Proyecto 1- Inteligencia de Negocios.

Juan David Bautista.

Juan Diego Lugo.

Daniel Zambrano.

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. Organización y rol dentro de ella que se beneficia con la oportunidad definida
   Empresas de la industria del cine y entretenimiento:
   - Productoras
   - Plataformas de contenido

   Consumidores:
   - Cineastas
   - Clientes empresariales (Cinemas, Aerolíneas, etc.)

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.

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.

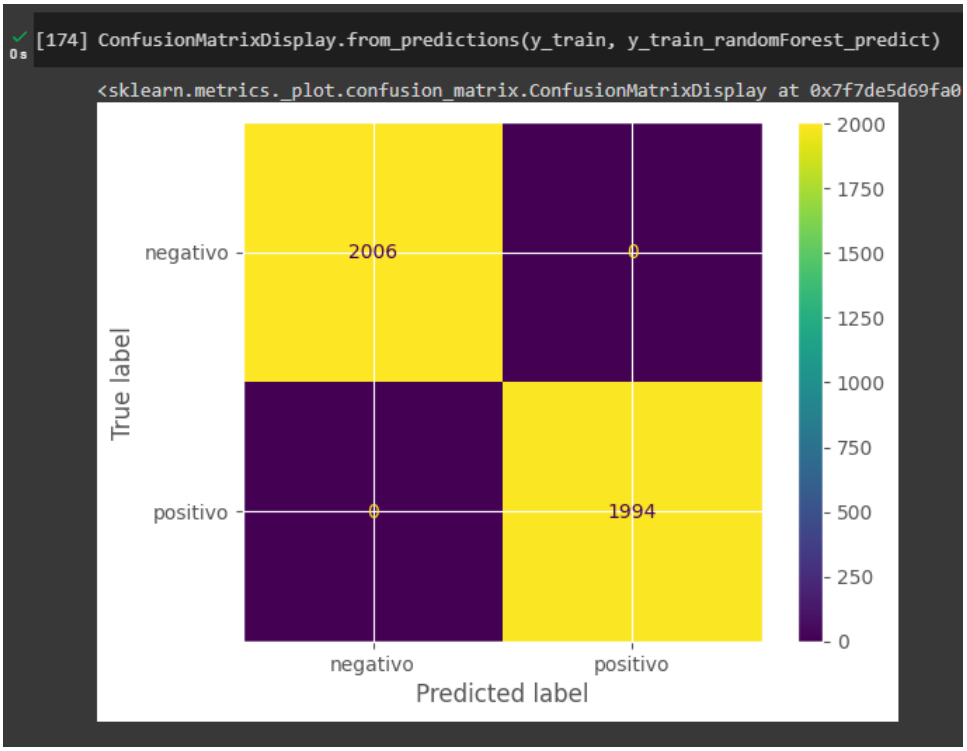

Imagen 1

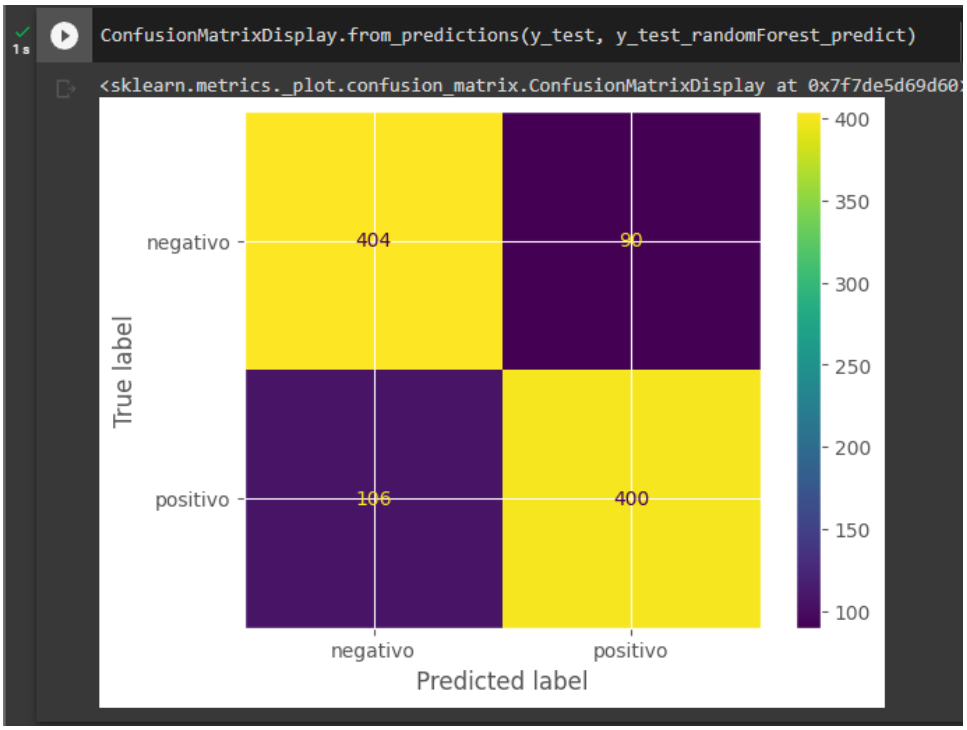

Imagen 2

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.

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.