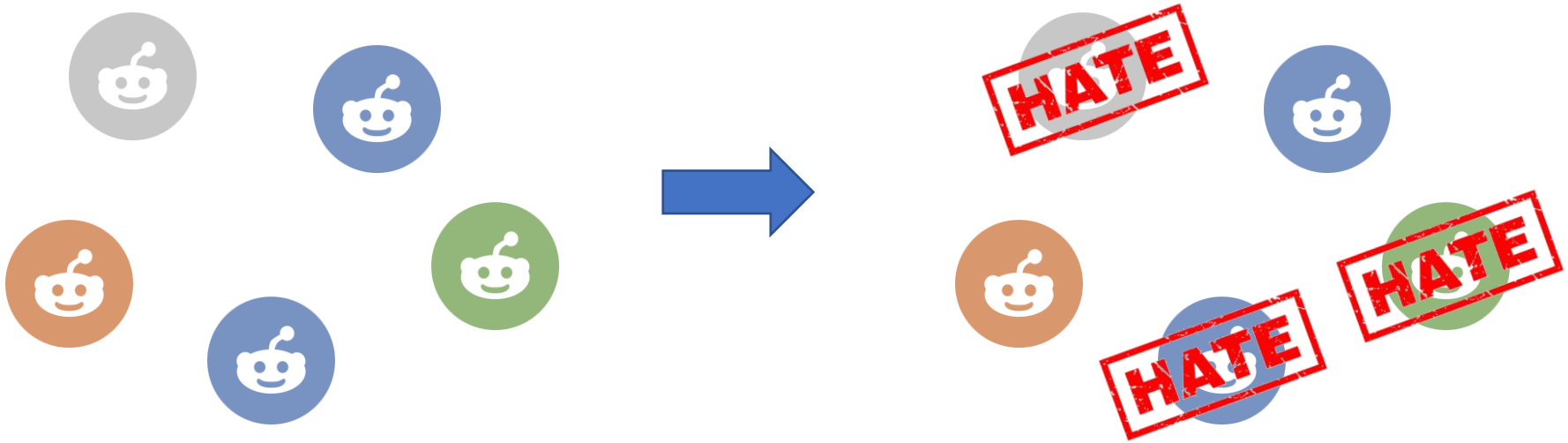
Procesamiento de Lenguaje Natural / TP

Detección de *hate speech* en medios sociales

Medios sociales & hate speech

Cuál es el objetivo?

Detectar aquellos comentarios de Reddit que contienen hate speech distinguiéndolos de aquellos que no!



Medios sociales & hate speech

Y los datos?

- **“A Benchmark Dataset for Learning to Intervene in Online Hate Speech”**
    - El objetivo del trabajo no fue detectar hate, sino diseñar estrategias de intervención con respuestas automáticas a conversaciones quqe tienen hate speech.

- Aproximadamente 5k conversaciones, con 22k comentarios.
    - Actualmente se encuentran disponibles alrededor de 4k con 18k comentarios.

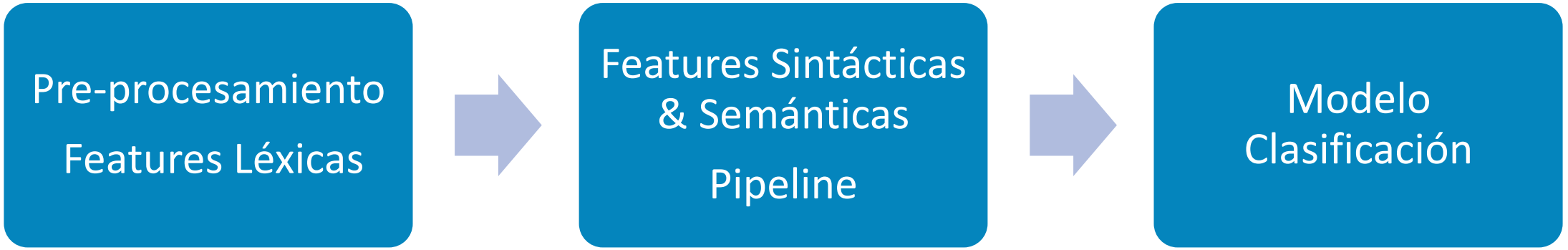
- Comentarios pertenecientes a 10 subreddits:
    - `r/DankMemes`
    - `r/Imgoingtohellforthis`
    - `r/KotakuInAction`
    - `r/MensRights`
    - `r/MetaCanada`
    - `r/MGTOW`
    - `r/PussyPass`
    - `r/PussyPassDenied`
    - `r/The Donald`
    - `r/TumblrInAction`

- Para cada subreddit, recolectarons los 200 posts más “hot”.
- Buscaron posts con keywords de hate.
- Recolectaron las conversaciones.
- No hay repetidos.

https://github.com/jing-qian/A-Benchmark-Dataset-for-Learning-to-Intervene-in-Online-Hate-Speech

Medios Sociales & Desinformación

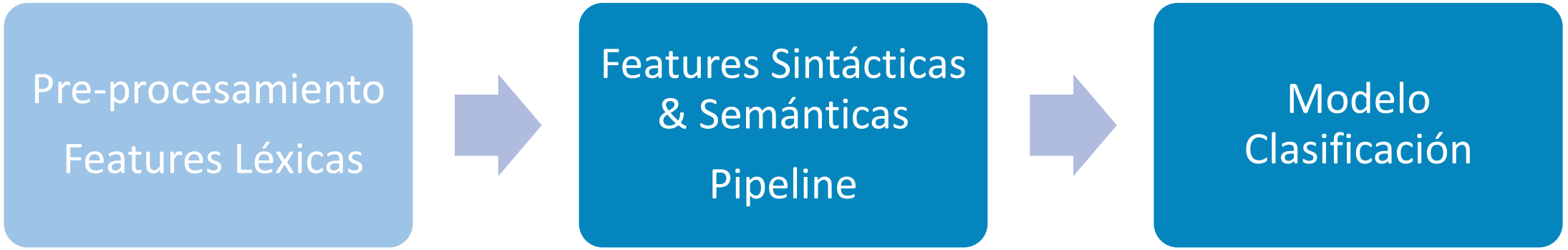
Los prácticos!



Semana 1 | Semana 2 | Semana 3

Medios Sociales & Desinformación

Qué hicieron hasta ahora?

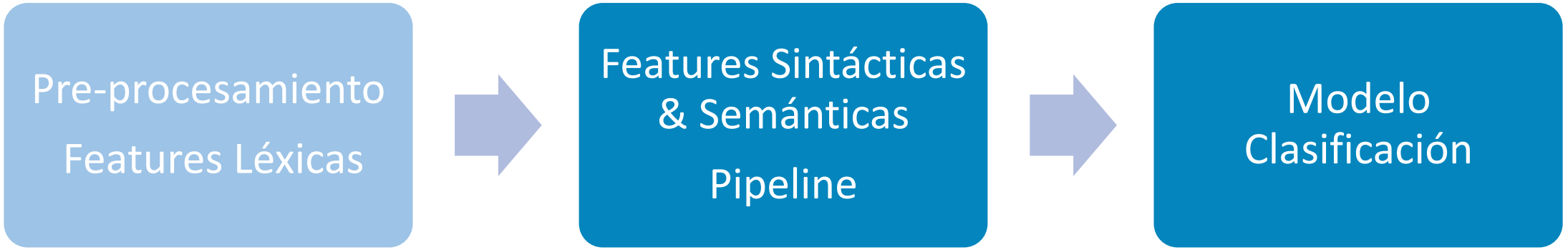


Con este práctico deberían haber:

- Procesado el json con los comentarios y almacenarlos en alguna estructura.
- Decidido si considerar el hilo de conversaciones.
- Elegido algunas características para representar los comentarios.
- Aplicado pasos de pre-procesamiento sobre el texto.
- Pensado en alguna estrategia para representar los comentarios (opcional).
- Calculado estadísticas sobre los comentarios (por ejemplo, palabras más frecuentes).

Medios Sociales & Desinformación

Qué hicieron hasta ahora?



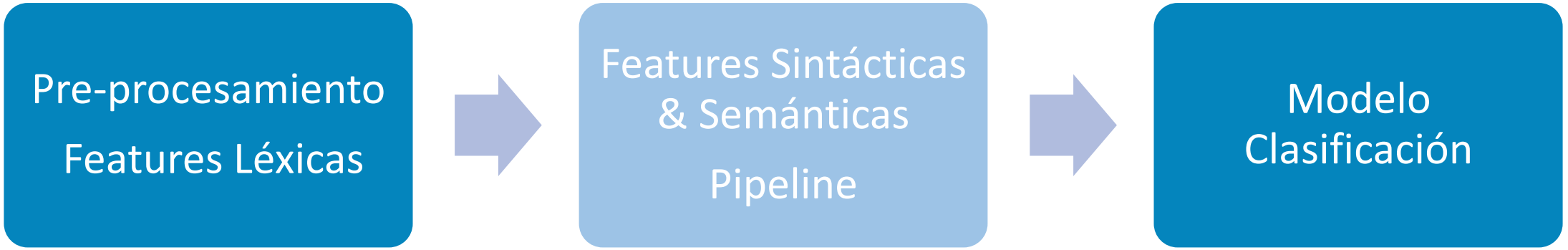
Con este práctico deberían haber:

- Procesado el json con los comentarios y almacenarlos en alguna estructura.
- Decidido si considerar el hilo de conversaciones.
- Elegido algunas características para representar los comentarios.
- Aplicado pasos de pre-procesamiento sobre el texto.
- Pensado en alguna estrategia para representar los comentarios (opcional).
- Calculado estadísticas sobre los comentarios (por ejemplo, palabras más frecuentes).

Recuerden, tienen tiempo hasta hoy para entregarlo!

Medios Sociales & Desinformación

Qué tienen que hacer?



- De todos los análisis que vimos hasta ahora, elegir al menos **dos** características nuevas para incorporar a los comentarios.
  - Por ejemplo, agregar el sentimiento, la emoción o etiquetas POS.
  - La selección de estas características debe quedar integrada con el procesamiento que hicieron en el TP 1.
- Definir la representación de los comentarios a utilizar.
  - Recordar que el objetivo final es entrenar un modelo de clasificación, con lo que la representación tiene que ser “amigable” con el posterior proceso de entrenamiento y test.
- Integración del procesamiento completo.
  - Desde la carga del dataset hasta la creación de la representación.

Medios Sociales & Desinformación

Qué tienen que entregar?

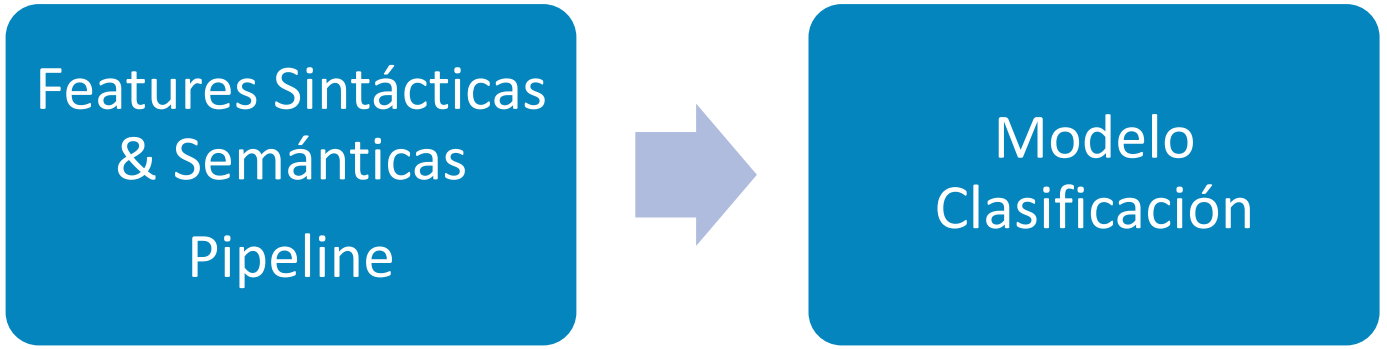
Features Sintácticas & Semánticas

Pipeline

- Notebook con:
  - Carga de dataset.
  - Selección de atributos. Mencionar brevemente por qué eligieron cada uno de los nuevos que hayan agregado.
  - Definición de la representación elegida. Explicar brevemente por qué la eligieron.
  - Integración del procesamiento completo.
    - Implementado como un Transformer de sklearn.
    - Implementado como un método a invocar que incluya el procesamiento.
    - Recordar que la estructura final debe ser amigable con la requerida para el entrenamiento del modelo de clasificación.

Medios Sociales & Desinformación

Qué tienen que entregar?



NO se entrega, se integra con el TP 3 :)

Procesamiento de Lenguaje Natural / TP

Detección de *hate speech* en medios sociales