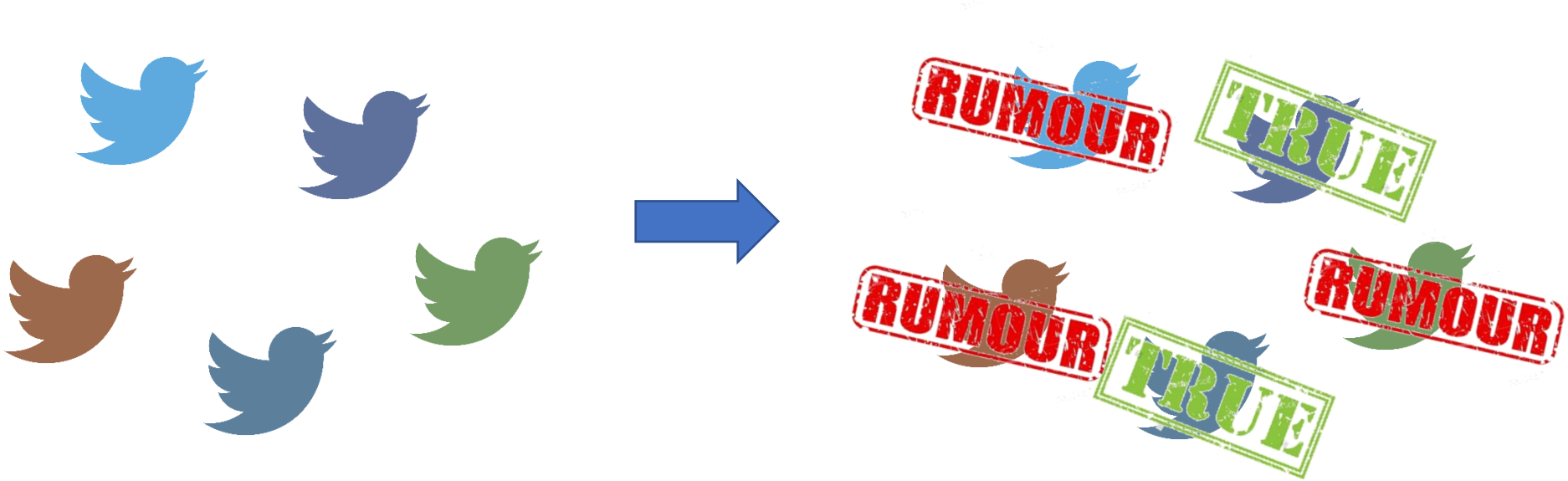
Procesamiento de Lenguaje Natural / TP

Detección de rumores en medios sociales

Medios Sociales & Desinformación

Cuál es el objetivo?

Detectar aquellos tweets que contienen información que todavía no ha sido verificada (los rumores), distinguiéndola de lo que no son rumores!



Medios Sociales & Desinformación

Y los datos?

- **Sydney siege.**
  - Un hombre armado tomó como rehenes a 10 clientes y 8 empleados en un local de Lindt en Sydney, Australia, el 15 de Diciembre de 2014.

Evento	Rumours	Non-rumours	Total
Sydney Siege	522 (42.8 %)	699 (57.2 %)	1.221

- Formato JSON → JavaScript Object Notation

- Por cada Tweet, se tiene:
  - Información completa.
  - Información completa de las reacciones.
  - Clase (Rumor o no rumor?)

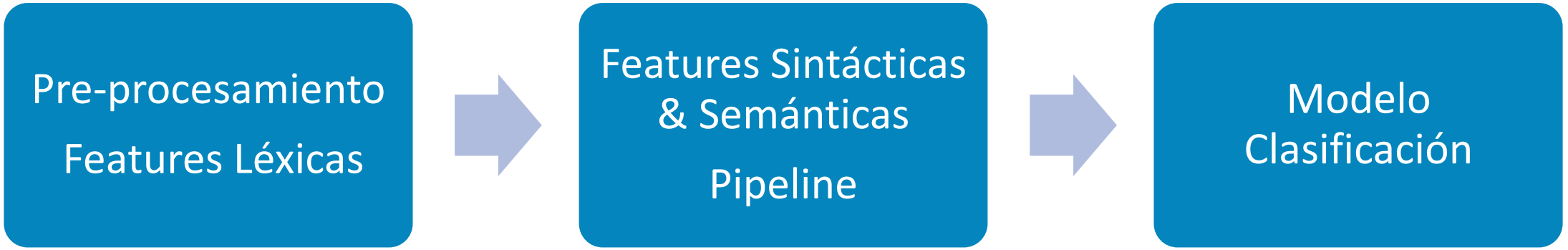
```
tweets/
├── rumours/
│   ├── tweet.json
│   ├── tweet1.json
│   ├── ...
│   ├── tweetn.json
│   ├── tweet_reacciones.json
│   ├── ...
│   └── tweetn_reacciones.json
└── non-rumours/
    ├── tweet.json
    ├── tweet1.json
    ├── ...
    ├── tweetn.json
    ├── tweet_reacciones.json
    ├── ...
    └── tweetn_reacciones.json
```

```
tweets/
├── tweets_class.csv
├── tweet.json
├── tweet1.json
├── ...
├── tweetn.json
├── tweet_reacciones.json
├── ...
└── tweetn_reacciones.json
```

```
tweets/
├── tweets.json
├── tweet1_reacciones.json
├── ...
└── tweetn_reacciones.json
```

Medios Sociales & Desinformación

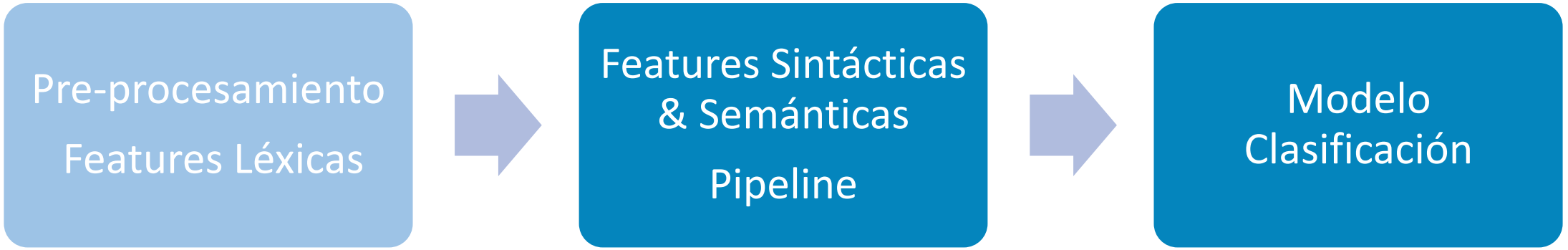
Los prácticos!



Semana 1 | Semana 2 | Semana 3

Medios Sociales & Desinformación

Qué hicieron hasta ahora?



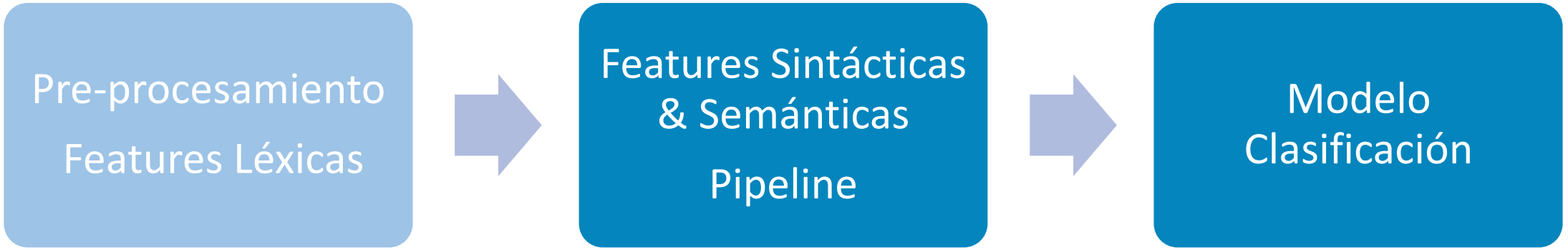
Semana 1 | Semana 2 | Semana 3

Con este práctico deberían haber:

- Procesado los json con los tweets y almacenarlos en alguna estructura.
- Decidido si considerar o no las reacciones.
- Elegido algunas características para representar los tweets.
- Aplicado pasos de pre-procesamiento sobre el texto.
- Pensado en alguna estrategia para representar los tweets (opcional).
- Calculado estadísticas sobre los tweets (por ejemplo, palabras más frecuentes)

Medios Sociales & Desinformación

Qué hicieron hasta ahora?



Semana 1 | Semana 2 | Semana 3

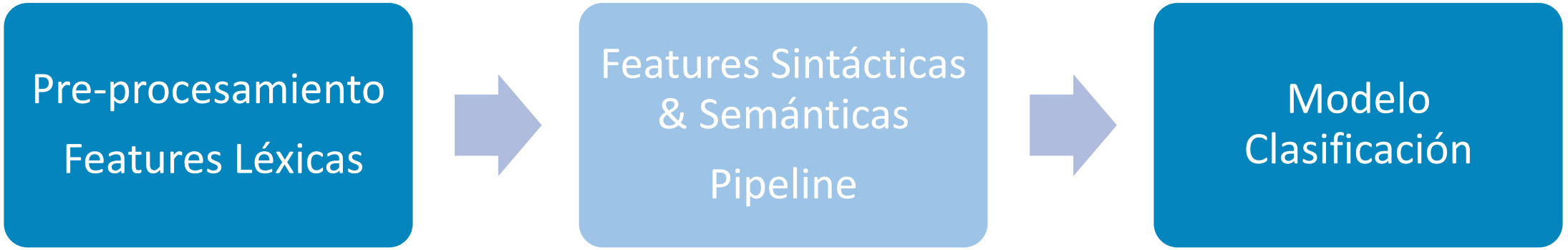
Con este práctico deberían haber:

- Procesado los json con los tweets y almacenarlos en alguna estructura.
- Decidido si considerar o no las reacciones.
- Elegido algunas características para representar los tweets.
- Aplicado pasos de pre-procesamiento sobre el texto.
- Pensado en alguna estrategia para representar los tweets (opcional).
- Calculado estadísticas sobre los tweets (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 tweets.
  - Por ejemplo, agregar el sentimiento de los tweets, la emoción o elegir solo utilizar un tipo de etiqueta POS.
  - La selección de estas características debe quedar integrada con el procesamiento que hicieron en el TP 1.
- Definir la representación de los tweets 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 de los tweets. Mencionar brevemente por qué eligieron cada uno de los nuevos que hayan agregado.
  - Definición de la representación elegida para los tweets. 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?

Features Sintácticas & Semánticas

Pipeline

Fecha de entrega: **15 de Agosto 2020**

- Notebook con:
  - Carga de dataset.
  - Selección de atributos de los tweets. Mencionar brevemente por qué eligieron cada uno de los nuevos que hayan agregado.
  - Definición de la representación elegida para los tweets. 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?

Features Sintácticas & Semánticas

Pipeline

Fecha de entrega: **15 de Agosto 2020**

La notebook debe poder ejecutarse sin errores y debe incluir los outputs generados!

- Notebook con:
  - Carga de dataset.
  - Selección de atributos de los tweets. Mencionar brevemente por qué eligieron cada uno de los nuevos que hayan agregado.
  - Definición de la representación elegida para los tweets. 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.

Procesamiento de Lenguaje Natural / TP

Detección de rumores en medios sociales