Proyecto 1- Inteligencia de Negocios.

Juan David Bautista.

Juan Diego Lugo.

Daniel Zambrano.

Parte 1

Entendimiento del negocio y enfoque analítico.

1. Oportunidad/problema Negocio
   Para el entendimiento del negocio y la posible oportunidad de negocios se trata de una plataforma que utilice un modelo predictivo de inteligencia artificial basado en reseñas obtenidas de redes sociales. De esta forma se puede ayudar a los usuarios a tomar decisiones sobre que películas ver. Además, la plataforma podría a partir de las películas reseñadas por los usuarios recomendarles películas similares a la que haya reseñado, o en caso de no haber hecho ninguna reseña recomendarle en base a las búsquedas que ha realizado.

   Esto implicaría que esta idea de negocio sea una página web de consulta pública donde la gente pueda crear un perfil, que esté ligado a datos relacionados con sus reseñas, películas vistas, géneros preferidos, etc. De esta manera el negocio puede crear vistas llamativas y sistema de premios para las personas que decidan realizar reseñas o llenar su perfil. Adicionalmente, se deberían abrir espacios para la retroalimentación de cada reseña y que las personas den puntos de credibilidad a los usuarios más populares o que más reseñas tengan.

   Finamente, la capitalización de este proyecto se encuentra en la venta y colaboración con industrias cinematográficas, las cuales encuentren útil la información que recopila el negocio. También, se podría presentar un sistema de publicidad para próximos largometrajes que se quieran sacar al mercado.

2. Enfoque analítico
   El enfoque que se usara para procesar el texto estructurado que ya tenemos, es el análisis de sentimientos en las palabras o estructuras identificadas en un texto pasado por parámetro.

Se hará un análisis en base a herramientas basadas en léxico que usa diccionarios para clasificar las estructuras del texto como positivas o negativas.

3. Definición de usuarios que se benefician:
   Los productores de películas: Necesitan entender las percepciones del público acerca de sus películas para poder tomar decisiones bien fundamentadas sobre cómo mejorarlas y cómo enfocar su estrategia de promoción.

   Académicos: Los expertos académicos utilizan la percepción del público sobre distintos géneros de películas, cineastas o intérpretes como tema de investigación para publicar estudios académicos o presentarlos en conferencias.

   Distribuidores de películas: Los distribuidores de películas emplean la percepción del público para elegir qué películas distribuir en diversas zonas geográficas y segmentos del mercado, con el objetivo de aumentar las oportunidades de triunfo y reducir la posibilidad de incurrir en pérdidas económicas.

   Investigadores de mercado: Los especialistas en investigación de mercado pueden emplear la herramienta para realizar estudios de mercado y obtener datos relevantes acerca de las inclinaciones del público en distintos países, zonas geográficas o grupos demográficos.

   Encargados de la toma de decisiones: Los encargados de la toma de decisiones pueden hacer uso de la aplicación para tomar decisiones bien fundamentadas acerca de qué películas financiar, distribuir o promocionar, basándose en las percepciones del público.

   Plataformas de streaming: Las plataformas de streaming pueden emplear la aplicación para sugerir películas a sus usuarios de acuerdo a sus preferencias y gustos.

   Agencias de publicidad y marketing: Las agencias de publicidad y marketing pueden hacer uso de la aplicación para obtener datos relevantes acerca de las inclinaciones del público y adaptar sus campañas publicitarias en función de ellos.

4. Técnicas y algoritmos utilizados
   Algoritmos de clasificación.
   - VADER & RoBERTa
   - Multinomial Naive Baye
   - Random forest

   Herramientas de tokenización y vectorización.
   - TF-IDF

Entendimiento y preparación de los datos.

1. Entendimiento de los datos
   Las reseñas de una película son importantes para la percepción de esta, pueden influir en la opinión que tiene el público sobre la película antes de verla. Las personas suelen confiar en las opiniones de los demás, por ello las reseñas son una forma de obtener información y una opinión sobre una película antes de decidir verla o no.

   Las reseñas pueden proporcionar una idea de la calidad de la película, su trama, los personajes, el ritmo, la dirección, el guion, la cinematografía, el sonido, entre otros aspectos. Las reseñas también pueden destacar los aspectos positivos y negativos de una película, lo que puede ayudar a las personas a decidir si les gustaría verla o no.

   Finalmente, las reseñas también pueden ser importantes para la percepción de la película después de haber sido vista. Es común que las personas compartan sus opiniones y comentarios sobre la misma, y estas opiniones pueden influir en la opinión de otros. Si las reseñas son negativas, esto puede influir en la percepción de la calidad de la película, mientras que las reseñas positivas pueden aumentar la popularidad de esta.

   Dado que como se mencionó anteriormente se tratan de opiniones que suelen estar sesgadas por las emociones causadas sobre las personas, sumado al hecho de que en internet no hay una rigurosa revisión ortográfica ni un sentido estricto de seguimiento de reglas gramaticales es necesario transformarlos a un formato más homogéneo.

2. Preparación de los datos
   Para la preparación de los datos se hizo uso de la librería re de Python para definir expresiones regulares, es así como se eliminan todos los caracteres especiales, a la par de los números y se reemplazan por espacios. Una vez eliminados se reemplazan todos los múltiples espacios por un único espacio dejándonos con una entrada mucho más uniforme.

   Ejemplo:

   Sin limpiar

   > "Es difícil contarle más sobre esta película sin estropearla.Lo disfruté porque no esperaba lo que estaba viendo, sino un drama sexual ordinario, así que ... es un thriller de PSCYHO-sexual, en el que nada es lo que parece.Cuenta con Emmanuelle Seigner, sin extraño al género (y a la desnudez) en la que su esposo, Polanski, la había dirigido.Y un rendimiento espeluznante (dije espeluznante / sí espeluznante) de Toreton (el actor de Bernard Tavernier).Parece que un Pascal Bruckner se encuentra con Roman Polanski (mejor que la luna amarga), como un chabrol que se fue por extravisos o thriller de Clouzot (he visto a alguien que menciona Les diaboliques), pero más cerca de Georges Franju's Les Yeux Sans Visage (ojos sin cara, laPadrino del Dr. Phibes y más).Una gema !Solo me temo que lo hicieran en un remake de Hollywood como lo hicieron con Nighwatch y la desaparición."

Limpio

es difícil contarle más sobre esta película sin estropearla lo disfruté porque no esperaba lo que estaba viendo sino un drama sexual ordinario así que es un thriller de pscyho sexual en el que nada es lo que parece cuenta con emmanuelle seigner sin extraño al género y a la desnudez en la que su esposo polanski la había dirigido y un rendimiento espeluznante dije espeluznante sí espeluznante de toreton el actor de bernard tavernier parece que un pascal bruckner se encuentra con roman polanski mejor que la luna amarga como un chabrol que se fue por extravisos o thriller de clouzot he visto a alguien que menciona les diaboliques pero más cerca de georges franju s les yeux sans visage ojos sin cara lapadrino del dr phibes y más una gema solo me temo que lo hicieran en un remake de hollywood como lo hicieron con nighwatch y la desaparición

Modelos aplicados.

1. VADER y RoBERTa(Daniel Zambrano)

El enfoque que se usará en última instancia para procesar el texto estructurado que ya tenemos, es el análisis de sentimientos en las palabras o estructuras ya identificadas.

En primera instancia se hizo un análisis en base a una herramienta basada en léxico que usa diccionarios para clasificar las estructuras del texto como positivas, negativas o neutrales.

Se utilizará VADER (Valence Aware Dictionary and sEntiment Reasoner) que es la herramienta de análisis de sentimientos basada en reglas y léxico escogida, que está específicamente entrenada con los sentimientos expresados en las redes sociales y funciona bien en textos de otros dominios. Esta herramienta nos permite trabajar con un enfoque en donde se usa una bolsa o diccionario gigante de palabras ya clasificadas como buenas, neutrales o malas para hacer la tarea de clasificar los tokens del texto estructurado en alguna de estas categorías y al final computar una fórmula para calcular los puntajes del texto en general en cada categoría.

En segunda instancia se hará un análisis de sentimientos en base a un modelo ya entrenado usando una tarea de *masked language modeling* (MLM) que es simplemente hacer que el modelo pueda predecir qué palabra debería llenar los espacios en blanco de una oración, recibe de entrada una máscara de texto como "la película estuvo una [MASCARA]", y retorna las posibles palabras que podrían llenar tal máscara.

El modelo usado se llama RoBERTa (Robustly Optimized BERT Pre-Training Approach) y es un modelo de transformadores pre-entrenado en un gran cuerpo de datos en inglés de manera auto supervisada. **RoBERTa** Está destinado principalmente a ajustarse en una tarea específica de modelado de lenguaje, que es nuestro caso es de sentimientos.

Haremos que también clasifique estructuras del texto en positivas, negativas o neutras pero, con la particularidad de que en este modelo ya se consideran las relaciones entre

palabras y frases, así como también el sarcasmo y más comportamientos exhibidos por las personas cuando escriben algún texto y, en principio se espera que el resultado de este análisis sea más preciso y/o confiable que el de VADER.

2. Multinomial Naive Baye(Juan Diego Lugo Sanchez)

El algoritmo de Multinomial Naive Bayes es un modelo de clasificación de aprendizaje que se utiliza comúnmente en problemas de clasificación de texto, se basa en el teorema de Bayes, que es una fórmula matemática que permite calcular la probabilidad de que ocurra un evento en función de la probabilidad de eventos relacionados. En este caso, se utiliza el teorema de Bayes para calcular la probabilidad de que una determinada reseña pertenezca a una clase de sentimiento (por ejemplo, positivo o negativo) en función de la frecuencia de las palabras en esa reseña y en las reseñas de entrenamiento en la misma clase.

Se asume que las características de entrada son independientes entre sí y que cada característica contribuye de manera independiente a la probabilidad de que una instancia pertenezca a una clase determinada. La técnica es conocida como "naive Bayes" (Bayes ingenuo), que asume que las palabras en una reseña son independientes entre sí, lo que significa que la presencia de una palabra en una reseña no está relacionada con la presencia de otras palabras en la misma reseña. Aunque esta suposición es en realidad rara vez cierta, el algoritmo de Multinomial Naive Bayes se ha demostrado que funciona bien en la práctica y es muy eficiente computacionalmente.

Al utilizar una distribución multinomial para modelar la frecuencia de cada palabra en un documento cada palabra se considera una característica y se cuenta el número de veces que aparece en cada clase. Luego, se utiliza el teorema de Bayes para calcular la probabilidad de que una instancia pertenezca a una clase determinada, dadas sus características. La frecuencia de cada palabra en una reseña sigue una distribución de probabilidad multinomial, lo que significa que se cuentan todas las ocurrencias de cada palabra en una reseña y se utiliza esa información para calcular la probabilidad de que esa palabra aparezca en una reseña de una clase de sentimiento determinada. A continuación, se multiplican las probabilidades de cada palabra en la reseña para obtener la probabilidad de que la reseña pertenezca a esa clase de sentimiento.

Al final del ejercicio de analisis Los resultados del análisis de sentimientos utilizando ROBERTA y VADER sobre el dataset fueron de 0.72 y 0.70 respectivamente. En el caso de ROBERTA, se obtuvo una precisión media de 0.72, con una precisión alta (0.91) para la clase 0.0 y una baja (0.52) para la clase 1.0. El recall fue alto (0.85) para la clase 1.0 y bajo (0.66) para la clase 0.0. En el caso de VADER, se obtuvo una precisión media de 0.70, con una precisión baja (0.55) para la clase 0.0 y alta (0.85) para la clase 1.0. El recall fue alto (0.65) para la clase 1.0 y bajo (0.79) para la clase 0.0. En ambos casos, se obtuvo un f1-score similar (0.71 y 0.74 respectivamente) y una accuracy del 70% y 72%. Se puede concluir que el análisis de sentimiento con ROBERTA obtuvo una precisión y recall mejores que el análisis con VADER.

3. Random Forest (Juan David Bautista Parra).

Para la tarea de clasificación, la tarea por Random Forest es una buena elección debido a que es un algoritmo de aprendizaje supervisado que genera múltiples árboles de decisión, que se encarga de predecir variables binarias. Al aplicar este modelo con los tokens que se obtuvieron por TF-IDF se obtuvieron valores de F1 de 100% en los datos de entrenamiento y 81.60% en los de prueba.

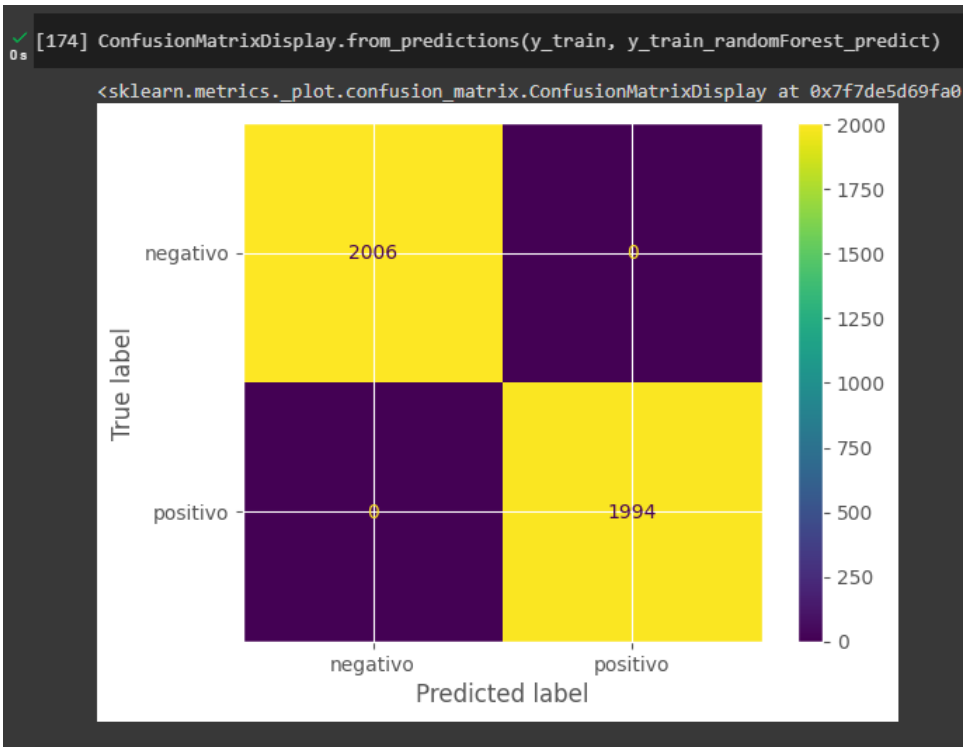

Imagen 1

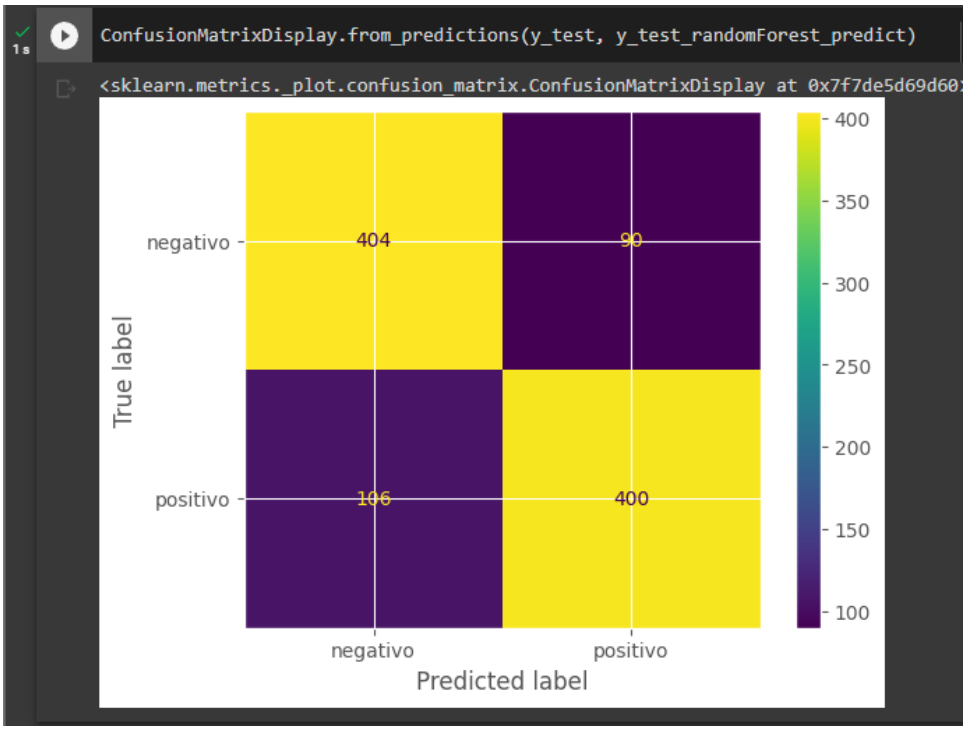

Imagen 2

Adicionalmente, se puede ver que hubo un total de 196 registros que fueron erróneamente clasificados y al hacer un análisis de estos se puede ver que los Falsos Positivos que el modelo clasificó se debe mayormente a que en los comentarios utilizaban palabras que pueden inducir a una falsa noción de positivismo, pero el contexto en el que lo utilizaban es distinto.

Para la búsqueda de los hiperparametros del modelo RandomForest se realizó un GridSearchCV con varios parámetros. Sin embargo, al ver los resultados del modelo obtenido por el GridSearch se observó que eran menos precisos que los hiperparametros por defecto. Por lo tanto, se concluye que los mejores hiperparametros para este modelo son 100 estimadores con una profundidad media de 1790.

Finalmente, los resultados de este modelo serian de gran ayuda para los negocios propuestos anteriormente. Tomando los resultados de las métricas de F1 se podría asegurar que el modelo puede ayudar a clasificar la mayoría de los casos en los que una persona tenga comentarios positivos respecto a un largometraje.

Validez de los resultados

Es fundamental en el ámbito de las investigaciones estadísticas tomar en cuenta varios elementos clave para poder determinar la credibilidad y precisión de las conclusiones obtenidas y así tomar decisiones bien informadas. Algunos de estos elementos son el tamaño de la muestra, la dirección de los datos, el método de muestreo utilizado y el tipo de análisis estadístico llevado a cabo.

Método de muestreo:

importante destacar que el método de muestreo simple es una técnica ampliamente utilizada y altamente confiable en estudios estadísticos. Este método se fundamenta en la suposición de que cada individuo de una población tiene igual probabilidad de ser seleccionado para formar parte de la muestra, lo que conduce a la obtención de una muestra representativa y precisa. Al emplear esta técnica, se aumenta significativamente la posibilidad de obtener resultados precisos y confiables que puedan ser generalizados con exactitud a toda la población. Por lo tanto, como estadístico, considero que el uso del método de muestreo simple es fundamental para llevar a cabo estudios estadísticos rigurosos.

Tamaño de la muestra:

La selección del tamaño de muestra óptimo es crucial en el diseño de investigaciones estadísticas, dependiendo de los objetivos del estudio y la población en cuestión. Para garantizar la precisión de los resultados, es importante que la muestra sea representativa de la población objetivo y que el tamaño de muestra sea suficiente. En este caso, se justifica el uso de una muestra de 4800 datos para el entrenamiento del modelo, lo que permite obtener una comprensión útil de las opiniones generales del público hacia una película. Es importante tener en cuenta que los resultados son solo una muestra y no representan necesariamente la opinión de toda la población.

Medida del sentimiento:

En esta investigación se evaluó la opinión sobre una película dividiéndola en dos categorías: positiva y negativa. Para ello, se utilizó un análisis estadístico basado en la frecuencia, el cual permitió determinar cuántas veces se seleccionó cada una de las dos categorías en la muestra. Es importante señalar que se empleó un análisis de correlación para examinar la relación entre la actitud positiva y negativa, lo que proporcionó información valiosa sobre cómo la audiencia percibía el producto, servicio o película en cuestión. Es fundamental tener en mente que la selección de la medida a utilizar en el estudio dependerá de los objetivos planteados y de la información que se pretenda obtener, siguiendo las directrices establecidas en el proyecto.

Análisis estadístico:

En el proyecto que se llevó a cabo, se utilizó una técnica de análisis estadístico para evaluar variables categóricas o nominales. Que consistió en medir el sentimiento expresado por el público mediante opciones positivas y negativas para encontrar la frecuencia de esto. Así pues, permitió determinar la cantidad de veces que cada opción fue seleccionada dentro de la muestra. Esto proporcionó información valiosa acerca de la opinión del público con respecto al objeto de estudio.

Es importante mencionar que esta técnica es muy útil en la investigación de mercados y estudios de opinión pública. A través de ella, se puede obtener información precisa y confiable acerca de las preferencias y opiniones de la audiencia, lo que a su vez puede ser utilizado para la toma de decisiones estratégicas en diferentes campos.

Finalmente, el proyecto en cuestión cumple con las condiciones necesarias para ser considerado una evaluación estadística rigurosa y confiable. Esto se debe a que se

utilizaron técnicas estadísticas adecuadas para la medición y análisis de variables categóricas o nominales, lo que permitió obtener información valiosa acerca de la percepción y opinión del público.

En primer lugar, se utilizó el análisis de frecuencia para determinar la cantidad de veces que cada opción fue seleccionada dentro de la muestra. Esto permitió obtener información precisa acerca de la frecuencia con la que los participantes eligieron opciones positivas y negativas, lo que a su vez proporcionó una idea clara acerca de la opinión del público respecto al objeto de estudio.

Es imprescindible mencionar que, el proyecto en cuestión utilizó una muestra aleatoria de 4800 datos. Esto significa que se seleccionaron los participantes de forma aleatoria, lo que garantiza que todos los individuos de la población tengan las mismas posibilidades de ser seleccionados y, por lo tanto, que la muestra sea representativa de la población en general.

Además, se utilizó un método de muestreo simple para la selección de los participantes. Este método es uno de los más sencillos y comunes en la investigación de mercados y estudios de opinión pública, y consiste en seleccionar los participantes de forma aleatoria y sin reemplazo. Esto garantiza que todos los individuos tengan las mismas posibilidades de ser seleccionados y que la muestra sea representativa de la población en general.

La utilización de una muestra aleatoria y de un método de muestreo simple es fundamental para garantizar la validez y fiabilidad de los resultados obtenidos. Si la muestra no fuera aleatoria, los resultados podrían verse afectados por sesgos y no serían representativos de la población en general. Del mismo modo, si se utilizara un método de muestreo más complejo o no adecuado, la selección de los participantes podría estar sesgada y los resultados no serían fiables.

Resultados.

Como conclusión, al negocio le conviene incentivar mucho a sus usuarios a escribir sobre películas y concentrarse en cómo se sintieron al verla. Ya que esto les puede servir mucho para recopilar mucha más información y de esta manera fortalecer los modelos que se presentaron anteriormente, ya que la cantidad de datos es muy poca y no está muy clara. Además, el Negocio debería considerar la limitación de caracteres en sus comentarios ya que muchos están desbalanceados en la longitud de sus textos.

Finalmente, Las métricas que más se acercaron a lo deseado fueron los modelos de Random Forest con una precisión del 81,60% y Multinomial Naive Baye con una precisión de 83%. Ambas opciones beneficiarían al negocio para determinar qué tipo de comentario se realizó. Esta información les podrá ayudar a clasificar de manera más segura cuales comentarios positivos se quieren mostrar y si se muestran o no los negativos. Además, muchas veces en los comentarios negativos se encuentra contenido no sensitivo para todo el público.

PROYECTO – ETAPA 2

Repartición de puntos:

- Juan David Bautista: 33

- Daniel Zambrano: 33
- Juan Diego Lugo: 33

Proceso de automatización del proceso de preparación de datos

Para el preparación y entendimiento de los datos se trabajó con la misma metodología de la etapa 1. Es decir, no añadimos ni quitamos más columnas, definimos las stopwords, usamos tokenizer para dividir las palabras en tokens y seleccionamos el mismo modelo para vectorizar (TfidfVectorizer).

Algoritmo a realizar: Random Forest

Construcción del Modelo:

Descripción de la función: La función tiene como objetivo crear una aplicación web de análisis de sentimientos que permita a los usuarios enviar reseñas de películas y recibir predicciones sobre el sentimiento expresado en sus reseñas (positivo o negativo). La aplicación se construirá utilizando Angular para el front-end, FastAPI para el back-end y un pipeline de Scikit-learn para realizar predicciones.

Requisitos: Para implementar esta función, se deben cumplir los siguientes requisitos:

Angular CLI y Node.js para el desarrollo front-end

FastAPI y Uvicorn para el desarrollo back-end

Scikit-learn y bibliotecas relacionadas para el entrenamiento y predicción del modelo

Python 3.x para ejecutar la aplicación FastAPI

Un conjunto de datos de reseñas de películas y sus correspondientes etiquetas de sentimiento para entrenar el pipeline de Scikit-learn

Arquitectura del sistema: La arquitectura del sistema consta de tres componentes principales:

Aplicación front-end Angular: proporciona una interfaz de usuario para enviar reseñas de películas y mostrar las predicciones de sentimiento.

Aplicación back-end FastAPI: maneja las solicitudes entrantes de la aplicación Angular, ejecuta el pipeline de Scikit-learn para realizar predicciones y devuelve los resultados de las predicciones.

Pipeline Scikit-learn: procesa los datos de entrada y realiza predicciones de sentimiento utilizando un modelo de aprendizaje automático entrenado.

Pasos de implementación:

a. Preparación de datos y entrenamiento del modelo: preprocesar el conjunto de datos de reseñas de películas limpiando el texto, tokenizando las reseñas y convirtiéndolas en representaciones numéricas. Entrenar un modelo de aprendizaje automático utilizando el pipeline de Scikit-learn y guardar el pipeline entrenado en un archivo con joblib.

b. Configuración de la aplicación FastAPI: instalar FastAPI y Uvicorn, crear un archivo app.py y definir el punto final /predict que acepta una reseña de película como entrada, ejecuta el pipeline de Scikit-learn para realizar una predicción y devuelve el resultado de la predicción.

c. Carga y uso del pipeline de Scikit-learn: cargar el pipeline guardado en la aplicación FastAPI y asegurarse de que todas las funciones o clases personalizadas, como el tokenizador, estén importadas y disponibles en la aplicación FastAPI.

d. Aplicación front-end Angular: crear una aplicación Angular que se comunique con el back-end FastAPI. Implementar un formulario para enviar reseñas de películas, enviar solicitudes al punto final /predict y mostrar los resultados de predicción recibidos al usuario.

Manejo de errores: implementar el manejo de errores para problemas potenciales, como datos de entrada no válidos, carga del pipeline de Scikit-learn y errores de comunicación entre la aplicación Angular y el back-end FastAPI.

Pruebas y validación: validar la implementación y funcionalidad de la función mediante pruebas unitarias, pruebas de integración y pruebas de extremo a extremo.

Implementación: implementar la aplicación front-end Angular en un proveedor de alojamiento estático, como Netlify

Acceso por medio de API

Para esta parte es menester aclarar que hemos desarrollado una aplicación web, con su Front-End y Back-End respectivos, en las los frameworks Angular y Java-

Springboot. Para que un usuario tenga acceso a la información de nuestro modelo o quiera subir una reseña, nuestro Api, es un Api rest que consulta y guarda infomación en una base de datos PostgresSQL atravez del framework Springboot. Por esto mismo el Usuario deberá registrarse en la página web y poder observar allí los comentarios que se han realizado y realizar reseñas nuevas. Para que nuestro modelo los clasifique. El API se conectará …...

Descripción del usuario/rol:

Para efectos de calidad en nuestra aplicación se crearon dos usuarios con caracterizcas diferentes. El primero siendo el usuario comentador quien va a colocar reviews a las peliculas que el quiera, y va a ayudar a entrenar nuestro algoritmo para que sea más robusto, además, recopilaremos sus datos para poder generar perfiles y vender esta información a nuestro segundo tipo de usuario. El segundo usuario es el cineasta, el cual podrá observar todas las estadistica y reseñas que se publiquen en nuestra página.

Esta aplicación es muy útil para recopilar y recolectar todos los datos con los que un cineasta quiere trabajar,también, es amigable con el reseñador ya que genera un ambiente de libertad de expresión.

Descripción de la aplicación

La aplicación tiene dos modos de registrarse uno para cineastas y otro para reseñadores, en el momento que se registran ambos roles pueden ver las reseñas que han escrito y podran ver la clasificación del modelo.