**1. Descripción del escenario del proyecto**

El proyecto consiste en un análisis de datos, especializado en el sector audiovisual, de la película Gladiator 2. Para la realización del análisis se ha desarrollado un sistema de extracción de información y análisis de sentimientos de reseñas sobre la película.

El objetivo es identificar las opiniones del público en relación con los siguientes aspectos:

* Opinión general sobre la película
* Opinión sobre el guión
* Opinión sobre los actores
* Opinión sobre la producción

Del conjunto de reseñas se extraerán las entidades relevantes que se clasificarán automáticamente según el sentimiento asociado.

**2. Corpus**

El corpus de análisis se ha obtenido de forma sintética mediante *prompt engineering*, debido a que las API´s de las plataformas que facilitan las reseñas sobre películas, como *IMDB*, *FilmAffinity* o *Rotten Tomatoes*, son de pago y tampoco permiten el *web scraping.*

Prompt: “*Actúa como un crítico especialista en cine. Accede a la página IMDB y analiza las reseñas asociadas a la película Gladiator II. Ahora ofréceme, en formato tabla, 40 reseñas inventadas por ti que reflejen de manera fiel el sentimiento de las reseñas analizadas.*”

En el corpus de análisis se ha añadido una columna llamada “Continent” que contiene las etiquetas de EEUU y EUROPA para evaluar las valoraciones de cada uno de los aspectos analizados en estos lugares.

Esta misma técnica se ha utilizado para la creación del corpus de entrenamiento, sin embargo, se le ha pedido que analice diferentes películas entre las que se han incluido algunas con contenido histórico como Troya o Centurión, puesto que el análisis de películas históricas y épicas puede variar en relación a películas de tipo contemporáneo. Además, se le ha pedido que asocie un sentimiento positivo(POS), negativo(NEG) y neutro(NEU) a cada una de las reseñas, para disponer de manera rápida de un número elevado de reseñas etiquetadas para el entrenamiento del modelo.

**3. Nombre de las tareas**

**3.1 Tratamiento de los datos**

Los datos obtenidos se han procesado para eliminar el ruido. Se han eliminado correos electrónicos, nombres de usuario, emoticonos, puntuación, se han convertido en minúsculas y se han tokenizado mediante la librería NLTK, utilizando “Word\_tokenize”. No se han filtrado las stopwords, puesto que son importantes para un proyecto de análisis de sentimiento.

**3.2 Definición de diccionario**

**Recursos definidos:**

Para conseguir una lista de entidades adecuada al objetivo del proyecto se ha recurrido al marco semántico del mundo cinematográfico. Este marco nos presenta las palabras necesarias para el análisis de un objeto audiovisual. Para obtener un diccionario de entidades adecuado al propósito del proyecto se ha sometido el conjunto de datos de un proceso de “bolsa de palabras” para determinar la relevancia de los términos más utilizados y que sean más adecuados para la extracción de entidades.

De acuerdo con los resultados obtenidos se ha elaborado el siguiente diccionario de entidades.

Entidad: Gladiator II

Variantes: secuela, continuación, película

Tipo semántico: PELÍCULA

Etiqueta: GENERAL

Ejemplo: Como fanático de "Gladiator", esperaba mucho de esta secuela. Aunque tiene momentos épicos, la trama se siente un poco forzada. Hay secuencias bien logradas, pero el guion no tiene el mismo impacto emocional que la primera película. Paul Mescal ofrece una actuación convincente, pero no logra transmitir la misma carga emocional que Russell Crowe en la original. Es entretenida, pero no memorable.

**Salida esperada: esperaba mucho de esta secuela**

Entidad: Paul Mescal, Pedro Pascal, Denzel Washington

Tipo semántico: PERSONA

Etiqueta: ACTOR

Ejemplo: Como fanático de "Gladiator", esperaba mucho de esta secuela. Aunque tiene momentos épicos, la trama se siente un poco forzada. Hay secuencias bien logradas, pero el guion no tiene el mismo impacto emocional que la primera película. Paul Mescal ofrece una actuación convincente, pero no logra transmitir la misma carga emocional que Russell Crowe en la original. Es entretenida, pero no memorable.

**Salida esperada: Paul Mescal ofrece una actuación convincente**

Entidad: Producción

Variantes:Ridley Scott, dirección, producción, visual, escenas, efectos especiales coreografías, música, batallas

Tipo semántico: OBJETO FÍLMICO

Etiqueta: PRODUCCIÓN

Ejemplo: Como fanático de "Gladiator", esperaba mucho de esta secuela. Aunque tiene momentos épicos, la trama se siente un poco forzada. Hay secuencias bien logradas, pero el guion no tiene el mismo impacto emocional que la primera película. Paul Mescal ofrece una actuación convincente, pero no logra transmitir la misma carga emocional que Russell Crowe en la original. Es entretenida, pero no memorable.

**Salida esperada: Hay secuencias bien logradas**

Entidad: Guion

Variantes: trama, historia, narrativa, guión, personajes,desarrollo

Tipo semántico: OBJETO ESCRITO

Etiqueta: GUION

Ejemplo: Como fanático de "Gladiator", esperaba mucho de esta secuela. Aunque tiene momentos épicos, la trama se siente un poco forzada. Hay secuencias bien logradas, pero el guion no tiene el mismo impacto emocional que la primera película. Paul Mescal ofrece una actuación convincente, pero no logra transmitir la misma carga emocional que Russell Crowe en la original. Es entretenida, pero no memorable.

**Salida esperada: La trama se siente un poco forzada. El guion no tiene el mismo impacto emocional que la primera película.**

**3.3 Extracción de entidades**

Realizado con el modelo “es\_core\_news\_lg" de la librería Spacy, personalizado con *EntityRuler*.

**3.4 Entrenamiento del modelo**

En este proceso se han seleccionado los modelos más destacados de la librería *scikit-learn* para un problema de clasificación *single label.*

* Regresión Logística (*Logistic Regression,LoR*)
* Naive Bayes (*Multinomial NB*)
* Clasificador de Vectores de Soporte (*Support Vector Classifier, SVC*).
* Clasificador de Vectores de Soporte con Kernel lineal ( *LSVC* ).
* Árboles de decisión (*Decision Tree Classifier*,*DTC*)

**3.5 Procesamiento de los datos**

En esta fase del proyecto se han seleccionado los textos para cada etiqueta y se han procesado con el modelo entrenado SVC. Esto ha dado una serie de resultados que se han graficado utilizando la librería matplotlib para su posterior análisis.

**4. Perfiles profesionales necesarios**

Para un proyecto de estas características es necesario contar con **expertos en procesamiento del lenguaje natural** que sepan reconocer patrones, los marcos semánticos necesarios para la actividad y para la definición de un diccionario que permita extraer la información de la manera más exacta posible.

Es necesario que existan **anotadores** para el desarrollar los corpus de entrenamiento de una manera rigurosa y exacta para evitar problemas en el modelado de los datos.

Este modelado y su posterior evaluación deben llevarlo a cabo los **científicos de datos** que realizan las tareas de escalamiento y estandarización de los datos, en caso de que fuese necesario;así como la selección de características en problemas de regresión.

Por otra parte, sería necesaria la intervención de un **analista de datos** para la correcta interpretación de los datos obtenidos.

Dependiendo del destino de este proyecto podrían ser necesarios **desarrolladores de software** para su implementación en alguna página web o aplicación.

**5. Conclusiones**

El mayor problema al que nos podemos enfrentar es encontrar los datos necesarios para el desarrollo del proyecto. Poseer datos estructurados es raro, puesto que la mayor parte de los datos en la web se encuentran de manera desestructurada. La inteligencia artificial puede ayudar en estos casos, pero las empresas son conscientes del valor que los datos han adquirido en los últimos tiempos y es difícil encontrar que los pongan a nuestra disposición de manera gratuita. Otro obstáculo al que nos podemos enfrentar es al de un nuevo lenguaje, el lenguaje de la programación y la cantidad de recursos y librerías útiles que existen y que pueden hacer que nos perdamos en un mar de recursos. Para que eso no suceda es imprescindible centrarse en las librerías y recursos necesarios para el desarrollo del trabajo y, si fuese necesario, ampliar los conocimientos progresivamente.