Análisis de Sentimientos

M. en C. Luis Norberto Zúñiga Morales

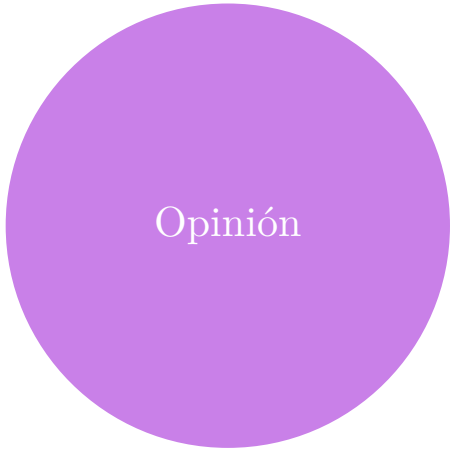
29 de marzo de 2022

Contenido

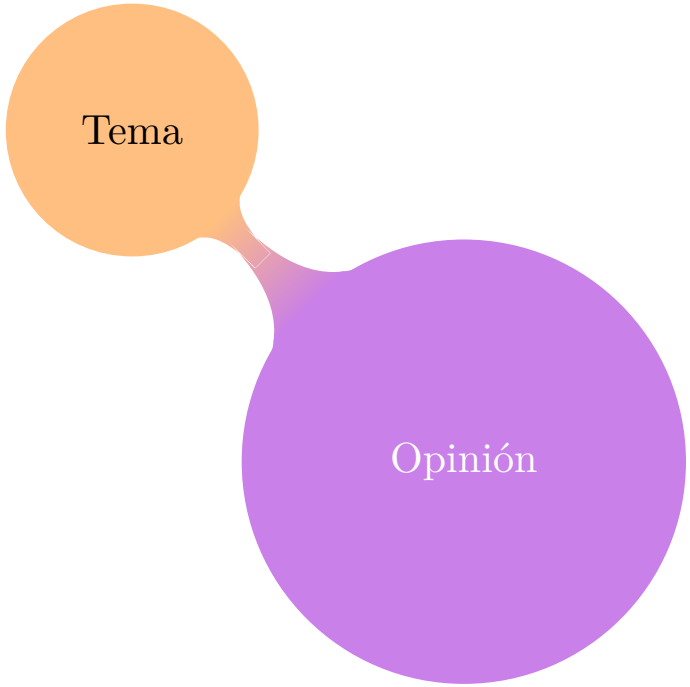
Sentimiento y Emoción

- El Análisis de Sentimientos [1] busca analizar el sentimiento presente en un texto para determinar las polaridades de las opiniones, juicios, pensamientos, actitudes y emociones expresadas por el usuario hacia distintos aspectos de un tema.
- Una opinión [2] es una postura que una persona tiene hacia cierto tema, la cual puede reflejar una emoción.
- Las emociones [3] son complejos estados sentimentales que activan reacciones físicas y fisiológicas que afectan la conducta y el pensamiento.

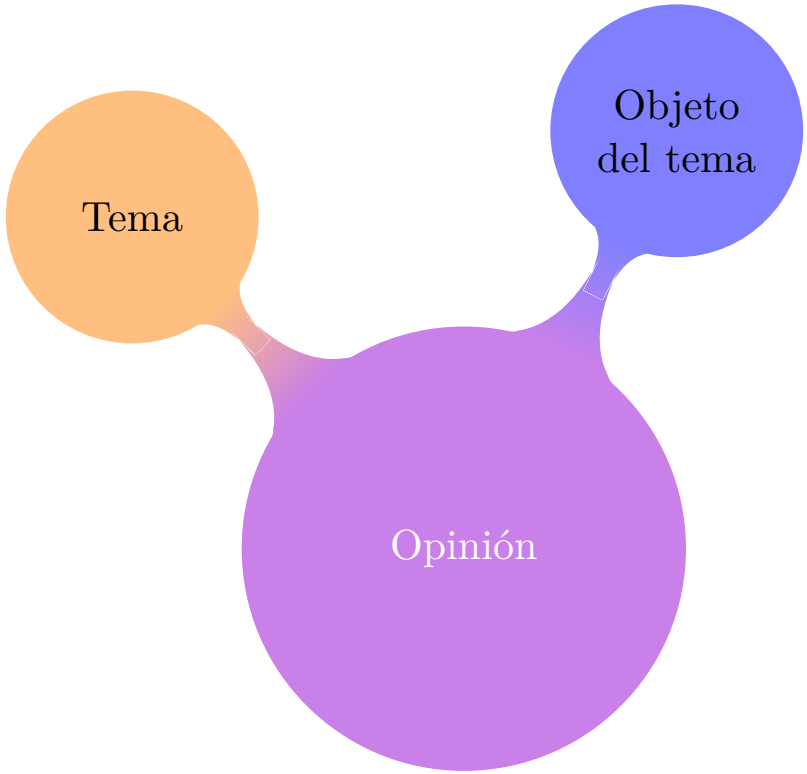
Introducción



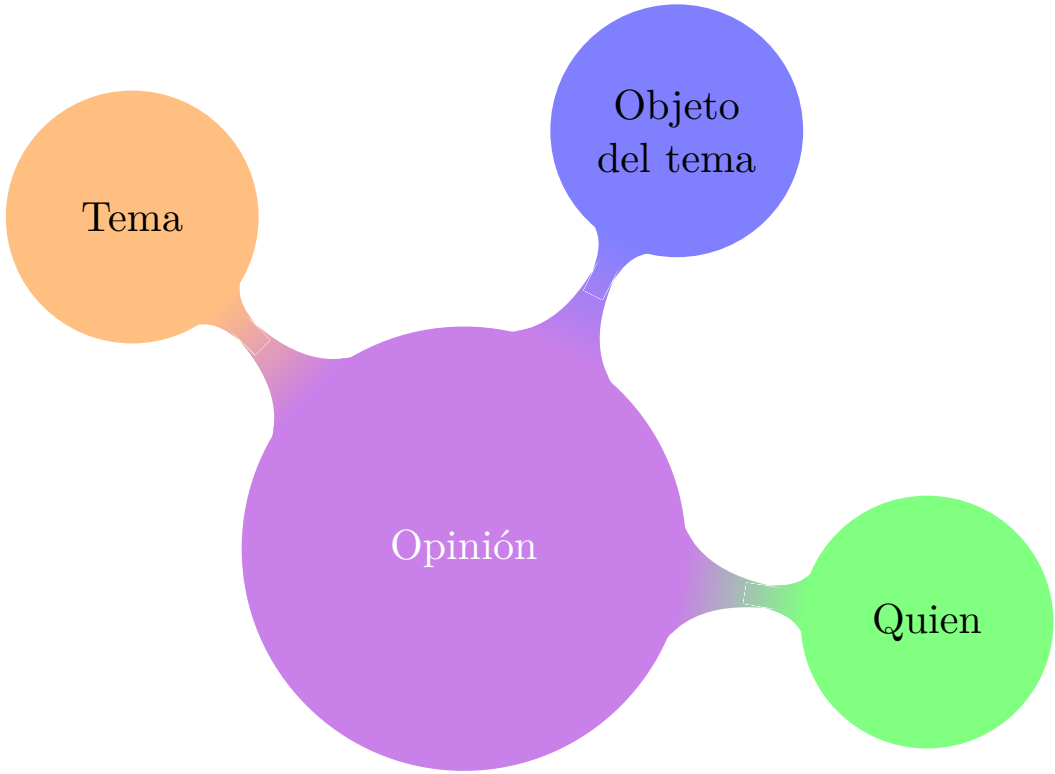
Introducción



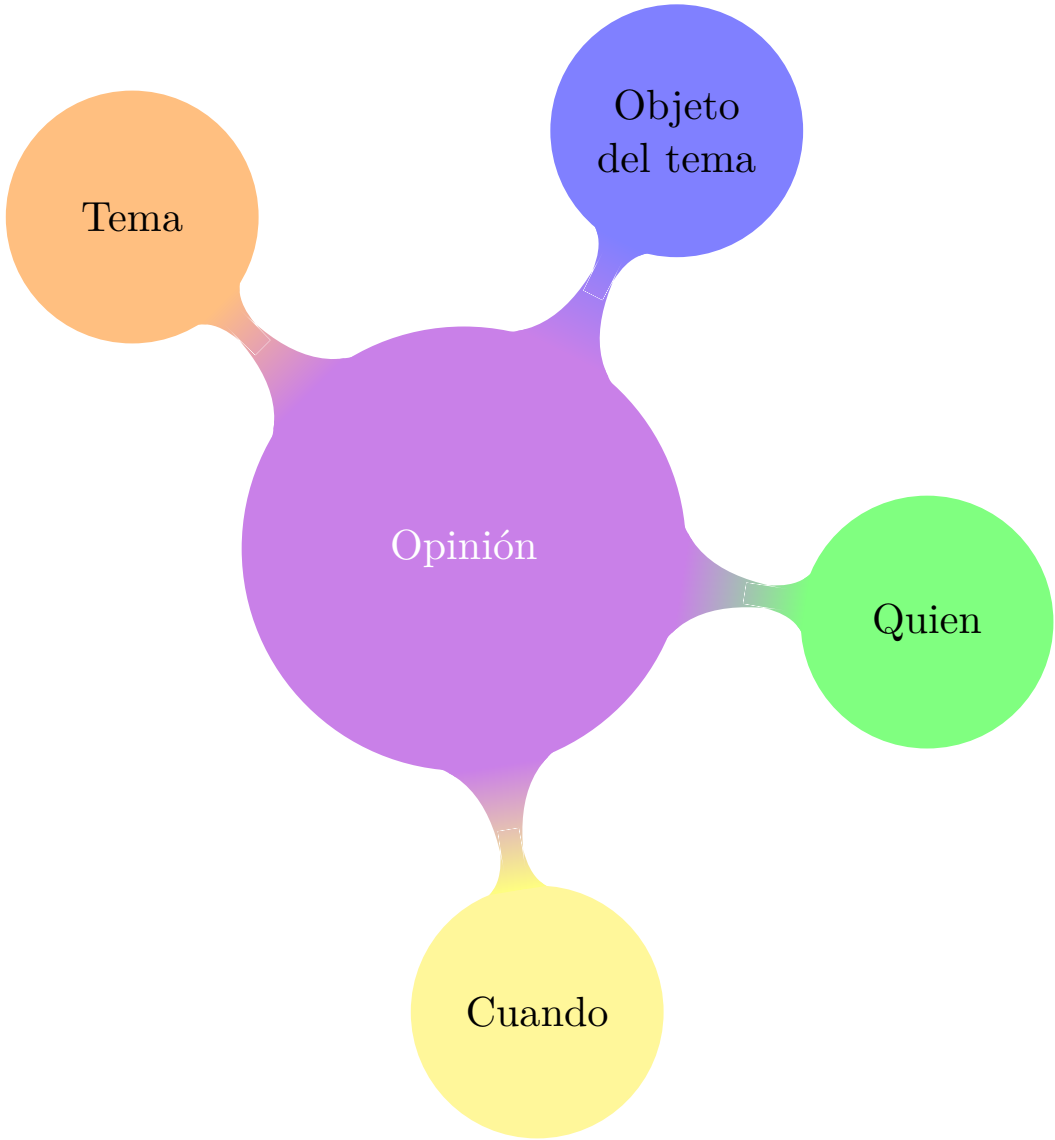
Introducción



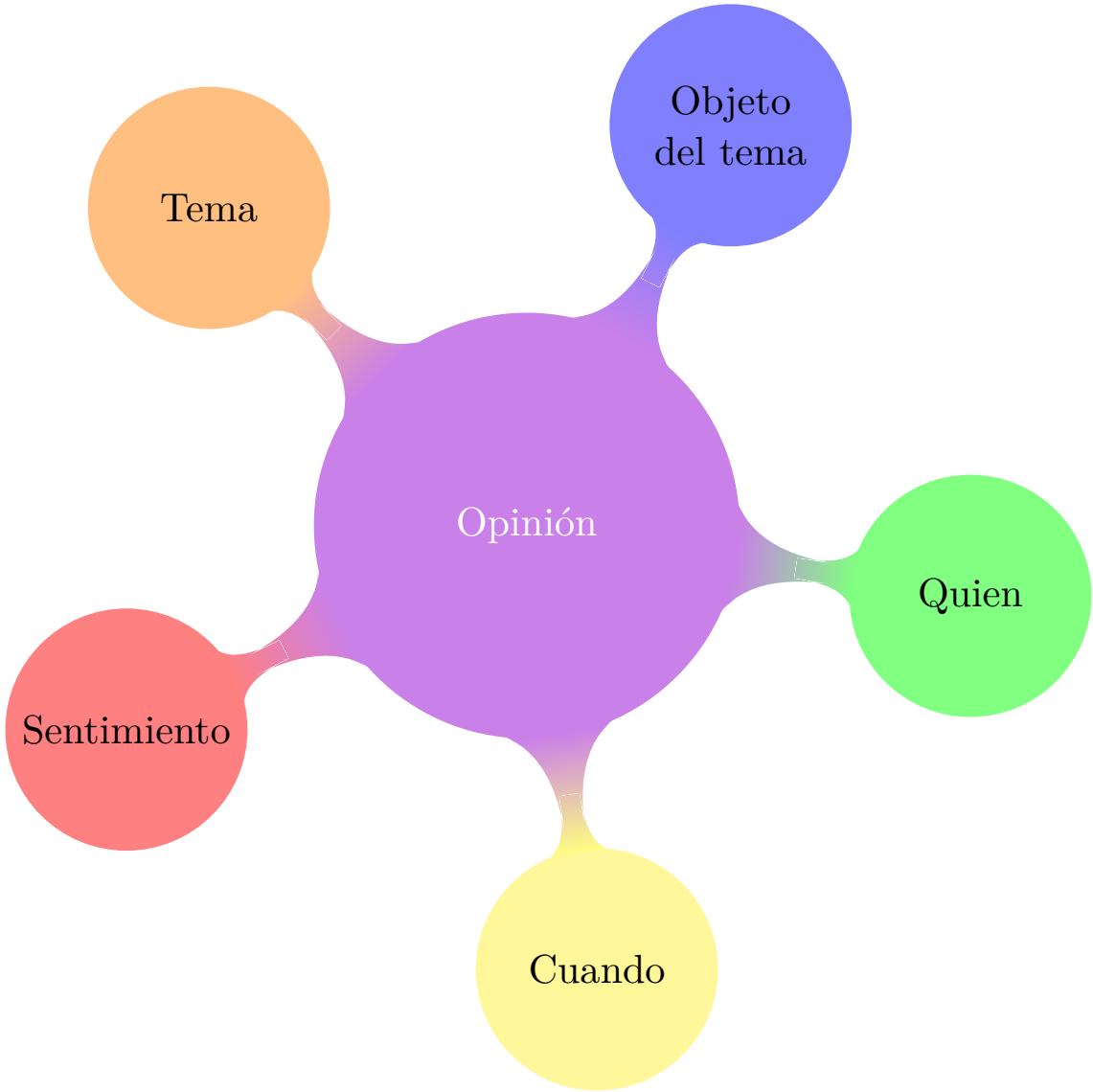
Introducción



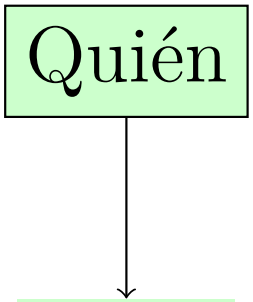
Introducción



Introducción

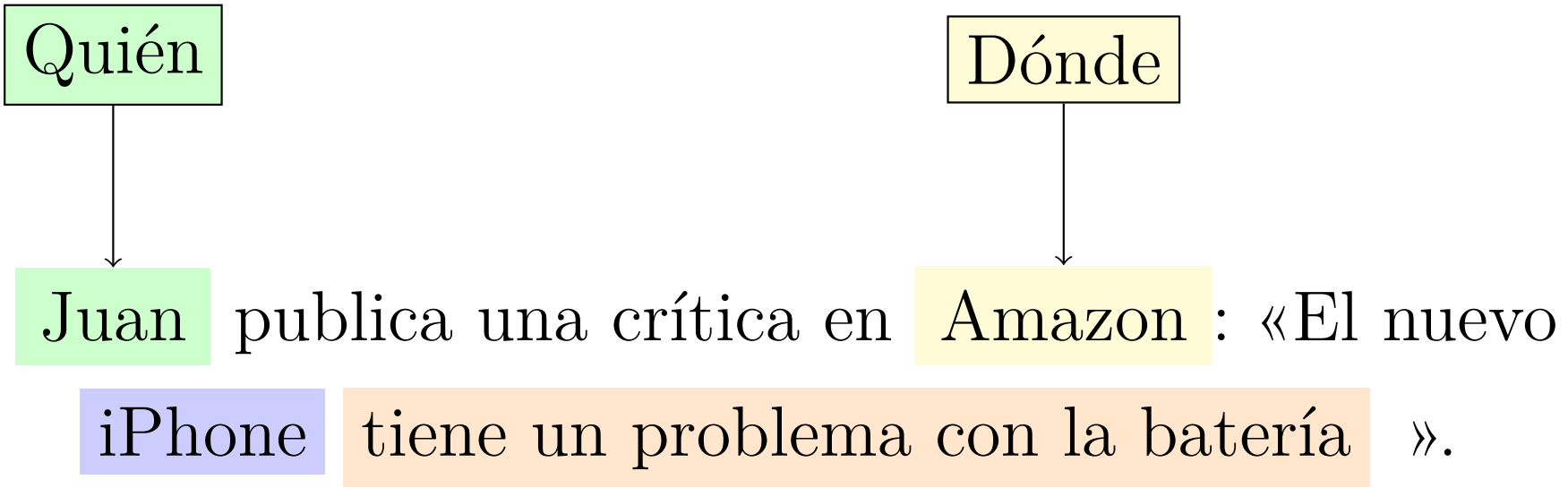


Componentes de una Opinión

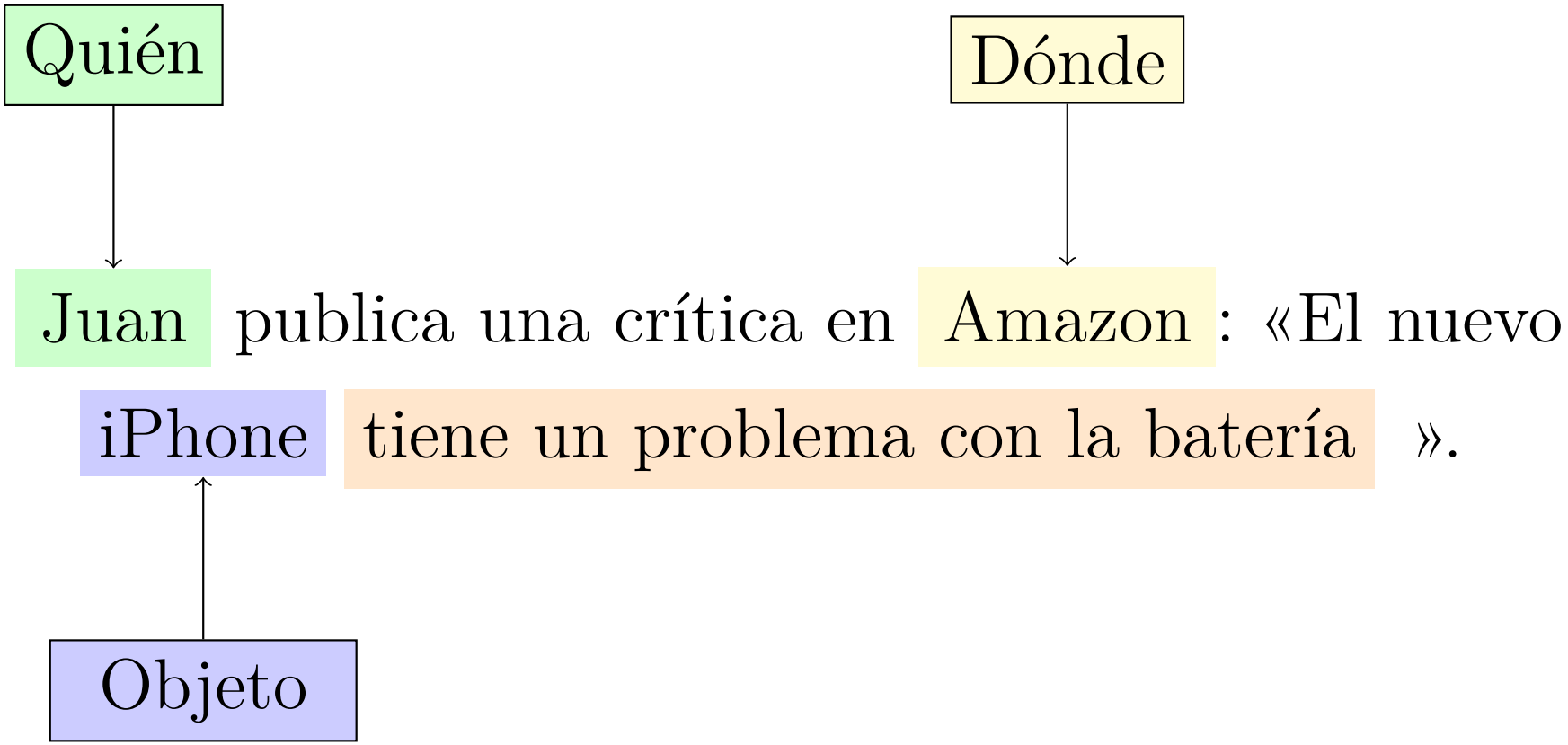


Juan publica una crítica en Amazon : «El nuevo iPhone tiene un problema con la batería ».

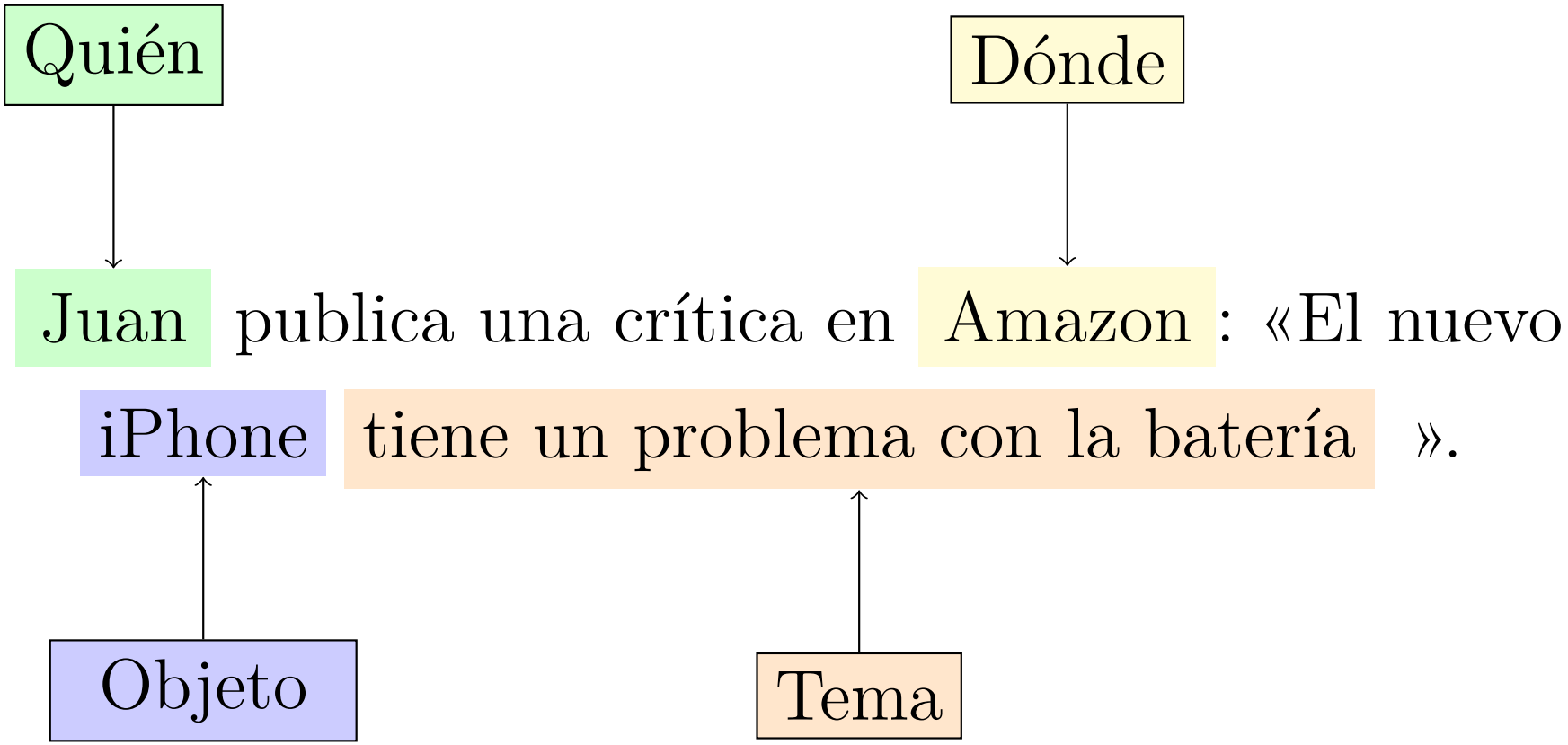
Componentes de una Opinión



Componentes de una Opinión



Componentes de una Opinión



Análisis de Sentimientos

- En el análisis de sentimientos se consideran dos clases de sentimientos: positivo (1) y negativo (-1).
- Es posible añadir una clase adicional, la neutra (0).
- Es común agregar una clase adicional que engloba todo el contenido basura.
- Agregar cierto grado de negatividad o positividad: -3,-2,-1,0,1,2,3.

Ejemplos en Twitter

-“El COVID19 no existe” — La vecina -“La vecina ya no existe” — El COVID19

Gracias a Dios he superado el covid, estoy listo para regresar. 🙏💛💙 #ClubAmerica 🦅 https://t.co/jsmHqwCuxV

Ahora que estamos en cuarentena por el #COVID19 ha aumentado el interés por tener plantas y cuidarlas. Lamentablemente no todos tenemos jardines, así que les voy a recomendar especies fáciles de cuidar y con las que puedes hacer una selva dentro de tu casa.Abro hilo. https://t.co/Q0eedpvhKM

Ejemplos en Twitter

Ejercicio

Busquen una publicación en alguna red social digital para exponerla a la clase. ¿Cuál es el sentimiento plasmado en ella? ¿Existe más de un sentimiento?

Análisis de Emociones

- El Análisis de Emociones [4] se enfoca en clasificar las distintas emociones expresadas en documentos de texto, ya sea párrafos, enunciados o documentos completos.

Análisis de Emociones

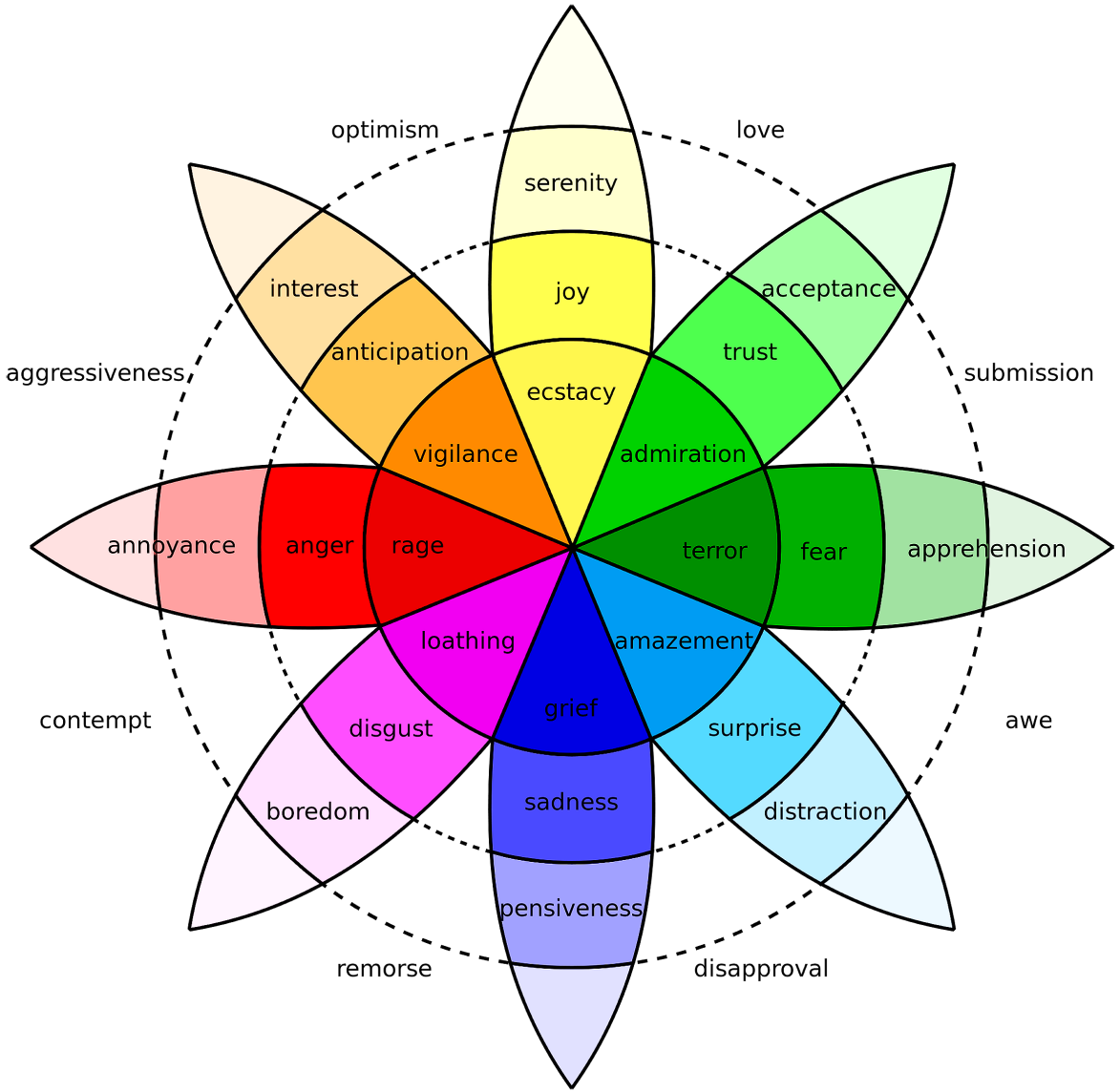


Figura: Rueda de emociones de Plutchik. [5]

Análisis de Emociones

Por otro lado, Ekman et al. [6] identifican seis emociones básicas:

- Ira (Anger)
- Disgusto (Disgust)
- Miedo (Fear)
- Felicidad (Happiness)
- Tristeza (Sadness)
- Sorpresa (Surprise)

Sentimiento vs Emoción

Pregunta

¿Cuál es la diferencia entre Análisis de Sentimiento y Análisis de Emoción?

Sentimiento vs Emoción

Pregunta

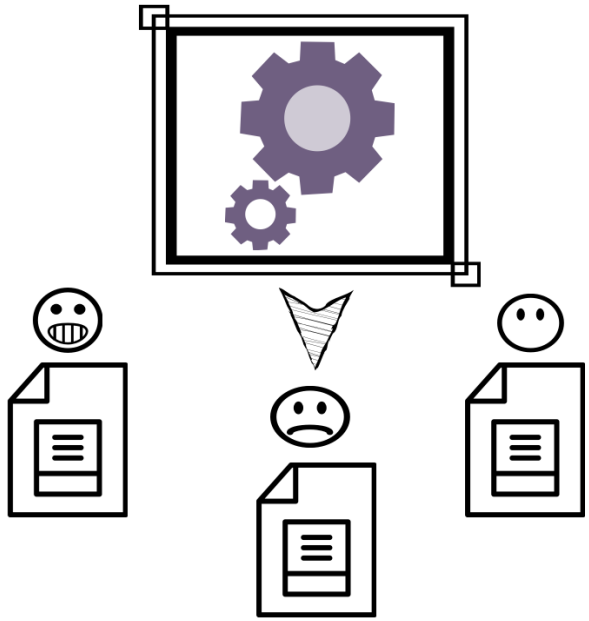
¿Cuál es la diferencia entre Análisis de Sentimiento y Análisis de Emoción?

Respuesta

- El sentimiento solo considera (usualmente) hasta tres clases.
- La emoción, aunque más difícil de modelar, se suelen considerar seis clases.

Introducción

Nuestro interés radica en extraer el sentimiento reflejado en la opinión.



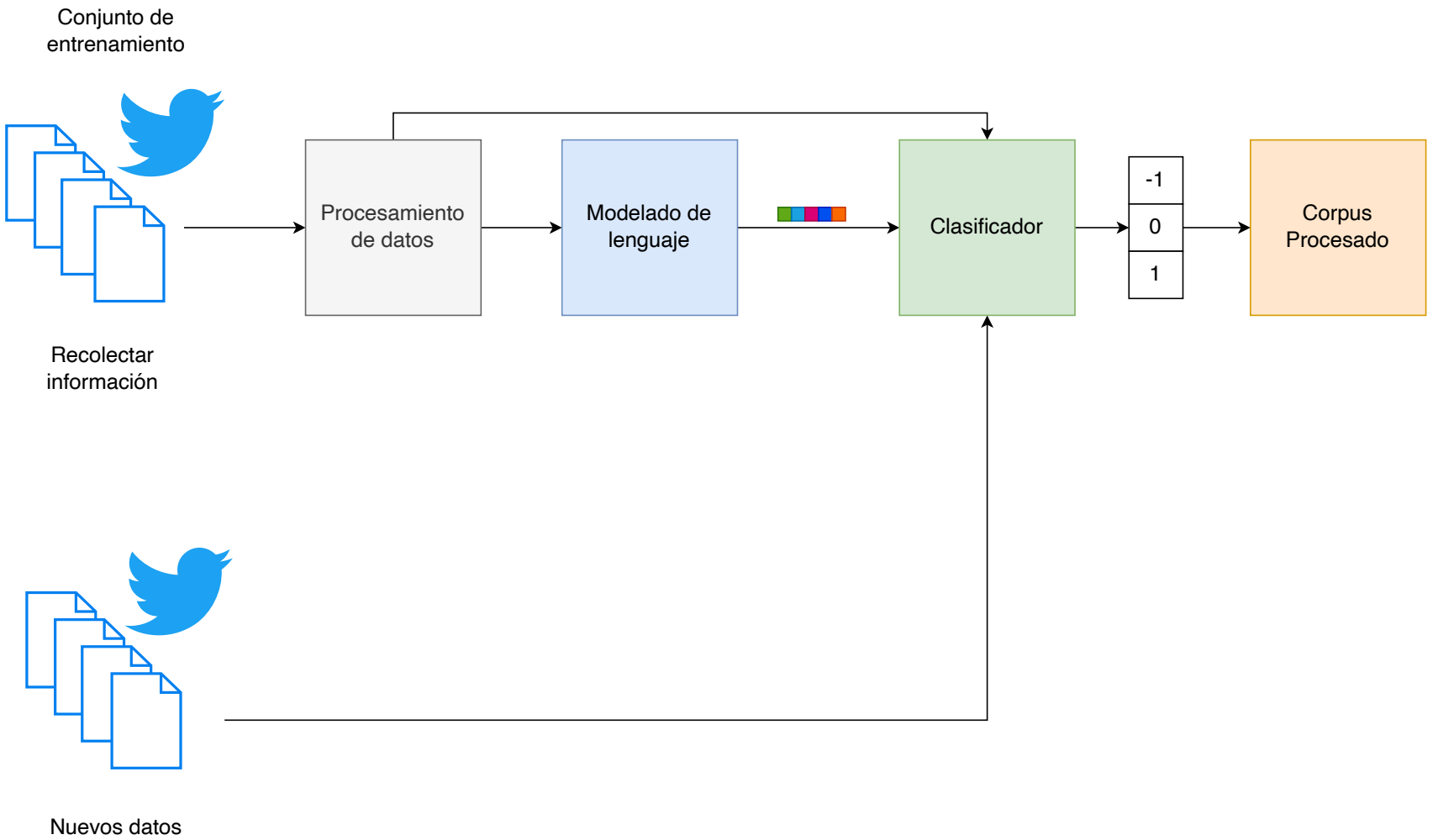
Objetivos

O1. Construir un método para detectar mensajes en Twitter adecuados para la investigación.

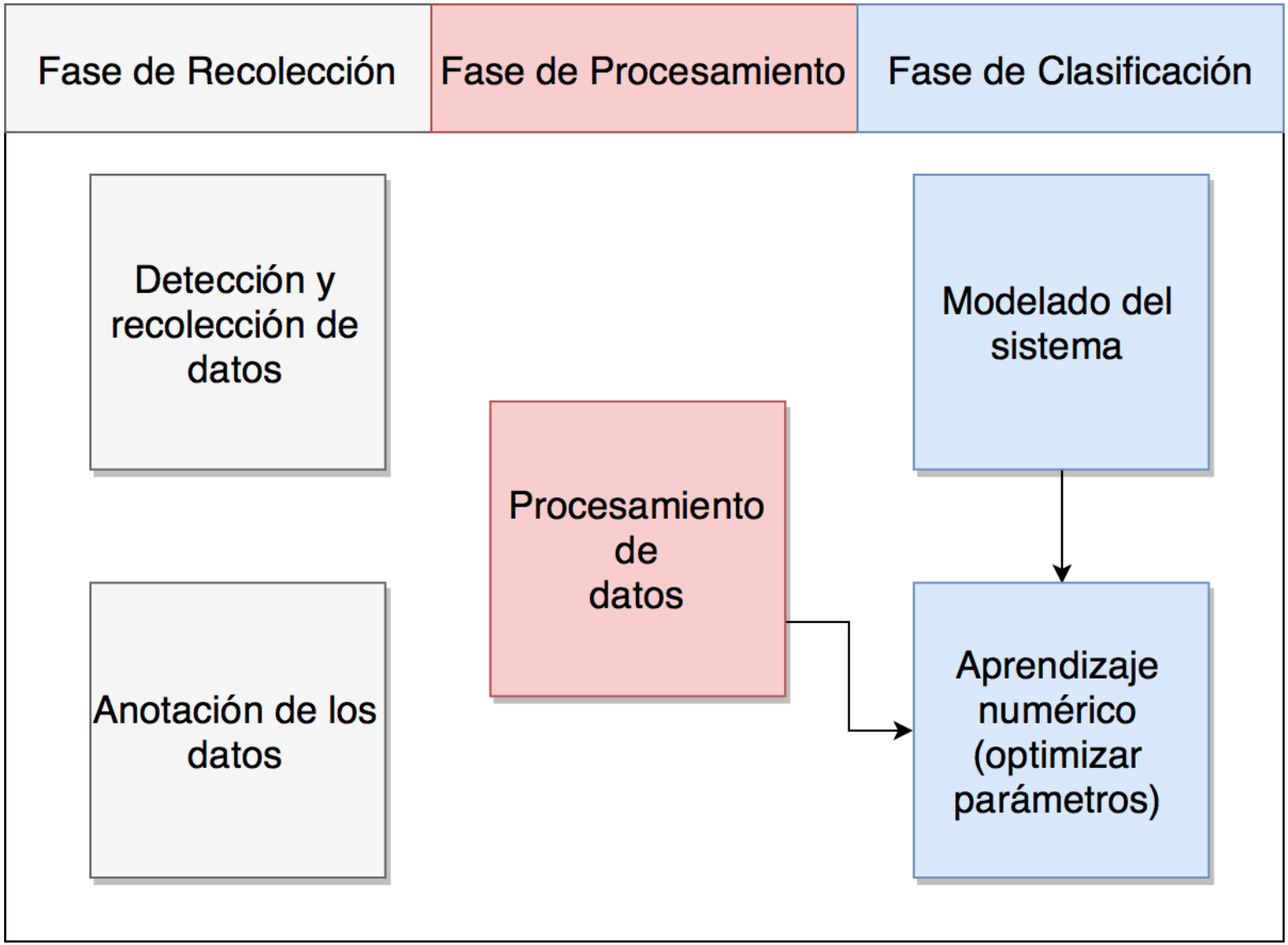
O2. Automatizar la clasificación de los mensajes considerando al sentimiento plasmado en ellos.

O3. Analizar la información obtenida y sacar conclusiones.

Idea General - Análisis de Sentimientos



Componentes del Análisis de Sentimiento



Recolección de Información

Pregunta

¿Cómo pueden definir su conjunto de datos?

Recolección de Información

Pregunta

¿Cómo pueden definir su conjunto de datos?

Respuesta

¡Haciendo una consulta (query) por medio de la API!

Recolección de Información

Pregunta

¿Cómo pueden definir su conjunto de datos?

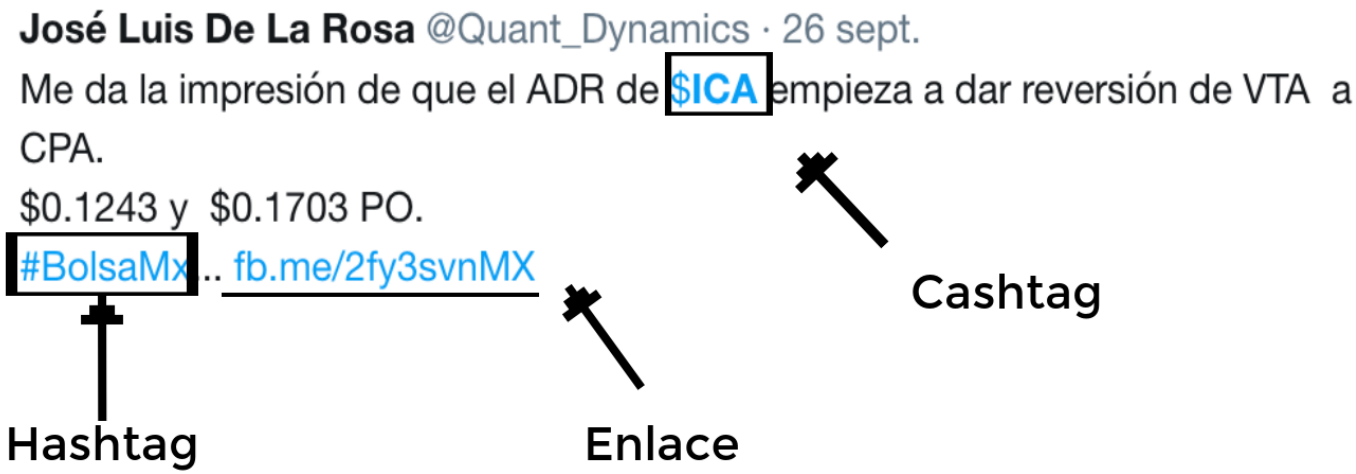
Respuesta

¡Haciendo una consulta (query) por medio de la API!

Veamos la documentación de Twitter sobre cómo construir consultas.

Recolección de Información

Es buena idea hacer un estudio preliminar, i.e., explorar qué información se puede encontrar. Para identificar tópicos de interés se puede usar el cashtag[1] y el hashtag para detectar mensajes útiles.



[1] El cashtag es la combinación del símbolo '$' y la etiqueta de una empresa listada en algún mercado accionario.

Recolección de Información

Por ejemplo, para cada tweet se obtienen los siguientes campos:

- `id_str` : El identificador único del tweet.
- `text` : El texto o contenido del tweet.
- `user.screen_name` : el objeto user tiene varios atributos, donde screen_name es el nombre en pantalla del usuario que publica el tweet.
- `created_at` : la fecha en que se publica el tweet utilizando el meridiano de Greenwich como huso horario estándar para todas las fechas.

Recolección de Información

Una vez recolectada la información se procede a anotar manualmente la clase de cada tweet de la siguiente manera:

- Positivo: El tweet en cuestión refleja un sentimiento positivo sobre el tema de estudio. Se representa mediante el número 1.
- Neutral: El tweet no presenta sentimiento alguno. Se representa mediante el número 0.
- Negativo: El tweet en cuestión refleja un sentimiento negativo hacia el tema de estudio. Se representa mediante el número -1.

Recolección de Información

¿Cómo debe guardarse el conjunto de datos (los tweets)?

Recolección de Información

¿Cómo debe guardarse el conjunto de datos (los tweets)?

Tienen dos opciones:

- Guardar los datos en un archivo csv.
- Guardarlo en una base de datos.

Recolección de Información

Usualmente se filtran publicaciones con las siguientes características:

- los retweets, o re-publicar un mensaje hecho por otro usuario, representado como RT
- tweets vacíos o sin texto
- tweets en idiomas ajenos al idioma principal del estudio
- tweets fuera de una región geográfica en particular
- tweets cuyo contenido no tenga sentido

Procesamiento de Datos

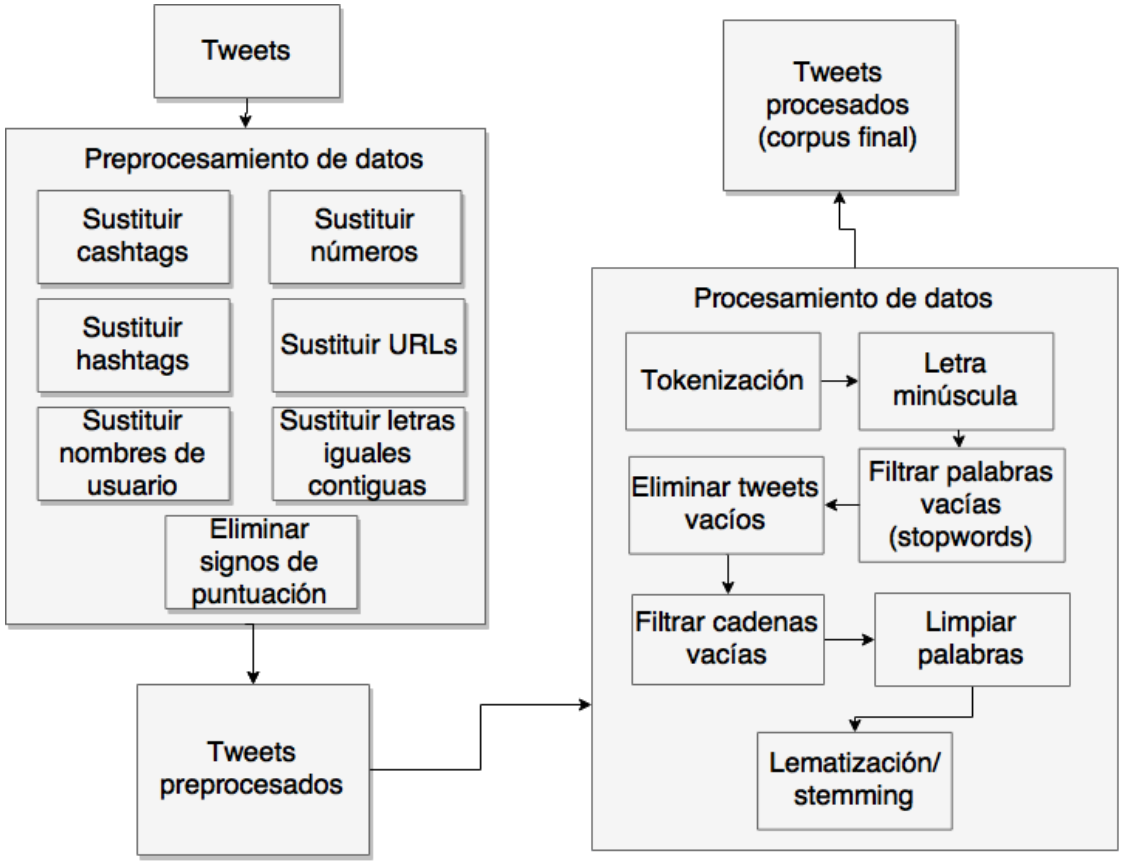


Figura: Módulo de Procesamiento

Sustituir cashtags realiza una sustitución de cashtags por la etiqueta ctag.

Ejemplo (sustitución de cashtags)

#Amediasesion el $IPC retrocede 0.45 %. Avanzan $LALA $AC $ELEKTR Bajan $OHLMEX $VOLAR $WALMEX ¿RSI sobre compra o s? https://t.co/CtY94rqGUc

↓

#Amediasesion el ctag retrocede 0.45 %. Avanzan ctag ctag ctag Bajan ctag ctag ctag ¿RSI sobre compra o s? https://t.co/CtY94rqGUc

Procesamiento de Datos

Sustituir números reemplaza números junto con símbolos asociados a ellos por la etiqueta xyz.

Ejemplo (sustitución de números)

#Amediasesion el $IPC retrocede 0.45 %. Avanzan $LALA $AC $ELEKTR Bajan $OHLMEX $VOLAR $WALMEX ¿RSI sobre compra o s? https://t.co/CtY94rqGUc

↓

#Amediasesion el ctag retrocede xyz. Avanzan ctag ctag ctag Bajan ctag ctag ctag ¿RSI sobre compra o s? https://t.co/CtY94rqGUc

Sustituir hashtags reemplaza hashtags[2] por la etiqueta htag.

Ejemplo (sustitución de hashtags)

#Amediasesion el $IPC retrocede 0.45 %. Avanzan $LALA $AC $ELEKTR Bajan $OHLMEX $VOLAR $WALMEX ¿RSI sobre compra o s? https://t.co/CtY94rqGUc

↓

htag el ctag retrocede xyz. Avanzan ctag ctag ctag Bajan ctag ctag ctag ¿RSI sobre compra o s? https://t.co/CtY94rqGUc

[2] Identificador formado por el símbolo '#' seguido de cualquier cadena de texto

Sustituir URLs reemplaza los enlaces presentes en los tweets a otras páginas de internet por la etiqueta url.

Ejemplo (sustitución de URLs)

#Amediasesion el $IPC retrocede 0.45 %. Avanzan $LALA $AC $ELEKTR Bajan $OHLMEX $VOLAR $WALMEX ¿RSI sobre compra o s? https://t.co/CtY94rqGUc

↓

htag el ctag retrocede xyz. Avanzan ctag ctag ctag Bajan ctag ctag ctag ¿RSI sobre compra o s? url

Por razones de privacidad, se sustituyen las menciones de usuario, cuya estructura es '@' junto con el nombre del usuario, por la etiqueta user mediante la función Sustituir nombre.

Ejemplo (sustitución de nombre de usuario)

@El_Nando84 Coincido, yo creo que viene flojo, de $Nemak afecta la baja en venta de autos, $Alpek el precio de petr...

↓

user Coincido, yo creo que viene flojo, de $Nemak afecta la baja en venta de autos, $Alpek el precio de petr...

La función Sustituir letras repetidas se eliminan repeticiones de letras contiguas en las palabras.

Ejemplo (sustitución de letras repetidas contiguas)

Gooool, Reemplazar

↓

Gool, Reemplazar

Procesamiento de Datos

La función Eliminar signos de puntuación excluye los siguientes signos de puntuación del texto del documento:

¿¡”#%’()*+,-$/=@&[]_ˆ{}|...

Ejemplo

@El_Nando84 Coincido, yo creo que viene flojo, de $Nemak afecta la baja en venta de autos, $Alpek el precio de petr...

↓

@El_Nando84 Coincido, yo creo que viene flojo, de $Nemak afecta la baja en venta de autos, $Alpek el precio de petr

Procesamiento de Datos

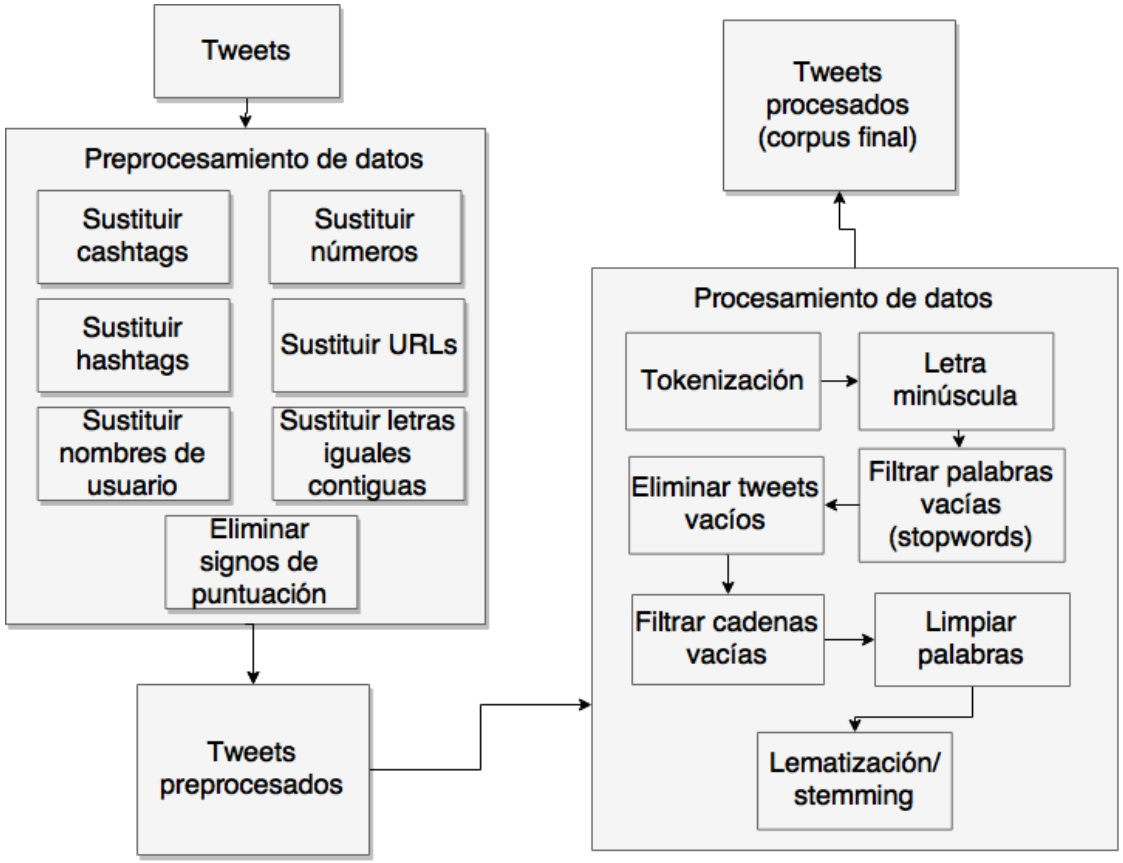


Figura: Módulo de Procesamiento

Procesamiento de datos

La tarea de tokenización separa y aísla cada palabra en el documento. Para ésto, se utiliza la librería Natural Languaje Toolkit (NLTK) de Python.

Ejemplo

htag el ctag retrocede xyz. Avanzan ctag ctag ctag Bajan ctag ctag ctag RSI sobre compra o s? url

↓

[‘htag’, ‘el’, ‘ctag’, ‘retrocede’, ‘xyz’, ‘.’, ‘Avanzan’, ‘ctag’, ‘ctag’, ‘ctag’, ‘Bajan’, ‘ctag’, ‘ctag’, ‘ctag’, ‘RSI’, ‘sobre’, ‘compra’, ‘o’, ‘s’, ‘?’, ‘url’]

Después se covierte todas las palabras a letra minúscula.

Ejemplo (letra minúscula)

[‘htag’, ‘el’, ‘ctag’, ‘retrocede’, ‘xyz’, ‘:’, ‘Avanzan’, ‘ctag’, ‘ctag’, ‘ctag’, ‘Bajan’, ‘ctag’, ‘ctag’, ‘ctag’, ‘RSI’, ‘sobre’, ‘compra’, ‘o’, ‘s’, ‘?’, ‘url’]

↓

[‘htag’, ‘el’, ‘ctag’, ‘retrocede’, ‘xyz’, ‘:’, ‘avanzan’, ‘ctag’, ‘ctag’, ‘ctag’, ‘bajan’, ‘ctag’, ‘ctag’, ‘ctag’, ‘rsi’, ‘sobre’, ‘compra’, ‘o’, ‘s’, ‘?’, ‘url’]

Posteriormente se filtran las palabras vacías, palabras sin significado como artículos, pronombres, preposiciones, etc[3]. Además, en éste paso se eliminan todas las etiquetas generadas en el preprocesamiento.

Ejemplo (palabras vacías)

[‘ctag’, ‘el’, ‘ctag’, ‘retrocede’, ‘xyz’, ‘.’, ‘avanzan’, ‘ctag’, ‘ctag’, ‘ctag’, ‘bajan’, ‘ctag’, ‘ctag’, ‘ctag’, ‘rsi’, ‘sobre’, ‘compra’, ‘o’, ‘s’, ‘?’, ‘url’]

↓

[‘retrocede’, ‘.’, ‘avanzan’, ‘bajan’, ‘rsi’, ‘sobre’, ‘compra’, ‘s’, ‘?’]

[3]Una lista de palabras vacías se puede encontrar en la siguiente liga: http://snowball.tartarus.org/algorithms/spanish/stop.txt

El proceso de limpar palabras busca eliminar las etiquetas de sutitución generadas en el preprocesamiento 'pegadas' a las palabras.

Ejemplo (limpiar palabras)

ctagexpandir, ejercicioctag

↓

expandir, ejercicio

El proceso de eliminar cadenas vacías busca filtrar dichos elementos representados en Python como ‘’, es decir, una cadena vacía.

La última tarea es la reducción de las palabras ya sea a su lema o a su raíz. La lematización consiste en reducir una palabra flexionada a su forma sin flexionar:

Ejemplo (lematización)

trabajamos → trabajar
trabajo → trabajo

El stemming busca reducir una palabra a su raíz.

Ejemplo (stemming)

trabajo, trabajamos, trabajar → trabaj

Procesamiento de Datos

Algunas tareas adicionales que pueden ser necesarias (heurísticas computacionales):

- Corrección de errores ortográficos.
- Expandir abreviaturas y contracciones.
- Considerar negaciones.
- Etiquetado gramatical para realizar Lematización.

Construcción del Clasificador

Recordemos qué es Machine Learning...

Construcción del Clasificador

Pregunta

¿Por qué debemos estudiar Machine Learning?

Construcción del Clasificador

Pregunta

¿Por qué debemos estudiar Machine Learning?

Respuesta

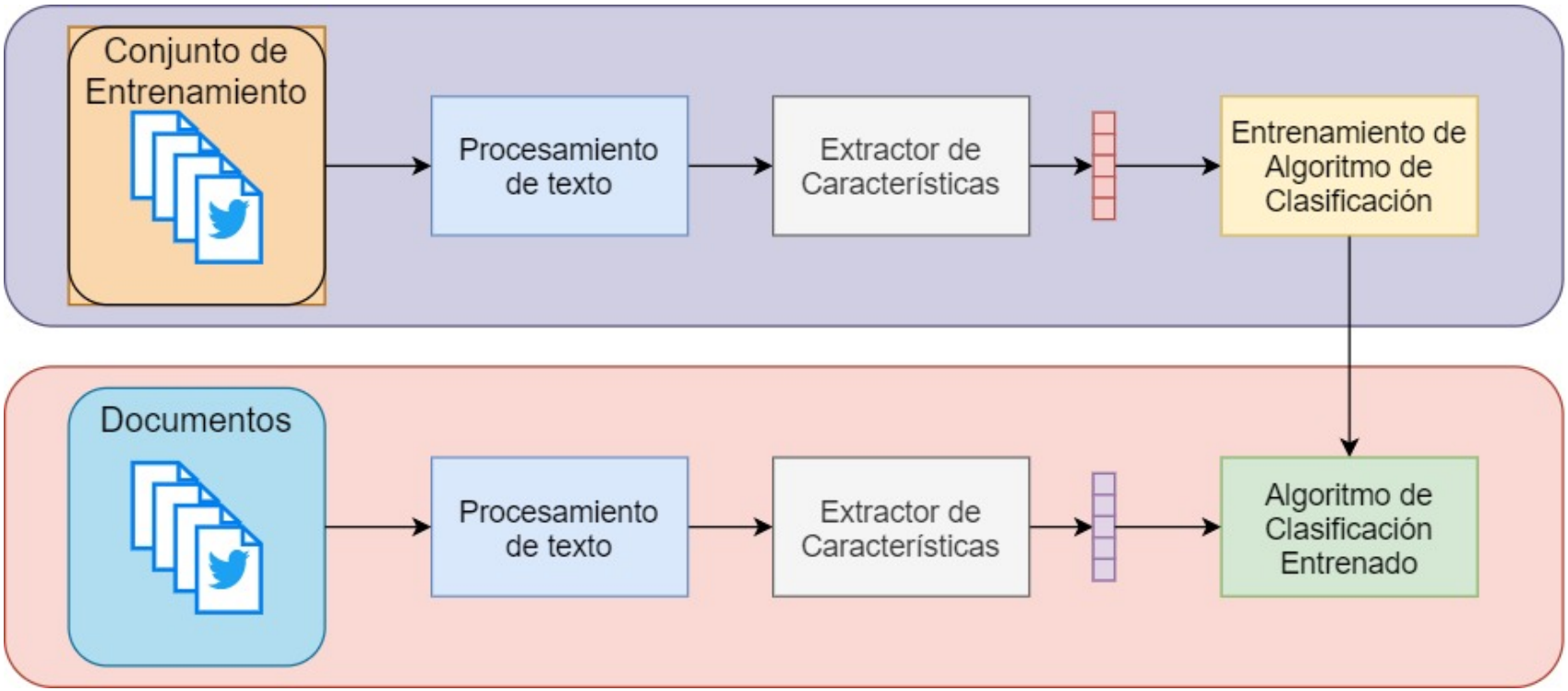
El Machine Learning, en la tarea de Análisis de Sentimientos, nos permite clasificar de forma automática los mensajes en una de tres (o dos) clases posibles. ¡Muy conveniente!

Construcción del Clasificador

Para la construcción del clasificador de sentimientos de los tweets se utilizan técnincas de Aprendizaje Numérico o Machine Learning. Proponemos utilizar el siguiente modelo de aprendizaje:

- Máquinas de Vectores de Soporte

Diagrama del Proceso



Modelado de Lenguaje

Pregunta

¿Cómo se les ocurre transformar texto a números?

Modelado de Lenguaje

Pregunta

¿Cómo se les ocurre transformar texto a números?

Respuesta

¡Con modelado de lenguaje! Exploremos algunas opciones...

Construcción del Clasificador

Repasemos que llevamos hasta el momento:

1. Tweets por medio de la API.
2. Un método para procesar los datos.
3. Exploración del conjunto de datos.
4. Un subconjunto de tweets anotado manualmente.
5. Una forma de modelar lenguaje.
6. Una propuesta de algoritmo de clasificación para anotar masivamente los tweets.

Construcción del Clasificador

¿Qué sigue?

- Entrenar el clasificador con nuestro conjunto de entrenamiento...

Construcción del Clasificador

Referencias I

[1] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. Image and Vision Computing, 65:3–14, 2017.

[2] F. A. Pozzi, E. Fersini, E. Messina, and B. Liu. Challenges of sentiment analysis in social networks. In Sentiment Analysis in Social Networks. Elsevier, 2017.

[3] E. Cambria, A. Livingstone, and A. Hussain. The Hourglass of Emotions, pages 144–157. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[4] Lei Zhang and Bing Liu. Sentiment analysis and opinion mining. In Encyclopedia of Machine Learning and Data Mining. Springer Science+Business Media, 2016.

[5] Robert Plutchik. A psychoevolutionary theory of emotions. Social Science Information, 21(4-5):529–553, jul 1982.

Referencias II

[6] Paul Ekman, E. Richard Sorenson, and Wallace V. Friesen. Pancultural elements in facial displays of emotions. Science, 164:86–88, 1969.