Drug Review Dataset
Examen general de conocimientos

López Velasco Jossé Armando

Marzo 14, 2022

1 Introducción

1.1 Análisis previo

El dataset utilizado para este examen general consta de 2 archivos tsv, el análisis que se muestra a continuación fue realizado sobre el conjunto de entrenamiento (drugsComTrain_raw.tsv)[1]. Este archivo presenta la siguiente lista de atributos:

- drugName: Nombre del medicamento
- condition: Enfermedad o padecimiento que el paciente presenta
- review: Reseña en formato texto para el medicamento prescrito.
- rating: Calificación en una escala del 1-10 que el paciente otorga al medicamento
- date: Fecha en la cual la reseña fue capturada
- usefulCount: No especificado

	drugName	condition	review	rating	date	usefulCount
206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9.0	2012-05-20	27
95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8.0	2010-04-27	192
92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5.0	2009-12-14	17
138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8.0	2015-11-03	10
35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9.0	2016-11-27	37

Figure 1: Muestra del dataset

El tamaño total del archivo es de 161297 registros, de los cuales la distribución de reseñas por medicamento se puede apreciar en la figura 2

El padecimiento más frecuente en la población corresponde a **Birth Control** con 28788 reseñas para distintos medicamentos, seguido de depresión, dolor, ansiedad, acné, etc. Los cuales se muestran en 3+ Existe una correlación rela-

Levonorgestrel	3657
Etonogestrel	3336
Ethinyl estradiol / norethindrone	2850
Nexplanon	2156
Ethinyl estradiol / norgestimate	2117
Ethinyl estradiol / levonorgestrel	1888
Phentermine	1543
Sertraline	1360
Escitalopram	1292
Mirena	1242
Implanon	1102
Gabapentin	1047
Bupropion	1022
Venlafaxine	1016
Miconazole	1000
Medroxyprogesterone	995
Citalopram	995
Lexapro	952
Bupropion / naltrexone	950
Duloxetine	934

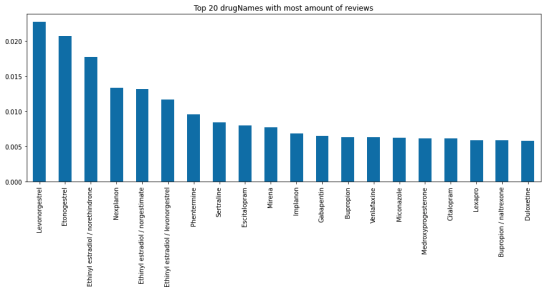

Figure 2: Top 20 medicamentos con mas reseñas

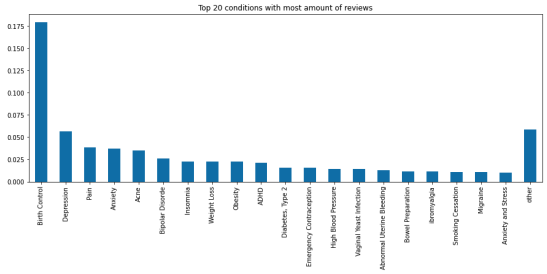

Figure 3: Top 20 condiciones con mas reseñas

	rating	usefulCount
rating	1.000000	0.234185
usefulCount	0.234185	1.000000

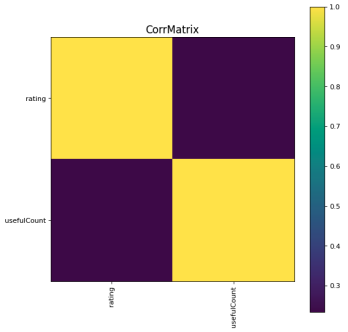

Figure 4: Matriz de correlación

tivamente alta entre las dos variables (rating, usefulCount) como se aprecia en la figura 4

Durante toda la historia que se tiene del medicamento **Levonorgestrel** se aprecia una aceptación buena para padecimientos como **Birth Control y**

[1]https://archive.ics.uci.edu/ml/machine-learning-databases/00462/

Emergency Contraception, sin embargo su aceptación se encuentra en lugares críticos para finales del 2017 para el padecimiento **Abnormal Uterine Bleeding**, esto se puede apreciar de una manera más clara en la figura 5 Al-

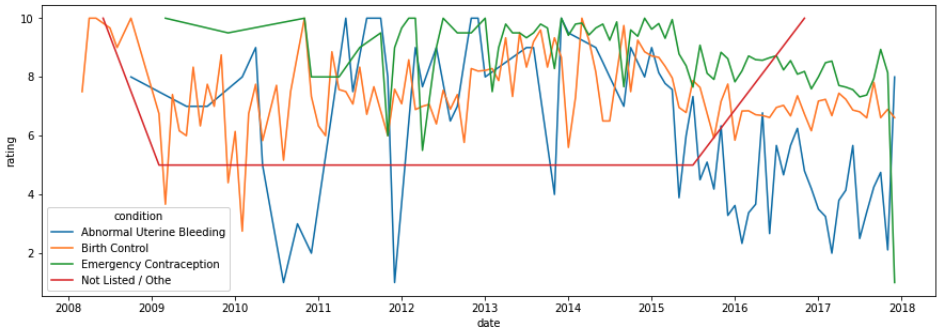

Figure 5: Historia mensual de los medicamentos con mejor calificación

gunos casos críticos que se pueden apreciar es la caída en la aceptación de los medicamentos para ciertas condiciones, como es el caso de **ansiedad y depresión**; ambos han estado presentando una caída en su promedio de rating mensual, para el caso de la **depresión** obtuvo una puntuación máxima de 9 a mediados del 2008 hasta una puntuación de 6 presentada a finales del 2017. Algo similar ocurre con la **ansiedad** cuya tendencia bajista es notoria desde el 2015 pasando de una aceptación de 9.5 hasta un mínimo de 6 para finales del 2017. Lo podemos apreciar de una manera más clara en la figura 6 El rating de

Figure 6: Promedio mensual de rating para depresión y ansiedad

los medicamentos data una condición ha seguido bajando durante los últimos años 7

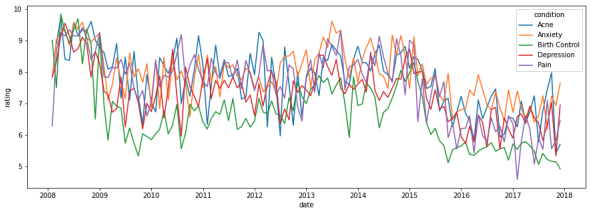

Figure 7: Promedio mensual de rating para top condiciones

2 Análisis del modelo de clasificación para NPS

Dado que tenemos el rating en una escala del 0-10 podemos construir un modelo que identifique promotores y detractores.

2.1 NPS

NPS (Net Promoter Score) es un indicador utilizado para medir la experiencia del usuario, gracias a este escore podemos identifciar qué tan probable es que el medicamento sea recomendado por el consumidor. Se calcula con base a estas sencillas reglas:

- Promotores: Puntuación de 9 o 10
- Neutros: Puntuación de 7 u 8
- Detractor: Puntuación de 0-6

2.2 Resultados NPS model

Se utilizó una regresión logística con el fin de obtener un modelo rápido del cual podíamos obtener información útil sobre el comportamiento del texto para una predicción de la clase promotor.

Sobre el archivo de train se realizó un split del 20% para test, los resultados que se muestran a continuación refieren a este subconjunto de datos.

2.2.1 AUC

El modelo obtuvo un auc en el conjunto de prueba del **0.73** y la siguiente matriz de confusión 8

2.2.2 SHAP values

Las palabras que aportan más a una predicción positiva son [year, life, great, work] 9

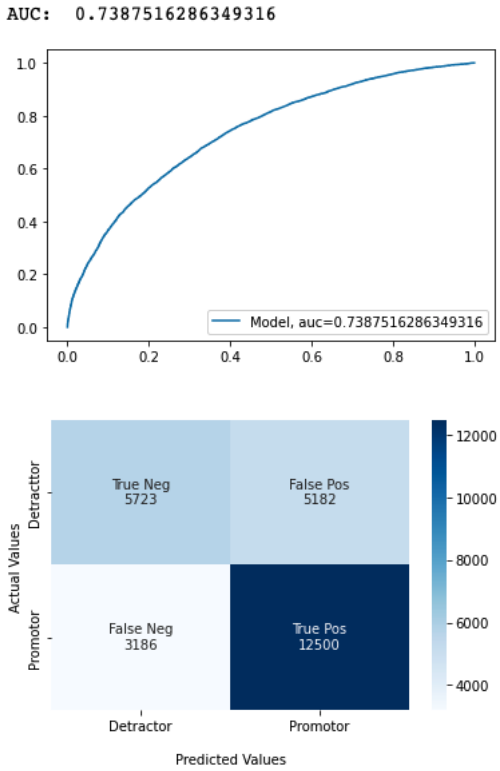

Figure 8: AUC y matriz de confusión para NPS

3 RNN

3.1 LSTM

3.2 Método

Para realizar la tarea central del examen se utilizó una LSTM **?** la cual es una red neuronal recurrente ya que es bien sabido que esta arquitectura funciona muy bien con secuencias de datos debido a sus conexiones en "reversa". Esta arquitectura está compuesta básicamente de

- Una celda
- Compuerta de entrada
- Compuerta de salida
- Compuerta del olvido

El flujo de datos que se llevó a cabo fue el que se describe a continuación:

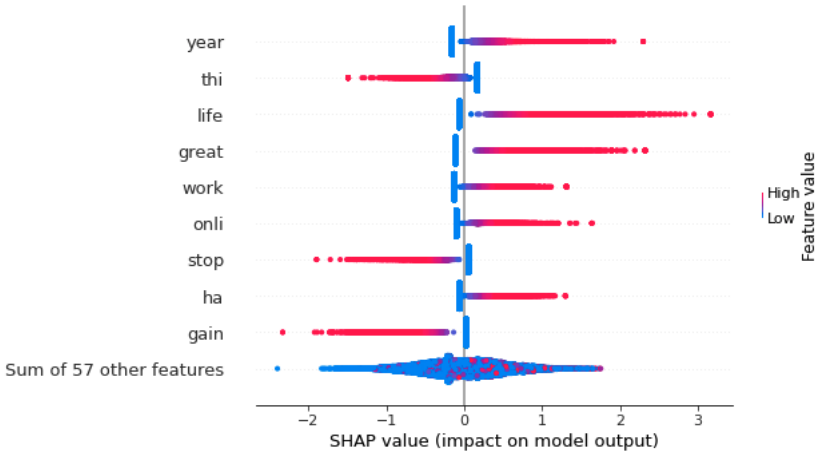

Figure 9: Shap values para modelo de nps

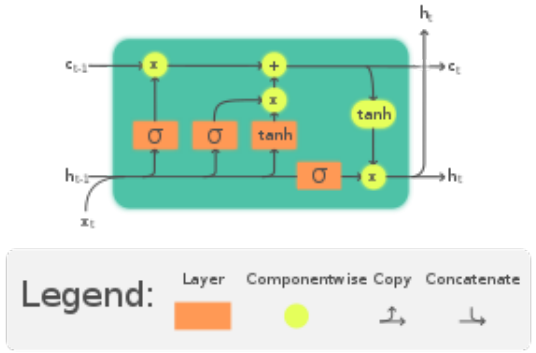

Figure 10: arquitectura de una LSTM

3.2.1 1.- Limpiar el texto

En este punto se realizaron sub-tareas necesarias para mejorar el rendimiento de la red, ya que si no se lleva un proceso adecuado podríamos terminar con un corpus demasiado grande que no aporte información necesaria al modelo. El punto de interés debe radicar en llegar a una representación vectorial que será alimentada a la LSTM. Para esto se realizó:

- Eliminación de "stop words": Este es un proceso necesario ya que palabras como [it, this, that, I, He, etc] no nos aportan mucha información sobre el punto central de la reseña, y además suelen representar la mayor cantidad de frecuencias debido a que son muy necesarios para conectar ideas.
- Eliminación de caracteres especial

- Stemming: Reducir la palabra a su raíz
- Vectorizar: Transformar una oración en un vector

Para el último punto utilizamos TFidVectorizer, el cual vectoriza el texto con base en la frecuencia en el documento y la frecuencia del término, de esta manera puedo limitar el mínimo número de apariciones de un término y obtener información de todo el corpus. Básicamente lo elegí debido al tamaño del dataset, el cual hacía muy tardado un entrenamiento en vectores más grandes. Cada vector resultante tenía un tamaño de [61,1], el cual iba a ser alimentado a la LSTM

3.2.2 Arquitectura

La LSTM utilizada es la que se muestra en :

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 61, 32)            1952
_________________________________________________________________
lstm_1 (LSTM)                (None, 4)                 592
_________________________________________________________________
dense_1 (Dense)              (None, 11)                55
=================================================================
Total params: 2,599
Trainable params: 2,599
Non-trainable params: 0
```

Figure 11: arquitectura utilizada

La primer capa contiene una capa que hace un embedding de nuestra entrada en un vector de tamaño (32,), de esta manera lo pasamos a 4 unidades LSTM. La salida de esta pasa por una capa densa que se activa con una función softmax a fin de generar una etiqueta por ejemplo.

4 Resultados

Se obtuvo un auc de 0.76 en el conjunto de validación y un f1 score de 0.31. En **?** podrá encontrar todos los resultados y una comparación con un modelo Naive multiclase con el mismo pre-procesamiento de datos.

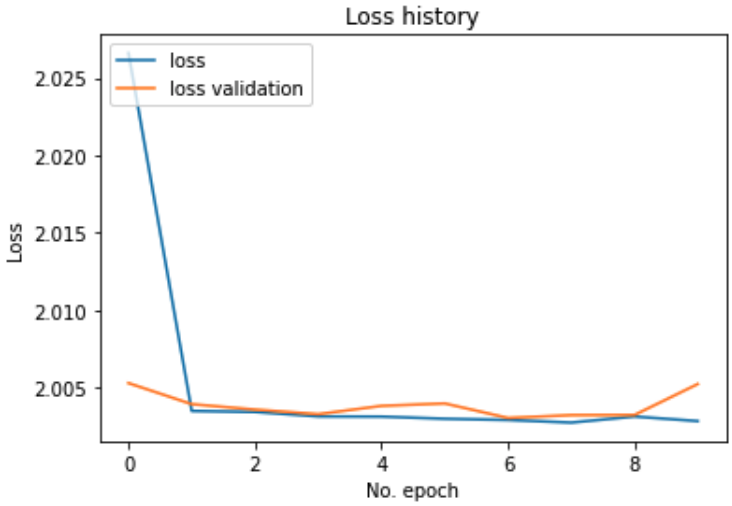

Figure 12: Loss plot