Trabajo de fin de grado. Grado en Ingeniería Informática.

Análisis de Sentimientos

Caso practico analizando las opiniones de los usuarios de museos

Autor: Jesús Sánchez de Castro

Tutores:
Victoria María Luzón García
Salvador García López

Índice:

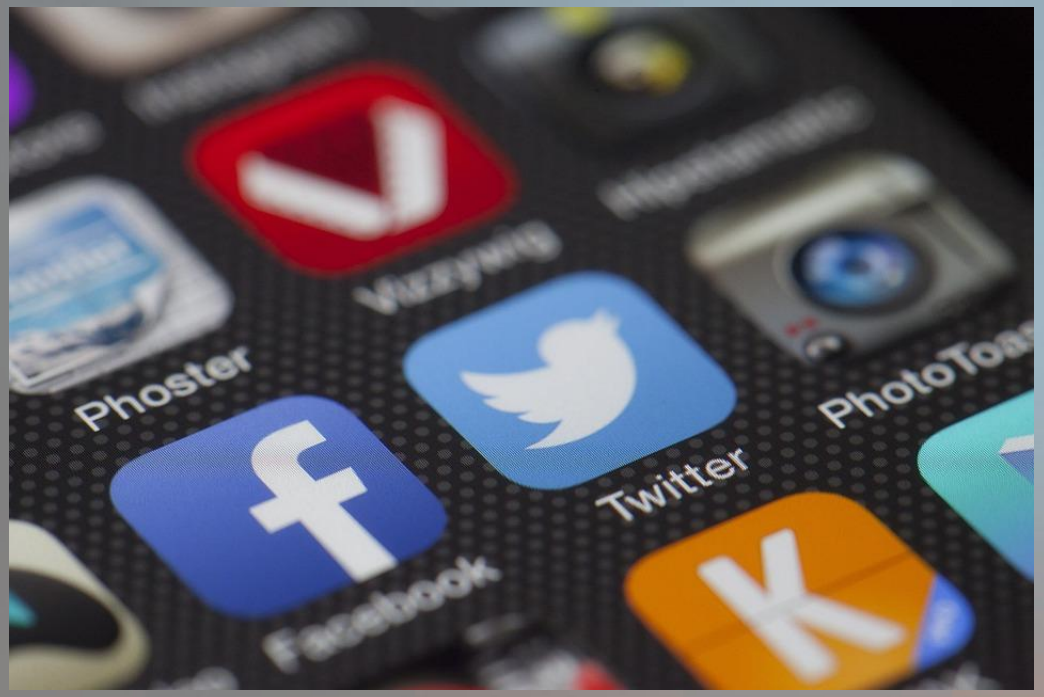
Índice:

Objetivos

- Estudio del problema del Análisis de datos.
- Empleo de Web Scraping para obtener las bases de datos.
- Estructuración de los datos y extracción de características.
- Balanceo de clases.
- Entrenamiento de algoritmos de aprendizaje automático.
- Extracción de conclusiones a partir de los resultados.

Índice:

Motivación: Web 2.0 y Minería de Texto



La Web 2.0 ha tenido un gran crecimiento generando enormes volúmenes de datos.

Una gran cantidad de estos datos son **texto**.

¿Qué es la minería de texto?

Proceso de extraer **conocimiento** o patrones interesantes y no triviales de documentos de texto **no estructurados**.

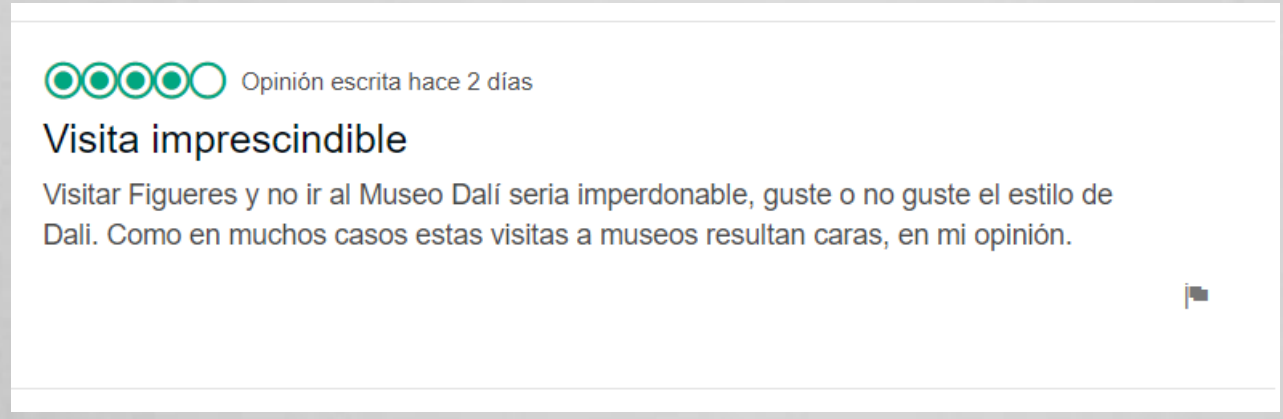
La estructuración de los datos es una tarea compleja.

Índice:

Análisis de sentimientos

Campo de estudio que analiza las **opiniones** de la gente hacia una **entidad** y sus **características**.

- ¿Qué es una opinión?
- ¿Qué es una entidad?
- ¿Qué es una característica?



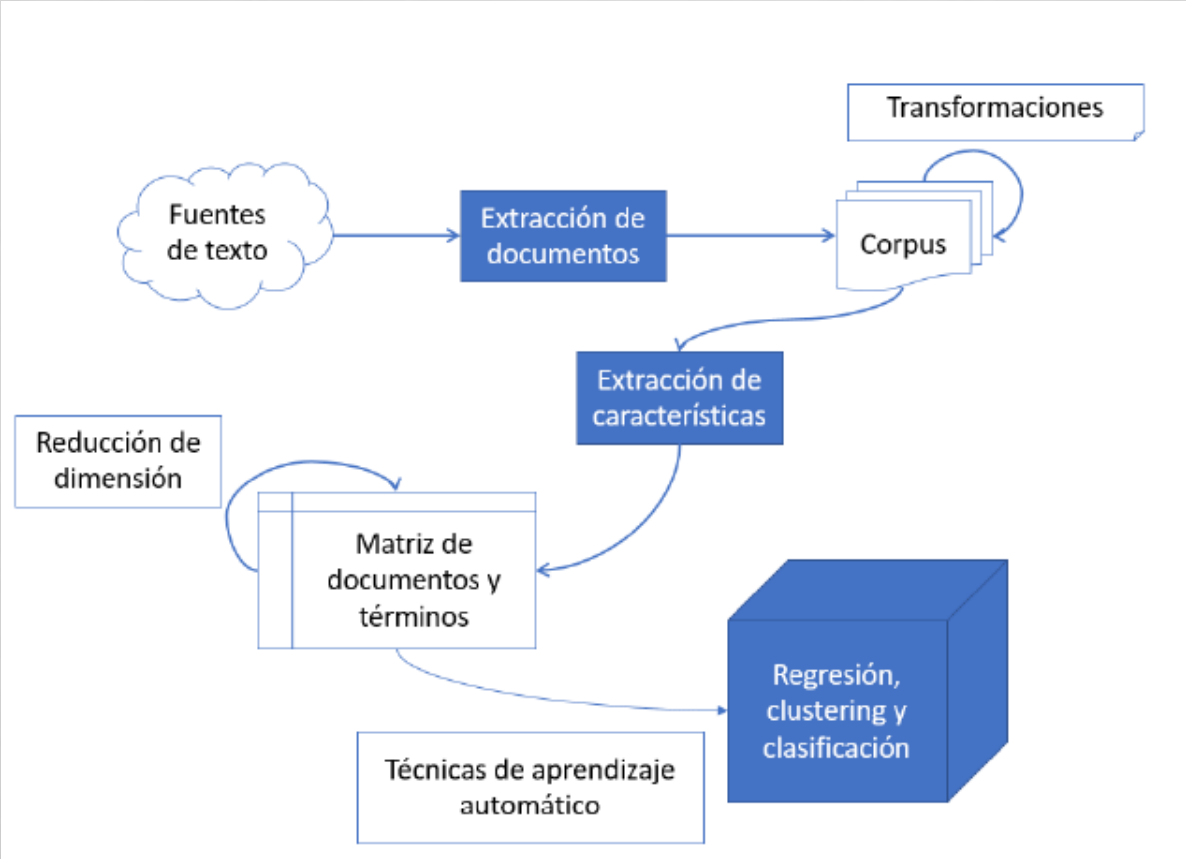
Análisis de sentimientos

Una opinión es una quíntupla (e, a, s, h, t):

- **e** es una entidad.
- **a** es un aspecto o característica.
- **s** es un sentimiento ligado a una opinión.
- **h** es la persona que da su opinión sobre una entidad.
- **t** es el momento en el que se da la opinión.

Análisis de sentimientos

El análisis de sentimientos tiene el siguiente proceso:

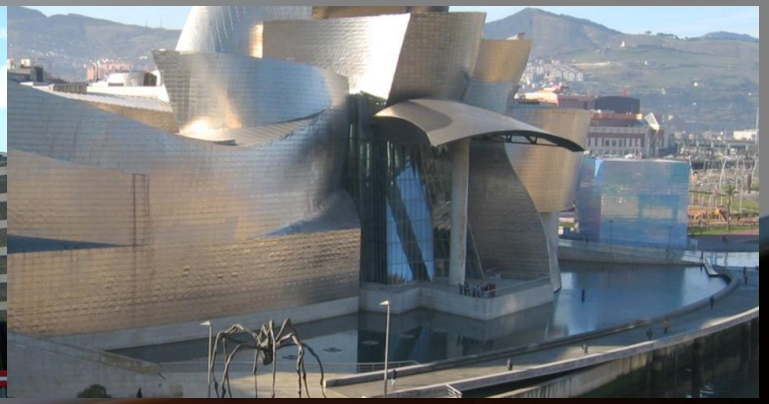


Índice:

Descripción del problema

- Se necesita un sistema **automático** que analice miles de opiniones y extraiga información relevante que permita tomar **decisiones**.
- Tanto **organizaciones** como **individuos** se benefician de las opiniones ajenas para tomar sus propias decisiones.
- Se aplica el **análisis de sentimientos** en el dominio del turismo y el viaje. Centrándonos dentro del turismo cultural, en los **museos**.

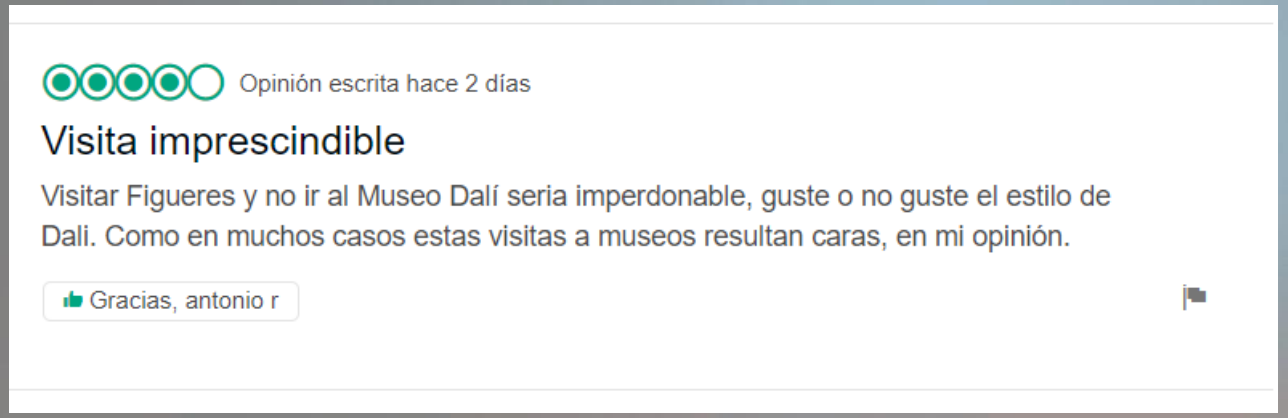




Índice:

Conjuntos de datos

Los datos se han extraído de la web de viajes y turismo **TripAdvisor.** Para ello se emplea **web scraping** para extraer información de los ficheros HTML que se descargan de la web.



Para el web scraping se ha utilizado el paquete de R “Rvest”.

Se descargan y analizan las opiniones de miles de visitantes de los 5 museos estudiados en este trabajo:

- Museo Reina Sofía, **11.670** opiniones.
- Museo del Prado, **41.733** opiniones.
- Teatro-Museo Dalí, **5.461** opiniones.
- Museo Thyssen-Bornemisza, **12.234** opiniones.
- Museo Guggenheim, **11.883** opiniones.

Conjuntos de datos

Características destacables de los datos:

- Etiqueta de clase:
  - SentimentValue, sentimiento del experto, es decir, el **usuario de TripAdvisor**.
  - SentimentCore, sentimiento máquina, inferido por el sistema **CoreNLP**.
- Datos con un gran desbalanceo de clases -> **Oversampling**

¿Cómo calculamos el sentimiento experto?

Índice:

Extracción de características

Unigram Feature Selection Method (UFSM)

```
UNIGRAM FEATURE SELECTION METHOD:

Dividir datos en 2 subconjuntos: positivo y negativo.

Para cada subconjunto hacer:
    1) Preprocesar el texto (palabras vacías, stemming, puntuación y
    números). Extraer unigramas.
    2) Calcular tfidf para cada unigrama,
    3) Calcula el valor medio de tfidf para todas las reseñas.
    4) Seleccionar los 500 unigramas con mayor tfidf de ambos conjuntos.
    5) Unir los conjuntos y eliminar unigramas comunes.
Fin para.

Construir Matrix Término-Documento.
```

¿Qué es el ***tfidf***?

$$tfidf = tf_{i,j} * idf$$

Medida numérica que expresa la relevancia de una palabra en un documento en una colección de documentos.

Empleo del paquete de R “tm” para el proceso de estructuración.

Extracción de características

Bigram Feature Selection Method (BFSM)

```
BIGRAM FEATURE SELECTION METHOD:

Dividir datos en 2 subconjuntos: positivo y negativo.

Para cada subconjunto hacer:
    1) Preprocesar el texto (palabras vacías, stemming, puntuación y
    números). Extraer bigramas.
    2) Calcular tfidf para cada unigrama,
    3) Calcula el valor medio de tfidf para todas las reseñas.
    4) Seleccionar los 500 unigramas con mayor tfidf de ambos conjuntos.
    5) Unir los conjuntos y eliminar unigramas comunes.
Fin para.

Construir Matrix Término-Documento.
```

¿Por qué utilizar pares de palabras?

Pueden aportar información que un unigrama no es capaz de aportar.

Métodos extraídos del TFM de Ana Valdivia

Índice:

Aprendizaje automático

Aprendizaje **supervisado** y no supervisado

Aplicar lo aprendido de **ejemplos etiquetados** con la clase correspondiente para **predecir** la clase de nuevos ejemplos.

Clasificación, regresión y clustering

Predecir el valor de una variable categórica **Y** en función de las características **X = (x1, x2, …, xn)**

Árbol de decisión c4.5, SVM y XGBoost

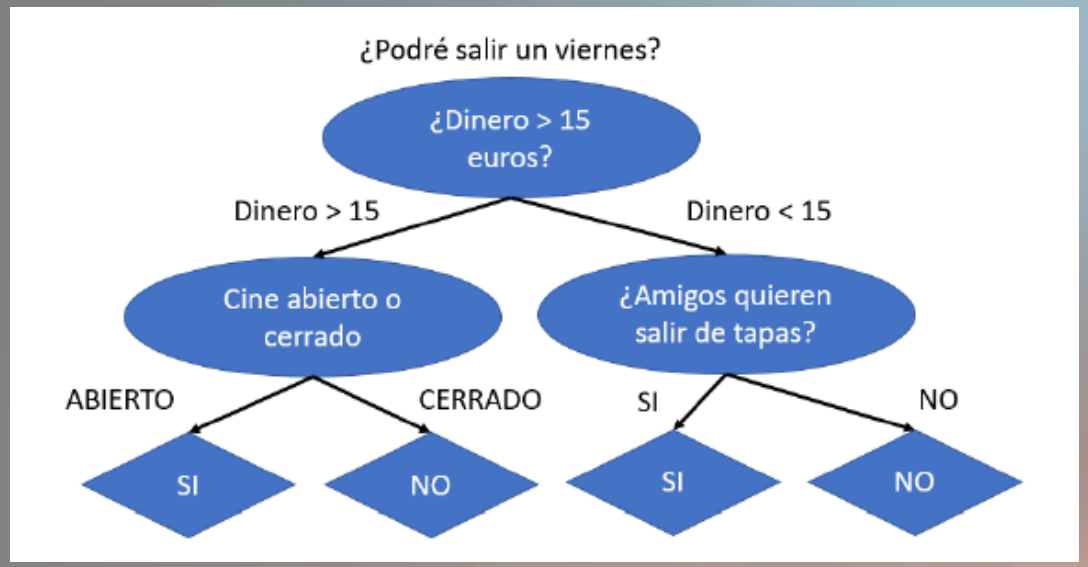
Algoritmos de clasificación.

Para el entrenamiento de algoritmos se ha empleado el paquete de R “caret”.

C4.5

Árbol de decisión que emplea la **ganancia de información modificada** como método de selección de las características elegidas a la hora de crear el árbol.

Existen otros métodos y árboles de decisión.



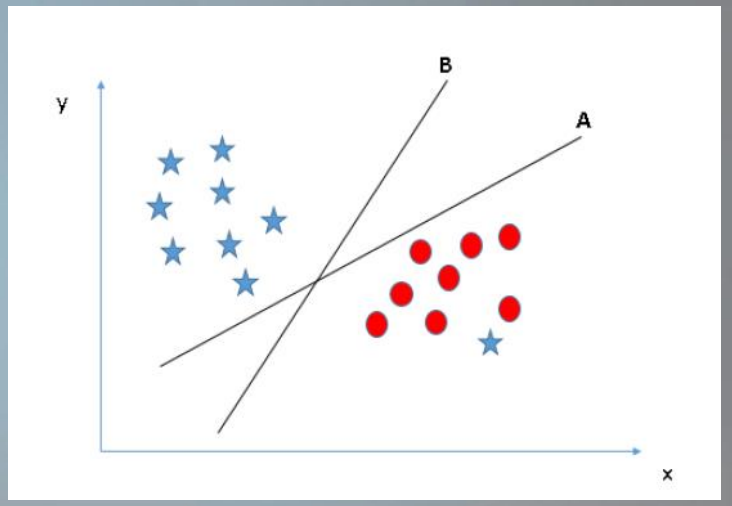

```
1) Si (dinero > 15) y (cine abierto) entonces (salir si)
2) Si (dinero > 15) y (cine cerrado) entonces (salir no)
3) Si (dinero < 15) y (amigos_quieren si) entonces (salir si)
4) Si (dinero < 15) y (amigos_quieren no) entonces (salir no)
```

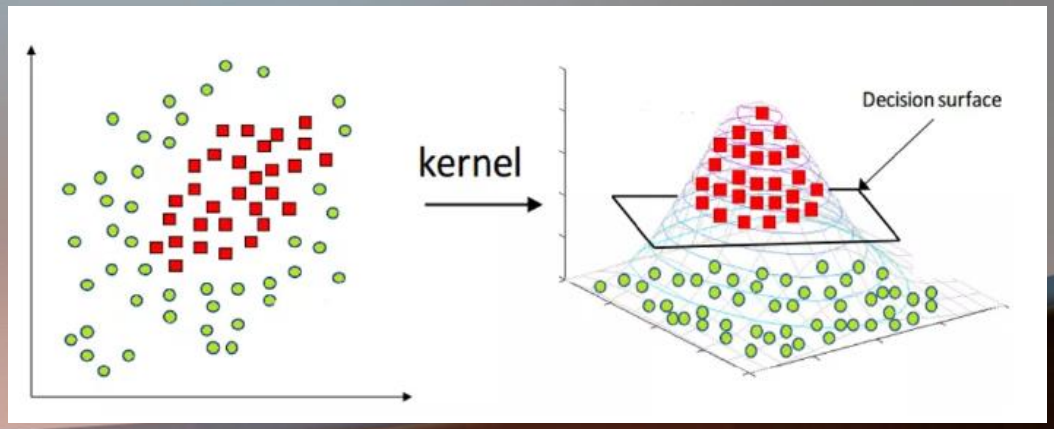
SVM

Técnica que trata de separar con una recta lo más larga posible los datos para poder diferenciar si una instancia pertenece a una clase u otra.

Esta recta de separación es llamada **hiperplano**.

Este algoritmo es capaz de cambiar la dimensionalidad con que representamos los datos para buscar un hiperplano que los separe correctamente.





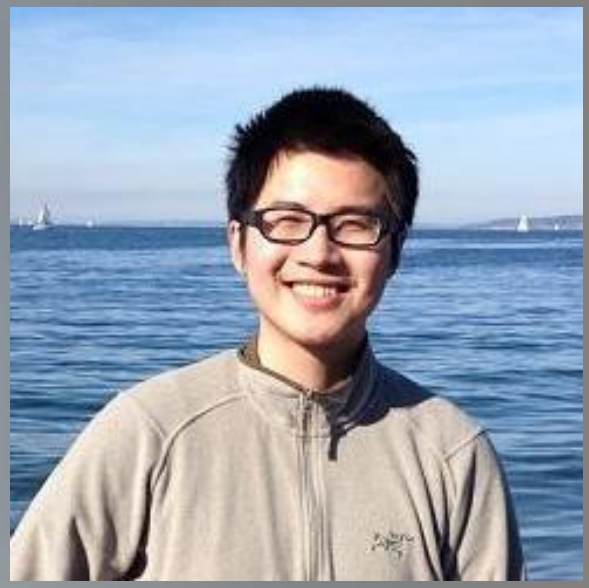
XGBoost

Técnica que emplea la salida de diferentes árboles de decisión para obtener una mejor solución.

Características:

- Algoritmo con muy **buenos resultados** que está dominando recientemente la plataforma Kaggle.
- **Rápido.**
- **Buen rendimiento**.

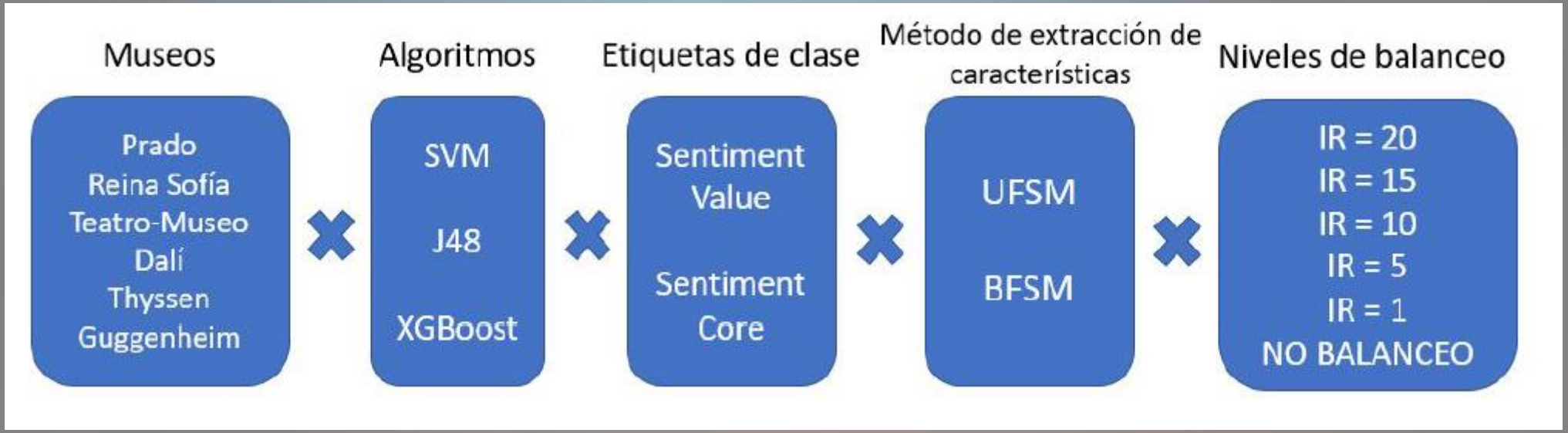
Creador: Tianqi Chen



Índice:

Experimentación

Posibles combinaciones para la creación de conjuntos de datos de entrenamiento.



Índice:

Conclusión y trabajos futuros.

Conclusiones:

- Sentimiento experto vs Sentimiento máquina.
- Desbalanceo y oversampling.
- Los unigramas obtienen mejores resultados que los bigramas.
- Los bigramas necesitan de un mejor preprocesamiento.
- Opiniones negativas debido a consejos para otros visitantes.
- La etiqueta clase ha obtenido mejores resultados que la etiqueta de sentimiento máquina.
- Todos los algoritmos han logrado experimentos con muy buenos resultados pero el más robusto ha sido SVM.

Conclusión y trabajos futuros.

Trabajos futuros:

- Empleo de técnicas de extracción de características más complejas.
- Estudio de los parámetros de los algoritmos para obtener mejores resultados.
- Estudio y empleo de técnicas de oversampling más complejas.
- Empleo de Subgroup Discovery.
- Uso de opiniones neutras en el estudio del análisis de sentimientos.

Consideraciones finales

- Licencias:
  - Memoria: Creative Commons Attribution-NonCommercial 4.0 International
  - Código: GNU General Public License v3.0

Todo el contenido del proyecto está en el repositorio de Github: https://github.com/Yussoft/TFG-

Gracias por su tiempo

¿Preguntas?