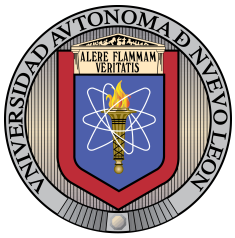

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Procesamiento y Clasificación de Datos

Gloria Samanta Servín García
1731703

PROFESORA

Mayra Cristina Berrones Reyes

Tarea2

2 de junio de 2022

Introducción

El propósito de esta actividad consiste en seguir el preprocesamiento de sus datos, realizar un análisis de sentimiento, haciendo comparación entre las diferentes librerías que se mencionan en clase. Se tomará en cuenta la discusión de resultados y su correcto análisis. Agregar en las conclusiones cual es la mejor librería que usaron y por qué.

Planteamiento

Originalmente, se intentó trabajar con diversas bases de datos buscando poder adaptar estas al código que se tiene de ejemplo, sin embargo estos intentos fracasaron por lo que la entrega de esta actividad se termina presentando con la información de Amazon Food Reviews buscando no dejar sin practicar ni analizar el tema de interés. Se dejará adjunto el código con el último intento personal que consistía en reseñas de IMDB.

Descripción de las técnicas y métodos empleados

Utilizando Google Colaboratory, tomamos de la cuenta personal de drive el archivo .csv con las reseñas. En una primera instancia revisamos una vista de la tabla para asegurarnos que el documento es el adecuado y es viable para trabajar, una vez que hayamos confirmado esto es posible continuar con el proceso de los datos.

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...

Figura 1: Dataset

Análisis de sentimiento usando TextBlop

TextBlob es una biblioteca de Python para procesar datos textuales. Proporciona una API simple para sumergirse en tareas comunes de procesamiento de lenguaje natural (NLP), como el etiquetado de partes del discurso, la extracción de frases nominales, el análisis de sentimientos, la clasificación, la traducción y más.

Utilizando la función **drop** reducimos el data set para que contenga únicamente las columnas que nos interesan así como reducimos nuestra n de 568454 a 560000.

	Score	Time	Summary	Text
0	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...

Figura 2: Dataset reducido

La limpieza de nuestros datos consiste en retirar todos los caracteres especiales y numéricos para dejar solamente el alfabeto.

	Score	Time	Summary	Text	Cleaned Reviews
0	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...	I have bought several of the Vitality canned d...
1	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...	Product arrived labeled as Jumbo Salted Peanut...
2	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...	This is a confection that has been around a fe...
3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...	If you are looking for the secret ingredient i...
4	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...	Great taffy at a great price There was a wide ...

Figura 3: Limpieza del dataset

Se procede a separar el escrito por palabras y transformarlas todas a minúsculas para poder identificar las "stop words"que son palabras utilizadas en el cuerpo del documento pero no tienen significancia alguna vistas de manera individual. Al contar con las palabras relevantes se continua a seleccionar la representante de todas las formas flexionadas de una misma palabra para así finalizar el preprocesamiento de los datos.

	Score	Time	Summary	Text	Cleaned Reviews	POS tagged
0	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...	I have bought several of the Vitality canned d...	[(bought, v), (several, a), (Vitality, n), (ca...
1	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...	Product arrived labeled as Jumbo Salted Peanut...	[(Product, n), (arrived, v), (labeled, v), (Ju...
2	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...	This is a confection that has been around a fe...	[(confection, n), (around, None), (centuries, ...
3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...	If you are looking for the secret ingredient i...	[(looking, v), (secret, a), (ingredient, n), (...
4	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...	Great taffy at a great price There was a wide ...	[(Great, n), (taffy, n), (great, a), (price, n...

Figura 4: Part of Speech del dataset

Del dataset, se toman las columnas correspondientes al texto y a la lematización realizada

	Text	Lemma
0	I have bought several of the Vitality canned d...	buy several Vitality can dog food product fi...
1	Product arrived labeled as Jumbo Salted Peanut...	Product arrive label Jumbo Salted Peanuts pe...
2	This is a confection that has been around a fe...	confection around century light pillowy citr...
3	If you are looking for the secret ingredient i...	look secret ingredient Robitussin believe fi...
4	Great taffy at a great price. There was a wid...	Great taffy great price wide assortment yumm...
...	...	...
8449	I expected more flavor from W. Puck. Though th...	expect flavor W Puck Though coffee taste goo...
8450	I've tried all Vanilla coffee K-cups and by fa...	try Vanilla coffee K cup far best Wolf Gang ...
8451	I am not a huge coconut fan, so I wasn't excit...	huge coconut fan excite try one sip definite...
8452	This is my daughter's favorite K-cup product. ...	daughter favorite K cup product Mild medium ...
8453	WOW! That's what my husband said when he taste...	WOW husband say taste Wolfgang Puck Jamaica ...

Figura 5: Lematización

Analisis de sentimiento usando TextBlob

Trabajando con el texto y su lematización, se encuentra la polaridad Negativa, Neutral o Positiva de los comentarios mediante el uso de las funciones **getPolarity** y **analysis**

	Text	Lemma	Polarity	Analysis
0	I have bought several of the Vitality canned d...	buy several Vitality can dog food product fi...	0.466667	Positive
1	Product arrived labeled as Jumbo Salted Peanut...	Product arrive label Jumbo Salted Peanuts pe...	0.216667	Positive
2	This is a confection that has been around a fe...	confection around century light pillowy citr...	0.187000	Positive
3	If you are looking for the secret ingredient i...	look secret ingredient Robitussin believe fi...	0.150000	Positive
4	Great taffy at a great price. There was a wid...	Great taffy great price wide assortment yumm...	0.458333	Positive

Figura 6: Polaridad de los comentarios

Se realiza un conteo del total de reseñas con las que se cuentan según la polaridad

```
Positive    7527
Negative     784
Neutral      143
```

Figura 7: Conteo por polaridad

Para finalizar este análisis, se generó una gráfica de pastel para mostrar la distribución de la polaridad.

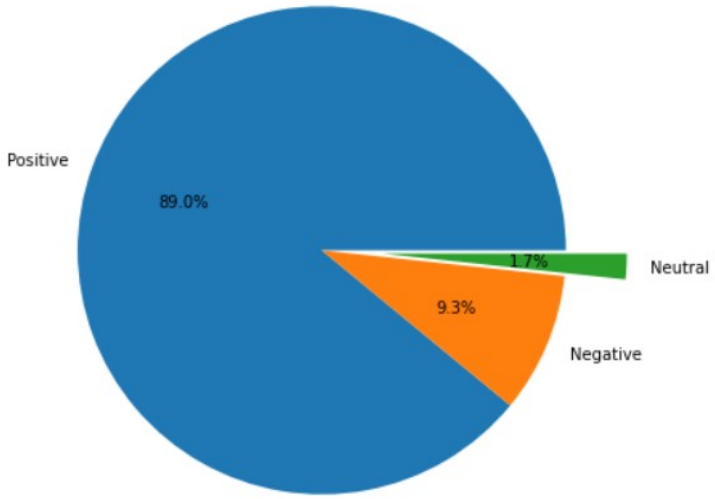

Figura 8: Gráfica de pastel para la polaridad

Análisis de sentimiento usando VADER

VADER usa una combinación de Un léxico de sentimiento es una lista de características léxicas(por ejemplo, palabras) que generalmente se etiquetan de acuerdo con su orientación semántica como positiva o negativa. VADERno solo habla sobre la puntuación de Positividad y Negatividad, sino que también nos dice qué tan positivo o negativo es un sentimiento.

Tomando de nuestra tabla el texto, su lematización así como la polaridad y análisis realizados previamente, se calcula el sentimiento y análisis correspondiente a VADER.

	Text	Lemma	Polarity	Analysis	Vader Sentiment	Vader Analysis
0	I have bought several of the Vitality canned d...	buy several Vitality can dog food product fi...	0.466667	Positive	0.9246	Positive
1	Product arrived labeled as Jumbo Salted Peanut...	Product arrive label Jumbo Salted Peanuts pe...	0.216667	Positive	-0.1027	Neutral
2	This is a confection that has been around a fe...	confection around century light pillowy citr...	0.187000	Positive	0.8532	Positive
3	If you are looking for the secret ingredient i...	look secret ingredient Robitussin believe fi...	0.150000	Positive	0.4404	Neutral
4	Great taffy at a great price. There was a wid...	Great taffy great price wide assortment yumm...	0.458333	Positive	0.9468	Positive

Figura 9: Análisis VADER

Se procede a realizar el conteo de reseñas por polaridad

```
Positive    7002
Neutral     1219
Negative     233
```

Figura 10: Conteo por polaridad VADER

Para apoyo visual, se generó una gráfica de pastel para mostrar la distribución de la polaridad VADER.

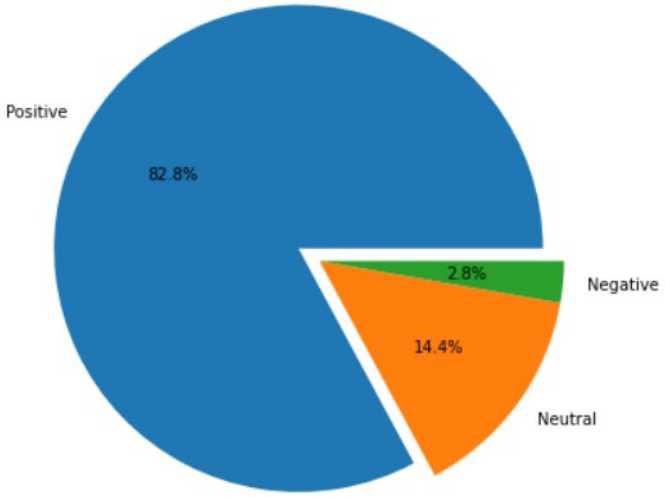

Figura 11: Gráfica de pastel para la polaridad VADER

Análisis de sentimiento usando SentiWordNet

SentiWordNet es un recurso léxico en el que cada sinset de WordNet está asociado a tres puntajes numéricos Obj(s), Pos(s) y Neg(s), que describen qué tan objetivos, positivos y negativos son los términos contenidos en el sinset. Un uso típico de SentiWordNet es enriquecer la representación del texto en aplicaciones de minería de opiniones (OM), agregando información sobre las propiedades relacionadas con el sentimiento de los términos en el texto.

Se toma la tabla previamente trabajada con el método VADER para agregar la polaridad calculada con este método

	Text	Lemma	Polarity	Analysis	Vader Sentiment	Vader Analysis	SWN analysis
0	I have bought several of the Vitality canned d...	buy several Vitality can dog food product fi...	0.466667	Positive	0.9246	Positive	Positive
1	Product arrived labeled as Jumbo Salted Peanut...	Product arrive label Jumbo Salted Peanuts pe...	0.216667	Positive	-0.1027	Neutral	Negative
2	This is a confection that has been around a fe...	confection around century light pillowy citr...	0.187000	Positive	0.8532	Positive	Neutral
3	If you are looking for the secret ingredient i...	look secret ingredient Robitussin believe fi...	0.150000	Positive	0.4404	Neutral	Positive
4	Great taffy at a great price. There was a wid...	Great taffy great price wide assortment yumm...	0.458333	Positive	0.9468	Positive	Positive

Figura 12: Análisis SentiWordNet

Se procede a realizar el conteo de reseñas por polaridad

```
Positive    6557
Negative    1585
Neutral      312
```

Figura 13: Conteo por polaridad SentiWordNet

Finalmente, se generó una gráfica de pastel para mostrar la distribución de la polaridad SentiWordNet

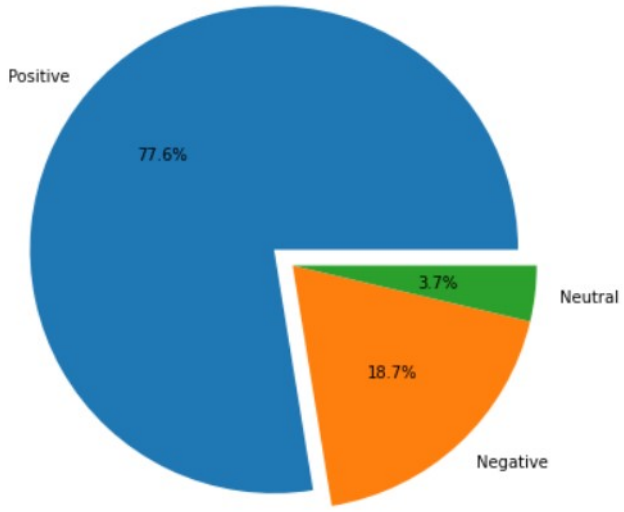

Figura 14: Gráfica de pastel para la polaridad SentiWordNet

Conclusiones

Tomando como referencia las tres gráficas de pastel, podemos notar que aunque se trabajó con la misma data, los resultados para la polaridad varían entre un análisis y otro

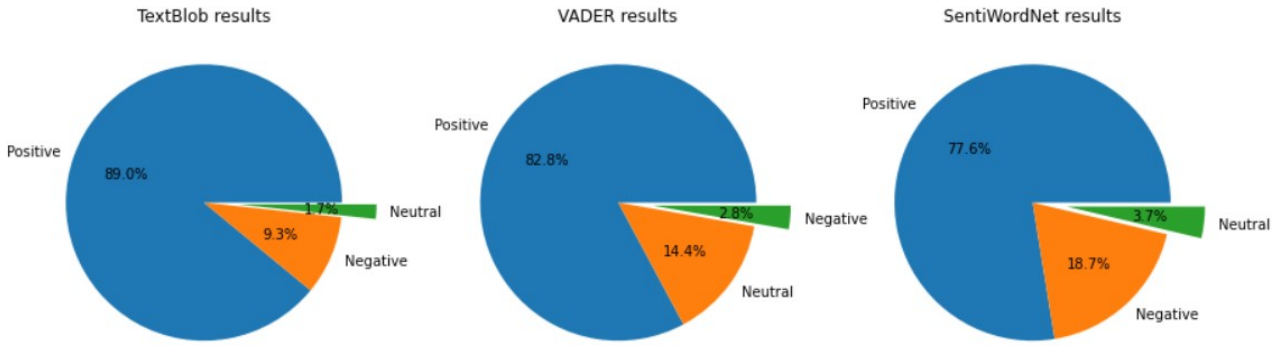

Figura 15: Gráficas de pastel para la polaridad

Si bien para los tres casos, en la mayoría de los comentarios son positivos, hay una diferencia significativa para el valor negativo de TextBlob con respecto a los demás, teniendo 5 unidades de distancia con respecto

a VADER y casi el doble contra SentiWordNet. Es aquí donde se debe determinar cuál de los análisis implementados en esta práctica se tomará como el más fiable.

Tomando en cuenta la asignación de valores numéricos que genera el análisis VADER para hacer la asignación de polaridad aún más sensible y específica hacia un sentimiento en específico, se selecciona este análisis como el mejor para este tipo de procesos.

Github

ClarasDinner
https://github.com/ClarasDinner/MCD_Procesamiento