

# Análisis de asociaciones en tiempo real en Twitter

Pedro Jesús López Abenza

Universidad de Granada

Trabajo Final del Master  
Master en Ciencia de Datos e Ingeniería de Computadores



# Overview

## 1 Introducción y Objetivos

## 2 Descripción del análisis

- Recolección de mensajes de Twitter
- Análisis de Sentimientos
- Estructuración del Texto
- Association Stream Mining

## 3 Flujos de datos

- Comparación: USA vs. ESPAÑA
- Flujo de datos: España

## 4 Análisis de resultados

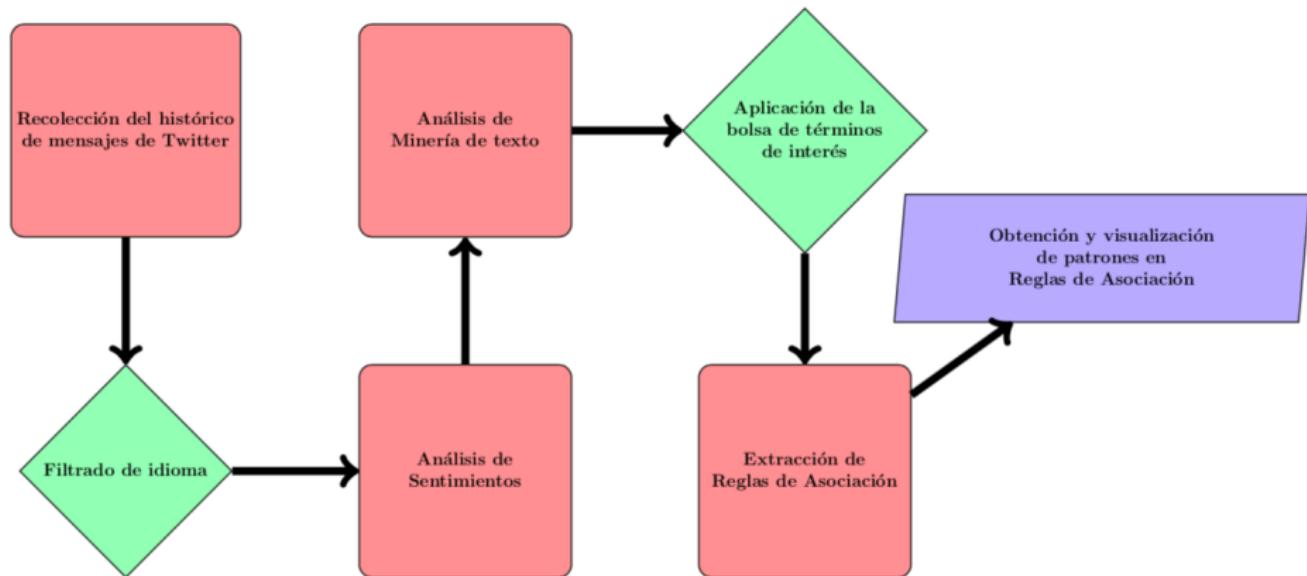
- Estudio del fenómeno de Concept drift
- Análisis del tiempo de cálculo

## 5 Conclusiones

# Introducción y Objetivos

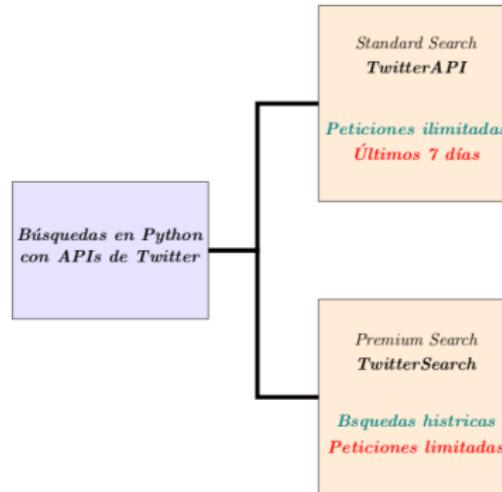
- Estudio del problema desde un contexto propio de Minería de flujo de datos.
- Analizar y estructurar la información contenida en mensajes de Twitter para poder extraer conocimiento de ellos.
- Búsqueda de patrones en las reglas de asociación extraídas a partir de estos *tweets*.

# Descripción del análisis



# Recolección de mensajes de Twitter

- Herramientas API de Twitter disponibles en *Python*.
- Dos enfoques, según la antigüedad de los datos.



# Análisis de Sentimientos

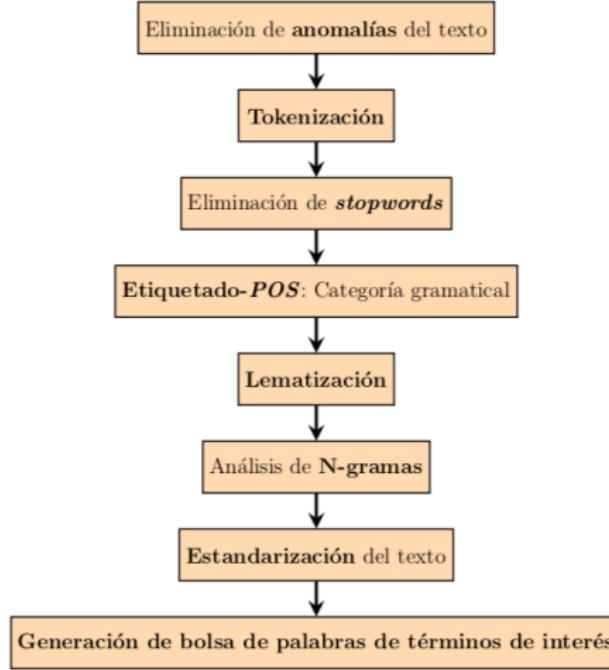
- El sentimiento de los mensajes extraídos de Twitter se presenta como atributo de interés.
- Dificultad adicional ante la brevedad que caracteriza a los *tweets*.
- Se requiere de un *software* externo para su cálculo.



- Utiliza varios *lexicons* de palabras, emoticonos y expresiones.
- Obtiene el sentimiento del texto en base a los valores de positividad/negatividad de las palabras que lo componen.
- Tres sentimientos posibles: Positivo, Negativo y Neutro.

# Estructuración de Texto

Análisis de **Minería de texto** realizado con *SpaCy*.



# Association Stream Mining

- Búsqueda de **reglas de asociación**. Dependencias entre términos: Cuando **X** aparece, **Y** también lo hace.

TÉRMINO → SENTIMIENTO

- Aplicación en un enfoque de **Minería de flujo de datos**:
  - Estudio sin conocer toda la base de datos.
  - Capaz de superar los retos propios de la MFD.
- Dos algoritmos utilizados:
  - *IncMine*
  - *Fuzzy-CSarAFP*
- Todas las reglas de asociación no son igual de válidas:  
**Filtrado** ⇒ Medida de calidad eficiente.

## Association Stream Mining: MOA-IncMine

- Extracción de *itemsets* cerrados frecuentes (FCIs)
- **Aprendizaje incremental** del flujo de datos.
- **Método aproximado**. Se introducen los semi-FCIs.
- Ventana deslizante de tamaño fijo sensible a transacciones.
- Método *batch* para la extracción de FCIs basado en CHARM
- Extracción de Reglas de asociación: Requiere de un análisis adicional *offline*.

## *Association Stream Mining: Fuzzy-CSarAFP*

- **Algoritmo genético** para la extracción de reglas de asociación difusas en MFD.
- Aprendizaje *online* del flujo de datos.
- Basado en un modelo metaheurístico del tipo *Michigan-LCSs*.
- Extracción **directa** de las reglas de asociación.
- Capaz de adaptar rápidamente el conocimiento adquirido en cambios de concepto.

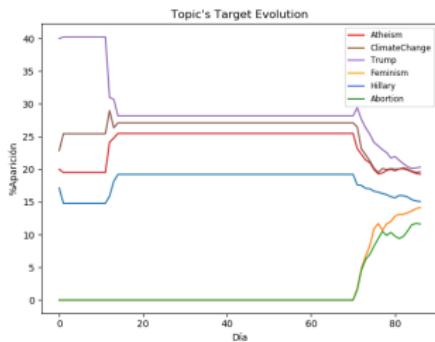
## Association Stream Mining: Filtrado

- **Problema** ⇒ Reglas interesantes pueden presentar valores bajos de soporte y/o confianza.
- En estudios de Twitter, es muy común encontrar soportes bajos. Se requiere otra medida de calidad válida.
- Filtrado de reglas en base a medida eficiente. **LIFT**: Cociente entre soporte de la regla y el producto de los soportes de los antecedentes.

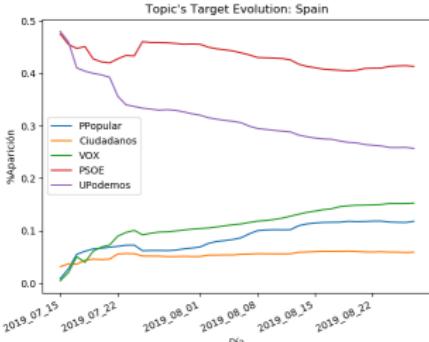


# Comparación de los flujos de datos

USA



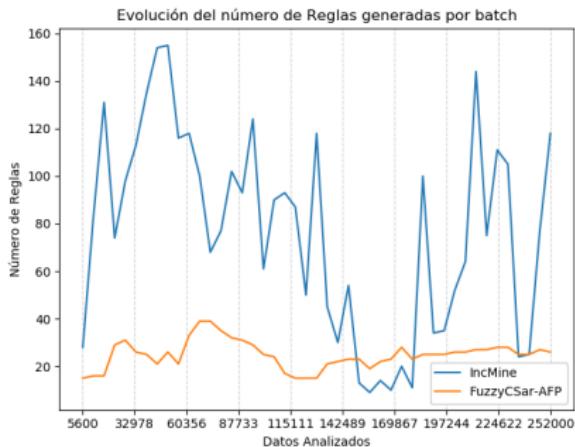
ESPAÑA



15 JULIO

**250147 TWEETS**

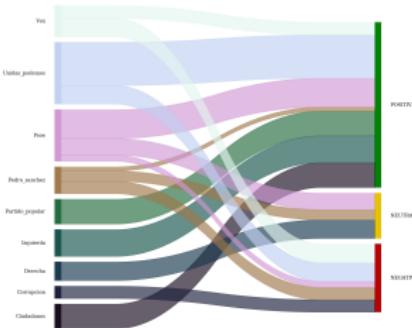
28 AGOSTO



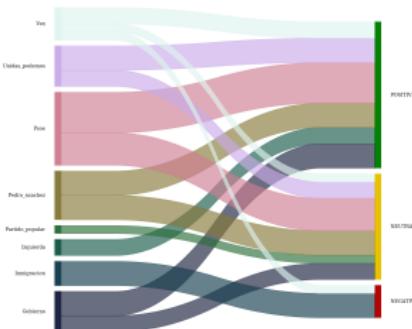
Estudio dinámico del comportamiento de las relaciones propuestas por las reglas generadas por los algoritmos.

# Concept drift: IncMine

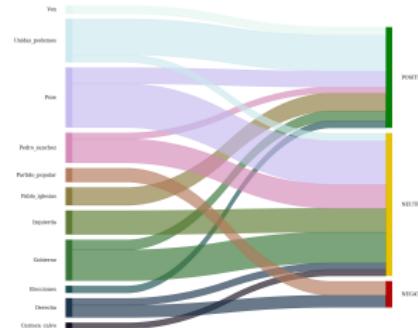
IncMine SPAIN: Sankey Diagram. 5600 datos analizados



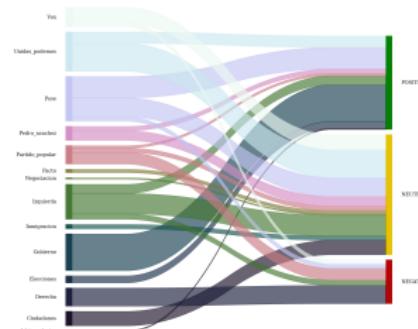
IncMine SPAIN: Sankey Diagram. 173600 datos analizados



IncMine SPAIN: Sankey Diagram. 89600 datos analizados

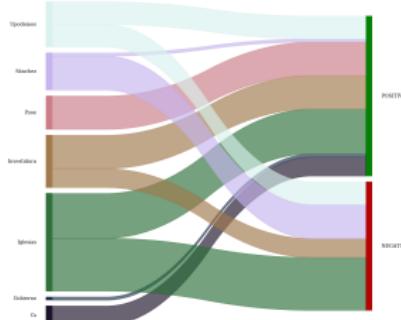


IncMine SPAIN: Sankey Diagram. 250147 datos analizados



## *Concept drift: Fuzzy-CSarAFP*

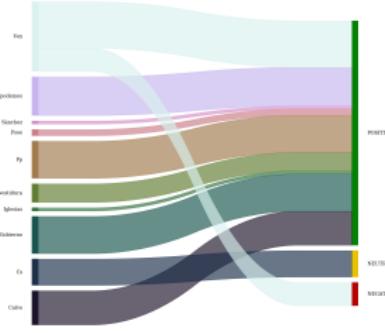
FuzzyCSar SPAIN: Sankey Diagram. 5600 datos analizados



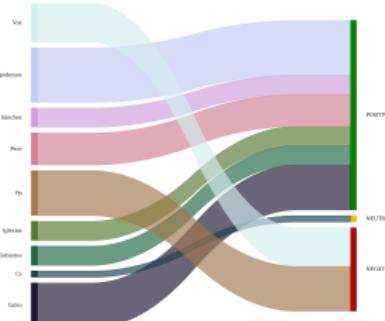
FuzzyCSar SPAIN: Sankey Diagram. 173600 datos analizados



FuzzyCSar SPAIN: Sankey Diagram. 89600 datos analizados

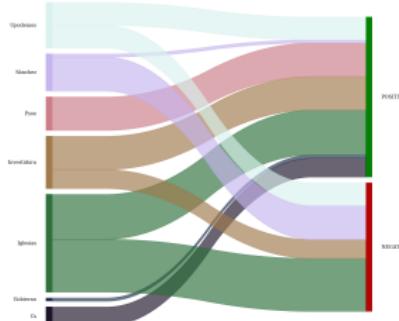


FuzzyCSar SPAIN: Sankey Diagram. 250147 datos analizados



# Concept drift: Fuzzy-CSarAFP

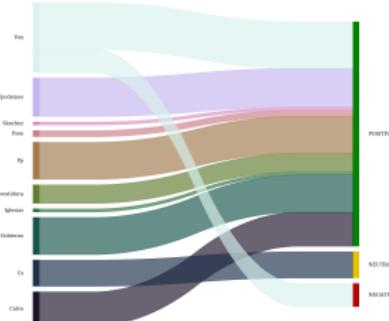
FuzzyCSar SPAIN: Sankey Diagram. 5600 datos analizados



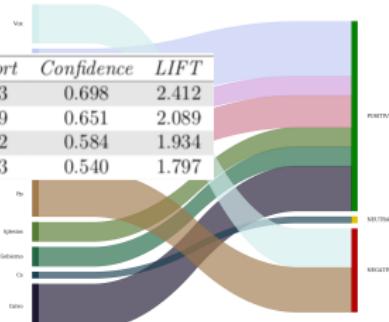
FuzzyCSar SPAIN: Sankey Diagram. 173600 datos analizados



FuzzyCSar SPAIN: Sankey Diagram. 89600 datos analizados



FuzzyCSar SPAIN: Sankey Diagram. 250147 datos analizados



# Análisis del tiempo de cálculo

| Tarea aplicada                            | Tiempo medio(segs) |
|---|--------------------|
| <b>Estructuración del texto</b>           |                    |
| Análisis de Sentimientos                  | 0.3395             |
| Comprobación de Idioma                    | 0.0250             |
| Análisis de Texto                         | 0.1899             |
| Estandarización del Texto                 | 0.0001             |
| <b>Extracción de Reglas de asociación</b> |                    |
| Extracción RAs ( <i>Fuzzy-CSarAFP</i> )   | 0.0005             |
| Extracción <i>FCIs</i> ( <i>IncMine</i> ) | 0.0001             |
| Extracción RAs ( <i>IncMine</i> )         | 0.0001             |
|   | ~ 0.55             |

- Estructuración del texto: Gran parte del tiempo.
- Extracción de reglas: Muy rápida.

# Conclusiones

- Estudio de un problema real de Minería de flujo de datos.
- Algoritmos de extracción de reglas con resultados algo distintos. Pueden ser complementarios.
- Presencias constantes y notables de fenómenos de *concept drift*.
- Análisis de tiempo de cálculo por dato correcto.

