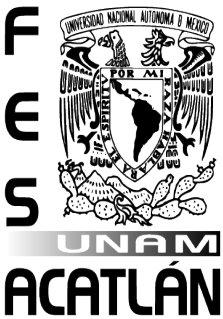

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FES ACATLÁN

ASIGNATURA

DIPLOMADO EN CIENCIA DE DATOS
MODULO IV: TÉCNICAS COGNITIVAS E INTRODUCCIÓN A BIG DATA

PROYECTO

IMPLEMENTACIÓN DE REDES NEURONALES ARTIFICIALES EN UN PROBLEMA DE CLASIFICACIÓN DE OPINIONES DE CLIENTES DE COMERCIO ELECTRÓNICO.

PROFESOR

LIC. ÓSCAR DANIEL ACOSTA GONZÁLEZ

ALUMNO

CRUZ RAMÍREZ NICOLÁS

FECHA DE ENTREGA

31 DE JULIO DE 2022

Índice general

Capítulo 1

Introducción y planteamiento del problema

1.1. Comercio electrónico

El comercio electrónico es un modelo de negocio basado en la compra, venta y comercialización de productos y servicios a través de medios digitales (paginas web, redes sociales, entre otros).

Mediante estas herramientas, los clientes pueden tener mayor acceso a los productos y/o servicios sin importar el lugar y el momento en el que se encuentren. La mayor parte del comercio electrónico consiste en la compra y venta de productos o servicios entre personas y empresas. Entre las ventajas del comercio electrónico podemos encontrar:

- Dar a conocer la marca y tener mayor oportunidad de ventas.
- Diversificar la oferta de productos y/o servicios.
- Contar con un horario comercial las 24 horas del día, los 7 días de la semana sin limitaciones geográficas.
- Personalizar la comunicación con los clientes y diseñar estrategias de ventas específicas para atender sus necesidades.
- Dar atención a diversos tipos de clientes al mismo tiempo.
- Ofrecer a los clientes diferentes formas de pago.
- Implementar y desarrollar estrategias de marketing enfocadas al tipo de clientes a través de descuentos, cupones, promociones especiales.
- Recolección de datos, ya que en internet recolectar información de los clientes es más sencillo que en el comercio tradicional.

Si los clientes del comercio electrónico se encuentran satisfechos con los servicios y la calidad de los productos también pueden recomendarlos, ya sea en plataformas de opinión o entre sus conocidos.

Por otra parte, también existen ciertas desventajas que pueden provocar una mala experiencia de compra para el cliente, entre las cuales están:

- Fallos en el sitio web.
- Seguridad en la forma de pago.
- Proceso de envío.
- Mercado más competitivo.
- El cliente no puede probar el producto antes de comprarlo.

Conociendo estas ventajas y desventajas los empresarios pueden actuar en consecuencia y diseñar estrategias de negocio que limiten los inconvenientes y potencien los puntos fuertes.

1.2. Planteamiento del problema

La compañia ABC es una pequeña empresa dedicada al comercio electrónico, específicamente a la venta de ropa. Para monitorear el nivel de satisfacción de sus clientes se optó por la siguiente estrategia: después de que uno de sus clientes realiza una compra, se les invita a llenar un pequeño formulario donde pueden compartir sus opiniones sobre su experiencia de compra y el producto, además, una de las preguntas importantes para la empresa es conocer si el cliente recomendaría comprar en la tienda en línea a otros usuarios.

Con esta idea en mente, se ha recabado la opinión de los clientes a lo largo de varios meses. Ahora se desea tomar como insumo los datos de tipo texto para predecir si el cliente se sintió satisfecho y recomendaría los productos del comercio.

1.3. Objetivos

Se plantean los siguientes objetivos como un paso inicial para mejorar la calidad del servicio:

- Realizar un análisis de sentimientos para clasificar la opinión de los clientes sobre su compra en línea.
- Obtener un modelo con un alto poder predictivo que nos permita predecir si el cliente recomendaría a otros usuarios realizar compras en línea en la tienda ABC.

Capítulo 2

Marco teórico y metodología

2.1. Redes Neuronales

Las redes neuronales artificiales son un modelo computacional evolucionado a partir de diversas aportaciones científicas que están registradas en la historia. Consiste en un conjunto de unidades denominadas neuronas artificiales, que se encuentran conectadas entre sí y que se transmiten señales. La información de entrada atraviesa la red neuronal, donde se somete a diversas transformaciones y finalmente produce diversos valores en la capa de salida.

Las redes neuronales son un modelo que tiene sus orígenes en la década de los 50's, pero que a partir del desarrollo de hardware sofisticado que permite realizar los cálculo necesarios, se han vuelto extremadamente poderosas a partir de la década de los 90's. A continuación se muestra una pequeña lista a manera de resumen con el desarrollo de las redes neuronales a través del tiempo:

- 1958 – Perceptron
- 1965 – Multilayer Perceptron
- 1980's
  - Neuronas Sigmoidales
  - Redes Feedforward
  - Backpropagation
- 1989 – Convolutional neural networks (CNN) / Recurent neural networks (RNN)
- 1997 – Long short term memory (LSTM)
- 2006
  - Deep Belief Networks (DBN): Nace deep learning
  - Restricted Boltzmann Machine
  - Encoder / Decoder = Auto-encoder
- 2014 – Generative Adversarial Networks (GAN)

2.2. Redes Neuronales Recurrentes: LSTM

Las Long short term memory son un tipo de Red Neuronal Recurrente. Esta arquitectura permite conexiones "hacia atrás" entre las capas. Esto las hace buenas para procesar datos de tipo temporal. En 1997 se crearon las LSTM que consisten en unas celdas de memoria que permiten a la red recordar valores por períodos cortos o largos.

Una celda de memoria contiene compuertas que administran como la información fluye dentro o fuera. La puerta de entrada controla cuando puede entran nueva información en la memoria. La puerta de "olvido" controla cuanto tiempo existe y se retiene esa información. La puerta de salida controla cuando la información en la celda es usada como salida de la celda. La celda contiene pesos que controlan cada compuerta. El algoritmo de entrenamiento -conocido como backpropagation-through-time optimiza estos pesos basado en el error de resultado. Las LSTM se han aplicado en reconocimiento de voz, de escritura, text-to-speech y otras tareas.

Cuando se realizan tareas de minería de texto, muchas tareas de preprocesamiento y modelado se enfocan en crear datos de manera secuencial. Ejemplos de estas tareas son la eliminación de stopwords, etiquetado gramatical (conocido como pos tagging). Estos son métodos que intentan volver a los datos más facilmente ."entendibles"por el modelo.

Dado que una LSTM puede memorizar una secuencia de datos, también puede eliminar información inutil, y dado que sabemos que los datos de tipo texto contienen mucha información que no será usada, esta puede ser eliminada por la LSTM por lo cual el tiempo de cálculo se reduce.

2.3. Metodología CRISP-DM

En la implementación de un proyecto de Ciencia de Datos exiten diferentes metodologías que nos permiten dividir nuestras actividades en etapas, organizar de mejor manera el trabajo y obtener resultados confiables. Dichas metodologías son sumamente útiles ya que nos permitirán tener un flujo de trabajo claro y los resultados serán replicables en cualquier otra situación. Entre las metodologías más populares se encuentra la metodología SEMMA creada por el SAS Institute y también la metodología CRISP-DM. Para este proyecto usaremos la metodología CRISP-DM la cual consta de las siguiente etapas:

- Entendimiento del negocio
- Entendimiento de los datos
- Preparación de los datos
- Construcción del modelo
- Evaluación
- Despliegue en producción

Las metodologías SEMMA y CRISP-DM comparten la misma esencia, estructurando el proyecto de en fases que se encuentran interrelacionadas entre sí, convirtiendo el proceso de Data Mining en un proceso iterativo e interactivo. La metodología SEMMA se centra más en las características técnicas del desarrollo del proceso, mientras que la metodología CRISP-DM, mantiene una perspectiva más amplia respecto a los objetivos empresariales del proyecto, por este motivo es la metodología elegida para la realización de este proyecto, a excepción de última estapa que consta del despliegue en producción.

Capítulo 3

Desarrollo

3.0.1. Entendimiento del negocio

Como se mencionó en la parte introductoria de este reporte, la empresa ABC se dedica a la venta de ropa en línea, y una de sus prioridades para continuar con su crecimiento es la satisfacción de sus clientes, ya que esto además de traer beneficios económicos, puede servir como publicidad ya que estos pueden recomendar la página de compras, y atraer la curiosidad de nuevos compradores. Además de que es mucho menos costoso conservar a los clientes habituales que realizar gastos de publicidad para atraer nuevos.

Con esta idea en mente, como primera aproximación, se desarrollo un formulario donde los clientes que realizan comprar en la tienda en línea pueden compartir su opinión sobre el proceso de compra, de la calidad de sus productos, su nivel de satisfacción, etc.

3.0.2. Entendimiento de los datos

Los datos con los que se trabajará se obtuvieron de Kaggle.com donde se pueden encontrar bajo el nombre "Women's E-Commerce Clothing Reviews". De acuerdo a la descripción de los datos, estos son reales pero se han enmascarado para mantener la anonimidad de los cleintes y la compañia. Además, cualquier referencia a la compañia que se hubiera incluido en el dataset, se ha reemplazado por la palabra retailer". La datos son de tipo no estructurado, ya que, si bien están contenidos en un archivo separado por comas, el contenido de las columnas es de tipo texto y numérico. En la figura 3.1 se puede observar la descarga de Kaggle.com.

La base consta de 23,486 filas y 10 columnas. Una de lsa columnas más importantes es la que contiene las opiniones de los clientes que realizaron compras en la tienda. También se recabó la edad de los clientes, el titulo de su opinión o review, así como el nombre del departamento y división de la tienda, el tipo de producto, etc. A continuación se presenta una descripción de las columnas con las que se cuenta:

- Clothing ID: Identificador de la prenda vendida.
- Age: Edad del cliente que ha realizado la compra.

- Title: Titulo de la reseña del cliente.
- Review Text: Cadena de texto con la opinión del cliente.
- Rating: Variable ordinal que representa la calificación que le da el cliente a su compra. Puede ser un entero del 1 al 5.
- Recommended IND: Variable binaria con la cual el cliente indica si recomienda el producto o no. Puede ser 1 o 0.
- Positive Feedback Count: Contador para cuantificar el número de pulgares arriba que recibió la contraseña del cliente.
- Division Name: Variable categórica para indicar la división del producto comprado.
- Department Name: Variable categórica para indicar el departamento a donde pertenece el producto comprado.
- Class Name: Variable categórica para indicar la clase de producto comprado.

Como buscamos una solución enfocada en el análisis de texto, para efectos de este breve trabajo se usará únicamente la columnas con la opinión de los clientes, y se tomará como variable objetivo si el cliente recomienda el producto o no.

	A	B	C	D	E	F	G	H	I	J	K
1		Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
2	0	767	33		Absolutely wonderful - silky and sexy and comfortable	4	1	0	Initmates	Intimate	Intimates
3	1	1080	34		Love this dress! it's sooo pretty. i happened to find it in a store, and i'm glad i did bc i n	5	1	4	General	Dresses	Dresses
4	2	1077	60	Some major design flaws	I had such high hopes for this dress and really wanted it to work for me. i initially ordered	3	0	0	General	Dresses	Dresses
5	3	1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, flirty, and fabulous! every time i wear it, i get nothin	5	1	0	General Petite	Bottoms	Pants
6	4	847	47	Flattering shirt	This shirt is very flattering to all due to the adjustable front tie. it is the perfect length to we	5	1	6	General	Tops	Blouses
7	5	1080	49	Not for the very petite	I love tracy reese dresses, but this one is not for the very petite. i am just under 5 feet tall	2	0	4	General	Dresses	Dresses
8	6	858	39	Cagrcoal shimmer fun	I aded this in my basket at hte last mintue to see what it would look like in person. (store p	5	1	1	General Petite	Tops	Knits
9	7	858	39	Shimmer, surprisingly goes with lots	I ordered this in carbon for store pick up, and had a ton of stuff (as always) to try on and u	4	1	4	General Petite	Tops	Knits
10	8	1077	24	Flattering	I love this dress. i usually get an xs but it runs a little snug in bust so i ordered up a size.	5	1	0	General	Dresses	Dresses
11	9	1077	34	Such a fun dress!	I'm 5'5" and 125 lbs. i ordered the s petite to make sure the length wasn't too long. i typic	5	1	0	General	Dresses	Dresses
12	10	1077	53	Dress looks like it's made of cheap material	Dress runs small esp where the zipper area runs. i ordered the sp which typically fits me	3	0	14	General	Dresses	Dresses
13	11	1095	39		This dress is perfection! so pretty and flattering.	5	1	2	General Petite	Dresses	Dresses
14	12	1095	53	Perfect!!!	More and more i find myself reliant on the reviews written by savvy shoppers before me a	5	1	2	General Petite	Dresses	Dresses
15	13	767	44	Runs big	Bought the black xs to go under the larkspur midi dress because they didn't bother	5	1	0	Initmates	Intimate	Intimates
16	14	1077	50	Pretty party dress with some issues	This is a nice choice for holiday gatherings. i like that the length grazes the knee so it is c	3	1	1	General	Dresses	Dresses
17	15	1065	47	Nice, but not for my body	I took these out of the package and wanted them to fit so badly, but i could tell before i pu	4	1	3	General	Bottoms	Pants
18	16	1065	34	You need to be at least average height, or taller	Material and color is nice. the leg opening is very large. i am 5'1 (100#) and the length h	3	1	2	General	Bottoms	Pants
19	17	853	41	Looks great with white pants	Took a chance on this blouse and so glad i did. i wasn't crazy about how the blouse is pl	5	1	0	General	Tops	Blouses
20	18	1120	32	Super cute and cozy	A flattering, super cozy coat. will work well for cold, dry days and will look good with jean	5	1	0	General	Jackets	Outerwear
21	19	1077	47	Stylish and comfortable	I love the look and feel of this tulle dress. i was looking for something different, but not ov	5	1	0	General	Dresses	Dresses
22	20	847	33	Cute, crisp shirt	If this product was in petite, i would get the petite. the regular is a little long on me but a	4	1	2	General	Tops	Blouses
23	21	1080	55	I'm tom!	I'm upset because for the price of the dress, i thought it was embroidered! no, that is a pri	4	1	14	General	Dresses	Dresses
24	22	1077	31	Not what it looks like	First of all, this is not pullover styling. there is a side zipper. i wouldn't have purchased it	2	0	7	General	Dresses	Dresses
25	23	1077	34	Like it, but don't love it.	Cute little dress fits tts. it is a little high waisted. good length for my 5'9 height. i like the dr	3	1	0	General	Dresses	Dresses
26	24	847	55	Versatile	I love this shirt because when i first saw it, i wasn't sure if it was a shirt or dress. since it is	5	1	0	General	Tops	Blouses

Figura 3.1: Datos originales descargados de Kaggle.com

3.0.3. Preparación de los datos

Una de las características más sobresalientes de los modelos de redes neuronales es que parten de la premisa de que la selección o procesamiento de las variables puede ser como tal una parte de la arquitectura de la red, es decir, en los modelos de aprendizaje de máquina tradicional, se requiere de un proceso de limpieza de datos, remoción de valores NA, valores atípicos, etc, sin embargo para Deep Learning se intenta que este proceso se realice de manera automática por la red, por ejemplo, en las redes neuronales convolucionales, las capas de convolución juegan un papel de selección de características, para después alimentar a las tradicionales redes densamente conectadas. Este enfoque es muy utilizado en deep learning, sin embargo para efectos de este proyecto se usará un procesamiento del texto de las opiniones para potenciar los resultados de la clasificación.

Como primer paso se puede analizar el porcentaje de valores missing en el dataset. Como se observa en la figura 3.2 la variable opinión cuenta con menos del 4 % de valores missing, por lo cual procederemos a eliminar las filas que tengan missing en la variable de opinion.

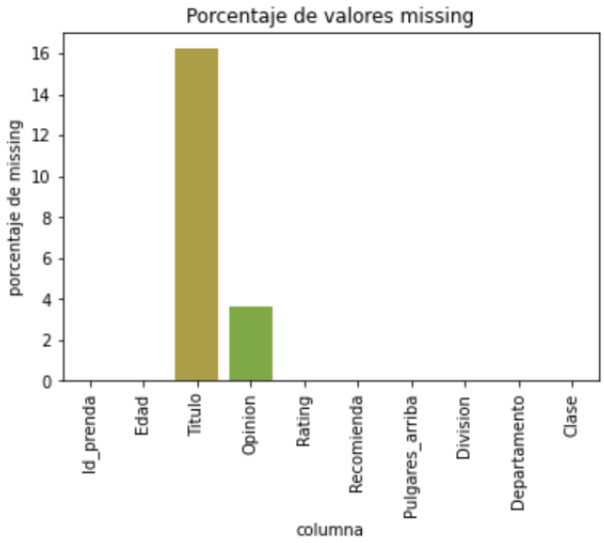

Figura 3.2: Porcentaje de valores missing para cada variable

Otro aspecto interesante a analizar es es la distribución de la variable objetivo. En la figura 3.3 se observa que existen un desbalance en el dataset, sin embargo este no llega a ser tan extremo, por lo cual debemos tener en consideración que además del accuracy, debemos inspeccionar las métricas como F1 y AUC.

Otra de las prácticas importantes para limpiar el texto es remover las denominadas stopwords, estas son palabras de uso común que no aportan información sobre la opinión de los cliente. Algunos ejemplos de Stop Word son: ante, antes, aún, aunque, aquí, arriba, atrás así, bajo, bastante, cabe, conmigo, bien, casi, cierto, como, debajo, ahí, ajeno, algo, algún, ambos, aquello. Básicamente son conjunciones, preposiciones, adverbios y artículos.

Posteriormente podemos explorar el contenido de las opiniones a través de una nube de palabras. En la figura 3.4 podemos observar como las palabras más repetidas tienen que ver con vestido, amor, genial, ajuste, hermoso, etc. Esto tiene mucho más sentido si recordamos que las reseñas son realizadas en su mayoría por mujeres, quienes tienden a usar este tipo de vocabulario.

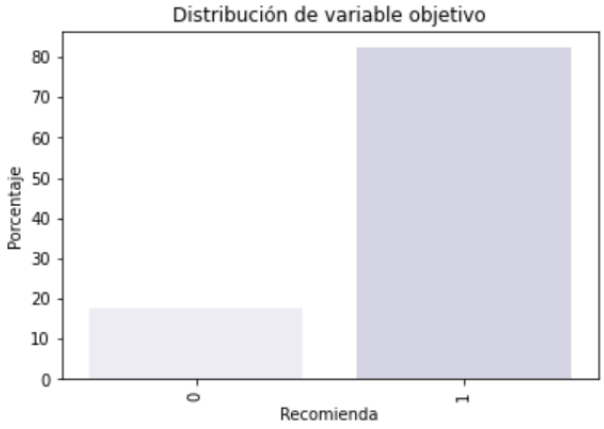

Figura 3.3: Distribución de clientes que recomiendan el producto vs clientes que no lo recomiendan

Figura 3.4: Nube de palabras con las opiniones de los clientes.

Posteriormente se realizar la vectorización del texto con las opiniones, es decir, representar de manera vectorial una de las opiniones. Esto nos ayudará a poder usar como input estos vectores, toda vez que los modelos de redes neuronales deben recibir como entradas matrices de valores finitos no nulos.

De igual manera, se debe realizar la partición de los datos en entrenamiento y prueba. Generalmente en la industria se usa una partición de tamaño 70 % para entrenamiento y el resto para la prueba, por lo cual es la que se seguirá en este proyecto. Además de los conjutos mencionados anteriormente, se suele utilizar un tercer conjunto denomindao conjunto de validación, cuyos resultados son los que se toman como una aproximación más cercana a los que ocurrirá en un ambiente productivo. En este caso se tomaron 500 tuplas como conjunto de validación. Dichas tuplas no serán utilizadas sino hasta la fase final de la evaluación del modelo.

3.0.4. Construcción del modelo

Para la etapa de implementación del modelo vamos a usar el modelado secuencial, es decir, se definirá cada capa de la red de manera secuencial para desde la capa de input hasta la capa de salida, esto lo haremos a través de KerasTensorflow ya que cuenta con comandos amigables e intuitivos que nos permiten fácilmente perosnalizar una red.

Un punto a recalcar es el uso de una capa de Embeding como la primera capa oculta de la red, una breve explicación de porque es conveniente usarlas es la siguiente: la capa de Embedding es un método de clustering, y como todo método de clustering matemático lo que pretende es, agrupar palabras similares en grupos homogeneos y que dichos grupo sean lo más heterogeneos entre si, unos de otros. Es decir, las palabras similares van juntas y están lo más separadas posible de las palabras no similares. Además de representar el texto como números al igual que hace One Hot Encoding, la capa de Embeding permite realizar una reducción de la dimensionalidad: este método reduce la dimensionalidad del texto haciendo posible de entrenar conjuntos de datos más grandes. También permite la creación de un contexto: al aplicar un método de clustering que diferencia unas palabras de otras, es capaz de proporcionarle a la red neuronal información valiosa.

Posteriormente a la capa de embeding, se usará una capa LSTM con 100 neuronas y posteriormente una capa con 2 neuronas de salida. Un resumen de la arquitectura de la red neuronal se muestra en la figura 3.6

```
Model: "sequential_4"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_3 (Embedding)     (None, 250, 50)           2500000

 spatial_dropout1d_3 (Spatia  (None, 250, 50)          0
 lDropout1D)

 lstm_3 (LSTM)               (None, 100)               60400

 dense_6 (Dense)             (None, 100)               10100

 dense_7 (Dense)             (None, 2)                 202

=================================================================
Total params: 2,570,702
Trainable params: 2,570,702
Non-trainable params: 0
_________________________________________________________________
```

Figura 3.5: Arquitectura de la red neuronal implementada.

Para la fase de entrenamiento se entrenó a 100 épocas, con un tamaño de lote de 1,000, y usando un parámetro de tolerancia de 15 para la detención temprana. Luego de realizar el entrenamiento en varias corridas, la dentención temprana lo detenía luego de no más de 50 épocas, ya que no encontró mejoras que superará el nivel de tolerancia, siendo uno de los mejores resultados un accuracy del 93.9 % en entrenamiento y un 89.3 % en en conjunto de prueba. Estos resultados no son excelentes pero llegan a ser aceptables para un modelo simple y cuyo tiempo de entrenamiento no rebasa los 20 minutos. En la figura 3.6 se puede observar el comportamiento del accuracy y función de pérdida en una corrida de entrenamiento.

A continuación veremos como se comporta con el conjunto de validación.

3.0.5. Evaluación

Cuando se implementa un modelo hay 3 conjuntos de datos fundamentales:

- Conjunto de datos de entrenamiento: son los datos que entrenan los modelos
- Conjunto de datos de validación: selecciona el mejor de los modelos entrenados
- Conjunto de datos de test: Nos ofrece el error real cometido con el modelo seleccionado

En esta etapa usaremos las 500 tuplas que se reservaron para la validación. Con ella se calcularon la matriz de confusión la cual se muestra en la figura 3.8.

De la matriz de confusión podemos concluir varias cosas, con u accuracy del 83 % en los datos de validación el modelo no se está desempeñando muy bien. Al usar un

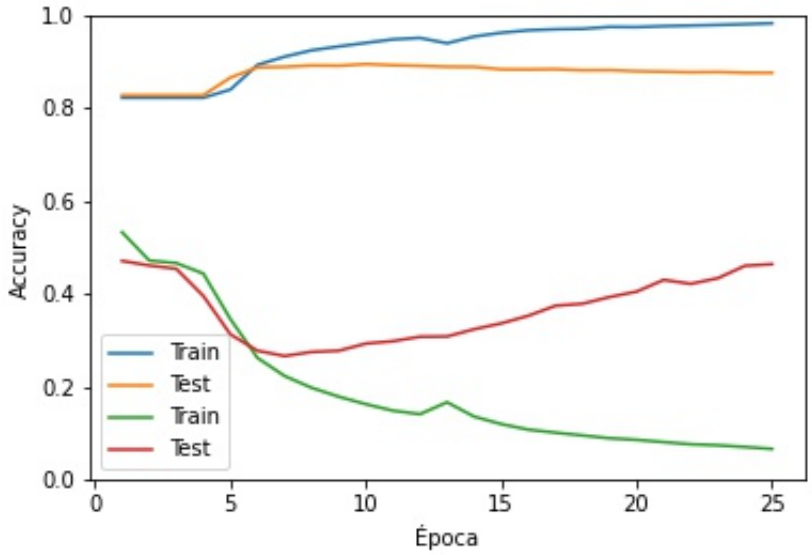

Figura 3.6: Resultados de la fase de entrenamiento y prueba.

número de neuronas reducido en cada capa, supuse que la complejidad del modelo se vería reducida y no habría tanto sobreajuste, sin embargo en esta etapa vemos que no es así, el modelo no generaliza de la mejor forma, podemos trabajar en eso aún más para aminorar el impacto del desbalance de la variable objetivo.

	precision	recall	f1-score	support
No Recomienda	0.55	0.76	0.64	98
Recomienda	0.93	0.85	0.89	402
accuracy			0.83	500
macro avg	0.74	0.80	0.76	500
weighted avg	0.86	0.83	0.84	500

Figura 3.7: Reporte de clasifiación para cada clase.

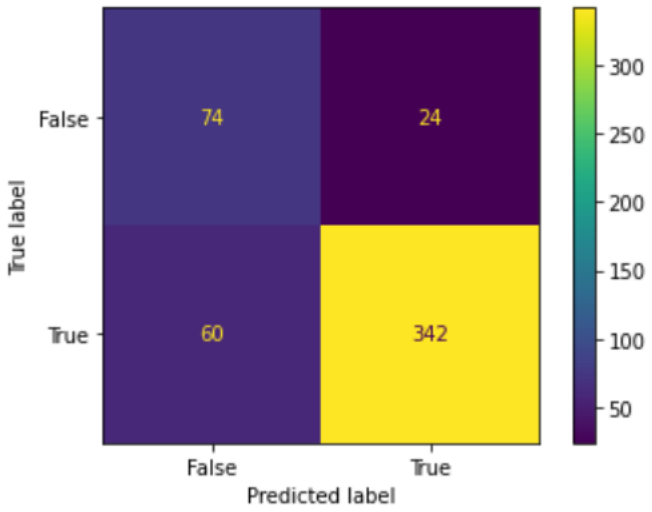

Figura 3.8: Matriz de confusión con los datos de validación.

Capítulo 4

Conclusiones

Con el modelo de red neuronal LSTM se puede implementar un clasificador para los comentarios de los clientes de un negocio de ropa en línea. Al clasificar los comentarios de nuevos clientes, podemos identificar si el cliente quedo satisfecho con su compra o por el contrario es un cliente que probablemente no vuelva a comprar en el portal de la empresa ABC. Al modelar como variable respuesta si el cliente recomienda o no comprar el producto, se puede estimar que tan conformes están nuestros clientes, y este mismo análisis se puedes realizar por departamente, por temporada, por tipo de ropa, etc. Si existen productos que se asocian a votos negativos habría que revisar el material de los mismos, las tallas, el servicio de entrega de paquetería, a fin de disminuir la inconformidad de los cliente. Para el caso de las opiniones negativas, que son un número menor, se puede implementar un seguimiento sobre esos casos particulares.

Capítulo 5

Fuentes

[1] López Gaona A., Avilés Rosas G, Minería de datos con R. Las Prensas de Ciencias. 1er Edición. 2021.

[2] Simon O. Haykin. Neural Networks and Learning Machines. Pearson. 3ra Edición. 2008.

[3] https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews

[4] https://es.stackoverflow.com/questions/476714/para-qu %C3 %A9-sirve-la-capa-embeddign-en-un-modelo-lstm