Análisis De Sentimiento

AMÉRICA VICTORIA RAMÍREZ CÁMARA
Facultad de Ciencias Físico Matemáticas
Universidad Autónoma de Nuevo León
Nuevo León, México
america.ramirezcm@uanl.edu.mx

I. Introducción

El análisis de sentimientos (o minería de opinión) es una técnica de procesamiento del lenguaje natural (PNL) utilizada para determinar si los datos son positivos, negativos o neutros. El análisis de sentimientos a menudo se realiza en datos textuales para ayudar a las empresas a monitorear el sentimiento de marca y producto en la retroalimentación del cliente, y entender las necesidades del cliente.

El análisis de sentimientos se centra en la polaridad de un texto (positivo, negativo, neutro) pero también va más allá de la polaridad para detectar sentimientos y emociones específicas (enojado, feliz, triste, etc), urgencia (urgente, no urgente) e incluso intenciones (interesado v. no interesado).

Dependiendo de cómo quiera interpretar los comentarios y consultas de los clientes, puede definir y adaptar sus categorías para satisfacer sus necesidades de análisis de sentimientos.

Dado que los seres humanos expresan sus pensamientos y sentimientos más abiertamente que nunca, el análisis de sentimientos se está convirtiendo rápidamente en una herramienta esencial para monitorear y entender los sentimientos en todo tipo de datos.

El análisis automático de la retroalimentación de los clientes, como las opiniones en las respuestas a las encuestas y las conversaciones en las redes sociales, permite a las marcas aprender lo que hace felices o frustrados a los clientes, para que puedan adaptar los productos y servicios a las necesidades de sus clientes.

Los beneficios generales del análisis de sentimientos incluyen:

- Clasificación de datos a escala: ¿Te imaginas ordenar manualmente a través de miles de tweets, conversaciones de atención al cliente o encuestas? Hay demasiados datos de negocio para procesarlos manualmente. El análisis de sentimientos ayuda a las empresas a procesar grandes cantidades de datos no estructurados de una manera eficiente y rentable.
- Análisis en tiempo real: El análisis de sentimientos puede identificar problemas críticos en tiempo real, por ejemplo, ¿se está intensificando una crisis de relaciones públicas en las redes sociales? ¿Está un cliente enojado a punto de batir? Los modelos de análisis de sentimientos pueden ayudarle a identificar inmediatamente este tipo de situaciones, para que pueda tomar medidas de inmediato.
- Criterios coherentes: Se estima que la gente solo está de acuerdo alrededor del 60-65% del tiempo cuando se determina el sentimiento de un texto en particular. Etiquetar el texto por sentimiento es altamente subjetivo, influenciado por experiencias personales, pensamientos y creencias. Mediante el uso de un sistema centralizado de análisis de sentimientos, las empresas pueden aplicar los mismos criterios a todos sus datos, ayudándoles a mejorar la precisión y obtener mejores conocimientos.

En esta tarea, se realiza un análisis de sentimientos utilizando la limpieza de información, stopwords, lematización, y tres librerías en particular llamadas: textblob, VADER y sentiwordnet.

II. Planteamiento del problema

El dataset analizado es: "rotten_tomatoes_critic_reviews.csv", en el cual, contiene todas las críticas para la película disponible en Rotten Tomatoes. El dataset mencionado fué descargado de la página Kaggle, sin embargo, la fuente oficial es de la página "Rotten Tomatoes", a continuación se expone la definición de estos conceptos:

- Stopwords: Las palabras de parada son cualquier palabra en una lista de parada (o stoplist o diccionario negativo) que se filtra (es decir, se detiene) antes o después del procesamiento de los datos del lenguaje natural (texto). No existe una lista universal única de palabras clave utilizadas por todas las herramientas de procesamiento del lenguaje natural, ni ninguna regla acordada para identificar las palabras clave, y de hecho no todas las herramientas incluso utilizan dicha lista. Por lo tanto, cualquier grupo de palabras puede ser elegido como las palabras clave para un propósito determinado. La "tendencia general en los sistemas de [recuperación de información] a lo largo del tiempo ha sido desde el uso estándar de listas de parada bastante grandes (200-300 términos) a listas de parada muy pequeñas (7-12 términos) a ninguna lista de parada en absoluto.
- Lematización: Es una técnica en la recuperación de datos en los sistemas de información (RDSI), esta técnica sirve para reducir variantes morfológicas de la formas de una palabra a raíces comunes o lexemas; con el fin de mejorar la habilidad de los motores de búsqueda para mejorar las consultas en documentos. Básicamente, este consiste en remover el plural, el tiempo, o los atributos finales de la palabra.
- Textblob: Utiliza NLTK (Natural Language ToolKit) y la entrada contiene una sola frase, La salida de TextBlob es la polaridad y la subjetividad. El puntaje de polaridad se

encuentra entre (-1 a 1) donde -1 identifica las palabras más negativas tales como 'disgusto', 'horrible', 'patético', y 1 identifica las palabras más positivas como 'excelente', o 'mejor'. Puntaje de subjetividad se encuentra entre (0 y 1), Muestra la cantidad de opinión personal, Si una oración tiene alta subjetividad i.e. cerca de 1, se asemeja a que el texto contiene más opinión personal que información fáctica. Como TextBlob es un analizador de sentimientos basado en Lexicón tiene algunas reglas predefinidas o podemos decir diccionario de palabras y peso, donde tiene algunas puntuaciones que ayudan a calcular la polaridad de una oración. Es por eso que los analizadores de sentimientos basados en Lexicón también se llaman "Analizadores de sentimientos basados en reglas".

- VADER: VADER (Valence Aware Dictionary and Sentiment Reasoner) es otro analizador de sentimientos basado en Lexicon que tiene reglas predefinidas para palabras o léxicos. El VADER no solo dice que el léxico es positivo, negativo o neutro, sino que también dice cuán positiva, negativa o neutra es una oración. La salida de VADER viene en un diccionario de Python en el que tenemos cuatro claves y sus valores correspondientes. 'neg', 'neu', 'pos', y 'compound' que significa Negativo, Neutro y Positivo respectivamente. La puntuación compuesta es una puntuación indispensable que se calcula mediante la normalización de las otras 3 puntuaciones (negativo, neu, pos) entre -1 y +1. Los criterios de decisión son similares a TextBlob -1 es para los más negativos y +1 es para los más positivos.
- Sentiwordnet: SentiWordNet es un recurso léxico en el que cada conjunto de términos de WordNet está asociado a tres puntuaciones numéricas Obj(s), Pos(s) y Neg(s), describiendo cuán objetivos, positivos y negativos son los términos contenidos en el conjunto de términos.Un uso típico de SentiWordNet es enriquecer la representación de texto en aplicaciones de minería de opinión (OM), agregando información sobre las propiedades relacionadas con el sentimiento de los términos en el texto. OM es una subdisciplina reciente en la encrucijada de la recuperación de información y la lingüística computacional que se refiere no al tema de un documento, sino a la opinión que expresa. OM tiene un rico conjunto de aplicaciones, que van desde el seguimiento de las opiniones de los usuarios sobre los productos o sobre los candidatos políticos, como se expresa en los foros en línea, a la gestión de las relaciones con los clientes. Con el fin de ayudar a la extracción de opiniones del texto, la investigación reciente ha intentado determinar automáticamente la 'polaridad PN' de los términos subjetivos, es decir. identificar si un término que es un marcador de contenido opinionado tiene una connotación positiva o negativa. La investigación para determinar si un término es efectivamente un marcador de contenido obstinado (un término subjetivo) o no (un término objetivo) ha sido, en cambio, mucho más escasa. SentiWordNet es el primer recurso léxico que proporciona un nivel de detalle tan específico (el sentido de la palabra representado por un synset) y una cobertura tan amplia (todos los más de 115.000 synsets de WordNet). El método utilizado para desarrollar SentiWordNet se basa en el análisis cuantitativo de las glosas asociadas a los synsets, y en el uso de las representaciones de términos vectoriales resultantes para la clasificación semisupervisada de synset. Las tres puntuaciones se obtienen combinando los resultados obtenidos por un comité de ocho clasificadores ternarios, todos ellos caracterizados por niveles de precisión similares pero con un comportamiento de clasificación diferente.

Estos conceptos se aplican en la solución propuesta para realizar un análisis de sentimiento de las críticas sobre tres películas estrenadas en diferentes años(2002, 2012 y 2017) con diferentes actores y actrices, contando la misma historia sobre un superheroe del MCU (Marvel Cinematic Universe): Spider-Man.

- Spider-Man(2002): Esa primera película que seguramente todos conocemos del hombre araña es Spider-Man, publicada en 2002 y protagonizada por Tobey Maguire. Peter Parker vive en casa de sus tíos que, tras la muerte de sus padres, lo criaron como a su propio hijo. Un día de visita a unos laboratorios, una araña radioactiva muerte a Peter dotándolo de todos sus poderes arácnidos. Así este se convierte en Spider-Man, teniendo que luchar contra todos los peligros de Nueva York y, en concreto, contra uno de sus némesis: el Duende Verde.
- The Amazing Spider-Man(2012): Hasta 2012 tuvimos que esperar para volver a ver al hombre araña en la gran pantalla con la entrega de The Amazing Spider-Man. En este caso Andrew Garfield encarnaba a Peter Parker pero aquí era una persona adolescente, más joven. Sus padres también fallecieron hace años y Peter intenta descubrir quién es hasta que encuentra un hilo del que tirar. Un secreto que su padre guardaba y que lo llevará a ser, nuevamente, el compañero y vecino Spider-Man. En ese momento tendrá que luchar contra su enemigo, Lagarto.
- Spider-Man Homecoming(2017): La primera vez que vimos a Tom Holland encarnar al hombre araña no fue en 2017 con la entrega de Spider-Man: Homecoming, sino que se presentó en las películas de los Vengadores como un nuevo Spider-Man más joven que nunca. Después de lo que vivió en las diferentes entregas de la tercera fase del UCM, Peter vuelve a casa con su tía May. Sin embargo, ahora cuenta con Tony Stark como mentor, que le ayudará a seguir su camino como superhéroe. Aunque este chico intenta vivir una vida normal como adolescente, la aparición de un nuevo villano llamado Buitre hará que Spider-Man tenga que hacer acto de presencia.

III. Solución Propuesta

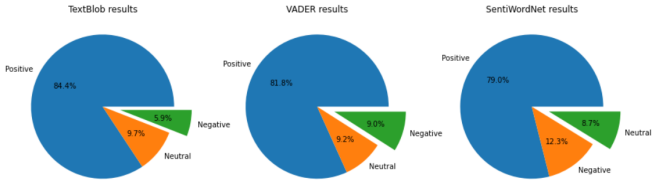
La solución propuesta para realizar este preprocesamiento de datos de texto es implementar; limpieza de información, stopwords, lematización y para el análisis de sentimiento se utilizaron tres librerías: textblob, VADER y sentiwordnet. Solo la limpieza de información se realizó de manera general, el resto del proceso se realizó tres veces, debido a que el análisis es para cada película.

- Limpieza de información: Para este proceso, lo que se realizó fue filtrar el archivo csv, mencionado previamente, mediante la columna ”rotten_tomatoes_link” (enlace desde el que se han adquirido los datos de las películas) para elegir las tres películas a analizar: m/spider_man_homecoming, m/the_amazing_spider_man y m/spiderman, creando un nuevo archivo llamado dataset.csv. Una vez guardada esta información en el nuevo archivo, se eliminaron registros duplicados y valores nulos de la columna ”review_content” (comentarios del crítico) en el mismo archivo. Luego se eliminaron las siguientes columnas: ’critic_name’, ’top_critic’, ’publisher_name’, ’review_type’, ’review_score’ ya que no son requeridos para el análisis. Finalmente, se sustraen solamente palabras con letras de la a-z y A-Z en la columna ”review_content”.
- Stopwords: Se eliminan las stopwords en idioma inglés utilizado la librería NLTK, se agregan las siguientes palabras: ’Amazing’, ’amazing’, ’AMAZING’. El motivo de agregar estas palabras fue porque en las críticas para la película ”The Amazing Spider-Man”, esta palabra es muy frecuente debido a que se menciona el título y no por ser una palabra positiva del crítico.
- Lematización: Se utiliza la función lematize utilizando la librería NLTK.
- Textblob: Se utiliza esta librería clasificando la polaridad de las críticas, después de los procesos mencionados anteriormente, como menores a 0 como negativas, iguales a 0 como neutras y mayores a 0 como positivas.
- VADER: Se utiliza esta librería clasificando la puntuación compuesta de las críticas, después de los procesos mencionados anteriormente, como menores a 0 como negativas, iguales a 0 como neutras y mayores a 0 como positivas.
- Sentiwordnet: Se utiliza esta librería clasificando las puntuaciones de las críticas, después de los procesos mencionados anteriormente, como menores a 0 como negativas, iguales a 0 como neutras y mayores a 0 como positivas.

IV. Análisis de resultados

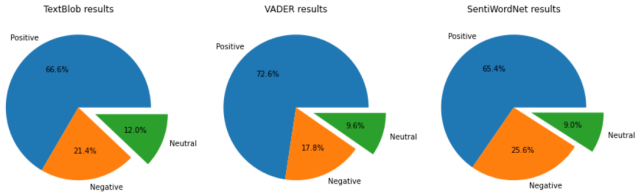
Para la película Spider-Man: Homecoming(2017) los resultados fueron:

Librería	Positivos	Negativos	Neutros
Textblob	329	23	38
VADER	319	35	36
Sentiwordnet	308	48	34



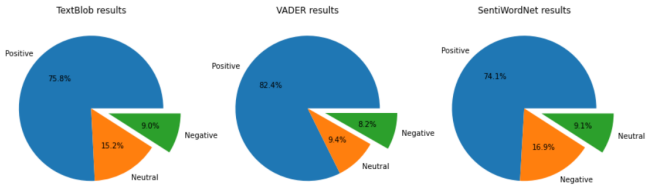
Para la película The Amazing Spider-Man(2012) los resultados fueron:

Librería	Positivos	Negativos	Neutros
Textblob	221	71	40
VADER	241	59	32
Sentiwordnet	217	85	30



Para la película Spider-Man(2002) los resultados fueron:

Librería	Positivos	Negativos	Neutros
Textblob	185	22	37
VADER	201	20	23
Sentiwordnet	180	41	22



V. Conclusiones

Para Spiderman Homecomming, los resultados utilizando las tres librerías fueron muy variados, a comparación de The Amazing Spiderman y Spiderman donde textblob y sentiwordnet resultaron casi parecidos. A pesar de que Spiderman tiene mucho menos críticas, los resultados de críticas positivas fueron mayores a los de The Amazing Spiderman. A pesar de que la clasificación del puntaje como mayores a 0 positivas, menores a 0 negativas e iguales a 0 neutras, para las tres librerías, VADER, resultó tener resultados muy diferentes a comparación de Textblob y Sentiwordnet. En mi opinión, me quedaría con Textblob y Sentiwordnet porque ambos cuentan con resultados parecidos a lo que me da entender que por coincidir con resultados similares en mas de una librería pueden ser más precisos.

References

[1] Github. (2022) ProcesamientoDeDatos. [Online]. Available: https://github.com/vickyrmz24/ProcesamientoDeDatos/blob/main/Procesamiento_de_datos___Tarea_2.pdf
[2] Github. (2022) ProcesamientoDeDatos. [Online]. Available: https://github.com/vickyrmz24/ProcesamientoDeDatos/blob/main/Tarea2.ipynb
[3] Github. (2012) ProcesamientoDeDatos. [Online]. Available: https://github.com/vickyrmz24/ProcesamientoDeDatos/blob/main/dataset.csv
[4] Stefano Leone. (2020) Rotten Tomatoes movies and critic reviews dataset. [Online]. https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset?resource=download&select=rotten_tomatoes_critic_reviews.csv
[5] Monkeylearn. (2022) Sentiment Analysis: A Definitive Guide. [Online]. Available: https://monkeylearn.com/sentiment-analysis/