



**uc3m | Universidad Carlos III de Madrid**

# Procesamiento de Lenguaje Natural aplicado a Sentiment Analysis

Isabel Segura Bedmar  
Grupo HuLAT,  
Departamento de Informática  
Leganés, 29 Noviembre 2019



# ¿Por qué Sentiment Analysis?



## ELECCIONES 2019

# Debate electoral: Las mentiras de los candidatos

EL MUNDO  
Madrid

Miércoles, 6 noviembre  
2019 - 10:17



Ver 444 comentarios



Los cinco candidatos que participaron anoche en el debate electoral

leares

Cataluña

Comunidad Valenciana

País Vasco



kalixx

06/11/2019 01:11

#408

bascal es un ser abyecto, un aprovechado que lleva viviendo de la marmandurria  
espera toda su vida, hasta que ha decidido que ya es mayor para montárselo por su  
uenta, bien acompañando de gentuza como él, de los que Espinosa, Monasterio o  
errano son buena muestra. Gracias a vuestra ceguera, ¿cómo se puede ser tan  
nto?, no sólo van a engordar su patrimonio sino que van a provocar un Tsunami,  
ucho más efectivo que el de los cachorros indepes. Yo tengo una abuela catalana,  
spero que pueda solicitar la nacionalidad cuando este estado casposo se rompa,  
o quiero quedar de vuestro lado.

Responder



Denunciar



kalixx

06/11/2019 01:03

#407

'ox, ¿el partido de los machos acomplejados? A juzgar por los comentarios así es.  
ais asco

Responder



Denunciar



pepeillo1

06/11/2019 01:01

#406

@Yabastadenanocabron #301 y si tienes hijos no tengan que sufrirlas, el ministro  
Aguilar el que sacó esa ley, creo que sufrió una denuncia, pero él no durmió en el  
calabozo, cualquier paria lo haría. Luego creo que retiró la denuncia. ¿Por qué será,  
que consiguió, fue falsa, no lo sabemos?

## DATA MINING &amp; MACHINE LEARNING



Booking.com

Alojamiento

Vuelos

Vuelo + Hotel

Alquiler de coches

Reserva ahora

Guarda para acordarte de reservar

Guardado en 899 listas

Igualamos el precio

## Buscar

Destino/Nombre del alojamiento:

Leganés

Fecha de entrada

Fecha de entrada

Fecha de salida

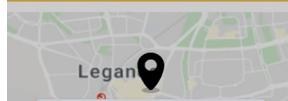
Fecha de salida

2 adultos

Sin niños 1 habitación

 Viajo por trabajo 

Buscar



Ver en el mapa

Google Map data ©2019 Inst. Geogr. Nacional

Cómo llegar a Tryp Madrid Leganes Hotel desde el Aeropuerto Adolfo Suárez Madrid - Barajas

Coche 25 minutos

Taxi 30 minutos

Parking disponible

Info y precios

Servicios

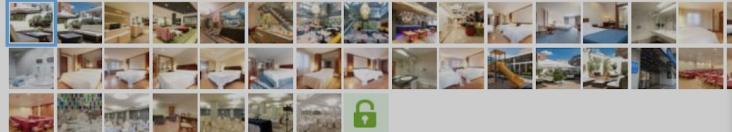
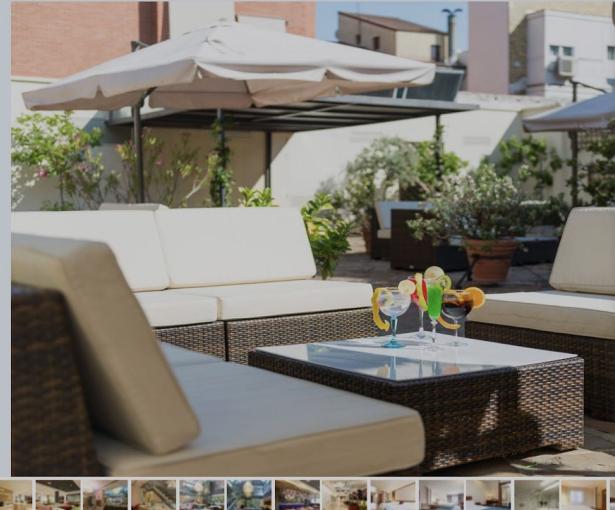
Léeme

A tener en cuenta

Comentarios

## Hotel Tryp Madrid Leganes Hotel

Avenida de la Universidad, 7, 28911 Leganés, España - Ubicación ideal. Mostrar mapa



El Tryp Leganes se encuentra a 12 km del centro de Madrid y cerca de la Universidad Carlos III. Además, cuenta con conexión Wi-Fi gratuita, cafetería y recepción 24 horas.

Las habitaciones del Tryp Madrid Leganes Hotel son luminosas y sencillas y disponen de TV, escritorio y aire acondicionado.

El Tryp sirve por las mañanas un desayuno bufé en el comedor con techo abovedado de cristal. Los huéspedes también podrán disfrutar de bebidas y aperitivos en la cafetería.

El hotel se encuentra a unos 2 km de las autopistas M-45 y M-50. Desde la estación de metro de Leganés, situada a 10 minutos a pie, se tardan 40 minutos en llegar al centro de Madrid. El parque de nieve y centro comercial de Xanadú está a 15 km del hotel.



Aurora  
España

6,7

Comentó en: 20 de noviembre de 2019

## Agradable

· La ubicacion

· La habitación estaba algo sucia y oía a tabaco, el baño muy pequeño.

Se alojó en: Habitación con cama grande  
Noviembre de 2019

Útil Poco útil



Mario  
España

7,5

Comentó en: 17 de noviembre de 2019

## Días de ocio en Madrid

· El desayuno es bueno y el personal atento y amable. La parada de tren a 5 minutos.

· La habitación era la más pequeña que he estado nunca. En las fotos de la descripción del hotel no salía. El Parque tenía las plazas muy justas y podía acceder cualquier y colarse en el hotel por el ascensor de dicho Parque

Se alojó en: Habitación con cama grande  
Noviembre de 2019

Útil Poco útil



Soraya  
España

8,8

Comentó en: 17 de noviembre de 2019

## Fabuloso

· Que no se oía nada

Se alojó en: Habitación con cama grande  
Noviembre de 2019

Útil Poco útil



Leganés ▾

## Restaurante Himalaya Tandoori

211 opiniones | N.º 2 de 213 Restaurantes en Leganés

Calle del Cipres 10 | Leganes, Madrid, 28918 Leganés, España



JunkieGirl  
Madrid, España  
 290 136

Opinión escrita ayer

Bien

La comida está bastante bien pero nos resultó cara para lo que ofrecen. Mesas muy juntas y salón bastante cutre. Esperábamos más.



**Fecha de la visita:** noviembre de 2019

Gracias, JunkieGirl



albertopiwit  
Fuenlabrada, España

Opinión escrita hace 3 días

**¡Por fin un buen sitio de comida india en la zona sur de Madrid!**

Hemos ido con mi mujer varias veces ¡y es una suerte tener que al fin haya un restaurante indio en la zona sur! Estaba harto de tener que ir siempre al centro de Madrid para comer comida de la India. Se lo recomiendo a cualquier... [Más](#)



13 1

**Fecha de la visita:** noviembre de 2019

Gracias, albertopiwit



Raul H  
 1

Opinión escrita hace 4 días

**Rico Pero muy excaso.**

La comida es aceptable, sabores indios muy alternativos a los de una dieta mediterránea, leímos la carta eh imaginamos que por los 12 (11'95) euros que cuestan sería una cantidad aceptable pero es demasiado excaso ( 4-5 cachos por ración) . Conclusión rico pero no... [Más](#)

**Fecha de la visita:** noviembre de 2019

Gracias, Raul H

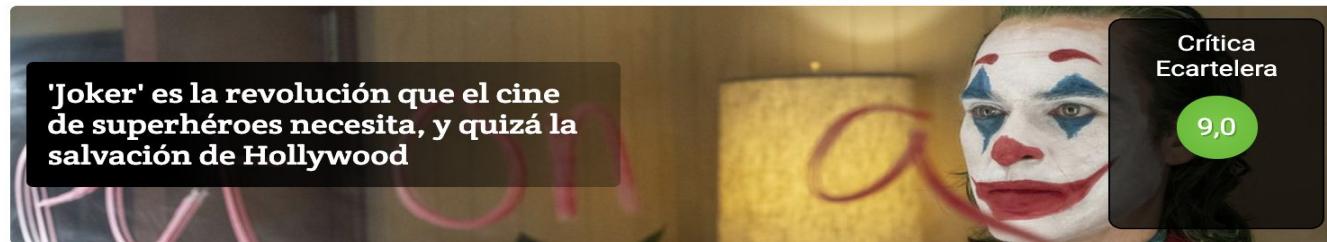




**e-cartelera** ES -

Car PORTADA NOTICIAS CARTELERA PELÍCULAS SERIES CRÍTICAS FOTOS VÍDEOS Foros

## Críticas de 'Joker'



**Openbank** Grupo Santander

**Hipoteca Fija Sin Comisiones**

\* Consulta condiciones en openbank.es/hipoteca-tipo-fijo. Cumpliendo condiciones.\*\* O de compra si este es menor

**1,55% TIN\*** **1,75% TAE\***

Plazo de 15 años Hasta el 50% del valor de tasación\*\*

Calcula tu cuota

[Críticas de los usuarios](#) [Críticas de los medios](#)



German17

5,9

**Regular**

▲0 ▼0 17 nov 2019

Sin spoilers No ha stado mal, pero tampoco es para darla un ocho. Esperemos que las personas se conciencen de lo visto en la peli Y no rechazar a los mas desfavorecidos, si no ayudar en lo que se pueda. Todos tenemos unas cualidades particulares y tenemos que respetar... [Ver mas](#)



Luly083

10

**Fantástica!**

▲0 ▼0 09 nov 2019

Sin spoilers Esta película ha hecho que me replanteé muchas cosas que hago diariamente y cómo pueden afectar a los demás. No da una visión clara de como es la población en algunos aspectos. Ojalá más películas así de buenas por parte de DC.... [Ver mas](#)



amazon.es prime

Todos los departamentos ▾

Enviar a Isabel  
Leganes 28914

Volver a comprar Amazon.es de Isabel Ofertas Prime Now y La Plaza de DIA Chollos Cheques regalo Vender Atención al Cliente Hogar y cocina Informátic

Amazon.es Los más vendidos Chollos Ofertas Productos Reacondicionados Lista de deseos Cheques regalo Amazon Prime Apps de Amazon Vender en Amazon Trabajar en Amazon

Rainbow Socks - Hombre Mujer Calcetines Deporte Colores de Algodón -... > [Opiniones de clientes](#)

## Opiniones de clientes

4,3 de 5 ✓

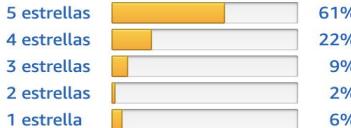


### Rainbow Socks - Hombre Mujer Calcetines Deporte Colores de Algodón - Pares - - Talla UE

por Rainbow Socks

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

620 valoraciones de clientes



[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 comodos, 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](#)

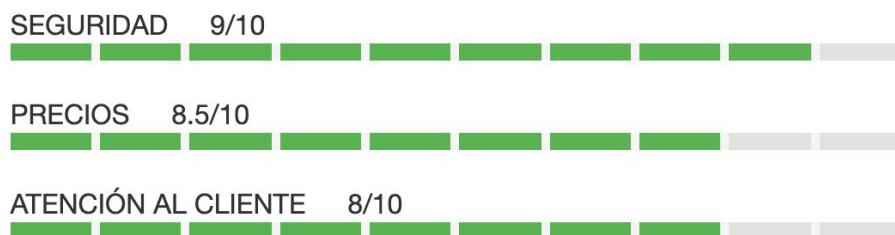
| [Informar de un abuso](#)



## Valoración de los clientes sobre Prosegur



8.5



Más alarmas

Estudio comparativo

### Marina Ceuta Sevilla

Pesimo servicio, nunca consiguieron que las imágenes de mi local llegaran a la app cuando lo solicitaba.No me dieron tranquilidad y cuando cerré el negocio tuve que pagar la permanencia de 2 años., Tuviera o no el negocio. Sin sentido cuando no existe el objeto a proteger.

Me gusta · Responder · 6 s

### Tomas Martinez-zaldivar Rodriguez

Tomas Martínez.  
Llevó sólo y medio con Prosegur. Lla he pagado todo el año por adelantado. He intentado darme de baja y no me lo devuelven.  
El motivo es que la aplicación del móvil no va. La mayoría de las veces no me avisa, por lo que nunca estoy tranquilo. Llevo casi un año detrás de que me lo solucionen y no hay manera. Sólo buenas palabras.  
En mi caso no lo recomiendo.

Me gusta · Responder · 26 s



### Bel Ch

Haz una reclamación y al consumidor o Juzgado.

Me gusta · Responder · 13 s

### Conchi De Ugarte Sanz

Es una auténtica vergüenza el servicio que están ofreciendo a una Anciana de 84 años que tiene el servicio de asistencia. Le retiraron a mi madre el dispositivo que tenía porque no funcionaba el día 23 de abril, sin avisarnos a las personas que nos tienen como responsables de contacto. El domingo 28 nos llaman a las hijas para avisarnos que no reciben comunicación del dispositivo desde hace 48h. Mi madre no nos había avisado y cuando intentamos averiguar porqué no habían dado un dispositivo de sustitución fue imposible. Desde el lunes 29 llevamos llamando, presentando incidencias, reclamaciones... Ver más

Me gusta · Responder · 29 s



### Comparaiso.es

Lamentamos lo ocurrido, Conchi. Comparaiso es un comparador de alarmas independiente, pero quedamos a tu disposición en el teléfono gratuito 91 076 95 99 para intentar solucionar cualquier incidencia de seguridad. Adicionalmente, en el siguiente enlace tienes más información sobre otros sistemas de teleasistencia: <https://www.comparaiso.es/alarmas/teleasistencia>  
Un saludo.

Me gusta · Responder · 28 s

### Jorge Ibañez

Pesimo servicio. Es inutil llamar. Siempre sale un disco y cuando por fin contesta un operador te pasa a otro que no te resuelve nada. Desaconsejo contratar con esta compañía



**Doctoralia**

p. ej. ginecólogo p. ej. Madrid

Iniciar sesión ¿Es usted un doctor?

Página De Inicio / Psicólogo / Madrid / Ana Lucas Prieto

**Ana Lucas Prieto** ✓  
Psicólogo, Psicólogo infantil  
Núm. Colegiado: M-21661  
★★★★★ 41 opiniones

[Reservar cita](#) [Enviar mensaje](#)

Consultas Precios **Opiniones(41)** Dudas solucionadas Experiencia

### Opiniones de pacientes

5 ★★★★★ Valoración global 41 opiniones

★★★★★ Puntualidad  
★★★★★ Atención  
★★★★★ Instalaciones

Todas las opiniones (41) Positivas (41) Neutras (0) Negativas (0)

A.V Paciente verificado • 7/11/19 [Mostrar más detalles](#)

Cita reservada en Doctoralia  
Visita psicología

Muy buena profesional.Trato excelente y totalmente recomendable para solucionar los problemas.Un acierto conocerla.

Reportar

I.A. Paciente verificado • 6/11/19 [Mostrar más detalles](#)

Cita reservada en Doctoralia  
Visita psicología

Perfecta la conexión con ella. Trabajo muy efectivo y con excelentes resultados. La evolución es grande.

Reportar

### Reservar cita

La reserva de cita es un servicio gratuito de Doctoralia. Fácil, rápido y seguro.

1 Dirección Mostrar en el mapa  
c/ Hilarion Eslava 55 planta 9 28050 Ma... Psico-salud

2 Hora de la cita

Hoy	Mañana	Sáb	Dom	Lun
21 Nov 09:50	09:50	-	-	11:30
-	10:40	-	-	14:00
-	11:30	-	-	14:50
-	12:20	-	-	16:50
-	13:10	-	-	19:20

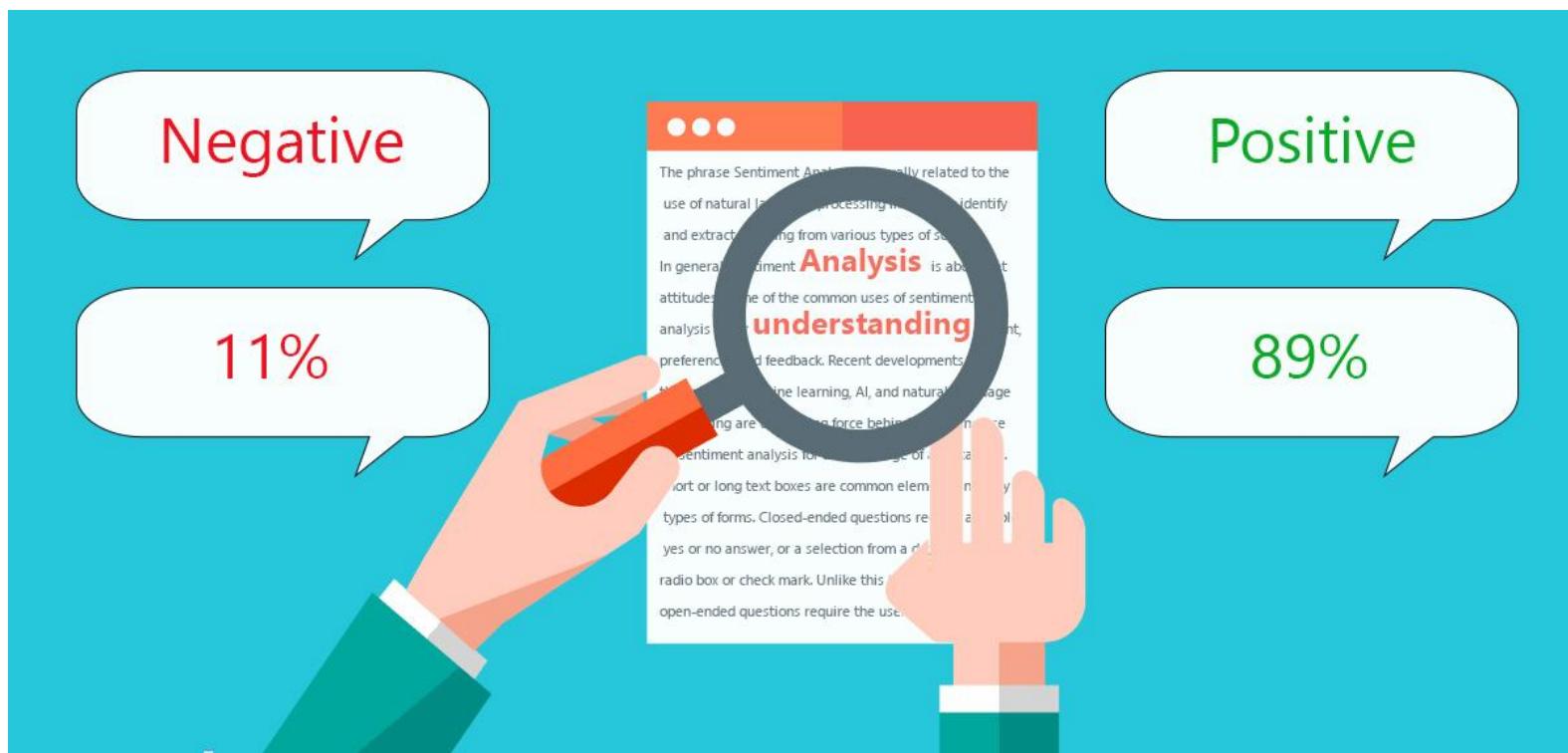
[Mostrar más horas](#)



# ¿Qué es Sentiment Analysis?



- Clasificación **automática** de la polaridad (positivo o negativo) de un documento.



<https://formtitan.com/Optimization-Tips/Online-form-with-built-in-Sentiment-Analysis>



## Críticas Booking:

-  – *Es el mejor hotel que he visitado nunca...*
-  – *Las habitaciones muy limpias. Buena relación calidad-precio.*
-  – *Espectacular , una verdadera maravilla!!*
-  – *No vale para nada. Todo muy sucio. El personal muy antipático.*
-  – *El hotel me ha parecido malísimo... Con la comida nos hemos llevado una decepción. Lo único bueno, el personal.*

## Lorazepam Rating Summary

**8.0 /10**

AVERAGE RATING

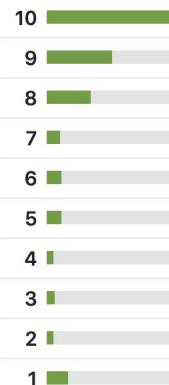
697 Ratings with 567 User Reviews

What next? Compare all 160 medications used  
in the treatment of [Anxiety](#).

[Share your Experience](#)

[Ask a Question](#)

### User Ratings



RyGuy

May 30, 2009

**Ativan (lorazepam):** "I've had no obvious side effects except dry mouth. I had severe panic attacks (12-15 a day) before taking Ativan. I couldn't work and thought I was dying. Since I started it I have had only one full blown panic attack in 5 years."

5.0

What this helpful? Yes No

133 · Report

Anonymous

April 26, 2008

**Ativan (lorazepam):** "a great med helps me alot"

10

What this helpful? Yes No

48 · Report

AnxiousPerson

March 11, 2008

**Ativan (lorazepam):** "Be careful when using ativan. Do not take doses more than 1 or 2 mg. It is easy to get carried away and addicted to Ativan. I find that it is a very effective treatment for anxiety when the RECOMMENDED DOSES are used though though."

7.0

What this helpful? Yes No

225 · Report

Jjkm

August 28, 2015

**Ativan (lorazepam):** "Wonderful takes stress away"

10

What this helpful? Yes No

33 · Report



# Sentiment Analysis (tipos)

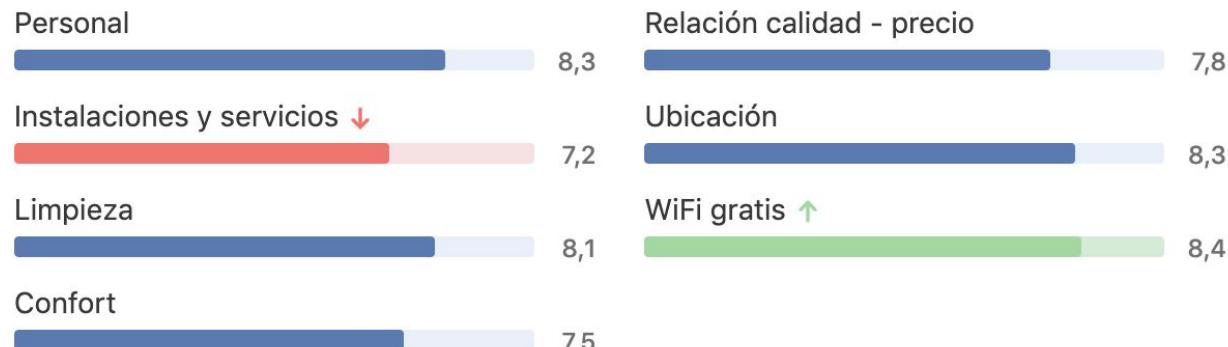
1. **Basada en documentos:** determinar el sentimiento expresado en una opinión, y representarlo
  - a. en dos o tres clases (positiva, negativa, neutra).
  - b. o mediante una puntuación (rating 1-10)
2. **Basada en aspectos:** determinar el sentimiento de una opinión sobre una característica específica.



# Sentiment Analysis (aspectos)

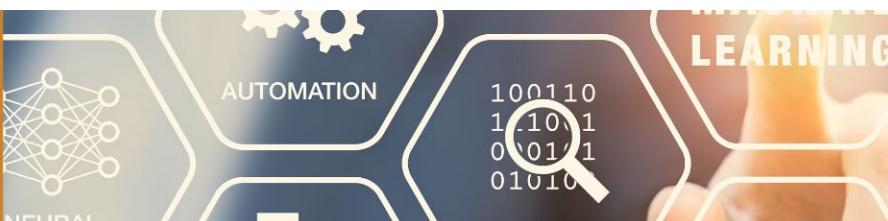
**Booking.com**

- 😊 · El desayuno es bueno y el personal atento y amable. La parada de tren a 5 minutos.
- 😢 · La habitación era la más pequeña que he estado nunca. En las fotos de la descripción del hotel no salía. El Parquing tenía las plazas muy justas y podía acceder cualquier y colarse en el hotel por el ascensor de dicho Parquing



↑ Puntuación alta para Leganés

↓ Puntuación baja para Leganés



# ¿Cómo?



- Enfoques basados Reglas + Diccionarios
- Enfoques basados en Aprendizaje Automático (Machine Learning)



# Enfoque reglas + diccionarios

- Usar diccionarios (listas) de palabras positivas y negativas (**sentiment lexicons**).



# Lexicones de Sentimiento en Español

- **ISOL** (<http://timm.ujaen.es/recursos/isol/>) lista con 2.509 palabras positivas y 4793 negativas.
- **ML-SentiCON Sentiment Spanish Lexicon**: inglés, español, catalán, vasco y gallego. <http://www.lsi.us.es/~fermin/ML-SentiCon.zip>.  
Para cada palabra, el recurso proporciona una estimación de la polaridad (de muy negativo -1,0 a muy positivo 1,0).
- **Sentiment Lexicon in Spanish**. Dos polaridades (positiva, negativa).  
<http://web.eecs.umich.edu/~mihalcea/downloads/SpanishSentimentLexicons.tar.gz>



# The Subjectivity Lexicon

- 8221 palabras anotadas con su polaridad (2718 positivas, 4912 negativas y 428 neutras).
- Idioma: Inglés
- Ejemplos:
  - good, adj, positivo
  - very, adv, neutral.
  - pain, nombre, negativo.
- [https://mpqa.cs.pitt.edu/lexicons/subj\\_lexicon/](https://mpqa.cs.pitt.edu/lexicons/subj_lexicon/)

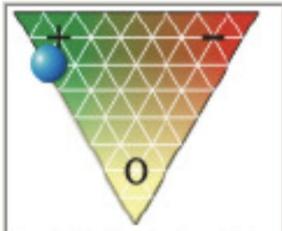
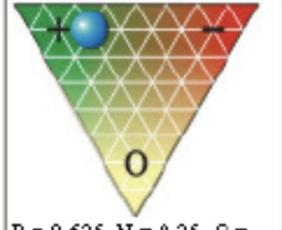
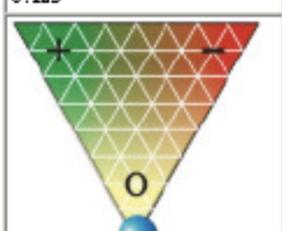


# SentiWordNet

show position

## Adjective

3 senses found.

 $P = 0.75, N = 0, O = 0.25$	<u>estimable(1)</u> <i>deserving of respect or high regard</i>
 $P = 0.625, N = 0.25, O = 0.125$	<u>honorable(5) good(4) respectable(2) estimable(2)</u> <i>deserving of esteem and respect; "all respectable companies give guarantees"; "ruined the family's good name"</i>
 $P = 0, N = 0, O = 1$	<u>computable(1) estimable(3)</u> <i>may be computed or estimated; "a calculable risk"; "computable odds"; "estimable assets"</i>



# Lexicones de Sentimiento en Inglés

- **General Inquirer:** <http://www.wjh.harvard.edu/~inquirer/homecat.htm>.

Contiene 1915 palabras positivas y 2291 negativas. También contiene otras clasificaciones más complejas: fuerte, débil, activo, pasivo, placer, dolor, virtud, vicio.

- **Bing Liu Opinion Lexicon**

(<http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar>): 2006 positivas y 4783 negativas.

- **81 diccionarios en distintos idiomas:**

<https://www.kaggle.com/rtatman/sentiment-lexicons-for-81-languages/d ata#>



# Enfoque reglas + diccionarios

- Usar diccionarios (listas) de palabras positivas y negativas (**sentiment lexicons**).
- Contar el número de palabras positivas (#numPos) y el número de palabras negativas (#numNeg).



# Enfoque reglas + diccionarios

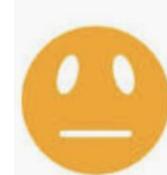
Si  $\#numPos > \#numNeg$ :

Asignar polaridad positiva



Si  $\#numPos = \#numNeg$ :

Asignar polaridad neutra



Si  $\#numPos < \#numNeg$ :

Asignar polaridad negativa



## Lorazepam Rating Summary

**8.0** /10

AVERAGE RATING

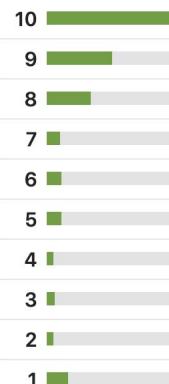
697 Ratings with 567 User Reviews

What next? Compare all 160 medications used  
in the treatment of [Anxiety](#).

[Share your Experience](#)

[Ask a Question](#)

### User Ratings



RyGuy

May 30, 2009

**Ativan (lorazepam):** "I've had no obvious side effects except dry mouth. I had severe panic attacks (12-15 a day) before taking Ativan. I couldn't work and thought I was dying. Since I started it I have had only one full blown panic attack in 5 years."

5.0

What this helpful? Yes No

133 · Report

Anonymous

April 26, 2008

**Ativan (lorazepam):** "a great med helps me alot"

10

What this helpful? Yes No

48 · Report

AnxiousPerson

March 11, 2008

**Ativan (lorazepam):** "Be careful when using ativan. Do not take doses more than 1 or 2 mg. It is easy to get carried away and addicted to Ativan. I find that it is a very effective treatment for anxiety when the RECOMMENDED DOSES are used though though."

7.0

What this helpful? Yes No

225 · Report

Jjkm

August 28, 2015

**Ativan (lorazepam):** "Wonderful takes stress away"

10

What this helpful? Yes No

33 · Report

Negativa: 1, 2, 3, 4  
Neutra: 5, 6,  
Positiva: 7, 8, 9, 10



Drug: Azithromycin

**Rating: 10.0**

Condition: bacterial infections

*"Very good response. It is so useful for me. "*

#numPos = 2 > #numNeg = 0





Drug: Macrobid

**Rating: 1.0**

Condition: Urinary Tract Infection

"Awful medicine, the worst. Headache the first night, leg and back pain. Pain got worse and worse."

#numPos = 0 < #numNeg = 7





Drug: Nardil

**Rating: 10.0**

Condition: Depression

*"This medication is brilliant. Completely got rid of my depression and social phobia. Was laying on the bed all day thinking of suicide and death. After I started Nardil my life completely changed. "*

#numPos = 1 < #numNeg = 4, ¿¿¿???

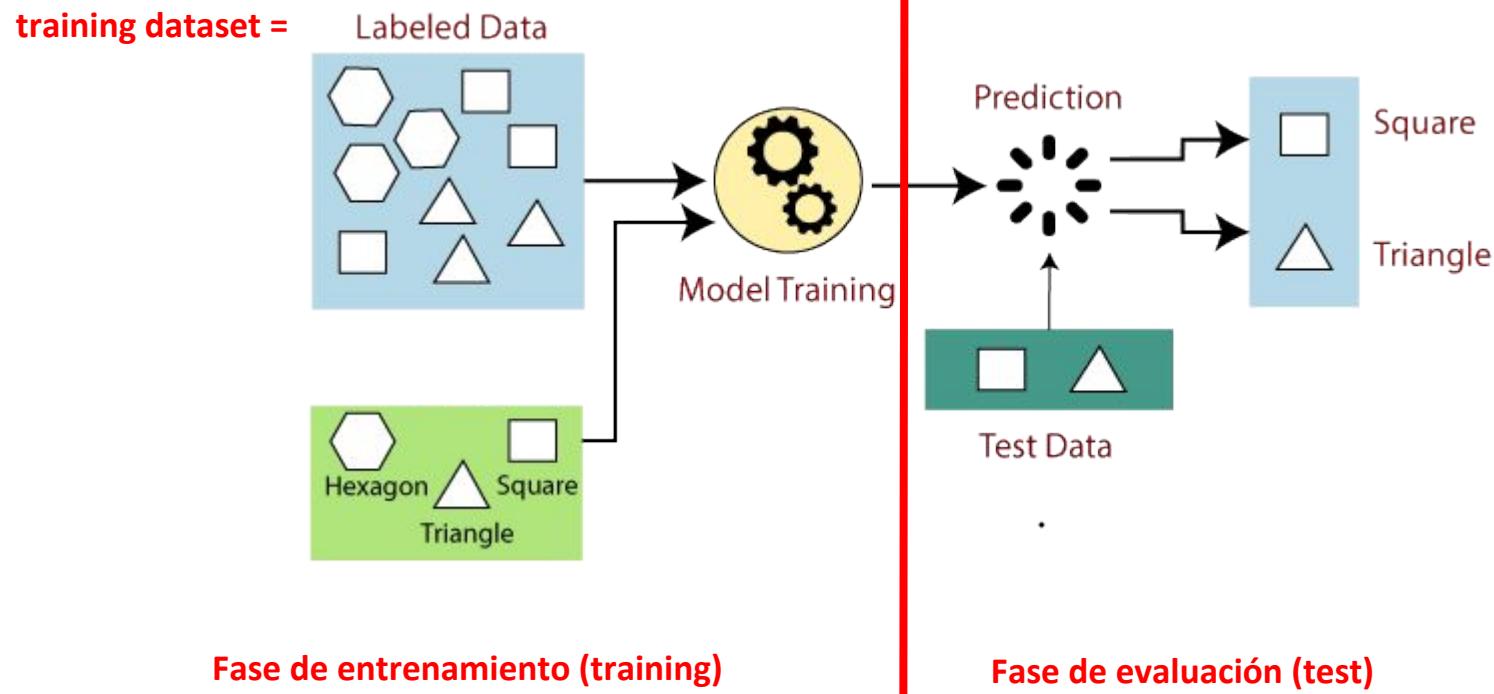




- Enfoques basados Reglas + Diccionarios
- **Enfoques basados en Aprendizaje Automático (Machine Learning)**

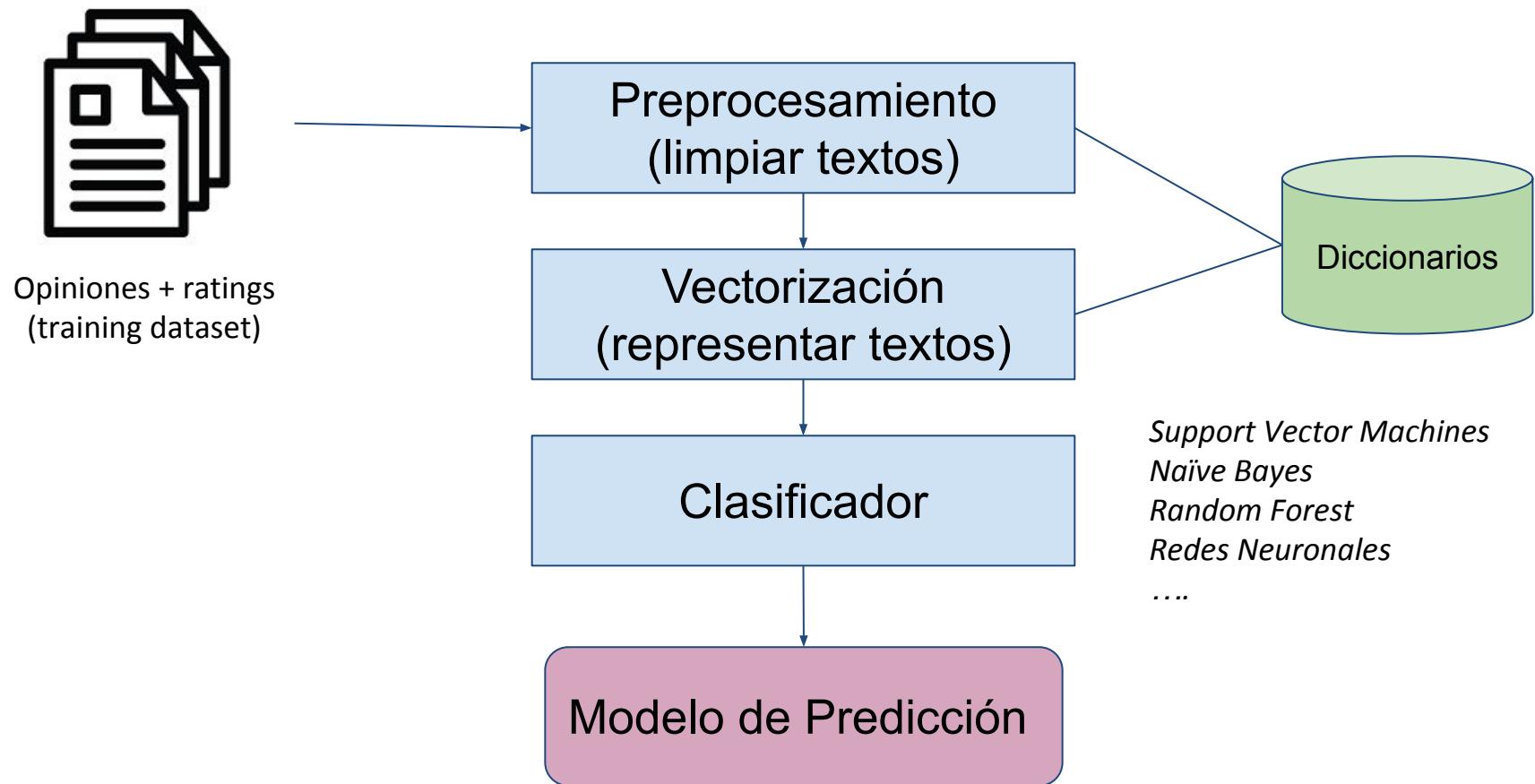


# Aprendizaje automático supervisado





# Arquitectura (fase de entrenamiento)



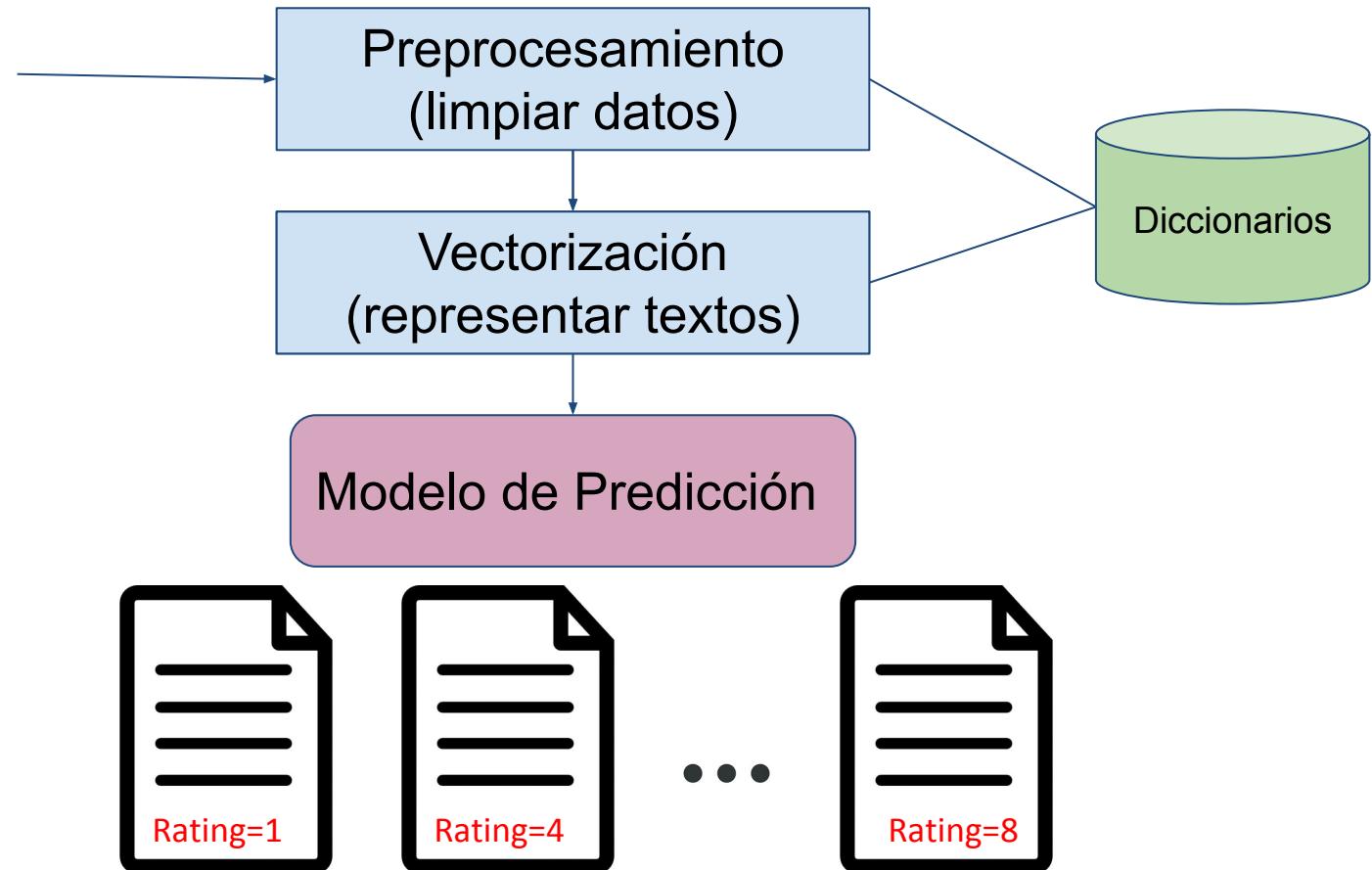


# Arquitectura (fase de evaluación)



Opiniones  
(test dataset)

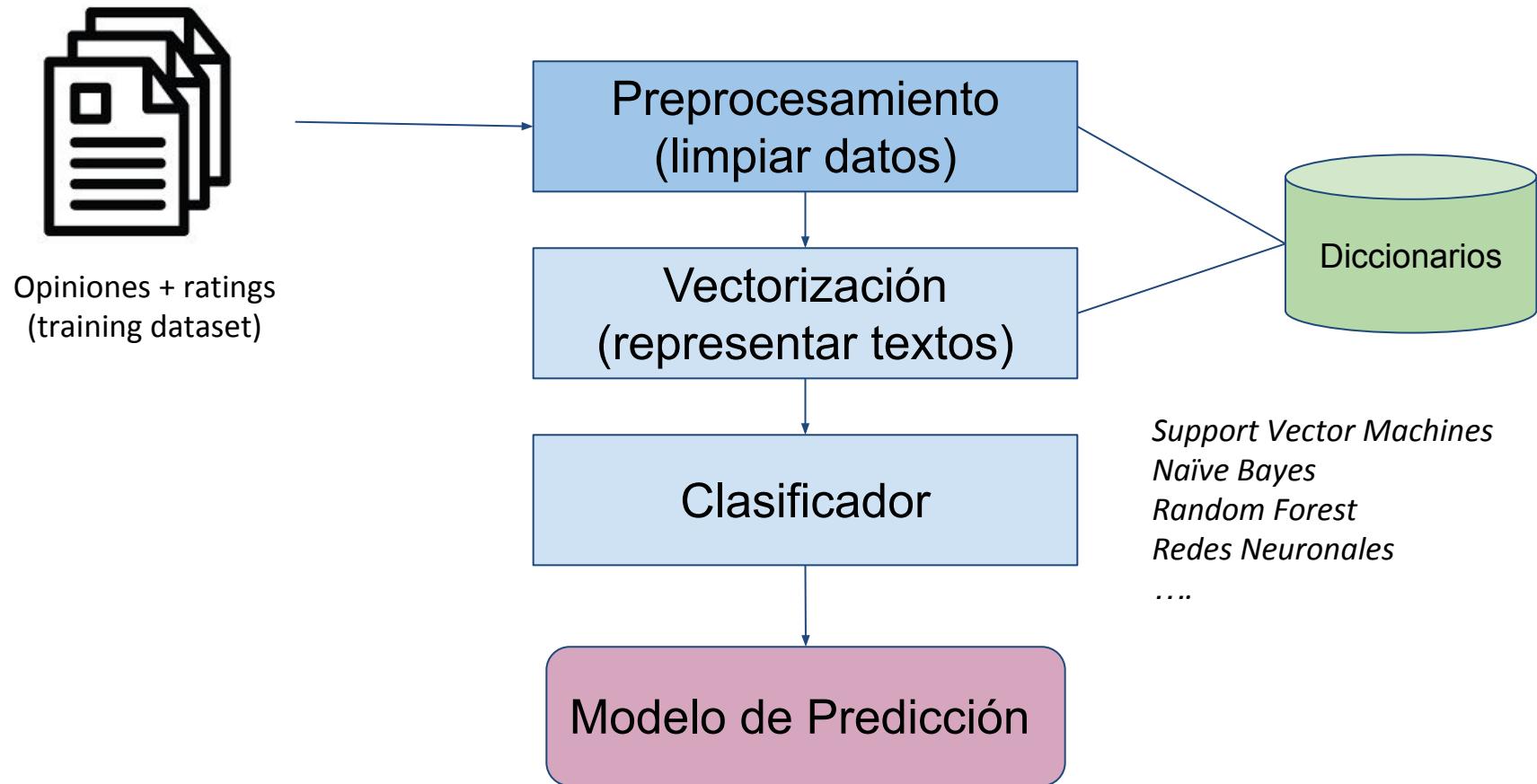
Sin ratings!!!



*Predictión de la puntuación para cada opinión (texto) del test dataset*



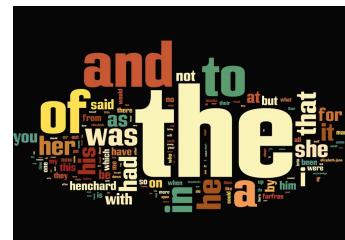
# Arquitectura (fase de entrenamiento)





# 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-*

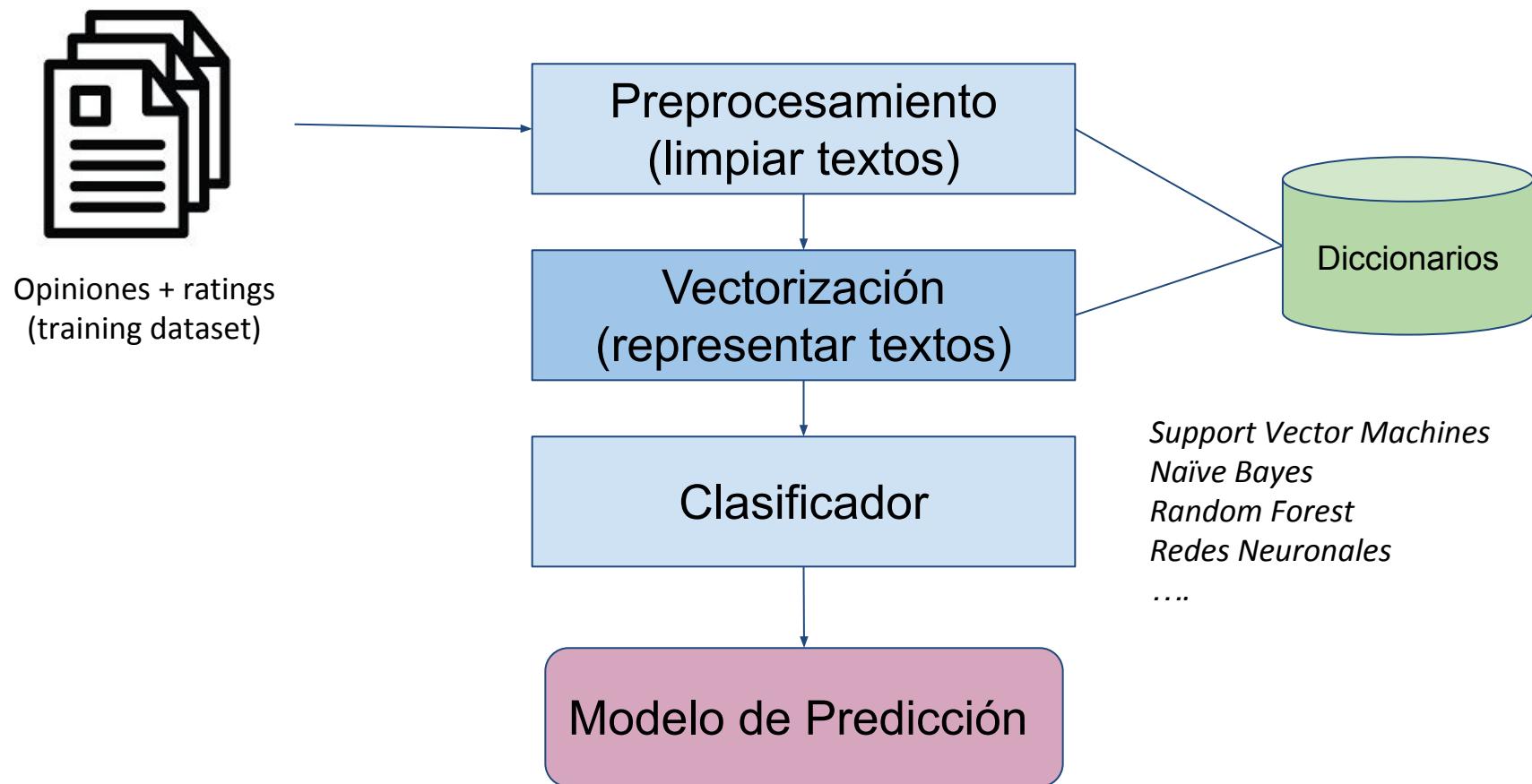


# Recursos

- Lematizador online:  
<http://www.gedlc.ulpgc.es/investigacion/scogeme02/lematiza.htm>
- Stemmer online: [\*https://snowballstem.org/demo.html\*](https://snowballstem.org/demo.html)



# Arquitectura (fase de entrenamiento)





# Vectorización

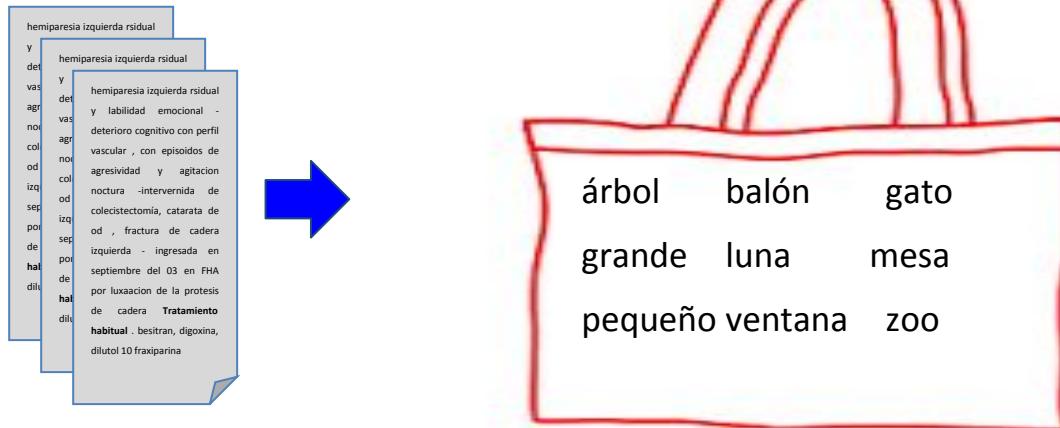
- Cada texto es representado como un conjunto de características (ej, número de palabras positivas, número de palabras negativas).
- Transformar textos en vectores de números.
- Los enfoques más utilizados son:
  - **Bolsa de Palabras.**
  - **TF-IDF**
  - **Word Embeddings**



# 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.

$$\text{tf}_{x,y} \times \log \left( \frac{N}{\text{df}_x} \right)$$

..

## TF-IDF

Term  $x$  within document  $y$

$\text{tf}_{x,y}$  = frequency of  $x$  in  $y$

$\text{df}_x$  = number of documents containing  $x$

$N$  = total number of documents



# 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”*

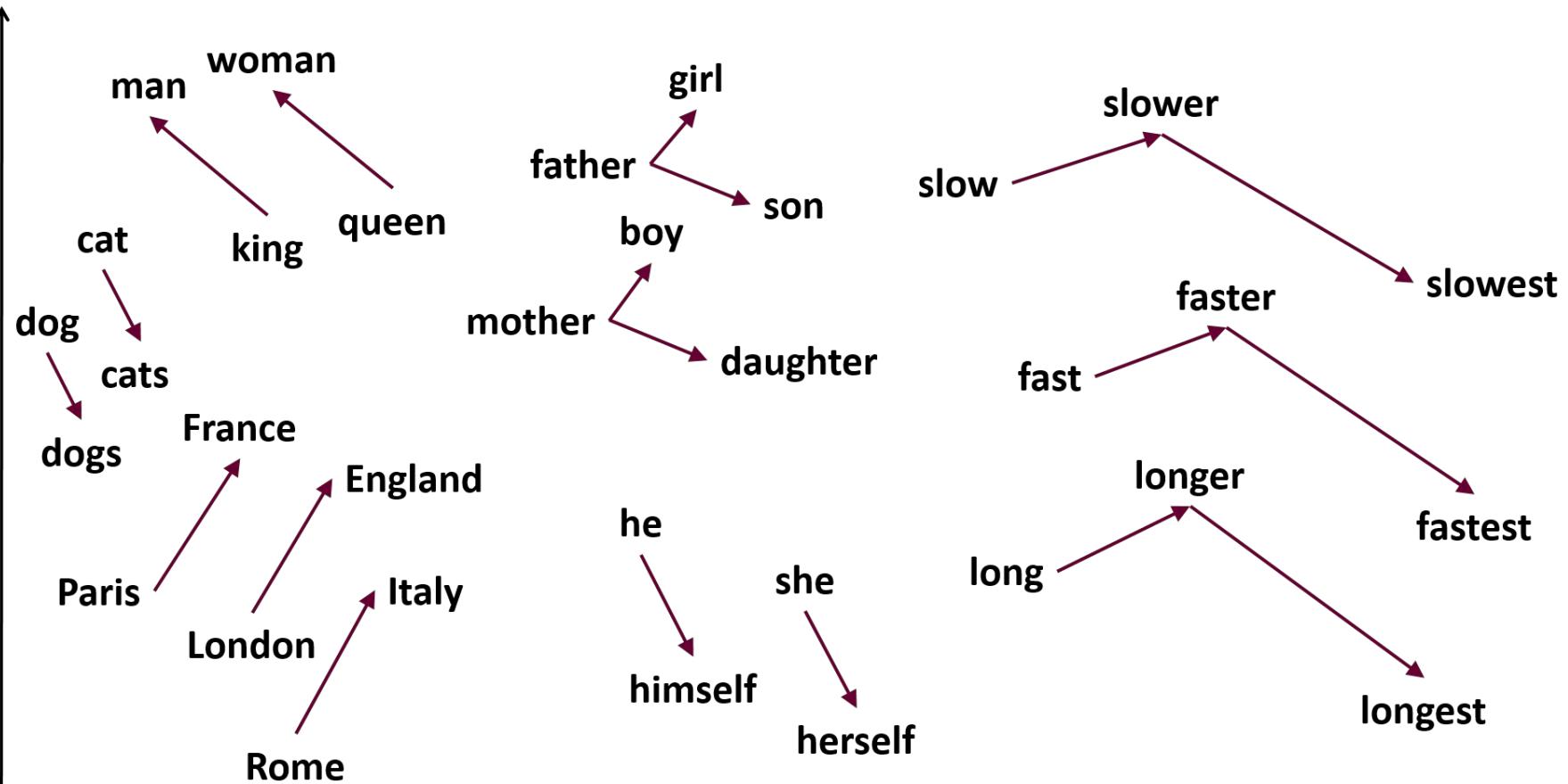


# Vectorización: Word Embeddings

- 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/](http://bionlp-www.utu.fi/wv_demo/)

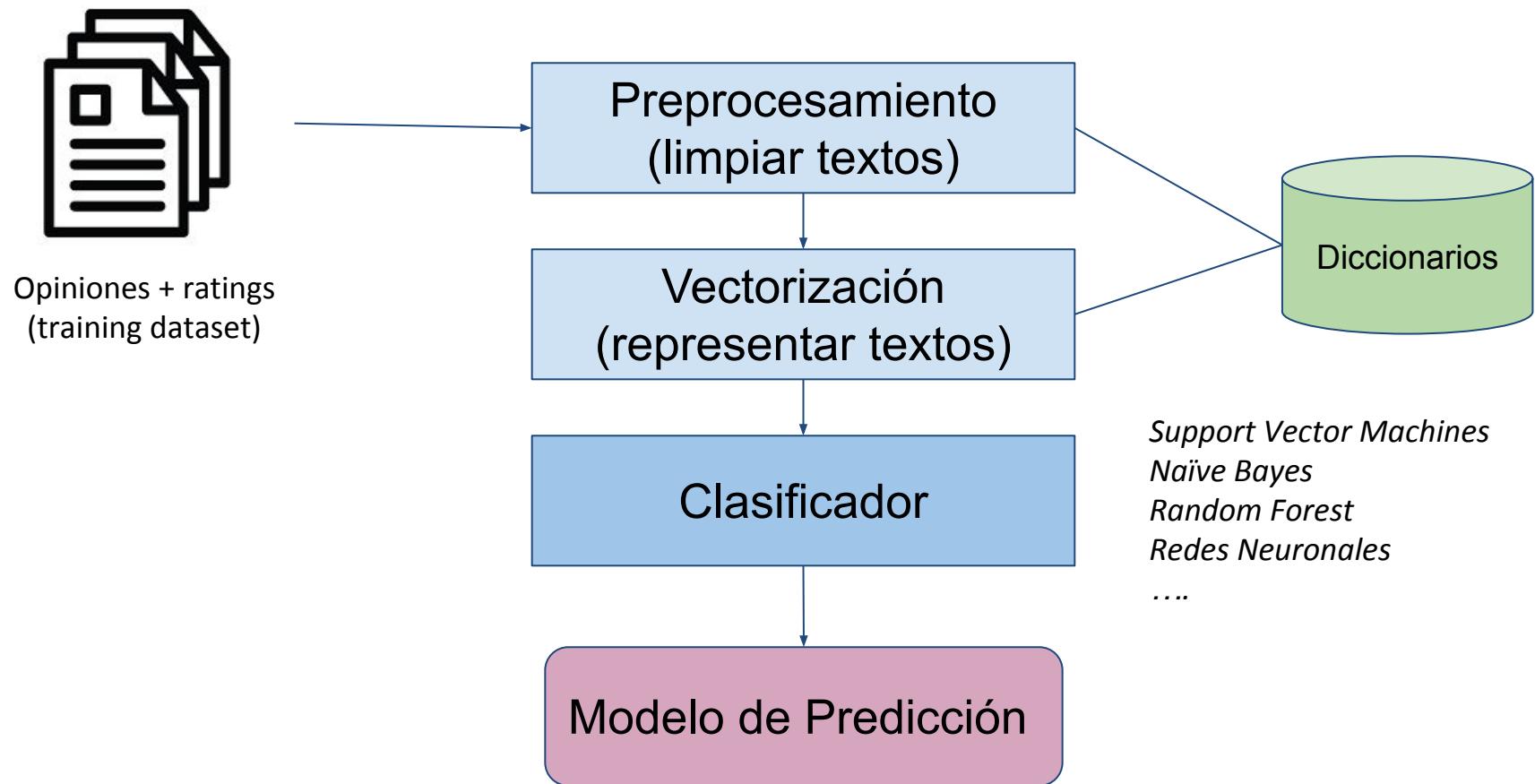


# Word Embeddings





# Arquitectura (fase de entrenamiento)





# Algoritmos de Clasificación

- Support Vector Machines.
- Naïve Bayes.
- k-Nearest Neighbors (kNN).
- MultiLayer Perceptron (MLP).
- Deep Learning models.
- ...



# Algoritmos: Naïve Bayes

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

C={Positivo, Negativo}

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

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

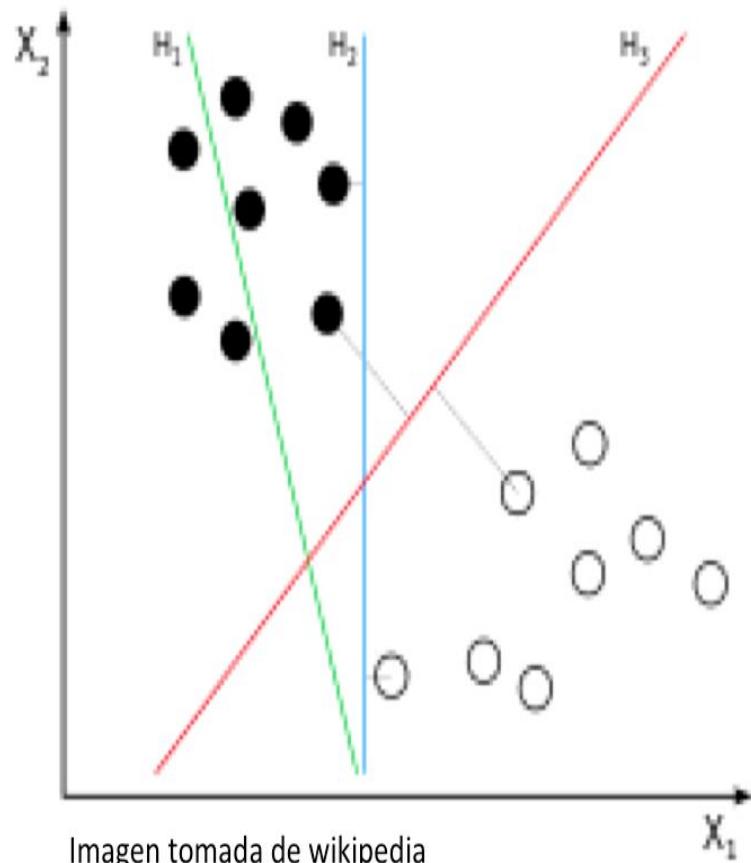
$$= \operatorname{argmax}_{c \in C} P(x_1, x_2, \dots, x_n | c)P(c)$$

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



- Positivo
- Negativo

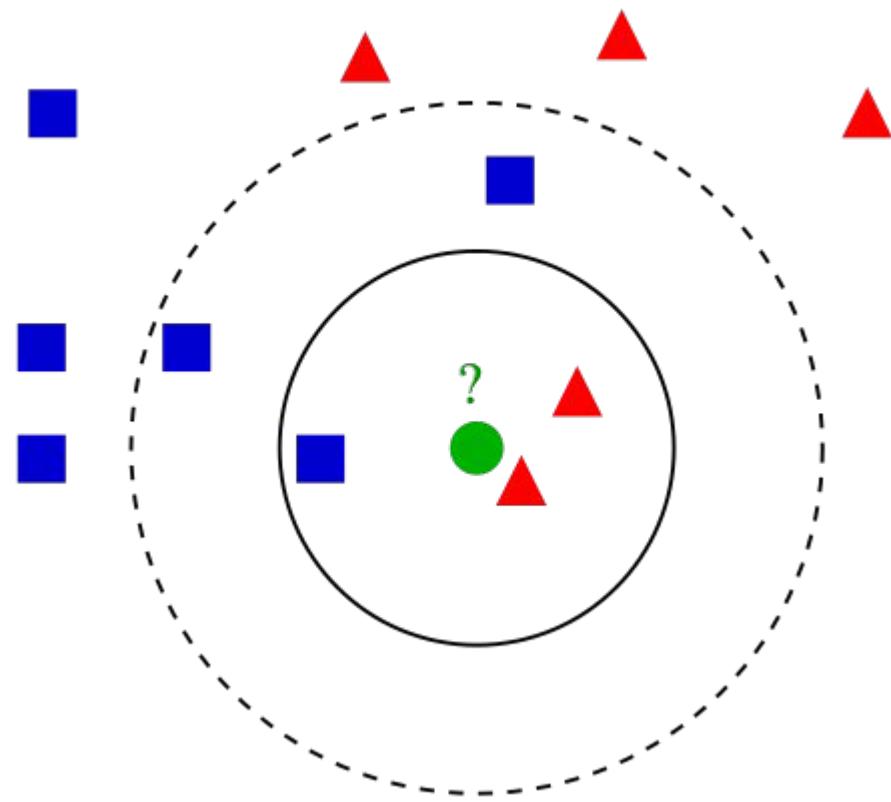
# Algoritmos: SVM





# Algoritmos: k-Nearest Neighbor

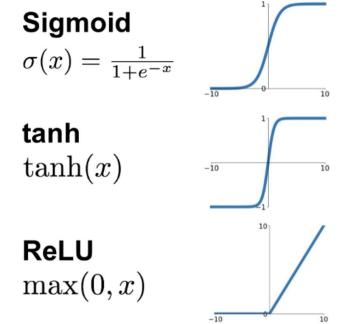
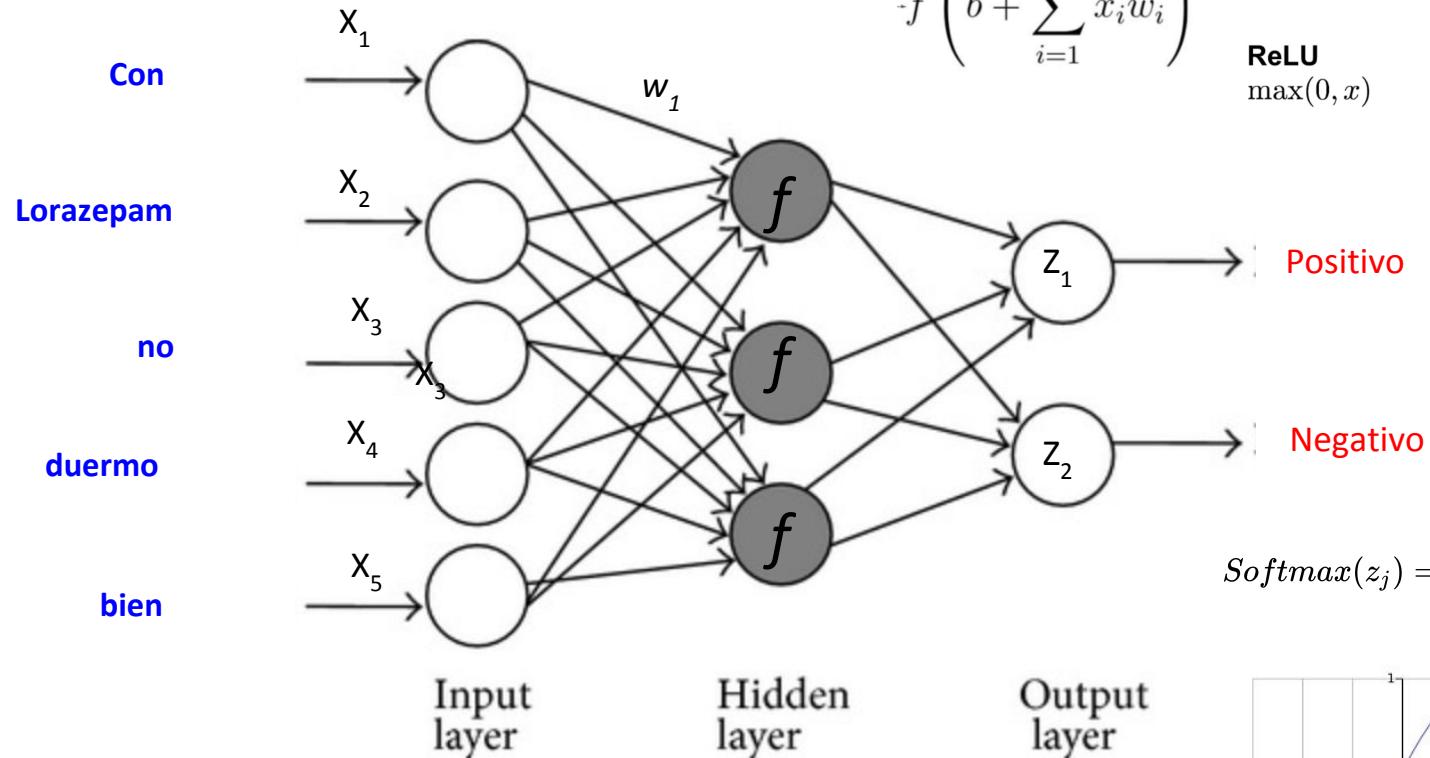
- Positivo
- ▲ Negativo



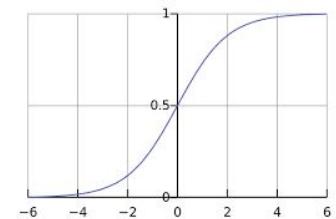
[https://es.wikipedia.org/wiki/K\\_vecinos\\_m%C3%A1s\\_pr%C3%B3ximos](https://es.wikipedia.org/wiki/K_vecinos_m%C3%A1s_pr%C3%B3ximos)



# MultiLayer Perceptron (MLP)



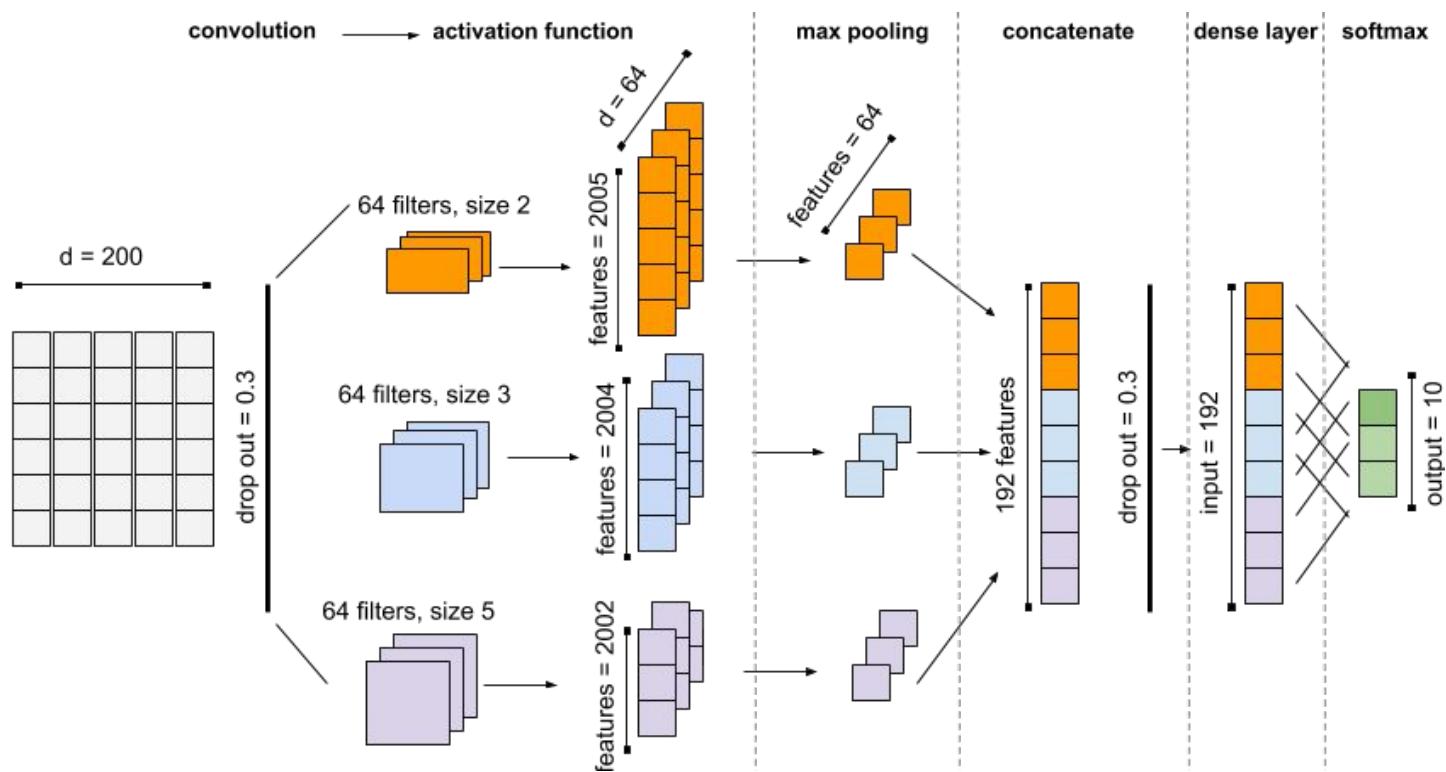
$$\text{Softmax}(z_j) = \frac{\sum e^{z_j}}{\sum_{k=1}^K e^{z_k}}$$





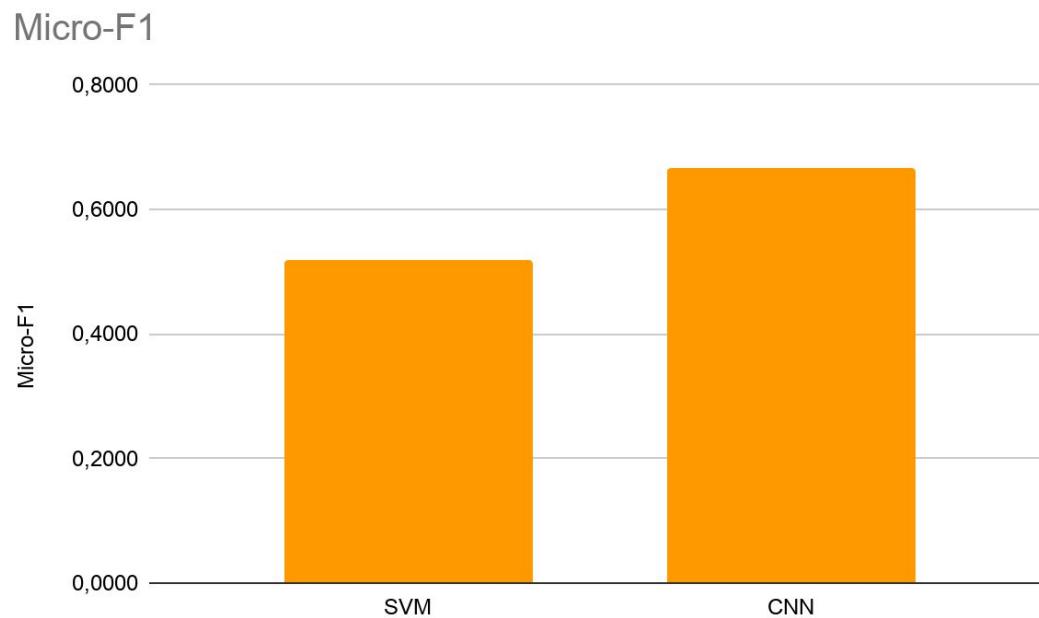
# Algoritmos: Deep Learning (CNN)

Con  
Lorazepam  
ya  
no  
tengo  
ansiedad





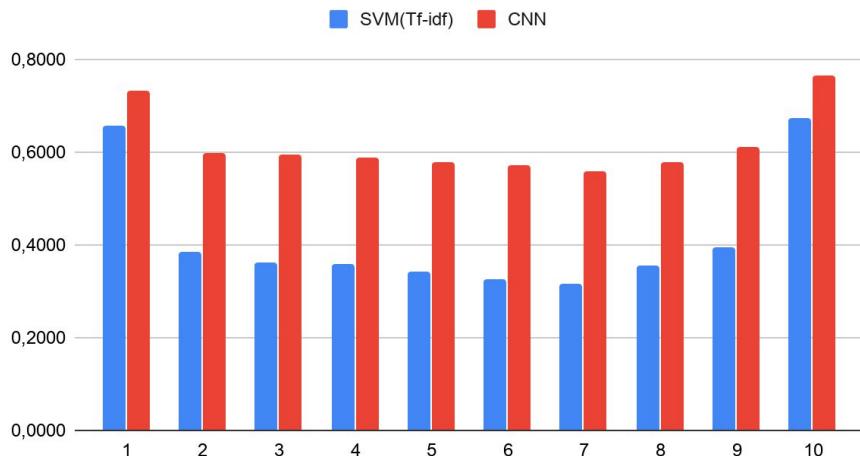
# Resultados - Drug Reviews



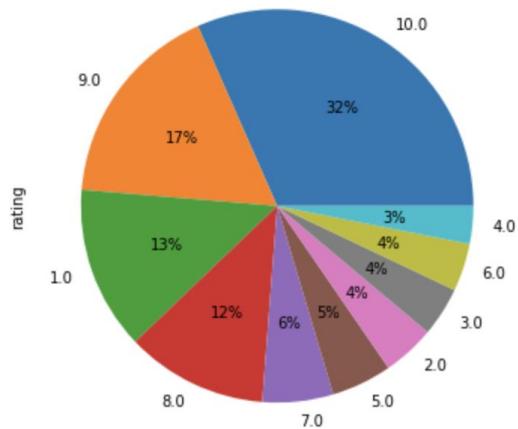


# Resultados - Drug Reviews

Comparativa SVM vs CNN



Distribución de drug reviews por grado de satisfacción





# Algunos consejos prácticos

- Muchísimos datasets para trabajar 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>)
- Diccionarios semánticos: ISOL, SentiWordNet
- Lenguaje Python (Google Colab).
- Librerías NLP: Spacy, NLTK.
- Librerías Word Embeddings: gensim.
- Librerías Análisis de datos y visualización: pandas, scipy, numpy, Matplotlib.
- Librerías Machine learning: scikit-learn.
- Librerías Deep Learning: keras, PyTorch.
- Disfrutar!!!.



# Recursos Útiles

## Text Classification Evaluation

<https://www.youtube.com/watch?v=TdkWIxGoiak>

## Recursos:

<https://www.w3.org/community/sentiment/wiki/Datasets>

<https://github.com/isegura/BasicNLP>

<https://github.com/isegura/SentimentAnalysis>



# Resumen

- Sentiment Analysis es una tarea de clasificación de textos.
- Lexicones de sentimientos recursos importantes.
- 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.