



EDUCACIÓN

SECRETARÍA DE EDUCACIÓN PÚBLICA



TECNOLÓGICO
NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación
y Desarrollo Tecnológico

Tesis de Maestría

Revaloración de la satisfacción de clientes basada en
la polaridad de reseñas

presentada por

Lic. Armando Miranda Molina

como requisito para la obtención del grado de
Maestro en Ciencias de la Computación

Director de tesis

Dr. Noé Alejandro Castro Sánchez

Codirector de tesis

Dr. Máximo López Sánchez

Cuernavaca, Morelos, México. Enero de 2023.

Cuernavaca, Morelos, **23/enero/2023**

OFICIO No. DCC/019/2023
Asunto: Aceptación de documento de tesis
CENIDET-AC-004-M14-OFICIO

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial de ARMANDO MIRANDA MOLINA, con número de control M20CE041, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis de grado titulado "REVALORACIÓN DE LA SATISFACCIÓN DE CLIENTES BASADA EN LA POLARIDAD DE RESEÑAS", y hemos encontrado que se han atendido todas las observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.



NOÉ ALEJANDRO CASTRO SÁNCHEZ
Director de tesis



MÁXIMO LÓPEZ SÁNCHEZ
Codirector de tesis

JAVIER ORTIZ HERNÁNDEZ
Revisor



DANTE MÚJICA VARGAS
Revisor

C.c.p. Depto. Servicios Escolares.
Expediente / Estudiante
JGGS/ibm



Cuernavaca, Mor.,
No. De Oficio:
Asunto:

26/enero/2023
SAC/034/2023
Autorización de
impresión de tesis

ARMANDO MIRANDA MOLINA
CANDIDATO AL GRADO DE MAESTRO EN CIENCIAS
DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **“REVALORACIÓN DE LA SATISFACCIÓN DE CLIENTES BASADA EN LA POLARIDAD DE RESEÑAS”**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

ATENTAMENTE

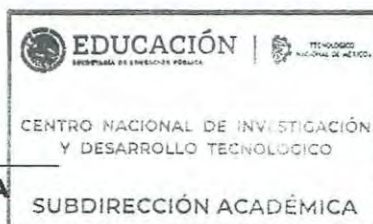
Excelencia en Educación Tecnológica®

“Conocimiento y tecnología al servicio de México”

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

C. c. p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/RMA



Resumen

Plataformas de comercio electrónico como Amazon México ofrecen para cada uno de sus productos opiniones de clientes, calificaciones en estrellas y una valoración general que resume las impresiones que los clientes han tenido de un determinado producto. Por lo tanto, la valoración de los usuarios es uno de los principales factores que influye en la decisión del cliente antes de seleccionar algún producto. Este hecho demuestra que la adquisición de un artículo va más allá de dar un clic y realizar la compra.

Sin embargo, en gran cantidad de artículos puede observarse una discordancia de la opinión expresada en el producto con la calificación dada en estrellas. Debido a que las personas pueden publicar libremente su propia experiencia con el artículo o producto, en la mayoría de los casos no se puede garantizar la calidad de sus opiniones. Por lo tanto, la idea central con este trabajo de investigación es identificar la discrepancia y proponer una revaloración que reajusta la calificación dada por las estrellas para hacerlas coincidir con el grado de satisfacción indicado en los comentarios de cada uno de los clientes de productos de la plataforma web de Amazon México mediante un método automático utilizando la minería de opiniones.

Se llevaron a cabo tres experimentos utilizando los algoritmos de aprendizaje automático Bosques Aleatorios, Regresión Logística, Naive Bayes y Máquina de Soporte Vectorial y se creó un algoritmo heurístico o basado en reglas. Los mejores resultados se obtuvieron con el algoritmo de Máquina de Soporte Vectorial con una precisión del 49%, mientras que para el algoritmo basado en reglas se obtuvo el 40%.

Tabla de contenido

Página

Lista de figuras.....	9
Lista de tablas.....	11
Capítulo 1 Introducción	1
1.1 Planteamiento del problema.....	2
1.2 Justificación.....	3
1.3 Objetivos	4
1.3.1 General.....	4
1.3.2 Particulares.....	4
1.4 Alcances y limitaciones.....	4
1.4.1 Alcances.....	4
1.4.2 Limitaciones.....	4
Capítulo 2 Marco teórico.....	6
2.1 Opiniones de clientes	6
2.2 Función de las reseñas de Amazon	6
2.3 Calificaciones de estrellas	7
2.3.1 Definición	7
2.3.2 Calificaciones en las extremidades de estrellas	7
2.4 Toma de decisiones en el proceso de compra	7
2.5 Características de la reseña de un producto	7
2.6 Análisis de sentimientos.....	8
2.6.1 Nivel de aspecto.....	8
2.6.2 Léxico afectivo.....	8
2.6.3 Polaridad	8
2.6.4 Reseña de productos	9
2.6.5 Subjetividad en opiniones.....	9

2.7	Aprendizaje automático.....	9
2.7.1	Método híbrido.....	9
2.7.2	Método supervisado	9
2.8	Corpus lingüístico	9
2.9	Preprocesamiento	9
2.9.1	Lematización.....	10
2.9.2	Tokenización.....	10
Capítulo 3 Estado del arte.....		11
3.1	Primer grupo (Análisis de reseñas y minería de opinión).....	11
3.1.1	Análisis de opinión de los productos de Amazon utilizando un algoritmo de aprendizaje automático conjunto (Sadhasivam & Kalivaradhan 2019).....	11
3.1.2	Análisis de opinión para revisión de productos (Sultana et al., 2019).....	12
3.1.3	Análisis de opinión sobre reseñas de productos en línea (Bose et al., 2018)	13
3.1.4	Análisis de sentimiento a nivel de aspecto en datos de comercio electrónico (Vanaja et al., 2018)	14
3.1.5	Análisis de sentimiento en tiempo real en la aplicación de comercio electrónico (Jabbar et al., 2019).....	15
3.1.6	Análisis de sentimiento en twitter usando un léxico y multiplicación de polaridad (Kusrini & Mashuri 2019)	17
3.1.7	Elección del método de preprocesamiento de texto más óptimo para el análisis de opiniones: caso tweets de iPhone (Resyanto et al., 2020)	18
3.1.8	Estudio comparativo de enfoques de aprendizaje automático para revisiones de Amazon (Rathora et al., 2018).....	21
3.1.9	Explicando las estrellas: análisis de sentimientos basado en aspectos de las reseñas de clientes en línea (Binder et at., 2019).....	22
3.1.10	Extracción de patrones de cambio de polaridad de cualquier corpus en función de la anotación natural (Yang et al., 2020).....	23

3.1.11 Minería de opiniones a nivel de aspecto para reseñas de hoteles en idioma Myanmar (Hnin et al., 2018)	24
3.1.12 Minería de opiniones/análisis de sentimientos para reseñas de usuarios (Mathew et al., 2020)	25
3.1.13 Relación entre el sentimiento del cliente y las calificaciones de los clientes en línea para hoteles: un análisis empírico (Geetha et al., 2018)	26
3.1.14 Una técnica de categorización de polaridad de sentimientos para reseñas de productos en línea (Kausar et al., 2019)	29
3.1.15 Un estudio de la tarea de análisis de sentimientos y sus desafíos (Pandey et al., 2019)	30
3.2 Tablas comparativas del primer grupo de trabajos.....	31
3.3 Segundo grupo (soluciones de procesamiento de lenguaje natural)	34
3.3.1 Analizadores multilingües en FreeLing (Padró 2010).....	34
3.3.2 Interacción discursiva y fórmulas de tratamiento en las respuestas de los hoteles a las opiniones de viajeros (Sanmartín 2018).....	35
3.3.3 Procesamiento práctico del lenguaje natural (Vajjala et al., 2020).....	37
3.3.4 Receta de procesamiento de lenguaje natural (Kulkarni et al., 2019)	41
3.3.5 Web scraping con Python (Mitchell 2018).....	42
3.4 Tabla de soluciones y métodos de lenguaje natural del segundo grupo de trabajos	44
Capítulo 4 Desarrollo del método de solución.....	46
4.1 Generación de corpus	46
4.2 Etapa de preprocesamiento.....	52
4.3 Algoritmo Basado en Reglas.....	54
4.4 Evaluación de polaridad	62
Capítulo 5 Pruebas y resultados	63
5.1 Análisis de la relación entre calificaciones de estrellas y los comentarios	63
5.2 Corpus de comentarios	72

5.3	Valores de las características textuales	73
5.4	Métricas de evaluación.....	73
5.5	Experimento 1: Algoritmos de aprendizaje automático.....	75
5.5.1	Primera prueba con 1-grama.....	77
5.5.2	Segunda prueba con 2-gramas	80
5.5.3	Comparación de las pruebas	83
5.6	Experimento 2: Algoritmo Basado en Reglas.....	84
5.7	Experimento 3: Comparación del algoritmo propuesto con algoritmos de aprendizaje automático.....	87
5.8	Reasignación de estrellas	90
Capítulo 6 Interfaz web del proyecto de tesis.....		94
6.1	Prueba piloto	99
Conclusiones		106
	Trabajos futuros	108
	Producción científica	108
Referencias.....		110

Lista de figuras

Figura 1.1 Comentario de la compra de una mochila.	1
Figura 1.2 Comentario de la compra de un maquillaje.	1
Figura 1.3 Comentario de la compra de un reloj inteligente.	1
Figura 1.4 Comentario de la compra de una báscula corporal.	1
Figura 1.5 Comentario de la compra de un disco en estado sólido.	2
Figura 3.1 Niveles de análisis de sentimientos.	14
Figura 3.2 Arquitectura propuesta del sistema.	16
Figura 3.3 Algoritmo para la identificación de polaridades negativas.	16
Figura 3.4 Ciclo de desarrollo en el PNL.	18
Figura 3.5 Precisión de cada método de preprocesamiento.	19
Figura 3.6 Ocho combinaciones de preprocesamiento con la máxima precisión.	20
Figura 3.7 Patrones basados en reglas lingüísticas.	25
Figura 3.8 Dendrograma de clúster para categoría de bajo presupuesto.	27
Figura 3.9 Dendrograma de clúster para categoría premium.	28
Figura 3.10 Reducción y extracción de oraciones a nivel de aspecto.	30
Figura 3.11 Reconocimiento de entidad con nombre.	30
Figura 3.12 Correferencia.	31
Figura 3.13 Clases de datos lingüísticos en FreeLing 3.0.	34
Figura 3.14 Clases de preprocesamiento en FreeLing 3.0.	35
Figura 3.15 Respuesta del responsable del hotel de inferior categoría.	36
Figura 3.16 Respuesta del responsable del hotel de inferior categoría.	36
Figura 3.17 Respuesta del responsable del hotel de elevada categoría.	36
Figura 3.18 Respuesta del responsable del hotel de elevada categoría.	37
Figura 3.19 Secciones estructuradas.	37
Figura 3.20 Bolsa de palabras para un corpus de cuatro documentos.	38
Figura 3.21 N-gramas para un corpus de cuatro documentos.	39
Figura 3.22 TF-IDF aplicado a un corpus con dos documentos.	40
Figura 3.23 Función para realizar procesamiento a los tweets.	42
Figura 3.24 Uso del Web Scraping con la librería BeautifulSoup.	43
Figura 3.25 Sitio web de productos.	43

Figura 3.26 Lista de productos extraída.....	43
Figura 4.1 Definición puntual del método de solución.....	46
Figura 4.2 Diagrama de flujo para realizar la generación de comentarios.	47
Figura 4.3 Organización del corpus.....	51
Figura 4.4 Servicios de análisis disponibles para cada lengua.	53
Figura 5.1 Valoración de comentarios.....	63
Figura 5.2 Matriz de confusión.....	74
Figura 6.1 Página principal.....	94
Figura 6.2 Enlace de detalles.....	95
Figura 6.3 Url absoluta.....	95
Figura 6.4 Producto seleccionado de Amazon.....	96
Figura 6.5 Ver todos los comentarios.....	96
Figura 6.6 Opiniones de clientes.....	97
Figura 6.7 Parámetros del algoritmo.....	98
Figura 6.8 Proceso finalizado correctamente.....	98
Figura 6.9 Archivo de salida (primera parte).....	99
Figura 6.10 Archivo de salida (segunda parte).....	99
Figura 6.11 Producto de la página Amazon.....	100
Figura 6.12 Reseñas globales.....	100
Figura 6.13 Ejecución del método de solución.....	101
Figura 6.14 Método de solución.....	102
Figura 6.15 Opiniones de clientes (originales).....	104
Figura 6.16 Opiniones de clientes (reassignadas).....	104

Lista de tablas

Tabla 3.1 Resultados de rendimiento de diferentes modelos de clasificación correspondientes a diferentes partes del habla.....	13
Tabla 3.2 Ocho tipos de cambio de polaridad.....	23
Tabla 3.3 Clasificaciones por estrellas.....	29
Tabla 3.4 Tabla comparativa de trabajos de análisis de reseñas y minería de opinión (primera parte)	32
Tabla 3.5 Tabla comparativa de trabajos de análisis de reseñas y minería de opinión (segunda parte)	33
Tabla 3.6 Soluciones de procesamiento de lenguaje natural	45
Tabla 4.1 Campos que componen un comentario.....	48
Tabla 4.2 Cantidad de comentarios por país.....	49
Tabla 4.3 Departamentos que componen el corpus.....	50
Tabla 4.4 Distribución de comentarios en los diferentes departamentos.....	50
Tabla 5.1 Distribución de comentarios por clase.....	72
Tabla 5.2 Resultados del conjunto de pruebas.....	77
Tabla 5.3 Comparación de evaluaciones originales y reasignadas (prueba 1).	78
Tabla 5.4 Comentarios etiquetados por el algoritmo (prueba 1).	79
Tabla 5.5 Resultados del conjunto de pruebas.....	81
Tabla 5.6 Comparación de evaluaciones originales y reasignadas (prueba 2).	82
Tabla 5.7 Comentarios etiquetados por el algoritmo (prueba 2).	82
Tabla 5.8 Resultados del conjunto de pruebas.....	84
Tabla 5.9 Comparación de evaluaciones originales y reasignadas.....	85
Tabla 5.10 Comentarios etiquetados por el algoritmo.....	86
Tabla 5.11 Nueva distribución por clase de comentarios.....	87
Tabla 5.12 Resultados del experimento 1.....	88
Tabla 5.13 Resultados del experimento 2.....	88
Tabla 5.14 Opiniones carentes de subjetividad.....	89
Tabla 5.15 Opiniones subjetivas.....	90
Tabla 6.1 Cantidad de comentarios por país.....	103
Tabla 6.2 Opiniones objetivas.....	105

Capítulo 1 Introducción

Todos los días en internet se generan grandes volúmenes de información, la gente intercambia opiniones sobre los productos que adquieren en las diferentes plataformas de línea como lo son: Amazon, Ebay, Mercado libre etc. El comercio electrónico se ha convertido en uno de los principales factores para la compra de productos entre clientes y empresas a nivel mundial. Sin embargo, de acuerdo a Castelli et al. (2017, pág. 116) “Un estudio demostró que el 75% de los clientes revisan las reseñas de un producto antes de realizar la compra”. Este hecho demuestra que la adquisición de un artículo va más allá de dar un clic y realizar la compra. Generalmente los clientes necesitan información sobre diferentes aspectos antes de adquirir un producto. Pandey et al (2019, pág. 1), menciona que las reseñas de un artículo están formadas por cinco características que son la entidad de un producto, el aspecto de la entidad, la opinión de la entidad, el título de la opinión y la fecha de la publicación. Cada una de estas características lleva a la siguiente relación explicada:

- 1) Revisión y retroalimentación de productos. Este tipo de comentarios son utilizados o tomados en cuenta por otro tipo de usuarios durante la búsqueda de algún producto.
- 2) Evaluación de estrellas. De acuerdo a la evaluación del producto (experiencia del cliente con el artículo), se asigna una calificación de estrellas.

En algunos casos, existen diferentes tipos de revisiones de productos donde se observa una inconsistencia entre las evaluaciones dadas a través de los comentarios y la calificación asignada con las estrellas. A continuación, de la Figura 1.1 a la Figura 1.5, se presentan ejemplos extraídos de la página web Amazon México que describen estas inconsistencias:



Figura 1.1 Comentario de la compra de una mochila.

La Figura 1.1 representa un comentario de la compra de una mochila con una valoración de cinco estrellas, siendo que normalmente las opiniones de cinco estrellas hablan muy bien del producto, pero en este caso el comentario muestra lo contrario.

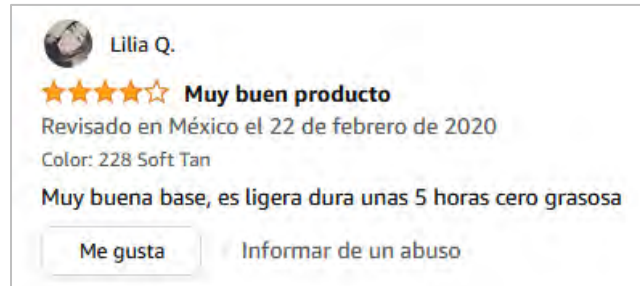


Figura 1.2 Comentario de la compra de un maquillaje.

La Figura 1.2 representa un comentario de la compra de un maquillaje con una valoración de cuatro estrellas. Generalmente, este tipo de comentarios se caracterizan porque hablan bien del producto no obstante, en algún punto de la opinión existe una parte negativa.



Figura 1.3 Comentario de la compra de un reloj inteligente.

La Figura 1.3 representa un comentario de la compra de un reloj inteligente con una valoración de tres estrellas. Regularmente los comentarios de tres estrellas están acompañados tanto de palabras positivas como negativas.



Figura 1.4 Comentario de la compra de una báscula corporal.

La Figura 1.4 representa un comentario de la compra de una báscula corporal con una valoración de dos estrellas. Por lo regular, los comentarios de dos estrellas incorporan una experiencia negativa y solo se destaca una parte positiva del producto.

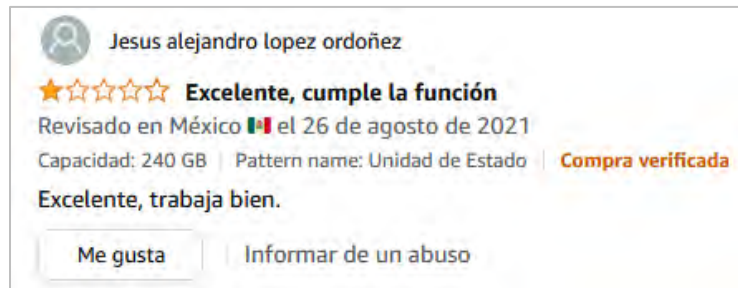


Figura 1.5 Comentario de la compra de un disco en estado sólido.

La Figura 1.5 representa un comentario de la compra de un disco en estado sólido con una valoración de una estrella. Frecuentemente los comentarios de una estrella son totalmente negativos, en otras palabras, la experiencia que se ha tenido con el producto es pésima.

Considerados los argumentos expuestos con anterioridad, esta investigación plantea realizar un análisis de texto de las opiniones sobre los productos de ventas en línea que proporcionan sitios web como Amazon México mediante técnicas de procesamiento de lenguaje natural por sus siglas (PLN), donde un primer análisis parece sugerir trabajar con un algoritmo híbrido que está formando por un enfoque semántico y un enfoque supervisado.

1.1 Planteamiento del problema

Sitios web de comercio electrónico como Amazon México, ofrecen para cada uno de sus productos opiniones de clientes, calificaciones en estrellas y una evaluación general que resumen las impresiones de los consumidores sobre los artículos calificados donde a mayor número de estrellas más positiva será la opinión del usuario. Sin embargo, en gran cantidad de productos puede observarse una discordancia de la opinión expresada en el producto con la calificación en estrellas. Dicha discordancia puede darse de dos maneras: por una parte, un producto puede recibir opiniones muy parecidas de acuerdo a la experiencia que se ha tenido con el artículo pero calificaciones distintas en estrellas y, por otro lado, un producto puede recibir una alta o baja valoración de estrellas pero la opinión muestra lo contrario.

La idea central con este trabajo de investigación es identificar la discrepancia y proponer una revaloración que reajusta la calificación dada por las estrellas para hacerla coincidir con el grado de satisfacción indicado en los comentarios de cada uno de los clientes permitiendo así obtener una polaridad simétrica entre la opinión y la calificación de cada producto.

1.2 Justificación

A continuación, se especifican los siguientes puntos que justifican esta investigación:

- Las ventas en línea desde hace algunos años comenzaron a tomar fuerza, sin embargo, no existe un control de como las personas puedan expresar la experiencia del producto para asignar una correcta reseña. No existe una forma de garantizar que el usuario genere un texto coherente de acuerdo con la calificación acompañada de su comentario.
- Al hablar de la minería de opinión o análisis de textos, se busca encontrar un patrón en el texto mediante técnicas de análisis de sentimientos. En esta investigación, el análisis de ese patrón es precisamente la polaridad de las opiniones de algún producto en particular que ha sido adquirido en algún sitio web en línea. Sin embargo, la mayoría de investigaciones que se han llevado a cabo se realizan en el idioma inglés.
- El análisis de sentimientos se puede aplicar en la mayoría de los negocios electrónicos donde el cliente proporciona comentarios. Desde el punto de vista en sitios de ventas en línea, es importante saber si los clientes están satisfechos con los productos adquiridos. Este tipo de información está disponible en los comentarios críticos de los usuarios. Sin embargo, no es posible analizar cada comentario manualmente para corroborar que tanto la calificación de las estrellas, como las revisiones de las reseñas, tienen coherencia. Por lo tanto, se requiere un sistema de análisis de sentimientos que clasifique automáticamente la opinión de los usuarios como positivas, negativas o neutrales. La idea principal es entrenar un algoritmo que tenga la capacidad de entender la opinión humana extraída de diferentes fuentes de sitios web.
- Los clientes buscan la retroalimentación de productos. La idea es realizar una revaloración que reajuste la calificación de las estrellas de los comentarios que se generen a diario de los sitios web de comercio electrónico. Por lo cual, se creará un método que genere una estructura interna y su función será implementar una capa de abstracción para trabajar con textos.

1.3 Objetivos

En esta sección se describen el objetivo general y los particulares para este tema de tesis.

1.3.1 General

Revalorar la satisfacción que los clientes de la plataforma de comercio electrónico Amazon indican mediante la escala de estrellas analizando la polaridad de sus reseñas a través de la minería de opiniones.

1.3.2 Particulares

- Crear un corpus de comentarios con su evaluación de estrellas de la plataforma web Amazon México para determinar la polaridad de cada comentario.
- Comprender cómo las calificaciones de estrellas se asocian con los comentarios, es decir, que características de las opiniones distinguen un comentario que está relacionado con una valoración.
- Determinar qué porcentaje de comentarios mantienen una relación coherente con su calificación a través de la implementación de un algoritmo de aprendizaje automático.
- Validar que exista consistencia en el ajuste de las revaloraciones de todos los comentarios con el método computacional propuesto.

1.4 Alcances y limitaciones

1.4.1 Alcances

- El modelo descargará las opiniones de los productos, ya sea por un número de paginaciones predefinidas o el total de opiniones.
- El modelo permitirá trabajar con comentarios que estén en otro idioma.

1.4.2 Limitaciones

- Los comentarios con sarcasmo e ironía no serán tratados en el desarrollo de esta investigación. Esta clase de opiniones requiere un estudio profundo, mismo que por los tiempos y alcances de esta investigación no serán abordados.

- Los comentarios no asociados directamente a la opinión y experiencia hacia el producto que se califica no serán tratados. De acuerdo al sitio web de Amazon México en el apartado de *ayuda y atención al cliente*¹, se explica que se debe evaluar únicamente estos aspectos.
- La autenticidad de los comentarios de la plataforma de Amazon México no podrá comprobarse, por lo que de existir reseñas falsas, estas serían procesadas por el método de solución desarrollado.
- Las reseñas actualizadas por los clientes no serán abordadas. Dentro de la página web de Amazon México existe un apartado para *editar las opiniones*². En otras palabras, este tipo de comentarios mantienen dos reseñas de un mismo producto en una sola opinión con una sola valoración.

El capítulo 2 marco teórico muestra los conceptos que sirven de base para esta investigación, el capítulo 3 estado del arte muestra los trabajos que se consideran más importantes relacionados con este tema de tesis, el capítulo 4 desarrollo del método de solución presenta las cuatro fases que componen la definición puntual del método de solución, el capítulo 5 pruebas y resultados muestra los experimentos que sirvieron para determinar si los comentarios que se encuentran clasificados según el número de estrellas que los clientes asignaron a los productos que adquirieron, no guardan congruencia entre la opinión del cliente y la evaluación en la estrella, el capítulo 6 interfaz web del proyecto de tesis muestra la implementación de una interfaz web con el propósito de observar los resultados del proyecto de investigación de una forma más simple.

1

https://www.amazon.com.mx/gp/help/customer/display.html?ref_=hp_left_v4_sib&nodeId=G3UA5WC5S5UUKB5G

2

https://www.amazon.com.mx/gp/help/customer/display.html?ref_=hp_left_v4_sib&nodeId=G5VY882PY3GFTN6Z

Capítulo 2 Marco teórico

A continuación, se presentan las definiciones fundamentales que abordan el contenido de este capítulo.

2.1 Opiniones de clientes

Las opiniones de los clientes son útiles para que otros usuarios obtengan más información acerca del producto y decidan si es el adecuado para ellos. A continuación, se presentan algunos tipos de opiniones no permitidas:

- Una opinión de alguien que tenga un interés financiero directo o indirecto en el producto.
- Una opinión de alguien que, presuntamente, comparta una estrecha relación personal con el propietario, autor o artista del producto.
- Una opinión del fabricante del producto en la que finja ser un comprador imparcial.
- Múltiples opiniones negativas acerca del mismo producto de un único cliente.
- Una opinión a cambio de una recompensa monetaria.
- Una opinión de un juego a cambio de créditos adicionales en el juego.
- Una opinión negativa de un vendedor acerca del producto de un competidor.
- Una opinión positiva acerca del álbum de un compañero a cambio de recibir una opinión positiva por su parte.

Por lo tanto, las opiniones de clientes tienen que proporcionar a los usuarios, comentarios genuinos de otros compradores sobre los productos³.

(https://www.amazon.com.mx/gp/help/customer/display.html?ref_=hp_left_v4_sib&nodeId=G3UA5WC5S5UUKB5G).

2.2 Función de las reseñas de Amazon

Las revisiones de productos en línea que se encuentran en Amazon, se han caracterizado por contener un estilo de escritura personal. El propósito es brindar información a los posibles consumidores, aunque no todas las reseñas cumplen esta función de la misma manera. (Skalicky 2013).

3

2.3 Calificaciones de estrellas

2.3.1 Definición

Para cada revisión, siempre hay una calificación de estrellas que asigna un usuario en función de su experiencia con un producto en particular. Por lo tanto, Amazon también contiene calificaciones de estrellas, cada vez que el cliente comparte opiniones. (Kausar et al 2019).

2.3.2 Calificaciones en las extremidades de estrellas

Una calificación muy baja (una estrella), indica una opinión extremadamente negativa del producto, una calificación muy alta (cinco estrellas), refleja una opinión extremadamente positiva del producto y, una calificación de tres estrellas refleja una opinión moderada (indiferencia), en otras palabras, es una serie de comentarios positivos y negativos que se anulan entre sí (ambivalencia). (Mudambi et al 2016).

2.4 Toma de decisiones en el proceso de compra

Un aspecto importante que puede afectar el proceso de toma de decisiones del consumidor, es la cantidad contenida de información disponible en las reseñas en línea. Las revisiones en línea deben contener suficiente información para que el consumidor conozca las características (los pros y los contras del producto). Sin embargo, también se ha observado que cuando el consumidor está expuesto a demasiada información, se produce una sobrecarga cognitiva. Por lo tanto, mucha información hace que el consumidor se sobrecargue y, por ende, obstaculiza la capacidad de toma de decisiones. (Kaushik et al 2018).

2.5 Características de la reseña de un producto

Las reseñas de un producto se representan con cinco características e_i , a_{ij} , s_{ijkl} , h_k , t_l , donde:

- e_i . Representa la entidad de un producto, es decir, si se trata de un producto o servicio. Por lo general un cliente expresa la entidad en la primera parte de la oración.
- a_{ij} . Representa el aspecto de la entidad. En otras palabras, son todas aquellas características, atributos o propiedades que representa un producto.
- s_{ijkl} . Representa la opinión de la entidad, la cual ya se cataloga como positiva, negativa o neutral.
- h_k . Es el titular de la opinión que representa al cliente de la revisión.

- t_1 . Denota el momento en que la revisión es publicada por el cliente. (Pandey et al 2019).

2.6 Análisis de sentimientos

El análisis de sentimientos, también llamado minería de opinión, es el campo de estudio que analiza las opiniones, sentimientos, evaluaciones, valoraciones, actitudes y emociones de las personas hacia entidades tales como productos, servicios, organizaciones, individuos, problemas, eventos, temas y sus atributos. Representa un gran espacio problemático. También hay muchos nombres y tareas ligeramente diferentes, por ejemplo, análisis de sentimientos, minería de opinión, extracción de opiniones, minería de sentimientos, análisis de subjetividad, análisis de afectos, análisis de emociones, minería de revisión etc. (Bing Liu 2012).

2.6.1 Nivel de aspecto

En el análisis de sentimientos, el nivel de aspecto mira directamente la opinión en lugar de observar documentos u oraciones. La salida de este nivel será la entidad, aspecto de esa entidad, Por ejemplo, *“Samsung j7 tiene la mejor calidad de cámara”*. Ahora en este nivel, la cámara es un aspecto de la entidad *“Samsung j7”*. (Jagdale et al 2019).

2.6.2 Léxico afectivo

El indicador de sentimiento más importante son las palabras de sentimiento. Estas son palabras que se usan comúnmente para expresar sentimientos positivos o negativos. Por ejemplo, bueno, hermoso y sorprendente son palabras de sentimiento positivo, en cambio malo y terrible son palabras de sentimiento negativo. Las palabras de opinión son importantes para el análisis de opinión por razones obvias. Una lista de tales palabras se llama léxico del sentimiento o léxico de opinión. (Kusrini et al 2019).

2.6.3 Polaridad

Permite medir cuál es el resultado del sentimiento (positivo o negativo). En el caso del procesamiento de lenguaje natural (PLN), se introduce un valor neutro para clasificación de oraciones. (Torres et al 2018).

2.6.4 Reseña de productos

Las opiniones se pueden clasificar para definir todo tipo de revisiones positivas, negativas o neutrales y se expresan mediante el uso de forma textual, calificación de estrellas y emoji. (Jabbar et al 2019).

2.6.5 Subjetividad en opiniones

Una oración subjetiva expresa algunos sentimientos, puntos de vista o creencias personales. (Bing Liu 2012).

2.7 Aprendizaje automático

Es una rama de la inteligencia artificial que se utiliza para proporcionar a los sistemas la capacidad de aprender y mejorar automáticamente de la experiencia previa sin ser programado explícitamente. (Saranya et al 2019).

2.7.1 Método híbrido

Incluye la combinación de categorías supervisadas y no supervisadas que se utiliza para realizar minería de opinión. (Saranya et al 2019).

2.7.2 Método supervisado

Estos son enfoques de aprendizaje automático, en los que el clasificador se entrena en función de un conjunto de características recopiladas, utilizando los datos de entrenamiento etiquetados. (Saranya et al 2019).

2.8 Corpus lingüístico

Se puede llamar corpus a cualquier colección que contenga más de un texto (corpus como cuerpo textual). El corpus es un conjunto amplio y estructurado del uso de la lengua. El corpus es una colección de textos que se supone que son representativos de un idioma determinado, los cuales se convierten en repositorios de información. (Pérez et al 2019).

2.9 Preprocesamiento

Transformaciones que se efectúan sobre los datos crudos con el objetivo de obtener algo uniforme y sin elementos que se puedan considerar ruidosos para el mejor manejo del texto. Como ejemplos, la eliminación de puntuaciones, caracteres, números, ligas y ciertas palabras que se consideran sin contenido. (Casasola et al 2019).

2.9.1 Lematización

Consiste en determinar la forma, clase o categoría gramatical de cada palabra de una oración. Es una técnica de reducción, permite detectar variantes morfológicas de un mismo. Por ejemplo: palabras como “computo”, “computadoras”, “computable”, “computación” son variantes del término computar, y reemplazadas por el término raíz o lema. (Cruz 2017).

2.9.2 Tokenización

Es el proceso de separar una secuencia de cadenas en individuos como palabras, palabras clave, frases, símbolos y otros elementos conocidos como tokens. Los tokens pueden ser palabras individuales, frases o incluso oraciones completas. En el proceso de tokenización, algunos caracteres como los signos de puntuación se descartan. Los tokens funcionan como entrada para diferentes procesos, como el análisis y la minería de texto. (Haque et al 2018).

Capítulo 3 Estado del arte

Este capítulo presenta el estado del arte, el cual se divide en dos grupos:

El primer grupo de trabajos está formado por quince artículos, los cuales hablan acerca del análisis de las reseñas junto con sus valoraciones y polaridades de los comentarios donde se evalúan generalmente en una escala de una a cinco estrellas en la plataforma de Amazon México, la minería de opinión la cual se encarga de identificar y extraer opiniones dentro del texto, finalmente los algoritmos de aprendizaje supervisado donde se ponen a prueba y se reportan los que mejores resultados han presentado en la literatura.

El segundo grupo de trabajos se conforma de cinco artículos, que incluye soluciones de procesamiento de lenguaje natural como estructuras de datos, servicios de análisis lingüísticos, análisis de textos usando expresiones regulares, etiquetado de parte del discurso y representación del texto en su forma numérica, útiles para la realización de la minería de opiniones.

3.1 Primer grupo (Análisis de reseñas y minería de opinión)

3.1.1 Análisis de opinión de los productos de Amazon utilizando un algoritmo de aprendizaje automático conjunto (Sadhasivam & Kalivaradhan 2019)

Esta investigación muestra que a pesar de que existen diferentes enfoques en el análisis de sentimientos, como Naive Bayes, Regresión Logística y Máquina de Soporte Vectorial, al clasificar las revisiones son efectivos, pero no en términos de precisión. Cuando la precisión es muy baja, el algoritmo particular no puede decidir la revisión como positiva, negativa o neutra.

Por lo tanto, se propone un enfoque de conjunto (Ensemble Machine Learning Algorithm), también conocido como votación por mayoría o clasificador de votación. Este tiene la capacidad de combinar algoritmos y la salida se proporciona en función de la votación de los predictores (valores de todos los algoritmos). Aquí el enfoque conjunto se utiliza combinando Naive Bayes, Regresión Logística y Máquina de Soporte Vectorial. Así mismo, el método propuesto producirá una mayor precisión cuando en comparación con los algoritmos existentes que producen menos precisión que la propuesta. El método de conjunto generalmente proporciona una mejor precisión, ya que combina los algoritmos que ayudan a funcionar de manera efectiva. Este método proporciona un mejor rendimiento que el que se podría obtener de cualquiera de los algoritmos solos. Cada modelo

predice para cada instancia, y la predicción de salida final es el que recibe más de la mitad de los votos. Si ninguna de las predicciones obtiene más de la mitad de los votos, el método de conjunto no podría hacer una predicción estable para esta instancia.

Los datos deben someterse a un procesamiento previo para eliminar frases, palabras y símbolos no deseados. La siguiente fase continuó con el etiquetado POS (partes del discurso). Aquí, se extraen todas las palabras y frases POS necesarias, como verbos, adverbios y adjetivos. Estas palabras serán consideradas con los puntajes ponderados. Después, los datos de entrenamiento se recopilarán y clasificarán utilizando Naive Bayes, Regresión Logística y Máquina de Soporte Vectorial. Finalmente, se implementa un enfoque Ensemble combinando todos los métodos anteriores. Una vez que se realiza la clasificación se vota la mayoría entre ellos (precisión), y el clasificador por votación produce una mayor precisión que el sistema existente.

En esta investigación se propuso un enfoque de ensamble que permite combinar varios algoritmos, a diferencia de otras revisiones en la literatura se observa que este enfoque permite traer una mayor precisión que los algoritmos por sí solos, lo cual es un excelente punto de partida a tomar en consideración con futuras investigaciones.

3.1.2 Análisis de opinión para revisión de productos (Sultana et al., 2019)

Los trabajos existentes que muestra este artículo como antecedentes manejan enfoques de aprendizaje supervisado y métodos para la clasificación y extracción de polaridades, a saber, extracción de sustantivos, verbos, adjetivos, adverbios y diccionarios léxicos como sentiwordnet, wordnet etc. Los investigadores mencionaron que, una fase es la combinación de diferentes partes del habla (partes de la oración) y por lo tanto, producen diferentes tasas de precisión. Así mismo, se llevó a cabo las pruebas con 50,000 revisiones en el conjunto de datos de Stanford disponible libremente⁴. El objetivo general fue hacer pruebas combinando las diferentes partes del habla, para este caso verbos, adjetivos, adverbios, junto con los clasificadores como: Naive Bayes, Regresión Logística, Soporte Vectorial y Árbol de Decisión, mostrados en la siguiente Tabla 3.1:

⁴ <https://www.lexalytics.com/technology/sentiment-analysis>

Tabla 3.1 Resultados de rendimiento de diferentes modelos de clasificación correspondientes a diferentes partes del habla.

Partes del habla considerados	Clasificadores			
	Naive Bayes	Regresión Logística	Soporte Vectorial	Árbol de Decisión
Adjetivo	83.81764	84.29752	81.92482	82.15142
Verbo	80.96507	81.28499	78.65902	78.04585
Adverbio	79.49880	81.43161	79.89869	78.55238
Adjetivo + Verbo	89.85500	88.60000	88.75000	87.87500
Adjetivo + Adverbio	89.85500	88.470000	88.66000	83.82500
Verbo + Adverbio	89.85500	87.255000	86.95000	86.40500
Adjetivo + Verbo + Adverbio	89.85500	89.575000	89.36000	87.78500

Como se observa en la Tabla 3.1, trabajar con la combinación de adjetivos, verbos y adverbios, queda demostrado que los clasificadores aumentaron su precisión.

3.1.3 Análisis de opinión sobre reseñas de productos en línea (Bose et al., 2018)

En este trabajo se experimentó con 568,454 reseñas de alimentos de 74,258 productos y 256,059 usuarios en Amazon durante un periodo de diez años. Seleccionaron los seis productos y usuarios más populares, se utilizó el léxico emocional NRC que se puede clasificar en ocho emociones básicas (ira, miedo, confianza, anticipación, tristeza, sorpresa, asco y alegría) y dos sentimientos (positivo y negativo). Como paso 1, se llevó a cabo el procesamiento de datos, paso 2, análisis de revisión de productos, a continuación, se detalla cada uno.

Paso 1:

- Eliminar las URL.
- Eliminar los símbolos, signos de puntuación y números.
- Sustituir cualquier carácter no UTF-8 por espacio.
- Reemplazar los emoticonos con su sentimiento.
- Cambiar texto a minúsculas.

- Reemplazar las palabras por su raíz.
- Eliminar los comentarios duplicados.

Paso 2:

- En el segundo paso se clasificó el texto revisado en ocho categorías de emoción. En cada categoría, la palabra con mayor frecuencia identificará al más perspicaz de los clientes. La mayoría de los clientes usaron las palabras clave; “bueno”, “maravilloso”, “sabor”, “orgánico” y “olor”. En esta investigación, se analizó la emoción de los productos en función de la opinión de los clientes de Amazon. La mayoría de las revisiones tienen más polaridad positiva como “alegría”, “confianza” que “miedo”, “enojo”, y “asco” para polaridad negativa.

Este artículo realizó la clasificación de sentimientos o emociones utilizando el léxico de emociones NRC. A diferencia de otras investigaciones, se propuso además de evaluar los resultados de las opiniones de los clientes, trabajar con las 8 emociones y dos de sentimientos (positiva y negativa). Este análisis ayuda a identificar el sentimiento de los clientes del mismo modo en el enfoque centrado en el producto encontramos el producto mejor revisado por número de clientes. También se analizó la emoción del número de clientes hacia el producto.

3.1.4 Análisis de sentimiento a nivel de aspecto en datos de comercio electrónico (Vanaja et al., 2018)

Los investigadores introdujeron muchos algoritmos sobre análisis de sentimientos. A pesar de que se han realizado muchas mejoras, hay áreas que necesitan progresar como el rendimiento, la extracción de datos. El análisis de sentimientos no es más que la clasificación de los comentarios.

La Figura 3.1 muestra los diferentes niveles de análisis de sentimientos.

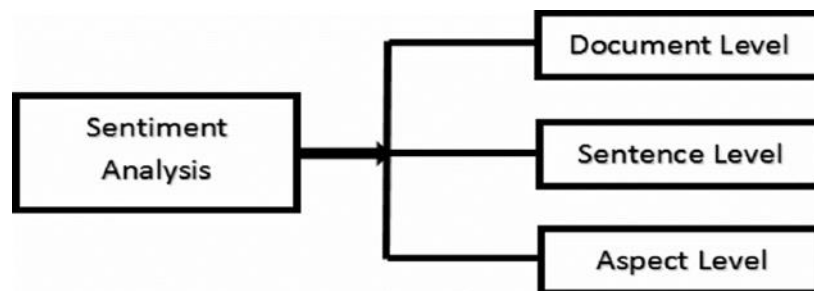


Figura 3.1 Niveles de análisis de sentimientos.

De acuerdo a la Figura 3.1 el análisis de sentimientos básicamente tiene tres etapas:

- 1) Nivel de documento.
- 2) Nivel de oración.
- 3) Nivel de aspecto.

A continuación, solo se detallará la tercera etapa:

Este nivel es el más importante cuando se trata de extraer la polaridad de textos en sitios de internet como Amazon. Esta investigación se enfoca precisamente en esta etapa, ya que permite profundizar y predecir información positiva, negativa y neutral de las opiniones de los clientes. El nivel de aspecto se puede lograr en tres pasos:

- **Identificación:** Se refiere principalmente a identificar los pares objetivo de emoción en las oraciones.
- **Clasificación:** Se refiere a la clasificación de pares de objetivos de emoción identificados.
- **Agregación:** Los valores clasificados se agregan para obtener una visión general.

Al igual que en otras investigaciones se propone hacer una comparación entre el algoritmo Naive Bayes por sus siglas en inglés (NB) y el algoritmo de Máquina de Soporte Vectorial por sus siglas en inglés (SVM). En la literatura los enfoques que se basan en redes neuronales como recurrentes o convolucionales, la mayoría se miden con el algoritmo de clasificación SVM porque este presenta los mejores resultados sin embargo, aquí el algoritmo Naive Bayes resultó ser el que mejor predice la polaridad de las opiniones de los clientes. Considerando lo anterior y las mejoras del análisis de sentimientos, se tiene la idea de que las tres etapas o niveles del análisis de sentimientos es sin duda el nivel basado en aspectos el que mejores resultados dará y los diferentes algoritmos de clasificación llevarán a tener diferentes resultados.

3.1.5 Análisis de sentimiento en tiempo real en la aplicación de comercio electrónico (Jabbar et al., 2019)

Se implementó el análisis de sentimientos en dos niveles de categorización: nivel de revisión y nivel de oración. Al igual que los demás trabajos en la literatura se busca predecir la polaridad del texto junto la evaluación de las estrellas. El trabajo propuso dos fases que se detallan a continuación:

- 1) Primera fase: Se construyó un modelo que realiza el análisis de opinión en tiempo real mediante una interfaz web que se muestra en la Figura 3.2.

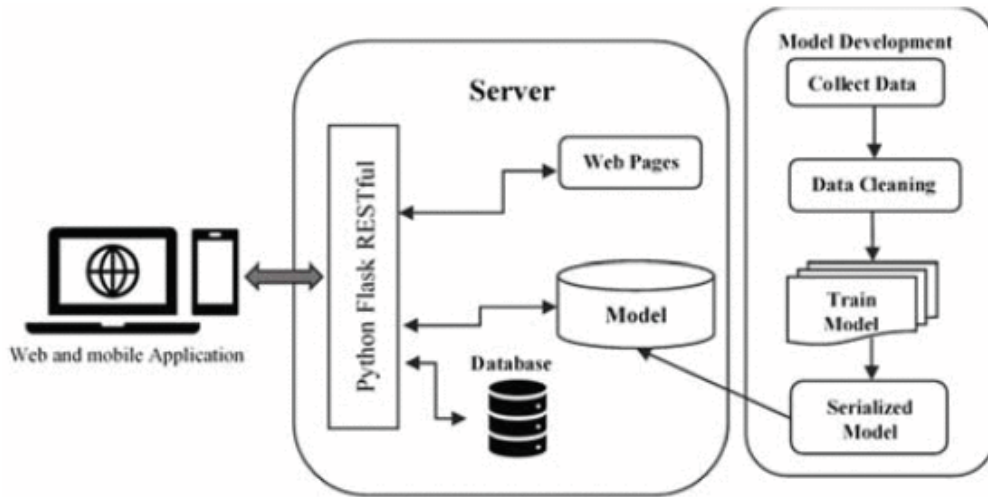


Figura 3.2 Arquitectura propuesta del sistema.

- 1) Segunda fase: Propuso la identificación de frases negativas también llamados prefijos negativos, que están en función de identificar los adjetivos y los verbos utilizados en la opinión de los usuarios. En la Figura 3.3 se muestra el algoritmo de identificación de prefijos negativos para identificar tales frases:

Algorithm 1 Negative polarities Identification

```

Require: Tagged Sentences, Negative Prefixes
Ensure: NOA Phrases, NOV Phrases
for every Tagged Sentence do
    for  $s/s + 1$  as every word/tag pair do
        if  $s + 1$  is a Negative Prefix then
            if there is an adjective tag or a verb tag in next pair then
                NOA Phrases  $\leftarrow (s, s + 2)$ 
                NOV Phrases  $\leftarrow (s, s + 2)$ 
            else
                if there is an adjective tag or a verb tag in the pair after
                    next then
                        NOA Phrases  $\leftarrow (s, s + 2, s + 4)$ 
                        NOV Phrases  $\leftarrow (s, s + 2, s + 4)$ 
                    end if
            end if
        end if
    end for
end for
return NOA Phrases, NOV Phrases
    
```

Figura 3.3 Algoritmo para la identificación de polaridades negativas.

La idea principal de este algoritmo de la Figura 3.3 es clasificar las partes negadas de la oración en función de los verbos y adjetivos. Como ejemplo, cuando un usuario escribe: “*la actitud de servicio al cliente no está mal*”, si no se tomaran en cuenta los prefijos de la negación de verbos (NV) y negación de adjetivos (NA), muy probablemente la clasificación de polaridad no sería la correcta porque la palabra “*mal*” hace referencia a un adjetivo para este ejemplo y su polaridad es negativa sin embargo, el prefijo de la negación cambia la polaridad a nivel de oración.

3.1.6 Análisis de sentimiento en twitter usando un léxico y multiplicación de polaridad (Kusrini & Mashuri 2019)

El enfoque en este artículo va hacia la dirección del análisis semántico para determinar su polaridad. Las reglas semánticas de una oración pueden afectar la polaridad. Por lo tanto, es necesario realizar una investigación en profundidad. El procesamiento del lenguaje natural por sus siglas en inglés (PNL), es una rama del campo de la inteligencia artificial. Este campo busca proporcionar un enlace para la interacción con los humanos a través del lenguaje natural que requiere tecnología computacional y lingüística para construirlo.

En la literatura más actual se comienza a combinar los enfoques léxicos con los enfoques de aprendizaje automático supervisado y no supervisado. Ambos enfoques tienen sus propias características. Los enfoques supervisados por su parte trabajan con cierto tipo de información y se da la respuesta de esta. Por otro lado, los enfoques no supervisados aprenden de cierto tipo de información, pero no se conoce su respuesta. Ambos enfoques a través de la literatura han traído buenos resultados, sin embargo, la base fundamental es una correcta investigación del análisis de sentimientos que se muestra en la Figura 3.4.

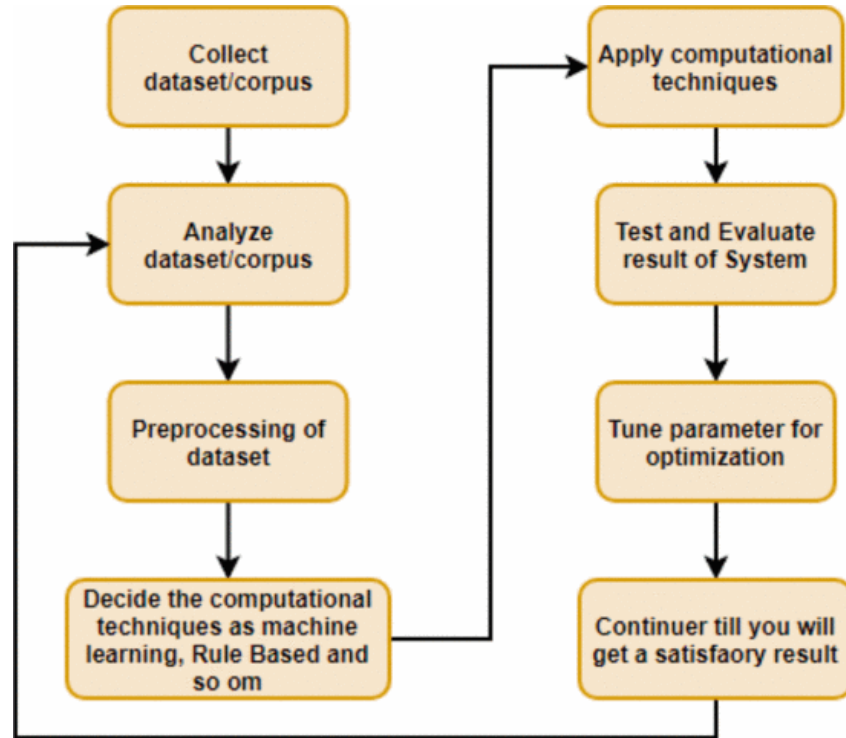


Figura 3.4 Ciclo de desarrollo en el PNL.

Analizar y entender el ciclo de la Figura 3.4 mostrada, da la idea que gran parte de éxito buscado en la polaridad de sentimientos, se basa precisamente en este diagrama. Pues es la parte medular de esta investigación.

3.1.7 Elección del método de preprocesamiento de texto más óptimo para el análisis de opiniones: caso tweets de iPhone (Resyanto et al., 2020)

El preprocesamiento es la etapa inicial en varias tareas de lenguaje natural y de acuerdo a la revisión sistemática se eligen algunas técnicas conforme al tema en desarrollo. Sin embargo, como se menciona en este artículo, rara vez se realizan estudios previos sobre la selección o combinación de métodos más efectivos para resolver parte del análisis de opiniones. Por lo tanto, se realizó un experimento utilizando tweets que hablan sobre el iPhone donde se eligieron nueve técnicas de procesamiento y la combinación de ellas. La Figura 3.5 representa el primer objetivo de esta investigación que consistió en clasificar dos mil quinientos tweets con polaridades positivas y negativas aplicando el algoritmo Naive Bayes, donde se ejecutaron métodos de preprocesamiento del texto individualmente mostrando la precisión de cada uno:

- Palabras a minúsculas: Consiste en reemplazar las letras de una oración en letras minúsculas.
- Alargamiento expresivo: Elimina el exceso de caracteres repetidos de una palabra.
- Manejo de emoticonos: Los símbolos de los emoticonos se convierten en palabras para analizarlos fácilmente.
- Remoción de URL: Elimina el enlace URL de una oración.
- Manejo de jerga: Palabras cortas que tienen el mismo significado.
- Manejo de puntuación: Consiste en separar palabras de la puntuación. Este método de preprocesamiento funciona para idiomas como el inglés que manejan palabras con contracciones.
- Eliminación de palabras vacías: Elimina conjunciones, pronombres de personas y otras palabras que no tienen un significado importante.
- Derivación: Consiste en cambiar la palabra que tiene un afijo en una palabra básica que tiene el mismo significado.
- Tokenización: Consiste en dividir una oración en unigramas, es decir, a nivel de palabra.

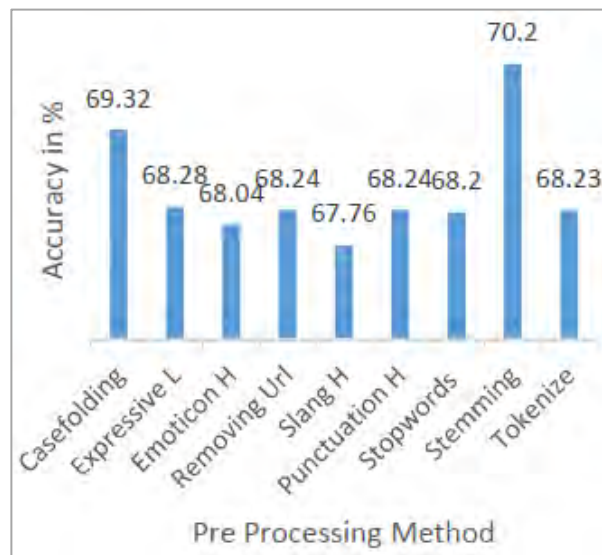


Figura 3.5 Precisión de cada método de preprocesamiento.

La siguiente Figura 3.6 representa el segundo objetivo. Consistió en combinar ocho métodos de preprocesamiento:

- Método A. Remoción de url, manejo de emoticonos y lematización.
- Método B. Manejo de emoticonos y lematización.
- Método C. Remoción de url, manejo de emoticonos, palabras a minúsculas, manejo de puntuación y lematización.
- Método D. Remoción de url, manejo de emoticonos, palabras a minúsculas y lematización.
- Método E. Manejo de emoticonos, palabras a minúsculas, manejo de puntuación y lematización.
- Método F. Manejo de emoticonos, palabras a minúsculas y lematización.
- Método G. Remoción de url, manejo de emoticonos, palabras a minúsculas, expresión alargada, manejo de puntuación y lematización.
- Método H. Remoción de url, manejo de emoticonos, palabras a minúsculas, expresión alargada y lematización.

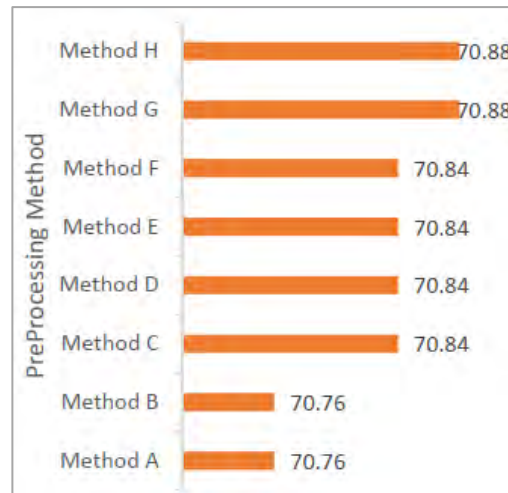


Figura 3.6 Ocho combinaciones de preprocesamiento con la máxima precisión.

Por lo tanto, la Figura 3.6 muestra que, tanto al *Método G* como al *Método H*, son las combinaciones con la máxima precisión y permite observar la importancia del conjunto de métodos en la etapa inicial del preprocesamiento de texto.

3.1.8 Estudio comparativo de enfoques de aprendizaje automático para revisiones de Amazon (Rathora et al., 2018)

Este documento se enfoca en examinar la eficiencia de tres técnicas de aprendizaje automático: (Máquina de Soporte Vectorial (MSV), Naive Bayes (NB) y Máxima Entropía (ME)) para clasificar las revisiones en línea que utilizan métodos de aprendizaje supervisado. Se han extraído revisiones de comentarios usando la API de Amazon. Las revisiones se dividen en positivas, neutras y negativas. Esto no solo es útil para consumidores que desean buscar las reseñas de productos antes de la compra sino también, para empresas que desean observar la reacción del público a sus productos. A continuación, se explica el método propuesto.

- 1) **Características de reducción:** Muchas palabras de relleno o las palabras de detención como "una", "un", "la", "es", utilizadas en una revisión, no especifican ningún sentimiento, por lo que se filtran.
- 2) **Vector de aspecto:** El conjunto de datos consta de 9,500 críticas positivas, 9,500 críticas negativas y 2,500 críticas neutras y se calculan de la siguiente manera:
 - **Unigramas:** Se obtuvieron 21,508 características y cada característica tiene pesos iguales.
 - **Unigramas ponderados:** Se ponderan las palabras clave positivo/bueno y negativo/malo más que los otros aspectos.
- 3) **Entrenamiento y pruebas:** Se eligieron aleatoriamente 3,000 revisiones para el conjunto de datos de prueba.
- 4) **Resultados:** Finalmente se muestran los resultados:
 - **Unigramas:**
 - a) Máquina de Soporte Vectorial: 62.8
 - b) Máxima Entropía: 60.21
 - c) Naive Bayes: 66.84
 - **Unigramas ponderados:**
 - a) Máquina de Soporte Vectorial: 81.20
 - b) Máxima Entropía: 70.35
 - c) Naive Bayes: 77.42

En la literatura se han hecho pruebas con estos clasificadores. En esta ocasión, se probó con una técnica de procesamiento de lenguaje natural que permite trabajar con n-gramas. Como se observa, la exactitud en los clasificadores no fue muy precisa sin embargo, al darle un peso a algunas características en los n-gramas en otras palabras, asignarles una ponderación se obtuvieron mejores resultados.

3.1.9 Explicando las estrellas: análisis de sentimientos basado en aspectos de las reseñas de clientes en línea (Binder et al., 2019)

En la literatura el enfoque del análisis de sentimientos en el aprendizaje supervisado la tarea es predecir la polaridad del texto pero, a su vez carece de un modelo para explicar las clasificaciones de estrellas generales. Por lo tanto, la capacidad de explicar e interpretar las evaluaciones de los clientes, expresadas por las calificaciones de estrellas generales asignadas, es un campo de investigación importante e interesante. Para contribuir a ambas brechas de investigación, se propone un modelo probit ordenado generalizado que utiliza sentimientos basados en aspectos como variables independientes para explicar las calificaciones generales de estrellas de las revisiones de clientes en línea como variables dependientes, a continuación, se detalla el procedimiento en dos pasos:

- 1) Primero, los sentimientos basados en aspectos deben extraerse de las revisiones en el conjunto de datos. Este paso es necesario para aplicar el modelo probit ordenado generalizado propuesto. Para extraer sentimientos del texto existen métodos bien establecidos. Dado que, los enfoques de aprendizaje supervisado requieren el etiquetado manual de una gran cantidad de revisiones, se aplicaron diccionarios de opinión separados. Esto permitió tener en cuenta las diferentes orientaciones de los sentimientos en función del aspecto mencionado.
- 2) Segundo, se establece un modelo subyacente para preferencias continuas en lugar de calificaciones discretas de estrellas. Luego, se utiliza una transformación no lineal de las preferencias subyacentes en la escala de calificación. Más precisamente, se determina el valor promedio del sentimiento general sobre todas las reseñas agrupadas por la clasificación de estrellas asignada.

Básicamente este enfoque explica las calificaciones generales de estrellas de las reseñas en línea, la idea es ayudar tanto a las empresas como a los clientes al evaluar el producto del porque se

asignaron esas calificaciones específicas para precisar mejor, se busca conseguir una simetría con las evaluaciones de los textos y la calificación general de estrellas de un producto cualquiera.

3.1.10 Extracción de patrones de cambio de polaridad de cualquier corpus en función de la anotación natural (Yang et al., 2020)

Hoy en día hay muchos textos de opinión. Determinar si un espacio de texto es positivo o negativo, es una tarea fundamental en el análisis de sentimientos. En algunos casos, la tarea de clasificación de polaridad se complica por el cambio de polaridad que es un fenómeno en el análisis de sentimientos que denota invertir, atenuar o cancelar la polaridad de una pista de sentimiento (palabra o frase) en una oración. El cambio de polaridad más común es usar "no" para negar el sentimiento. Por ejemplo:

- 1) Ella no es feliz ahora.
- 2) No me gusta comer manzanas

Algunos tipos de cambio de polaridad son menos obvios que "no". Por ejemplo:

- 1) Si está triste, beberá.
- 2) Esta tienda solía ser muy buena.
- 3) Este plato parece muy delicioso.

Encontrar patrones de cambio de polaridad en un corpus no es fácil. Actualmente, la construcción manual es la forma principal de obtener patrones de cambio de polaridad lo que lleva mucho tiempo.

A diferencia de otros artículos que profundizan en las entidades basadas en aspectos se presenta un enfoque que puede extraer patrones de cambio de polaridad de cualquier corpus. La idea es detectar los cambios de polaridad para mejorar el análisis de sentimientos en los textos. La Tabla 3.2 muestra ocho tipos de cambio de polaridad.

Tabla 3.2 Ocho tipos de cambio de polaridad.

Tipos de cambio de polaridad	Ejemplos
Negación explícita	no, no, sin
Transición de contraste	pero, desafortunadamente

Negación implícita	evitar, apenas
Falsa impresión	mira, parece
Probabilidad	probablemente
Contrafactual	debería haría
Excepción	lo único
Hasta	hasta

A continuación, se muestran ejemplos de oraciones y se demuestra porque es importante trabajar con las palabras polares y no solo con los enfoques a nivel de aspecto:

- 1) Es estúpido y no es inteligente en absoluto.
- 2) La habitación es ruidosa, necesitamos un lugar tranquilo.
- 3) Es un día triste, pierdo interés por todo.

En la oración (1), claramente, "*estúpido*" es negativo e "*inteligente*" positivo; podemos considerar "*estúpido*" como la anotación de polaridad (negativa) para toda la oración. Entonces, debe haber un mecanismo de cambio de polaridad para permitir que ocurra "*inteligente*" en una circunstancia negativa; de lo contrario la oración sería extraña: "*Es estúpido e inteligente*". En la oración (2), toda la frase se anota como negativa (ruidosa), se usa para cambiar la palabra positiva "*sin molestias*", y "*necesita*" significa que falta (algo bueno). En la oración (3), toda la oración se anota como negativa "*triste*" y "*pérdida*" se usa para cambiar la palabra positiva "*interés*".

3.1.11 Minería de opiniones a nivel de aspecto para reseñas de hoteles en idioma Myanmar (Hnin et al., 2018)

El método propuesto se enfoca en las reseñas de hoteles escritas en idioma Myanmar, e identifica características como: aspectos de servicio y comida por mencionar algunos ejemplos para clasificarlos en comentarios positivos, negativos y neutrales. Por lo tanto, el objetivo se centra en extraer las opiniones a nivel de aspecto utilizando patrones con algunas reglas lingüísticas. En la Figura 3.7 puede observarse la creación de expresiones regulares que hacen uso de reglas gramaticales donde la letra "A", representa el aspecto, las letras "OW", representan la opinión de palabra y la letra "M", representa el modificador. De esta manera, es posible aplicar expresiones

regulares para obtener los diferentes aspectos de las opiniones de productos para este proyecto de tesis.

Pattern#1: $A^+, (M^*, OW)^+$ Pattern#2: $(M^*, OW)^+, A^+$ Pattern#3: $(A, (M^*, OW))^+$ Pattern#4: $((M^*, OW), A)^+$ Pattern#5: $(A, (M^*, OW), A)$ Pattern#6: $(M^*, OW)^+$
--

Figura 3.7 Patrones basados en reglas lingüísticas.

A continuación, se presenta un ejemplo con la siguiente oración:

“A pesar de que el servicio es bueno la comida no es deliciosa”

Para identificar las diferentes características de la frase anterior es necesario realizar la segmentación de múltiples aspectos, donde se presentan los siguientes puntos:

- 1) En este caso los aspectos a identificar son *comida* y *servicio*
- 2) Es necesario encontrar que aspecto modifica la palabra de opinión, por ejemplo:

La sentencia *servicio* es modificada por *bueno*

La sentencia *comida* es modificada por *no es deliciosa*

De acuerdo a la Figura 3.7 para identificar las diferentes características de una opinión o reseña con sus aspectos, es necesario aplicar patrones basados en reglas lingüísticas.

3.1.12 Minería de opiniones/análisis de sentimientos para reseñas de usuarios (Mathew et al., 2020)

En este proyecto, se utiliza el método K-Nearest-Neighbour (Vecinos más Cercanos) (NKK) que es una técnica de aprendizaje supervisado no paramétrico, en el que se intenta clasificar un punto de datos a una categoría determinada con la ayuda del conjunto de entrenamiento. Es un algoritmo simple que almacena todos los casos disponibles y clasifica nuevos casos por mayoría de votos por sus K vecinos. Utiliza funciones de distancia para clasificar el punto de datos a una categoría, pueden ser Euclidiana, Manhattan, Minkowski, levenshtein y Hamming. Si $K = 1$, el caso se asigna a la clase de su vecino más cercano. El algoritmo propuesto se lleva a cabo en dos análisis:

- 1) Después de la primera fase, se asigna un puntaje positivo a cada revisión.
- 2) Esto se pasa para el segundo análisis y se proporciona una opinión de revisión neutral.

La puntuación se modifica si es necesario. Se determina un archivo de salida con su puntaje positivo. Las cadenas se dividen en conjuntos de palabras que se analizan contra un archivo TXT, que contiene palabras diferentes con las clasificaciones. Luego, se calcula un puntaje basado en este análisis y se forma el sentimiento del puntaje de análisis.

La precisión del sistema propuesto es de aproximadamente 82 por ciento. La precisión alcanzada por el sistema propuesto es de hasta 86 por ciento. Cuando los métodos de clasificación se aplican en los mismos conjuntos de datos, se muestra que el método de clasificación K-NN proporciona más precisión (~ 83.65%) en comparación con el método de clasificación Bayesiano ingenuo que proporciona el resultado de precisión (~ 75.77%).

A diferencia de otros algoritmos de aprendizaje automático, el algoritmo K-NN demuestra una mayor precisión que el clasificador Bayesiano ingenuo y da una clara idea, de que no solo los algoritmos más populares deben ser considerados, aquí esta una muestra de esto.

3.1.13 Relación entre el sentimiento del cliente y las calificaciones de los clientes en línea para hoteles: un análisis empírico (Geetha et al., 2018)

El objetivo de estudio es analizar si los sentimientos percibidos en las opiniones de los clientes son consistentes con las calificaciones en hoteles de bajo presupuesto y premium. La investigación inicia con la selección de 20 hoteles de categoría baja y 20 de categoría alta, ubicados en Goa (uno de los destinos más importantes del panorama turístico de la India.). Posteriormente se condujeron a las siguientes preguntas de investigación:

- ¿Los sentimientos de las opiniones de los clientes pueden explicar las calificaciones que asignan a los hoteles?
- ¿La relación entre los sentimientos de las opiniones de los clientes y las calificaciones de los hoteles es coherente en todas las categorías en hoteles de bajo presupuesto y premium?

Dicho lo anterior, se realizó lo siguiente:

- 1) Procesamiento y limpieza de los comentarios en hoteles. Remoción de puntos, números etc.

- 2) Análisis exploratorio. Se observaron que las palabras con los términos más frecuentes en ambas categorías fueron: “*hotel*”, “*bueno*”, “*personal*”, “*servicio*”, etc. Sin embargo, no era posible visualizar las relaciones entre palabras por lo tanto, se genera una matriz de documentos de términos utilizando la distancia euclidiana para realizar un agrupamiento jerárquico donde, los diferentes conglomerados formados por términos basados en qué tan fuertemente están correlacionados en términos de uso, y cómo los conglomerados más pequeños a su vez se combinan para formar conglomerados más grandes, como se observa en la Figura 3.8 y Figura 3.9:

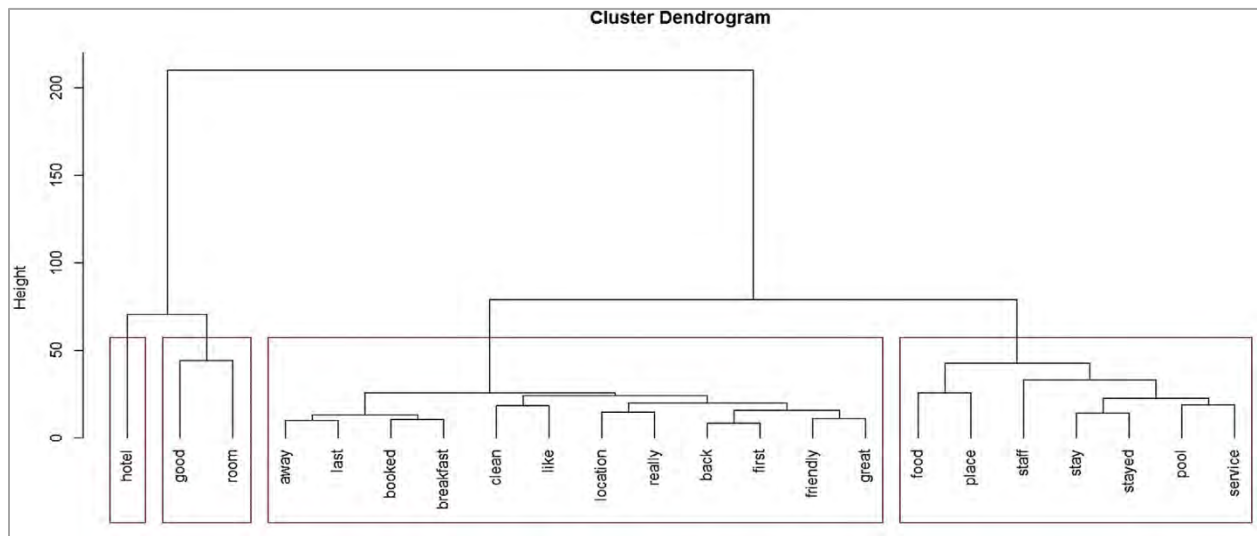


Figura 3.8 Dendrograma de clúster para categoría de bajo presupuesto.

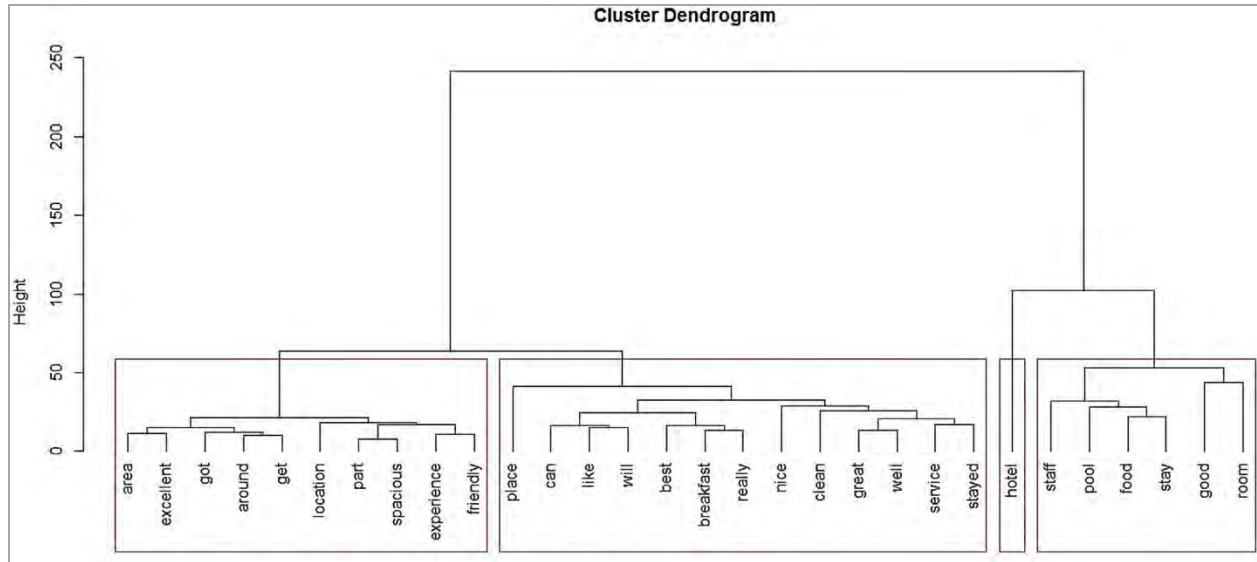


Figura 3.9 Dendrograma de clúster para categoría premium.

3) Análisis de datos predictivos. Para responder las preguntas, se llevó a cabo una regresión lineal simple donde, la polaridad del sentimiento del cliente es la variable independiente y la calificación es la variable dependiente. Este proceso utilizó un léxico de sentimientos basados en palabras, de las cuales se consideran “positivas” cuando son mayores a 1.5, “neutrales” si son mayores o iguales a 1 y menores o iguales a 1.5, “negativas” si son menores que 1. Para determinar la polaridad de cada revisión se utilizó el clasificador Naive Bayes entrenado con el léxico antes mencionado, posteriormente se aplicó la siguiente fórmula: *polaridad del sentimiento del cliente = palabras positivas / palabras negativas de la revisión*. Aplicada esta información, se resumen los resultados del análisis de regresión simple:

- a. Hoteles de baja categoría: El coeficiente de determinación o R cuadrado tuvo un valor de 0.44.
- b. Hoteles de alta categoría: El coeficiente de determinación o R cuadrado tuvo un valor de 0.21.

A partir del análisis de regresión se encuentra que en hoteles de baja categoría el 44% de las opiniones de los clientes explican la calificación asignada, mientras que en hoteles de lujo solo se explica el 21%. La variación en las calificaciones de los hoteles se explica por el sentimiento del cliente en mayor medida en el caso de los hoteles de categoría económica. Por lo tanto, esta

investigación demuestra que este tipo de análisis puede ser aplicado al corpus de comentarios y determinar si existe una relación de las opiniones en Amazon con la calificación en estrellas.

3.1.14 Una técnica de categorización de polaridad de sentimientos para reseñas de productos en línea (Kausar et al., 2019)

Esta investigación considera que para evaluar la clasificación de estrellas la polaridad positiva, negativa y neutral, debe ser manejada por cinco tipos de polaridades: fuertemente negativas, negativas, neutrales, positivas y fuertemente positivas. El primer paso para evaluar las clasificaciones de cinco estrellas es averiguar el rango que va de la más alta a la más baja calificación. Dentro de las investigaciones que se llevaron a cabo en este artículo, se indica que el rango altamente positivo y altamente negativo va desde el -1 hasta el +1. La Tabla 3.3 muestra las clasificaciones de estrellas junto con los valores de polaridad.

Tabla 3.3 Clasificaciones por estrellas.

Clasificaciones de estrellas	Valores de polaridad	Clase
1	-1 al -0.5	Fuertemente negativa
2	-0.5 al 0	Negativa
3	0	Neutral
4	0 al 0.5	Positiva
5	0.5 al 1	Fuertemente positiva

3.1.15 Un estudio de la tarea de análisis de sentimientos y sus desafíos (Pandey et al., 2019)

La idea principal es centrar la información en el nivel de aspecto donde se trabajan con tres dependencias fundamentales en la reducción y extracción de características de la siguiente Figura 3.10.



Figura 3.10 Reducción y extracción de oraciones a nivel de aspecto.

A continuación, se detalla la Figura 3.10:

- En esta tarea se pasa una oración como entrada a un algoritmo, la oración se procesa para identificar todos los sustantivos presentes en ella. Considere el siguiente ejemplo:

“La Canon Power Shot es una gran cámara para principiantes. Es fácil de usar y es de muy buena calidad. Los gráficos son geniales y toma la imagen rápidamente.”

En la revisión, la entidad objetivo es la cámara Cannon y la revisión mencionó el nombre del producto que es "Canon Power Shot". Un ejemplo de la herramienta Stanford CoreNLP NER, para la revisión anterior se muestra en la Figura 3.11.

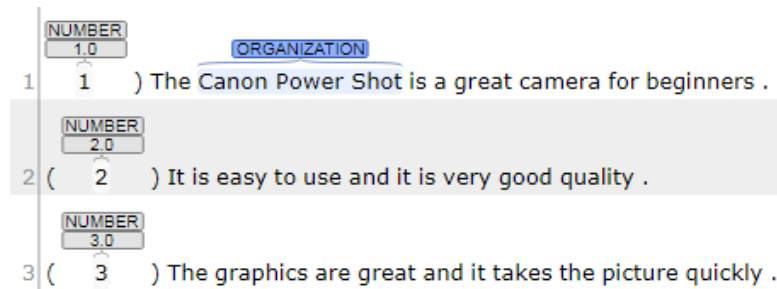


Figura 3.11 Reconocimiento de entidad con nombre.

- Correferencia. Se refiere a mantener la referencia de los aspectos cuando el cliente evalúa a más de un artículo. Los escenarios posibles se observan cuando se adquiere un producto y además este viene con diferentes componentes. Considere el siguiente ejemplo:

"El portátil era muy bueno y va muy rápido."

En esta frase, el pronombre "El" se refiere a la entidad de sustantivo "laptop". La Figura 3.12 muestra la resolución de correferencia de la herramienta Stanford CoreNLP NER.

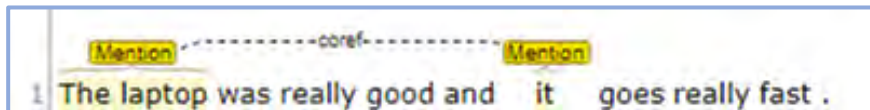


Figura 3.12 Correferencia.

- Clasificación de la subjetividad. La tarea de clasificación de subjetividad se realiza antes de la tarea de clasificación de opiniones. Por lo tanto, la clasificación subjetiva se realiza en primer lugar, de esta manera, puede mejorar la precisión de la clasificación.

3.2 Tablas comparativas del primer grupo de trabajos

Tabla 3.4. Muestra los trabajos relacionados respecto al conjunto de datos que utilizaron en los comentarios (Amazon, Twitter, Stanford, repositorios web). Además, se presenta la escala de polaridad que se empleó desde fuertemente negativo, negativo, neutral, positivo y fuertemente positivo. Asimismo, se compara el nivel de análisis de sentimientos (documento, oración y aspecto), las características de polaridad (sustantivos, verbos, adjetivos y adverbios) y la primera parte del preprocesamiento.

Tabla 3.5. De igual manera, muestra los trabajos relacionados con las técnicas y herramientas del procesamiento del lenguaje natural y tipos de clasificadores que se han usado en el estado del arte como Naive bayes, Árboles de decisiones, Bosques aleatorios, Regresión logística etc.

Es posible observar que en este trabajo de investigación se utilizan las cuatro características de polaridad que son verbos, adjetivos, adverbios y sustantivos, donde en la literatura se combinan estas pero no las cuatro, además se propone llevar a cabo la revaloración de estrellas ya que en la mayoría de la literatura solo se manejan las polaridades positivas, negativas y neutrales.

Tabla 3.4 Tabla comparativa de trabajos de análisis de reseñas y minería de opinión (primera parte)

Trabajos	Conjunto de datos				Clases de polaridad	Nivel de análisis de sentimientos	Características de polaridad	Preprocesamiento																		
	API		Scraping	Corpus				Eliminación																		
	Repositorio web	Amazon						Twitter	Universidad de Stanford	Fuertemente negativo	Negativo	Neutral	Positivo	Fuertemente positivo	Revaloración de estrellas	Documento	Oración	Aspecto	Verbos	Adjetivos	Adverbios	Sustantivos	Palabras de detección	Espacios en blanco	Etiquetas de nueva línea	Etiquetas HTML
Sadhasivam et al., 2019		✓												✓	✓	✓			✓						✓	
Sultana et al., 2019					✓	✓		✓						✓											✓	✓
Bose et al., 2018	✓					✓		✓											✓			✓			✓	✓
Vanaja et al., 2018	✓					✓	✓	✓					✓		✓			✓	✓							
Jabbar et al., 2019	✓					✓		✓				✓	✓						✓							
Kusrini et al., 2019			✓		✓	✓	✓	✓						✓					✓			✓		✓		
Resyanto et al., 2020			✓			✓		✓											✓			✓	✓	✓		
Rathora et al., 2018		✓				✓	✓	✓											✓							
Binder et al., 2019	✓					✓	✓	✓	✓	✓			✓													
Yang et al., 2020	✓					✓		✓					✓	✓				✓	✓							✓
Hnin et al., 2018				✓		✓	✓	✓					✓													
Mathew et al., 2020	✓					✓	✓	✓					✓						✓							
Geetha et al., 2018	✓					✓	✓	✓											✓	✓					✓	✓
Kausar et al., 2019	✓					✓	✓	✓	✓	✓			✓						✓	✓	✓	✓	✓	✓	✓	
Pandey et al., 2019	✓					✓	✓	✓	✓				✓						✓						✓	
Propuesta de tesis			✓			✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Tabla 3.5 Tabla comparativa de trabajos de análisis de reseñas y minería de opinión (segunda parte)

Trabajos	Procesamiento de datos no estructurados										Clasificadores supervisados									
	Técnicas y herramientas de procesamiento de lenguaje natural																			
	Revisión ortográfica	Etiquetado de emoticonos	Etiquetador POS	Extracción de entidad nombrada	Dependencia de dominio	Clasificación de subjetividad	Resolución de referencia conjunta	SentiWordNet	Léxico afectivo	Léxico de emociones NCR	Negación de verbos	Negación de adjetivos	Naive Bayes	Árbol de decisión	Bosques aleatorios	Máquina de soporte vectorial	Aumento de gradiente	Máxima entropía	K vecinos más cercanos	Regresión logística
Sadhasivam et al., 2019	✓	✓	✓									✓								
Sultana et al., 2019												✓	✓		✓					✓
Bose et al., 2018		✓					✓		✓											
Vanaja et al., 2018			✓				✓					✓			✓					
Jabbar et al., 2019			✓							✓	✓									
Kusrini et al., 2019								✓				✓			✓					
Resyanto et al., 2020		✓										✓								
Rathora et al., 2018												✓			✓		✓			
Binder et at., 2019								✓				✓								
Yang et al., 2020			✓					✓				✓		✓	✓				✓	✓
Hnin et al., 2018			✓					✓		✓	✓								✓	
Mathew et al., 2020																		✓		
Geetha et al., 2018							✓	✓				✓								
Kausar et al., 2019			✓									✓	✓	✓	✓	✓				
Pandey et al., 2019				✓	✓	✓	✓													
Propuesta de tesis	✓	✓	✓			✓		✓		✓	✓	✓		✓	✓					✓

3.3 Segundo grupo (soluciones de procesamiento de lenguaje natural)

3.3.1 Analizadores multilingües en FreeLing (Padró 2010)

Python es uno de los lenguajes de programación más utilizados en el campo de la inteligencia artificial, donde en el área de procesamiento de lenguaje natural, existen librerías para trabajar como: NLTK, TextBlob, Gensim, Spacy etc., se enfocan principalmente al idioma inglés y en algunos casos incorporan soporte para otras lenguas sin embargo, no todos los métodos que se incluyen en cada una de las librerías tienen soporte al cien por ciento para trabajar con otros idiomas, de forma más precisa con el lenguaje español. Por tal motivo, para el desarrollo de esta investigación de tesis se utiliza la librería FreeLing, además de ser código abierto cuyo objetivo es brindar herramientas para el procesamiento del lenguaje natural la interfaz está diseñada para trabajar en el idioma español, lo que permite generar avances más rápidos en proyectos de investigación. FreeLing está compuesto por dos módulos, el primero son las clases de datos lingüísticos que permite contener palabras, etiquetas morfológicas, frases, árboles sintácticos, párrafos etc. A continuación, en la Figura 3.13 se muestra un diagrama UML con la relación de cada una.

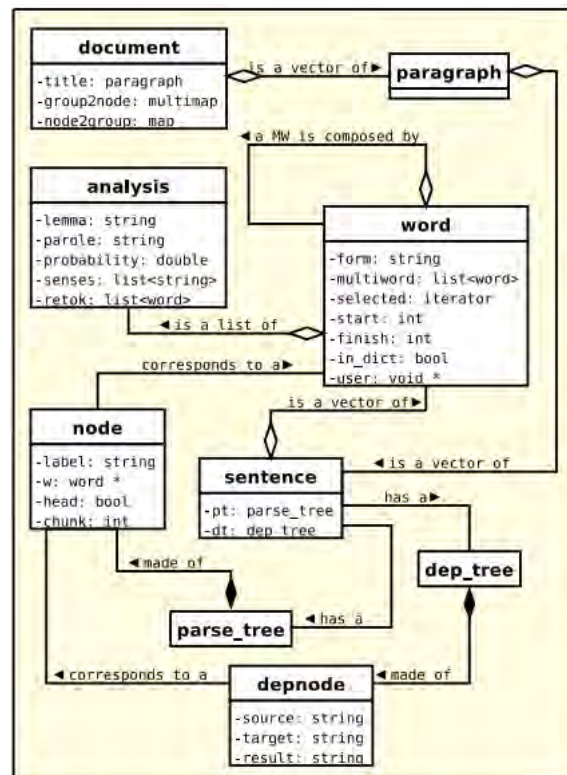


Figura 3.13 Clases de datos lingüísticos en FreeLing 3.0.

El segundo módulo son las clases de preprocesamiento, una vez que han sido almacenados los datos lingüísticos, estos se transforman y se les añade información adicional. La Figura 3.14 muestra un diagrama con la relación de las clases:

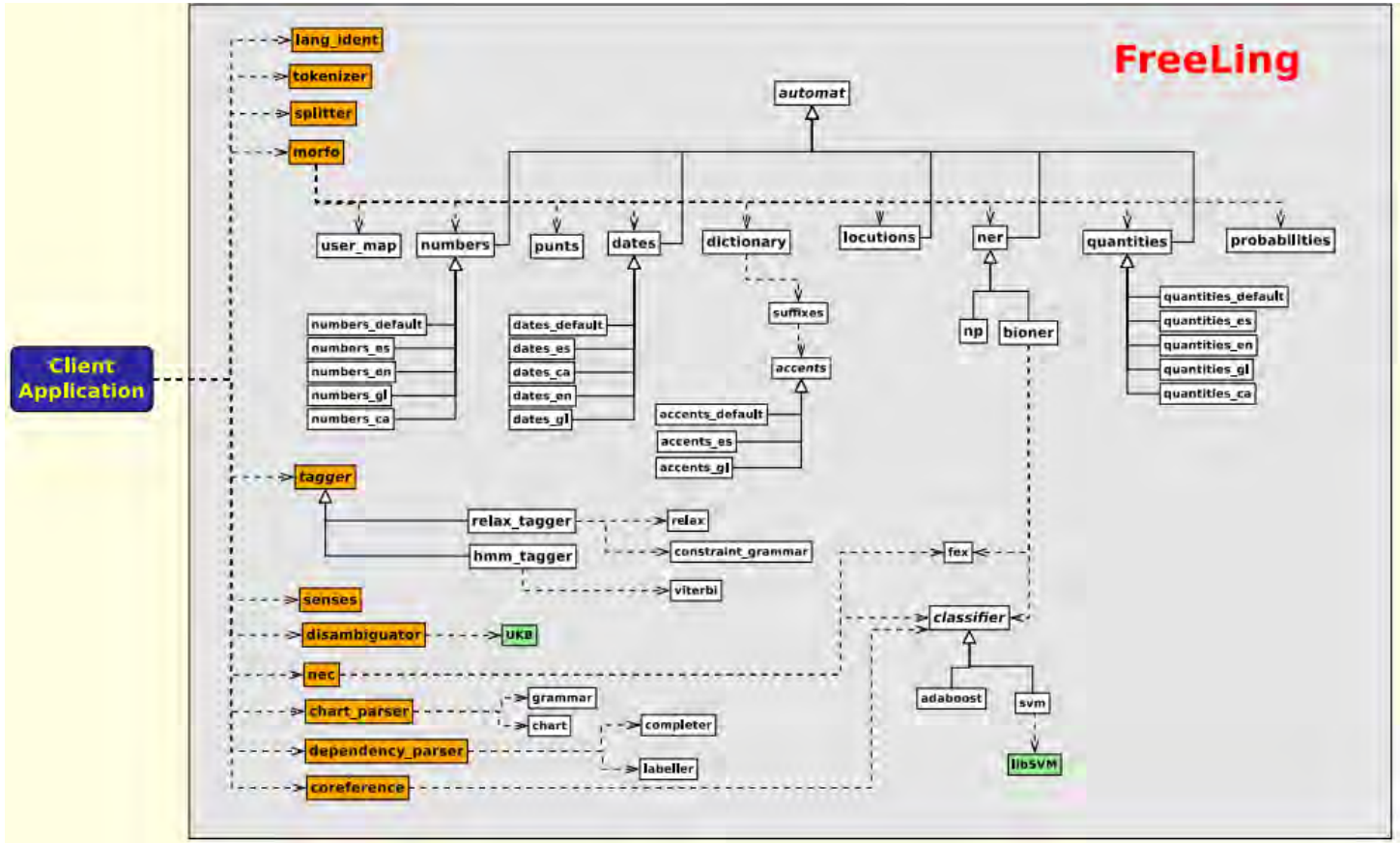


Figura 3.14 Clases de preprocesamiento en FreeLing 3.0.

Como se puede observar en la Figura 3.13 y Figura 3.14, FreeLing es una librería que permite desarrollar potentes aplicaciones y en el método general de solución propuesto más adelante, se menciona que clases de preprocesamiento se implementan.

3.3.2 Interacción discursiva y fórmulas de tratamiento en las respuestas de los hoteles a las opiniones de viajeros (Sanmartín 2018)

Este artículo se centra en los viajeros que han sido clientes o consumidores de las prestaciones del alojamiento en un hotel, donde emiten opiniones positivas o negativas y como los responsables de hoteles responden con fórmulas para el tratamiento en las respuestas. Para comprobar qué tipo de fórmulas aparecen se ha diseñado un corpus constituido por respuesta de 40 establecimientos

hoteleros: 20 establecimientos de España (ubicados en la Comunidad Valenciana) y 20 de Argentina (ubicados en Buenos Aires), clasificados en dos tipos de establecimientos: 10 de elevada categoría (5 y 4 estrellas) y 10 de inferior categoría de la página web “*www.tripadvisor.es*”. Se parte de la hipótesis de que en los de elevada categoría existirá cierta distancia con el cliente frente a los hoteles de menor categoría.

Para el tratamiento de las fórmulas en la respuesta emitida por los propios establecimientos de alojamiento turístico el marketing digital recomienda contestar a todas las opiniones, sean negativas o positivas, con rapidez para evitar la influencia en otros clientes y asegurar una imagen de responsabilidad, para recuperar la confianza y personalizar cada una de las respuestas. A continuación, de la Figura 3.15 a la Figura 3.18, se muestran como ejemplo las respuestas de los responsables de hoteles a los viajeros que dejan opiniones:

X Hoteles, *Propietario* en X Valencia, respondió a esta opinión

Lamento leer este comentario. Y además lo que más *lamento* es que tienes toda la razón. *Tuvimos* un serio problema con el personal de limpieza que *nos* ocasionó tener fallos puntuales que ya *hemos* solventado. De todo corazón. LO SENTIMOS!!!! (Respuesta. ESP. 1 estrella)

Figura 3.15 Respuesta del responsable del hotel de inferior categoría.

Estimada , lamentamos que *lo haya pasado* tan mal , nos esforzaremos mas para la proxima vez que *puedas* visitarnos , Sin que se mal entienda , quiero *explicarte* que los hostels alquilan solo las camas y a *tu* lado puede estar durmiendo un desconocido , en un hotel *vos alquilas* la habitacion , que puede ser con baño privado o sin baño privado , la proxima vez *te* convendria estar en un cuarto con baño privado , cuesta un poquito mas y es mucha la diferencia . Saludos Jose (Respuesta. ARG. 1 estrella)

Figura 3.16 Respuesta del responsable del hotel de inferior categoría.

Estimado Juan, Muchas gracias por sus comentarios. *Nos alegra* saber que ha disfrutado de las comodidades y de la ubicación del hotel pero *nos* gustaría tener mas detalle en referencia al trato de la Recepción. *Estoy enviando* un mensaje privado para que pueda contactarme. Lo *esperamos* en otra oportunidad. Cordialmente, C. V. Gerente de Recepción (Respuesta. ARG. 5 estrellas)

Figura 3.17 Respuesta del responsable del hotel de elevada categoría.

Estimado cliente, Le *agradecemos* su tiempo en escribir una opinión en Tripadvisor. *Agradecemos* su fidelidad a H hotel y ser miembro del programa de puntos X. *Lamentamos* que la decoración art deco del hotel no sea de su agrado. Para *nosotros* la seguridad de nuestros clientes es esencial y por tanto antes de entregar copias de llaves de habitaciones o activar zonas de acceso a clientes es imprescindible identificarse con *nuestro* personal de recepción, *lamentamos* que esto haya sido un inconveniente durante su estancia. No dude en *contactarme* para su próxima visita y me *ocupare* personalmente de sus peticiones para que su estancia este acorde a sus expectativas.

Saludos Cordiales, C. V. Guest Service Manager (Respuesta. ESP. 5 estrellas)

Figura 3.18 Respuesta del responsable del hotel de elevada categoría.

Como se puede observar en la Figura 3.15 y Figura 3.16, los establecimientos de inferior categoría prefieren combinar la tercera persona (usted) con la segunda persona (tú o vos). Por el contrario, en las Figura 3.17 y Figura 3.18, los hoteles de categoría superior optan por usar la tercera persona de formalidad (usted). Este hecho demuestra que los hoteles de elevada e inferior categoría, distinguen características lingüísticas para diferenciar la distancia o proximidad entre la respuesta y la opinión de clientes. Por lo tanto, para el desarrollo de esta tesis deberán explicarse las calificaciones de estrellas asociadas con las características de los comentarios.

3.3.3 Procesamiento práctico del lenguaje natural (Vajjala et al., 2020)

Describe gran parte de métodos para trabajar con procesamiento de lenguaje natural como puede observarse en la Figura 3.19:

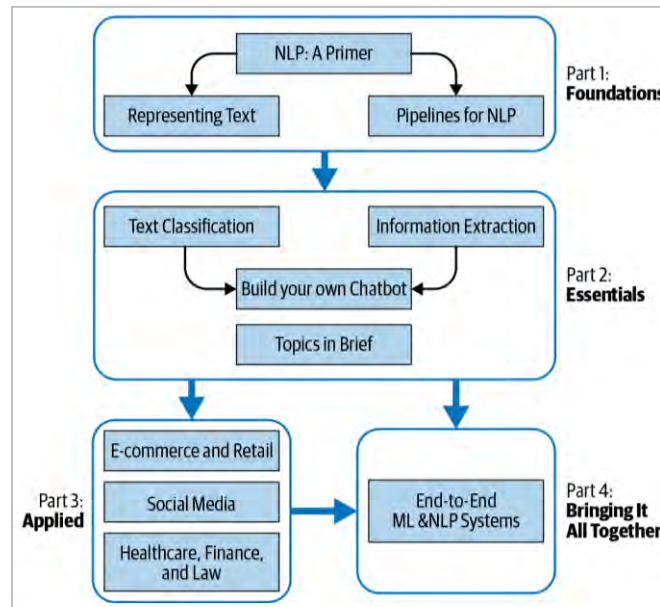


Figura 3.19 Secciones estructuradas.

Sin embargo, para esta investigación se busca conocer que técnicas se emplean para llevar a cabo la representación del texto en su forma numérica de las cuales se observaron los siguientes métodos:

- Bolsa de palabras. Es una técnica clásica para la representación del texto que se utiliza comúnmente en el procesamiento de lenguaje natural. La idea es representar el texto como una bolsa (colección) de palabras ignorando el orden y el contexto. Dicho de otra manera, el texto que pertenece a una clase dada en el conjunto de datos se caracteriza por un conjunto único de palabras. Si dos piezas de texto tienen casi las mismas palabras, entonces pertenecen a la misma bolsa (clase). Así, al analizar las palabras presentes en un texto se puede identificar la clase (bolsa) a la que pertenece. En este método a cada palabra w del vocabulario del corpus, se le asigna un ID, que representa un entero único w_{id} entre 1 y $|V|$, para palabras diferentes, donde V es el conjunto del vocabulario del corpus. Cada documento del corpus se convierte luego en un vector de $|V|$ dimensiones donde en el i -ésimo componente del vector, $i = w_{id}$, es simplemente el número de veces que la palabra w aparece en el documento. A continuación, se ilustra un ejemplo de la Figura 3.20, donde se utiliza una librería de Python llamada “*CountVector*” y se muestran los resultados sobre un corpus de cuatro documentos:

```

1. Corpus: ['bolsa de palabras', 'palabras de bolsa', 'bolsa para oraciones', 'palabras para texto']
2. Vocabulario del corpus: {'bolsa': 0, 'de': 1, 'palabras': 3, 'para': 4, 'oraciones': 2, 'texto': 5}
3. Representación para el documento 'Bolsa de palabras': [[1 1 0 1 0 0]]
4. Representación para el documento 'Palabras de bolsa': [[1 1 0 1 0 0]]
5. Representación para el documento 'Existen muchas frases': [[0 0 0 0 0 0]]

```

Figura 3.20 Bolsa de palabras para un corpus de cuatro documentos.

La Figura 3.20 representa el resultado de aplicar la bolsa de palabras a un corpus de cuatro documentos. Como se puede observar en el punto uno: se muestra el texto de cada oración, el punto dos: muestra el entero único que se mapea con el vocabulario del corpus, el punto tres y cuatro como ejemplo, se ingresan dos oraciones para determinar cómo se representan las dimensiones en 0 (ceros) y 1 (unos). Finalmente, en el punto cinco: se ingresa una oración con palabras que no forman parte del vocabulario, por lo tanto, las dimensiones del vector se llenan con 0 (ceros) porque no son parte de la bolsa de palabras. Nota: cabe mencionar que la librería de Python la asignación de ID comienza a partir de cero.

- N-gramas. Como se observó en la representación del esquema anterior las palabras son tratadas como unidades independientes. No hay noción de frases u ordenamiento. El enfoque de n-gramas intenta remediar esto. Lo hace dividiendo el texto en palabras contiguas (tokens). Cada documento del corpus se representa mediante un vector de longitud $|V|$. Este vector contiene los recuentos de frecuencia de los n-gramas presentes en el documento y cero para los n-gramas que no estén presentes. A continuación, se muestra un ejemplo con los resultados de la Figura 3.21, donde se aplica el mismo corpus formado por cuatro documentos utilizando la librería de igual forma “*CountVector*”, la cual permite trabajar con n-gramas:

```

1. Vocabulario del corpus:
{'bolsa de': 0, 'de palabras': 3, 'palabras de': 4, 'de bolsa': 2, 'bolsa para': 1, 'para oraciones': 6, 'palabras para': 5, 'para texto': 7}
2. Representación para el documento 'Bolsa de palabras': [[1 0 0 1 0 0 0]]
3. Representación para el documento 'Palabras de bolsa': [[0 0 1 0 1 0 0]]
4. Representación para el documento 'Existen muchas frases': [[0 0 0 0 0 0 0]]
    
```

Figura 3.21 N-gramas para un corpus de cuatro documentos.

La Figura 3.21 representa el resultado de aplicar n-gramas a un corpus de cuatro documentos. Como se mencionó anteriormente, los n-gramas dividen el texto en palabras contiguas, dicho de otra forma, los n-gramas pueden ser representados como: unigramas, bigramas, trigramas etc., para este ejemplo se utilizaron bigramas (2-gramas). En el punto uno: se puede observar el vocabulario donde las palabras se muestran de manera contigua de dos en dos, en los puntos dos y tres, se ingresan dos oraciones para determinar cómo se representan las dimensiones en 0 (ceros) y 1 (unos). Finalmente, en el punto cuatro: se ingresa una oración con palabras que no forman parte del vocabulario.

- TF-IDF. En los enfoques anteriormente mencionados todas las palabras del texto se tratan como igualmente importantes, no existe la noción de que algunas palabras son más importantes que otras. TF-IDF (termino de frecuencia-frecuencia inversa de documentos), tiene como objetivo cuantificar la importancia de una palabra dada en relación con otras palabras en el documento y el corpus. Este enfoque se utiliza para la recuperación de información, para extraer documentos relevantes de un corpus.

El método TF-IDF se plantea de la siguiente manera: si una palabra w aparece muchas veces en un documento d_i , pero no aparece mucho en el resto de los documentos d_j en el

corpus, entonces la palabra w debe ser de gran importancia para el documento d_i . La importancia de w debería aumentar en proporción a su frecuencia en d_i , pero al mismo tiempo, su importancia debería disminuir en proporción a la frecuencia de la palabra en otros documentos d_j del corpus. Matemáticamente, esto se captura usando dos cantidades TF e IDF:

$$TF(t, d) = \frac{\text{(Número de ocurrencias del término } t \text{ en un documento } d)}{\text{(Total del número de términos en el documento } d)} \quad (3.1)$$

$$IDF(t) = \log_e \frac{\text{(Número total de documentos en el corpus)}}{\text{(Número de documentos con el término } t)} \quad (3.2)$$

La Ecuación 3.1 mide la frecuencia con la que aparece un término o palabra en un documento determinado. Dado que, los diferentes documentos del corpus pueden tener diferentes longitudes, un término puede aparecer con mayor frecuencia en un documento más largo que en un documento más corto. Se divide el número de apariciones por la longitud del documento.

La Ecuación 3.2 mide la importancia del término en un corpus. Al calcular TF, todos los términos reciben la misma importancia (ponderación). Finalmente, la puntuación TF-IDF es un producto de estos dos términos. Por lo tanto, la puntuación TF-IDF = TF * IDF. A continuación, se presenta un ejemplo en la Figura 3.22 donde se aplica el termino de frecuencia-inversa de documentos, con una librería de Python llamada “*TfidfVectorizer*” a un corpus con dos oraciones:

```

1. Corpus:
'Esta película es buena', 'Esta película es mala'
-----
2. Vocabulario del corpus:
['buena', 'es', 'esta', 'mala', 'película']
-----
3. Valores TF:
[[1 1 1 0 1]
 [0 1 1 1 1]]
-----
4. Valores TF-IDF:
[[0.63009934 0.44832087 0.44832087 0.         0.44832087]
 [0.         0.44832087 0.44832087 0.63009934 0.44832087]]
    
```

Figura 3.22 TF-IDF aplicado a un corpus con dos documentos.

La Figura 3.22 representa el resultado de aplicar TF-IDF a un corpus con dos documentos. Sin embargo, si se aplican las Ecuaciones 3.1 y 3.2 se obtendrán diferentes resultados, esto es debido a que existen variaciones en la fórmula de TF-IDF. La librería “*TfidfVectorizer*”, aplica las Ecuaciones 3.1 y 3.2 de la siguiente manera:

$$tf(w, d) = f_d: \text{Número de ocurrencias del término } w \text{ en un documento } d \quad (3.3)$$

$$idf(w, D) = \left(\ln \frac{1 + D}{1 + df(d, w)} \right) + 1 \quad (3.4)$$

La Ecuación 3.3 solo considera el número de ocurrencias del término w en un documento d . Mientras que en la Ecuación 3.4 de acuerdo a la documentación de la librería Python se suma el número uno para evitar una división por cero. Después se aplica el producto de TF * IDF, para finalmente llevar a cabo el proceso de normalización con la siguiente Ecuación 3.5:

$$v_{norm} = \frac{v}{\sqrt{v_1^2 + v_2^2 + \dots + v_n^2}} \quad (3.5)$$

3.3.4 Receta de procesamiento de lenguaje natural (Kulkarni et al., 2019)

Plantea diferentes problemas junto con la solución donde ayudan a comprender e implementar de manera más eficiente las técnicas básicas y avanzadas de procesamiento del lenguaje natural. Además, se abordan enfoques como la heurística aplicada al texto en crudo en otras palabras, expresiones regulares que ayudan a obtener características como se muestra en la siguiente Figura 3.23 donde se plantea una solución para extraer contenido no deseado así como el reemplazo de algunos caracteres:

```

2 def processRow(row):
3     import re
4
5     #Removes unicode strings like "\u002c" and "x96"
6     tweet = re.sub(r'(\u[0-9A-Fa-f]+)', r'', tweet)
7     tweet = re.sub(r'^\x00-\x7f', r'', tweet)
8     #convert any url to URL
9     tweet = re.sub(r'((www\.[^\s]+)|(https?://[^\s]+))', 'URL', tweet)
10    #Convert any @Username to "AT_USER"
11    tweet = re.sub(r'@[^\s]+', 'AT_USER', tweet)
12    #Remove additional white spaces
13    tweet = re.sub(r'[\s]+', ' ', tweet)
14    tweet = re.sub('[\n]+', ' ', tweet)
15    #Remove not alphanumeric symbols white spaces
16    tweet = re.sub(r'^\w', ' ', tweet)
17    #Removes hastag in front of a word
18    tweet = re.sub(r'#([^\s]+)', r'\1', tweet)
19    #Replace #word with word
20    tweet = re.sub(r'#([^\s]+)', r'\1', tweet)
21    #Remove :( or :)
22    tweet = tweet.replace(':', '')
23    tweet = tweet.replace(':', '(')
24    #remove numbers
25    tweet = ".join([i for i in tweet if not i.isdigit()])"
26    #remove multiple exclamation
27    tweet = re.sub(r"(!)\1+", ' ', tweet)
28    #remove multiple question marks
29    tweet = re.sub(r"(?)\1+", ' ', tweet)

```

Figura 3.23 Función para realizar procesamiento a los tweets.

La Figura 3.23 muestra una función para llevar a cabo parte del preprocesamiento a los tweets, donde se hacen uso de las expresiones regulares, de las cuales se observan: remoción de caracteres extraños, espacios en blanco, símbolos no alfanuméricos con espacios en blanco, números, etc.

3.3.5 Web scraping con Python (Mitchell 2018)

Describe diferentes técnicas y librerías escritas en lenguaje de programación Python como: Request, Spider y Selenium para extraer componentes de una página web. Uno de sus objetivos es la extracción de texto. A continuación, el fragmento de código que se muestra en la Figura 3.24 hace una petición a la página web⁵ que se ilustra en la Figura 3.25 donde se imprime una lista de productos la cual puede observarse en la Figura 3.26:

⁵ <http://www.pythonscraping.com/pages/page3.html>

```

1 from urllib.request import urlopen
2 from bs4 import BeautifulSoup
3 html = urlopen('http://www.pythonscraping.com/pages/page3.html')
4 bs = BeautifulSoup(html, 'html.parser')
5 for child in bs.find('table',{'id':'giftList'}).children:
6     print(child)
7

```

Figura 3.24 Uso del Web Scraping con la librería BeautifulSoup.

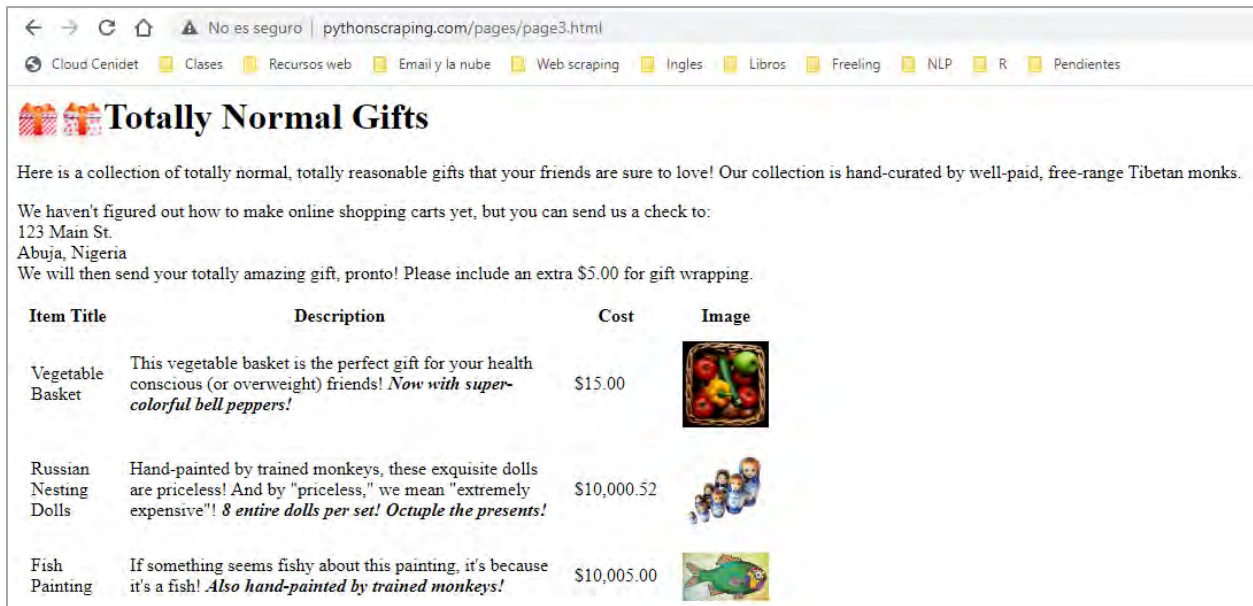


Figura 3.25 Sitio web de productos.

```

C:\Windows\system32\cmd.exe
<tr><th>
Item Title
</th><th>
Description
</th><th>
Cost
</th><th>
Image
</th></tr>

<tr class="gift" id="gift1"><td>
Vegetable Basket
</td><td>
This vegetable basket is the perfect gift for your health conscious (or overweight) friends!
<span class="excitingNote">Now with super-colorful bell peppers!</span>
</td><td>
$15.00
</td><td>

</td></tr>

<tr class="gift" id="gift2"><td>
Russian Nesting Dolls
</td><td>
Hand-painted by trained monkeys, these exquisite dolls are priceless! And by "priceless," we mean "extremely expensive"!
<span class="excitingNote">8 entire dolls per set! Octuple the presents!</span>
</td><td>
$10,000.52

```

Figura 3.26 Lista de productos extraída.

Los datos extraídos de la Figura 3.26, muestran que no tienen un formato. Cuando se realiza Web Scraping, básicamente la información obtenida viene con datos no deseados y como se explicó anteriormente, es necesario llevar a cabo el uso de expresiones regulares para procesar los elementos que se deseen eliminar o mantener.

3.4 Tabla de soluciones y métodos de lenguaje natural del segundo grupo de trabajos

Tabla 3.6. Presenta soluciones y métodos que se emplean en el procesamiento de lenguaje natural y que son útiles para la minería de opiniones. Los métodos que se emplean para el procesamiento de lenguaje natural, así como las distintas soluciones, se obtuvieron de revisión sistemática y principalmente de libros enfocados a este tema.

Tabla 3.6 Soluciones de procesamiento de lenguaje natural

Trabajos	FreeLing			Opiniones positivas y negativas	Representación del texto en su forma numérica			Soluciones de procesamiento de lenguaje natural					Extracción de texto		
	Estructuras de datos y servicios de análisis lingüístico	Clases de almacenamiento de datos lingüísticos	Clases de procesamiento	Características lingüísticas entre la respuesta y la opinión de clientes	Bolsa de palabras	N-gramas	TF-IDF	Extraer texto de la web	Análisis de texto usando expresiones regulares	Exploración y procesamiento de datos de texto	Conversión de texto en características	Etiquetado de parte del discurso	Request	Spider	Selenium
Padró 2010	✓	✓	✓												
Sanmartín 2018				✓											
Vajjala et al., 2020					✓	✓	✓								
Kulkarni et al., 2019								✓	✓	✓	✓	✓			
Mitchell 2018													✓	✓	✓
Propuesta de tesis	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓			✓

Capítulo 4 Desarrollo del método de solución

El desarrollo del método de solución se divide en cuatro grandes fases que se explican en cada uno de los siguientes puntos de la Figura 4.1:

- 1) Creación de corpus.
- 2) Etapa de preprocesamiento.
- 3) Algoritmo Basado en Reglas.
- 4) Reasignación de estrellas.

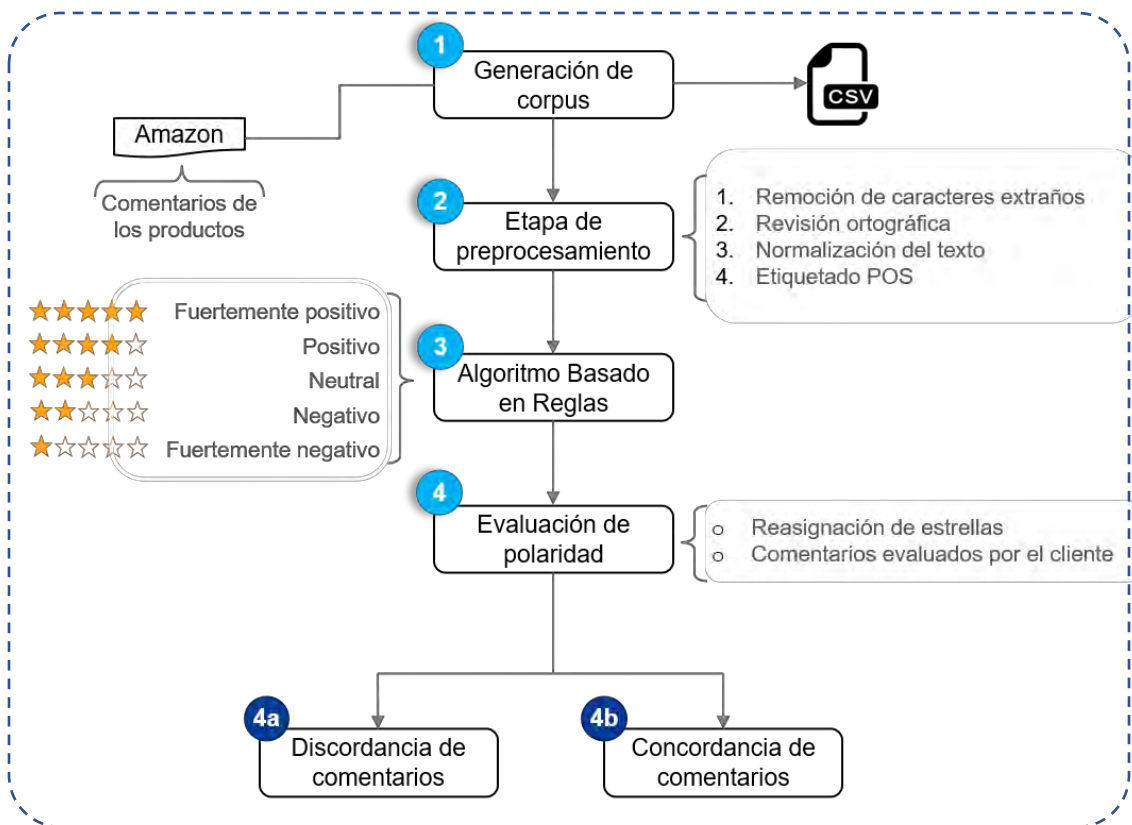


Figura 4.1 Definición puntual del método de solución.

4.1 Generación de corpus

Se decidió crear el corpus de comentarios de la plataforma de Amazon México, ya que es una de las principales empresas de ventas en línea y está presente alrededor del mundo en países como: México, España, Estados Unidos, Italia, Francia, Canadá, China, Japón, etc. Cuenta con una variedad muy amplia de artículos que van acompañados de reseñas y una valoración numérica que los clientes asignaron a los productos que adquirieron. Por tal motivo, en muchas investigaciones en el estado del arte, suelen encontrarse grandes cantidades de corpus de comentarios en esta

plataforma. Sin embargo, no existe un corpus lingüístico de propósito específico para el idioma español basado en comentarios donde se capturen todas las opiniones de un determinado producto o artículo, incluyendo las reseñas que estén en otros idiomas y posteriormente traducidas.

Por otra parte, según Haque T. et al. (2018, pág. 5) “Los sitios de comercio electrónico tienen sus limitaciones para proporcionar datos públicamente”. De acuerdo a Ali Salmony et al. (2021, pág. 136) “La mayoría de los sitios como Facebook, Twitter y Amazon ofrecen una API que facilita el acceso a información, como publicaciones o comentarios disponibles para los investigadores con algunas restricciones. Esta API requiere tener una cuenta o convertirse en asociado de estos sitios web”. Por tal motivo, para llevar a cabo la generación del corpus se utilizó una herramienta conocida como “Web Scraping”, de acuerdo a Mitchell R. (2018, pág. ix) “Su función es recopilar y procesar grandes cantidades de datos rápidamente de sitios web”. Además, con esta técnica es posible seleccionar y extraer cualquier elemento visible dentro de una página web sin restricción alguna.

A continuación, se muestra un diagrama de flujo para observar el proceso de la generación del corpus: Figura 4.2.

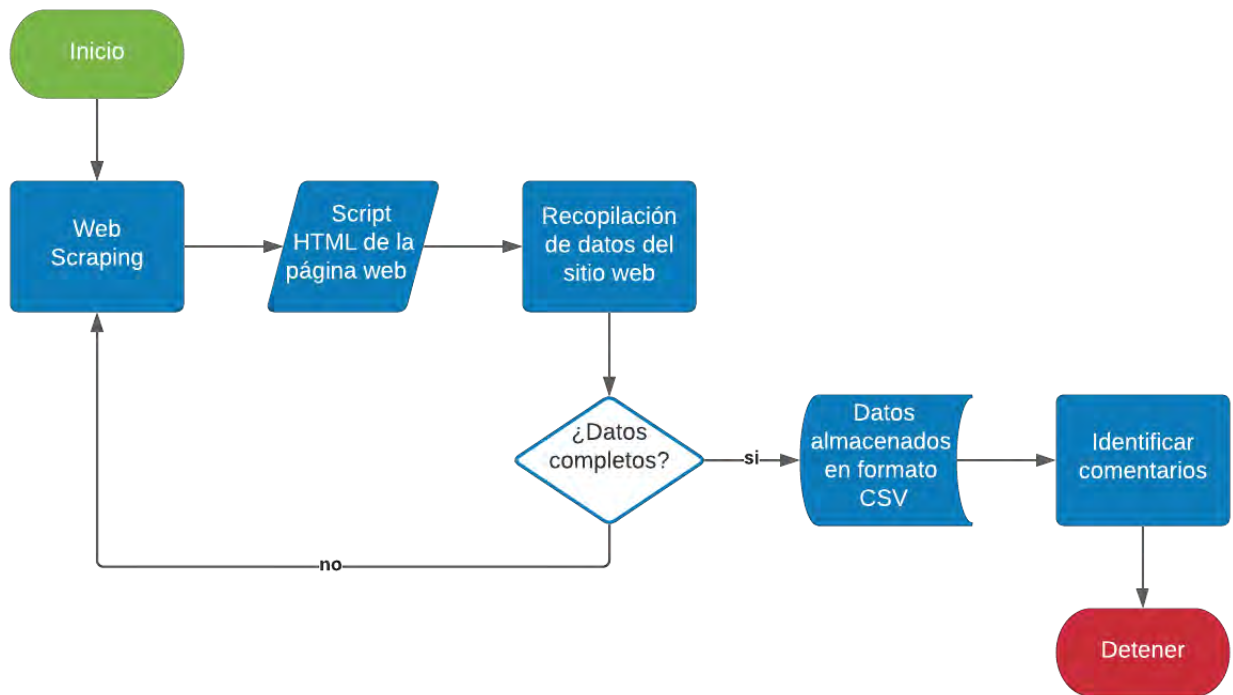


Figura 4.2 Diagrama de flujo para realizar la generación de comentarios.

El proceso de la Figura 4.2 cuenta con seis pasos para la generación del corpus los cuales serán explicados a continuación:

- 1) El primer paso hace uso de la librería llamada Selenium⁶, la cual se encarga de hacer la petición al servidor de Amazon México.
- 2) El segundo paso se encarga de mostrar los datos de la página web, en este caso son los archivos HTML los cuales contienen la información que devuelve el servidor.
- 3) El tercer paso es el proceso de extracción de información. En otras palabras, se lleva a cabo la descarga de los comentarios sin embargo, a pesar de estar en el dominio de Amazon México, existen revisiones escritas en otros idiomas y que se traducen para también utilizarlas.
- 4) El cuarto paso determina lo siguiente: si la información del producto se ha completado continua al siguiente paso, de lo contrario regresa al paso número dos.
- 5) El quinto paso se encarga de construir el corpus para un determinado producto en formato CSV.
- 6) El sexto paso aplica un algoritmo de detección de idiomas para garantizar que las reseñas se asocien con el lenguaje (en este caso el idioma español). Se utilizó la librería “fastText” que se implementa en Joulin, A. et al. (2016).

El corpus de comentarios cuenta con 1,005,364 reseñas y está formado por los siguientes campos que se muestran en la Tabla 4.1:

Tabla 4.1 Campos que componen un comentario.

<i>usuario</i>	Nombre de la persona que emite la reseña
<i>país</i>	Lugar donde se genera la reseña
<i>idioma</i>	Idioma en el que originalmente está escrito el comentario
<i>fecha</i>	Fecha en que se genera el comentario
<i>título</i>	Título de la opinión
<i>estrellas</i>	Calificación que dan al comentario

⁶ <https://www.selenium.dev/>

<i>comentario</i>	Reseña de algún producto en particular
<i>beneficiados</i>	A cuantas personas resultó útil el comentario
<i>artículo</i>	Nombre del producto
<i>departamento</i>	Nombre del departamento que pertenece el artículo
<i>lenguaje</i>	Garantiza que las reseñas se asocien con el idioma

A continuación, se describe la información acerca de su organización que comprende la Tabla 4.2, Tabla 4.3 y Tabla 4.4.

Tabla 4.2 Cantidad de comentarios por país.

País	Cantidad de comentarios
Alemania	28,477
Australia	3,709
Brasil	9,284
Canadá	58,965
España	27,369
Estados Unidos	516,306
Francia	15,206
India	18,287
Italia	36,822
Japón	7,251
México	225,204
Países Bajos	183
Polonia	9
Reino Unido	58,212

Suecia	80
--------	----

Tabla 4.3 Departamentos que componen el corpus.

Departamento	Cantidad de comentarios
Accesorios de Viaje	135,632
Belleza y Cuidado Personal	242,446
Electrónicos	479,794
Muebles de Hogar	147,492

Tabla 4.4 Distribución de comentarios en los diferentes departamentos.

Estrellas	Cantidad de comentarios	Porcentaje
5	660,913	66
4	127,106	12
3	67,838	7
2	47,642	5
1	101,865	10

Con la descripción de las tablas anteriores se muestra un ejemplo en la siguiente Figura 4.3 de la organización de las carpetas del corpus:

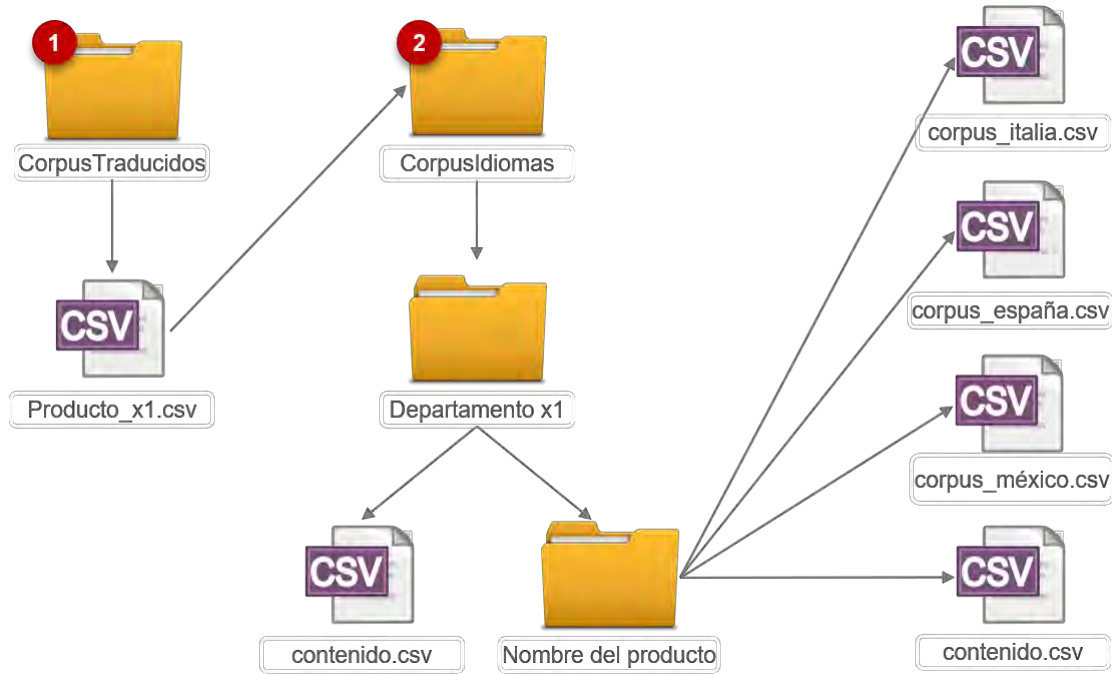


Figura 4.3 Organización del corpus.

Como se puede observar en la Figura 4.3, existen dos carpetas principales de trabajo (*CorpusTraducidos* y *CorpusIdiomas*), las cuales al momento de ejecutar el código, es posible cambiar el nombre, pero para este ejemplo se mantendrán los que aparecen en la Figura 4.3. La carpeta 1 se encarga de almacenar los productos que se requieran descargar, cada uno con su nombre y extensión en formato “CSV” (ejemplo *Producto_x1.csv*). La carpeta 2 genera un documento con el nombre del departamento que se le pase al código (ejemplo *Departamento x1*). Dentro de esta carpeta hay un archivo llamado *contenido.csv*, el cual registra los diferentes productos que pertenecen al mismo departamento, de esta forma se previene volver a extraer el mismo artículo. Posteriormente, se crea una carpeta con el nombre del producto extraído de Amazon (ejemplo *Nombre del producto*). Dentro de esta carpeta se hace uso del archivo “*Producto_x1.csv*” para generar otro documento con el nombre “*contenido.csv*”, cuya finalidad es registrar los diferentes archivos “CSV” (ejemplo *corpus_méxico.csv*, *corpus_italia.csv*, etc.) los cuales representan las reseñas de todos los comentarios por producto pero separados por países, en este caso se generan los archivos con el nombre del país.

Una vez demostrado el funcionamiento de este punto (primera etapa) del método de solución, se entiende que como información de entrada, recibe los comentarios de la página de Amazon México donde estos datos serán enviados a la “*Etapa de preprocesamiento*”.

4.2 Etapa de preprocesamiento

De acuerdo a Gelbukh et al. (2010, pág. 26), el preprocesamiento “se ocupa de los aspectos técnicos, algorítmicos y matemáticos”. Como información de entrada recibe los datos extraídos de la página web Amazon México, donde se realiza lo siguiente:

- Eliminación de caracteres especiales.
 - a) Saltos de línea.
 - b) Más de dos espacios en blanco.
 - c) Símbolos como \grave{a} , $?$, $¡$, $\#$, \grave{c} , $¡$, \ll , \gg etc.

- Eliminación de palabras vacías. Se lleva a cabo el etiquetado POS por sus siglas en inglés *part of speech* (partes del discurso), para identificar la categoría gramatical de las palabras. Por lo tanto, los términos que no formen parte del conjunto de los verbos, adverbios, adjetivos y sustantivos, se consideran palabras vacías. Ejemplo: “*El producto no llegó, tuve que pedir reembolso*” se convierte “*Producto no llegó, pedir reembolso*”.

- Revisión ortográfica.
 - a) Language-tool⁷. Herramienta que corrige la ortografía y la gramática e incluye una API por sus siglas en inglés *Application Programming Interface* (interfaz de programación de aplicaciones) para ser usada en el desarrollo con Python. Sin embargo, es posible incluir un algún recurso léxico, donde se entiende tanto los diccionarios de diversos tipos: monolingües, bilingües, explicativos, diccionarios de sinónimos, tesauros, enciclopedias, diccionarios de colocaciones, etc., como los corpus (Gelbukh y Sidorov, 2010, pág. 27).
 - b) Caracteres alargados. En algunas ocasiones se genera la representación de caracteres repetidos (ejemplo “*felizzzzz*” se convierte en “*feliz*”), que se resuelven por medio de las expresiones regulares.

⁷ <https://languagetool.org/es>

- Tokenización. Aparición concreta de cada palabra separada por comas. Ejemplo: “Informe ejecutivo del tercer semestre” se convierte en “Informe”, “ejecutivo”, “del”, “tercer”, “semestre”.
- Normalización del texto. Es el proceso de transformación de texto, donde se aplican diferentes métodos. En este caso se realizan los siguientes procedimientos:
 - a) Palabras a minúsculas.
 - b) Dígitos a números (6 a seis).
 - c) Lematización del texto.

Como se detalló anteriormente, los pasos que se llevan a cabo a excepción de la revisión ortográfica se realizan implementando la librería *Freeling*⁸, cuyo objetivo es proporcionar una amplia gama de servicios de análisis lingüísticos para diversos idiomas, además ofrece a los desarrolladores de aplicaciones de procesamiento de lenguaje natural, funciones de análisis y anotaciones lingüísticas de textos, con la siguiente reducción del coste de construcción de dichas aplicaciones (Padró, 2010, pág. 15). En la siguiente Figura 4.4 se puede observar los servicios disponibles para cada lengua:

	as	ca	cy	en	es	gl	it	pt	ru
Tokenization	X	X	X	X	X	X	X	X	X
Sentence splitting	X	X	X	X	X	X	X	X	X
Number detection		X		X	X	X	X	X	X
Date detection		X		X	X	X		X	X
Morphological dictionary	X	X	X	X	X	X	X	X	X
Affix rules	X	X	X	X	X	X	X	X	
Multiword detection	X	X	X	X	X	X	X	X	
Basic named entity detection	X	X	X	X	X	X	X	X	X
B-I-O named entity detection				X	X	X			
Named Entity Classification				X	X				
Quantity detection		X		X	X	X		X	X
PoS tagging	X	X	X	X	X	X	X	X	X
WN sense annotation		X		X	X				
UKB sense disambiguation		X		X	X				
Shallow parsing	X	X		X	X	X		X	
Full/dependency parsing	X	X		X	X	X			
Coreference resolution					X				

Figura 4.4 Servicios de análisis disponibles para cada lengua.

⁸ <https://nlp.lsi.upc.edu/freeling/node/1>

Después de aplicar la etapa de preprocesamiento de los comentarios extraídos de Amazon México, la información generada se envía al módulo de “*Algoritmo Basado en Reglas*”.

4.3 Algoritmo Basado en Reglas

El objetivo de este algoritmo es poder detectar los productos que guardan una discordancia de la opinión expresada en el artículo con la calificación en la estrella y proponer una revaloración, es decir, identificar el número de estrellas real utilizando técnicas de procesamiento de lenguaje natural.

Este algoritmo recibe como información de entrada los comentarios que han pasado por la etapa de preprocesamiento. A continuación, se describen las características que conforman este algoritmo:

- Con el conocimiento acerca del dominio del problema se generaron 17 reglas. Estas aplican heurística para determinar la polaridad de un texto basado en las frecuencias de la polaridad de palabras.
- Estas frecuencias se determinan con la combinación de n-gramas, para este caso se utilizaron, 2-gramas, 3-gramas y 4-gramas.
- Los n-gramas requieren el uso de un recurso léxico. En este caso se utilizó un léxico afectivo propuesto por Castro N., Baca Y. & Martínez A. (2015), el cual incorpora un conjunto de 1,224 palabras positivas y 2,119 palabras negativas.
- Las reglas junto con los n-gramas requieren un trato especial en algunas oraciones con las palabras positivas y negativas, ya que estas pueden cambiar o invertir la polaridad. De acuerdo a Fang X., & Zhan J. (2015) describen la siguiente oración donde se puede observar la polaridad invertida:
 - “*El altavoz incorporado también tiene sus usos, pero hasta ahora **nada revolucionario***”

La palabra *revolucionario* tiene una polaridad positiva. Sin embargo, la frase "*nada revolucionario*" da un sentimiento negativo. A esto se le conoce como la negación de verbos y la negación de adjetivos. Para poder trabajar con estos dos elementos Fang X., & Zhan J. (2015) describen los prefijos de negación: ‘no’ y ‘nada’. Por otro lado, esta investigación que se basa en comentarios en idioma español, se llevaron a cabo pruebas

para determinar otros prefijos de negación como las palabras: 'no', 'nunca', 'ni', 'jamás', 'tampoco', 'nada' y 'ninguno'. Por ejemplo:

- *“La mochila es de baja a mediana calidad. **No cumple** con las especificaciones, no es impermeable, con una lluvia ligera se mojaron todos los tickets que tenía dentro. Es pequeña y viene muy arrugada ya que hacen muchos dobleces para meterla en un reducido empaque.”*
- *“Justo a los 4 meses de uso dejó de funcionar, comenzó por no ser reconocido esporádicamente, luego fue más frecuente hasta que ya no respondió, lo probé en otros equipos y tampoco lo reconoce, Amazon solo da dos meses de garantía, la página del proveedor te pide llenar un formulario y **nunca te confirman** si recibieron tu petición, pésimo servicio de Amazon y de Adata”.*
- *“Me mandaron una báscula de comida en lugar de la de la foto, tenga cuidado asegúrense bien del proveedor, según tenía buenos reviews y por eso confíe, pero **ni alcazo**”*
- *“El producto se ve bueno, pero jamás llegó y nunca me avisaron cuál fue el problema, y con el vendedor ni hablar **jamás lo logre** contactar, el equipo de Amazon fue quien me ayudó con el reembolso”*
- *“Yo tuve esta Smart Band y la 4, son productos defectuosos ya que en la ducha hacen falso contacto y se activa la pantalla como si la estuvieras manipulando. El cálculo de pasos es incorrecto, hay días que me dice haber caminado 8,000 pasos sin siquiera haber caminado la mitad. El sensor de pulso **tampoco es fiable**. Tuve que devolver ambos productos, les recomiendo productos Garmin a mí no me han fallado.”*
- *“El audífono derecho dejó de funcionar, no son **nada cómodos** para hacer ejercicio ya que se salen constantemente y tienes que estar atrapando los para que no se caigan”*

- *“He comprado dos controles y ya **ninguno sirve**. Cómo puedo hacer para que los reparen?”*

Asimismo, este tipo de prefijos de negación también tienen su lado opuesto, es decir pueden negar una palabra con polaridad negativa e invertirla a positiva. Por ejemplo:

- *“Me gustó mucho, viene bien empaquetado para que **no se maltrate**. Super fácil de montar, sencillo y elegante.”*

Nota: La palabra **maltrate** (*del verbo maltratar*) de acuerdo al léxico, tiene polaridad negativa, pero el prefijo **no** al ser negativo la convierte a una polaridad positiva.

- *“La compré para mi Nintendo Switch, hasta ahorita corre bien los juegos, sin ningún problema, no noto ralentización y **tampoco me dio problemas** al detectar la memoria, al insertar, fue detectada.”*

Nota: La palabra **problemas** de acuerdo al léxico tiene polaridad negativa, pero el prefijo **tampoco** al ser negativo, la convierte a una polaridad positiva.

Otro tipo de prefijos identificados fueron: ‘sin’ y ‘poco’, por ejemplo:

- *“Una silla se quebró de las cuatro uniones con las patas a la primera sentada **sin esfuerzo** alguno, a los pocos días el resto de las sillas se quebraron también, son de muy mala calidad y resistencia.”*

Nota: La palabra **esfuerzo**, de acuerdo al léxico, tiene polaridad positiva, pero el prefijo **sin**, la niega.

- *“Es un buen producto, muy funcional, las caratulas muy variadas. Lo único que no me gustó es que es un **poco grande**.”*

Nota: La palabra **grande** de acuerdo al léxico, tiene una polaridad positiva pero el prefijo **poco** la niega.

En el caso particular del prefijo ‘sin’ también en algunas oraciones se presenta su lado opuesto. Por ejemplo:

- *“Muy fácil de usar, dejó uno en el estuche, se conecta rápido y **sin falla**.”*

Nota: La palabra **falla** (*del verbo fallar*) de acuerdo al léxico, tiene polaridad negativa pero el prefijo **sin** al ser negativo, la convierte a una polaridad positiva.

También se identificó una palabra especial (**pena**) que dependiendo del contexto puede cambiar la polaridad. Por ejemplo:

- *“Buena calidad. Cumple su proposito...Un poco caro, pero **vale** la **pena**.”*

Nota: La palabra **pena** de acuerdo al léxico tiene una polaridad negativa, pero si la precede la palabra **vale**, cambia la polaridad. Sin embargo, existe la parte opuesta como se muestra a continuación.

- *“Nada que ver. No tiene nada de exacto. Ni siquiera se acerca **no vale** la **pena** la inversión.”*

Esta investigación también muestra que en algunas oraciones, los verbos negativos funcionan como prefijos y pueden negar tanto a verbos como a sustantivos con polaridades positivas. Por ejemplo:

- *“A dos años de haberla comprado, ha **disminuido** su **rendimiento**, el precio es bastante elevado, he tenido mejores experiencias con memorias de otras marcas con mejor calidad a un precio mucho menor.”*

Nota: La palabra **rendimiento** de acuerdo al léxico, tiene una polaridad positiva pero el prefijo **disminuido** (*del verbo disminuir*) cambia la polaridad a negativa.

- *“A seis meses de uso la esponja del asiento **perdió** su **consistencia**, pero lo realmente preocupante fue que el respaldo se venció, la contra donde se atornilla el respaldo, que está en el asiento, rompió la lámina de manera que lo sujetaba, yo solo peso 70 Kg, lo cual no es mucho para lo que dice el anuncio que soporta la silla.”*

Nota: La palabra **consistencia** de acuerdo al léxico, tiene una polaridad positiva pero el prefijo **perdió** (*del verbo perder*) cambia la polaridad a negativa.

- *“Bien, la memoria tiene buena capacidad, velocidad y se ve elegante. Pero despues de un mes de uso todos los dispositivos **dejaron** de **reconocerla**. TODA MI*

INFORMACION SIN RESPALDAR se perdió ahí, me siento molesto por que pense que una memoria de tal magnitud seria sinonimo de calidad. Espero me hagan el reembolso, faltaban pocos dias para que ya no calificara.”

Nota: La palabra **reconocerla** (*del verbo reconocer*) de acuerdo al léxico, tiene una polaridad positiva pero el prefijo **dejaron** (*del verbo dejar*) cambia la polaridad a negativa.

- *“La silla llegó rota del respaldo por lo que solicité un cambio, la segunda unidad fallaba para ajustar la altura por lo que solicité una devolución”*

Nota: La palabra **ajustar** de acuerdo al léxico, tiene una polaridad positiva pero el prefijo **fallaba** (*del verbo fallar*) cambia la polaridad a negativa.

- Con el uso de los n-gramas y las reglas, el algoritmo devuelve dos conjuntos como salida:
 - El primero son las frecuencias de palabras positivas.
 - El segundo son las frecuencias de palabras negativas.
- La información devuelta es adquirida por una función cuyo propósito es aplicar una base matemática para identificar la polaridad de una oración. Geetha M., Singha P., & Sinha S. (2018) proponen la siguiente Fórmula 4.1:

$$\text{polaridad del sentimiento del cliente} = \frac{\text{palabras positivas}}{\text{palabras negativas}} \quad (4.1)$$

Esta fórmula maneja los siguientes valores de corte: Se consideran “positivas” cuando son mayores a 1.5, “neutrales” si están entre 1 y 1.5, “negativas” si son menores que 1.

Haciendo un pequeño paréntesis, la mayoría de los trabajos en el estado del arte que aplican análisis de sentimientos en la minería de opinión, cuando trabajan con comentarios que son evaluados en una escala de 1 a 5 estrellas, la gran mayoría combinan 1 y 2 estrellas como negativas, 3 estrellas como neutrales y, 4 y 5 estrellas como positivas. Muy pocos artículos trabajan con los 5 niveles. Estos trabajos hacen uso de un recurso léxico llamado SentiWordNet, el cual maneja un léxico afectivo y asigna a cada palabra un peso. Aplican una fórmula matemática que maneja 5 tipos de cortes y llevan a cabo la clasificación de los comentarios.

La Fórmula 4.1 solo maneja 3 tipos de cortes, sin embargo, con el estudio de los comentarios es posible clasificar las opiniones en una escala de 1 a 5 estrellas utilizando la fórmula 4.1: valores positivos corresponden a cuatro o cinco estrellas, neutrales a tres estrellas y negativos a una o dos estrellas. Donde, los comentarios de cinco estrellas reciben únicamente palabras positivas, los de cuatro estrellas recibe un mayor número de palabras positivas que negativas, los de tres estrellas mantienen el mismo número de palabras positivas como negativas, las de dos estrellas recibe un mayor número de palabras negativas que positivas y de una estrella recibe únicamente palabras negativas. A continuación, se presentan 5 ejemplos para observar las clasificaciones.



- 1) *“Es una base muy ligera pero muy buena. No la calificaría como la mejor pero si es aceptable aun precio bastante bueno.”*

Este comentario corresponde a una opinión que fue evaluada con 4 estrellas. Al ingresar al Algoritmo Basado en Reglas, se obtendrá como salida:

palabras positivas: ['ligero', 'bueno', 'aceptable', 'bueno']

palabras negativas: ['no-mejor']

Al aplicar la fórmula 4.1, el resultado es 2 y de acuerdo al corte, los valores mayores a 1.5 son positivos. Pero se puede observar que este corte, tolera palabras negativas y, por lo tanto, no es un comentario de cinco estrellas, pues se espera que la experiencia con el producto maneje solo palabras positivas.



- 2) *“La brocha para maquillaje no es tan suave, se siente ligeramente picazón al estar esparciendo la base. Las brochas pequeñas son muy prácticas.”*

Este comentario corresponde a una opinión que fue evaluada con 3 estrellas. Al ingresar al Algoritmo Basado en Reglas, se obtendrá como salida:

palabras positivas: ['esparcir', 'práctico']

palabras negativas: ['no-suave']

Al aplicar la fórmula 4.1, el resultado es 1 y de acuerdo al corte, los valores que están entre 1 y 1.5 son neutrales.



- 3) *“La mesa es muy bonita y esta bien terminada, el problema son las patas, no se mantienen fijas, lo que hace que se incline mas de un lado que del otro, por lo tanto es inestable, las "tuerquitas" que trae para fijar las patas no sirven de nada. No me gusto, la regrese!”*

Este comentario corresponde a una opinión que fue evaluada con 2 estrellas. Al ingresar al Algoritmo Basado en Reglas, se obtendrá como salida:

palabras positivas: ['bonito', 'bien', 'terminar']

palabras negativas: ['problema', 'no', 'inestable', 'no-gustar', 'no-mantener']

Al aplicar la fórmula 4.1, el resultado es 0.43 y de acuerdo al corte, los valores que son menores a 1, son negativos. Pero se puede observar que este corte, tolera palabras positivas y, por lo tanto, no es un comentario de una estrella, pues se espera que la experiencia con el producto maneje solo palabras negativas.



- 4) *“Excelente servicio y producto, fue un regalo para mi hijo y está feliz!!!”*

Este comentario corresponde a una opinión que fue evaluada con 5 estrellas. Al ingresar al Algoritmo Basado en Reglas, se obtendrá como salida:

palabras positivas: ['excelente', 'regalo', 'feliz']

palabras negativas: []

Este comentario se trata de un caso especial, y debido a que solo manejas palabras positivas, no se aplica la fórmula 4.1, simplemente se clasifica como 5 estrellas.



- 5) *“Si eres de México, no recomiendo hacer la compra, desafortunadamente me pasó igual que a otros que ya hicieron la compra, me llegó incompleta la silla, la caja estaba demasiado maltratada, y como se ve en la foto, la piel de la silla estaba rota. Además no vale la pena por el tiempo de espera, la esperé un mes y llegó mal.”*

Este comentario corresponde a una opinión que fue evaluada con 1 estrella. Al ingresar al Algoritmo Basado en Reglas, se obtendrá como salida:

palabras positivas: []

palabras negativas: ['desafortunadamente', 'incompleto', 'demasiado', 'maltratar', 'no', 'pena', 'mal', 'no-recomendar']

Este comentario se trata de un caso especial, y debido a que solo manejas palabras negativas, no se aplica la fórmula 4.1, simplemente se clasifica como 1 estrella.



- 6) *“Porque al momento de ponerla en mi Switch me dice que tengo espacio de 366 GB si se supone que la memoria es de 400?”*

Este comentario corresponde a una opinión que fue evaluada con 2 estrella. Al ingresar al Algoritmo Basado en Reglas, se obtendrá como salida:

palabras positivas: []

palabras negativas: []

Este comentario se trata de un caso especial, y debido a que no maneja palabras que contengan polaridad, no se aplica la fórmula 4.1, simplemente se clasifica como 0 estrellas (sin polaridad o carente de subjetividad).

Después de aplicar el Algoritmo Basado en Reglas a los comentarios procesados, la información generada se envía al módulo de “Evaluación de polaridad”.

4.4 Evaluación de polaridad

Como información de entrada recibe la reasignación de estrellas propuesto por el algoritmo Basado en Reglas y las calificaciones que originalmente otorgaron los clientes a los productos. En otras palabras, se lleva a cabo la comparación uno a uno, es decir, se verifica si las revaloraciones de los comentarios propuesta por el algoritmo coinciden o son iguales a las calificaciones que los clientes cedieron, en caso de ser afirmativo se obtendría la suma de las opiniones que guardan una congruencia de la opinión expresada del cliente respecto a su calificación. En caso de ser negativo, se calcularía la suma total de comentarios menos la suma de los comentarios más congruentes y como resultado se obtendría la cantidad de comentarios que guardan discordancia.

Capítulo 5 Pruebas y resultados

5.1 Análisis de la relación entre calificaciones de estrellas y los comentarios

Antes de abordar los experimentos es necesario explicar cómo las calificaciones de estrellas se asocian con los comentarios. En otras palabras, que características de las opiniones distinguen un comentario que está relacionado con una calificación el cual puede tener una valoración de 1 hasta 5 estrellas. De acuerdo con la investigación realizada en la literatura los trabajos relacionados con la minería de opinión, la gran mayoría solo trabajan con tres tipos de polaridades (positivas, negativas y neutrales) y muy pocos con cinco (fuertemente positivas, positivas, neutrales, negativas y fuertemente negativas). Estos últimos no describen cómo las calificaciones de estrellas se asocian con las reseñas. Como se observa en la siguiente Figura 5.1 que describe una evaluación general que resumen las impresiones de los clientes sobre algún determinado artículo, se pueden apreciar estos cinco tipos de polaridades.



Figura 5.1 Valoración de comentarios.

Por lo tanto, en esta investigación se procedió a estudiar una gran cantidad de reseñas del corpus de comentarios, con el objetivo de poder determinar qué características utilizan los clientes para distinguir (otorgar) comentarios con valoraciones entre cinco, cuatro, tres, dos y una estrella. A continuación, se describen algunos ejemplos, donde se explican estas opiniones utilizando un léxico afectivo, según Kusrini et al. (2019, pág. 366) “Son palabras que se usan comúnmente para expresar sentimientos positivos o negativos y son importantes para el análisis de opinión por razones obvias”.

Nota: Para poder detectar las palabras afectivas de un comentario, es necesario aplicar la lematización, de acuerdo a Cruz (2017, pág. 18) “Consiste en determinar la forma, clase o categoría gramatical de cada palabra de una oración”. Es una técnica de reducción, permite detectar variantes morfológicas de un mismo. Por ejemplo: palabras como “computo”, “computadoras”, “computable”, “computación” son variantes del término computar, y reemplazadas por el termino raíz o lema.

Por tal motivo, los siguientes ejemplos se mantienen originalmente como se extrajeron del corpus de comentarios para su mayor entendimiento. Asimismo, las palabras que se identificaron con polaridad positiva, se presentan de color azul y las palabras negativas de color rojo. Cabe mencionar que en algunas ocasiones las palabras con polaridades positivas pueden cambiar a una polaridad negativa. Esto sucede principalmente por la negación de verbos y adjetivos, lo que causa que se invierta la polaridad de una palabra. Esta información es presentada por (Fang et al., 2015) en el artículo “Análisis de opinión utilizando datos de reseñas de productos”.

- **Comentarios de 5 estrellas:** Este tipo de opiniones se caracterizan porque en el cuerpo de la reseña, solo incorporan palabras positivas, donde relatan la experiencia fuertemente positiva que han tenido con el producto adquirido.

- “Fue para *regalo* de mi mamá Y ella súper *emocionada* por lo práctica y *eficiente* que es Alexa.”

De acuerdo al léxico afectivo, se identificaron tres palabras con polaridad positiva. Como se puede observar, en el cuerpo de la reseña, esta no contiene en absoluto algo negativo.

- “Tarjeta muy *buena*, funciona de *maravilla* y súper rápida la use en mi android y mas *contento* que nunca.”

De acuerdo al léxico afectivo, se identificaron tres palabras con polaridad positiva. Al igual que el ejemplo anterior, la experiencia con este producto, es fuertemente positiva.

- **Comentarios de 4 estrellas:** Este tipo de opiniones son positivas, pero a diferencia de comentarios de 5 estrellas, estos se caracterizan porque en el cuerpo de la reseña, también

incorporan palabras negativas. Sin embargo, éstas últimas aparecen, cuando la experiencia con el producto, no ha sido cien por ciento satisfactoria.

- *“Es ligero y por lo mismo **no** te **da** una sensación de **seguridad** (estabilidad) pero en general es **funcional**. Incluye el manual con las **diferentes** posiciones en las cuales **puedes** usarlo. Los ventiladores muy silenciosos aunque de baja potencia. Relación **calidad**-precio es **bueno**.”*

De acuerdo al léxico afectivo, se identificaron cinco palabras con polaridad positiva y tres con polaridad negativa. Como se explicó anteriormente, es un comentario positivo, pero no fuertemente positivo, ya que la experiencia con el producto no fue lo suficientemente satisfactoria para recibir cinco estrellas.

- *“**Buen** producto que funciona **bien**, la aplicación dice muchas cosas **útiles**. Pero **no** debería tener ads porque ya pagué por el producto.”*

De acuerdo al léxico afectivo, se identificaron tres palabras con polaridad positiva y una con polaridad negativa. Como se observa, es un comentario positivo, pero que incorpora una parte negativa, por lo cual, su valoración fue de cuatro estrellas.

- **Comentarios de 3 estrellas:** Las reseñas de tres estrellas, generalmente se caracterizan porque en el cuerpo del comentario, tanto las palabras con polaridades positivas y negativas se encuentran balanceadas como ocurre en el primer ejemplo, pero en el segundo ejemplo, la diferencia es mínima.

- *“El color es **bonito**, pero al ser la punta muy gruesa, particularmente a mi se me hace **difícil** pintas el borde de los labios.”*

De acuerdo al léxico afectivo, se identificó una palabra con polaridad positiva y una con polaridad negativa. Es un comentario balanceado en este sentido y por lo tanto fue valorado con tres estrellas.

- *“La balanza funciona muy **bien** y la app es muy **facil** de usar, pero desde que la compre antes de subirme ya tiene 25 gramos en la cuenta que **no** he **podido** quitar, pero de ahí en fuera todo **bien**.”*

De acuerdo al léxico afectivo, se identificaron tres palabras con polaridad positiva y dos con polaridad negativa. Como se puede apreciar, este comentario tiene una carga positiva y una negativa casi en la misma proporción, por lo tanto, también esta opinión fue evaluada con tres estrellas.

- **Comentarios de 2 estrellas:** Las opiniones de dos estrellas de acuerdo a esta investigación, se caracterizan porque tienen una mayor carga de polaridades negativas que positivas, en otras palabras, ocurre lo contrario a comentarios de cuatro estrellas, es decir, la experiencia con el producto fue negativa, pero no fuertemente negativa, ya que en el cuerpo de la opinión, puede apreciarse que en algún momento mencionan algo positivo del producto.
 - “Las 4 sillas se encontraban **flojas** después de algunos meses. **1 colapsó** mientras estaba sentado. **No** es posible apretarlas más porque se va **dañando** la madera en una zona **inaccesible**. Es un **mal** diseño estructural a pesar de que tienen **buen** diseño estético.”

De acuerdo al léxico afectivo, se identificaron seis palabras con polaridad negativa y una con polaridad positiva. Por lo tanto, es un comentario que fue evaluado con dos estrellas, pese a la mala experiencia, se destaca una parte positiva del producto.

- “Es **buen** producto, sin **embargo** después de un tiempo resulta **doloroso** el tenerlo en la oreja... **No** es **compatible** con video llamada y a veces se desconecta de mi teléfono.”

De acuerdo al léxico afectivo, se identificaron cuatro palabras con polaridad negativa y una con polaridad positiva. Ocurre lo mismo que en el ejemplo anterior, solamente se destaca una parte positiva de la mala experiencia con el artículo.

- **Comentarios de 1 estrella:** Los comentarios de una estrella, se caracterizan porque en toda la opinión, no existen elementos que hablen bien en algún momento del artículo, por lo cual, esta clase de opiniones reciben la valoración de una estrella, pues la experiencia con el producto es fuertemente negativa.
 - “Los pedi y los tuve que devolver porque la cancelación de ruido **no** funcionaba, los volví a pedir y ahora se desconectaba de la nada, los volví a pedir y otra vez **no** funciona la cancelación de ruido y el modo ambiente, estoy **harto**.”

De acuerdo al léxico afectivo, se identificaron dos palabras con polaridad negativa. Es un tipo de comentario donde se relata la mala experiencia que tuvo un cliente con el producto adquirido y en ningún momento se menciona algo positivo del artículo.

- “*No me gusto, es de muy mala calidad, no es del material que dicen en su descripción, se ve y se siente bastante corriente.*”

De acuerdo al léxico afectivo, se identificaron seis palabras con polaridad negativa. Es un comentario fuertemente negativo, al igual que en el ejemplo anterior. La diferencia es que, este comentario tiene un número mayor de palabras negativas, pero al no incorporar algo positivo con la experiencia del producto, este recibe una calificación de una estrella.

De acuerdo con los cinco tipos de polaridades que se investigaron en el corpus de comentarios y, como es que las calificaciones de estrellas se asocian con las opiniones, se encontró un artículo en la literatura que muestra una fórmula para clasificar la polaridad del sentimiento del cliente en positiva, neutral y negativa la cual fue propuesta por (Geetha et al., 2018, pág. 48). Dicha fórmula se describe a continuación:

$$\text{polaridad del sentimiento del cliente} = \frac{\text{palabras positivas}}{\text{palabras negativas}} \quad (5.1)$$

Como se puede observar, esta fórmula 5.1 se basa en las frecuencias de la polaridad de palabras, es decir, dividir el número total de palabras positivas entre negativas, además, comprende tres tipos de cortes para posteriormente llevar a cabo la clasificación del sentimiento del cliente:

- Polaridad positiva: Los valores mayores a 1.5, se consideran comentarios con polaridad positiva.
- Polaridad neutral: Los valores mayores o iguales a 1 y menores o iguales a 1.5, se consideran comentarios con polaridad neutral.
- Polaridad negativa: Los valores menores a 1, se consideran comentarios con polaridad negativa.

De acuerdo con la información explicada, para obtener la clasificación del sentimiento del cliente por medio de la fórmula 5.1 y junto con esta investigación, donde se comprobó con ejemplos como

es que las calificaciones de estrellas se asocian con los comentarios en una escala de 1 a 5, se pudo corroborar que cuando estas opiniones pasan por la Fórmula 5.1, de acuerdo a los resultados matemáticos, tienen una estrecha relación con las características que utilizan los clientes para distinguir comentarios con valoraciones entre cinco, cuatro, tres, dos y una estrella. Para demostrar esto, a continuación se toman en cuenta los mismos ejemplos, pero en esta ocasión, se abordarán primero los comentarios de cuatro, tres y dos estrellas, dejando al final los comentarios de cinco y una estrella.

- **Comentarios de 4 estrellas:**

- “Es ligero y por lo mismo **no** te **da** una sensación de **seguridad** (estabilidad) pero en general es **funcional**. Incluye el manual con las **diferentes** posiciones en las cuales **puedes** usarlo. Los ventiladores muy silenciosos aunque de baja potencia. Relación **calidad**-precio es **bueno**.”

$$\text{polaridad del sentimiento del cliente} = \frac{5}{3} = 1.67$$

- “**Buen** producto que funciona **bien**, la aplicación dice muchas cosas **útiles**. Pero **no** debería tener ads porque ya pagué por el producto.”

$$\text{polaridad del sentimiento del cliente} = \frac{3}{1} = 3$$

Los resultados de la Fórmula 5.1 en ambos ejemplos, se consideran comentarios positivos. De acuerdo al primer corte, aquellos valores superiores a 1.5, son reseñas con polaridad positiva. Por lo tanto, la fórmula 5.1 clasifica comentarios positivos sin importar la intensidad, es decir, no distingue entre opiniones de cuatro y de cinco estrellas, pues ambas se consideran positivas, sin embargo, se puede observar que en la parte del denominador, la fórmula 5.1 tolera palabras con polaridad negativa, de acuerdo a esto y al análisis realizado en el corpus de comentarios, una opinión de estas características, es aquella que tiene una mayor frecuencia de palabras positivas que negativas y al ingresar este tipo de reseñas, el resultado siempre será mayor a 1.5, por esta razón, tanto en la fórmula 5.1 como en el análisis, se consideran como comentarios de cuatro estrellas.

- **Comentarios de 3 estrellas:**

- “El color es **bonito**, pero al ser la punta muy gruesa, particularmente a mi se me hace **difícil** pintar el borde de los labios.”

$$\text{polaridad del sentimiento del cliente} = \frac{1}{1} = 1$$

- “La bascula funciona muy **bien** y la app es muy **facil** de usar, pero desde que la compre antes de subirme ya tiene 25 gramos en la cuenta que **no** he **podido** quitar, pero de ahí en fuera todo **bien**.”

$$\text{polaridad del sentimiento del cliente} = \frac{3}{2} = 1.5$$

Los resultados de la Fórmula 5.1 en ambos ejemplos, se consideran comentarios neutrales. De acuerdo al segundo corte, aquellos valores mayores o iguales a 1 y menores o iguales a 1.5, se consideran comentarios con polaridad neutral. Para este caso, los ejemplos están dentro de este rango. Por lo tanto, una opinión de estas características, es aquella que tiene por lo regular, las mismas frecuencias de palabras positivas y negativas, siempre y cuando se cumpla que, el número de palabras negativas no sea mayor al de palabras positivas, pues así se estipula en el corte de la fórmula 5.1 para comentarios neutros. Por esta razón, tanto en la fórmula 5.1 como en el análisis, se consideran como comentarios de tres estrellas.

- **Comentarios de 2 estrellas:**

- “Las 4 sillas se encontraban **flojas** después de algunos meses. **1 colapsó** mientras estaba sentado. **No** es posible apretarlas más porque se va **dañando** la madera en una zona **inaccesible**. Es un **mal** diseño estructural a pesar de que tienen **buen** diseño estético.”

$$\text{polaridad del sentimiento del cliente} = \frac{1}{2} = 0.5$$

- “Es **buen** producto, sin **embargo** después de un tiempo resulta **doloroso** el tenerlo en la oreja... **No** es **compatible** con video llamada y a veces se desconecta de mi teléfono.”

$$\text{polaridad del sentimiento del cliente} = \frac{1}{4} = 0.25$$

Los resultados de la Fórmula 5.1 en ambos ejemplos, se consideran comentarios negativos. De acuerdo al tercer corte, aquellos valores menores a 1, se consideran comentarios con polaridad negativa. Por lo tanto, la fórmula 5.1 clasifica comentarios negativos sin importar la intensidad, es decir, no distingue entre opiniones de dos y de una estrella, pues ambas se consideran negativas, sin embargo, se puede observar que en la parte del numerador, la fórmula 5.1 tolera palabras con polaridad positiva, de acuerdo a esto y al análisis realizado en el corpus de comentarios, una opinión de estas características, es aquella que tiene una mayor frecuencia de palabras negativas que positivas y al ingresar este tipo de reseñas, el resultado siempre será menor a 1, por esta razón, tanto en la fórmula 5.1 como en el análisis, se consideran como comentarios de dos estrellas.

Hasta este punto de la investigación, se ha comprobado junto con la literatura como distinguir comentarios que son evaluados con cuatro, tres y dos estrellas. Además, se dejó en claro, que la fórmula 5.1 no distingue que tan positiva o negativa puede ser una opinión, pero precisamente, con ayuda de la fórmula 5.1 y las características que se describieron anteriormente para cada caso de comentarios, se pueden clasificar perfectamente en cuatro, tres y dos estrellas. Para comentarios de cinco y de una estrella, es un caso especial, por tal motivo se discuten a continuación:

- **Comentarios de 5 estrellas:**

- “Fue para **regalo** de mi mamá Y ella súper **emocionada** por lo práctica y **eficiente** que es Alexa.”
- “Tarjeta muy **buena**, funciona de **maravilla** y súper rápida la use en mi android y mas **contento** que nunca.”

Para comentarios de cinco estrellas se mencionó, que se caracterizan por incorporar en el cuerpo de la reseña, únicamente palabras con polaridad positiva, lo que convierte la experiencia del cliente con el producto como cien por ciento positiva. Esta es la pequeña brecha que existe entre comentarios de cinco y de cuatro estrellas. Por tal motivo, la fórmula 5.1 es funcional para comentarios de cuatro

estrellas. Por ello, al no existir palabras con polaridad negativa en un comentario, no es necesario observar los resultados de la fórmula 5.1, ya que se entiende que es una reseña de cinco estrellas.

- **Comentarios de 1 estrella:**

- *“Los pedi y los tuve que devolver porque la cancelación de ruido **no** funcionaba, los volví a pedir y ahora se desconectaba de la nada, los volvi a pedir y otra vez nos funciona la cancelación de ruido y el modo ambiente, estoy **harto**.”*
- *“**No** me **gusto**, es de muy **mala calidad**, **no** es del material que dicen en su descripción, se ve y se siente bastante **corriente**.”*

Para comentarios de una estrella se mencionó, que se caracterizan por incorporar en el cuerpo de la reseña, únicamente palabras con polaridad negativa, lo que convierte una mala experiencia del cliente con el producto. Esta es la pequeña brecha que existe entre comentarios de una y de dos estrellas. Por tal motivo, la fórmula 5.1 es funcional para comentarios de dos estrellas. Por ello, al no existir palabras con polaridad positiva en un comentario, no es necesario observar los resultados de la fórmula 5.1, ya que se entiende que es una reseña de una estrella.

Una vez explicado el análisis de la relación entre calificaciones de estrellas y los comentarios, a continuación, se presentan experimentos y resultados que se realizaron con algoritmos de aprendizaje automático y la creación de un algoritmo heurístico o basado en reglas. Los experimentos sirvieron para determinar si los comentarios que se encuentran clasificados según el número de estrellas que los clientes asignaron a los productos que adquirieron no guardan congruencia entre la opinión del cliente y la evaluación en la estrella.

El primer experimento se realizó con Bosques Aleatorios, Regresión Logística, Naive Bayes y Máquina de Soporte Vectorial.

El segundo experimento hace uso del Algoritmo Basado en Reglas.

El tercer experimento, pone en marcha nuevamente los anteriores, pero con una pequeña variación donde se muestra la comparación de los resultados obtenidos.

Finalmente se llevó a cabo una prueba piloto, donde se implementó una interfaz web la cual hace uso de cada uno de los módulos del método de solución, donde se obtiene la calificación general dada a un producto y posteriormente se calcula la calificación propuesta por el mejor algoritmo seleccionado en los experimentos. Es importante aclarar, que no se utilizaron evaluadores, ya que como objetivo es poder modelar el fenómeno tal cual aparece en la realidad, es decir, aprender sobre la manera en que los clientes reseñan los comentarios y les asocian sus estrellas. Precisamente, estas calificaciones se toman como las etiquetas de las opiniones.

5.2 Corpus de comentarios

El corpus de comentarios presentado en el capítulo 4 cuenta con 1,005,364 reseñas el cual está conformado por diferentes países traducidos al español, sin embargo, solo se tomarán en cuenta las opiniones que pertenezcan a México. Esto debido a que este subconjunto estaría cumpliendo con los objetivos y alcances para este trabajo de investigación. El corpus para México cuenta con 24,125 reseñas y en la siguiente Tabla 5.1 puede observarse que porcentaje pertenece por clase de estrellas:

Tabla 5.1 Distribución de comentarios por clase.

Estrellas	Cantidad de comentarios	Porcentaje
5	128,796	73
4	21,893	12
3	9,332	5
2	4,825	3
1	12,175	7

Teniendo en cuenta la información de la Tabla 5.1 la distribución de comentarios por clase se encuentra totalmente desbalanceada, por tal motivo, para llevar a cabo los experimentos se tomó el 3% que corresponde al total de comentarios de la clase minoritaria con 4,825 reseñas.

5.3 Valores de las características textuales

De acuerdo al estado del arte en la minería de opinión, para poder detectar la polaridad de una oración es necesario extraer cuatro características las cuales son: sustantivos, verbos, adjetivos y adverbios. Estas propiedades (características textuales) deben ser transformadas en valores numéricos los cuales se definen con la Fórmula 5.2.

$$\text{Características textuales} = TF \times IDF \quad (5.2)$$

“Esta medida se utiliza a menudo como un factor de ponderación en la recuperación de información y la minería de texto” (Valdez, 2019, pág. 11). Es importante mencionar que TF-IDF recibe como información de entrada N-gramas, que “Son secuencias de elementos tal como aparecen en un documento” (Gelbukh, A., & Sidorov, G. 2013, pág. 49). Nota: esta medida se aplica únicamente para el experimento 1, debido a que se emplean algoritmos de aprendizaje automático antes mencionados, mientras que para el experimento 2, no es necesario ya que utiliza un léxico afectivo en lugar de la Fórmula 5.2.

5.4 Métricas de evaluación

Para poder cuantificar el desempeño de un modelo, existen métricas de rendimiento que pueden utilizarse para medir la relevancia de un modelo como la precisión, la exhaustividad y el valor F. Antes de ver los detalles de estas métricas de evaluación, es necesario hablar de la matriz de confusión, una matriz que dispone del rendimiento de un algoritmo de aprendizaje. La matriz de confusión es simplemente una matriz cuadrática que informa del recuento de las predicciones de Verdadero positivo (TP, del inglés, True positive), Verdadero negativo (TN, del inglés, True negative), Falso positivo (FP, del inglés False positive) y Falso negativo (FN, del inglés, False negative) de un clasificador, como se muestra en la Figura 5.2 (Raschka, & Mirjalili, 2019, pág. 257).

		Clase predicha	
		<i>P</i>	<i>N</i>
Clase real	<i>P</i>	Verdadero positivo (TP)	Falso negativo (FN)
	<i>N</i>	Falso positivo (FP)	Verdadero negativo (TN)

Figura 5.2 Matriz de confusión.

Como resultado de la explicación de la matriz de confusión presentada en la Figura 5.2, se procede a describir las tres métricas de evaluación previamente mencionadas:

- **Precisión:** Mide la exactitud de un clasificador, cuántos de los documentos recuperados son correctos. Una mayor precisión significa menos falsos positivos, mientras que una menor precisión significa más falsos positivos. La precisión (*P*) es la proporción de números de instancias correctamente clasificados del total. Puede definirse como la Fórmula 5.3 (Haque, Nawal, & Muhammad, 2018, pág. 4).

$$precisión = \frac{TP}{TP + FP} \quad (5.3)$$

- **Exhaustividad:** Calcula la sensibilidad de un clasificador, cuántos datos positivos devuelve. Una mayor cobertura significa menos falsos negativos mientras que una menor cobertura significa más falsos negativos. Puede definirse como la Fórmula 5.4 (Haque, Nawal, & Muhammad, 2018, pág. 4).

$$exhaustividad = \frac{TP}{TP + FN} \quad (5.4)$$

- **Valor F:** La combinación de las Fórmulas 5.3 y 5.4 producen métricas únicas conocidas como medida *F* y es la media armónica ponderada de precisión y exhaustividad. Puede definirse como la Fórmula 5.5 (Haque, Nawal, & Muhammad, 2018, pág. 4).

$$valor\ F = \frac{2 \times precisión \times exhaustividad}{precisión + exhaustividad} \quad (5.5)$$

5.5 Experimento 1: Algoritmos de aprendizaje automático

Las razones por la cual se decide realizar el primer experimento con algoritmos de aprendizaje automático es por un lado, que la mayoría de los trabajos en la literatura se enfocan al idioma inglés con las polaridades positivas, neutrales y negativas. Muy pocos con las polaridades fuertemente positivas, positivas, neutrales, negativas y fuertemente negativas. La siguiente razón es que se utilizan evaluadores para etiquetar los comentarios y posteriormente entrenarlos. Otro motivo se debe, a que los pocos trabajos existentes en idioma español no trabajan con plataformas que manejen comentarios de productos como lo es Amazon México, por tal motivo, en el método de solución se crea un algoritmo basado en reglas para obtener una línea base, es decir, se propone un método capaz de analizar los cinco tipos de polaridades en idioma español. De igual forma no requiere etiquetadores humanos, ya que el objetivo general de este algoritmo es reducir la asimetría entre la polaridad de las reseñas y las valoraciones numéricas dadas por los clientes de productos de la plataforma web de Amazon México utilizando la minería de opiniones. Sin embargo, al no utilizar evaluadores humanos, es necesario comparar los resultados de este algoritmo contra algo, en otras palabras, los algoritmos de aprendizaje automático ayudarían a determinar que tan bien se desempeña el algoritmo propuesto en el método de solución.

Este experimento demuestra que existen comentarios que no guardan congruencia con la opinión del producto y la valoración que otorgan los clientes en una escala de 1 a 5 estrellas. Por lo cual, se utilizaron las opiniones del corpus de comentarios y sus calificaciones como etiquetas para entrenar los algoritmos. Así mismo, en unas tablas se presentan las valoraciones originales y las reasignadas por el algoritmo con el mejor rendimiento obtenido por medio de las métricas de clasificación previamente mencionadas.

Los algoritmos utilizados para este experimento fueron Bosques Aleatorios, Regresión Logística, Naive Bayes y Máquina de Soporte Vectorial.

Antes de abordar estos algoritmos es necesario hacer la representación del texto, para este caso se utilizó TF-IDF, además, se comentó que esta medida recibe como información de entrada N-gramas, por lo tanto, este experimento se basa en dos pruebas la primera utiliza TF-IDF con 1-grama, la segunda utiliza TF-IDF con 2-gramas.

En ambas pruebas se tomaron en cuenta 24,125 opiniones del corpus de comentarios de los cuales por cada clase de estrellas corresponden 4,825 reseñas con la intención de balancear la información, posteriormente se llevó a cabo el preprocesamiento.

A continuación, se menciona el método que se utilizó para la optimización de los parámetros de ajuste, también conocidos como hiperparámetros. Principalmente existen dos técnicas, la primera recibe el nombre de búsqueda en cuadrícula, según Belete et al. (2021, pág. 3) “Se puede describir como una exploración exhaustiva o un método de fuerza bruta que prueba todas las combinaciones de hiperparámetros dados a la configuración de la cuadrícula”. La segunda recibe el nombre de búsqueda aleatoria, de acuerdo a Elgeldawi et al. (2021, pág. 7) “Es una técnica en la que se utilizan combinaciones aleatorias de los hiperparámetros para encontrar la mejor solución para el modelo en consideración. Por ejemplo, en lugar de rotar las 100,000 muestras solo se verifican 1,000 muestras aleatorias de conjuntos de hiperparámetros”. De acuerdo a estas dos técnicas, “Aunque la búsqueda de cuadrículas es un enfoque potente para encontrar el conjunto de parámetros óptimo, la evaluación de todas las posibles combinaciones de parámetros es también muy costosa computacionalmente hablando. Un enfoque alternativo para muestrear diferentes combinaciones de parámetros es la búsqueda aleatoria” (Raschka, S., & Mirjalili, V. 2019, pág. 255). Por tal motivo, se decide utilizar la técnica de búsqueda aleatoria y a continuación, se presentan los valores de los parámetros de ajuste de cada algoritmo mencionado anteriormente:

Combinación de hiperparámetros:

- Bosques Aleatorios:
 - o `n_estimators = 1000`
 - o `min_samples_split = 10`
 - o `min_samples_leaf = 1`
 - o `max_features = 'sqrt'`
 - o `max_depth = 90`
 - o `criterion = 'gini'`
 - o `bootstrap = False`
 - o `random_state = 42`
- Regresión Logística:
 - o `Solver = 'lbfgs'`
 - o `Penalty = 'l2'`
 - o `max_iter = 2500`

- o `random_state = 42`
- Naive Bayes:
 - o `Alpha = 1`
 - o `fit_prior = True`
- Máquina de Soporte Vectorial:
 - o `C = 1`
 - o `Kernel = 'rbf'`
 - o `Degree = 1`
 - o `gamma = 'scale'`

Para el entrenamiento y las pruebas los datos se dividieron en 75 y 25 por ciento que corresponden a 18,093 y 6,032 comentarios.

5.5.1 Primera prueba con 1-grama

Considera dos campos del corpus de comentarios:

- El primero corresponde con los comentarios procesados los cuales ingresan a TF-IDF como información de entrada en forma de 1-grama.
- El segundo corresponde con las valoraciones de las opiniones.

Una vez que TF-IDF devuelve los comentarios como valores, se ingresan a los algoritmos de aprendizaje supervisado y las calificaciones de las opiniones como etiquetas. El tiempo de ejecución fue de 3 minutos. A continuación, se muestran los resultados en la Tabla 5.2.

Tabla 5.2 Resultados del conjunto de pruebas.

Algoritmos utilizados	Estrellas	Métricas de clasificación		
		precisión	exhaustividad	valor-F
Bosques Aleatorios	1	0.54	0.67	0.60
	2	0.39	0.33	0.36
	3	0.34	0.29	0.31
	4	0.45	0.36	0.40
	5	0.59	0.73	0.65
Regresión Logística	1	0.59	0.63	0.61
	2	0.40	0.37	0.39
	3	0.33	0.31	0.32
	4	0.42	0.41	0.41
	5	0.64	0.69	0.66

Naive Bayes	1	0.58	0.61	0.59
	2	0.39	0.37	0.38
	3	0.32	0.31	0.31
	4	0.40	0.39	0.39
	5	0.64	0.66	0.65
Máquina de Soporte Vectorial	1	0.60	0.63	0.61
	2	0.41	0.40	0.41
	3	0.33	0.34	0.34
	4	0.42	0.39	0.41
	5	0.67	0.67	0.67

De acuerdo a los resultados de la Tabla 5.2 Máquina de Soporte Vectorial reportó ligeramente mejores métricas que los demás algoritmos. Por lo tanto, para esta prueba del experimento 1, se presenta la Tabla 5.3 con las valoraciones originales del corpus de comentarios y las reasignaciones propuestas, pero, considerando únicamente el 25 por ciento de la información que comprende la parte de pruebas del algoritmo que tuvo los mejores resultados:

Tabla 5.3 Comparación de evaluaciones originales y reasignadas (prueba 1).

Estrellas	Cantidad de comentarios (valoración original)	Cantidad de comentarios (valoración reasignada)
5	1,207	1,212
4	1,207	1,130
3	1,206	1,236
2	1,206	1,190
1	1,206	1,264

En la Tabla 5.3 se observa que varios comentarios de la columna valoración original, cambiaron de calificación como se muestra en la columna valoración reasignada. De acuerdo a esta información, 2,941 comentarios fueron los más congruentes, en otras palabras, son opiniones que están correctamente asociadas a las calificaciones de estrellas, sin embargo, 3,091 reseñas no guardan congruencia entre la opinión del cliente y la valoración en la estrella. Por lo tanto, en la

Tabla 5.4 se muestran algunos ejemplos de estos comentarios para observar que etiquetas se les asignó, pero antes se describe cada campo.

- 1) No. Número del comentario.
- 2) Comentario. Reseña de algún producto en particular.
- 3) Estrellas. Calificación que dan al comentario.
- 4) Artículo. Nombre del producto.
- 5) Predicción. Valor predicho que asigna el algoritmo de Máquina de Soporte Vectorial.

Tabla 5.4 Comentarios etiquetados por el algoritmo (prueba 1).

No	Comentario	Estrellas	Artículo	Predicción
1	Son totalmente perfectos, me encantaron. Vienen con sus gomitas de repuesto chicas, medianas y grandes y lo mejor es que la propia marca te ofrece 2 años de garantía, sin necesidad de comprar el que te ofrece Amazon. Ojo, cualquier garantía chequen la, tomen en cuenta que no entra por descuido tuyo.	5	Auricular Inteligente	5
2	Si vale la pena, me gustó mucho, las funciones aunque no muy precisas (llega a contar más pasos o algunos movimientos los considera como pasos) si te ayudan a ver un panorama de tu día a día y llevar un mejor control de tus rutinas	4	Reloj Inteligente	4
3	Pésimo, mancha horrible y lo peor que no se acepta para devolución ❖	1	Labial	1
4	Satisfecha con mi compra	4	Reloj Inteligente	5
5	"Si hablamos del producto Alexa, es bueno como asistente. No he podido dominarla 100% ya que no me deja seleccionar días en específico simplemente es toda la semana o solo el fin de semana. (Para alarmas y rutinas) Es confuzo como enlazar amazon music aunque e supone que es en automatico el enlace"	5	Bocina Inteligente	2
6	Después de 2 meses de uso un audífono dejó de funcionar ❖	2	Auricular Inteligente	1

7	Es mentira el echo de que es impermeable el agua le traspasa sin ninguna resistencia	3	Mochila	4
8	"Está exageradamente sencillo, los materiales son buenos. La pantalla no es HD, tiene muy pocas funciones útiles, la interfaz no se puede modificar, y está muy complicado insertar tanto el SIM como una micro SD"	2	Reloj Inteligente	4
9	La parte de los barrotos es inservible. Se desperdicia espacio	4	Escritorio	3

A continuación, se discute la Tabla 5.4. Las opiniones 1-3 son comentarios que mantienen una concordancia con la evaluación que los clientes ceden, además, puede apreciarse que las etiquetas asignadas por el algoritmo son las mismas. Para las opiniones 4-6 los comentarios son discordantes, por lo tanto, el algoritmo propone una nueva valoración de forma correcta. Finalmente las opiniones 7-9 se observa que las reasignaciones propuestas por el algoritmo no son las correctas, debido a que se toman las valoraciones como etiquetas para entrenar al algoritmo, pero al no haber sido analizadas las reseñas por etiquetadores humanos muchos de estos comentarios no guardan congruencia con la valoración, y por tal motivo, también existe la posibilidad que el algoritmo se equivoque.

5.5.2 Segunda prueba con 2-gramas

Considera dos campos del corpus de comentarios:

- El primero corresponde con los comentarios procesados, los cuales ingresan a TF-IDF como información de entrada en forma de 2-gramas.
- El segundo corresponde con las valoraciones de las opiniones.

Los datos devueltos por TF-IDF se ingresan a los algoritmos de aprendizaje supervisado y las valoraciones de los comentarios como etiquetas. El tiempo de ejecución fue de 6 minutos con 32 segundos. A continuación, se muestran los resultados en la Tabla 5.5.

Tabla 5.5 Resultados del conjunto de pruebas.

Algoritmos utilizados	Estrellas	Métricas de clasificación		
		precisión	exhaustividad	valor-F
Bosques Aleatorios	1	0.52	0.65	0.58
	2	0.39	0.26	0.31
	3	0.36	0.25	0.30
	4	0.43	0.30	0.35
	5	0.47	0.80	0.60
Regresión Logística	1	0.57	0.60	0.59
	2	0.39	0.40	0.40
	3	0.33	0.28	0.30
	4	0.43	0.41	0.42
	5	0.62	0.67	0.64
Naive Bayes	1	0.58	0.61	0.59
	2	0.40	0.39	0.39
	3	0.33	0.30	0.31
	4	0.42	0.45	0.44
	5	0.65	0.64	0.65
Máquina de Soporte Vectorial	1	0.60	0.59	0.59
	2	0.38	0.41	0.39
	3	0.32	0.34	0.33
	4	0.44	0.39	0.41
	5	0.66	0.65	0.65

Los resultados mostrados en la Tabla 5.5 reportan que Máquina de Soporte Vectorial es ligeramente mejor que los demás algoritmos. Por lo tanto, para esta prueba se presenta la Tabla 5.6 con las valoraciones originales del corpus de opiniones y las reasignaciones propuestas, donde se considera únicamente el 25 por ciento de la información que comprende la parte de pruebas del algoritmo con mejores resultados:

Tabla 5.6 Comparación de evaluaciones originales y reasignadas (prueba 2).

Estrellas	Cantidad de comentarios (valoración original)	Cantidad de comentarios (valoración reasignada)
5	1,207	1,188
4	1,207	1,085
3	1,206	1,284
2	1,206	1,284
1	1,206	1,191

Al igual que en la anterior prueba también puede apreciarse en la Tabla 5.6 que muchos de los comentarios se han movido respecto de su valoración original. De acuerdo a esta información, 2,867 comentarios fueron los más congruentes, en otras palabras, son opiniones que están correctamente asociadas a las calificaciones de estrellas, sin embargo, 3,165 reseñas no guardan congruencia entre la opinión del cliente y la valoración en la estrella. En la Tabla 5.7 se presentan ejemplos para observar las calificaciones propuestas por el algoritmo.

Tabla 5.7 Comentarios etiquetados por el algoritmo (prueba 2).

No	Comentario	Estrellas	Artículo	Predicción
1	Me gustó que es un reloj muy amigable, fácil de programar y de buen tamaño.	5	Reloj Inteligente	5
2	Por algún motivo estoy teniendo archivos que no son fluidos al 100% según las especificaciones de ambos productos el dron dónde estoy grabando y esta tarjeta deberían de ser compatibles y suficientes sin embargo no ha sido el caso Seguiré intentando a ver si es cuestión de configuración pero hasta el momento no	3	Micro SD	3
3	La recibí y muy mala nunca marco mi peso, siempre marco 14 kg y mi esposa 10 kg obvio no tenemos ese peso	1	Báscula Corporal	1
4	Excelente la verdad estoy muy agusto con la compra	4	Reloj Inteligente	5

5	"Mala inversión, un rimel muy caro y mis pestañas si bajaron, sólo es publicidad. Me quedo con Prosa, económico y de mejor calidad. No gasten su dinero. En mi pic, el comparativo Prosa vs Sky High"	3	Rímel	1
6	Después de unos meses, dejó de funcionar.	2	Báscula Corporal	1
7	No es lo que esperaba	1	Mesa	2
8	Buen material, buenos acabados y llego en un tiempo inimaginable de un día para otro!! La compre un domingo y el día lunes ya la tenía en mi casa!!!	5	Mesa	4
9	"Las sillas estan padres, pero para armar es un gran esfuerzo y la calidad no es la esperada. Ya que me llegaron el lunes, y al apretarlas ya se rompieron 2 de 4 sillas... PESIMO ARTICULO...."	1	Silla	3

La información presentada en la Tabla 5.7 puede observarse que las opiniones 1-3 si presentan concordancia con la calificación que el cliente da al producto y además, las etiquetas propuestas por el algoritmo resultan las mismas. Las opiniones 4-6 proponen una nueva valoración a estos comentarios que no son congruentes con su calificación original. Sin embargo, los comentarios 7-9 manejan una congruencia entre la opinión y la valoración original, pero no así, con las etiquetas que les asigna el algoritmo.

5.5.3 Comparación de las pruebas

Como se ha mostrado se puede observar que en ambas pruebas los resultados son muy semejantes. Además, se demostró que existen comentarios que no guardan congruencia con la valoración y son reclasificados correctamente por el algoritmo de Máquina de Soporte Vectorial. Para concluir este primer experimento, se elige la primera prueba, donde se emplea 1-grama ya que es ligeramente mejor que la segunda que utiliza 2-gramas en cuanto al desempeño de las métricas.

5.6 Experimento 2: Algoritmo Basado en Reglas

Los dos experimentos realizados anteriormente demuestran la existencia de comentarios que son discordantes con la opinión del producto y la valoración que da un cliente en una escala de 1 a 5 estrellas. A diferencia del anterior experimento, este implementa un algoritmo Heurístico que requiere un conocimiento del dominio del problema para generar las reglas. En otras palabras, la función de estas reglas es detectar las palabras que tengan polaridades positivas y negativas (en el capítulo 4 Desarrollo del método de solución, sección 4.3 Algoritmo Basado en Reglas, se discutió su funcionamiento). Así mismo, en una tabla se presentarán las valoraciones originales y las reasignadas por el algoritmo por medio de las métricas de clasificación.

De la misma forma, se considera el mismo conjunto de datos, es decir, se toman 24,125 opiniones del corpus de comentarios de los cuales por cada clase de estrellas corresponden 4,825 reseñas. Esta información se divide en 75 por ciento el restante que corresponde al 25 por ciento se utiliza para generar las métricas de clasificación.

De los campos del corpus solamente se consideran los comentarios procesados. Como se había mencionado anteriormente, este experimento no requiere el uso de TF-IDF, debido a que emplea un léxico afectivo en su lugar. Con esta información es posible generar una matriz de confusión donde se comparen las calificaciones que originalmente otorgaron los clientes que adquirieron un producto contra las que propone el algoritmo. El tiempo de ejecución fue de 46 segundos. A continuación, se presenta la Tabla 5.8 donde se muestran estos resultados.

Tabla 5.8 Resultados del conjunto de pruebas.

Algoritmo utilizado	Estrellas	Métricas de clasificación		
		precisión	exhaustividad	valor-F
Algoritmo Heurístico o Basado en Reglas (propio)	0	0	0	0
	1	0.39	0.43	0.41
	2	0.39	0.32	0.35
	3	0.22	0.29	0.25
	4	0.24	0.31	0.27
	5	0.66	0.57	0.61

Como puede observarse en la Tabla 5.8 en la columna *estrellas* se anexa una clase adicional (la clase cero). Originalmente las clases de comentarios se clasifican en una escala del 1 a 5, sin

embargo, el algoritmo detecta aquellos comentarios carentes de subjetividad. De acuerdo a Fang X., & Zhan J. (2015, pág. 13) “Una oración de sentimiento es aquella que contiene al menos una palabra positiva o negativa”. Por consiguiente, al no existir estos elementos en algunos comentarios (palabras con polaridad), estos son clasificados en una nueva clase. Por otro lado, cuando se obtienen las métricas de evaluación anteriormente mencionadas para esta nueva clase (clase cero), los valores son cero debido a que en el corpus de comentarios no existen estas calificaciones. A continuación, se muestra la Tabla 5.9 donde se comparan las valoraciones originales del corpus de opiniones y las reasignaciones propuestas por este algoritmo donde se considera únicamente el 25 por ciento de la información que comprende la parte de pruebas.

Tabla 5.9 Comparación de evaluaciones originales y reasignadas.

Estrellas	Cantidad de comentarios (valoración original)	Cantidad de comentarios (valoración reasignada)
5	1,207	1,436
4	1,207	904
3	1,206	901
2	1,206	1,502
1	1,206	1,065
0	-	224

En la Tabla 5.9 la cantidad de comentarios evaluados originalmente tuvieron una mayor reasignación de estrellas a diferencia del experimento 1. De acuerdo a esta información, 2,297 comentarios fueron los más congruentes, en otras palabras, son opiniones que están correctamente asociadas a las calificaciones de estrellas, sin embargo, 3,735 reseñas, no guardan congruencia entre la opinión del cliente y la valoración en la estrella. En la siguiente Tabla 5.10 se muestran ejemplos para poder observar estas diferencias pero antes se describe cada campo.

- 1) No. Número del comentario.
- 2) Comentario. Reseña de algún producto en particular.
- 3) Estrellas. Calificación que dan al comentario.
- 4) Artículo. Nombre del producto.

5) Reasignación. Valor de estrellas reasignadas por el Algoritmo Basado en Reglas.

Tabla 5.10 Comentarios etiquetados por el algoritmo.

No	Comentario	Estrellas	Artículo	Reasignación
1	Muy buena compra	5	Reloj Inteligente	5
2	Está incómodo, se escucha muy bajo muy caro para la basura que es	1	Auricular Inteligente	1
3	En general están bastante bien por el precio, son básicamente unas sillas de plástico más estéticas que las comunes, si no las vas a ocupar para algo realmente exigente, están bastante bien, el instructivo no sirve de mucho pero hay varios vídeos en YT, si estás pensando en pedir las no dudes son bastante buenas en cuanto al precio	4	Silla	4
4	Muy buena relación calidad precio, si lo recomiendo	4	Báscula Corporal	5
5	No detecta bien muchas funciones con Alexa y no te deja descargar varias aplicaciones y luego con la caja un poco dañada	4	Dispositivo de Streaming	2
6	No llevaba ni una semana y dos de los cierres fallaron.	3	Mochila	1
7	Buen producto, no costó trabajo conectarlo. Llegó en la fecha acordada y a muy buen precio	5	Dispositivo de Streaming	4
8	ya tengo mucho tiempo con ella y aun funciona perfectamente no he tenido ninguna clase de falla o deterioro, super recomendada	5	Micro SD	3
9	Deja la piel con buena textura, un olor agradable y se ve natural.	5	Maquillaje	4
10	La mía llegó en idioma Inglés y chino como la cambie a español	2	Reloj Inteligente	0
11	Pasable	3	Maletín	0
12	Más o menos	3	Mesa	0

De acuerdo a las opiniones 1-3 de la Tabla 5.10 son comentarios congruentes con la calificación que reciben y por supuesto el algoritmo las clasifica con la misma valoración. Para las opiniones 4-6 de la Tabla 5.10 no manejan congruencia con las evaluaciones, pero el algoritmo las reasigna correctamente. Los comentarios 7-9 de la Tabla 5.10 a pesar de presentar una correcta valoración por parte del cliente el algoritmo se equivoca, ya que les reasigna una distinta. Finalmente los comentarios 10-12 de la Tabla 5.10 son carentes de subjetividad y a diferencia del primer experimento, una porción de estos comentarios se mueve a la clase cero.

5.7 Experimento 3: Comparación del algoritmo propuesto con algoritmos de aprendizaje automático

Una vez demostrada la existencia de comentarios que no guardan congruencia con la valoración en la estrella, se lleva a cabo un tercer experimento con el propósito de comparar los anteriores modificando unos cambios y por medio de tablas comparar los resultados y elegir el mejor algoritmo.

Este experimento lleva a cabo los siguientes cuatro puntos:

- 1) Se ejecuta nuevamente el experimento 1, pero tomando en cuenta el algoritmo de Máquina de Soporte Vectorial con 1-grama el cual fue ligeramente mejor que los demás.
- 2) Se ejecuta de igual forma el experimento 2.
- 3) Como el Algoritmo Basado en Reglas detecta aquellas opiniones que son carentes de subjetividad (886 opiniones de comentarios objetivos), estos se eliminan del corpus quedando un total de 23,239 reseñas y la distribución por clase de comentarios se muestra en la Tabla 5.11.

Tabla 5.11 Nueva distribución por clase de comentarios.

Estrellas	Comentarios procesados
5	4679
4	4694
3	4631
2	4659

1	4576
---	------

Puede apreciarse en la Tabla 5.11 que a pesar de haberse eliminado algunos comentarios, se presenta un desbalance muy tenue, lo cual no debería afectar los resultados.

- 4) Con los campos de la Tabla 5.11 es posible comparar exactamente los mismos datos para los experimentos 1 y 2. Finalmente se toman opiniones que no fueron vistas por el algoritmo de Máquina de Soporte Vectorial y los resultados de este serán comparados con los resultados del Algoritmo Basado en Reglas.

Para el entrenamiento y las pruebas en el caso del algoritmo de Máquina de Soporte Vectorial los datos se dividieron en 75 y 25 por ciento que corresponden a 17,429 y 5,810 comentarios. De igual forma esta información se aplica al Algoritmo Basado en Reglas con el propósito de comparar el 25 por ciento. A continuación, se presentan los resultados en la Tabla 5.12 y

Tabla 5.13 que cubren los puntos 1, 2 y 3.

Tabla 5.12 Resultados del experimento 1.

Algoritmo utilizado	Estrellas	Métricas de clasificación		
		precisión	exhaustividad	valor-F
Máquina de Soporte Vectorial	1	0.56	0.63	0.60
	2	0.39	0.38	0.38
	3	0.35	0.35	0.35
	4	0.46	0.40	0.43
	5	0.67	0.69	0.68

Tabla 5.13 Resultados del experimento 2.

Algoritmo utilizado	Estrellas	Métricas de clasificación		
		precisión	exhaustividad	valor-F
Algoritmo Heurístico o Basado en Reglas (propio)	1	0.42	0.43	0.43
	2	0.40	0.31	0.35
	3	0.23	0.29	0.25
	4	0.26	0.35	0.30
	5	0.70	0.57	0.63

Como resultado en la comparación de la Tabla 5.12 y

Tabla 5.13 se puede observar que Máquina de Soporte Vectorial sigue siendo ligeramente mejor. Hasta este punto se podría concluir que de elegir un algoritmo, sería este, sin embargo, esto no es del todo cierto ya que Máquina de Soporte Vectorial por medio de TF-IDF recibe la representación del texto, es decir, a las palabras se les asigna un peso de acuerdo a su importancia. Con estos pesos el algoritmo entrena y aprende a clasificar los comentarios con sus correspondientes etiquetas y por medio de las pruebas junto con las métricas se puede observar el desempeño, sin embargo, no todos los comentarios tienen una correcta clasificación. En cambio, el Algoritmo Basado en Reglas considera otros factores para llevar a cabo la clasificación. Precisamente el punto número 4, enseguida explicado, demuestra la diferencia de los resultados entre estos dos algoritmos. A continuación, se muestra una pequeña lista en forma de tabla con los comentarios que el Algoritmo Basado en Reglas clasificó como opiniones carentes de subjetividad y por supuesto que predicciones le da el algoritmo de Máquina de Soporte Vectorial pero antes se describe cada campo de la Tabla 5.14.

- 1) No. Número del comentario.
- 2) Comentario. Reseña de algún producto en particular.
- 3) Estrellas. Calificación que dan al comentario.
- 4) Artículo. Nombre del producto.
- 5) Predicción. Valor predicho que asigna el algoritmo de Máquina de Soporte Vectorial.

Tabla 5.14 Opiniones carentes de subjetividad.

No.	Comentario	Estrellas	Artículo	Predicción
1	Se hace como una pasta con este color	1	Labial	3
2	la mesa de trabajo es muy chiquita	1	Escritorio	4
3	El tono seleccionado en la imagen se ve rojo quemado pero en físico es rosa.	2	Labial	3
4	Porque al momento de ponerla en mi Switch me dice que tengo espacio de 366 GB si se supone que la memoria es de 400?	2	Micro SD	5
5	Mi Mac se volvió lenta	3	NVMe	2

6	Armarla fue toda una travesía.	3	Silla	4
7	Necesito factura urgente	4	Silla	3
8	Ese producto lo uso habitualmente! Lo compre en AMAZON porque es más barato ❖	4	Rímel	1
9	A PRUEBA DE SUDOR	5	Maquillaje	5
10	Los compré para mi hermanito	5	Auricular Inteligente	1

De acuerdo con la información de la Tabla 5.14 en el campo de los comentarios no se manejan palabras que se identifiquen como un sentimiento o una polaridad. Del mismo modo no hay una congruencia con las valoraciones que ceden a los comentarios ni tampoco las etiquetas que clasifica el algoritmo.

5.8 Reasignación de estrellas

De acuerdo con el objetivo general, se crea un método automático para reducir la asimetría entre la polaridad de las reseñas y las valoraciones numéricas dadas por los clientes de productos de la plataforma web de Amazon México utilizando la minería de opiniones. Por lo tanto, a continuación se presenta la Tabla 5.15 con comentarios que manejan polaridad, donde se discute como se lleva a cabo la reasignación de estrellas, pero antes se describe cada campo de la Tabla 5.15.

- 1) No. Número del comentario.
- 2) Comentario. Reseña de algún producto en particular.
- 3) Estrellas. Calificación que dan al comentario.
- 4) Artículo. Nombre del producto.
- 5) Predicción. Valor predicho que asigna el algoritmo de Máquina de Soporte Vectorial.
- 6) Reasignación. Valor de estrellas reasignadas por el Algoritmo Basado en Reglas.

Tabla 5.15 Opiniones subjetivas.

No.	Comentario	Estrellas	Artículo	Predicción	Reasignación
1	La silla es muy incómoda, la esponja del asiento es muy delgada. Solo la utilice por 3 semanas, ya	1	Silla	2	1

	que se rompieron los dos descansabrazos. No la recomiendo.				
2	La mochila: bien, resistente, buen material, grande, amplia, buena cantidad de bolsas y cierres. Lo malo: No traía los cables para audio y USB. No es vital, pero sí debe traerlos.	1	Mochila	4	4
3	Casi no pinta. Hay que aplicar 3 o 4 capas para notar el color. No lo volvería a comprar.	2	Maquillaje	2	1
4	Decepcionado, es tamaño infantil, las medidas que anuncian(30x19x47) no son reales (30x17x42)	2	Mochila	1	1
5	Baratos y con razón, no funcionan bien en general, el volumen máximo es bajo y de mala calidad pero lo peor es que en cuanto estas cerca de otros celulares se empieza a cortar la señal y se escucha entrecortado	3	Auricular Inteligente	2	1
6	Que no funciona a la semana dejó de funcionar	3	Micro SD	1	1
7	Cumplió con mis expectativas	4	Micro SD	5	5
8	Parecen buenos, el único problema es que es un gigantesco problema la configuración...es ridículamente complicada, de ahí en fuera todo genial	4	Auricular Inteligente	4	3
9	Muy buena fidelidad de sonido	5	Auricular Inteligente	5	5
10	Son muy buenos en cuanto a calidad de sonido pero para ejercicio al no tener certificación ipXX no me da confianza	5	Auricular Inteligente	3	2

Con la información de la Tabla 5.15 pueden observarse algunos datos interesantes de los algoritmos como se describen a continuación:

- Los valores de las columnas de la *predicción* y *reasignación* que se encuentran de color azul, significan que han clasificado por igual al comentario, pero en la mayoría de los casos proponen una nueva valoración.
- Los valores de las columnas de la *predicción* y *reasignación* que se encuentran de color rojo, significan que han clasificado diferente y en la mayoría proponen una nueva valoración. A continuación, se aborda este punto precisamente para explicar estas diferencias.

- a) Comentario número 1 de la Tabla 5.15: “*La silla es muy incómoda, la esponja del asiento es muy delgada. Solo la utilice por 3 semanas, ya que se rompieron los dos descansabrazos. No la recomiendo.*”

Al ingresar al Algoritmo Basado en Reglas, la salida es:

palabras positivas: []

palabras negativas: ['incómodo', 'solo', 'romper', 'no-recomendar', 'descansar']

Se determina que la clase es 1. Puede observarse la palabra **recomendar**, la cual tiene una polaridad positiva, sin embargo, por la regla que trabaja la negación de verbos, esta cambia su polaridad a negativa.

- b) Comentario número 3 de la Tabla 5.15: “*Casi no pinta. Hay que aplicar 3 o 4 capas para notar el color. No lo volvería a comprar.*”

Al ingresar al Algoritmo Basado en Reglas, la salida es:

palabras positivas: []

palabras negativas: ['no', 'no']

Se determina que la clase es 1. No existen palabras positivas.

- c) Comentario número 5 de la Tabla 5.15: *“Baratos y con razón, no funcionan bien en general, el volumen máximo es bajo y de mala calidad pero lo peor es que en cuanto estas cerca de otros celulares se empieza a cortar la señal y se escucha entrecortado”*

Al ingresar al Algoritmo Basado en Reglas, la salida es:

palabras positivas: []

palabras negativas: ['malo', 'peor', 'cortar', 'calidad', 'no-bien']

Se determina que la clase es 1. Pueden observarse las palabras *calidad* y *bien*, las cuales tienen una polaridad positiva, sin embargo, por las reglas que trabajan la negación, estas cambian su polaridad a negativa.

- d) Comentario número 8 de la Tabla 5.15: *“Parecen buenos, el único problema es que es un gigantesco problema la configuración...es ridículamente complicada, de ahí en fuera todo genial”*

Al ingresar al Algoritmo Basado en Reglas, la salida es:

palabras positivas: ['bueno', 'único', 'genial']

palabras negativas: ['problema', 'problema', 'complicar']

Al aplicar la base matemática se determina que la clase es 3. Tienen el mismo número de palabras positivas como negativas.

- e) Comentario número 10 de la Tabla 5.15: *“Son muy buenos en cuanto a calidad de sonido pero para ejercicio al no tener certificación ipXX no me da confianza”*

Al ingresar al Algoritmo Basado en Reglas, la salida es:

palabras positivas: ['bueno', 'calidad', 'ejercicio']

palabras negativas: ['no', 'dar', 'no-confianza']

Al aplicar la base matemática se determina que la clase es 2. Tiene una carga mayor de palabras negativas que positivas.

Capítulo 6 Interfaz web del proyecto de tesis

Se implementó una interfaz web (con el propósito de observar los resultados del proyecto de investigación de una forma más simple) basada en el framework Django, que permite el desarrollo de sitios web seguros y mantenibles. Además, entre las ventajas se destaca principalmente que puede ser ejecutado en cualquier sistema operativo. A continuación, se describe la interfaz web, así como su funcionamiento:

La página principal hace énfasis con el concepto del análisis de sentimientos, posteriormente describe el trabajo de esta investigación como se muestra en la Figura 6.1:



Figura 6.1 Página principal.

En el enlace de *detalles* de la Figura 6.1, se explica como el algoritmo puede ser ejecutado mediante la *Url Absoluta* o *Url Relativa* de acuerdo con la Figura 6.2:



Figura 6.2 Enlace de detalles.

Los enlaces tanto de la *Url Absoluta* como la *Url Relativa* de la Figura 6.2, recibe 5 parámetros a continuación explicados.

- **URL:** Dirección web de Amazon que hace referencia hacia algún producto. Para el caso de la *Url Absoluta*, es suficiente con elegir el producto deseado, copiar la url como se muestra en la Figura 6.3.



Figura 6.3 Url absoluta.

En el caso de la *Url Relativa* de la Figura 6.2 una vez seleccionado los comentarios del producto, es posible filtrarlos por alguna estrella en particular e incluso haber hecho paginaciones de las opiniones y, en el punto deseado copiar la url como se observa de la Figura 6.4 a la Figura 6.6.



Figura 6.4 Producto seleccionado de Amazon.

Como se puede apreciar en la Figura 6.4 es necesario seleccionar un producto.



Figura 6.5 Ver todos los comentarios.

Después de haber seleccionado el producto se hace un desplazamiento hacia abajo, hasta llegar a la parte que dice *Ver todas las opiniones* como se muestra en la Figura 6.5 posteriormente dar clic a ese enlace.



Figura 6.6 Opiniones de clientes.

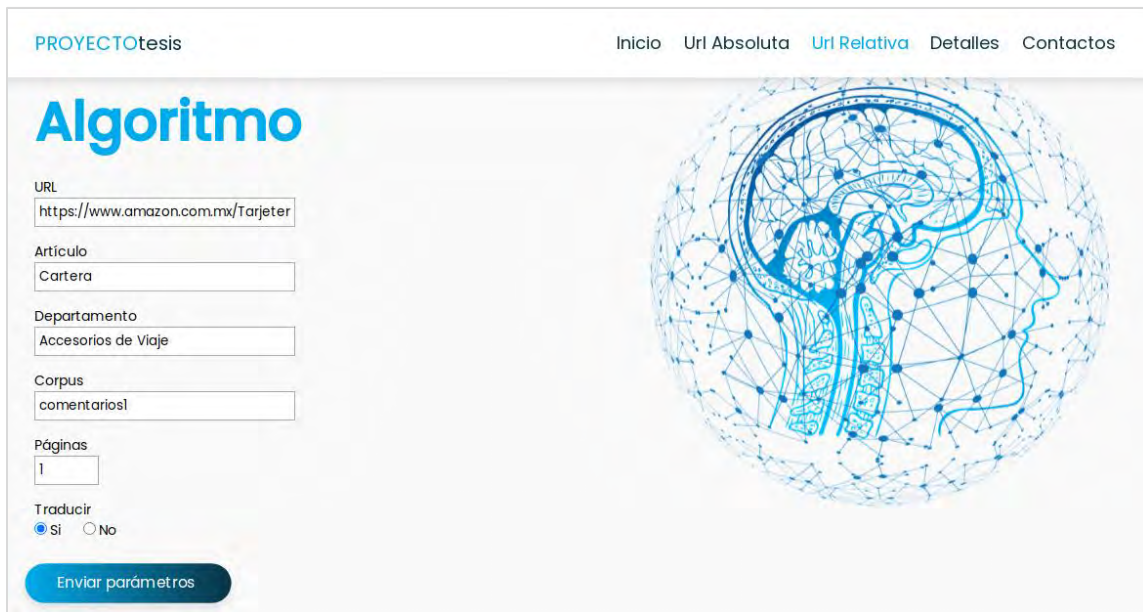
En la Figura 6.6 de lado izquierdo, se pueden observar las calificaciones que los clientes ceden al producto. Hasta este punto es posible hacer clic derecho en alguna estrella y copiar la url. Este resultado corresponde a una url relativa ya que de esta manera es posible filtrar las valoraciones por estrella e incluso hacer paginaciones, y posteriormente copiar la url.

Nota: No se puede repetir la misma url, para tal caso, es necesario borrar la información que está en la carpeta *ProyectoTesis/ProyectoTesisApp/CorpusIdiomas* + los archivos que se quieran descargar nuevamente. Hasta aquí, se finaliza este punto, y se continua con los siguientes parámetros que recibe la aplicación web.

- **Artículo:** Nombre del producto seleccionado.
- **Departamento:** Nombre del departamento al que pertenece el producto.
- **Corpus:** Nombre donde se almacenarán todos los comentarios extraídos de la página web Amazon México. Nota. El nombre no puede ser el mismo para cada ejecución del algoritmo, para tal caso, debe eliminarse el archivo que está en la carpeta *ProyectoTesis/ProyectoTesisApp/CorpusTraducidos* + el nombre del archivo a eliminar.
- **Páginas:** Representa las paginaciones que se deseen recorrer de un producto Amazon. Recibe un intervalo de 1 como mínimo y 5 como máximo. Tiene el valor de 1 por defecto.

- **Traducir:** Al descargar comentarios en Amazon, es posible encontrar algunos que están en otros idiomas. Para tal caso, activar esta opción, permite traducir opiniones al idioma español, sin importar que estén escritos en algún otro lenguaje.

Los parámetros explicados deben pasarse a cualquiera de los dos enlaces, dependiendo por supuesto, si se trata de una url absoluta o relativa. Como muestra la Figura 6.7.



The screenshot shows a web form titled 'Algoritmo' on the 'PROYECTOtesis' website. The navigation menu includes 'Inicio', 'Url Absoluta', 'Url Relativa', 'Detalles', and 'Contactos'. The form contains the following fields and options:

- URL:
- Artículo:
- Departamento:
- Corpus:
- Páginas:
- Traducir: Sí No
- Enviar parámetros:

The background features a stylized blue brain graphic composed of a network of nodes and lines.

Figura 6.7 Parámetros del algoritmo.

NOTA: Al enviar los parámetros anteriormente mencionados se redirigirá a una página con el nombre “*procesar_informacion*”. Aquí se imprimirá si el proceso terminó correctamente o hubo un error como se indica en la Figura 6.8.

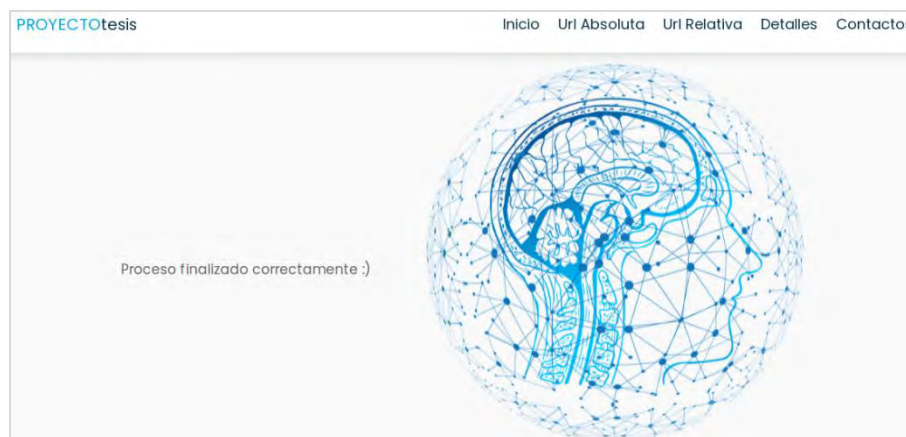


Figura 6.8 Proceso finalizado correctamente.

Además, en la carpeta *ProyectoTesis/ProyectoTesisApp/resultados* se creará un archivo con la extensión “csv” con el mismo nombre del corpus, pero este tendrá la reasignación de las evaluaciones de los comentarios que no guardan congruencia. Caso contrario tendrán la misma valoración. Para poder corroborar esta información es necesario ir a la carpeta de resultados y hacer clic en el archivo deseado como se muestra en la Figura 6.9 y Figura 6.10.

identificador	usuario	pais	idioma	fecha	titulo	estrellas	comentario
1	KNV	México	Español	23 de julio	Bonito para	5	Fue para un r
2	Aldenhy	México	Español	14 de sept	Excelente	5	El diseño est
3	Lillian Dec	México	Español	12 de novi	Muy prácti	5	Resistentes y
4	Luis alfons	México	Español	4 de mayo	excelente	5	muy practica
5	Chuchin	México	Español	30 de ener	El diseño	5	Muy práctica
6	Bertha Dur	México	Español	19 de dicie	Relación c	5	Buena Calide
7	Ricardo	México	Español	1 de dicien	Excelente	5	Muy buena c
8	King. J	Reino Unic	Inglés	5 de julio d	Es un titul	5	Llegó a tiemp
9	Leslie	Reino Unic	Inglés	14 de julio	Gran	5	Pequeño, cor
10	A. Raymor	Reino Unic	Inglés	30 de may	Gran	5	Valor por din

Figura 6.9 Archivo de salida (primera parte).

beneficiados	articulo	departamento	lenguaje	comentario_procesado	reasignacion
0	Cartera	Accesorios de	\Spanish	ser regalo gustar mucho	5
1	Cartera	Accesorios de	\Spanish	diseño estar parisino fin_	5
0	Cartera	Accesorios de	\Spanish	resistente bueno calidad p	5
0	Cartera	Accesorios de	\Spanish	muy practicar excelente c	5
1	Cartera	Accesorios de	\Spanish	muy práctico fin_ngrama	5
0	Cartera	Accesorios de	\Spanish	bueno calidad solo no env	3
0	Cartera	Accesorios de	\Spanish	muy bueno calidad fin_ng	5
0	Cartera	Accesorios de	\Spanish	llegar tiempo bueno relaci	5
0	Cartera	Accesorios de	\Spanish	pequeño fin_ngrama com	5
0	Cartera	Accesorios de	\Spanish	valor dinero fin_ngrama	5

Figura 6.10 Archivo de salida (segunda parte).

6.1 Prueba piloto

Con los ejemplos de la Tabla 5.15 donde se demostró que en algunos comentarios el Algoritmo Basado en Reglas, propone una correcta reasignación con aquellas opiniones que no guardan congruencia con la valoración que ceden los clientes al producto. El objetivo de esta prueba piloto es obtener la calificación general dada a un producto y calcular la calificación general por el Algoritmo Basado en Reglas. Por lo tanto, se procedió a descargar un nuevo artículo

completamente de la página web de Amazon⁹ incluyendo los que estén en otros idiomas cómo se observa en la siguiente Figura 6.11.



Figura 6.11 Producto de la página Amazon.

De acuerdo con la información de la Figura 6.11 puede observarse el nombre del producto (Meguiar's DKO G18211 Cera Ultimate En Pasta) y de lado izquierdo la cantidad de las calificaciones globales, así como el porcentaje correspondiente por clase de comentarios o estrellas. Sin embargo, si se desplaza un poco el cursor hacia la parte de abajo, puede encontrarse la siguiente información de la Figura 6.12.

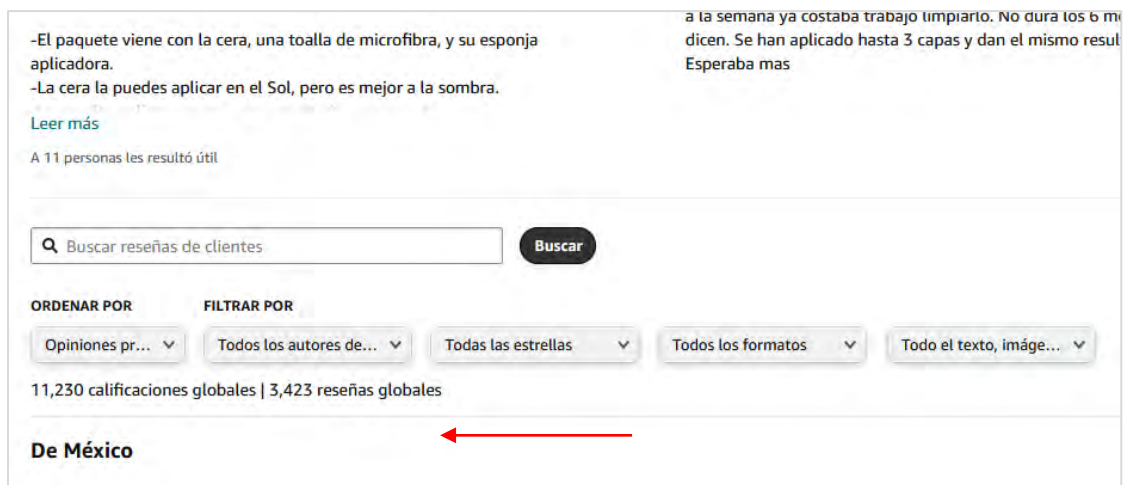


Figura 6.12 Reseñas globales.

9

https://www.amazon.com.mx/product-reviews/B004HCS4PK/ref=cm_cr_arp_d_show_all?ie=UTF8&reviewerType=all_reviews&page

La flecha de color rojo de la Figura 6.12 apunta hacia las *calificaciones globales* con un total de 11,230 pero también señala las *reseñas globales* con la cantidad de 3,423. En otras palabras, la porción de reseñas globales es la correspondiente proporción de comentarios para este producto. Una vez explicada esta información se extrajeron todos los comentarios incluyendo los que están en otros idiomas (traducidos al español). La razón por la cual se procedió a considerar todas las reseñas incluyendo las de otros idiomas, es que se requiere hacer una comparación con los porcentajes originales de este producto contra los propuestos por el Algoritmo Basado en Reglas. Por medio de la implementación de la interfaz web se procedió a ingresar como información de entrada del método de solución la dirección url de la página web de Amazon México como se muestra en la Figura 6.13.

The screenshot shows a web interface titled 'PROYECTOtesis' with a navigation menu: Inicio, **Url Absoluta**, Url Relativa, Detalles, Contactos. The main heading is 'Algoritmo'. The form contains the following fields and options:

- URL:
- Artículo:
- Departamento:
- Corpus:
- Páginas:
- Traducir: Si No
- Enviar parámetros:

On the right side of the form is a large graphic of a human brain composed of a network of blue nodes and lines, symbolizing neural networks or algorithms.

Figura 6.13 Ejecución del método de solución.

De acuerdo con la Figura 6.13 con los campos completados, se envía esta información al servidor y se ejecuta el método de solución como se describe en la siguiente Figura 6.14.

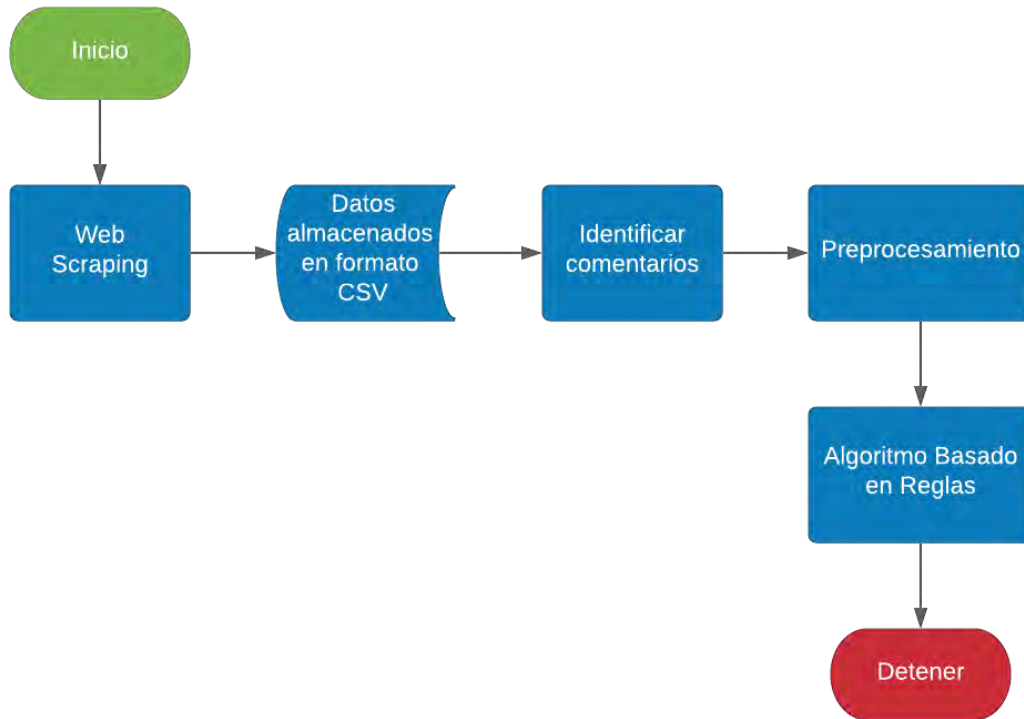


Figura 6.14 Método de solución.

La Figura 6.14 muestra los 5 módulos por los cuales debe ir pasando la ejecución de la prueba piloto. A continuación, se describe cada uno:

- **Web Scraping:** Recibe como información de entrada por parte de la aplicación web, la url que hace referencia a un producto de cera para autos de la marca Meguiar’s, que corresponde al departamento de *Cuidado de Coche y Moto*. Al finalizar la extracción de los comentarios, se descargó un total de 3,423 reseñas.
- **Datos almacenados en formato CSV:** Las opiniones descargadas se almacenan en un formato con la extensión *csv*. Hasta este punto, se ha generado el corpus de comentarios.
- **Identificar comentarios:** Este módulo se encarga de identificar las opiniones que se encuentren en idioma español, asignándoles la etiqueta *Spanish* por lo cual de las 3,423 reseñas 522 opiniones se identificaron en un idioma diferente al español, ahora se cuenta con un total de 2,901 comentarios. Para saber el porqué, véase el artículo que se presentó en la Séptima Jornada de Ciencia y Tecnología que se titula “*Generación de corpus de comentarios de Amazon en español*”.
- **Preprocesamiento:** Al ingresar el corpus de comentarios a esta fase, se realiza lo siguiente:
 - Tokenización.

- Eliminación de caracteres especiales.
- Revisión ortográfica.
- Eliminación de palabras vacías.
- Normalización del texto.
- Etiquetado post.

Nota: 33 reseñas quedaron eliminadas debido a que en el cuerpo de las diferentes opiniones, carecían de características como verbos, adverbios, adjetivos y sustantivos de acuerdo al estado del arte, quedando un total de 2,868 opiniones. Por lo tanto, en la Tabla 6.1 se describe la cantidad real de comentarios, después de haber pasado por el módulo de preprocesamiento. Esta tabla está formada por el país al que pertenece cada comentario y por supuesto la cantidad de opiniones:

Tabla 6.1 Cantidad de comentarios por país.

País	Cantidad
Alemania	43
Australia	1
Canadá	14
España	40
Estados Unidos	2,332
Francia	2
India	2
Italia	38
México	126
Reino Unido	270

- **Algoritmo Basado en Reglas:** Ahora se procede a enviar la información preprocesada al Algoritmo Basado en Reglas y finalmente se presenta la Figura 6.15 donde se podrá

observar las proporciones originales por tipo de comentarios y en la Figura 6.16 las reasignadas por el algoritmo:



Figura 6.15 Opiniones de clientes (originales).



Figura 6.16 Opiniones de clientes (reasignadas).

De acuerdo con la información de la Figura 6.15 estas son las calificaciones que originalmente los clientes cedieron al producto, para este caso, tiene una calificación general 4.5 de 5 estrellas (véase el siguiente enlace de Amazon donde se explica cómo obtienen la evaluación general¹⁰). En el caso de la Figura 6.16 puede observar una etiqueta que se titula **0 estrellas**. Esto debido a que el Algoritmo Basado en Reglas también detecta aquellos comentarios que son carentes de

¹⁰ https://sellercentral.amazon.com.mx/gp/help/external/G761?language=es_MX

subjetividad, por lo tanto, las porciones se reparten en 6 estrellas y el cálculo general se ve penalizado y como resultado se obtiene 4.5 de 6 estrellas. De acuerdo a esta información 1,483 comentarios fueron los más congruentes en otras palabras, son opiniones que están correctamente asociadas a las calificaciones de estrellas, sin embargo, 1,385 reseñas no guardan congruencia entre la opinión del cliente y la valoración en la estrella. Por lo tanto, en la Tabla 6.2 Opiniones objetivas. puede observarse algunos ejemplos de comentarios que se identificaron como opiniones objetivas:

Tabla 6.2 Opiniones objetivas.

Calificación original	Comentario	Calificación reasignada
5	Para encerar mi auto	0
5	la colección definitiva de meguiars es épica	0
5	¡Es una lata de cera de pasta de coche! Tendrá que ver cómo serán los resultados a finales de esta semana.	0
5	Gran	0
4	Por encima del promedio	0
5	Gran cosa.	0
4	Como se anuncia	0
5	Funciona como se esperaba	0
4	hace su trabajo	0
5	Siempre hace que mi coche parezca una exhibición de sala de exposición.	0

Conclusiones

Se observó aproximadamente entre un 30% a 40% de discordancia en los comentarios, por tal motivo, se implementó una estrategia para reasignar la calificación dada por estrellas. Además, para llevar a cabo este procedimiento se desarrolló un algoritmo para la generación de un corpus de comentarios basado en la plataforma de Amazon México. El corpus de comentarios cuenta con 1,005,364 reseñas de los productos de accesorios de viaje, belleza, electrónicos y muebles de hogar de las cuales, una parte comprende las opiniones que están escritas en idioma español, y otras se derivan de la traducción de idiomas como Alemania, Australia, Brasil, Canadá, España, Estados Unidos, Francia, India, Italia, Japón, Países Bajos, Polonia y Reino Unido. Asimismo, un comentario comprende los campos de usuario, país, idioma, fecha, título, estrellas, comentario, beneficiados, artículo, departamento y lenguaje.

Como se mencionó en el capítulo 5 de pruebas y resultados, no se utilizaron evaluadores ya que en el corpus de comentarios los clientes que han adquirido un producto otorgan una valoración que se sitúa en una escala de 1 a 5 estrellas. Por tal motivo, estas calificaciones se tomaron como las etiquetas de las opiniones para utilizarlas en los experimentos, sin embargo, en las métricas de evaluación que se utilizaron no se tuvo un buen desempeño, esto debido a que en esta investigación se observó una gran cantidad de comentarios analizados manualmente que no guardan congruencia con la valoración asignada y por tal motivo el haber implementado etiquetadores humanos que reclasificaran las valoraciones de los comentarios discordantes, las métricas tendría un mejor desempeño.

Por otra parte, en los comentarios existe una gran cantidad de faltas ortográficas que pueden resolverse con librerías ya definidas para este propósito, pero que en algunos casos las palabras son irreconocibles, por lo tanto, este tipo de algoritmos corrigen de forma errónea. Por tal motivo, no es una tarea sencilla trabajar con esta clase de comentarios y reasignarles una correcta valoración.

Otro punto importante a tomar en cuenta es que no existen reglas en la plataforma de Amazon México que ayuden a los clientes que han adquirido un producto y que de acuerdo a su experiencia obtenida puedan otorgar una correcta calificación en la estrella, ya que existe una cantidad importante de comentarios que aunque son muy cortos son iguales, pero descritos por un cliente

diferente y que pueden ceder diferentes estrellas. Por esta razón fue necesario hacer un estudio manualmente de las opiniones y por medio de los tres cortes de la fórmula matemática se consiguió definirlos en alguna estrella en particular, ya que en la mayoría del estado del arte solo trabajan con tres tipos de polaridades: positivas, negativas y neutrales.

Esto añadió como desafío el desarrollo de un algoritmo Basado en Reglas, por los siguientes motivos:

- Los trabajos que abordan el análisis de sentimientos en idioma español, en su gran mayoría, solo trabajan con polaridades positivas, negativas y neutrales.
- Otros trabajos igualmente relacionados con el idioma español, utilizan herramientas como Vader o TextBlob haciendo uso de la traducción del español al inglés, ya que estas librerías no trabajan en este idioma. Igualmente solo trabajan con polaridades positivas, negativas, neutrales y en el caso de TextBlob la subjetividad.
- Los trabajos que implementan los cinco niveles de polaridad que son las estrellas que los clientes asignan a un producto usan un enfoque tradicional, es decir, utilizan etiquetadores humanos para posteriormente entrenar con algoritmos de aprendizaje supervisado y obtener resultados.
- Por parte de los trabajos en idioma inglés utilizan la librería SentiWordNet y diferentes fórmulas matemáticas que les permiten manejar cinco tipos de cortes para asignar una valoración a los comentarios. De igual manera etiquetan manualmente los comentarios para entrenarlos posteriormente con algoritmos de aprendizaje supervisado.

Por tal motivo, se decidió proponer un algoritmo de este tipo, el cual con el conocimiento acerca del dominio del problema se generaron 17 reglas. Estas aplican heurística para determinar la polaridad de un texto basado en las frecuencias de la polaridad de palabras, donde estas frecuencias se determinan con la combinación de n-gramas, para este caso se utilizaron, 2-gramas, 3-gramas y 4-gramas con la ayuda de un léxico afectivo. Asimismo, no requiere tener previamente las valoraciones o calificaciones de un producto analizadas por un etiquetador, ya que tiene la capacidad de asignar una valoración que va en una escala de 1 a 5 estrellas. También puede detectar aquellos comentarios carentes de subjetividad e incluso trabajar con un corpus de comentarios de otras plataformas (parecidas a Amazon) siempre y cuando se valoren comentarios con 5 tipos de estrellas.

Trabajos futuros

El Algoritmo Basado en Reglas puede ser mejorado y eso conlleva a tener un mejor desempeño. A continuación, se describen los puntos que pueden ser atendidos:

- Se hizo un estudio para generar las reglas, sin embargo, el mayor potencial de estas, se basa en los 2-gramas y 3-gramas, para los 4-gramas solo se aplican dos reglas, debido a la complejidad que hay entre las características de las palabras, es decir, cuando se deben activar y cuando no. Por lo tanto, el estudio de los 4-gramas sería interesante profundizarlo.
- El léxico afectivo que se utiliza, podría complementarse con otros, ya que el lenguaje es muy extenso y existen comentarios donde la reasignación propuesta por el algoritmo no es la mejor y cabe la posibilidad que en el léxico no se estén considerando algunas palabras que sí manejan una polaridad positiva o negativa.
- Existen comentarios que manejan palabras que al ser leídas por un humano se interpretan con polaridad, pero que sin embargo no la tienen, por ejemplo: *más*, *menos*, *funcionar* y *esperaba*, aparecen en opiniones como:
 - “*El producto de **funcionar***”.
 - “*Es lo que **esperaba***”.
 - “***Esperaba** más del producto*”.
 - “***Más o menos***”.

Como se puede apreciar, ninguna de estas palabras representa una polaridad pero su estudio posiblemente podría considerarlo, ya que el algoritmo detecta este tipo de reseñas como carentes de subjetividad.

- El algoritmo genera una incorrecta reasignación en la estrella, en comentarios donde la experiencia del producto fue pésima, y posteriormente comienzan a narrar una excelente experiencia con un artículo del mismo departamento adquirido previamente.

Producción científica

Las publicaciones realizadas con este trabajo de investigación fueron:

Armando Miranda Molina, Noé Alejandro Castro-Sánchez, Máximo López Sánchez, Dante Mújica Vargas y Javier Ortiz Hernández (2021). Generación de corpus de

comentarios de Amazon en español (Séptima Jornada de Ciencia y Tecnología Aplicada). Centro Nacional de Investigación y Desarrollo Tecnológico.

Armando Miranda Molina, Noé Alejandro Castro-Sánchez y Máximo López Sánchez (2021). Caracterización de comentarios de Amazon clasificados por estrellas (Coloquio de Lingüística Computacional). Universidad Nacional Autónoma de México.

Referencias

- Ali Salmony, M. Y., & Rasool Faridi, A. (2021). Supervised Sentiment Analysis on Amazon Product Reviews: A survey. 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM).
- Belete, D. M., & Huchaiah, M. D. (2021). Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *International Journal of Computers & Applications*, 1–12. <https://doi.org/10.1080/1206212x.2021.1974663>.
- Binder, M., Heinrich, B., Klier, M., Obermeier, A., & Schiller, A. (2019) Explaining the stars: aspect-based sentiment analysis of online customer reviews, in proceedings of the 27th European conference on Information systems (ECIS), Stockholm & Uppsala, Sweden, June 8-14, 2019. ISBN 978-1-7336325-0-8 research papers, (pp. 1-13).
- Bose, R., Dey, K. R., Roy, S., & Sarddar, D. (2018, agosto 31) Sentiment analysis on Online product reviews, conference: ICT4SD 2018 – Goa, At Goa, India.
- Castelli, M., Manzoni, L., Vanneschi, L., & Popovič A. (2017, mayo 5) An expert system for extracting knowledge from customers' reviews: the case of Amazon.com, Inc., doi: 10.1016/j.eswa.2017.05.008.
- Casasola, E., Pimentel, A., Sierra, G., Cámara, E., 3 & Marín, G. (2019) Análisis comparativo de las características computacionales en los sistemas modernos de análisis de sentimiento para el español, doi: 10.26342/2019-62-8.
- Cruz, L., A. (2017) Caracterización de comentarios emitidos por consumidores de productos y servicios (tesis de maestría), Centro Nacional de Investigación y Desarrollo Tecnológico, (pp. 11, 26-53).
- Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter tuning for machine learning algorithms used for Arabic sentiment analysis. *Informatics (MDPI)*, 8(4), 79. <https://doi.org/10.3390/informatics8040079>.
- Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, 2(1). <https://doi.org/10.1186/s40537-015-0015-2>.

- Geetha, M., Singha, P., & Sinha, S. (2018). Relationship between customer sentiment and online customer ratings for hotels - An empirical analysis. *Tourism Management*, 61, 43–54.
- Gelbukh, A., & Sidorov, G. (2013). *Construcción no lineal de n-gramas en la lingüística computacional*, México: Sociedad Mexicana de Inteligencia Artificial, 2013. 166 p, ISBN 978-607-95367-9-4.
- Gelbukh, A., & Sidorov, G. (2010) *Procesamiento automático del español con enfoque en recursos léxicos grandes*, Instituto politécnico Nacional México.
- Haque, T. U., Saber, N. N., & Shah, F. M. (2018). Sentiment analysis on large scale Amazon product reviews. 2018 IEEE International Conference on Innovative Research and Development (ICIRD).
- Hnin, C., Naw, N., & Win, A. (2018) *Aspect Level Opinion Mining for Hotel Reviews in Myanmar Language*, Singapore, Singapore, doi: 10.1109/AGENTS.2018.8460040.
- Jabbar, J., Urooj, I., Sheng, W., & Azeem, N. (2019) Real-time sentiment analysis on e-commerce application, IEEE 16th International Conference on Networking, Sensing and Control (ICNSC), doi: 10.1109/ICNSC.2019.8743331.
- Jagdale, S., Shirsat, S., & Deshmukh, N. (2019) Sentiment analysis on product reviews using machine learning techniques, *Advances in intelligent systems and computing*, doi: 10.1007/978-981-13-0617-4_61.
- Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., y Mikolov, T. (2016). FastText.zip: Compressing text classification models. In arXiv [cs.CL]. <http://arxiv.org/abs/1612.03651>.
- Kausar, S., Huahu, X., Ahmad, W., Shabir, M., & Ahmad, W. (2019, diciembre 30) A sentiment polarity categorization technique for Online product reviews, Published in: *IEEE Access* (Volume: 8), doi: 10.1109/ACCESS.2019.2963020.
- Kaushik, K., Mishra, R., Rana, N. P., & Dwivedi, Y. K. (2018). Exploring reviews and review sequences on e-commerce platform: A study of helpful reviews on Amazon.in. *Journal of Retailing and Consumer Services*, 45, 21–32. <https://doi.org/10.1016/j.jretconser.2018.08.002>.

- Kulkarni, A. & Shivananda A. (2019) Natural Language Processing Recipes Editorial APRESS.
- Kusrini, & Mashuri, M. (2019, mayo 15) Sentiment analysis in twitter using lexicon based and polarity multiplication, International Conference of Artificial Intelligence and Information Technology (ICAIT), doi: 10.1109/ICAIT.2019.8834477.
- Liu. B. (2012, abril 22) Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers.
- Mathew, L., James, A, D., Shaji, A., Sebastian, M, S. (2020, febrero 17) Opinion mining / sentiment analysis for user reviews.
- Mitchell, R. (2018). Web scraping with python: Collecting more data from the modern web (2nd ed.). *O'Reilly Media*.
- Mirjalili, S. & S, Raschka. (2019). Aprendizaje automático con Python 2da edición, ISBN: 978-84-267-2720-6.
- Mudambi, & Schuff. (2010). Research note: What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly: Management Information Systems*, 34(1), 185. <https://doi.org/10.2307/20721420>.
- Padró, L. (2012). Analizadores Multilingües en FreeLing. *Linguamática*, 3(2), 13-20. Obtenido de <https://www.linguamatica.com/index.php/linguamatica/article/view/115>.
- Pandey, S., & Deorankar, A. (2019, febrero 20) A study of sentiment analysis task and it's challenges, Published in: 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), doi: 10.1109/ICECCT.2019.8869160
- Pérez, C., Rebollar, A., & Pérez, M. (2019) Construcción de un corpus lingüístico a partir de tweets tomando como base un Tema específico, Jornada de Ciencia y Tecnología Aplicada Tecnológico Nacional de México/CENIDET vol. 2., (pp. 15-19).
- Torres, E., Jaimes, R., Hidalgo, O., & Mora, S. (2018, marzo 30) Influencia de redes sociales en el análisis de sentimiento aplicado a la situación política en Ecuador, doi: 10.29019/enfoqueute.v9n1.235.
- Rathora, S., Agarwalb, A., Dimric, P. (2018) Comparative study of machine learning approaches for amazon reviews, in *Procedia Computer Science*, doi: 10.1016/j.procs.2018.05.119

- Resyanto, F., Sibaroni, Y., & Romadhony, A. (2020) Choosing The Most Optimum Text Preprocessing Method for Sentiment Analysis, Semarang, Indonesia, Indonesia, doi: 10.1109/ICIC47613.2019.8985943.
- Ryan, M. (2018) Web scraping with Python, Editorial O'REILLY.
- Valdez, R., Sánchez, C., & Martínez, S. (2019). Distribución Semántica para el Análisis y la Clasificación de Documentos de Texto, Jornada de Ciencia y Tecnología Aplicada Tecnológico Nacional de México/CENIDET vol. 2., (pp. 11-14).
- Vanaja, S., & Belwal, M. (2018, julio 12) Aspect-level sentiment analysis on e-commerce data, International Conference on Inventive Research in Computing Applications (ICIRCA), doi: 10.1109 / ICIRCA.2018.8597286.
- Sadhasivam, J., Kalivaradhan, B, R. (2019, febrero 5) Sentiment analysis of amazon products using ensemble machine learning algorithm, International Journal of Mathematical, Engineering and Management Sciences, Vol. 4, doi: 10.33889/IJMEMS.2019.4.2-041.
- Sanmartín Sáez, J. (2018). Interacción discursiva y fórmulas de tratamiento en las respuestas de los hoteles a las opiniones de viajeros. *Onomázein Revista de Lingüística Filología y Traducción*, (add), 119–141.
- Saranya, K. Roslin, L., & Priyan H. (2019) Sentiment based comments rating approach, doi: 10.15680/IJIRCCE.2019. 0702097.
- Sultana, N., Chandra, S., Kumar, P., & Alam, S. (2019) Sentiment analysis for product review, Article in International Journal of Soft Computing, Vol. 9, doi: 10.21917/ijsc.2019.0266.
- Skalicky, S. (2013). Was this analysis helpful? A genre analysis of the Amazon.com discourse community and its “most helpful” product reviews. *Discourse Context & Media*, 2(2), 84–93. <https://doi.org/10.1016/j.dcm.2013.04.001>.
- Yang, X., Cai, Y., Ruan, Z., & Wang, T. (2020) Extracting polarity shifting patterns from any corpus based on natural annotation, ACM Transactions on Asian and Low-Resource Language Information Processing, Article No. 23 doi: 10.1145/3345518.
- Vjjala, S. (Ed.) (2020) Practical natural language processing, Editorial O'REILLY.