Proyecto 1

Inteligencia De Negocios

Jairo Nicolás Gómez
Paola Campiño
Felipe Duque

Análisis de Sentimientos en

Películas Abril 2 de 2023

Link a repositorio: https://github.com/Paolaaaaaa/Proyecto_1_Bi.git

Etapa 1: Construcción de modelos de analítica de textos.

1.1 Entendimiento del negocio y enfoque analítico

Oportunidad/problema Negocio	La oportunidad que vemos en este proyecto es poder ver y analizar las opiniones de los espectadores para mejorar los factores que son más criticados en las películas y mantener los aspectos más positivos y que son mejor resaltados por los espectadores.
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático)	Llegar a obtener un modelo con una alta precisión para analizar las reseñas sobre las películas y lograr clasificar de la mejor manera si son negativas o positivas.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	Este modelo podría llegar a tener un beneficio en empresas como lo puede ser cineColombia, Netflix y plataformas de streaming de películas o incluso páginas web de reseñas. Pues de esta forma muy rápidamente se podría llegar a una idea de cómo se ha recibido las películas entre el público para tomar decisiones de negocio como quitarla de taquilla, no recomendarla. En un cierto punto se podría hacer un análisis detallado de como la opinión de las personas podría cambiar en cuanto a una película al paso del tiempo.
Técnicas y algoritmos a utilizar	Utilizaremos una técnica de clasificación y los algoritmos de regresión logística, arboles de decisión y SVM (máquinas de vectores de soporte).

1.2 Modelado y evaluación:

a. Regresión logística (Felipe Duque):

El modelo de regresión logística para análisis de textos es una técnica de aprendizaje supervisado que se utiliza para predecir la probabilidad de que un texto

pertenezca a una o más categorías predefinidas. Este modelo es especialmente útil para el análisis de texto, ya que permite modelar la relación entre las palabras o características de un texto y la probabilidad de que ese texto pertenezca a una determinada categoría.

El modelo de regresión logística utiliza una función logística para modelar la relación entre las características de entrada del texto y la variable de salida categórica. Esta función logística transforma una variable continua en una variable binaria, es decir, la probabilidad de que el texto pertenezca a una categoría determinada.
El modelo de regresión logística se utiliza en análisis de textos porque permite la clasificación de grandes volúmenes de datos de texto en diferentes categorías de manera rápida y precisa. Por ejemplo, en este caso se puede utilizar para analizar opiniones de los clientes sobre una película.

b. Resultados: Regresión logística (Felipe Duque):

Luego de entrenar el modelo de regresión logística con el 70% (3500 registros) de los datos proporcionados fue testeado con el 30% restante (1500 registros) y lo primero que generamos es una matriz de confusión que arrojo los siguientes datos:

- El modelo predijo 594 registros positivos correctamente.
- El modelo predijo 668 registros negativos correctamente.
- El modelo predijo 143 registros positivo erróneamente.
- El modelo predijo 95 registros negativos erróneamente.

A partir de lo anterior encontramos las siguientes métricas cuantitativas para la evaluación de la efectividad del modelo:

- Para clasificar comentarios negativos obtuvimos:
    - Precisión: 86%
    - Recall: 81%
    - F1-score: 83%
    - 737 registros
- Para clasificar comentarios positivos obtuvimos:
    - Precisión: 82%
    - Recall: 88%
    - F1-score: 85%
    - 763 registros

Finalmente, en promedio el modelo tiene una precisión del 84%, un recall del 84% y un f1-score del 84% lo cual es indicativo de que se ha logrado un buen modelo de clasificación para categorizar los sentimientos sobre las películas entre positivo y negativo.

```
Informe de clasificación:  precision recall f1-score support

 0 0.86 0.81 0.83 737  1 0.82 0.88 0.85 763
```

```
 accuracy 0.84 1500  macro avg 0.84 0.84 0.84 1500 weighted avg 0.84
0.84 0.84 1500
```

a. Árbol de decisión (Paola Andrea Campiño):

Los arboles de decisión es un algoritmo de clasificación en un aprendizaje supervisado no paramétrico. La solución plantea un árbol, que consta de nodo raíz, ramas, nodos y hojas. Los nodos hojas representan los resultados posibles dentro del conjunto de datos, donde hay un numero de reseñas y un numero de entropía. Cabe resaltar que entre menor sea dicho número mejor son la clasificación de las reseñas.

b. Resultados: Árbol de decisión:

Para la implementación del árbol de decisión por este algoritmo se planteó 3 diferentes modelos. El primero con un 0.691 de presición en la que se define el número mínimo de elementos para generar un nodo de 2 elementos, usando como métrica la entropía y una altura por default. Las métricas relacionadas a este modelo se encuentran a continuación:

```
precision recall f1-score support

 0 0.68 0.73 0.70 498
 1 0.71 0.65 0.68 502

 accuracy 0.69 1000
 macro avg 0.69 0.69 0.69 1000
weighted avg 0.69 0.69 0.69 1000
```

Para el segundo modelo se intentó encontrar el mejor árbol posible del que salió que los parámetros deberían ser que la métrica sería entropía, la altura sería de 28 y que el número mínimo de elementos debía estar 4 del que si se llegó a las siguientes métricas:

```
precision recall f1-score support

 0 0.66 0.80 0.72 498
 1 0.75 0.59 0.66 502

 accuracy 0.69 1000
 macro avg 0.70 0.69 0.69 1000
weighted avg 0.70 0.69 0.69 1000
```

En base a lo anterior se puede decir que no es muy bueno usar árbol de decisión como modelo para hacer la clasificación, pues el porcentaje de error es de un 30% que es considerablemente alto.

Y Finalmente para el 3r modelo se hizo una prueba con todos los parámetros por

default los cuales, no generaron muchos cambios en términos del porcentaje de error. En si se considera que al hacer un modelo con árboles de decisión se puede notar que la precisión sigue estando en un 69% y un porcege de error del casi 31%.

a. SVM(Jairo Nicolás Gómez):

El algoritmo de máquinas de vectores de soporte es un algoritmo de aprendizaje supervisado, el cual se usó para la clasificación binaria y poder separar de la mejor forma posible dos clases diferentes de puntos de datos. Lo anterior con el objetivo de negocio en cuenta para mirar las películas con comentarios positivos y negativos, y ver cómo mejorar para un futuro con base en la retroalimentación de los espectadores. Después de hacer todo el proceso de entendimiento de datos y limpieza, y crear la matriz de palabras con el countvectorizer, se realizó un modelo el cual en la primera iteración arrojo resultados muy buenos, tomando 80% de registros como entrenamiento y 20% de registros de prueba.

b. Resultados SVM (Jairo Nicolás Gómez):

Lo anterior arrojo un 82% de precisión por lo que no fue necesario hacer un segundo modelo

```
Classification Report:  precision recall f1-score support

 0 0.83 0.80 0.81 485  1 0.82 0.84 0.83 515

 accuracy 0.82 1000  macro avg 0.82 0.82 0.82 1000 weighted avg 0.82
0.82 0.82 1000
```

En esta tabla se observa que:

Precisión: En ambas clases es del 82%, lo que indica que el modelo tiene buena capacidad para clasificar las instancias

Recall: El recall promedio es de 82%, lo que indica que el modelo tiene buena capacidad para detectar correctamente las instancias positivas

F1-Score: El modelo tiene un buen equilibrio entre la precisión y el recall pues es del 82% también.

1.3 Resultados en base a todos los modelos

En base a los resultados de los modelos anteriores podemos concluir que lo mejor sería usar el modelo de regresión logística, ya presenta un menor porcentaje de error cerca del 16% y una precisión de casi del 84%.

Cabe resaltar que las pruebas se hicieron con una muestra 1000 reseñas sobre la población, por lo que es muy posible que teniendo un mayor número de datos podría llegar a una precisión mayor. Es muy posible que sea necesario mejorar la limpieza de datos de manera que podamos llegar un error menor lo cual garantizaría una mayor confiabilidad sobre el modelo planteado.

Si bien el porcentaje de error es considerablemente alto, es importante enfatizar que se puede mejorar el modelo y a futuro podría llegar a ser de valor dentro de la organización de manera que se pueda tomar decisiones en cuanto a la recepción de una película nueva, incluso podría llegar a ser de valor para el descubrimiento de películas que podrían llegar a ser buenas.

1.4 Acta de reunión con la experta en estadística:

<table>
<tr><td colspan="2">Acta de Reunión</td></tr>
<tr><td>Fecha: 1/04/2023
Hora Inicio: 6:51pm</td><td>Hora de Finalización:
7:16pm</td></tr>
<tr><td>Asistentes:

Paola Andrea Campiño (Grupo 26)
Jairo Nicolas Gomez Mendoza (Grupo 26)
Valery Sharith Fonseca Pana (Experta en estadística)</td><td>Orden Reunión:

1. Presentación
2. Contexto
3. Presentar modelos
4. Presentación de métricas
5. Evaluación de métricas.</td></tr>
<tr><td colspan="2">Observaciones/Recomendación:</td></tr>
<tr><td colspan="2">- Hay que mejorar la interpretación tabla confusión, expresarlo en % de error. - En este caso se habla de una muestra y no una población. (Dentro de una población se seleccionó 5000 muestras)
- Hablar del margen de error.</td></tr>
<tr><td colspan="2">Firma:
Paola Andrea Campiño (Grupo 26)
Jairo Nicolas Gomez Mendoza (Grupo 26)
Valery Sharith Fonseca Pana (Experta en estadística)</td></tr>
</table>

1.5 Trabajo en Grupo:

01	Paola Campiño	Entendimiento de los datos
02	Paola Campiño	Limpieza de datos
03	Grupal	Modelos
04	Paola Campiño Jairo Nicolas	Reunión con experta

Etapa 2: Automatización y uso de modelos de análitica de textos

El objetivo de esta etapa se centra en el proceso de despliegue de la solución analítica en el ambiente de producción de una organización:

2.1 Proceso de automatización del proceso de preparación de datos, construcción del modelo, persistencia del modelo y acceso por medio de API:

En base al objetivo anterior planteado, para el desarrollo de esta aplicación se planteó que el valor que se pudiera dar a los clientes estaría relacionado con los modelos de análisis de sentimientos producidos en la etapa 1. En base a el resultado de los modelos anteriores se optó por tomar el modelo con mayor precisión para realizar este análisis que en este caso en concreto sería **regresión logística** con el que llegamos a una **precisión del 84%**.

Teniendo esta primera base se optó por generar pipelines que nos permitiera automatizar el modelo. De esta forma solo sería cuestión que cliente proporcionará la información en formato csv con al menos una columna con el nombre **review-es** que garantiza una respuesta de 1 y 0 s representando la reseña que sería positiva o negativa. En base a esto se ha generado un pipeline usando la librería de python Joblib que permite paralelizar operaciones complejas. Es así como el archivo **modelo.joblib** surge, en base al desarrollo de analitica de datos usando **regresión logística**.

Así mismo comienza a hablar de una aplicación la cual es desarrollada con FastApi , con el cual se desarrollan diferentes rutas con las que podamos interactuar con posibles usuarios. Entre las más importantes están **/index/** que desde ahí se genera un POST con el csv de los datos. El POST mencionado anteriormente es recibido desde la url **/upload-csv/** que nos permite recibir los csv persistirlos con el fin de proporcionar información haciendo uso de la

función **use_pipeline()** la cual toma los datos y usa el pipeline para devolver una lista con la clasificación de cada una de las reseñas [1..0] en donde 1 es positivo y 0 es negativo.

2.2 Desarrollo de la aplicación y justificación

Este aplicativo está dedicado a organizaciones que buscan mejorar la experiencia de usuario al momento de generar recomendaciones de una película, incluso a aquellas empresas que se dedican mostrar películas en cine de manera que de manera fácil se pueda llegar a traer al cine no solo peliculas nuevas sino que también películas que puedan llegar a ser buenas. Podría llegar a tener un rol en la toma de decisiones de negocio en cuanto al contenido/ servicios que se van a proveer en este caso qué películas se van a llegar a mostrar en taquilla.
Debido a que el manejo de archivos csv es mucho más fácil en computador, se afirma que la aplicación es exclusivamente web. Sin embargo, se ha considerado llegar a hacer que esta aplicación pueda llegar a ser mucho más que un analizador de sentimientos si no que también se considera que podría llegar a ser una página que agrupa reseñas en donde se pueda llegar a obtener información propia por medio de comentarios. Esto podría llegar a abrirnos a muchos más tipos de clientes, incluso a aquellos que no son empresas que ni tienen datos propios en relación a películas específicas. En un futuro esta aplicación podría llegar a ser móvil y mucho más apta para muchos más tipos de usuarios.

4. Acta de reunión con la experta en estadística:

<table>
<tr><td colspan="2">Acta de Reunión</td></tr>
<tr><td>Fecha: 29/04/2023
Hora Inicio: 3:30pm</td><td>Hora de Finalización:
4:00pm</td></tr>
<tr><td>Asistentes:

Paola Andrea Campiño (Grupo 26)
Valery Sharith Fonseca Pana (Experta en estadística)</td><td>Orden Reunión:

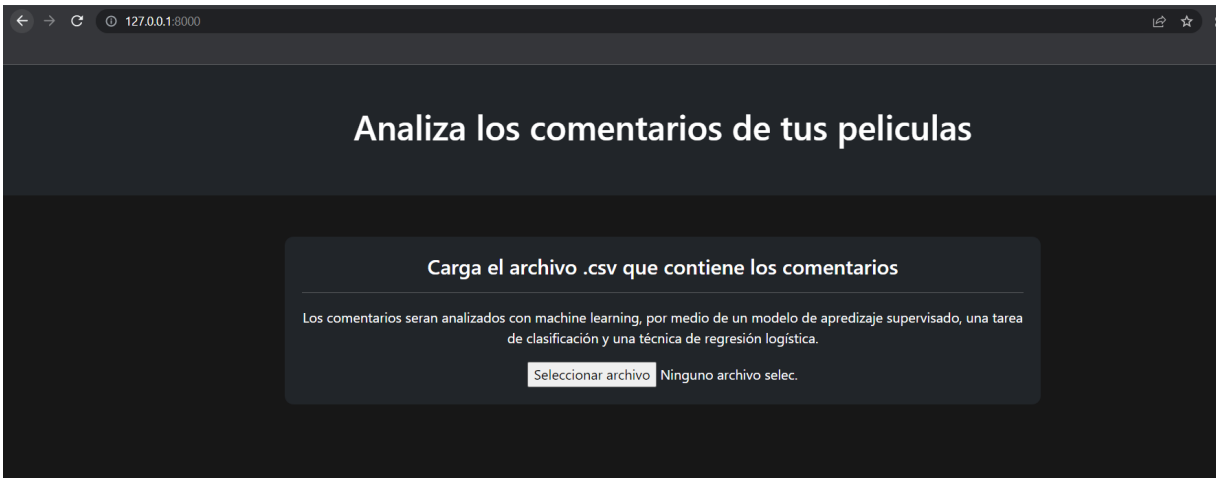
1. Actualización y precisión del modelo.

2. Evaluación del aplicativo.</td></tr>
<tr><td colspan="2">Observaciones/Recomendación:</td></tr>
<tr><td colspan="2">- Trabajar en pruebas por hipótesis, de manera que podamos validar un poco mejor el modelo.
- Mejorar el diseño de manera que sea mucho más claro que es lo que hace la página.</td></tr>
</table>

Firma:
Paola Andrea Campiño (Grupo 26)

2.3 Resultados y desarrollo de la aplicación y justificación:

Después de realizar todo este proceso se ha llegado a esta página, donde se pide al usuario que este suba un archivo.csv.



.
Después de unos segundos, esta página nos va a redirigir a gráficos que muestran el resultado del análisis:

Así mismo se puede ver la documentación/ este documento:

Validaciones por hipótesis:

En base a las recomendaciones se ha optado por validar por hipótesis por lo que se ha tomado opiniones muy exactas y poco generales para validar en qué puntos puede fallar. A continuación se puede ver un acercamiento a estas pruebas, aquellas que están en rojo son casos en los que el modelo está fallando ya que su respuesta no concuerda con la hipótesis planteada:

H: Hipotesis
R: Resultado del modelo

H0 'Muy buena película' = positiva
R0: 1 equivalente a positiva.

H1 'Me gusto mucho'= positiva
R1: 1 equivalente a positiva.

H2 'No me gusto la pelicula'= Negativa
R2: 1 equivalente a positiva.

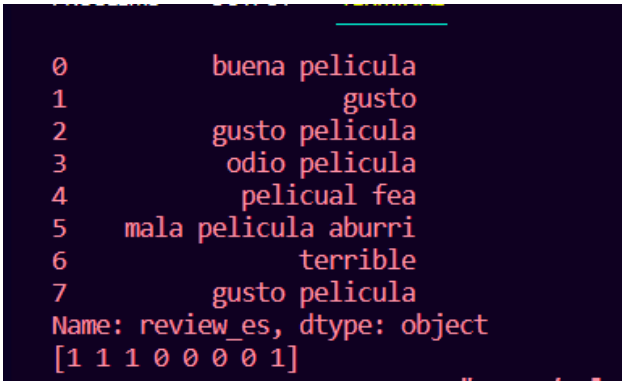
H3 'odio esta pelicula'= Negativa
R3: 0 equivalente a Negativa.

H4 'una pelicual muy fea'= Negativa
R4: 0 equivalente a Negativa.

H5 'mala pelicula me aburri' = Negativa
R5: 0 equivalente a Negativa.

H6 'terrible' = Negativa
R6: 0 equivalente a Negativa.

H7 'No me gusto mucho la pelicula' = Negativa
R7: 1 equivalente a positiva.



Por medio de las palabras que se analizan y las hipótesis se pudo concluir que las palabras como no pueden llegar a ser concluidas como stopwords por lo que han llegado a ser eliminadas del análisis y a llevar a resultados erróneos. Por lo que para la siguiente construcción del modelo se tiene que intentar buscar una manera más correcta de eliminar stopwords que podrían llegar a tener más significado.