Proyecto de análisis de sentimientos con Python
Inteligencia Artificial -Ingeniería del Software

Curso 2022/2023
Propuesta de trabajo
Ana Belén Romero-Losada

Objetivo y contexto

Se pretende un modelo de aprendizaje automático capaz de analizar el sentimiento de frases o textos y aplicarlo para predecir el sentimiento de *tweets*.
El análisis de sentimientos es una rama de la inteligencia artificial que se dedica a analizar las opiniones, emociones y actitudes expresadas en textos escritos, tales como comentarios en redes sociales, reseñas de productos o servicios, artículos de opinión, entre otros. Esta disciplina surge como respuesta a la necesidad de extraer información valiosa a partir de grandes cantidades de datos no estructurados generados por usuarios en internet.
Esta metodología se aplica en diversos campos, desde el *marketing* y la publicidad hasta la política y la salud mental. En el mundo empresarial, por ejemplo, se utiliza para conocer la satisfacción del cliente y mejorar la calidad de los productos y servicios ofrecidos. En política, puede ayudar a los candidatos a entender cómo se perciben sus discursos y propuestas, mientras que en salud mental puede ser utilizado para detectar signos tempranos de depresión o ansiedad.
Existen herramientas de procesamiento de lenguaje natural que permiten realizar tareas de detección de sentimientos. Este modelo asigna una polaridad a cada texto analizado, es decir, una etiqueta que indica si la opinión es positiva, negativa o neutra.
Durante el desarrollo de este proyecto se aprenderá a usar las herramientas ya existentes para *python* para generar un conjunto de datos de entrenamiento con el que desarrollar vuestro propio modelo de aprendizaje automático capaz de analizar el sentimiento detrás de *tweets*.

Objetivos específicos:

- Generar tu propio conjunto de datos de entrenamiento.
- Usar las herramientas ya existentes para el análisis de sentimientos.
- Usar distintas herramientas de procesamiento de lenguaje natural para depurar el texto recopilado.
- Proponer un modelo de aprendizaje automático que refine la detección de sentimientos, diferenciando entre “Muy feliz”, “Contento”, “Neutro”, “Molesto” y “*Hater*”.
- Usar el modelo desarrollado para clasificar el estado de ánimo de dos personajes públicos según sus *tweets*.

Metodología

1. **Recopilación de datos**. Se deben recopilar mas de 1000 *tweets* que puedan tener un sentimiento positivo, negativo o neutro. Procurar mantener en la medida de lo posible una proporción equilibrada. Para la recopilación de los datos se puede hacer de forma manual o utilizar la librería tweepy de *Python*, que permite acceder a la API de *Twitter* y buscar *tweets* que contengan ciertas palabras clave o *hashtags*. Los *tweets* recopilados se deben almacenar en un archivo de texto plano (*txt*) o en un archivo *CSV*.

2. **Eliminar las palabras que no aportan información.** Crear una función *limpiar_texto* que utilizando la librería *NLTK* (*Natural Language Toolkit*):

- Utilice técnicas de limpieza de texto para eliminar las palabras comunes y poco informativas (conocidas como *stop words*).
- Eliminar menciones, *hashtags*, *URLs* y cualquier otro símbolo extraño que pueda estar presentes en los *tweets*.
- Utilice técnicas de lematización (*stemming*) y *tokenización* para reducir las palabras a su forma base y así simplificar el vocabulario y reducir el número de características a considerar en el modelo.

3. **Etiquetado de datos.** Utilizando un paquete librería ya existente, se debe etiquetar las frases o textos recopilados en la etapa anterior con las etiquetas mencionadas en los objetivos. El resultado de este proceso debe almacenarse en un archivo *CSV* con dos columnas: una para la frase o texto y otra para la etiqueta correspondiente. **Deberá usarse el paquete librería *TextBlob* en los trabajos presentados en la 1ª convocatoria (Junio) y *VADER* para la 2ª y 3ª convocatoria (Julio y Noviembre).**

4. **Validación de la predicción realizada.** Revisar y corregir posibles errores en la etiquetación de las frases o textos de forma manual. Generar así un conjunto de datos etiquetados por un modelo ya existente y corregidos por un humano, que se usará como datos de entrenamiento y prueba en el desarrollo de nuestro modelo.

5. **Entrenamiento del modelo.** Utilizando diferentes algoritmos de aprendizaje automático, como los estudiados en esta signatura, se deben entrenar varios modelos con el conjunto de datos etiquetados. Se debe evaluar y comparar el rendimiento de los diferentes modelos utilizando técnicas como validación cruzada.

6. **Predicción de tweets.** Utilizando el modelo con mejor rendimiento, se deben predecir el sentimiento de los *tweets* del conjunto de prueba. Se debe evaluar la precisión de las predicciones y comentar los resultados.

7. **Análisis de *tweets* de personajes públicos.** Se deben recopilar los últimos 30 *tweets* de dos personas con influencia en redes sociales, alguna conocida por su mala fama como *hater* y otra con una valoración social más positiva. Utilizar el modelo desarrollado para predecir el sentimiento de cada *tweet*. Se debe generar un *pie chart* en *python* que muestre el estado de ánimo de cada personaje (el porcentaje de *tweets* clasificados con cada etiqueta) y comparar ambas proporciones. Discutir si la etiqueta de *hater* que le ha atribuido tiene fundamento.

Documentación y entrega

El trabajo deberá documentarse siguiendo un formato de artículo científico, con una extensión mínima de 6 páginas. En la página web de la asignatura se pueden encontrar plantillas donde se sugiere una estructura general. Estas plantillas siguen el formato de los *IEEE conference proceedings*, cuyo sitio web guía para autores ofrece información más detallada. El documento entregado deberá estar en formato *PDF*.

En el caso concreto de este trabajo, la memoria deberá al menos incluir: introducción, funcionamiento de los modelos de detección de sentimientos ya existentes y que aplicaciones tienen actualmente, depuración del texto, desarrollo del modelo y gráficas anímicas de cada uno de los personajes públicos elegidos como conclusiones. En ningún caso debe incluirse código en la memoria. La entrega del trabajo consistirá de la memoria del trabajo y el código implementado (cuadernos de *Jupyter*). Ambos deben subirse a la página de la asignatura en un único fichero comprimido *zip*.

Criterios de evaluación:
Para la evaluación del trabajo se tendrán en cuenta los siguientes criterios, considerando una nota total máxima de 4 puntos:

Memoria del trabajo (hasta 1.5 puntos): se valorará la claridad de las explicaciones, el razonamiento de las decisiones, el análisis y presentación de resultados y el correcto uso del lenguaje. La elaboración de la memoria debe ser original, por lo que no se evaluará el trabajo si se detecta cualquier copia del contenido.

Código fuente (hasta 1.5 puntos): se valorará la claridad y buen estilo de programación, corrección y eficiencia de la implementación y calidad de los comentarios. El código debe ser original, por lo que no se evaluará el trabajo si se detecta código copiado o descargado de internet.

Presentación y defensa (hasta 1 puntos): se valorará la claridad de la presentación y la buena explicación de los contenidos del trabajo, así como, las respuestas a las preguntas realizadas por la profesora.