PROCESAMIENTO DEL LENGUAJE NATURAL EN NOTICIAS DE PERIODICOS ESPAÑOLES

VERÓNICA RUIZ MÉNDEZ

vruiz@afi.es
https://www.linkedin.com/in/veronica-ruiz-mendez/

Grupos de Usuarios de R de Madrid
28 de Julio de 2021

Descripción del problema

Descripción del problema

WEB SCRAPPING NOTICIAS PERIODICOS

Indicadores temporales para informar de la evolución de éstos a lo largo del tiempo gracias a la información extraída de las noticias scrapeadas.

Extraer el **sentimiento de las noticias** scrappeadas haciendo uso de algoritmos de Machine Learning.

Estado del arte

Estado del arte

API The New York Times

headline	date	doc_type	material_type
Couture Creations for Dancing Bodies	2020-01-01 10:00:21	article	News
100 Years Ago, the Booziest January Suddenly Dried Up	2020-01-01 10:00:22	article	News
Elizabeth Warren Isnâ€™t Talking Much About â€˜Medicare for Allâ€™ Anymore	2020-01-01 10:00:22	article	News
Living In ... Bedminster, N.J.	2020-01-01 10:00:22	multimedia	Slideshow
In a Homecoming Video Meant to Unite Campus, Almost Everyone Was White	2020-01-01 10:00:25	article	News
Bedminster, N.J.: Horses, Golf and Presidential Visits	2020-01-01 10:01:24	article	News
China Moves to Steady Its Slowing Economic Growth	2020-01-01 10:12:44	article	News
Pete Buttigiegâ€™s Campaign Says It Raised $24.7 Million in the Fourth Quarter	2020-01-01 10:53:14	article	News

Análisis de Sentimientos en Noticias

Dashboard noticias – cotización IBEX

Análisis exploratorio

Análisis Exploratorio

Extracción de datos: Web Scrapping

PERIÓDICO ECONÓMICO

- Años 2012 – 2020
- 169.899 noticias
- Fecha, Título, Subtítulo

PERIÓDICO NACIONAL

- Años 2012 - 2020
- 753.752 noticias
- Fecha, Categoría, Autor, Título, Subtítulo

Análisis Exploratorio

Preparación de datos

```
df$texto <- paste0(df$titulo, " ", df$subtitulo)
```

Limpieza de textos

Nuevos emails falsos suplantan a Netflix e intentar robar nuestros datos.

-> emails falsos suplantan netflix intentan robar datos

Convertir a minúsculas y eliminar tildes

```
df$texto <- chartr("ÁÉÍÓÚ", "AEIOU", tolower(df$texto))
df$texto <- chartr("áéíóú", "aeiou", tolower(df$texto))
```

Eliminar signos de puntuación

```
df$texto <- removePunctuation(df$texto)
```

Análisis Exploratorio

Preparación de datos

Tokenización

Nuevos emails falsos suplantan a Netflix e intentar robar nuestros datos.

->[‘Nuevos’, ‘emails’, ‘falsos’, ‘suplantan’, ‘a’, ‘Netflix’, ‘e’, ‘intentan’, ‘robar’, ‘nuestros’, ‘datos’, ‘.’]

```
df_tokens <- df %>%
  unnest_tokens(word, texto)
```

fecha	titulo	subtitulo	texto	word
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	arrancan
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	las
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	rebajas
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	en
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	madrid
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	con
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	agresivos
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	descuentos
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	para
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	sortear
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	la
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	crisis
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	las
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	rebajas
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	de
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	invierno
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	arrancan

Análisis Exploratorio

Preparación de datos

Tokenización - stopwords

Nuevos emails falsos suplantan a Netflix e intentar robar nuestros datos.

->[‘Nuevos’, ‘emails’, ‘falsos’, ‘suplantan’, ‘Netflix’, ‘intentan’, ‘robar’, ‘datos’, ‘.’]

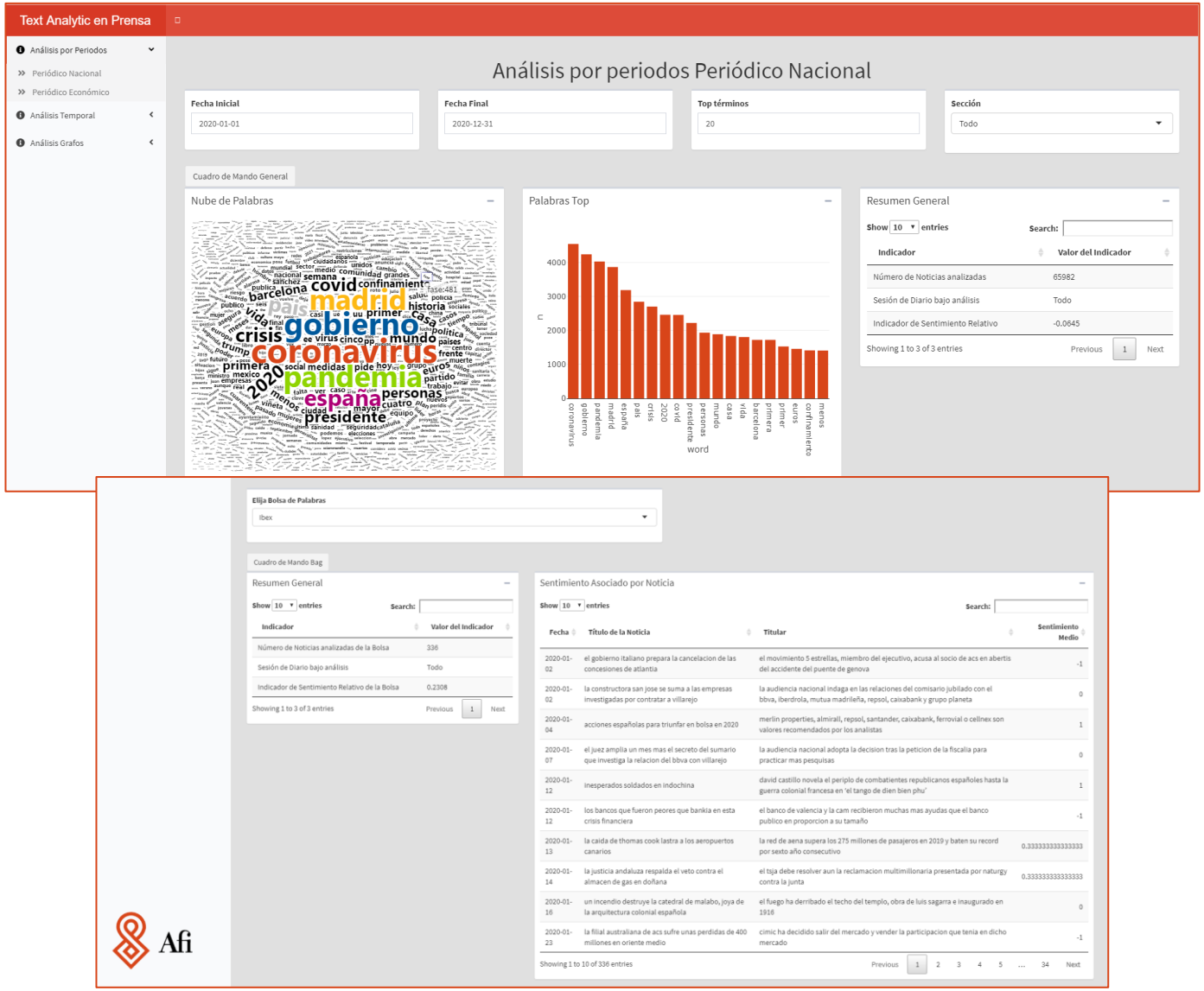
```
df_tokens <- df %>%
  unnest_tokens(word, texto)%>%
  anti_join(stopwords, by = 'word')
```

fecha	titulo	subtitulo	texto	word
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	arrancan
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	rebajas
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	madrid
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	agresivos
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	descuentos
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	sortear
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	crisis
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	rebajas
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	invierno
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	arrancan
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	madrid
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	cambio
2012-01-01	Arrancan las rebajas en Madrid con agresivos descuentos p...	Las rebajas de invierno arrancan en Madrid con el cambio d...	arrancan las rebajas en madrid con agresivos descuentos pa...	importantes

Análisis de indicadores temporales

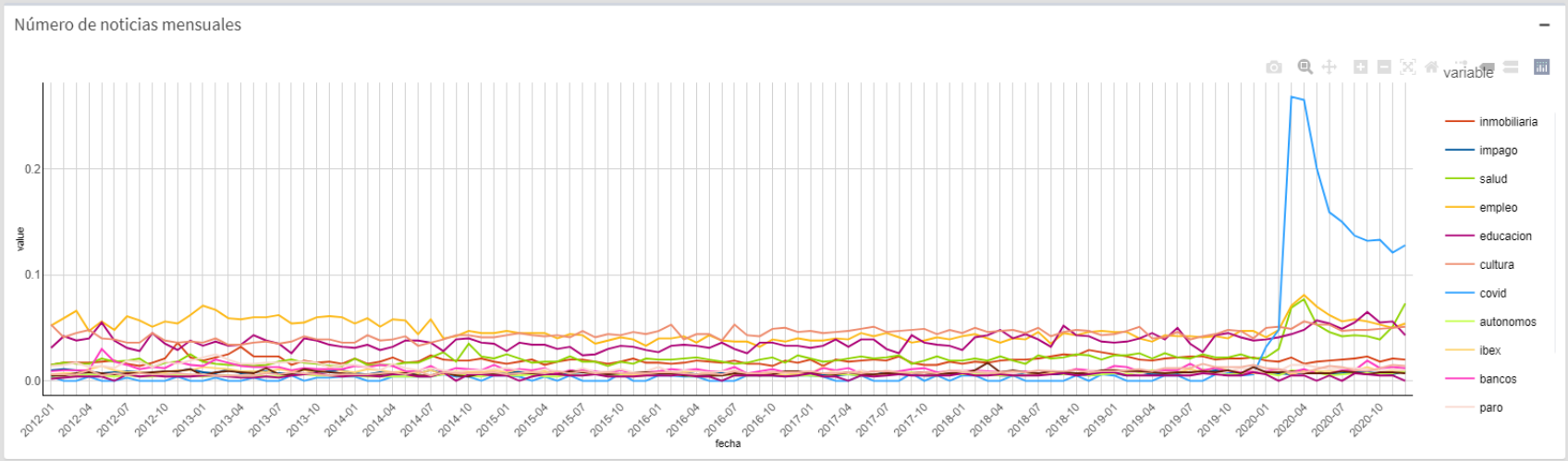
Análisis de indicadores temporales

Análisis por periodos



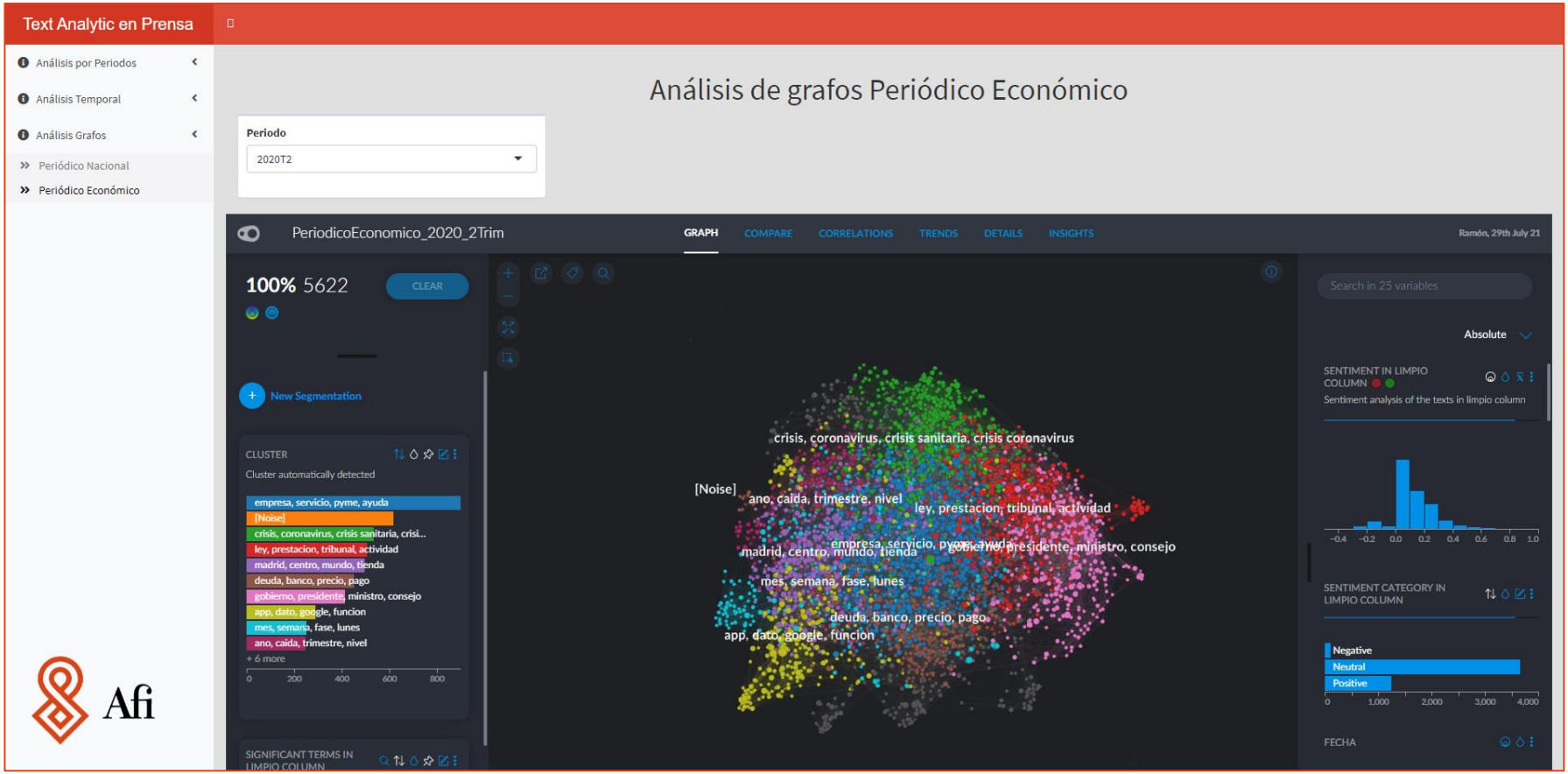
Análisis de indicadores temporales

Análisis temporal



Análisis de indicadores temporales

Análisis de grafos

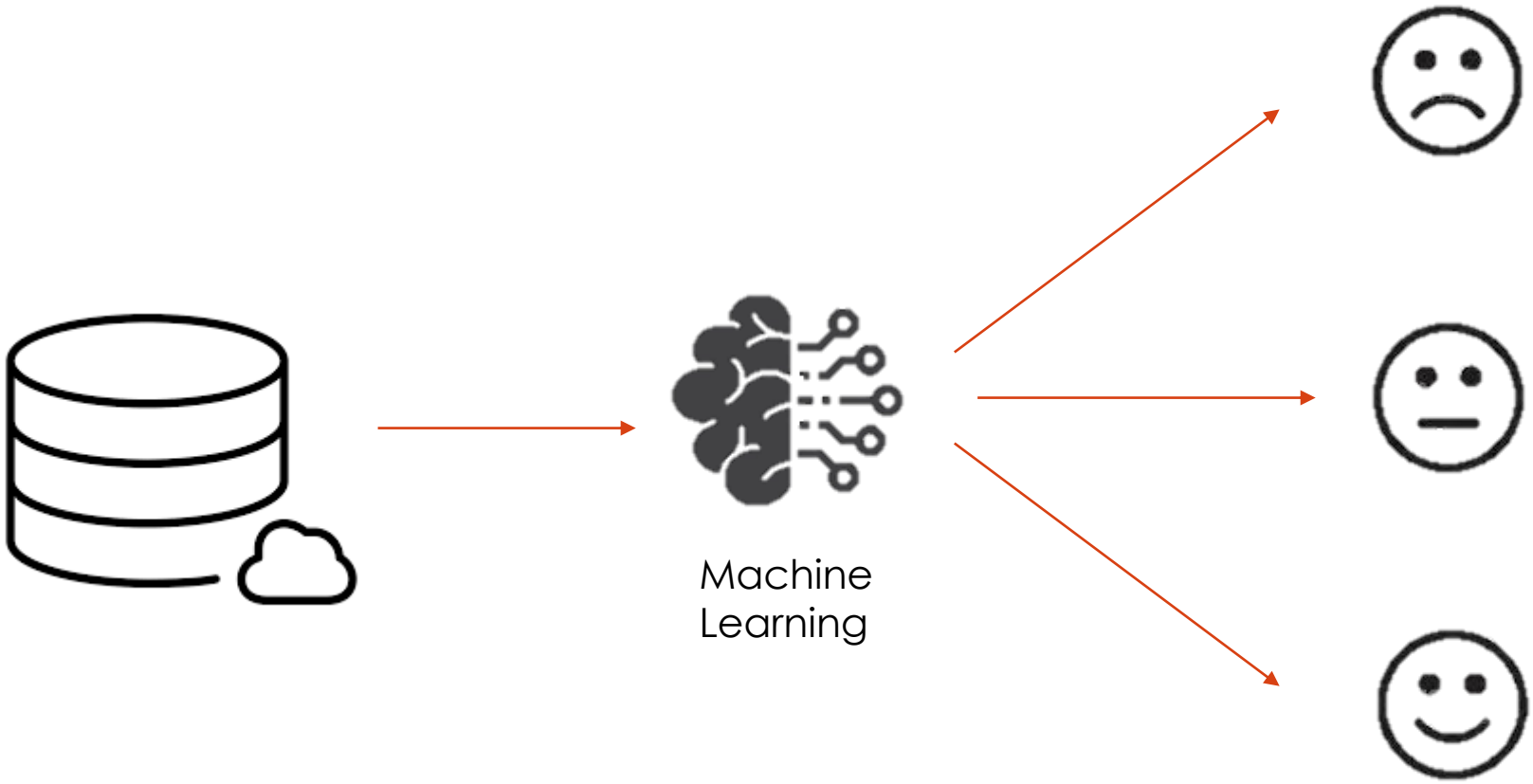


Modelos de Machine Learning

Modelos de Machine Learning

Objetivo

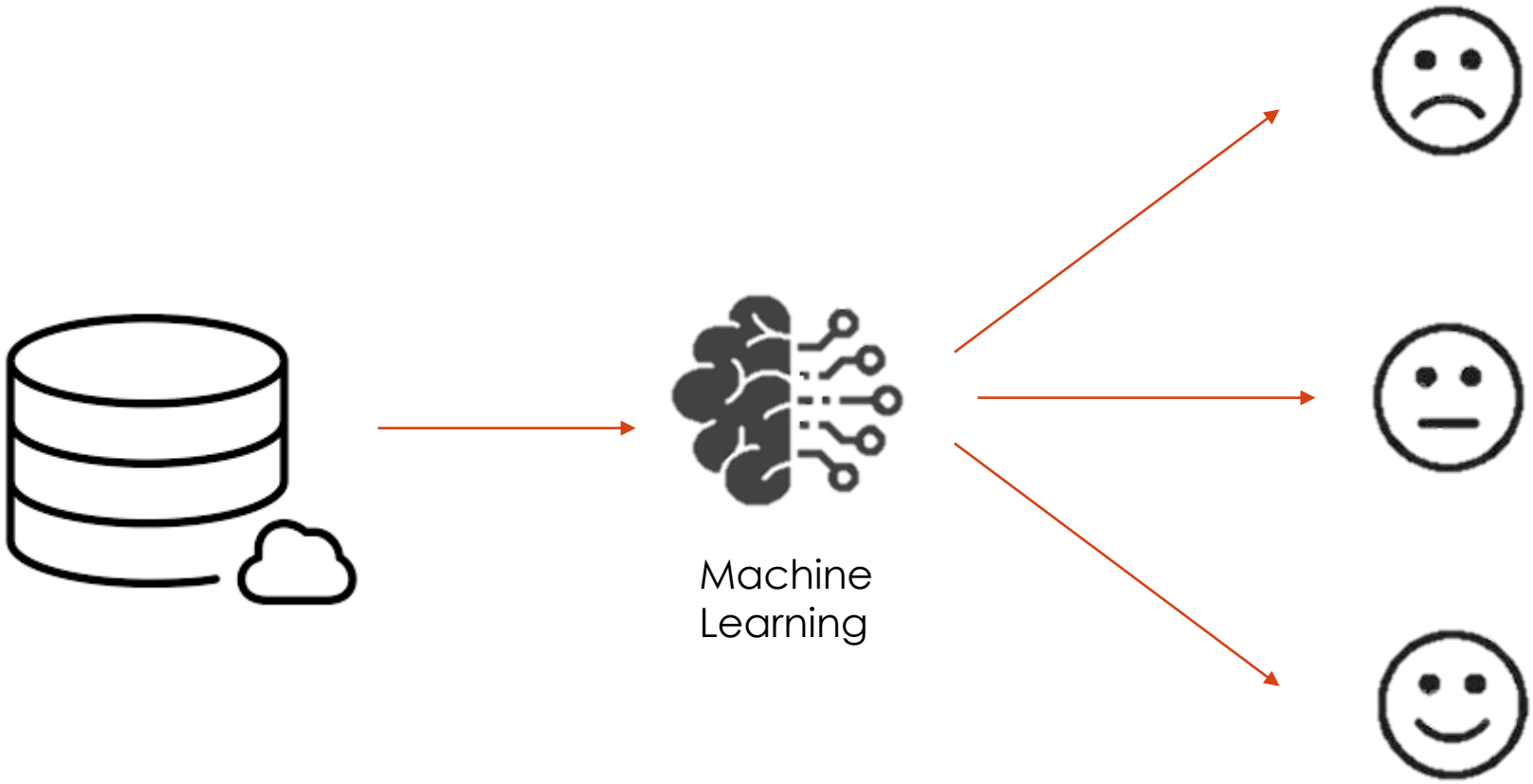
CLASIFICACIÓN: predecir una categoría



Modelos de Machine Learning

Objetivo

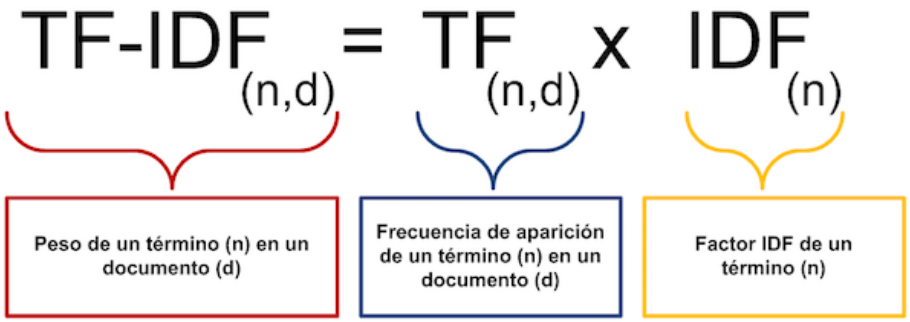
CLASIFICACIÓN: predecir una categoría ← **SUPERVISADO**



Análisis Exploratorio

Preparación de datos – preparación de datos para modelos

TF-IDF



Lematización

Seremos -> Lema: ser
Amigos -> Lema: amigo

N-gramas

- N=1 -> 'Esta es la frase' { Esta, es, la, frase
- N=2 -> 'Esta es la frase' { Esta es, es la, la frase
- N=3 -> 'Esta es la frase' { Esta es la, es la frase

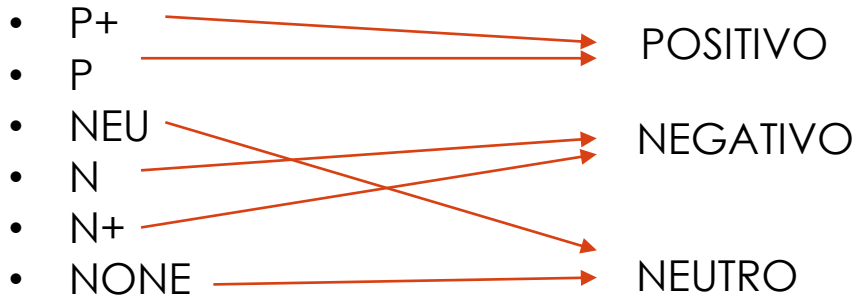
Análisis Exploratorio

Preparación de datos

Clasificación de noticias

+

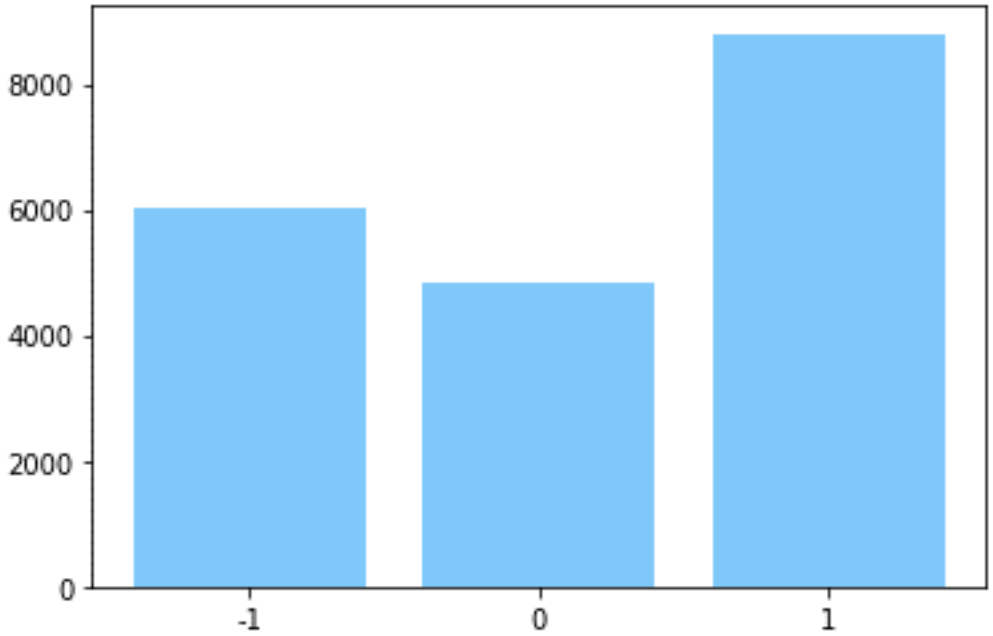
- P+
- P
- NEU
- N
- N+
- NONE

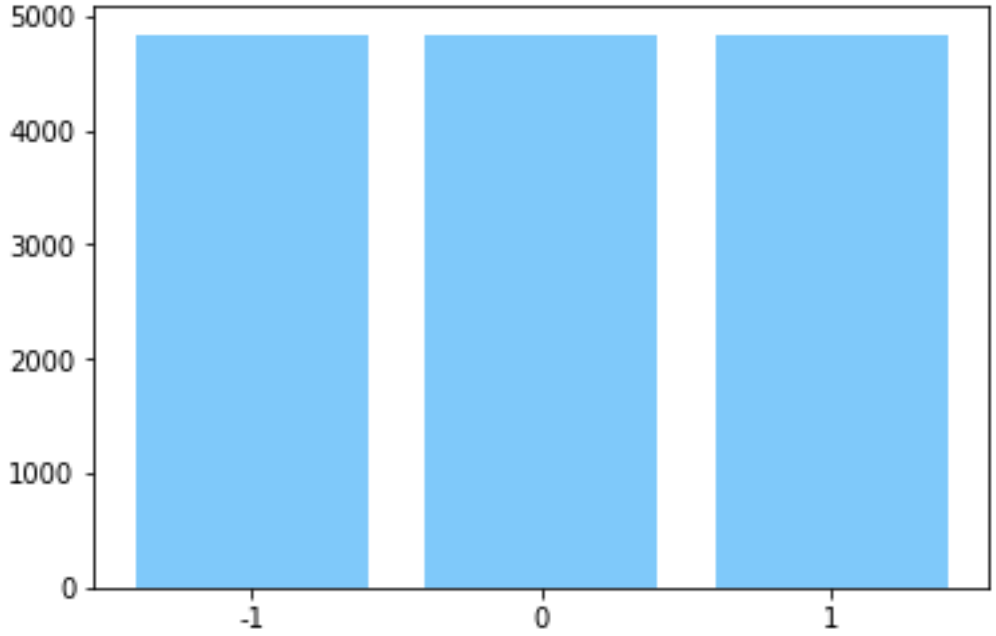


Análisis Exploratorio

Preparación de datos

Transformación de datos - balanceo

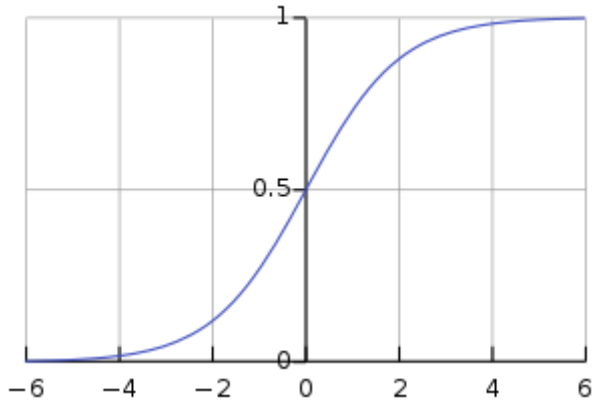




Modelos de Machine Learning

Regresión Logística

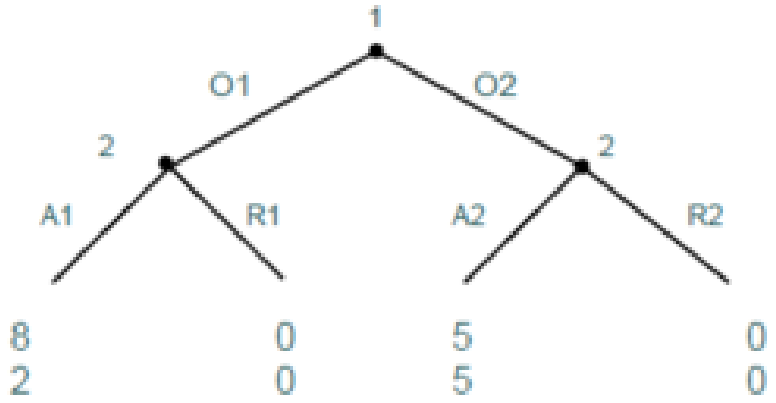
$$\Pi(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}$$



- ✓ Adaptar threshold
- ✓ Simplicidad
- ✓ Facilidad de interpretación

Modelos de Machine Learning

Árbol de Decisión

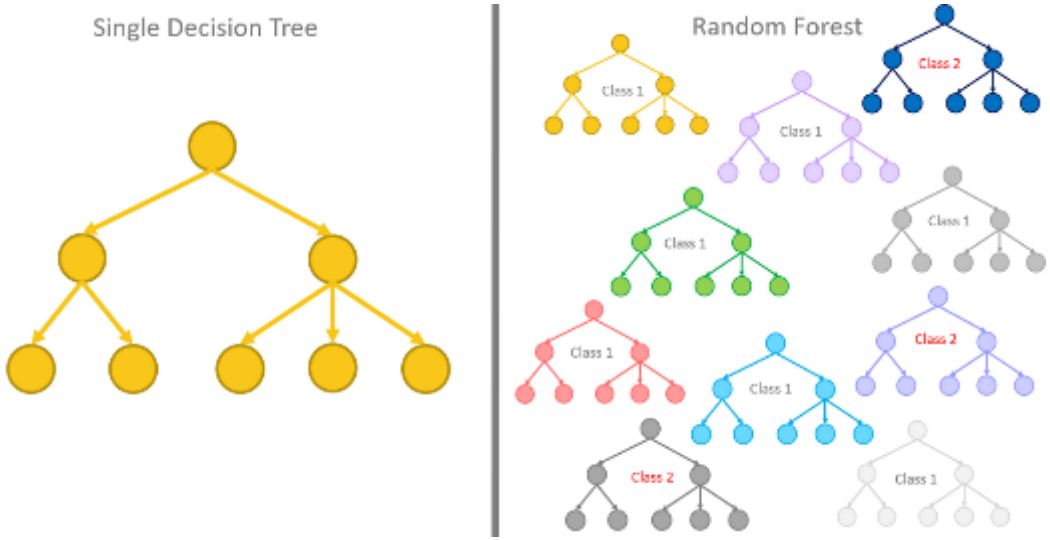


- ✓ Fácil de entender
- ✓ Fácil de interpretar

- × No garantizamos óptimo
- × Aprendizaje términos

Modelos de Machine Learning

Random Forest

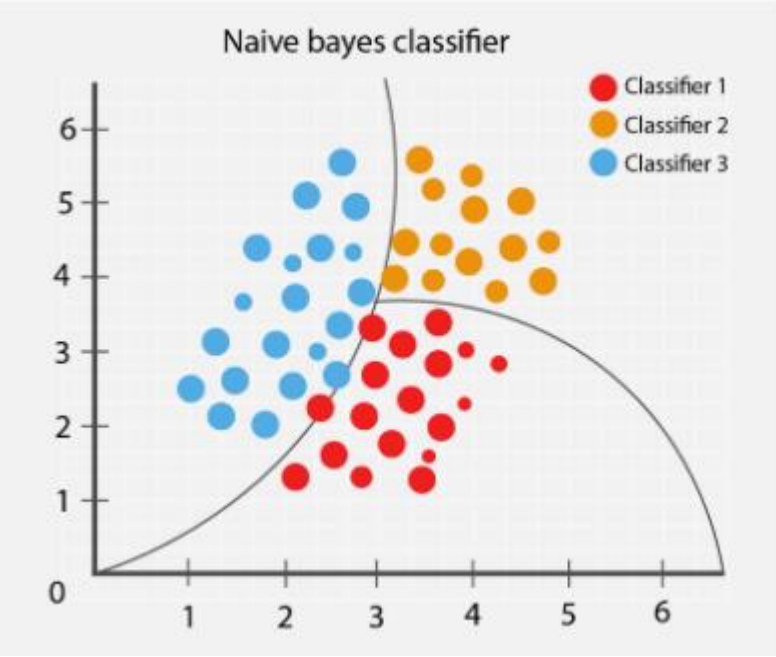


- ✓ Combinación de árboles de decisión
- × Pequeños cambios en train generan árboles muy distintos
- × Pierde interpretabilidad respecto árbol decisión

Modelos de Machine Learning

Naive Bayes

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

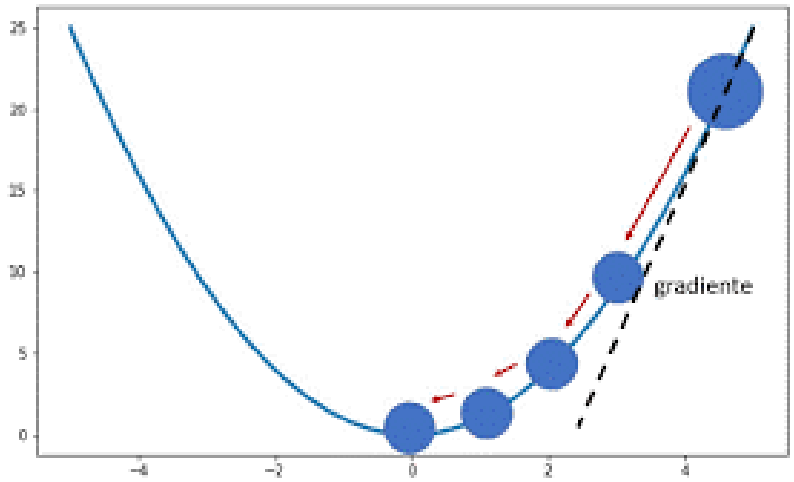


- ✓ Útil en detección de spam, análisis de sentimientos
- ✓ Fácil y rápido de implementar

- × Puede fallar ante características raras

Modelos de Machine Learning

Descenso del Gradiente Estocástico

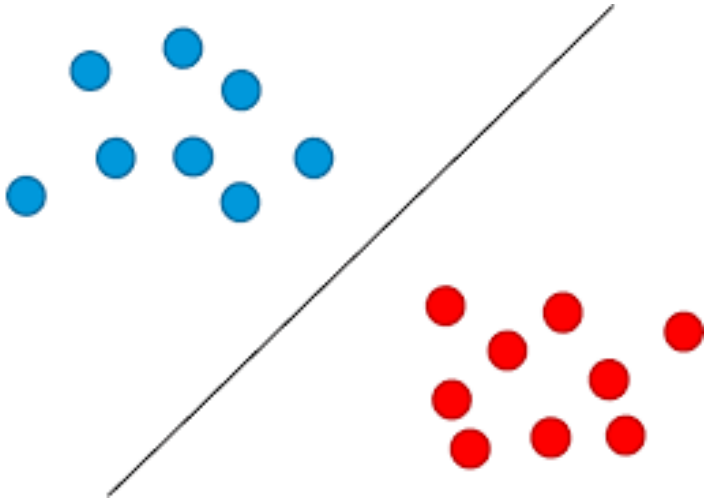


- ✓ Utilizado en problemas de clasificación de texto y procesamiento del lenguaje natural
- ✓ Eficiencia
- ✓ Facilidad de implementación

- × Sensible a hiperparámetros y número de iteraciones

Modelos de Machine Learning

Máquinas de Vectores soporte



- ✓ Robusto frente a ruido
- ✓ Útil en clasificación de texto
- × Difícil de interpretar
- × Gran tiempo de cómputo

Modelos de Machine Learning

Estrategias seguidas

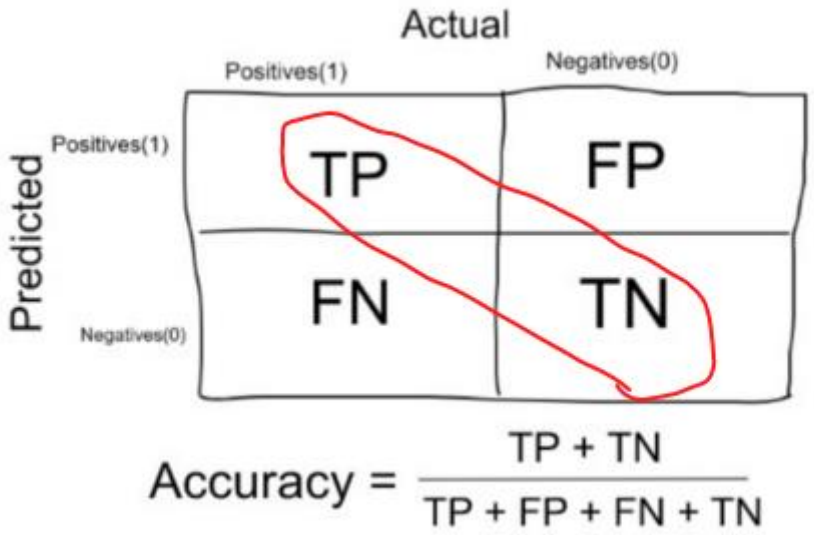
ESTRATEGIA I	ESTRATEGIA II	ESTRATEGIA III	ESTRATEGIA IV
Datos originales	Datos originales + Balanceo	Lematización	Lematización + Balanceo

ESTRATEGIA V	ESTRATEGIA VI	ESTRATEGIA VII	ESTRATEGIA VIII
N-gramas	N-gramas + Balanceo	N-gramas + Lematización	N-gramas + Lematización + Balanceo

Modelos de Machine Learning

Métrica

ACCURACY

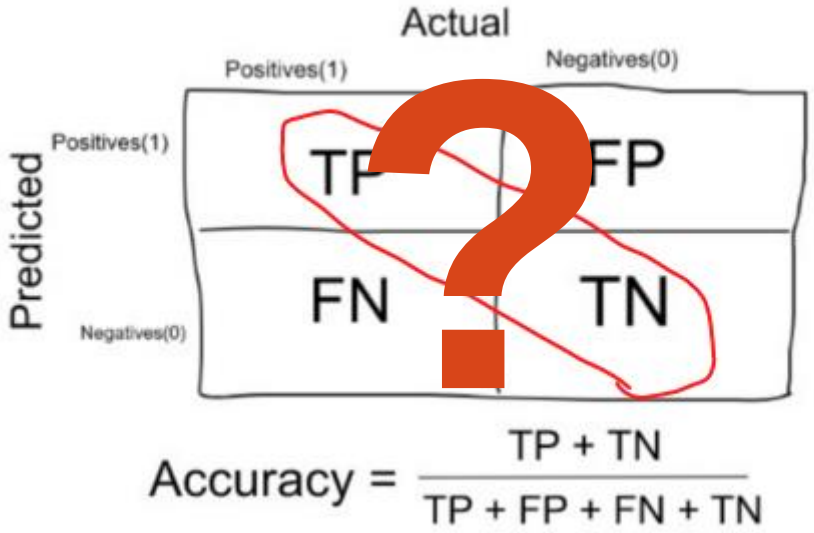


$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Modelos de Machine Learning

Métrica

ACCURACY

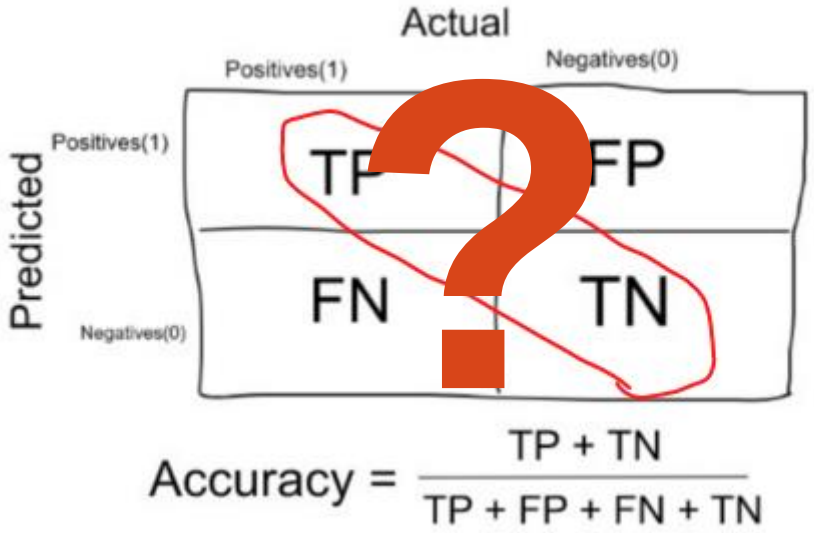


$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Modelos de Machine Learning

Métrica

ACCURACY



$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

KAPPA

$$K = \frac{P_0 - P_e}{1 - P_e}$$

Modelos de Machine Learning

Resultados

ESTRATEGIA I
SGD

Kappa train	0.60
Kappa test	0.42
Accuracy train	0.74
Accuracy test	0.62
Tiempo	20.05

ESTRATEGIA II
SGD

Kappa train	0.62
Kappa test	0.38
Accuracy train	0.75
Accuracy test	0.59
Tiempo	16.23

ESTRATEGIA III
SGD

Kappa train	0.31
Kappa test	0.28
Accuracy train	0.57
Accuracy test	0.55
Tiempo	12.59

ESTRATEGIA IV
SGD

Kappa train	0.46
Kappa test	0.30
Accuracy train	0.64
Accuracy test	0.53
Tiempo	13.67

ESTRATEGIA V
SGD

Kappa train	0.65
Kappa test	0.50
Accuracy train	0.79
Accuracy test	0.74
Tiempo	11.60

ESTRATEGIA VI
SVM

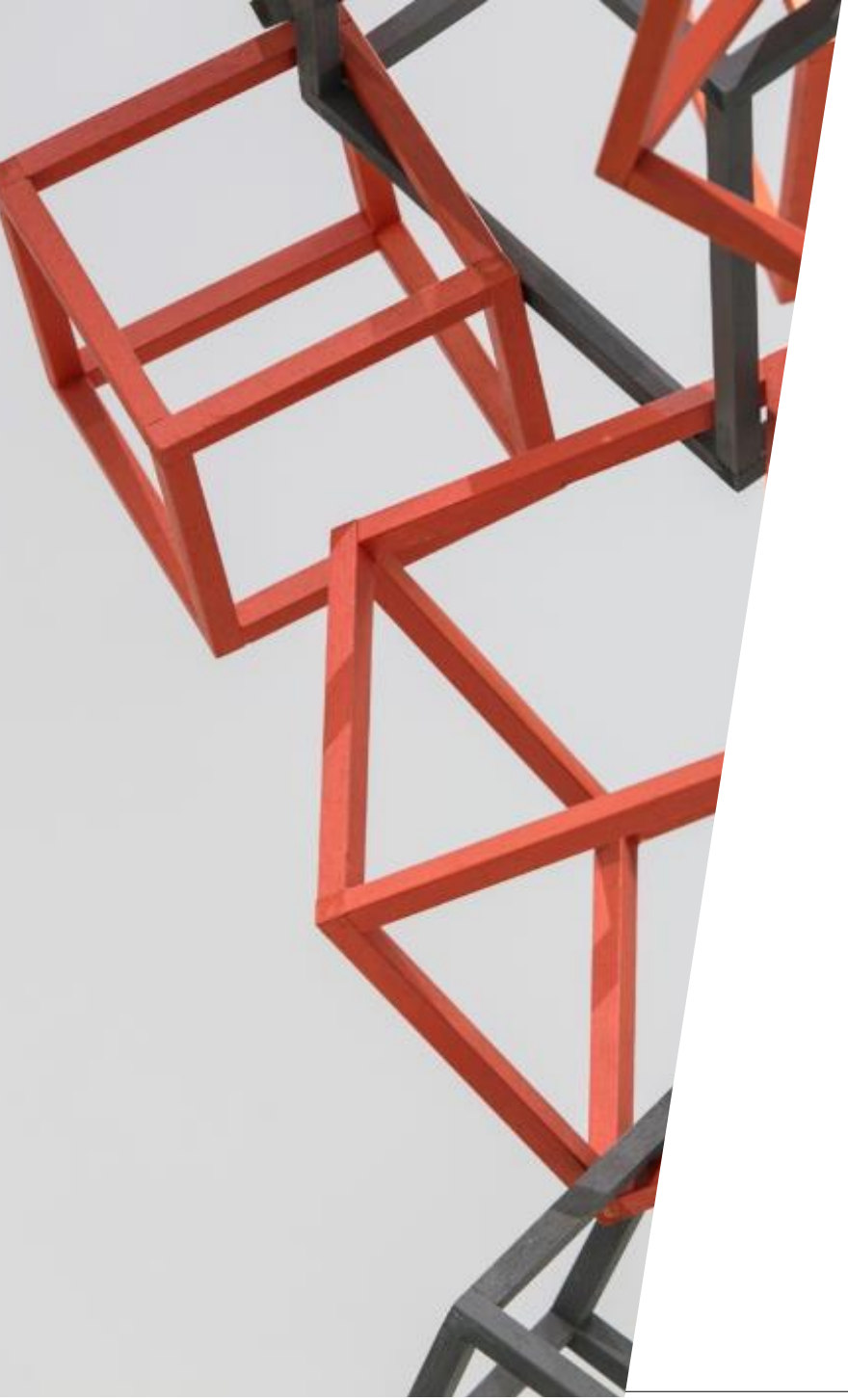
	SGD	SVM
Kappa train	0.63	0.61
Kappa test	0.51	0.5
Accuracy train	0.83	0.83
Accuracy test	0.45	0.8
Tiempo	7.57	0.55

ESTRATEGIA VII
REGRESIÓN LOGISTICA

Kappa train	0.67
Kappa test	0.48
Accuracy train	0.81
Accuracy test	0.74
Tiempo	19.19

ESTRATEGIA VIII
REGRESIÓN LOGISTICA

Kappa train	0.66
Kappa test	0.47
Accuracy train	0.85
Accuracy test	0.81
Tiempo	13.88



Dificultades encontradas

Dificultades encontradas

- Clasificación de noticias para obtener un set de datos grande.
- Número de palabras elegido como columnas de la matriz es muy influyente.
- Mucho trabajo manual de lectura y decisión que influye en los resultados.
- Gran cantidad de overfitting.
- En español hay muchas menos librerías y técnicas que en inglés.

Conclusiones

Conclusiones

- No obtenemos muy buenos resultados.
- Hay bastante overfitting en los modelos.
- Combinar las técnicas ha sido una buena idea, por lo que se podrían probar más técnicas tratando de conseguir mejores resultados.
- Los resultados pueden estar condicionados por los datos utilizados.
- Shiny es una buena herramienta para estudiar la evolución de los indicadores teniendo una visión sobre el número de noticias publicadas y el sentimiento asociado como complemento.

¡MUCHAS GRACIAS!

Afi