

## Caso Sentimientos en Twitter

Twitter es una red social que ha degradado a una comunidad llena de *Trolls* (trollear es la acción de hacer sentir mal o hacer enojar a alguien con bromas pesadas o comentarios fuera de lugar) o *Cuñados* (aquellos que comentan sobre cualquier asunto, queriendo aparentar ser más listos que los demás). El dicho “No alimentes al Troll” no se aplica en Twitter y estos se retroalimentan hasta el punto de conseguir que conversaciones de alto nivel se conviertan en un patio de escuela. Entre *Trolls* y *Cuñados* Twitter ya no es lo que era.

Gracias a las técnicas de análisis de texto (Text Analysis) podemos combatir los *Trolls* y *Cuñados* y devolver Twitter a sus años de esplendor. Text Analysis consiste en extraer información a partir de datos de lenguaje humano para comprender cómo otros seres humanos entienden el mundo.

En este Caso deberás realizar un análisis sentimental, sintáctico y gramatical de comentarios Twitter. La base de datos la puedes descargar desde eStudy (Caso Twitter dataset), la cual contiene un CSV de mensajes enviados a Twitter con las siguientes columnas:

1. Puntuación sentimental o polaridad (negativa, neutral, positiva) (por calcular)
2. Id del tweet
3. Fecha del tweet (Sat May 16 23:58:44 UTC 2009)
4. Búsqueda. Tweet resultado de una búsqueda. Si no hay búsqueda, el valor es NO\_QUERY
5. Usuario que ha tuiteado
6. Texto del tweet

Con estos datos se os propone que apliquéis técnicas analíticas y de visualización para responder a las siguientes preguntas. No hay restricciones acerca de las técnicas ni tecnologías a utilizar siempre y cuando los resultados sean reproducibles y estén debidamente justificados. No obstante, las siguientes librerías y códigos de ejemplo os pueden ser muy útiles para responderlas:

### Librería AFINN-165

<https://pypi.org/project/afinn/>

Propósito: Análisis sentimental para detectar textos con carácter positivo, negativo o neutral.

### Librería NLTK

<https://www.nltk.org/install.html>

Propósito: Trabajar con datos en lenguaje humano.

### Librería textstat

<https://pypi.org/project/textstat/>

Propósito: Calcular estadística a partir de datos en lenguaje humano.

**Librería NeatText**

<https://pypi.org/project/neattext/>

Propósito: Alternativa simple para limpiar datos textuales y preprocesamiento de texto.

**Librería NetworkX**

<https://networkx.org/>

Propósito: Creación, manipulación y estudio de la estructura, dinámica y funciones de redes complejas.

ANÁLISIS: Añade en cada respuesta una justificación de la(s) librería(s) y función(es) usadas.

1. Calcula la polaridad de cada tweet y su índice de legibilidad. Muestra una tabla resumen del *dataset* solo de aquellos 10 tweets que un estudiante de 3ºESO pueda entender, ordenados por legibilidad de mayor a menor, mostrando los 5 primeros tweets de mayor legibilidad y los 5 últimos tweets de menor legibilidad.
2. Analiza las frases de los tweets del *dataset*:
  - a. Número de frases totales.
  - b. Número de palabras distintas de más de 3 letras.
  - c. Máxima, mínima y media de longitud de las palabras.
  - d. La facilidad de lectura considerando las frases del tweet.
3. Detecta los usuarios *trolls* mostrando los 10 primeros. Son *trolls* aquellos usuarios con un porcentaje de mensajes negativos superior al 75%.
4. ¿Qué relación existe entre la tendencia de polaridad y legibilidad?
5. ¿Cuáles son los 3 usuarios más *influencers* considerando su grado de centralidad entre sus relaciones?

VISUALIZACIÓN: Añade conclusiones de interpretación para cada gráfico.

6. Crea un gráfico nube de palabras para las polaridades: positiva, negativa y neutral.
7. Alrededor de ciertas temáticas se generan conversaciones interesantes. Realiza una visualización agregada de la polaridad de las conversaciones para cada una de las palabras “kindle” y “sleep”.
8. Muestra la media de legibilidad de los usuarios más citados según cuartiles.
9. Muestra un grafo de las relaciones entre los usuarios del *dataset* identificando por colores el factor que creas más conveniente (trolls, influencers, legibilidad, polaridad...).
10. Muestra un grafo de las relaciones entre los usuarios *trolls* del *dataset*. Descartar usuarios no *trolls*.