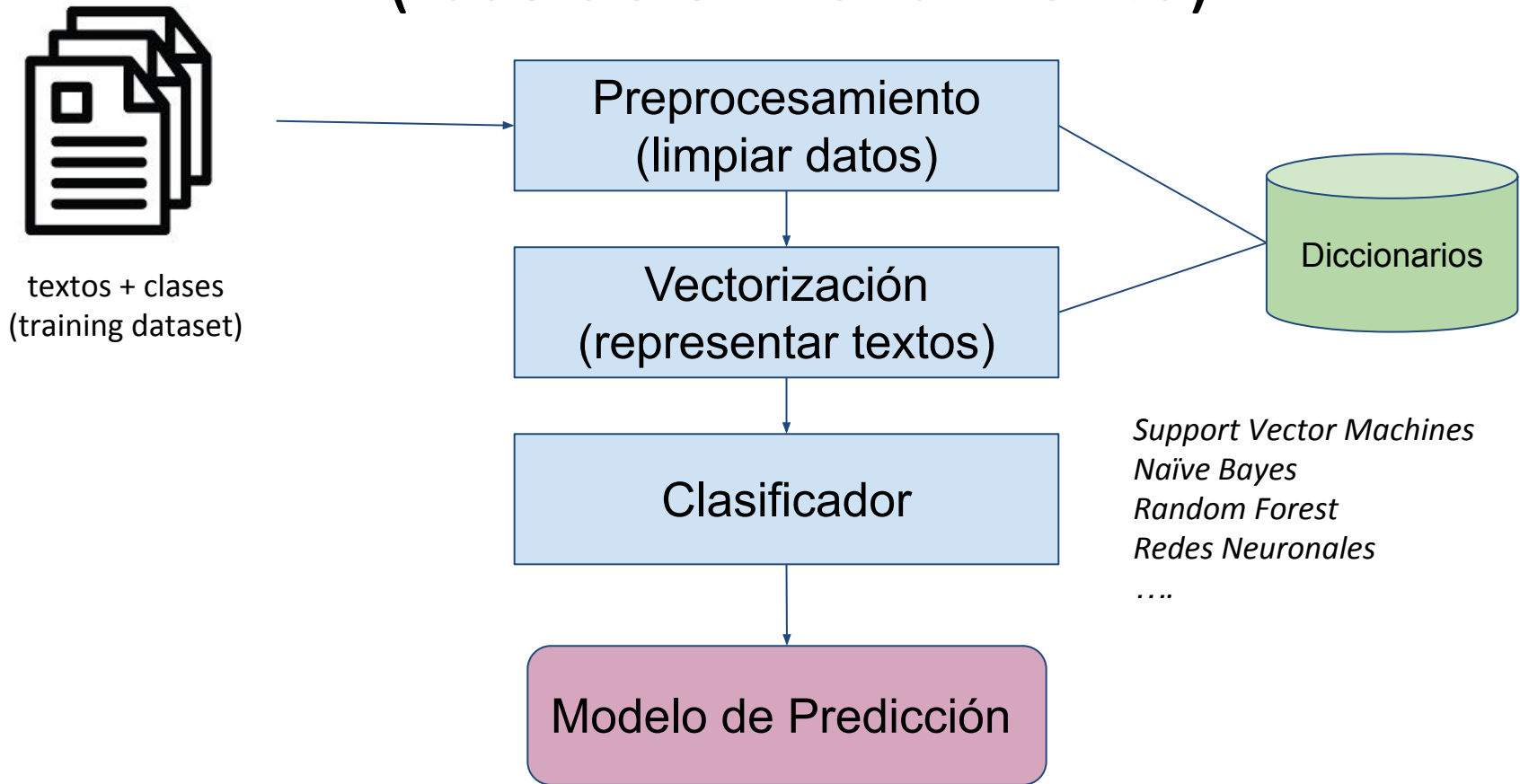
Comentar

Informar de un abuso

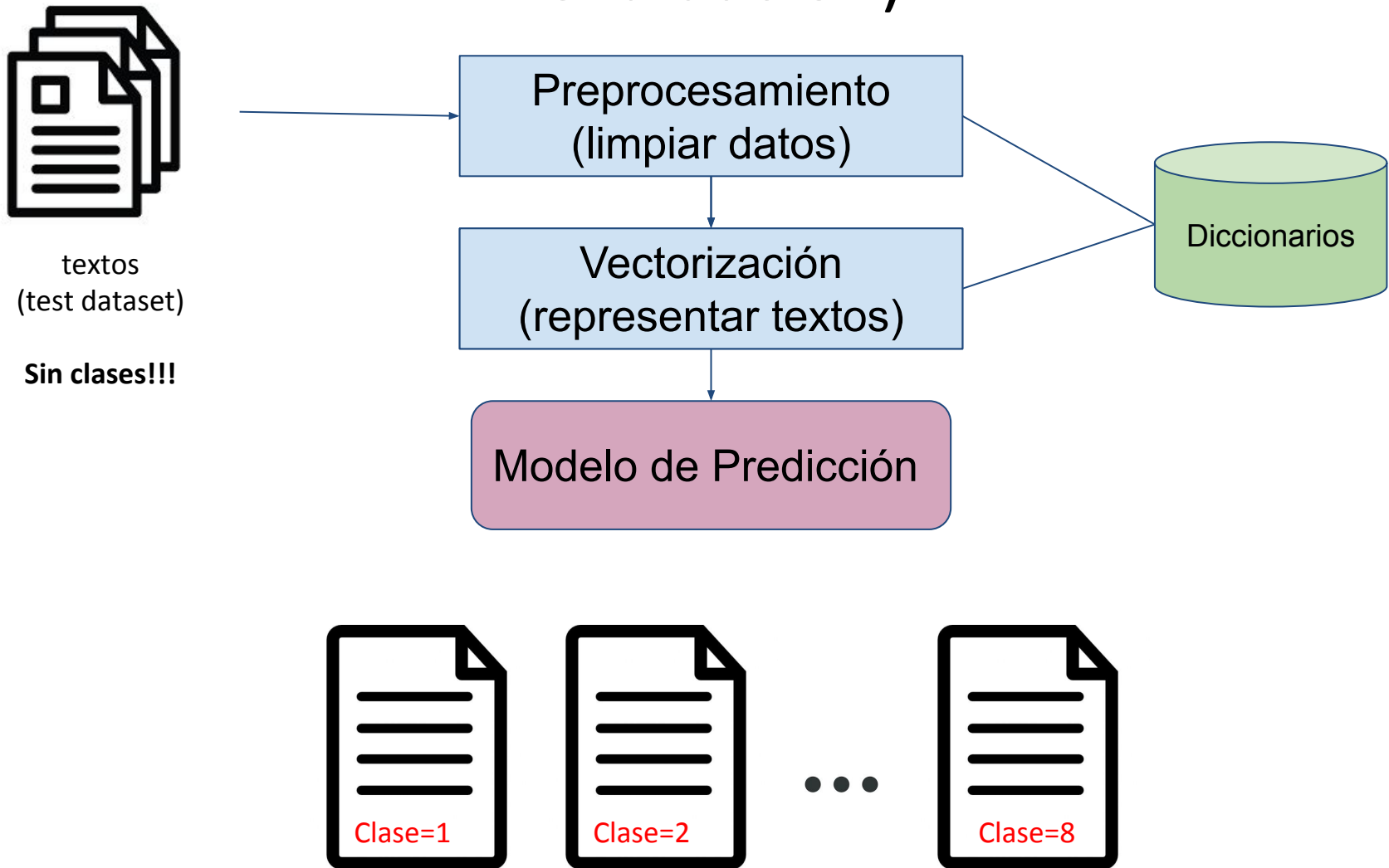
Más ejemplos

- Identificar mensajes obscenos o ofensivos (redes sociales).
- Detectar casos de anorexia.
- Detectar casos de posibles suicidios.
- Clasificación de notas clínicas de pacientes para estimar las probabilidades de metástasis.
- Identificación de cohortes de pacientes para ensayos clínicos.
-

Arquitectura Clasificación de Textos (fase de entrenamiento)

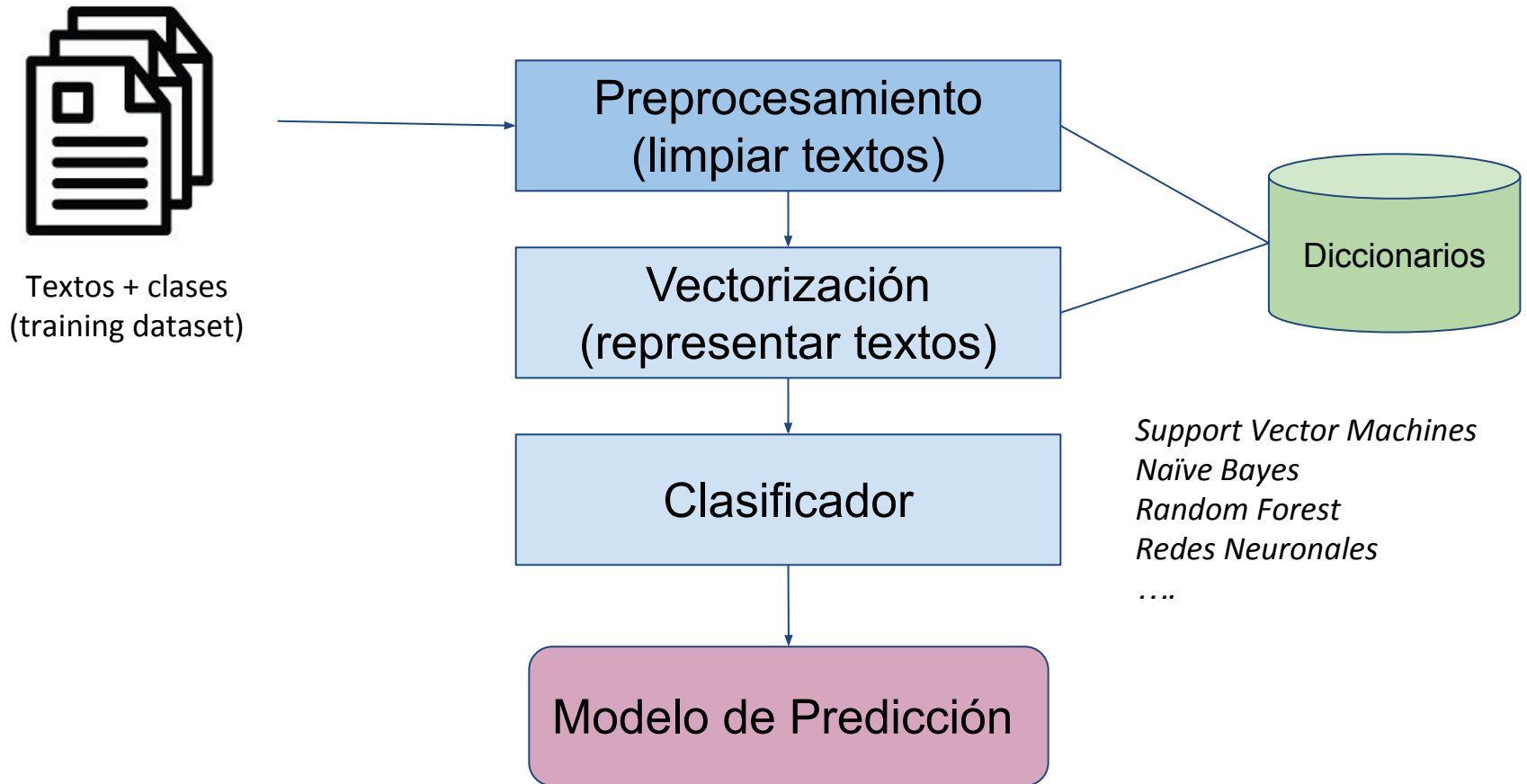


Arquitectura Clasificación de textos (fase de evaluación)

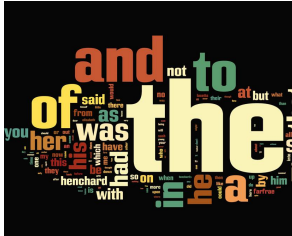


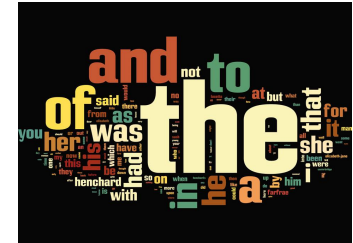
Predicción de las clases para cada texto del test dataset

Arquitectura (fase de entrenamiento)



Preprocesamiento

- Limpiar los textos de palabras innecesarias.
 - Las tareas más comunes son:
 - 1) Tokenización.
 - 2) Eliminar signos de puntuación y caracteres especiales.
 - 3) Eliminar stopwords
 - 4) Stemming o lematización.
- 

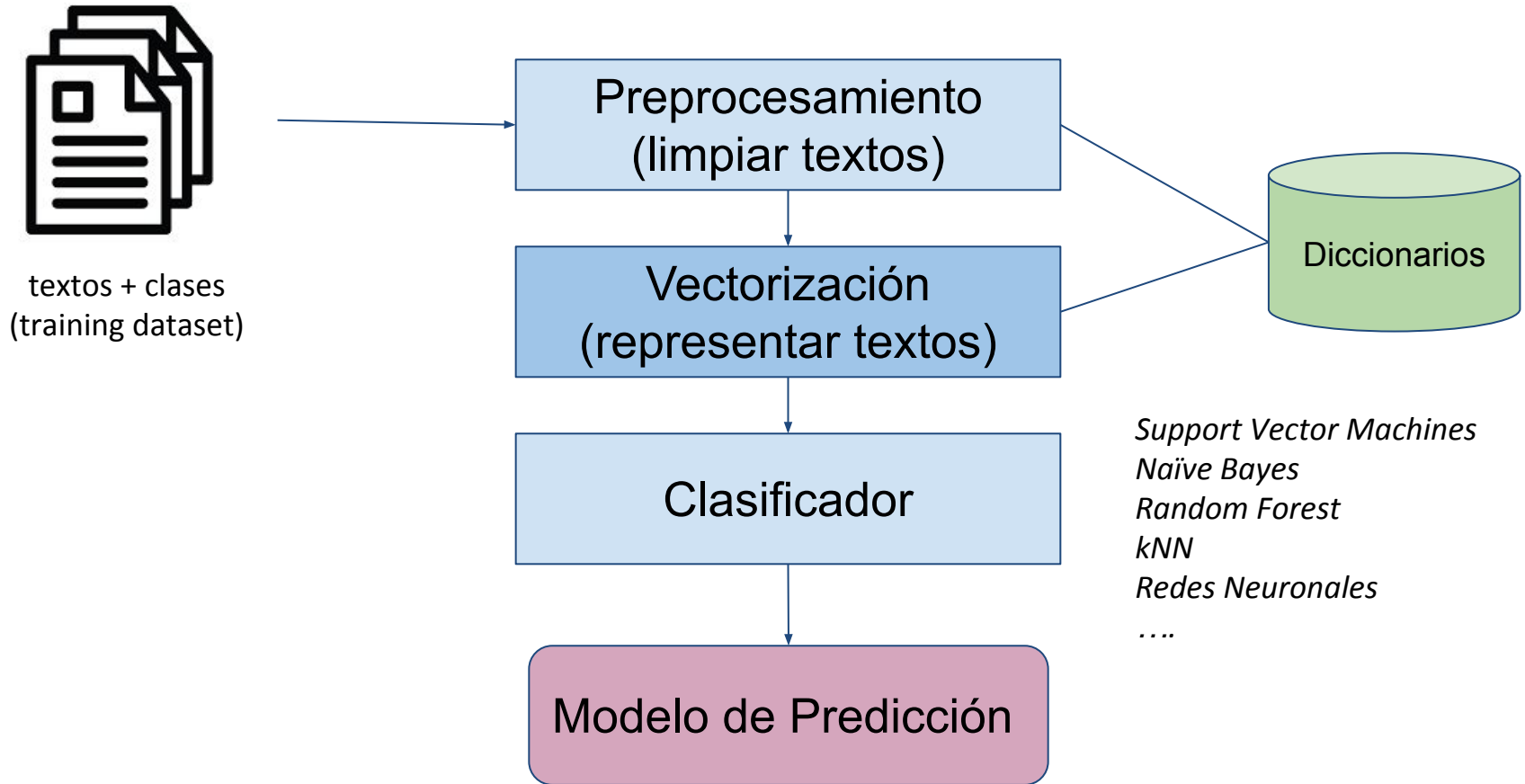


Preprocesamiento: Lematización vs Stemming

- Disminuyen la variabilidad léxica.
 - **Lematización:** obtener el lema o forma canónica de una palabra.
 - *caros, caras, carísimo -> caro*
 - **Stemming:** obtener la raíz (stem) de la palabra.
 - *alojamos, alojaremos, alojé -> aloj-*

- Lematizador online: <http://www.gedlc.ulpgc.es/investigacion/scogeme02/lematiza.htm>
- Stemmer online: <https://snowballstem.org/demo.html>

Arquitectura (fase de entrenamiento)



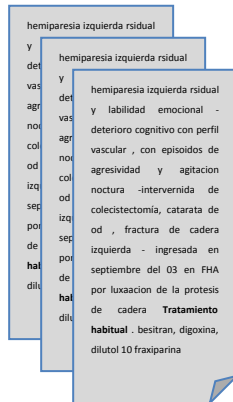
Vectorización

- Cada texto es representado como un conjunto de características (ej, número de palabras positivas, número de palabras negativas).
- Transformar textos en vectores de números.
- Los enfoques más utilizados son:
 - **Bolsa de Palabras.**
 - **TF-IDF**
 - **Word Embeddings**

Vectorización: Bolsa de Palabras (Bag of Words)

Pasos:

1. Preprocesamiento (eliminar stopwords y lematización).
2. Se obtiene el vocabulario (lista de palabras distintas) de todos los documentos.



Vectorización: Bolsa de Palabras

3. Cada documento se representa como un vector de las frecuencias de sus palabras.

D1: ~~El~~ gato grande ~~está en la~~ mesa ~~y el~~ gato pequeño ~~en la~~ ventana:

D2: ~~La~~ mesa ~~y la~~ ventana ~~son~~ pequeños:

D3: ~~La~~ luna ~~y el~~ árbol ~~son~~ grandes:

Vectores (features):

	árbol	balón	gato	grande	luna	mesa	pequeño	ventana	zoo
D1	0	0	2	1	0	1	1	1	0
D2	0	0	0	0	0	1	1	1	0
D3	1	0	0	1	1	0	0	0	0

Vectorización: TF-IDF

- Extensión del modelo de bolsa de palabras.
- Cada documento es representado con la TF-IDF de sus palabras.
- Esta métrica, TF-IDF, consigue **disminuir** el **peso** de las **palabras** que son muy **comunes** en toda la colección de documentos.

Vectorización: TF-IDF

- Term frequency - inverse document frequency.

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

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

<https://mropengate.blogspot.com/2016/04/tf-idf-in-r-language.html>

Vectorización:TF-IDF

D1: El gato grande está en la mesa y el gato pequeño en la ventana-

D2: La mesa y la ventana son pequeños-

D3: La luna y el árbol son grandes-

REPRESENTACIÓN USANDO BOLSA DE PALABRAS

	árbol	balón	gato	grande	luna	mesa	pequeño	ventana	zoo
D1	0	0	2	1	0	1	1	1	0
D2	0	0	0	0	0	1	1	1	0
D3	1	0	0	1	1	0	0	0	0

REPRESENTACIÓN USANDO TF-IDF

	árbol	balón	gato	grande	luna	mesa	pequeño	ventana	zoo
D1	0	0	0.95	0.17	0	0.17	0.17	0.17	0
D2	0	0	0	0	0	0.17	0.17	0.17	0
D3	0.47	0	0	0.17	0.47	0	0	0	0

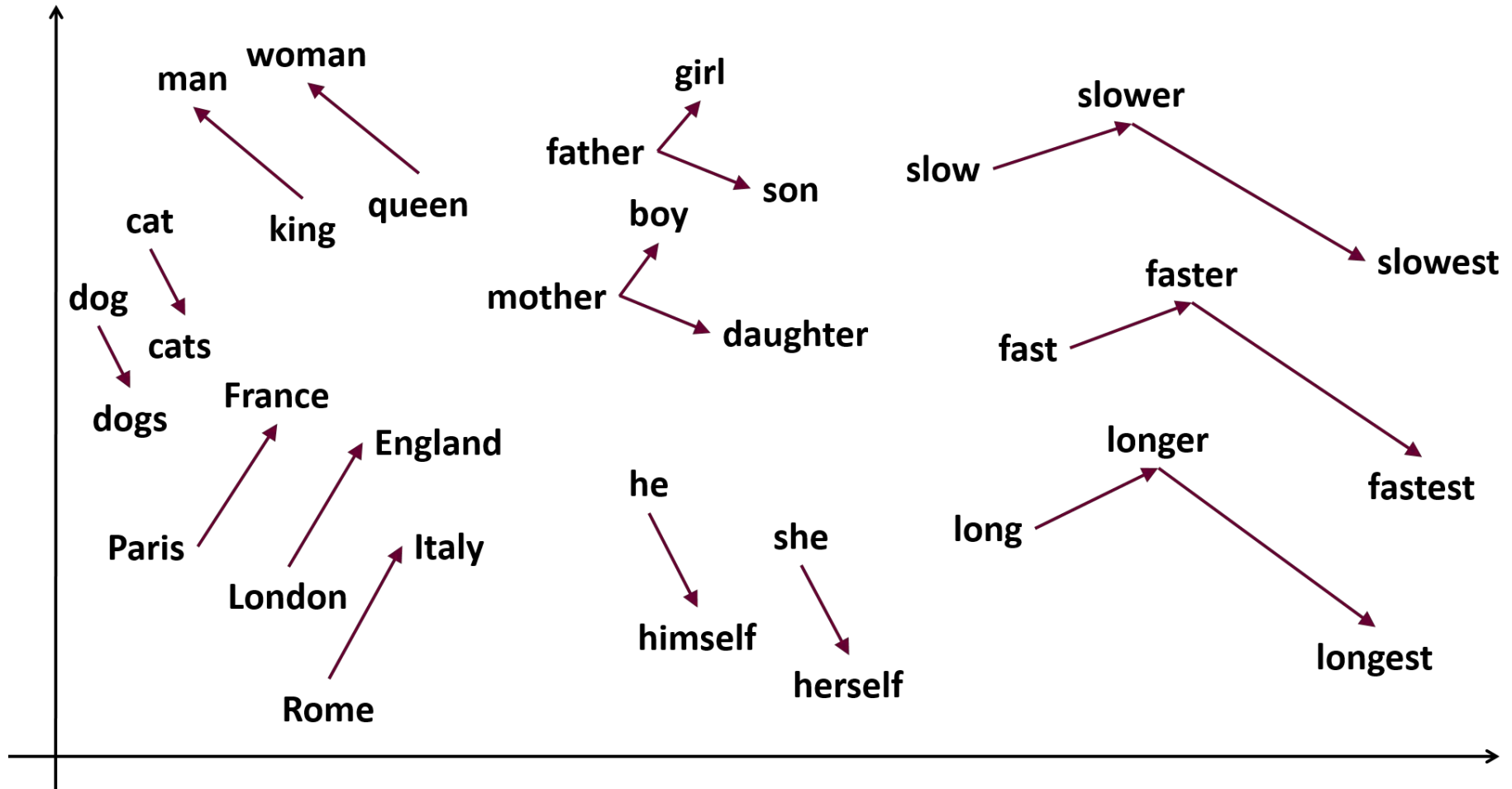
Limitaciones de Bolsa de Palabras y TF-IDF

- Vectores con una **alta dimensionalidad** y muy **dispersos**.
- NO capturan el **orden de las palabras**.
“El hotel era bueno y no era caro” != “El hotel era caro y no era bueno”.
- NO capturan **similitudes semánticas**:
“precio carísimo” ~ “importe prohibitivo”

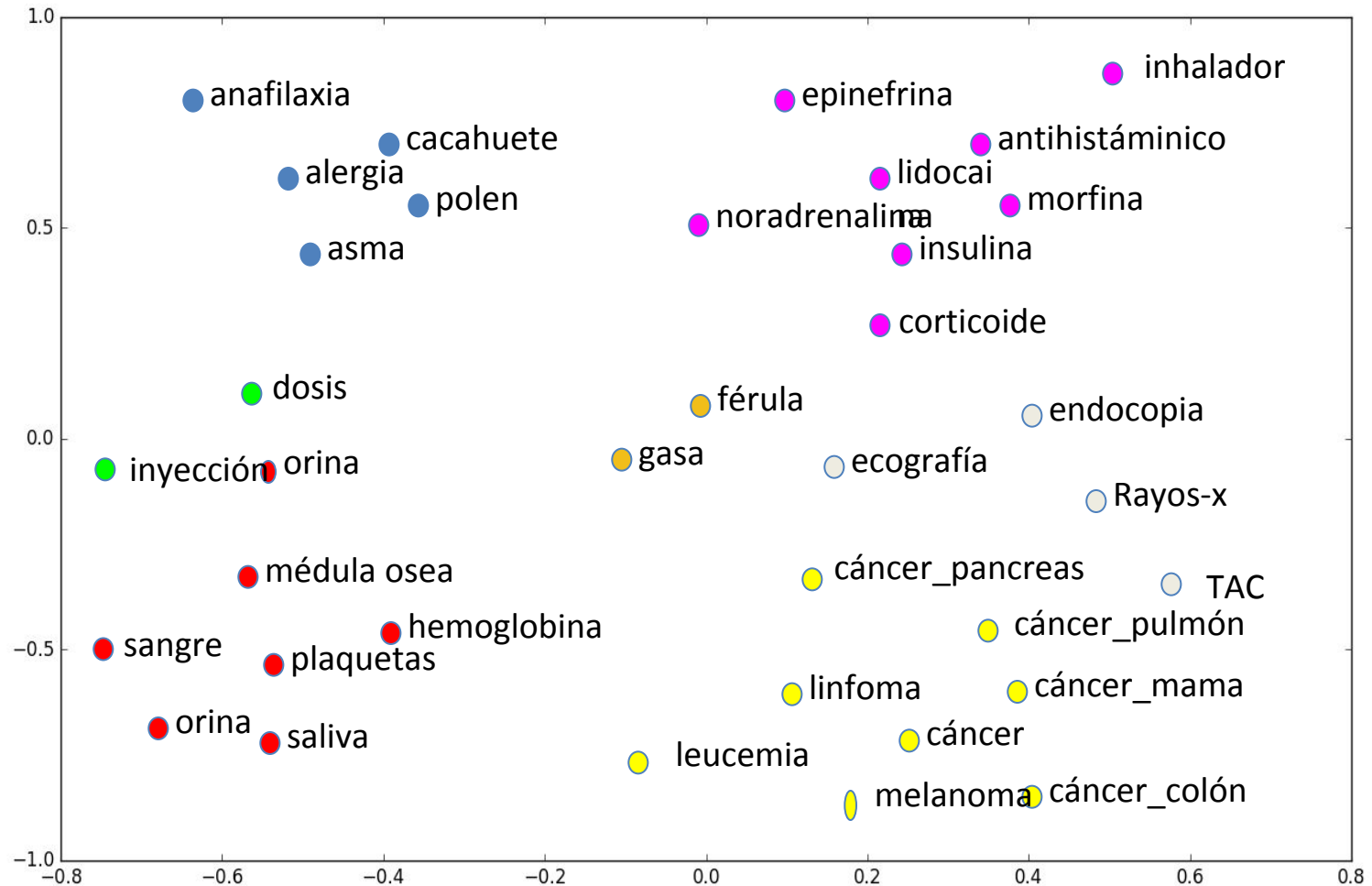
Vectorización: Word Embeddings

- Se construye a partir de una gran colección de documentos.
- Representación de palabras en vectores, capaces de capturar las relaciones semánticas y sintácticas entre las palabras.
- Herramientas: Word2Vec, FastText, Glove
- Demo: http://bionlp-www.utu.fi/wv_demo/

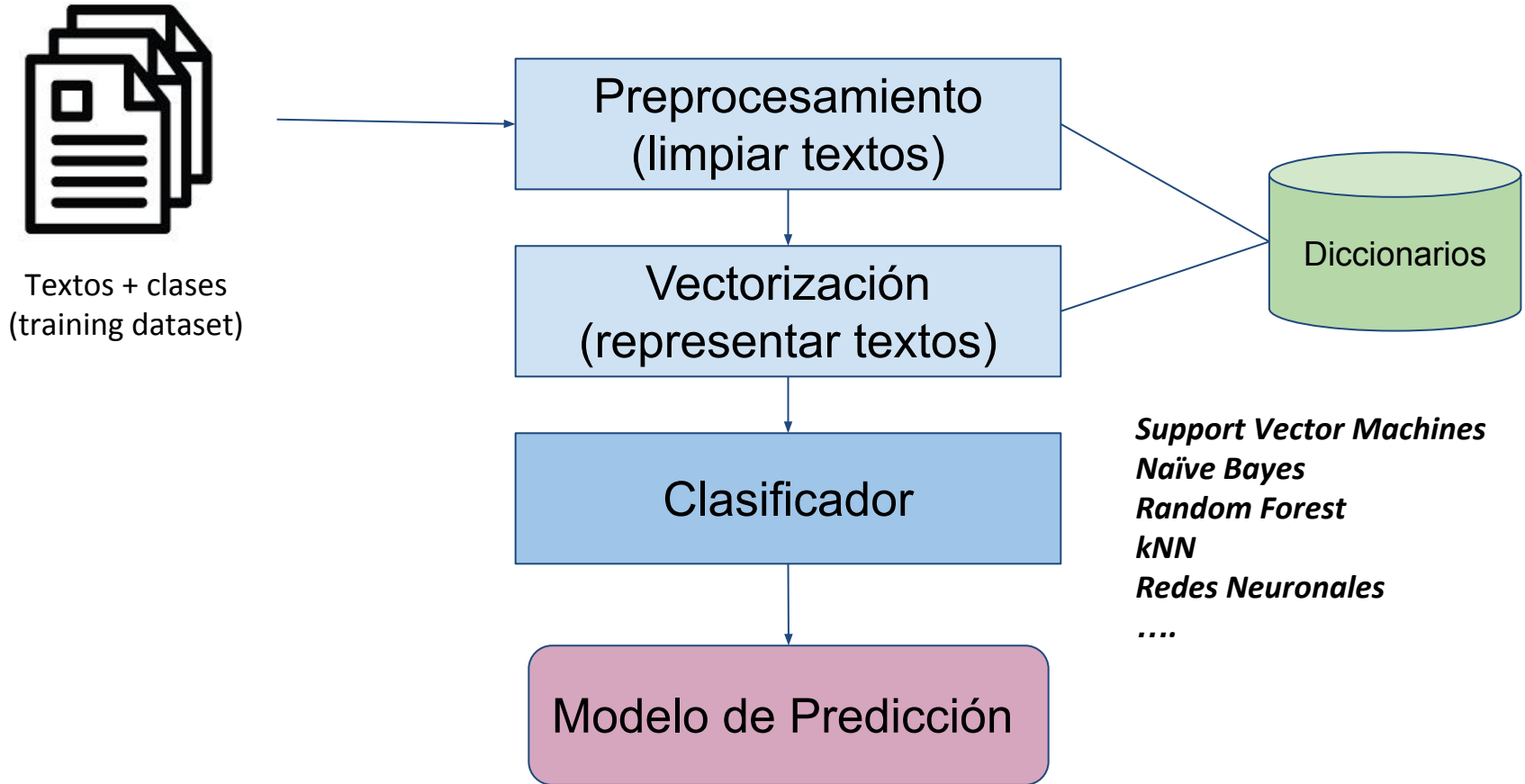
Word Embeddings



Word Embeddings



Arquitectura (fase de entrenamiento)



Algoritmos de Clasificación

- kNN
- Support Vector Machines
- Tree decisions and Random Forest
- MLP
- Deep Learning models.
- ...

Algoritmos: Naïve Bayes

$$C_{MAP} = \operatorname{argmax}_{c \in C} P(c | d)$$

$C = \{\text{Positivo}, \text{Negativo}\}$

$$= \operatorname{argmax}_{c \in C} \frac{P(d | c)P(c)}{P(d)}$$

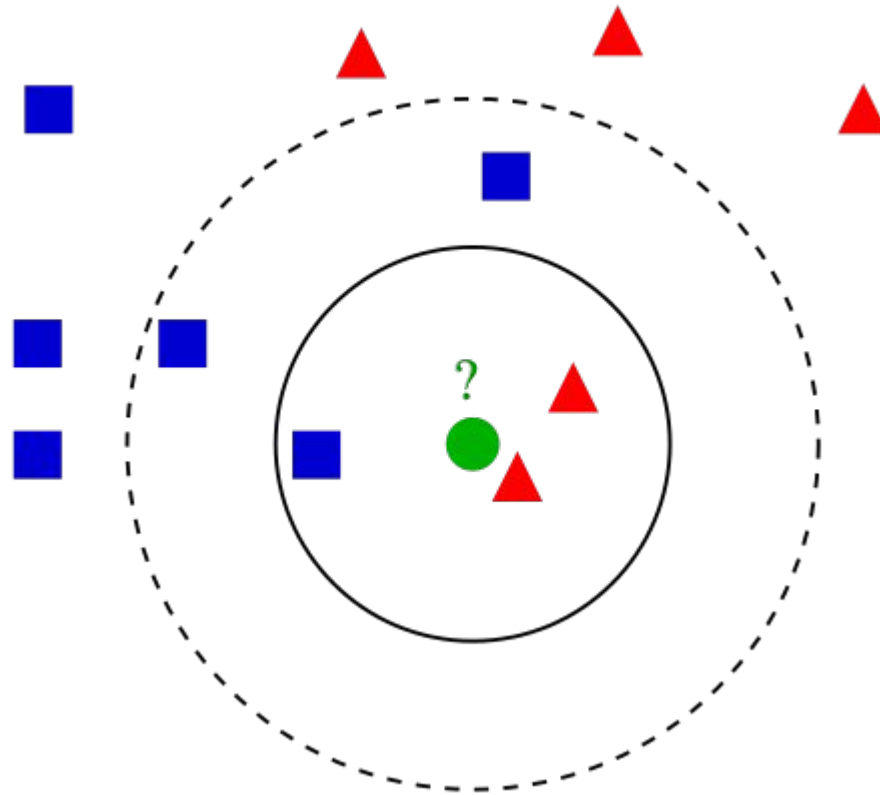
$$= \operatorname{argmax}_{c \in C} P(d | c)P(c)$$

$$= \operatorname{argmax}_{c \in C} P(x_1, x_2, \dots, x_n | c)P(c)$$

$$C_{NB} = \operatorname{argmax}_{c \in C} P(c_j) \prod_{x \in X} P(x | c)$$

Algoritmos: k-Nearest Neighbor

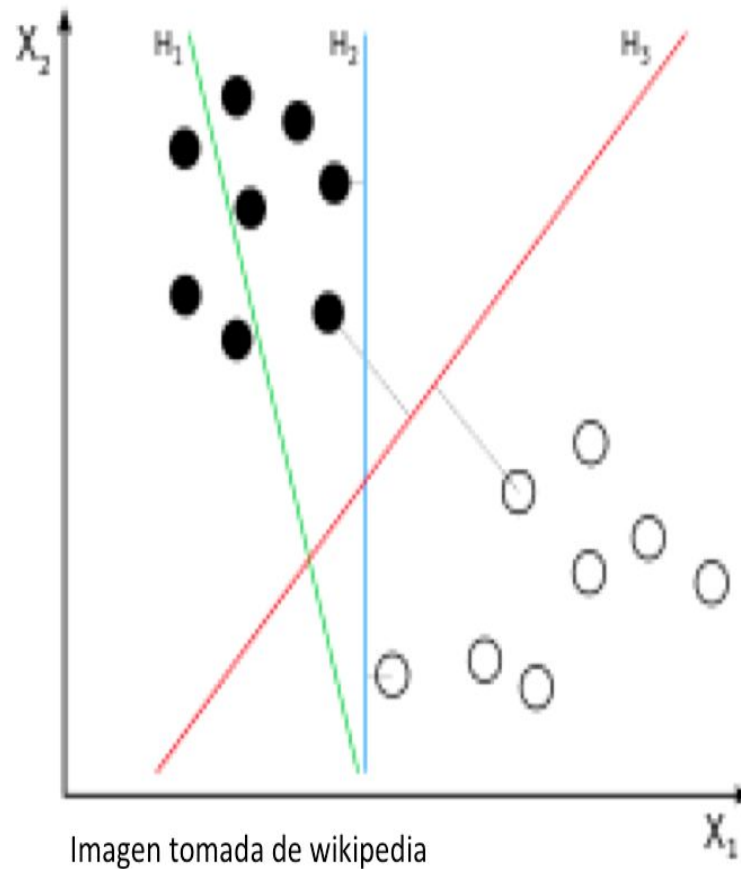
■ Positivo
▲ Negativo



https://es.wikipedia.org/wiki/K_vecinos_m%C3%A1s_pr%C3%B3ximos

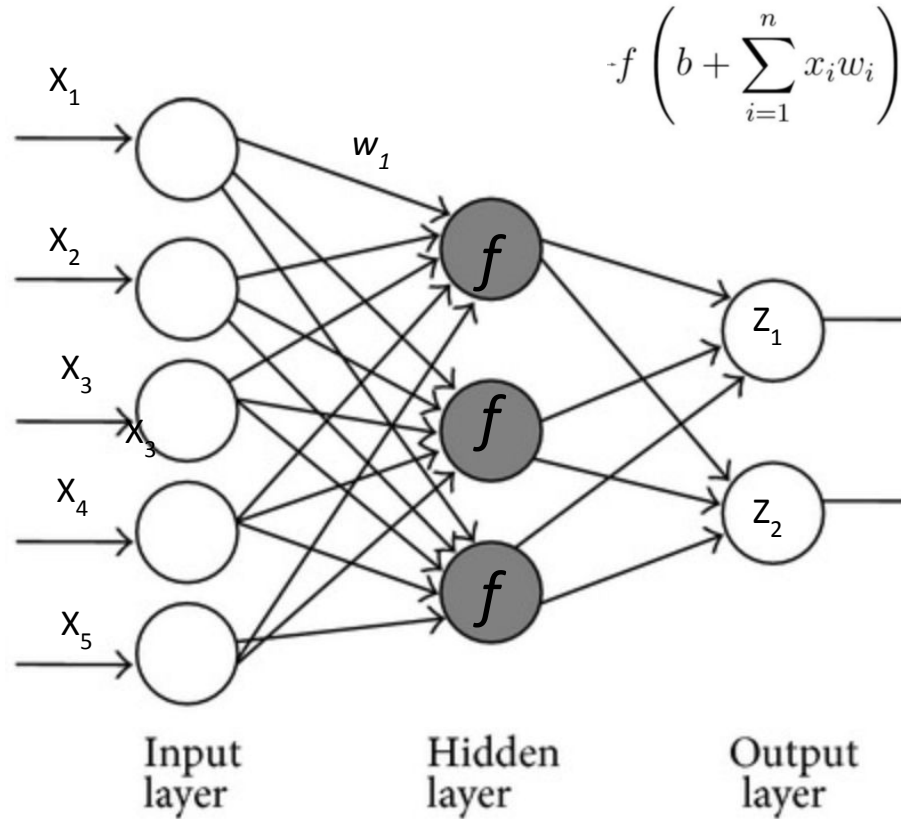
Algoritmos: SVM

- Positivo
- Negativo

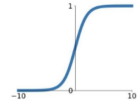


MultiLayer Perceptron (MLP)

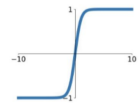
Con
Lorazepam
no
duermo
bien



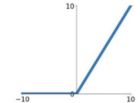
Sigmoid
 $\sigma(x) = \frac{1}{1+e^{-x}}$



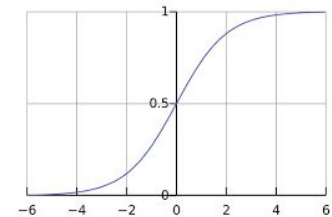
tanh
 $\tanh(x)$



ReLU
 $\max(0, x)$

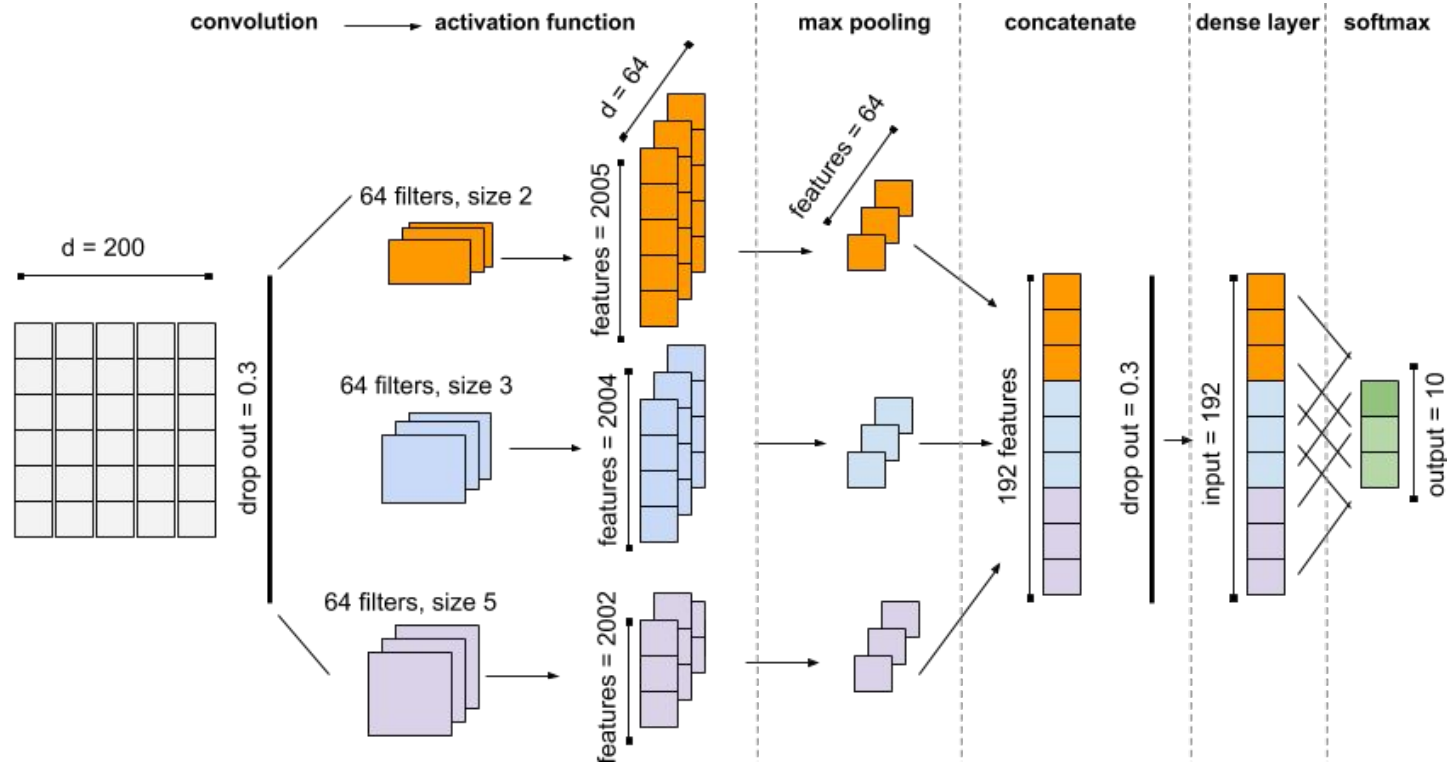


$$\text{Softmax}(z_j) = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}$$

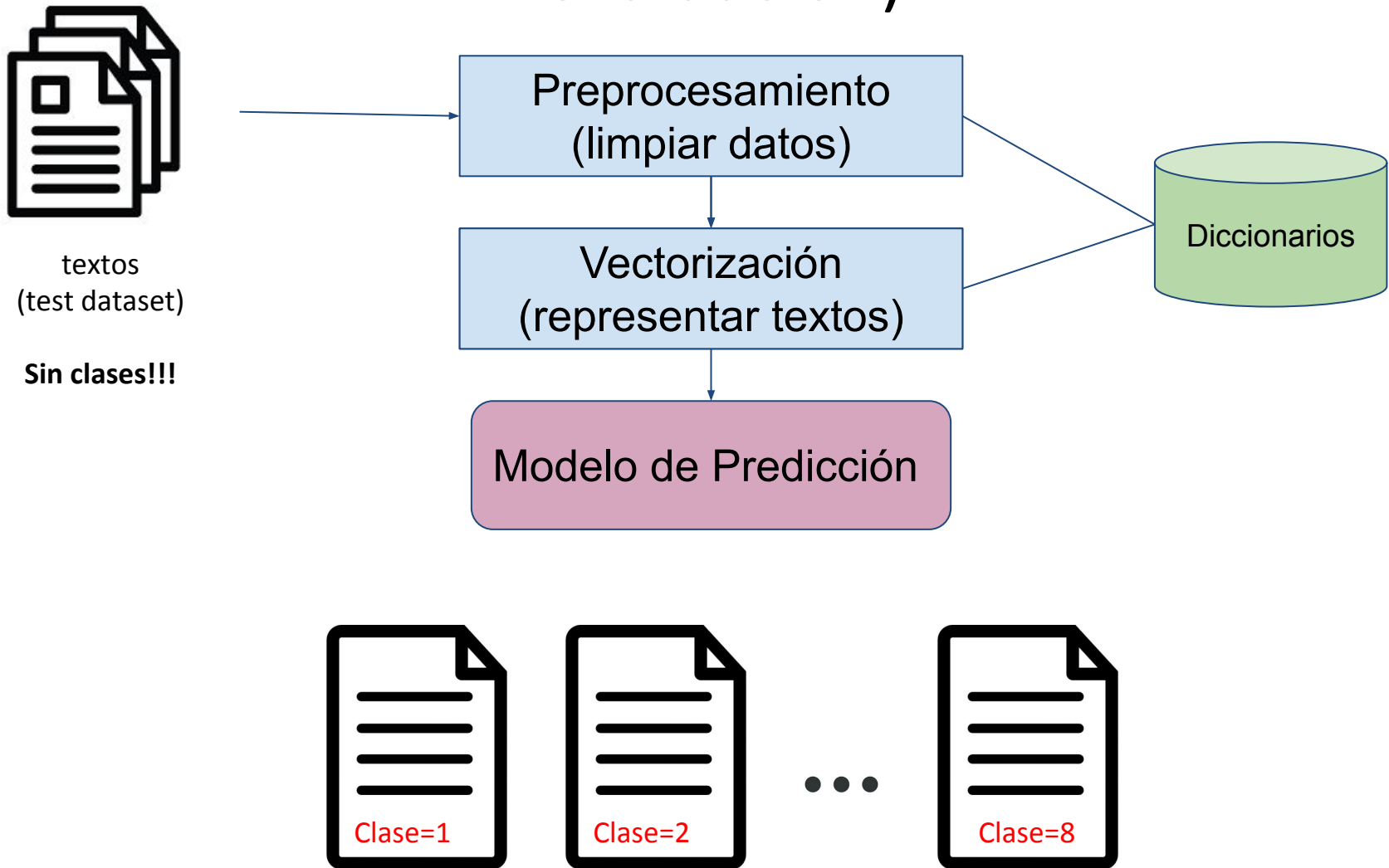


Algoritmos: Deep Learning (CNN)

Con
Lorazepam
ya
no
tengo
ansiedad



Arquitectura Clasificación de textos (fase de evaluación)



Predicción de las clases para cada texto del test dataset

¿Cómo evaluar?

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

<https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c>

¿Cómo evaluar? - Precision

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

<https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c>

¿Cómo evaluar? - Recall

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FN})$$

<https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c>

¿Cómo evaluar? - F1

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

$$\mathbf{F1 = 2 * Precision * Recall / (Precision + Recall)}$$

<https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c>

Algunos consejos prácticos

- Muchísimos datasets para trabajar en Text Classification (y en particular, en Sentiment Analysis!!!).
 - Español: TASS <http://www.sepln.org/workshops/tass/>, ...
 - Inglés: IMDB reviews (inglés), Rotten Tomatoes dataset, Twitter US Airline Sentiment, etc... (<https://data.world/datasets/sentiment>)
- Lenguaje Python (Google Colab).
- Librerías NLP: Spacy, NLTK, TextBlob..
- Librerías Word Embeddings: gensim.
- Librerías Machine learning y Deep Learning: scikit-learn, Keras, PyTorch.
- Librerías Análisis de datos y visualización: pandas, scipy, numpy, MatPlobLib.
- Disfrutar!!!.

Resumen

- La clasificación de textos consiste en asignar una categoría a un texto.
- Los textos son preprocesados y limpiados (tokenización, stopwords, lematización, etc).
- Los textos deben ser representados como vectores de números. Los enfoques más utilizados son bolsas de palabras, TF-IDF, y word embeddings.
- Una vez representados, utilizamos algún algoritmo para entrenar un modelo, que será aplicado sobre el dataset de test.

Resumen

- Enfoque supervisado =
 - training dataset: textos con sus clases. Utilizamos estos textos para entrenar un modelo de un algoritmo de clasificación.
 - test dataset: textos sin clases. El objetivo es asignar una clase a cada documento.
- Algoritmos: SVM, kNN, Deep Learning, etc.
- Evaluación: calculamos las métricas de precisión, recall y F1. Para ello se comparan las clases reales de los documentos del dataset de test y las clases asignadas por el modelo entrenado.