Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Maestría en Ciencia de Datos

Tarea 2 - Análisis de sentimiento de tweets realizados con el hashtag #TheSocialDilemma

Autor:
Lic. Leobardo García Reyes

Supervisado por:
Dr. Mayra Cristina Berrones Reyes

26 de mayo del 2022

Análisis de sentimiento de tweets realizados con el hashtag #TheSocialDilemma

Introducción

Las redes sociales son una fuente enorme de recopilación de información, en estas diferentes usuarios comparten sus opiniones sobre diferentes temas, divididos entre positivos, neutrales o negativos. Pero realizar un análisis manual es imposible, por lo que se requiere de técnicas más eficaces e inteligentes que pueda analizar y proporcionar la polaridad de estos datos textuales.

La técnica de análisis de sentimiento, se trata de una clasificación masiva de textos de manera automática, la cual clasificará dichos textos en positivo, neutral o negativo ocupando un procesamiento de lenguaje natural. Dentro de esta técnica, existen diferentes librerías que ayudan a clasificar, en este documento se abarcaran 3: TEXTOBLOB, VADER y SENTIWORDNET.

Se empleará un análisis de sentimiento a tweets realizados con el hashtag #TheSocialDilemma y con ayuda de las 3 librerías se clasificarán estos tweets y se compararan entre si para observar las diferencias.

Figura 1: Fuente: Imagen obtenida de internet[10]

Datos

El conjunto de datos que se seleccionó para aplicar preprocesamiento de texto y analizar, se obtuvo a través del sitio web "Kaggle"[2]. El conjunto de datos es una recopilación de tweets realizados con el hashtag #TheSocialDilemma después del lanzamiento del documental "The Social Dilemma" el 9 de septiembre del 2020.

El conjunto de datos se extrajo con TwitterAPI y consta de 20,068 tweets de usuarios de Twitter de todo el mundo. Se proporciona de manera tabulada de la siguiente manera (Fig. 2):

- **user_name:** Nombre del usuario.
- **user_location:** La ubicación definida por el usuario para el perfil de esta cuenta.
- **user_description:** La descrición del perfil definida por el usuario de su cuenta.
- **user_created:** Hora y fecha en que se creó la cuenta.
- **user_followers:** El número de seguidores que tiene actualmente una cuenta.
- **user_friends:** El número de amigos que tiene actualmente una cuenta.
- **user_favourites:** El número de favoritos que tiene actualmente una cuenta.
- **user_verified:** Cuando es verdadero, indica que el usuario tiene una cuenta verificada.
- **date:** Fecha y hora UTC en que se creó el tweet.
- **text:** El texto real del tweet.
- **hashtags:** Todos los demás hashtags publicados en el tweet junto con #TheSocialDilemma.
- **source:** Herramienta utilizada para publicar el tweet. Los tweets del sitio web de Twitter tienen un valor de fuente: web.
- **is_retweet:** Indica si el tweet ha sido retwitteado por el usuario autenticado.
- **Sentiment:** Indica el sentimiento del tweet, consta de tres categorías: positivo, neutral y negativo.

user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source	is_retweet	Sentiment
Mari Smith	San Diego, Calif	Premier Facebook	11/09/2007 22:22	579942	288625	11610	False	16/09/2020 20:55	@musicmadmarc	@Social	Twitter Web	False	Neutral
Mari Smith	San Diego, Calif	Premier Facebook	11/09/2007 22:22	579942	288625	11610	False	16/09/2020 20:53	@musicma		Twitter Web	False	Neutral
Varun Tyagi	Goa, India	Indian | Tech Solu	06/09/2009 10:36	257	204	475	False	16/09/2020 20:51	Go watch		Twitter for iF	False	Positive
Casey Conw	Sydney, New S	Head of Diversity	28/12/2012 21:45	11782	1033	12219	True	16/09/2020 20:51	I watched	['TheSocialD	Twitter for iF	False	Negative
Charlotte Pa	Darlington	Instagram Charlot	28/05/2012 20:43	278	387	5850	False	16/09/2020 20:51	The problem	['TheSocialD	Twitter for iF	False	Positive
Denny Hulm	Manchester, England		06/11/2009 15:20	336	616	4748	False	16/09/2020 20:50	#TheSocialDi	['TheSocialD	Twitter for iF	False	Positive
Serkan Hicranlı±		KÃ¼Ã§Ã¼k kÃ¼Ã	18/11/2017 14:31	120	841	546	False	16/09/2020 20:49	@harari_yuv	['TheSocialD	Twitter for iF	False	Neutral
Laura	Kent	Mother, optimist,	05/02/2009 20:39	696	444	10551	False	16/09/2020 20:48	Erm #TheSoc	['TheSocialD	Twitter for iF	False	Negative
Eugeneð¥«	South Africa	AfricanðŸŒ | Mus	25/10/2012 09:16	2180	1570	18692	False	16/09/2020 20:48	#TheSocialDi	['TheSocialD	Twitter for A	False	Positive

Figura 2: Conjunto de datos en disposición. Fuente: Imagen realizada por mi.

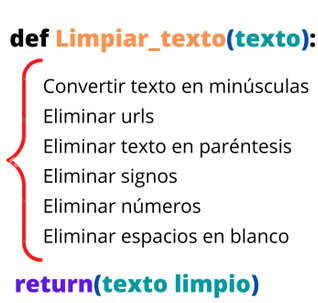
En la Fig. 3, se muestra la cantidad de tweets realizados durante casi un mes. En esta misma, se puede observar el aumento de tweets realizados durante 5 días, Llegando a su punto máximo el día 13-09-2020 con 2,005 tweets. Después de este pico fue decayendo y algo importante a destacar, es que a partir del día 24-09-2020, no existen tweets realizados con el hashtag #TheSocialDilemma hasta el día 09-10-2020.

Figura 3: Serie de tiempo de tweets realizados con el hashtag #TheSocialDilemma.

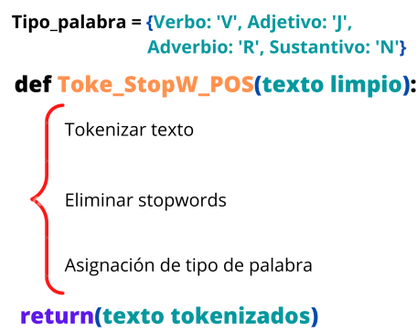
Metodología

Pre-procesamiento

En esta sección se describirá el preprocesamiento implementado en los tweets con el hashtag #TheSocialDilemma. El preprocesamiento, es utilizado para remover texto que no sea relevante para el futuro análisis como son las urls, signos, números, espacios en blanco, etc. Para esto, en la Fig. 4a, se explica la función empleada para limpiar el texto como primera parte.



(a) Función utilizada para realizar una primera limpieza de texto.



(b) Función utilizada para tokenizar, remover stopwords y lematizar.

Figura 4: Funciones utilizadas para limpiar, tokenizar, remover stopwords y lematizar. Fuente: Imágenes realizadas por mi.

Una vez realizado una primera limpieza, es importante remover las palabras que no aporten información adicional como son las stopwords. También, es importante dejar las palabras a su lema y para esto, se empleó el método de Lematización. La Lemmatization, incorpora información sobre la parte del discurso del término utilizando diccionarios y un análisis morfológico mas complejo.

Para aprovechar el potencial del método Lematización a cada palabra tokenizada, se le asignará una letra dependiendo su contexto (Fig. 4b). Las cuales serían: V - Verbo, J - Adjetivo, N - Sustativo, R - Adverbio, esto sirve para transformar y agrupar las palabras raíz de una mejor manera.

Una vez realizado todo este proceso de limpieza de texto, se decidió de igual forma, implementar un función que elimine aquellas palabras que tengan menos de 2 caracteres. También, se añadió al listado de stopwrods nuevas palabras como us, dont, im y números del 0 al 10 escritos en texto.

Resultados

Análisis de sentimiento

El análisis de sentimiento, es una de las tareas de procesamiento del lenguaje natural (NLP) más conocidas. Tiene como finalidad determinar la tendencia o actitud de un comunicador (Negativo, Neutral o Positivo) a través de la polaridad contextual de su escritura. Existen diferentes formas de realizar un análisis de sentimiento, entonces, para propósito del documento se escogieron 3 los cuales son: TextBlob, Vader y SentiWordNet.

TEXTBLOB

Esta librería se utiliza para asignar puntajes de polaridad y subjetividad a los textos. La polaridad es un valor que se encuentra entre [-1, 1], en donde -1 identifica a las palabras como negativas, 0 las identifica como neutrales y 1 como palabras positivas.Y la subjetividad, se encuentra entre [0,1] y cuantifica la cantidad de opinión personal e información fáctica contenida en el texto. Es decir, cuando es 1 significa que el texto contiene una opinión personal en lugar de información fáctica.

Si se gráfica la polaridad y la subjetividad (Fig. 5), se puede observar que se forma un cono, en donde hay más valores entre 0 y 0.5. Esto puede dar indicios a que hay más tweets positivos y neutrales. Lo mismo pasa con la subjetividad, hay más valores cercanos a 0, por lo que la mayoría son tweets fácticos.

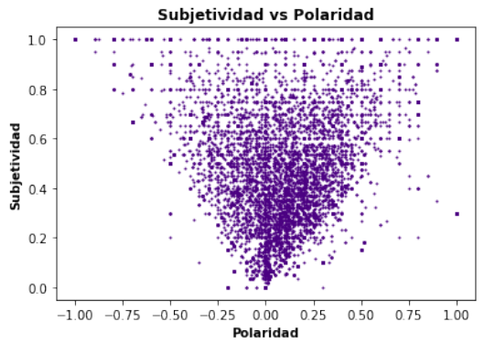


Figura 5: Gráfico entre Subjetividad y Polaridad.

VADER

VADER (Valence Aware Dictionary and Sentiment Reasoner) es otro analizador de sentimientos basado en Lexicon que tiene reglas predefinidas para palabras o léxicos.VADER tiene 4 categorías, positivo, neutral, negativo y compuesto. Este último, es una puntuación indispensable que se calcula normalizando las otras 3 puntuaciones de positivo, neutral y negativo y va entre [-1, 1] que indica qué tan positiva, negativa o neutral es el tweet.

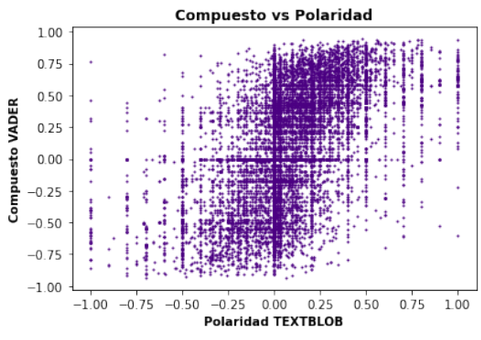
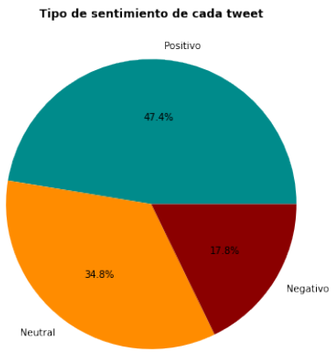


Figura 6: Gráfico entre Compuesto VADER y Polaridad TEXTBLOB.

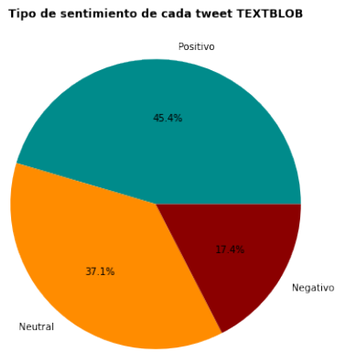
Si se gráfica el compuesto VADER y la polaridad TEXTBLOB (Fig. 6), se puede observar como se divide en 4 cuadrantes, en donde el cuadrante 1 y 3, es donde coinciden (positivo y negativo). Mientras que en los cuadrantes 2 y 4, es donde se confunden, hace recordar a la matriz de confusión.

SENTIWORDNET

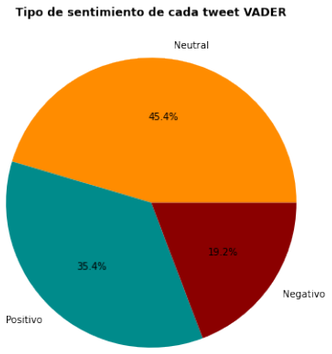
SentiWordNet opera de acuerdo al texto proporcionada por WordNet. Es decir, primero se lematiza las palabras para buscarlas en el diccionario y asignarles un valor de acuerdo si es positivo, negativo o neutral. Las tres puntuaciones oscilan entre los valores [0,1].



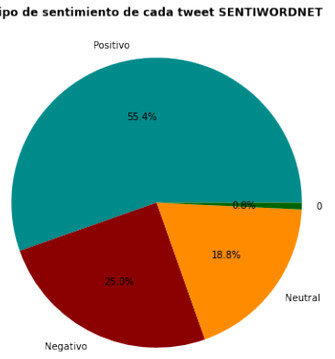
(a) Gráfico circular de tweets categorizados en positivo, neutral y negativo (Real).



(b) Gráfico circular de tweets categorizados en positivo, neutral y negativo (TEXTBLOB).



(c) Gráfico circular de tweets categorizados en positivo, neutral y negativo (VADER).



(d) Gráfico circular de tweets categorizados en positivo, neutral y negativo (SENTIWORDNET).

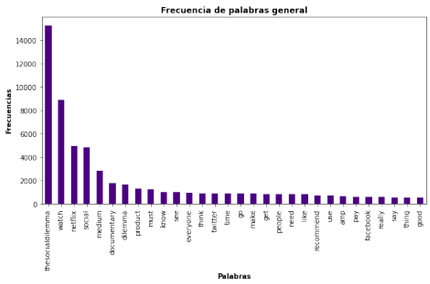
Figura 7: Comparación de tweets categorizados con la base real y los 3 métodos de análisis de sentimiento.

En la Fig. 7 se observa el porcentaje que representa cada categoría (positiva, neutral y negativa) utilizando los 3 métodos de análisis de sentimiento. Suponiendo que la categorización de la Fig. 7a es la real y esta bien (es la que se proporciona al descargar el conjunto de datos), se puede observar que el método TEXTBLOB (Fig. 7b) es el que más se parece. Seguido del método SENTIWORDNET (Fig. 7d), pero confunde entre negativo y neutral, además, aparece una etiqueta de 0 que representa el 0.8 % debido a que no halló donde categorizar unas palabras. Algo parecido sucede con el método VADER (Fig. 7c), ya que confunde positivo con neutral. En este ultimo, se especifico un compuesto de mayores o igual a 0.2 como positivos, menores o igual a -0.2 como negativo y neutral en otro caso.

En este ultimo (Fig. 7c), se especifico un compuesto de mayores o igual a 0.2 como positivos, menores o igual a -0.2 como negativo y neutral en otro caso.

Resultados generales

Como se puede apreciar en la Fig. 8, se puede obtener información de donde ver la película por la palabra "netflix", que la recomiendan por palabras como "wacht", "recommend", "like", "good". También, tal vez algunos expresaron su opinión después de verla por palabras como "facebook", "twitter", "really", "think", "illegal", "change".

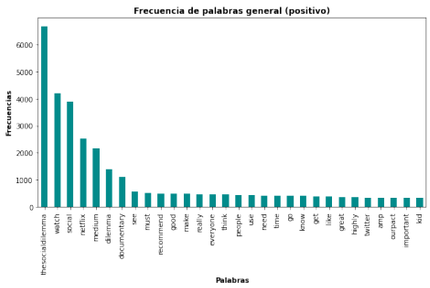


(a) Frecuencia de palabras de tweets realizados con el hashtag #TheSocialDilemma.

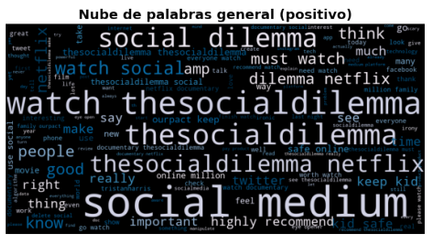
(b) Nube de palabras de tweets realizados con el hashtag #TheSocialDilemma.

Figura 8: Frecuencia y nube de palabras de tweets realizados con el hashtag #TheSocialDilemma.

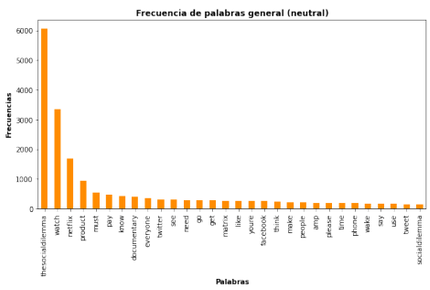
Se mencionó como es que el método VADER confundió neutral con positivo si se compara con la categorización original. En la Fig. 9 y la Fig. 10, no hay una gran diferencia entre cada una de sus categorías, la suposición que establezco, es que tal vez el método VADER tome más palabras neutrales pareciendo que se confundió.



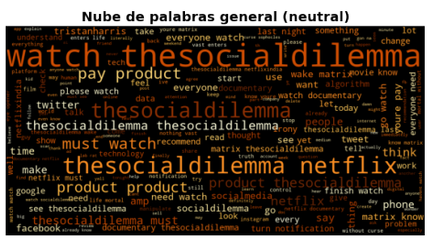
(a) Frecuencia de palabras de tweets realizados con el hashtag #TheSocialDilemma (positivo).



(b) Nube de palabras de tweets realizados con el hashtag #TheSocialDilemma (positivo).

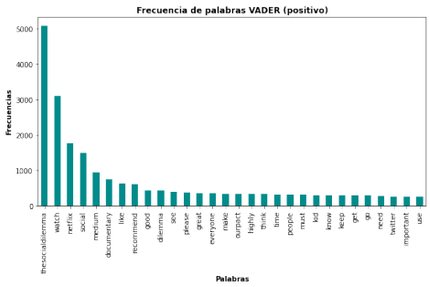


(c) Frecuencia de palabras de tweets realizados con el hashtag #TheSocialDilemma (neutral).



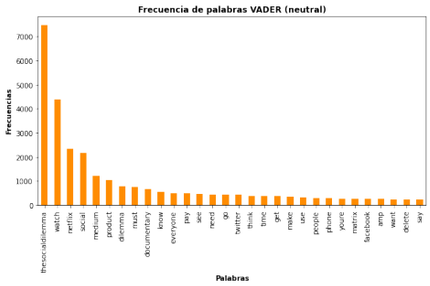
(d) Nube de palabras de tweets realizados con el hashtag #TheSocialDilemma (neutral).

Figura 9: Frecuencia y nube de palabras de tweets realizados con el hashtag #TheSocialDilemma (positivo y neutral).

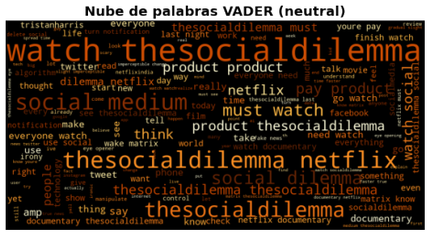


(a) Frecuencia de palabras de tweets realizados con el hashtag #TheSocialDilemma (VADER, positivo).

(b) Nube de palabras de tweets realizados con el hashtag #TheSocialDilemma (VADER, positivo).



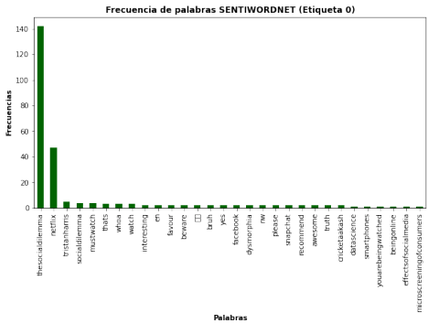
(c) Frecuencia de palabras de tweets realizados con el hashtag #TheSocialDilemma (VADER, neutral).



(d) Nube de palabras de tweets realizados con el hashtag #TheSocialDilemma (VADER, neutral).

Figura 10: Frecuencia y nube de palabras de tweets realizados con el hashtag #TheSocialDilemma (VADER, positivo y neutral).

Por último, algo que también llamo mi atención, fue el método SENTIWORDNET el etiquetado con 0. En la Fig. 11, se puede observar que son palabras que no categorizo, porque no logro asociarlas con positivo, neutral o negativo. Al parecer las palabras que están en esa categoría, son nombre de usuarios, palabras que están juntas, caracteres extraños y palabras de 2 caracteres.



(a) Frecuencia de palabras de tweets realizados con el hashtag #TheSocialDilemma (SENTIWORDNET, etiqueta 0).

(b) Nube de palabras de tweets realizados con el hashtag #TheSocialDilemma (SENTIWORDNET, etiqueta 0).

Figura 11: Frecuencia y nube de palabras de tweets realizados con el hashtag #TheSocialDilemma (SENTIWORDNET, etiqueta 0).

Para concluir el trabajo se realizó un modelo de aprendizaje automático utilizando Random Forest. Se obtuvo una exactitud del 82.51 %. Aquí se muestra la matriz de confusión:

	Negativo	Neutral	Positivo
Negativo	368	146	190
Neutral	34	1236	56
Positivo	59	217	1708

Conclusiones

Al comparar las 3 librerías utilizadas para realizar el análisis de sentimiento, se tiene que ser muy precavido y tener claro que se quiere analizar, ya que con cada una se puede obtener diferentes resultados y alguna puede ser más beneficiosa que otra.

Es muy importante tener un buen preprocesamiento, ya que en los resultados se puede ver reflejado, por ejemplo con SENTIWORDNET. Al no limpiar bien los tweets, encontró palabras que no logro clasificar, porque no las encontraba en el diccionario.

Así como este documento analizó los tweets de gente que escribió lo que sintió al ver la película "The Social Dilemma", se puede abarcar áreas más importantes en donde implique comentarios de gente al comprar un producto, contratar un servicio, etc.

Referencias

[1] GitHub https://github.com/Zarcklet/ProcesamientoClasificacionDatos

[2] Kash. (2021). The Social Dilemma tweets - text classification [Data set]. https://www.kaggle.com/datasets/kaushiksuresh147/the-social-dilemma-tweets.

[3] Ahmad, M. (2017). Machine Learning Techniques for Sentiment Analysis: A Review. International journal of multidisciplinary sciences and engineering. 8(3), 27-32. http://www.ijmse.org/Volume8/Issue3/paper5.pdf.

[4] Ahuja, S., & Dubey, G. (2017). Clustering and sentiment analysis on Twitter data. 2017 2nd International Conference on Telecommunication and Networks (TEL-NET). https://doi.org/10.1109/TEL-NET.2017.8343568.

[5] Baccianella, S. et al. (2010). Sentiwordnet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf.

[6] Elbagir, S. & Jing Yang (2019). Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment. Proceedings of the International MultiConference of Engineers and Computer Scientists 2019 (IMECS 2019). http://www.iaeng.org/publication/IMECS2019/IMECS2019_pp12-16.pdf.

[7] Guerini, M. (2013). Sentiment Analysis: How to Derive Prior Polarities from SentiWordNet. ARXIV. https://doi.org/10.48550/arXiv.1309.5843.

[8] Kumar, M. (20 de octubre del 2021). Sentiment Analysis with TextBlob and Vader. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/sentiment-analysis-with-textblob-and-vader/.

[9] Neethu, M. S., & Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). https://doi.org/10.1109/ICCCNT.2013.6726818.

[10] Netflix (2020). The social dilemma [Fotografía]. Jeff Orlowski. https://www.netflix.com/title/81254224.

[11] Ohana, B. & Tierney, B. (2009). Sentiment classification of reviews using SentiWordNet. 9th. IT&T Conference, Technological University Dublin, Dublin, Ireland, 22-23 October. https://doi.org/10.26438/ijcse/v6i11.719726.

[12] Praveen, J. & Prasanna, H. (2021). Sentiment Analysis:Textblob For Decision Making. International Journal of Scientific Research & Engineering Trends. 7(2). https://ijsret.com/wp-content/uploads/2021/03/IJSRET_V7_issue2_289.pdf.

[13] Shah, P. (27 de junio del 2020). Sentiment Analysis using TextBlob. Sentiment Analysis using TextBlob and its working! Medium. https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524.

[14] Shahul, ES. (3 de diciembre del 2021). Sentiment Analysis in Python: TextBlob vs Vader Sentiment vs Flair vs Building It From Scratch. NeptuneBlog. https://neptune.ai/blog/sentiment-analysis-python-textblob-vs-vader-vs-flair.

[15] Sharma, S. (30 de junio del 2021). Sentiment Analysis Using the SentiWordNet Lexicon. Medium. https://srish6.medium.com/sentiment-analysis-using-the-sentiwordnet-lexicon-1a3d8d856a10#:~:text=Sentiment%20Analysis%20is%20the%20computational,mining.