# Tarea 1

IRVING DANIEL ESTRADA LÓPEZ
*Facultad de Ciencias Físico Matemáticas*
*Universidad Autónoma de Nuevo León*
Nuevo León, México
irving.estradalo@uanl.edu.mx

## I. Introducción

El preprocesado de datos es toda acción que se lleva a cabo antes de que empiece el proceso de análisis de datos. Se tienen que transformar los datos crudos en un nuevo conjunto de datos que tenga la información valiosa de nuestros datos reales, en la mayoría de las veces tenemos que adecuar nuestro conjunto de datos para dicho análisis. Los conjuntos de datos vienen con ruido, datos nulos, incompatibilidad entre los datos, provienen de múltiples fuentes de información, contienen datos irrelevantes para nuestro análisis, entre otras cosas. Nos referimos a información valiosa como aquella que aporte significancia a nuestro análisis.

Debemos tener en cuenta los datos con los que estemos trabajando, así como el análisis que se quiere llevar a cabo, ya que esto dependerá del preprocesado que se tiene que realizar para hacer un buen análisis.

En este documento la intención es hacer un preprocesado de texto, hacer un conteo de palabras, extraer características importantes del conjunto de datos e identificar áreas de oportunidad, para en un futuro llevar acabo alguno de los siguientes análisis: clasificación de texto, análisis de sentimiento, análisis de tópicos o análisis de intención.

Primero comenzaremos con un análisis exploratorio de nuestras variables, de esta manera podemos perfilar hacia donde tiene que ir dirigido nuestro preprocesado. Continuando con un análisis de frecuencias para identificar características relevantes de los textos. Por último, en la parte final de este documento se identificarán áreas de oportunidad que se adecúen al conjunto de datos seleccionado. Se mencionarán posibles análisis que contrubuyan brindando información relevante del conjunto de datos. Hoy en día se genera una gran cantidad de datos diarios, de los cuales podemos extraer información valios, siendo el preprocesado un obstáculo para elaborar un buen análisis.

## II. Datos

El conjunto de datos con el que se trabajó en este artículo es de dominio público y fue obtenido de Kaggel. Nuestro conjunto de datos contiene 9,713 registros los cuales fueron extraídos de Flipkart utilizando Selenium y BeautifulSoup. Como podemos ver en la figura 1 el conjunto de datos contiene la valoración que le dan al producto Iphone SE, un comentario y la reseña de los clientes en el e-commerce de la India Flipkart.

Fig. 1.

| | Ratings | Comment | Reviews |
|---|---|---|---|
| **0** | 5 | Super! | Great camera for pics and videos Battery life ... |
| **1** | 5 | Must buy! | Great device. Let me tell the Pros..1. Superb ... |
| **2** | 5 | Great product | Who all loves older size i.e., 4.7 inch type s... |
| **3** | 5 | Simply awesome | This iPhone SE is the best phone ever you get.... |
| **4** | 5 | Classy product | This is my second iphone after iphone 4s. I've... |

En la tabla 1 podemos revisar los atributos de nuestro conjunto de datos junto a una descripción, es importante destacar que la variable Ratings va de 1 a 5. Siendo la calificación más baja 1 y siendo la mejor 5.

| Variable | Representación |
|---|---|
| Ratings | La calificación que los clientes le dieron al producto. |
| Comment | Comentario, es el título de la reseña. |
| Reviews | La reseña del cliente hacia el producto. |

TABLE I
Variables del Conjunto de datos

## III. Metodología

### A. Análisis Exploratorio

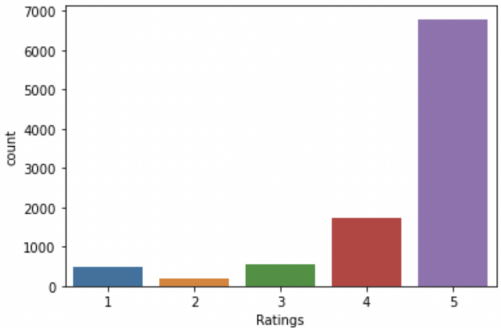
En esta sección extraeremos algunas características descriptivas de las variables y concluiremos en base a ellas. Dos de nuestras tres variables son textos, la variable Comment y la variable Reviews. Es importante mencionar que no hay datos faltantes, ya que si los hubiera deberíamos tomar la decisión si eliminar los registros o llevar acabo un preprocesado. La variable Ratings es de tipo entero, así que podemos extraer algunas características interesantes de ella.

Fig. 2.

| | Ratings |
|---|---|
| **count** | 9713.000000 |
| **mean** | 4.456399 |
| **std** | 1.032911 |
| **min** | 1.000000 |
| **25%** | 4.000000 |
| **50%** | 5.000000 |
| **75%** | 5.000000 |
| **max** | 5.000000 |

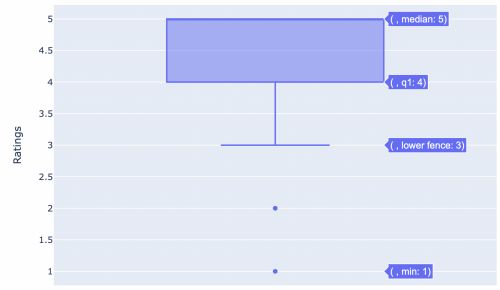
Como podemos darnos cuenta en la figura 2, el promedio de las calificaciones dadas al producto es 4.45, y viendo los cuartiles, más del 50% de nuestros datos cuentan con una calificación de 5. El 75% cuenta con una calificación de 4 o más. Esto nos indica que el producto tiene buenas valoraciones. De los 9,712 registros, 6,788 tiene 5 de calificación, casi el 70% de nuestras calificaciones.

Fig. 3.




Para verlo de una forma más ilustrativa tenemos la figura 3. Donde claramente podemos darnos cuenta de que la calificación que domina es el 5, siguiéndole la calificación 4. Los registros que cuenta con 1 como calificación son 474, menos del 5%. Indicándonos que por lo menos con los registros que tenemos los usuarios le dan en su mayoría una buena calificación.

Fig. 4.




Nos apoyamos también en la figura 4 y efectivamente el producto tiene calificaciones buenas. Incluso está tomando la calificación 1 y 2 como datos anómalos. Esto comprobando que el producto aparentemente para los usuarios es de buena calidad, al menos con los registros que tenemos.

Fig. 5.

```
'Great camera for pics and videos Battery life is good so far with some setting turn of
which i never use and when i use i turn those on and i use it in power saving mode all t
he time so a full day with light gaming of 1hr or more using camera for 1hr or more list
ening music in my car on youtube and Bluetooth on for couple of hours it gives me full d
ay of battery varing from 4/5 to 6/7 hours sot per dayPerformance is top notch plays eve
ry game and every task with easeVery premium phone look...READ MORE'
```

En nuestro conjuto de datos, en los registros de la variable Reviews tenemos el siguiente caso mostrado en la figura 5. Al final de cada una de las reseñas tenemos "...READ MORE" que esto proviene directamente de cuando se extrajeron los datos de Flipkart, esta es una de las cosas que debemos de adecuar en nuestro preprocesado.

### *B. Preprocesado*

Los procesos que se mencionarán a continuación se llevaron a cabo tanto en los registros de la columna Comment como en los de la columna Reviews.

Se llevarán acabo las siguientes acciones en el preprocesado de texto:

**Limpieza de texto:** Se comienza eliminando los patrones de ruido que trae consigo el texto, en algunas ocaciones por la fuente de datos en la extracción contiene patrones de ruido al inicio o al final de los textos, como mencionamos anteriormente, nuestra variable Reviews tiene en la parte final el texto "...READ MORE", el cual genera ruido dentro de nuestros elementos en nuestra variable. Utilizando la librería "re" y apoyándonos con un ciclo for podemos eliminar dicho ruido, para de esta forma se modifiquen todos nuestros registros y no hacerlo de forma individual.

**Remover emojis:** Hoy en día es común que las personas utilicen emojis, en este caso los removeremos y nos enfocaremos en las palabras. Hay librerías que nos pueden describir con palabras los emojis, es importante saberlo para preprocesados futuros.

**Remover URL:** En caso de existir alguna liga que nos lleve a un sitio web, no aporta alguna característica valiosa a los análisis posteriores.

**Remover signos de puntuación:** Debemos de quitar los signos de puntuación ya que en este caso no aportarán en el conteo de palabras, en otra situación como el análisis de sentimientos nos podrían brindar información valiosa. Esto se debe de hacer después de remover la URL para que no haya problemas, ya que la URL contiene signos de puntuación y todo está concatenado, de lo contrario existe la posibilidad de que se interprete como palabras.

**Conversión a minúsculas:** Para no tener problemas al momento de obtener la frecuencias de las palabras, convertimos todo nuestro texto en minúsculas. Si no hicieramos este paso, identificaría palabras distintas por el hecho de tener diferencia entre mayúsculas y minúsculas. En algunas ocaciones es importante no hacer este paso, debido a que las mayúsculas y minúsculas nos aportan información, como en el caso de análisis de sentimientos que se relacionan las mayúsculas como un enfasis en las palabras del usuario.

**Revisión de ortografía:** Este es un factor importante, tomando en cuenta que el sito de donde fueron extraídos la infromación es de la India, y su primer idioma no es el inglés. Existen librerías que nos ayudan a la revisión de ortografía, ya que muchas veces nuestros conjuntos de datos contienen faltas de ortografía las cuales pueden afectar nuestros análisis, utilizamos la librería de Python "pyspellchecker" la cual utiliza la distancia de Levenshtein para hayar la palabra más parecida en caso de contener algun error ortográfico.

**Tokenizar:** Este paso consiste en dividir el texto en las unidades que lo conforman, entendiendo por unidad el elemento más sencillo con significado propio para el análisis en cuestión, en este caso las palabras.

**Eliminación de stopwords:** Las stopwords son consideradas palabras que no tienen un significado por si solas, sino acompañan a las otras palabras. Es importante removerlas debido a que abarcan una gran cantidad en los textos y al

momento de obtener la frecuencia de nuestras palabras, generalmente aparecen en el top de palabras repetidas. También afectarían el costo computacional de todos nuestros procesos. Apoyándonos con la librería "nltk" se eliminaron las stopwords de todas nuestras listas. Más adelante se excluyeron algunas palabras que se consideraron como parte de las stopwords. Ya que el conjunto de datos y las reseñas son acerca de Iphone SE, se excluyeron palabras como "iphone", "se", "phone", entre otras. Esto debido a que al igual que las stopwords aparecerán con una gran frecuencia, impidiendo entrar al top de palabras información valiosa.

**Stemming o lematización:** Las palabras tienen muchas variaciones, es importante encontrar el estado base o estandarizar las palabras, para que se pueda llevar a cabo el análisis que se quiere lograr. El stemming es un algoritmo que busca el nucleo de la palabra pero en el inglés así como en otros idiomas existen excepciones, dandonos resultados erroneos. El lematizar es más complejo ya que busca el lema de la palabra que sería una serie de caractéres que forma una unidad semántica y que puede constituir una entrada en el diccionario. La librería "nltk" tiene ambas, sin embargo, ninguna de las dos esta libre de posibles errores en las conversiones o la absoluta llegada a los resultados deseados.

Una vez realizado el preprocesado conseguimos la estructura adecuada para comenzar con un conteo de palabras y analizar sus frecuencias, esto nos lleva a identificar palabras destacadas o relevantes que los usuarios mencionan del producto. Este conjunto de datos tiene la ventaja que podríamos filtrar por medio de calificaciones, para extraer información valiosa. Por ejemplo, en las malas calificaciones del producto en la variable Ratings podemos extraer las palabras que tienen mayor frecuencia, de este modo dándonos una perspectiva de qué es lo que más mencionan los usuarios que le dan una mala calificación al producto.

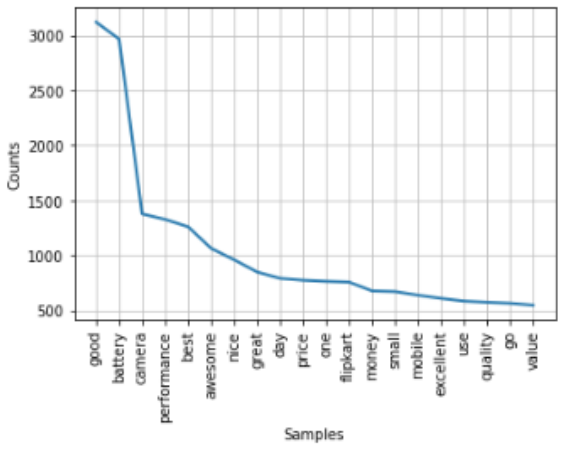
## IV. Resultados

A continuación, se presentan de forma gráfica las palabras con mayor repetición de forma general de las respectivas variables. Después de analizar los gráficos generales, nos concentraremos en las reseñas que dieron una baja calificación al producto, para de esta forma revisar cuales son las palabras que más se mencionan al momento de darle una mala calificación al Iphone SE. Estos gráficos no hubieran sido posible sin el preprocesado de datos que llevamos a cabo, recordando que el texto es de tipo de dato no estructurado. Con estos gráficos podemos destacar las palabras con mayor frecuencia en las reseñas de Flipkart.

### A. Variable Reviews

La figura 6 nos muestra las palabras con mayor repetición dentro de nuestras reseñas. El top 3 de palabras con mayor repetición es "good", "battery" y "camera", recordando la estadística descriptiva el producto tiene buenas valoraciones en su mayoría, así que existe la sospecha de que lo que más les gusta a los usuarios es la cámara y la batería de dicho producto o al menos es lo que más se menciona dentro de la mayoría de las reseñas. Sin embargo, no dejan de ser sospechas.

Fig. 6.



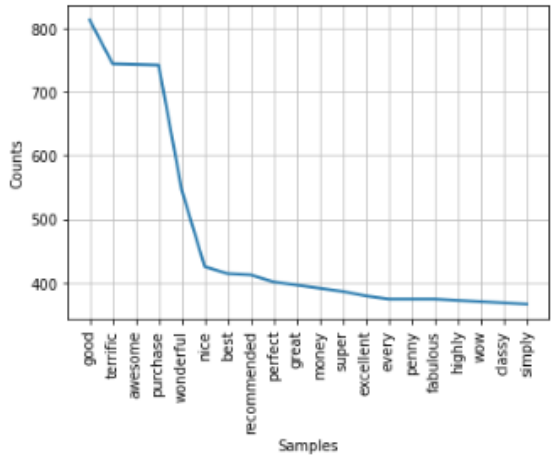

Después en la figura 7 tenemos la nube de palabras de todas las reseñas, este gráfico nos permite ver de una forma más ilustrativa la repetición de las palabras. Entre ellas podemos destacar palabras que hacen referencia a que este es un buen producto. Estos gráficos junto a la estadística descriptiva de la Ratings se complementan, indicando la calidad del producto.

Fig. 7.


### B. Variable Comment

En la figura 8 podemos apreciar que en la variable Comment tenemos de top 3 las palabras "good", "terrific" y "awesome", todas siendo positivas. La palabra más utilizada en los comentarios es la palabra "good". Tiene sentido que las palabras con mayor frecuencia hagan referencia a algo bueno, ya que como hemos estado revisando a través del artículo la mayoría de los registros de nuestro conjunto de datos nos señalan valoraciones positivas.

Fig. 8.



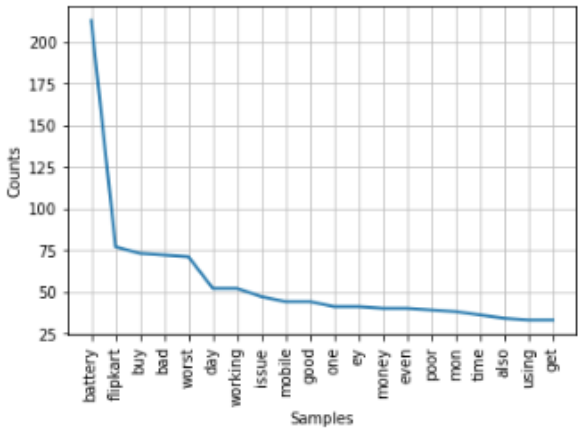

En la figura 9 tenemos la nube de palabras de la variable Comment, recordando que esta variable se podría considerar como el título de la reseña. Podemos identificar una vez más, palabras que contribuyen a la calidad del producto indicándonos que al menos con los registros que tenemos los usuarios en su mayoría lo consideran una buena compra.

Fig. 9.


## C. Variable Reviews (Mala calificación)

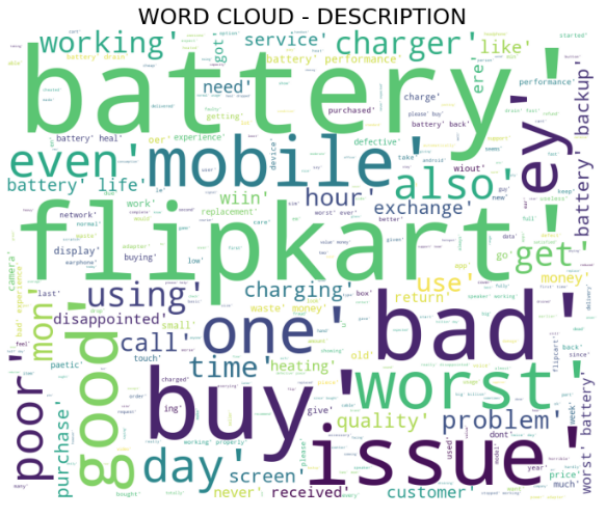
Como mencionamos anteriormente, si analizamos la frecuencia de las palabras dirigidas hacia las malas reseñas podríamos identificar características importantes a destacar. Con la figura 10 identificamos que el top 3 de palabras son “battery”, “flipkart” y “buy”. Estos resultados crean la sospecha de que los usuarios que le dan una mala calificación al producto han tenido algún detalle con la batería o incluso con el e-commerce flipkart, en muchas ocasiones los clientes para dar una calificación toman en cuenta el tiempo en que tarda en llegar el producto, que llegue en buen estado, entre otras cosas. La calificación del producto puede verse afectada por estos procesos. Por otro, lado se pudo tener algún detalle con la batería en cuanto a su duración o algún fallo que el usuario detectara.

Fig. 10.




En la figura 11 tenemos la nube de palabras, donde vemos de forma más visual que las palabras destacadas son “battery” y “flipkart”. Identificamos otras palabras relacionadas a un mal desempeño del producto como “issue” donde tendríamos que identificar cuál es el problema o con quién, ya que podría ser con el producto o con el e-commerce.

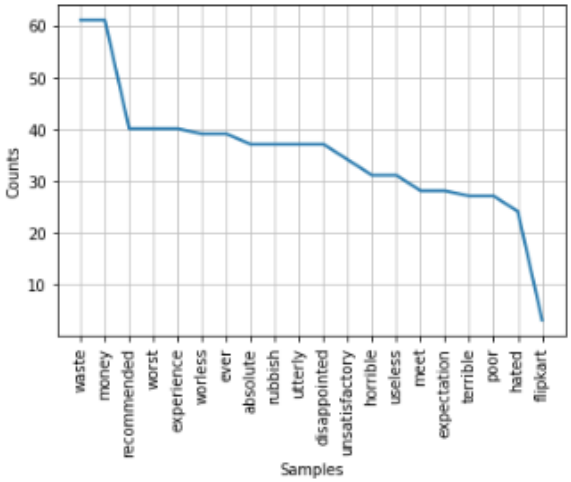
Fig. 11.




## D. Variable Comment (Mala calificación)

En la figura 12 podemos identificar en la variable Comment que tenemos de top 3 las palabras “waste”, “money” y “recommended”. Aquí pasa algo interesante, la palabra “recommended” por si sola pudiéramos decir que es positiva, sin embargo, al momento de hacer cualquier análisis no debemos de ser absolutos al interpretar los datos. Anteriormente en la estadística descriptiva nos dimos cuenta de que más de 50% de las valoraciones tenían una calificación de 5, la cual es la mejor calificación, sin embargo, en esta sección estamos con los comentarios que dieron la peor calificación. Existe la sospecha de que esta palabra va acompañada de otra, así formando una expresión negativa. Si recordamos, eliminamos las stopwords eliminando la palabra “not”. Existe la posibilidad que el complemento forme la expresión “not recommended”.

Fig. 12.




En la nube de palabras de la figura 13, podemos identificar de forma más visual otras palabras que hacen referencia a que el producto decepcionó al usuario, y otros elementos de disconformidad. No se aprecia ningún elemento del producto, como la batería o la cámara ya que la variable Comment se considera el título de la reseña, aparentemente no es usual que se incluya.

Fig. 13.


## V. Conclusión

El preprocesado es una etapa importante en el ciclo de análisis de datos. Esta etapa abarca un buen porcentaje de trabajo y es esencial, ya que nuestro análisis se puede ver afectado por un mal preprocesado. Los datos con los que trabajamos, en la mayoría de las ocasiones no se extraen con las características que el modelo requiere, pueden tener datos faltantes o pueden ser datos no estructurados, como lo vimos en este artículo. Las técnicas de preprocesados dependerán del tipo de dato y del análisis. En este caso estábamos trabajando con texto, y aunque hay preprocesados que se siguen sin importar la situación, hay otros que tiene un mayor impacto dependiendo de la situación. Nuestras reseñas no contenían una cantidad grande de emojis, sin embargo, en el caso de trabajar con los textos de un chat, es más común que la gente utilice los emojis que en las reseñas. Siguiendo con el ejemplo de los chats, es común utilizar abreviaturas de palabras o incluso expresiones, las cuales hay que tomar en cuenta. El entendimiento de nuestros datos, así como su fuente, es esencial para llevar a cabo un buen preprocesado y un buen análisis.

Como vimos a través del artículo, nuestro conjunto de datos es acerca de valoraciones del Iphone SE y el análisis que se quiera llevar a cabo tiene sus implicaciones en el preprocesado, en nuestro caso fue un análisis de frecuencia de palabras. Como resultado podemos decir que el conjunto de datos se presta para hacer una predicción de la calificación que el usuario le puede dar al producto según su reseña o comentario. También se pude hacer un análisis de sentimientos para identificar si la reseña es positiva o negativa, suponiendo que no tuviéramos la variable Ratings o incluso comparándo los resultados con la variable Ratings. Existe un área de oportunidad con este conjunto de datos para la parte de análisis de interacción, que hoy en día ha ocupado una gran relevancia para las compañías.

## References

[1] KAMAL DAS. (2021). Apple iPhone SE reviews ratings. Mayo del 2022, de Kaggle Sitio web: https://www.kaggle.com/datasets/kmldas/apple-iphone-se-reviews-ratings

[2] Irving Estrada. Github. 2022,Sitio web: https://github.com/Irving-Estrada/Procesamiento

[3] Aditya Beri. (2020). Stemming vs Lemmatization. Mayo del 2022, de Towards Data Science Sitio web: https://towardsdatascience.com/stemming-vs-lemmatization-2daddabcb221

[4] A. Famili, Wei-Min Shen, Richard Weber , Evangelos Simoudis . (1997). Data Preprocessing and Intelligent Data Analysis. Intelligent Data Analysis, 1, 3-23.

[5] Ms. Kranti Vithal Ghag, Dr. Ketan Shah. (2015). Comparative Analysis of Effect of Stopwords Removal on Sentiment Classification. IEEE International Conference on Computer, Communication and Control, 4.