Análisis de texto

Leslye Marisol Hernandez Bolaños $\|Matrícula$: 1819111

Universidad Autónoma de Nuevo León
Facultad de Ciencias Físico Matemáticas
Maestría en Ciencia de Datos
Procesamiento y Clasificación de Datos

marisol.hernandezblns@uanl.edu.mx

***Resumen*—Este trabajo presenta una comparación de análisis de texto entre dos fuentes con diferentes objetivos, el primero se basa en un conjunto de reseñas extraídas de la plataforma de ventas de e-commerce; Amazon, para la categoría de *All Beauty*, mientras el segundo análisis hace referencia al libro de la autora Louisa May Alcott, *"Little Woman"*.**
Utilizando técnicas de procesamiento de lenguaje natural, estadistíca descriptiva, se pretende presentar las principales diferencias entre ambos textos, en cuestión de la frecuencia de palabras, signos de puntuación, lematización y características que distinguen a un texto contra el otro, esto con el uso de paquetes y librerías del lenguaje Python.

***Index Terms*—Análisis de texto, Procesamiento de lenguaje natural.**

I. Introducción

El catedrático del Instituto Cervantes, Agustín Vera Luján describe al texto lingüístico como un acto de la comunicación por el que dos o mas personas se transmiten información, o a la unidad a través de la que se lleva a cabo tal acto de comunicación, ya sea, por oral o escrita.

Enfocándonos al tipo de texto escrito, este es contituído por un objetivo, lo que distingue su contenido, estructura y audiencia a la cual va dirigida, es así en como se pueden agrupar en diferentes tipos de textos:

- **Texto expositivo:** Contiene una estructura textual que está ligada al análisis y la síntesis, su principal objetivo es explicar o exponer algo en términos de su tema principal, considera las relaciones causa-efecto o de comparación, entre otros aspectos.
  Algunos ejemplos de este tipo son los Folletos, libros de texto, enciclopedias, manuales, exposiciones orales o museísticas, disertación, conferencia, ponencia, informes, apuntes, exámenes.
- **Texto narrativo:** Texto en el que predomina la narración, se distingue porque desarrolla una historia en un tiempo y un espacio determinado (real o imaginario) y además las acciones con personajes tienen un desarrollo dentro de la historia, estos personajes pueden ser personas, animales o cosas humanizadas.
  Algunos géneros del texto narrativos son: cuento, leyenda, novela, noticia, crónica, chiste, historieta, textos de historia, cartas familiares, biografía, videoclip, videojuego, testimonio, película de ficción o documental.
- **Texto argumentativo:** Se distingue por una organización textual centrada en el juicio y en la toma de posición respecto de algún asunto polémico. Este tipo de texto contiene la defensa de una un tema sustentada con argumentos, considera una crítica sobre un hecho, causa o circunstancia.
  Entre los géneros de este tipo están: asambleas, debates, texto político, texto publicitario, artículo de opinión, entrevista, cartón político, editorial, columna, reseña crítica, etc.

En el documento se enfocará en estos últimos dos, ya que, la novela "Little Women"pertence al tipo de texto narrativo, por otro lado, las reseñas son del tipo argumentativo, ambos si bien van al público en general, tienen diferentes enfoques y objetivos, los cuales, por medio de la estadítica descritiva se irán exponiendo.

II. Estructura, análisis estadístico

Las novelas siguen una forma de escritura estructurada, la que incluye dintigidos componentes como la trama, el desarrollo y la problemática.

Por otro lado, las reseñas utilizan una estructura diferente, generalmente está la introducción en la que se expone la perspectiva del redactor, después se encuentra el cuerpo de la reseña en donde se expone como más enfásis su argumento y por último se encuentra la conclusión, en donde resalta su opinión. Sin embargo, hoy en día la estructura de una reseña se ha ido modificando, ya que ésta se escribe en sitios dónde la estructura y la grámatica no es el principal objetivo sino que se le da más importancia al objetivo de la misma, es decir , al argumento.

II-A. Novela

La novela fue extraída de www.gutenberg.org, por medio de la librería gutenbergpy. Después de una limpieza, descodificación, se obtuvieron los siguientes resultados:

Métrica	No. de carácteres	Total de palabras en la oración
Promedio	59.764178	11.276781
Desviación	17.558014	3.701438
Mínimo	3	1
Máximo	71	19
1er Cuartil	63	10
Mediana	68	12
3er Cuartil	70	14

Tabla I
TABLA DE ESTADÍSTICA DESCRIPTIVA DE "LITTLE WOMEN ELABORACIÓN PROPIA.

No.	A	B	C	D	E
1	1	1	0	0	0
2	0	0	1	1	1
3	0	0	0	0	0

Tabla II
A: ANY PRESENTS GRUMBLED
B: PRESENTS GRUMBLED JO
C: SIGHED MEG LOOKING
D: SO DREADFUL TO
E: TO BE POOR

Palabras

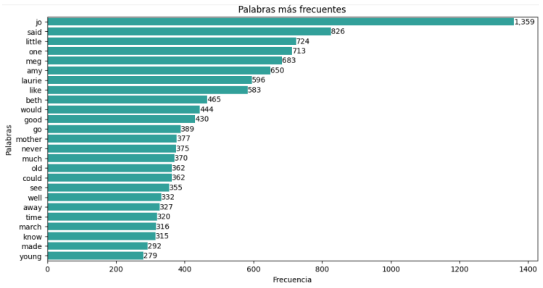

Figura 1. Top 20 de palabras más frecuentes en el texto - Elaboración propia.

Con la información de la figura 1, se observa en que dentro del top 20 de las palabras con mayor frecuencia en el texto están los nombres de los personajes, pues estos son los de mayor relevancia dentro de la obra. De igual forma, dentro del top 20, se encuentran verbos en pasado, pues la novela cuenta una historia que ya ha ocurrido.

Signos de puntuación

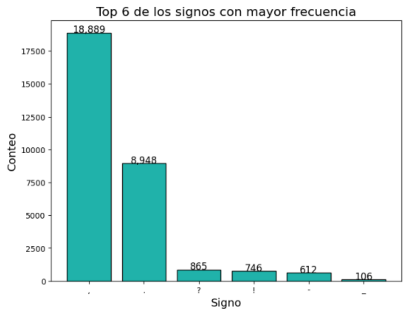

Figura 2. Top 6 de signos más frecuentes en el texto - Elaboración propia.

La figura 2, nos da una visualización del tipo de signos que se encuentran en el texto. Entre estos se encuentran los signos de coma, punto y signos de interrogación, ya que al ser un texto literario es fundamental el uso de los primeros dos signos mientras al existir diálogos entre personajes, es importante añadir los signos de admiración e interrogación para evocar el sentimiento al lector.

Grámatica

Usando el paquete *spacy*, se pueden extraer distintas características, por ejemplo, de la siguiente frase:

As young readers like to know 'how people look', we will take this moment to give them a little sketch of the four sisters, who sat knitting away in the twilight, while the December snow fell quietly.

- Verbos encontrados: 'like', 'know', 'look', 'take', 'give', 'sat', 'knitting', 'fell'
- Raíces de los verbos: 'like', 'know', 'look', 'take', 'give', 'sit', 'knit', 'fall'
- Desinencias aproximadas: 'ke', 'ow', 'ok', 'ke', 've', 'at', 'ng', 'll'
- Adverbio encontrado: 'quietly'
- Raíz del adverbio: 'quiet'
- Desinencia aproximada: 'ly'

Esta librería es de gran utilidad cuando se requiere analizar el texto y traducir a diferentes idiomas.

Tema principal

Con el paquete *sklearn* se pueden obtener los Unigramas, bigramas y trigramas, estos son importantes para obtener información del texto, como el tema principal, algunos sentimientos del personaje, etc. En la tabla II, se muestra un ejemplo de un trigrama, considerando el primer parráfo del primer capítulo de la obra. Se observa que las frases que más aparecen en la obra son las frases C,D y E, pues las podemos encontrar al menos dos veces en el extracto del texto. En la figura 3, se visualiza un unigrama, se pude destacar que en el primer parráfo la palabra que aparece al menos dos veces en este es Çhristmas", lo que se pude sugerir que el primer parráfo habla sobre la navidad.

Sentimiento.

Con ayuda del paquete *textblob*, se puede usar la función para extraer la polaridad, la cual está asociada a un sentimiento positivo (Valores cercanos a 1) y negativos (valores cercanos a -1) y subjetividad del texto, para este caso, se utilizó el mismo parráfo y se extrajeron los siguientes resultados:

- Polarity: -0.433
- Subjectivity: 0.6

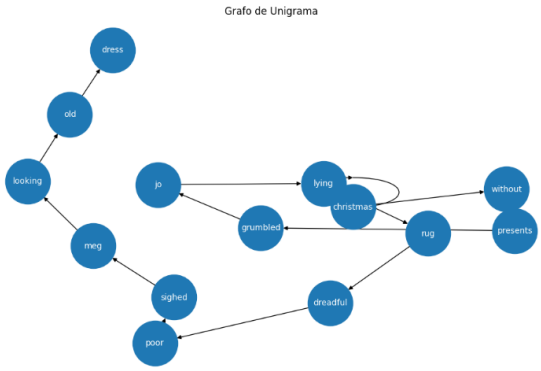

Figura 3. Unigrama, primer párrafo del primer capítulo - Elaboración propia.

II-B. ***Reseña***

La reseña fue extraída de https://amazon-reviews-2023.github.io, en csv. El extracto contiene una serie de columnas que hace referencia a reseñas que son escritas por usuarios en la plataforma de amazon, dentro de sus columnas se encuentran:

- *productId*: De tipo string, asocia el producto con un id único.
- *prodDescription*: De tipo string, describe a que producto hace referencia.
- *titleReview*: De tipo string, es la reseña que da el usuario.
- *rating*: De tipo numérico, tiene una escala del 1 al 5.

Para comprender mejor los datos del conjunto de reseñas, se muestra la figura 4. La cual muestra a los productos que recibieron reseñas más extensas (con mayor número de palabras), el porcentaje hace referencia a la frecuencia relativa considerando todas las reseñas extraídas de la categoría, que son 420,518 reseñas, mientras el número que aparece entre paréntesis es la cantidad de palabras que tiene el producto en sus reseñas.

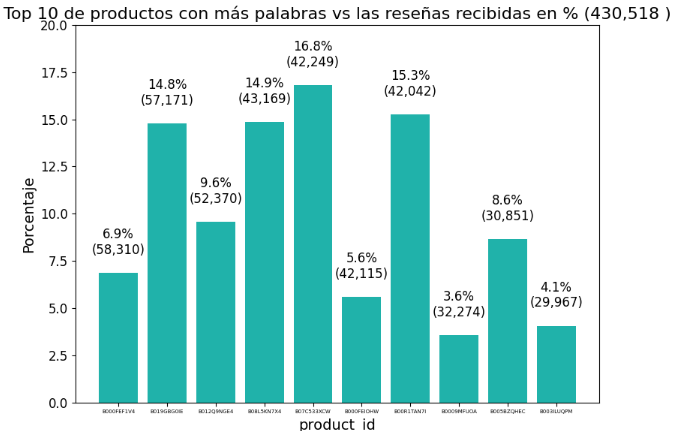

Figura 4. Top 10 de los productoo con más palabras dentro de sus reseñas - Elaboración propia.

Métrica	No. de carácteres	Total de palabras en la reseña
Promedio	196.23487	36.877759
Desviación	255.761862	47.526113
Mínimo	3	1
Máximo	15,034	2,594
1er Cuartil	61	11
Mediana	124	23
3er Cuartil	237	45

Tabla III
TABLA DE ESTADÍSTICA DESCRIPTIVA DE ”LITTLE WOMEN ELABORACIÓN PROPIA.

Palabras En la figura 5, se muestran las palabras más repetidas dentro de las reseñas, se puede resaltar que las que mayormente tienen relevancias son palabras como ”great”, ”hair”, ”product”, lo que se puede asumir que la mayoría de las reseñas son positivas de los productos.

Figura 5. Top 20 de palabras más frecuentes en el texto - Elaboración propia.

Signos de puntuación

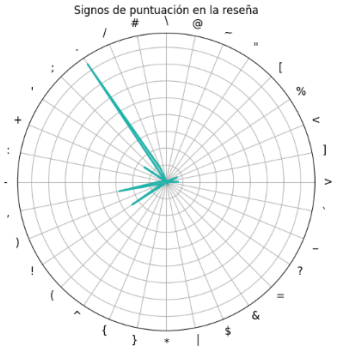

Figura 6. Signos en la reseña - Elaboración propia.

La figura 8, muestra los diferentes signos que se encuentran dentro de las reseñas. Entre estos se encuentran los signos de coma, punto y paréntesis, además a diferencia del texto literario, se exponen más signos de puntuación, pues estos pueden estar asociados a escribir alguna expresión de .emoji”

Emoji Un emoji es un pictograma que hace referencia a una emoción, hoy en día es utilizado dentro de las redes sociales o textos informales para dar más sentimiento al texto que se redacta. Dentro de las reseñas, se encuentran algunas que contienen diferentes emojis, los más repetidos estan asociados con emociones positivas.

No.	A	B	C	D	E
1	2	0	1	1	2
2	0	1	0	0	1
3	0	0	0	0	0

Tabla IV
A: NOT GREAT
B:THANK YOU
C: THE CONSISTENCY
D: THICK WOUND
E: THIS PRODUCT

Gramática

Considerando la siguiente frase: *It was very easy to use and worked very quickly This product worked amazingly well. It was very easy to use and worked very quickly.*

Con ayuda del paquete *spacy*, se pueden extraen las siguientes características:

- Verbos encontrados: 'use', 'worked', 'worked', 'use', 'worked'.
- Raíces de los verbos: 'use', 'work', 'work', 'use', 'work'.
- Desinencias aproximadas: 'se', 'ed', 'ed', 'se', 'ed'.
- Adverbios encontrados: 'very', 'very', 'quickly', 'amazingly', 'well', 'very', 'very', 'quickly'.
- Raíces de los adverbios: 'very', 'very', 'quickly', 'amazingly', 'well', 'very', 'very', 'quickly'.
- Desinencias aproximadas: 'ry', 'ry', 'ly', 'ly', 'll', 'ry', 'ry', 'ly'.

Esta librería es de gran utilidad cuando se requiere analizar el texto y traducir a diferentes idiomas.

Tema principal

Como se mencionó anteriormente, con el paquete *sklearn* se extraen los Unigramas, bigramas y trigramas y en la tabla IV, se muestra un ejemplo de un bigrama, considerando 3 reseñas elegidas aleatoriamente.

Se observa que las frases que más aparecen en las reseñas son las frases A y E, pues las podemos encontrar al menos dos veces en el extracto del texto.

En la figura 7, se visualiza un unigrama, se pude destacar que en el primer parráfo la palabra que aparece al menos dos veces en este es Çhristmas", lo que se pude sugerir que el primer parráfo habla sobre la navidad.

Sentimiento.

Con el paquete *textblob*, se extrajó la polaridad, la cual está asociada a un sentimiento positivo (Valores cercanos a 1) y negativos (valores cercanos a -1) y subjetividad del texto, para este caso, se utilizó la frase propuesta en el apartado de grámatica y se obtuvieron los siguientes resultados:

- Polarity: 0.5186
- SSubjectivity: 0.840

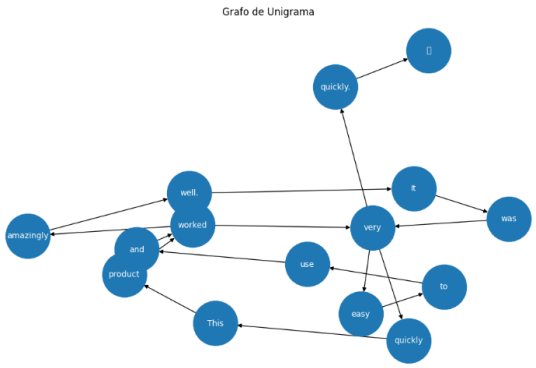

Figura 7. Unigrama de una reseña. - Elaboración propia.

III. PRINCIPALES DIFERENCIAS.

Por medio del análisis que se obtuvo en la sección anterior se pueden destacar distintas diferencias entre un texto y otro.

III-A. Propósito y audiencia

En primer lugar el próposito o el fin de un texto es muy distinto al otro, mientras la novela cuenta una historia que intenta tomar la atención del lector, la reseña tiene el fin de informar y criticar a un producto o una situación. Esto lo deja en claro la frecuencia de las palabras, mientras en la novela se distinguen una serie de nombres propios y verbos en pasado, en las reseñas se distinguen adjetivos que califican o describen al sustantivo.

La audiencia en ambos textos está ligada a los intereses del lector o usuario.

III-B. Gramática

La grámatica juega un papel muy importante en la construcción de la novela y de las reseñas. En la novela se observa en la sintaxis del texto, en los verbos que se utilizan y la puntuación presicia que tiene todo el texto mientras que la reseña tiene más variedad en los signos de puntuación. Un punto a resaltar es en la contracción de las palabras, en la figura 6 se puede observar en que uno de los signos más utilizados es el apóstrofe, pues este signo es comúnmente utilizado en el lenguaje informal y poco utilizado en el lenguaje informal, por lo que, en la obra literaria no aparece dentro del top 6.

El uso de verbos es indistinto y no depende de un texto en especifico, sin embargo, en la novela, la coherencia y la grmática forma un rol de alta relevancia que muy pocas veces se considera al escribir una reseña en un sitio web, por lo que típicamente en estas se encuentran errores de ortografía.

IV. CONCLUSIONES FINALES

Conforme avanza la tecnología y el e-commerce ligado a ésta, es más común encontrar reviews de todo tipo en internet.

Las personas se han ido acostumbrando en aceptar la grámatica de este tipo de textos, encontrar palabras contraídas, slangs o los hasta emojis que se vuelven una forma común al

leer un texto en la red, pues normalmente el enfoque de las reseñas está dirigido al valor de experiencia que aporta.

Si bien las diferencias entre un texto narrativo y un argumentativo son evidentes, cabe señalar en la facilidad que se tiene para detectar estas diferencias por medio de tecnologías y lenguajes de programación, como los usados en este trabajo. A través de este, se puede inducir el tema principal de cualquiera de los dos textos, por medio de la frecuencia de las palabras o también en que tiempo están escritos, esto por medio de los verbos y las librerías que extraen estos mismos. Asímismo se realizan análisis de sentimientos para detectar con que sentimiento está asociado el texto.

Este tipo de análisis de texto es crucial para la detección de patrones, detectar muletillas del redactor o también en un enfoque más avanzado, para traducir de un idioma a otro. Lo importante es darle el enfoque correcto y la interpretación adecuada.

Referencias

- Agustín Vera Luján (2024). "¿Qué es un texto?" Disponible en: https://www.cerasa.es/media/areces/files/book-attachment-3498.pdf.
- Gracida, M. Y. (2011). "Tipos de textos". En Textos Modelo. Portal Académico del CCH, UNAM. Disponible en: https://portalacademico.cch.unam.mx/alumno/tlriid2/unidad1/textosmodelo/tiposDeTextos.