Documento Proyecto Analítica de texto – Entrega 1

Entendimiento del negocio y enfoque analítico:

Oportunidad/problema Negocio	La oportunidad de negocio es que las empresas puedan obtener una retroalimentación de los usuarios de una página de películas o series, ya que al no saber cómo analizar o darles una clasificación a estos comentarios o "reviews", puedan construir una herramienta que les permita categorizarlos a fin de mejorar el servicio.
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático)	El objetivo analítico en este caso es que, mediante un modelo de machine learning se puedan clasificar los comentarios en las categorías de positivo o negativo basándose en las palabras que se presenten dentro de los mismos.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	Esta oportunidad beneficia a compañías en la industria del cine, tales como Cine Colombia, Netflix, Cuevana entre otras, ya que pueden aplicar el mismo proceso que nosotros realizaremos a sus comentarios y así realizar un análisis sobre estos, logrando el enfoque analítico.
Técnicas y algoritmos para utilizar	En primer lugar, se realizará un preprocesamiento de la la muestra para obtener más información acerca de estas reviews, y de esta manera producir una serie de tablas preparadas que alimente modelos de machine learning para el correcto entendimiento de la informacion recolectada. (Eliminación del ruido, Tokenización, Normalización, Transformación de campos, etc.) Y como algoritmos haremos uso de 3: Random Forest, GaussianNB y AdaboostClassifier.

Entendimiento y preparación de los datos:

El conjunto de datos está conformado por dos columnas principales:

Review_es: Contiene el texto de la reseña o comentario que realizaron acerca de una película. Tipo String y 5000 filas no nulas.

Sentimiento: Contiene el sentimiento asociado a la reseña o comentario y es negativo o positivo. Tipo String y tiene 5000 filas no nulas.

El conjunto de datos es un conjunto de reviews de películas que tienen las siguientes características:

- Duplicados: El conjunto de datos contiene solo 2 filas duplicadas, por lo tanto, no se ve tanta importancia en quitarlas, ya que existen 4000 datos en total en el dataset del train.

- Alta cardinalidad: La columna de review tienen un total de cardinalidad de 3999 lo cual nos indica que a excepción de una review todas las reviews son distintas

- Valores faltantes: No hay valores faltantes en ninguna de las dos columnas.

- Distribución uniforme: La columna de texto de la reseña está uniformemente distribuida, lo que indica que cada valor único aparece aproximadamente la misma cantidad de veces en el conjunto de datos.

El preprocesamiento de datos es una etapa importante en el procesamiento de lenguaje natural, ya que permite transformar los datos textuales en un formato más manejable y útil para el análisis. Esta tarea se realiza en tres etapas principales:

1. Eliminación del Ruido.
2. Tokenización.
3. Normalización.

La eliminación del ruido implica la eliminación de cualquier información que no sea relevante para el análisis, como etiquetas HTML, símbolos de puntuación, números, caracteres especiales, etc. Esto ayuda a reducir la cantidad de datos que el modelo tendrá que procesar y puede mejorar la precisión del análisis.

La tokenización se refiere al proceso de dividir el texto en palabras individuales o tokens. Esto ayuda a identificar las palabras clave y los patrones de lenguaje que pueden ser relevantes para el análisis. También ayuda a convertir el texto en un formato más estructurado que el modelo puede procesar.

La normalización se utiliza para hacer que el texto sea más uniforme en términos de formato y estilo de escritura. Esto puede incluir la eliminación de mayúsculas y minúsculas, la eliminación de palabras vacías, la corrección ortográfica, la lematización, la eliminación de prefijos y sufijos, entre otras técnicas. La normalización ayuda a reducir la complejidad del texto y a garantizar que los patrones de lenguaje se puedan identificar de manera más efectiva.

En conjunto, estas tres etapas son cruciales para preparar los datos de texto para el análisis. Al realizar estas tareas, se pueden eliminar ruidos innecesarios, dividir el texto en unidades más manejables y uniformizar el formato para que se pueda analizar con mayor precisión.

Además, se realiza un proceso denominado transformación de campos y esto permite en pocas palabras convertir las opiniones de los usuarios preprocesadas en un vector

de características numéricas. Específicamente permite convertir una colección de documentos de texto en una matriz de recuento de términos/documentos. Cada columna de la matriz representa una palabra en el vocabulario y cada fila representa una opinión de un usuario. Los valores en la matriz son el recuento de ocurrencias de cada palabra en cada opinión de usuario.

Finalmente, al momento de crear un modelo se dice que la preparación de datos es una etapa crucial para poder obtener un modelo preciso y efectivo. Es por eso que se utilizan herramientas como pipelines, que permiten automatizar y estandarizar el proceso de preprocesamiento de los datos y entrenamiento del modelo.

En este caso, utilizaremos nuestro "preparador" de datos que hemos construido previamente, el cual realiza una serie de tareas como tokenización, normalización, eliminación de ruido, lematización y eliminación de prefijos y sufijos. Luego, lo incluiremos en el pipeline para poder realizar la búsqueda del mejor modelo.

El pipeline nos permitirá unificar todo el proceso de preprocesamiento de los datos y entrenamiento del modelo, lo cual es especialmente útil cuando trabajamos con múltiples modelos y queremos comparar su rendimiento. Asimismo, al estandarizar el proceso, nos aseguramos de que los datos sean tratados de manera consistente y reducimos el riesgo de errores o de introducir sesgos en el proceso.

De esta manera, al utilizar el pipeline junto con nuestro preparador de datos, podremos realizar de manera eficiente la búsqueda del mejor modelo, que nos permitirá predecir con mayor precisión el sentimiento expresado en las reseñas de los clientes.

Modelado y evaluación: Para la construcción de los modelos se plantearon 3 posibles modelos que son de los principales de cara al contexto de procesamiento de texto y predicciones con esta (Todos de la librería sklearn):

- Random Forest
- Naive Bayes
- AdaBoostClassifier

Como se mencionó anteriormente se utilizaron pipelines para la construcción de estos modelos y se realizó un proceso de gridsearch para encontrar los mejores hiperparámetros posibles para estos.

Por lo tanto, se realiza una búsqueda de hiperparámetros para encontrar el transformador más afín al modelo y a los datos y se escoge entre tres: CountVectorizer(binary=True,lowercase=False), CountVectorizer(lowercase=False), T fidfVectorizer(lowercase=False).

Luego, de tener escogido el transformador, se realiza una búsqueda de hiperparámetros específicos para cada modelo.

Finalmente, se prueban con los datos de prueba el mejor modelo encontrado para cada tipo de modelo (en total tres) para poder de cara a estos resultados de accuracy escoger el modelo para predecir los sentimientos de los comentarios de las películas.

Resultados:

El resultado de la ejecución de todos los modelos fue el siguiente:

Modelo	Acc Training	Acc Test
Random Forest	100%	84%
GaussianNB	80%	58%
AdaBoostClassifier	100%	80%

Lo que nos permite concluir que el modelo que mejor se adapta a la muestra otorgada es el sistema Random Forest, debido a que es el que nos otorga una mayor precisión en la obtención de datos y clasificación de los mismos. Por este motivo será la opción final para cumplir los objetivos impuestos en este trabajo.

Para el largo plazo, se recomienda hacer un tratamiento de datos en los cuáles se eliminen las palabras que tienen una importancia parecida tanto para comentarios positivos como negativos y así poder construir modelos con un mejor rendimiento.

Describir los roles y las tareas realizadas por cada integrante del grupo

Actividad	Encargado	Tareas
Líder de proyecto	María José Cely	1. Informar a los integrantes las fechas y horas de las reuniones 2. Revisar el cumplimiento de las tareas asignadas 3. Toma de decisiones
Líder de negocio	Javier Serrano	1. Responsable de velar por la resolución del problema 2. Garantiza que el producto se puede comunicar de forma apropiada 3. Encargado de contactar al grupo de expertos en estadística para la evaluación del modelo
Líder de datos	Alejandro González	1. Gestionar los datos utilizados para el proyecto 2. Definir las tareas que cada integrante tiene respecto a los datos 3. Dejar los datos disponibles para el grupo
Líder de analítica	María José Cely	1. Revisión y gestión de analítica 2. Selección del mejor modelo de acuerdo con resultados del trabajo

Reunión de lanzamiento y planeación:	María José Cely	1. Definición de roles 2. Pactos de grupo 3. Definición de tiempos 4. Ideas para el trabajo
Reunión de ideación:	María José Cely	1. Revisión de tareas asignadas 2. Asignación de nuevos entregables 3. Comunicación con el chico de Estadística 4. Corrección de errores
Reuniones de seguimiento	María José Cely	1. Revisión de cumplimiento de asignaciones 2. Revisión de faltantes
Reunión de finalización	María José Cely	1. Revisión del proyecto completo 2. Revisar que se puede mejorar

Encargado	**Algoritmo**	**Tiempo**	**Puntaje**
María José Cely	Random Forest	3	33.33
Javier Serrano	GaussianNB	3	33.33
Alejandro González	AdaBoostClassifier	3	33.33

Puntos para mejorar para próximos trabajos:

1. Iniciar con más tiempo el trabajo
2. Coordinar de mejor manera los tiempos que cada uno de los integrantes dispone
3. Entender los problemas, planes e inconvenientes de las otras personas
4. Mejorar la comunicación grupal para evitar inconvenientes

Retos y soluciones presentados en el proyecto:

1. Estuvimos un poco cortos de tiempo, realmente la solución a este problema fue que se corrió la entrega por lo cual nos permitió terminar con éxito el trabajo
2. No le corría el programa a María José Cely, pero Javier Serrano le hizo el favor de verificar si estaba funcionando lo que hizo.
3. Creo que pueden ser los modelos presentados, que además de que solucionan un problema de que, al tener la entrada de cualquier comentario de cualquier usuario sobre una película, este comentario se podría clasificar positivo o negativo y así poder tomar acciones para el mejoramiento de una app o página web.