



Text Mining

Luis Gascó Sánchez, Ph.D.



Cronograma

Día 1:

1. Introducción
 1. Contexto histórico
 2. ¿Qué es el Text Mining?
 3. Librerías de programación para Text Mining
2. Técnicas y conceptos básicos de NLP
3. Representación numérica de documentos textuales



Cronograma

Día 2:

1. Técnicas de Text Mining:
 1. Flujo de los datos
 2. Clasificación
 3. Topic Modeling
2. Caso de estudio: Análisis de sentimiento



1. Técnicas de Text Mining

Flujo de los datos



Clasificación de textos



1. Técnicas de Text Mining – Clasificación de textos

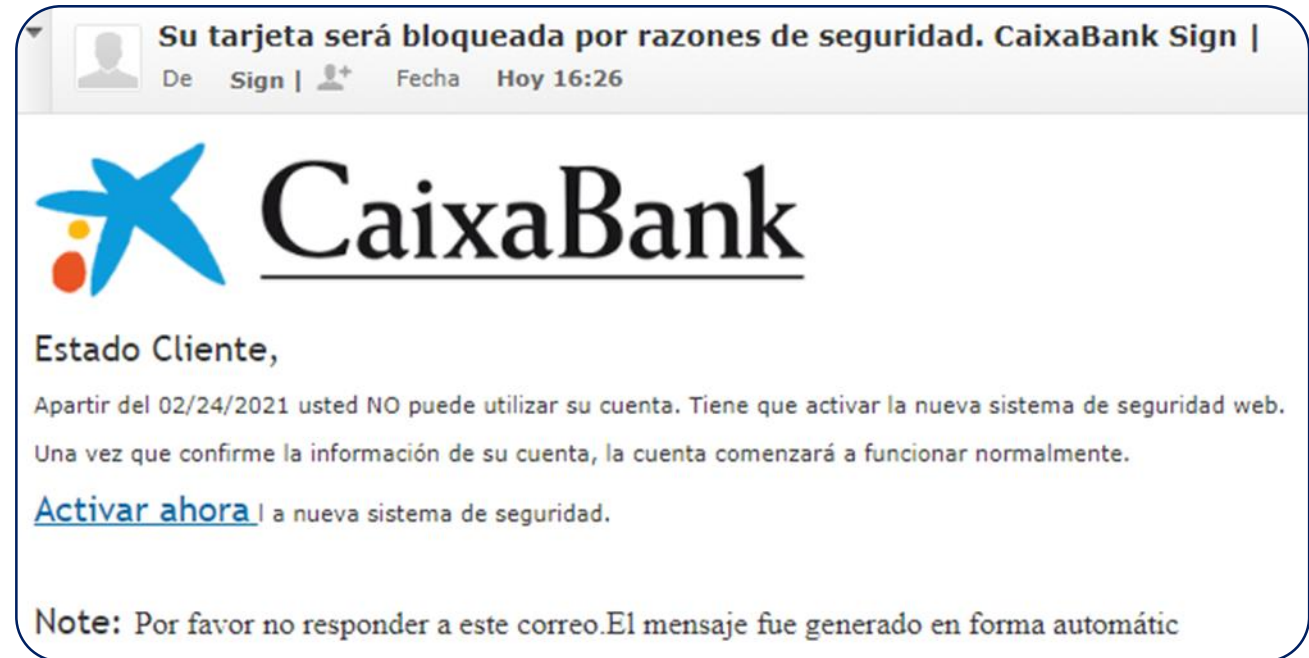
Usos

SPAM

Análisis de sentimiento

Discurso de odio

Fake news



1. Técnicas de Text Mining – Clasificación de textos


Usos


SPAM

Análisis de sentimiento

Discurso de odio


Fake news


**Hard Rock Cafe**
\$\$ · American (New), Burgers, Music Venues
1000 Universal Studios Blvd
Universal City, CA 91608


 1/11/2014


Lat night we had the best time at dinner that I think I've ever had! Our server Jessica was adorable and full of personality plus. She was attentive, entertaining and knowledgeable. She is a huge asset to your restaurant and deserves much kudos. We will definitely be back soon!


Was this review ...?


 Useful

 Funny

 Cool





**Comment from Scott B. of Hard Rock Cafe**
Business Manager

1/21/2014 · Hi Debra,

Thank you very much for such great review! We work really hard to offer the best food in the best possible environment, so I am happy to see that reflected on your last visit. I made sure that Jessica got the recognition she deserved. She was really happy to see your review.

Hope to see you again soon!

German Crespi
AGM [Read less](#)

1. Técnicas de Text Mining – Clasificación de textos

Usos

SPAM

Análisis de sentimiento

Discurso de odio

Fake news



1. Técnicas de Text Mining – Clasificación de textos

Usos

SPAM

Análisis de sentimiento

Discurso de odio

Fake news

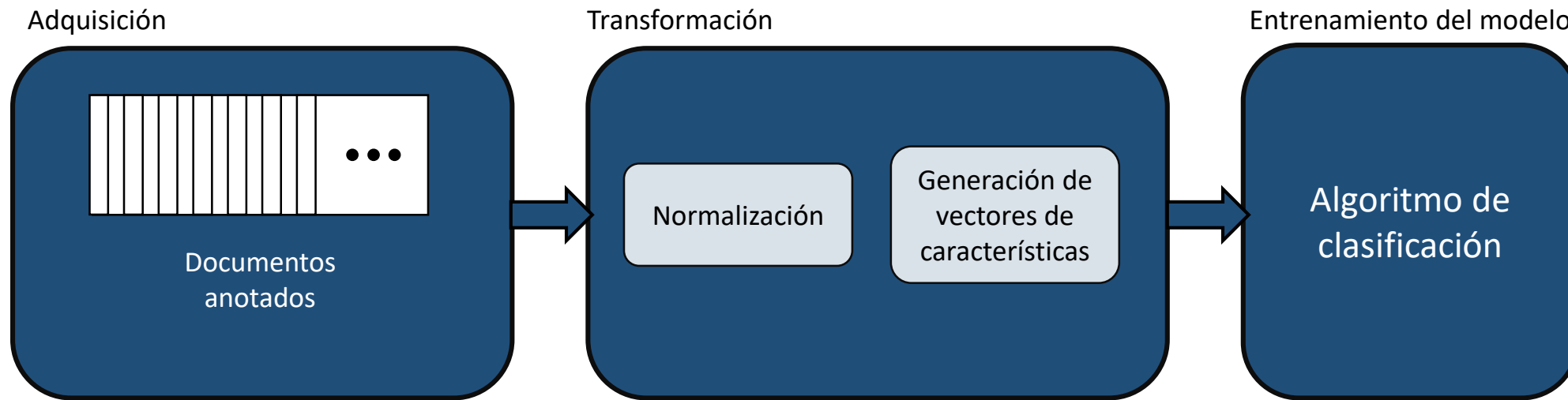


Google Fact Check Tools



1. Técnicas de Text Mining – Clasificación de textos

Flujo de los datos



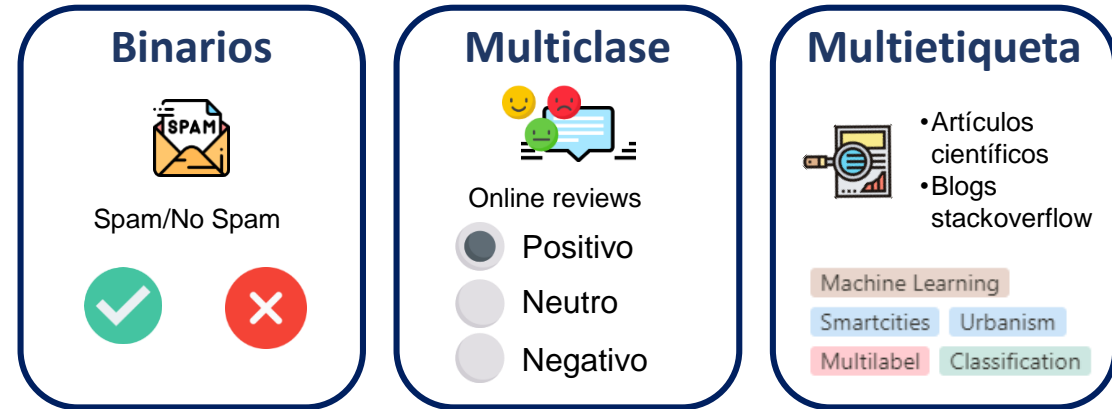
1. Técnicas de Text Mining – Clasificación de textos

Obtención de datos

Adquisición corpus anotados



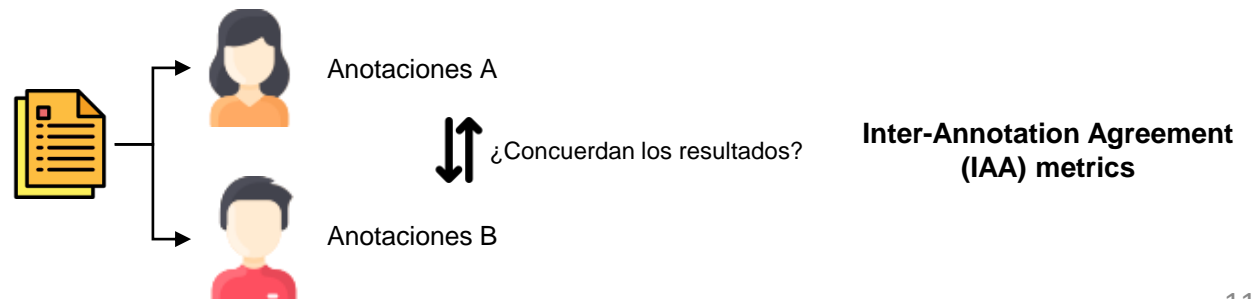
Tipos de datos anotados



¿Como anotar?



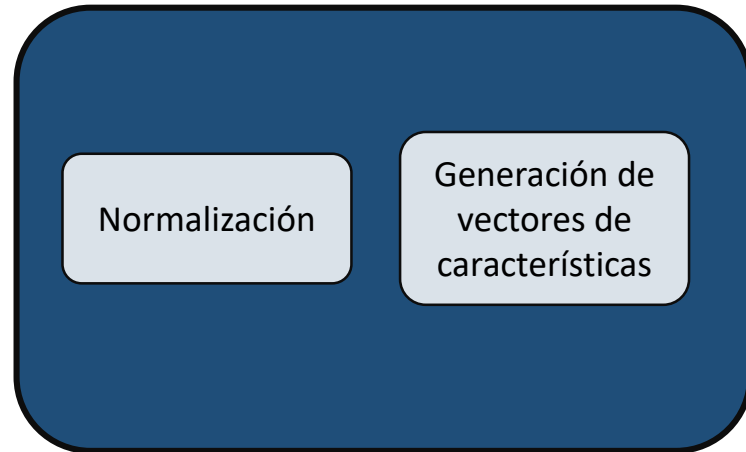
¿Calidad de anotación?



1. Técnicas de Text Mining – Clasificación de textos

Transformación

Transformación



Normalización



Reducir
dimensionalidad sin
bajar rendimiento

1. Transformación a minúsculas
2. Stemming/Lematización
3. Eliminación de palabras vacías y puntuación

Generación de vectores de características



Datos comprensibles
por el modelo de ML

1. TF-IDF incluyendo unigramas, bigramas, trigramas
2. Omitir n-gramas poco frecuentes
3. Incorporar información de embeddings

1. Técnicas de Text Mining – Clasificación de textos

Entrenamiento del modelo

Entrenamiento del modelo

Algoritmo de
clasificación

Entrenamiento

Naïve Bayes

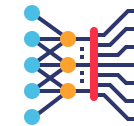
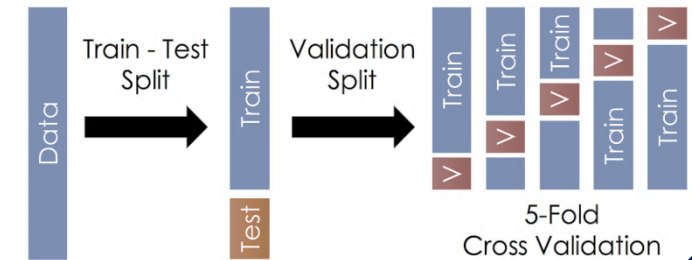
Regresor logístico

SVM

Validación



Machine Learning Tradicional



Redes Neuronales Profundas



Topic Modeling



1. 1. Técnicas de Text Mining – Topic Modeling

Introducción

Gran volumen de datos no etiquetados



P.ej: Cada día más de 500 millones de Tweets

+

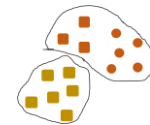
Proceso de anotación



No siempre hay tiempo para anotación (ni presupuesto)

=

Técnicas no supervisadas



- No se necesitan datos anotados
- Buenas para análisis descriptivos, que pueden ser suficientes

1. 1. Técnicas de Text Mining – Topic Modeling

Definición

Uso de modelos estadísticos para la extracción de los temas tratados en un corpus

¿Qué entendemos por “topic” en este contexto?

Conjunto de palabras que es más probable que aparezcan en un mismo contexto

1. 1. Técnicas de Text Mining – Topic Modeling

Topic Modeling

Corpus

Doc1: A María le encantan los animales. Disfruta mucho paseando a sus **perros** y montando a **caballo**. En ocasiones, cuida de los **gatos** de sus amigos. Suele alimentarles con las sobras de sus comidas siempre que sean proteínas como **pollo** y **ternera**. Nunca les da **pescado**.

Doc2: Él es experto gastronómico para la Guía Michelin. Ha asistido a los mejores restaurantes del mundo y ha probado la mejor carne de **ternera** y el mejor **sushi** del mundo en Japón

Doc3: El **perro** de Juan murió hace 6 meses. Él todavía no lo ha superado, por eso está planteando adoptar un **gato** y comprar un **conejo** para no sentirse tan solo.

Algoritmo de
Topic
Modeling

Topic 1

Perros
Caballo
Gatos
Conejo

Topic 2

Pollo
Ternera
Pescado
Sushi

Doc1: Topic 1 (70%) y Topic 2 (30%)
Doc2: Topic 2 (100%)
Doc3: Topic 1 (100%)

Nº de topics

1. 1. Técnicas de Text Mining – Topic Modeling

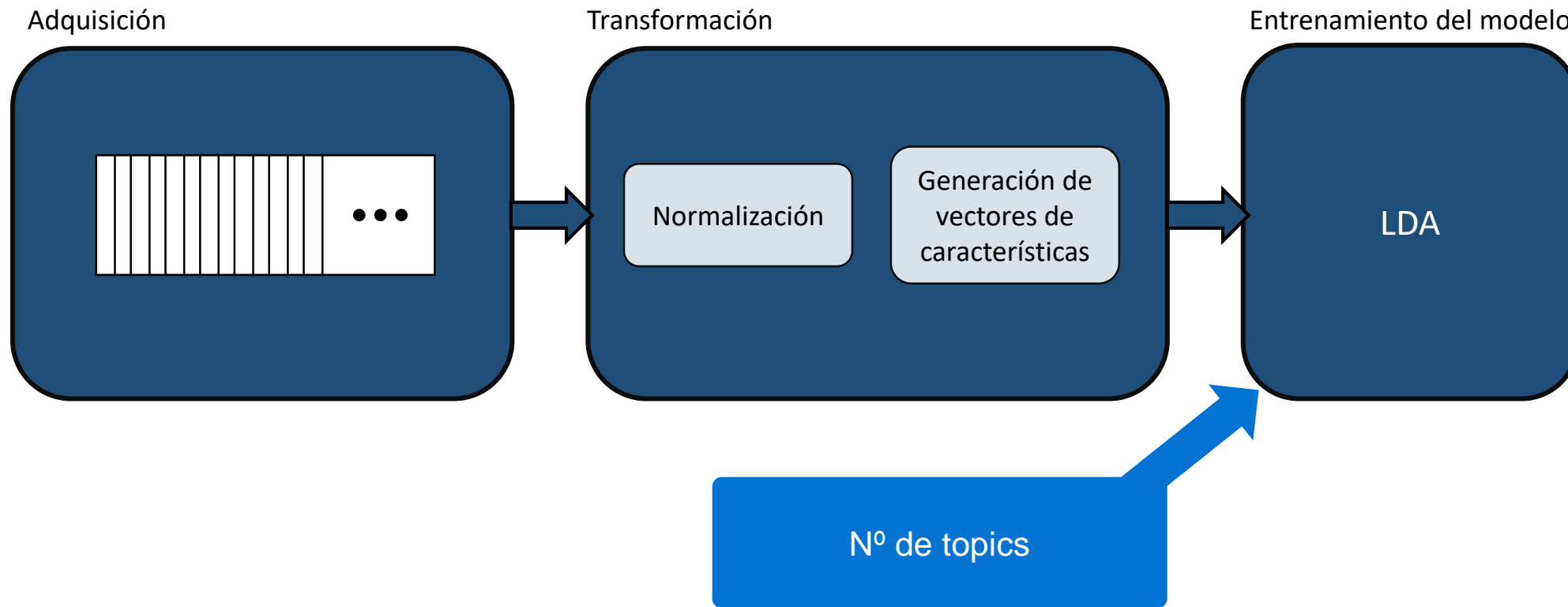
Latent Dirichlet Allocation (LDA)



- Creado en el año 2003 por David Blei, Andrew Ng y Michael Jordan
- Asume que los documentos tienen una estructura semántica implícita (los topics)
- Estos topics representados por palabras, podrían reconstruir cualquier texto del corpus:
- $\text{Un documento} = \text{Topic1} + \text{Topic2} + \dots + \text{TopicN}$

1. Técnicas de Text Mining – Topic Modeling

Flujo de los datos



1. Técnicas de Text Mining – Topic Modeling

Transformación

Normalización y vectorización

- Quitar puntuación
- Transformar a minúsculas
- Quitar stopwords
- Lematizar
- Generar unigramas, bigramas y/o trigramas

Nº Topics

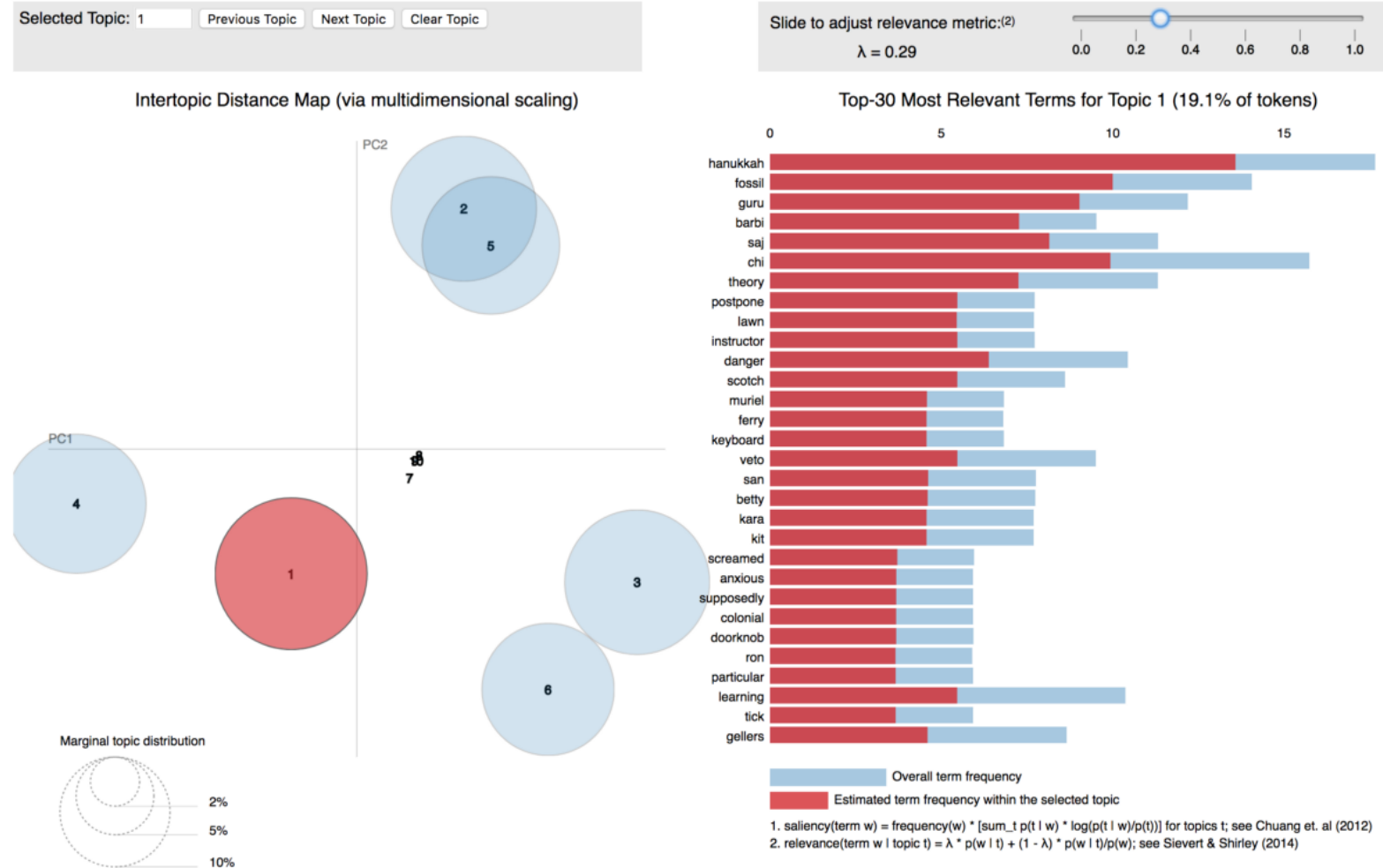
¿Cómo elegir el nº de topics?

Proceso iterativo y validación

1. Técnicas de Text Mining – Topic Modeling

Validación cualitativa

Librería pyLDAvis

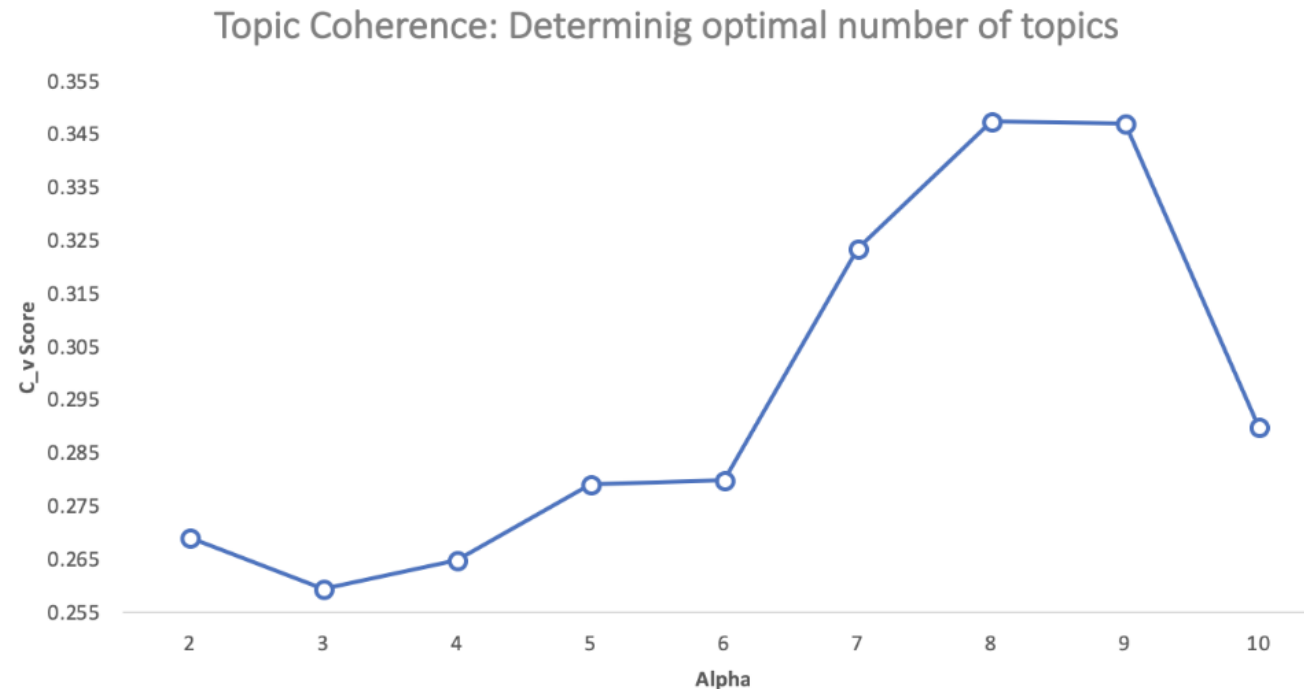


1. Técnicas de Text Mining – Topic Modeling

Validación cuantitativa

Similitud semántica entre topics

Medidas de coherencia



Análisis de sentimiento



2. Caso de estudio: Análisis de sentimiento

Introducción



Figure 2: Customer heat maps for three local businesses in Phoenix

2. Caso de estudio: Análisis de sentimiento

Niveles de análisis

Nivel de documento

Camera Good points:

- The lens. It's very well made - feels super solid, fast to focus and very light weight. See photo of it on my A7Riii
- The silver finish looks nice and feels very good - better quality than Fujifilm for example
- The start-up and autofocus is a little faster and more accurate than both my A7Riii and A6500
- The size is quite good but needs to be just a little taller - even for medium hands

Camera Bad points:

- The grip is a real step backwards - way too shallow to feel comfortable and looks REALLY cheap
- The size simply too small to grip, only 3 fingers can really sit properly
- The EVF - its really small (much smaller than my A7Riii) and the resolution seems very low - it's not a nice experience
- The price is ridiculous compared to say the Nikon Z6ii or Sony A7iii with better specs and build
- No custom buttons, means it's not as quick to operate in any mode other than full auto
- No built in flash - which my A6500 has, and is essential for smaller sensor and slower lens, because once you have the flash on APSC, you make up for the light loss (and more) from the smaller sensor.



Clasificación

2. Caso de estudio: Análisis de sentimiento

Niveles de análisis

Nivel de frase

Camera Good points:

- The lens. It's very well made - feels super solid, fast to focus and very light weight. See photo of it on my A7Riii
- The silver finish looks nice and feels very good - better quality than Fujifilm for example
- The start-up and autofocus is a little faster and more accurate than both my A7Riii and A6500
- The size is quite good but needs to be just a little taller - even for medium hands

Camera Bad points:

- The grip is a real step backwards - way too shallow to feel comfortable and looks REALLY cheap
- The size simply too small to grip, only 3 fingers can really sit properly
- The EVF - its really small (much smaller than my A7Riii) and the resolution seems very low - it's not a nice experience
- The price is ridiculous compared to say the Nikon Z6ii or Sony A7iii with better specs and build
- No custom buttons, means it's not as quick to operate in any mode other than full auto
- No built in flash - which my A6500 has, and is essential for smaller sensor and slower lens, because once you have the flash on APSC, you make up for the light loss (and more) from the smaller sensor.

Clasificación



6



1



3

Texto negativo

2. Caso de estudio: Análisis de sentimiento

Niveles de análisis

Nivel de entidad

Camera Good points:

- The lens. It's very well made - feels super solid, fast to focus and very light weight. See photo of it on my A7Riii
- The silver finish looks nice and feels very good - better quality than Fujifilm for example
- The start-up and autofocus is a little faster and more accurate than both my A7Riii and A6500
- The size is quite good but needs to be just a little taller - even for medium hands

Camera Bad points:

- The grip is a real step backwards - way too shallow to feel comfortable and looks REALLY cheap
- The size simply too small to grip, only 3 fingers can really sit properly
- The EVF - its really small (much smaller than my A7Riii) and the resolution seems very low - it's not a nice experience
- The price is ridiculous compared to say the Nikon Z6ii or Sony A7iii with better specs and build
- No custom buttons, means it's not as quick to operate in any mode other than full auto
- No built in flash - which my A6500 has, and is essential for smaller sensor and slower lens, because once you have the flash on APSC, you make up for the light loss (and more) from the smaller sensor.

Cámara



**Problema más complejo:
NLP + NER +Entity Linking**

¿Qué palabras expresan sentimiento?
¿Cómo saber si es positivo o negativo?



Negative

I'm **dissatisfied** with your customer service.
No one was able to help me with the **problems** I had with using your product.



Neutral

The product has multiple features that are suitable for users with different levels of experience.



Positive

I **really enjoy** how **easy** this product is to use and how it **successfully helps** my team complete their day-to-day tasks.

2. Caso de estudio: Análisis de sentimiento

Sentiment Lexicons

Los Léxicos de sentimientos son diccionarios de partículas de sentimientos que asocian cada palabra con un índice de positividad, negatividad o neutralidad

Basados en WordNet 3.0

- [SentiWordNet](#)
- [WordNet-Affect](#)

Diccionarios no jerárquicos

- [Bin Liu](#)
- EmoLex

Palabra	ind _{pos}	ind _{neg}	ind _{neut}
Bonito	1	0	0
Feo	0	1	0
Asequible	0.3	0.3	0.9



¿Y que pasa con los emojis?

2. Caso de estudio: Análisis de sentimiento

Sentiment Lexicons

Los emojis ayudan a transmitir sentimiento de forma pictórica

I ❤️ the new iPhone camera






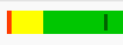

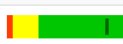



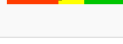


Sustitución por palabras

Emot library

```
EMOTICONS = {  
  u":-\\)": "Happy face or smiley",  
  u":\\)": "Happy face or smiley",  
  u":-\\)": "Happy face or smiley",  
  u":\\)": "Happy face or smiley",  
  u":-3": "Happy face smiley",  
  u":3": "Happy face smiley",  
  u":->": "Happy face smiley",  
  u":>": "Happy face smiley",
```

Diccionario de sentimientos de emojis

Emoji Sentiment Ranking

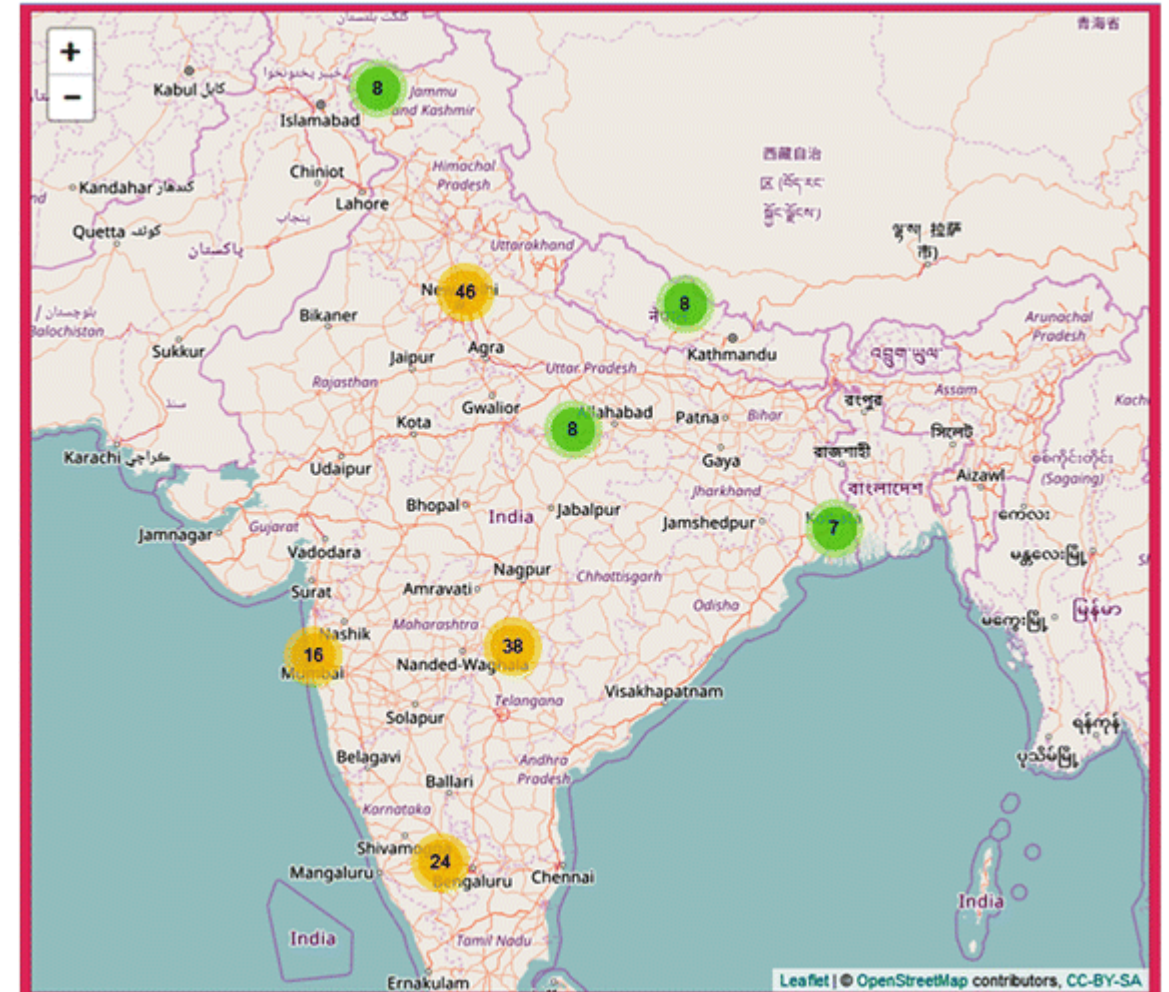
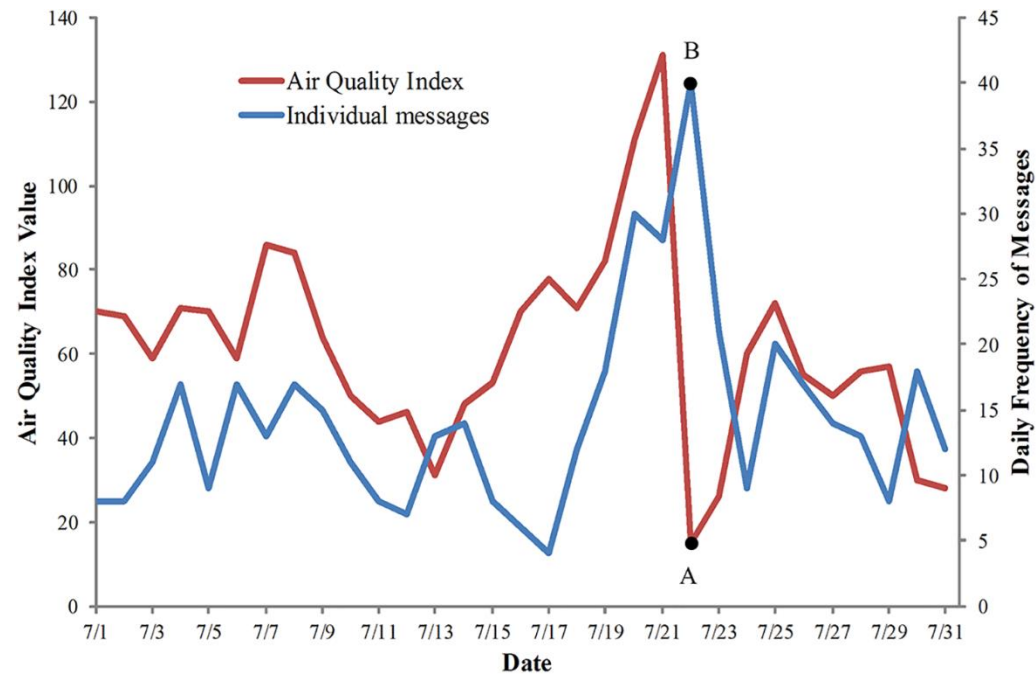
Image [twemoji]	Unicode codepoint	Occurrences [5...max]	Position [0...1]	Neg [0...1]	Neut [0...1]	Pos [0...1]	Sentiment score [-1...+1]	Sentiment bar (c.i. 95%)
	0x1f602	14622	0.805	0.247	0.285	0.468	0.221	
	0x2764	8050	0.747	0.044	0.166	0.790	0.746	
	0x2665	7144	0.754	0.035	0.272	0.693	0.657	
	0x1f60d	6359	0.765	0.052	0.219	0.729	0.678	
	0x1f62d	5526	0.803	0.436	0.220	0.343	-0.093	
	0x1f618	3648	0.854	0.053	0.193	0.754	0.701	
	0x1f60a	3186	0.813	0.060	0.237	0.704	0.644	

Análisis de Sentimiento en Twitter



2. Caso de estudio: Análisis de sentimiento

Análisis de sentimiento de Tweets



2. Caso de estudio: Análisis de sentimiento

Análisis de sentimiento de Tweets: Sobre Twitter



2. Caso de estudio: Análisis de sentimiento

Análisis de sentimiento de Tweets: Fases



2. Caso de estudio: Análisis de sentimiento

Análisis de sentimiento de Tweets: Fases

Obtención
de datos



Twitter API



Mayoría Retweets

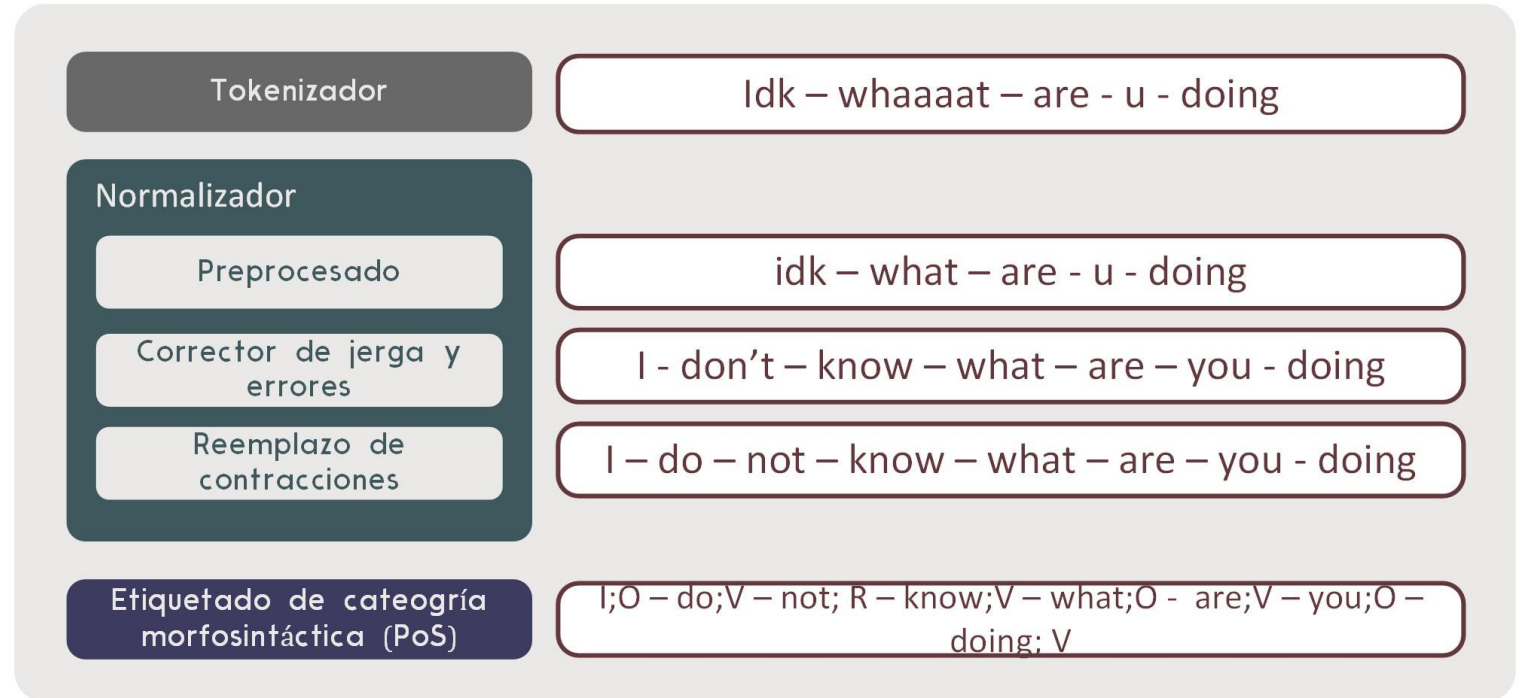


Multitud de emojis

2. Caso de estudio: Análisis de sentimiento

Análisis de sentimiento de Tweets: Transformación - Normalización

Transformación de
datos



2. Caso de estudio: Análisis de sentimiento

Análisis de sentimiento de Tweets: Transformación - Características



Transformación de
datos

TF-IDF

Características a partir de PoS

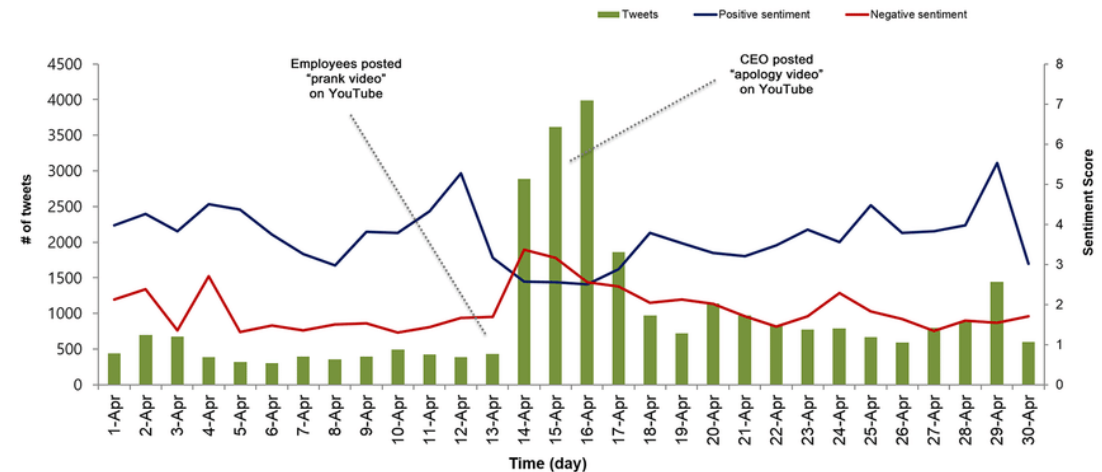
Características de sentimiento

Word Embeddings

2. Caso de estudio: Análisis de sentimiento

Análisis de sentimiento de Tweets: Transformación - Características

Análisis y/o creación de modelos



2. Caso de estudio: Análisis de sentimiento

Análisis de sentimiento de Tweets: Transformación - Visualización

