Análisis de sentimientos de reseñas para determinar la acogida de un producto utilizando técnicas de machine learning y data mining.

Julián Andrés Espitaleta Baldovino
Dept. Ingeniería de sistemas
Universidad del norte
Barranquilla, Colombia
jespitaleta@uninorte.edu.co

José Gabriel Maza Mendoza
Dept. Ingeniería de sistemas
Universidad del norte
Barranquilla, Colombia
jgmaza@uninorte.edu.co

Kelly Paola García de la Hoz
Dept. Ingeniería de sistemas
Universidad del norte
Barranquilla, Colombia
garciakp@uninorte.edu.co

Asesor

Eduardo Zurek Varela
Dept. Ingeniería de sistemas
Universidad del norte
Barranquilla, Colombia
ezurek@uninorte.edu.co

Tutor

Wilson Nieto Bernal
Dept. Ingeniería de sistemas
Universidad del norte
Barranquilla, Colombia
wnieto@uninorte.edu.co

***Abstract*— Reading multiple reviews can be tedious and concluding if a product has been liked or not is complicated, so it is necessary to implement a tool that analyzes all the reviews of a product and determines its polarity. The foregoing in order to speed up and improve decision-making about a product for those interested. During the development of the project, the strategy was designed and implemented using Machine learning and Data mining techniques to solve the problem. As a result, web scraping was applied to extract the reviews of an E-commerce, data was being visualized through Python libraries and transformed to classify them, thus allowing the polarity of a product to be obtained.**

Keywords—data mining, machine learning, sentiment analysis, classification.

I. INTRODUCCION.

A lo largo de las últimas dos décadas, hemos podido apreciar no solo el crecimiento exponencial de la cantidad de datos que generamos como usuarios de internet, sino también el refinado continuo de las técnicas que existen, así como el desarrollo de técnicas nuevas para procesar, analizar, cuantificar, y describir el volumen titánico de información. Para Blaid, Gupta y Chaplot en 2017 esto se debe a que las redes sociales y otras plataformas contienen grandes cantidades de información en forma de tweets, blogs, posts, etc, esto hace que encontrar información pertinente para un usuario sea una tarea más compleja.

Por consecuencia de lo anterior, en términos de experiencia de usuario ha crecido la expectativa subconsciente, de que los volúmenes inmensos de data que se generan en diferentes contextos sean utilizados de forma inteligente, ya sea, para que logren simplificar la experiencia de usuario, o para que generen un beneficio o conveniencia que no sería posible si no se pudiese trabajar con la mencionada cantidad de datos y contenido generado por el usuario. Es por esto que Kanwal et al. en 2021 mencionan la existencia de los sistemas de recomendaciones basadas en texto (RS), sistemas que prometen a encontrar información de manera más eficiente extrayendo información del texto, como a su vez estudios muestran el análisis de sentimientos como una manera efectiva de desarrollar la satisfacción del usuario a través de interpretación la data anteriormente mencionada (L. Yang, Y. Li, J. Wang and R. S. Sherratt, 2020).

En yuxtaposición, el gigantesco volumen de contenido generado por usuarios de un aplicativo hace que sea imposible que un usuario pueda consumir la totalidad del contenido que quiera o necesite consultar, con distintos fines. Se vuelve imposible leer cada comentario, cada reseña, cada artículo, cada nota, o cada crítica. Lo cual, puesto desde el punto de vista del usuario, hace imposible tomar decisiones respecto al contenido que se está viendo, o causa que el usuario tenga sesgos en su toma de decisiones según el contenido que le resulta plausible consultar (S. Hu, A. Kumar, F. Al-Turjman, S. Gupta, S. Seth and Shubham, 2020). Esto es importante, por ejemplo, en el área de compras de productos o servicios en línea, donde, a diferencia de una tienda física donde se puede apreciar el producto en persona, o en su defecto, es posible consultar a alguien físicamente para adquirir información sobre el producto. Por esto, los estudios de *Data Mining* se encuentran en pos de generar conversaciones funcionales que geneten entre el cliente, la tienda y a su vez el producto una disfrutable conversación a

través de expresiones automáticamente generadas que sinteticen la opinión general con no más información que la explícita (Y. Matsuyama, A. Saito, S. Fujie and T. Kobayashi, 2015).

Por supuesto, como ya se mencionó, con el avance en tecnologías de *Data Mining* y *Machine Learning* es posible sacar conclusiones completamente informadas, haciendo uso de toda la información que se tiene a nuestra disposición. A lo largo de este documento se explicará un ejemplo puntual, haciendo uso de Amazon como plataforma de e-commerce, para clasificar, analizar, y sacar conclusiones a partir de la experiencia de otros compradores. Se explicará la problemática en mayor detalle, así como la metodología que será utilizada para taclear el problema. Se seguirá con una implementación de la solución que plantearemos, y finalmente daremos nuestras conclusiones respecto a la metodología utilizada.

II. PROBLEMA.

Los comercios en línea suelen tener grandes cantidades de productos, los compradores deben navegar por multiples de ellos, incluso unos similares a otros, leyendo múltiples comentarios o reseñas sobre el producto para determinar si desean comprarlo o no ya que algunas veces conocer la calificación de un producto no es suficiente para convencerlos. Adicionalmente, durante este proceso de lectura de las reseñas, un comentario que un comprador dejó como positivo puede ser erróneamente interpretado como negativo, o bien un comentario negativo puede ser interpretado como positivo, es por eso que es necesario unificar este criterio.

Los grandes proveedores de productos en línea a debido a la latente necesidad de conocer la opinión general del cliente con el fin de mejorar sus productos o realizar promociones, se cree que las reseñas on-line ayudan incluso a formar la imagen del negocio y a promocionar las ventas (Huang et al., 2019). Sin embargo, para los vendedores leer múltiples reseñas e interpretarlas conlleva mucho tiempo, además se pueden cometer errores por ignorar algunas reseñas o clasificarlas erróneamente por intentar obtener resultados concluyentes de forma rápida.

Cabe destacar la importancia de estudios como el realizado por Huang et al. (2019) que aplican un modelo para extraer y filtrar el sentimiento de reseñas online, sin embargo, no ha sido aplicado para comercios en línea, por lo que invitan a los lectores a usar estas técnicas para llevar acabo un análisis de reseñas para un E-commerce. Igualmente, Baid, Gupta y Chaplot en 2017 realizaron un estudio donde analizan múltiples reseñas de películas utilizando distintas técnicas como Naive Bayes, K-Nearest neighbour y Random Forest, sin embargo,este estudio toma reseñas de multiples peliculas buscando el mejor metodo para clasificar, el reto de este proyecto está en extraer las reseñas directamente de un sitio web para un solo producto y así concluir la polaridad del mismo.

El propósito de este estudio, de teoría fundamentada en análisis de sentimientos, es desarrollar un algoritmo para determinar la acogida de un producto, que permita a compradores y vendedores de sistemas de ventas online (e-commerce) clasificar las reseñas de un producto de su escogencia de forma rápida y ágil, evaluadas bajo un mismo criterio.

III. JUSTIFICACIÓN.

La solución propuesta, analizará una gran cantidad de reseñas de un producto de forma ágil, utilizando un estándar o una misma métrica para evaluar todos los comentarios por igual, evitando así los errores humanos como no tener en cuenta alguno de los comentarios o interpretarlos de forma distinta a la intencionada. En relación a lo anterior, Hu et al. (2020) mencionan la dificultad que se presenta a la hora de descifrar las preferencias de los usuarios, más aún automatizar esta tarea, por lo que han ideado un modelo que incluso tiene en cuenta la credibilidad de un usuario al tener en cuenta su opinión, dándole un peso dentro de la clasificación final de las reseñas.

Las reseñas son una medida cualitativa a la que posibles y/o futuros compradores recurren para determinar la escogencia de lo que desean adquirir. Para los clientes, leer múltiples reseñas para un producto resulta ser bastante tedioso, más aún cuando se están comparando entre ellos y el potencial comprador debe leer varias reseñas de múltiples artículos. Para Kanwal et al (2019) este tipo de sistemas tienen la capacidad de encontrar información relevante en poco tiempo tan solo usando textos, los cuales en este caso serían reseñas.

Igualmente, la otra parte de los e-commerce, los vendedores, han comenzado a darse cuenta de la importancia de analizar las reseñas de sus productos ya que se cree que estos tienen un impacto significativo a la hora de formar una marca comercial, así como para promocionar sus ventas (Huang et al., 2019). Es importante recordar que algunos los vendedores tienen una gran variedad de productos, por lo que deben conocer cuáles son los que más les gustan a sus clientes y por lo tanto deberían promocionar más para aumentar sus ventas; o por el contrario deberían identificar los productos que no han tenido mucha aceptación y necesitarían considerar retirar de su catálogo o posiblemente dejar de producir. Así mismo, Yang et al. en 2020, afirman que realizar un análisis de sentimientos de multiples reseñas en plataformas e-commerce puede mejorar la satisfacción de los usuarios.

Es posible dar respuesta a esta problemática por medio de técnicas de *Data mining* y *Machine learning*, extrayendo las reseñas de un e-commerce y clasificándolas, determinando la polaridad de las mismas y del producto en general. Es por esto que investigadores, han decidido desarrollar herramientas que permitan determinar si un texto es positivo o negativo. Por ejemplo, Kausar et al., (2019) dan cuenta de un trabajo titulado "A Sentiment Polarity Categorization Technique for Online Product Reviews", donde exponen el desarrollo de una herramienta que categoriza una reseña en 5 sentimientos o polaridades teniendo como base los sentimientos positivos, negativos o neutros. En este estudio además se menciona la dificultad de poder realizar dicha categorización.

Este proyecto se realiza con el fin de obtener datos concluyentes sobre la polaridad de un producto permitiendo una mejora en la relación cliente-empresa. Se logrará identificar los productos de mayor preferencia facilitando así la ejecución de campañas de marketing o estudios de mercado. Además, el cliente podrá obtener un resumen de las reseñas permitiéndole observar más productos, agilizando la toma de decisiones a la hora de comprar.

IV. OBJETIVOS.

A. *Objetivo general*

Diseñar e implementar un programa para analizar la acogida de un producto a partir de la polaridad de las reseñas del mismo aplicando técnicas de machine learning y data mining.

B. *Objetivos específicos*

- Identificar los componentes claves para el diseño y la implementación relacionados con el análisis de sentimientos en textos a partir de la revisión sistemática de la literatura.
- Diseñar la arquitectura lógica de la solución para la implementación de algoritmos que permitan el análisis de sentimientos de compradores.
- Desarrollar un prototipo de la solución que encuentre las reseñas de un producto, analice los sentimientos de sus compradores y genere una conclusión a partir de este.
- Validar que el prototipo de la solución propuesta extraiga información sobre reseñas de productos, logre analizar la polaridad de los sentimientos de los compradores de un producto dado y genere una conclusión a partir de dicho análisis.

V. DISEÑO DE LA INVESTIGACIÓN.

El diseño de investigación es un plan estructurado de acción que reúne el conjunto de métodos y/o técnicas utilizadas para que el problema a resolver sea manejado eficiente. Por consiguiente, dicho diseño es fundamental para alcanzar los objetivos propuestos pues, es el que define los pasos a seguir para poder lograrlo. Teniendo en cuenta lo anterior, para el proyecto planteado se han propuesto 3 fases que permitirán alcanzar los objetivos y, por lo tanto, dar solución a la problemática a resolver.

A. *Investigación inicial.*

En esta fase, se propone una revisión sistémica de la literatura, que permita identificar los componentes claves para el diseño e implementación de proyectos relacionados con el análisis de sentimientos en textos y la elaboración de textos. Lo anterior, se realiza con el fin de comprender el problema y los conceptos a tratar a lo largo del proyecto y así poder plantear posibles soluciones que se puedan implementar.

Durante la revisión sistemática de la literatura se buscaron conceptos clave tales como review sentiment analysis, building lexicons, sentence generation y text-based recommendation system en fuentes de información como IEEE-XPLORE y ACM-DL aplicando filtros. A partir de lo anterior, se realizaría una tabla *(Anexo 1)* con literaturas que tratan problemas similares al planteado en este proyecto.

Fuentes de información sin filtros				
Fuente	*Review sentiment Analysis*	*Building lexicons*	*Sentence generation*	*Text-based recommendation system*
ACM DL	505,477	365,985	408,372	621,196
IEEE-XPLORE	2,322	380	1,160	61

Tabla 1. Resultados de la búsqueda en fuentes de información sin aplicar filtros

Fuentes de información, últimos 5 años, papers				
Fuente	*Review sentiment analysis*	*Building lexicons*	*Sentence generation*	*Text-based recommendation system*
ACM DL	9,836	6,912	7,530	11,878
IEEE-XPLORE	1,651	142	607	28

Tabla 2. Resultados de la búsqueda en fuentes de información con filtros

B. *Arquitectura de la solución.*

Teniendo en cuenta la información obtenida en la fase anterior, se realiza un análisis del problema identificado. Por medio de dicho análisis será posible plantear una solución factible, pero, además será posible observar que herramientas han sido utilizadas anteriormente para solucionar problemas similares al identificado y que, sería recomendable utilizar para obtener buenos resultados. Por consiguiente, es posible construir la arquitectura lógica y física del proyecto que definirán las herramientas y/o tecnologías a utilizar para el desarrollo de la solución y que se implementarán en la siguiente fase, recordando que estas arquitecturas deben ser intuitivas, flexibles y consistentes.

C. *Desarrollo, implementación y validación del prototipo.*

En esta última fase, teniendo en cuenta la arquitectura física y lógica de la solución planteada anteriormente, y considerando la naturaleza del problema a solucionar, se propone una metodología para el desarrollo, implementación y validación del prototipo del proyecto. La metodología CRISP-DM, es un modelo que se utiliza comúnmente en proyectos de desarrollo relacionados con *Data mining*, por lo que se ajusta a la solución propuesta. Por medio de este modelo, será posible comprender el problema, obtener los datos necesarios y prepararlos para ser utilizados, modelar la solución y evaluar que este funcione correctamente.

VI. METODOLOGÍA.

Debido a que el proyecto propuesto es una solución relacionada con *Machine learning* y *Data Mining*, hay que tener en cuenta que algunas preocupaciones desaparecen al momento de planificar, implementar y realizar pruebas. Por consiguiente, se propone hacer uso de la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), un modelo de proceso de minería de datos que describe una manera en la que los expertos en esta materia abordan el problema (Galán, 2015), para el diseño e implementación del proyecto.

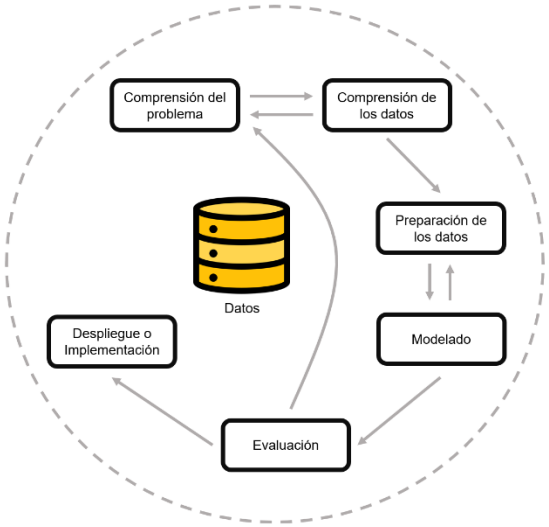

Fig 1: Metodología CRISP-DM

1) Fase de comprensión del problema

En la primera fase, se comprenderá el problema a tratar, teniendo en cuenta los objetivos y requisitos del proyecto, con el fin de obtener una idea de lo que se puede conseguir con los datos obtenidos o bien los resultados obtenidos.

2) Fase de comprensión de los datos

En la segunda fase de esta metodología, se exploran los datos disponibles y se decide de que fuente se van a extraer los datos que se utilizaran posteriormente. Para el desarrollo del proyecto, se hará uso de dos tipos de datos: datos para entrenamiento del modelo y datos empíricos que se obtendrán luego de entrenado el modelo, por consiguiente, el proceso mencionado anteriormente se deberá realizar ambos conjuntos de datos. Los primeros datos, se tomarán de bases de datos, o datasets, públicos que se pueden obtener en la plataforma web Kaggle, estos se utilizarán para entrenar el modelo antes de hacer uso de datos empíricos; mientras que los segundos datos se obtendrán por medio la técnica de *Web scrapping* a una página web de un e-commerce.

3) Fase de preparación de los datos

Los datos que se recopilaron en la fase anterior pasarán por un proceso de selección, limpieza y transformación donde se tratará de corregir o eliminar los errores en algunos valores y se tomará una decisión con los datos incompletos. Es posible realizar este proceso de preparación de los datos de manera ágil y sencilla por medio algunas de las librerías de Python. Al final de esta fase, se tendrá un conjunto de datos definitivo con aquellos datos que son relevantes para la investigación que se va a realizar, estos se guardarán en una base de datos temporal para ser utilizados posteriormente.

4) Fase de modelamiento

El conjunto de datos definitivo entra en la fase principal del proyecto donde se pasa al procesado y creación de un modelo. En esta fase, los datos se agruparán, clasificaran y analizaran incorporándolos en distintas herramientas tales como Keras, Sci-Kit Learn, Tensorflow y otras definidas en la arquitectura lógica de la solución.

5) Fase de evaluación

En esta fase, dentro de los resultados obtenidos se seleccionan los más pertinentes, estos pasan por un proceso en el que se realiza la evaluación e interpretación de los datos, teniendo en cuenta los resultados obtenidos en la fase de comprensión del problema, con el fin de poder entender el resultado obtenido. Dependiendo de estos, podría ser necesario tener que realizar una revisión del problema, donde se analizaran nuevamente los datos y el modelo para realizar los ajustes que sean necesarios con el fin de que el modelo sea perfeccionado. Adicionalmente, en esta fase se deberá comprobar y validar que el modelo cumple con los objetivos propuestos.

6) Fase de despliegue o implantación

En esta fase, luego de que el modelo ha sido construido y validado, se transforma el conocimiento aplicando el modelo a diferentes conjuntos de datos. Esta última fase de la metodología CRISP-DM no se tendrá en cuenta ya que, generalmente un proyecto de minería de datos no concluye en la implantación del modelo ya que se deben documentar y presentar los resultados de manera comprensible para el usuario (Galán, 2015)

VII. MARCO TEORICO.

En la actualidad, gran parte de las ventas de productos de interés se dan a través de plataformas e-commerce, tales como redes sociales y páginas web. Por consiguiente, el descifrar las preferencias de compra de los usuarios, sus gustos y disgustos se ha vuelto una necesidad para estas plataformas, dicho proceso es difícil incluso para los humanos, lo que hace que su automatización sea un trabajo muy complejo (Hu et al., 2020). De igual forma, las plataformas e-commerce, se han dado cuenta de la importancia de analizar las reseñas recibidas por parte de los clientes a través de la valoración sus productos, ya que estas, generan un impacto significativo para la plataforma o la marca, así como la promoción e incremento de sus ventas (Huang et al., 2019).

En base a lo anterior, una opción presente en la actualidad para la interpretación de dichas reseñas es el *Sentiment analysis*, el cual tiene como objetivo extraer opiniones o emociones principalmente del texto escrito (Martinis et al., 2022). A través de este método, es posible interpretar la intención de un usuario a la hora de escribir una reseña para un producto, bien sea positiva, negativa o neutra. Con respecto a lo anterior, Huang et al. (2019), exponen resultados

experimentales que han demostrado que es importante filtrar correctamente la información valiosa oculta dentro de las reseñas de los consumidores para predecir las ventas, mejorando incluso su precisión.

Así mismo, por medio del *Sentiment analysis*, el cual se ha utilizado anteriormente para interpretar distintos tipos de textos, es posible realizar un análisis sobre las reseñas de un e-commerce. En relación a lo anterior, Huang et al. (2019) han propuesto utilizar estos métodos para conseguir buenos resultados en el análisis de reseñas, aplicándolos para extraer y filtrar la información que se encuentra dentro de las valoraciones de los productos. Adicionalmente, mencionan que este modelo es posible aplicarlo para un e-commerce con el fin de analizar la conducta de sus reseñas en línea.

Por otro lado, por medio de la extracción, usando técnicas de *Web Scrapping*, y de clasificación de las reseñas es posible obtener palabras claves, que pueden llegar a determinar la polaridad de la información que se encuentra tanto como dentro de una reseña en particular, como de un grupo de estas. Al respecto, Tchalakova et al. (2011) hicieron uso del *Sentiment analysis* con el fin de encontrar una forma estadística de representar palabras y frases que se usan en reseñas ya sean negativas o positivas. De igual forma, destacan la importancia de tener en cuenta el contexto en que un texto ha sido escrito ya que, depender de *Sentiment lexicons* predefinidos podrían resultar en ambigüedades. Debido a esto, Bross y Ehrig (2010) se han propuesto generar unos *Sentiment lexicons* que, tuviesen en cuenta el contexto, haciendo uso del *Sentiment analysis* para enfocarse en dominar las reseñas de los usuarios sobre algunos productos, explorando el uso de algunas frases para clasificar las reseñas.

Teniendo en cuenta lo anterior, es posible extraer reseñas directamente de los e-commerce por medio de técnicas de *Web Scrapping*, y por medio de estas obtener algunas las palabras clave, o bien, *Sentiment lexicons*, que permitan entonces clasificar si una reseña en particular es positiva o negativa; o bien de un conjunto de reseñas de los clientes que han pedido un producto y determinar si dicho producto les ha gustado o no.

Evidentemente, el rápido desarrollo del internet y tecnologías las compras por internet han aumentado, y aplicar el *Sentimental analysis* a las reseñas de los productos es una forma de aumentar la satisfacción de los clientes y/o futuros compradores (Yang et al., 2020). Leer múltiples reseñas de un producto, clasificarlas y concluir el sentimiento que tienen los clientes hacia dicho producto puede ser tedioso y difícil de concretar tanto para vendedores como compradores, por lo que aplicar las técnicas y procesos mencionados anteriormente se ha vuelto una necesidad para el comercio en línea.

VIII. MARCO CONCEPTUAL.

A. *Clasificación*

La clasificación permite separar un grupo de cosas gracias a un criterio determinado, en el campo de la inteligencia artificial es posible clasificar información y datos más complejos gracias al *Data mining* y *Machine learning*. Por medio del *Sentiment Analysis*, es posible generar una clasificación por polaridad de los productos. En ese orden de ideas, dicha clasificación por polaridad tiene como objetivo encontrar una forma estadística de representar las palabras o frases que se utilizan en una reseña, teniendo en cuenta si estas expresan algo negativo o algo positivo (Tchalakova et al., 2011).

B. *Data Mining*

La minería de datos es la práctica de buscar automáticamente grandes cantidades de datos para descubrir patrones y tendencias que van más allá del simple análisis. La minería de datos utiliza algoritmos matemáticos sofisticados para segmentar los datos y evaluar la probabilidad de eventos futuros. La minería de datos también se conoce como *Knowledge Discovery in Data (KDD)*. (Oracle, 2018)

C. *Machine learning*

El Machine learning es un tipo de IA que permite a las máquinas aprender en forma directa de ejemplos y de experiencia, la cual es adquirida a partir de un conjunto de datos (conocido como dataset), a través de un entrenamiento. La programación tradicional está basada en reglas estáticas, las cuales establecen cómo resolver un problema, paso a paso. Por lo contrario, con machine learning se dispone una gran cantidad de datos para utilizar como ejemplo de cómo la tarea puede ser realizada, o para detectar patrones (The Royal Society, 2017). Este tipo de algoritmos tienen la capacidad de predecir nuevos casos en base a la experiencia aprendida a través de los datos que fueron usados para su entrenamiento (Gisela & García, 2014).

D. *Sentiment analysis*

El análisis de sentimientos también es conocido como la minería de opiniones, por medio de este es posible analizar los sentimientos que se encuentran en un texto por medio de las palabras que se utilizan dentro del mismo, logrando incluso determinar su polaridad. Es así como entonces, el *Sentimental analysis* ha podido llegar a mostrar la opinión y las emociones de las personas sobre productos o servicios, pues por medio de esta técnica posible conocer las reseñas sobre un producto son positivas, negativas o neutras (Kausar et al. 2020). En este orden de ideas, Yang et al. (2020) afirman que analizar los sentimientos de los compradores es de gran importancia para otros compradores y plataformas e-commerce, pues las reseñas influyen en las decisiones de compra de los usuarios.

E. *Sentiment lexicón*

A la hora de analizar un texto por medio del análisis de sentimientos, es necesario una colección de palabras con algún sentimiento asociado, se utiliza para mejorar las características de las opiniones en las reseñas (Yang et al., 2020). Con respecto a esto, Bross y Ehrig, (2010) afirman que una buena parte del *Sentiment analysis* es usar lexicones especiales que provean información para la orientación semántica (positiva, negativa o neutral).

F. *Web Scrapping*

El web scrapping es una técnica que se utiliza para extraer información de páginas web, automatizando el proceso de

recolección de datos. Las páginas web estan construidas por elementos, dichos elementos pueden ser un título, una imagen, un enlace, un párrafo, etc. Por medio del web scrapping, se puede acceder a los elementos de una página web, revisando el codigo fuente de la página web, extrayendo así sus atributos y contenido. Para López (2018), el web scrapping multiplica las posibilidades de recolectar información que, aunque está publicada en Internet, es inaccesible por no tener una estructura clara, estar dispersa o ser masiva.

IX. Arquitectura lógica de la solución

Para la realización de un prototipo para la solución, se hará uso de una arquitectura lógica basado en un Stack de Python. Primeramente, la data que se obtendrá para entrenamiento se obtendrá del repositorio Kaggle. La data se extrae en formato CSV y se almacenaran los datos sin procesar localmente. El lenguaje de Python tiene un ecosistema extenso lleno de herramientas e implementaciones aplicadas. Dentro de nuestro stack, se usarán las herramientas NumPy, SciPy, Pandas y Matplotlib para la exploración, análisis preliminar, y limpieza de datos, una vez hecho esto, se usará Pandas en combinación con SQLAlchemy para almacenar los datos en una Base de datos temporal, para agilizar el proceso de obtención de los datos limpios. Una vez se tenga un dataset estructurado, se pasa al procesado y creación de un modelo. Para esto se usan herramientas como Keras, Sci-KitLearn, y más importantemente, Tensorflow. Con estas herramientas construiremos un flujo de información que, a su vez, generará un modelo computacional que analizará el texto de los comentarios de nuestro dataset.

Finalmente, una vez hayamos entrenado el modelo, el siguiente paso será recolectar datos empíricos. Para esto, se utilizarán las herramientas Scrappy y Beautiful Soup 4, que permiten hacer scrapping a las páginas y obtener datos experimentales en los que aplicar el modelo. Con las variables de salida del modelo, se utilizará Matplotlib dara mostrar los datos gráficamente.

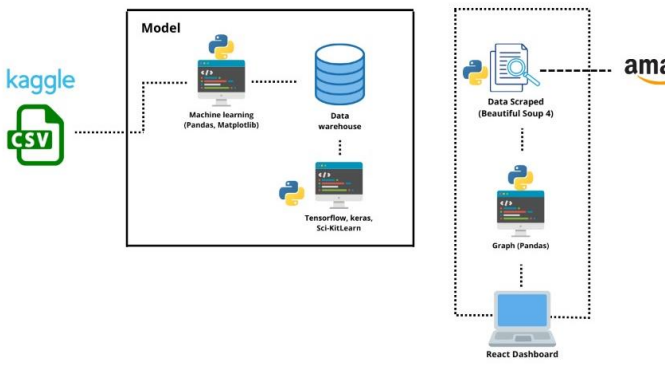

Fig 2. Arquitectura de la solución

X. Prototipo.

Para la implementación de prototipo , se hizo uso de la librería Pandas, librería clave en el estudio de datos, el algoritmo interno de esta librería requiere de una validación de la *data* (datos descargados del dataset, proporcionado por Kaggle). Es decir que, en primera instancia el algoritmo del prototipo consta de las siguientes validaciones y a su vez respectiva limpieza de datos, teniendo en cuenta que *data* es el dataset anteriormente mencionado:

```
# Clean the data
# Check for NA values
print("Number of NA values in each column:")
print(data.isna().sum())

# Check for duplicates
print("Number of duplicated rows: ", end='')
print(data.duplicated().sum())


# Remove NA values
print("Removing NA values... ", end='')
data.dropna(inplace=True)
print("Done!")
```

Fig 3. Código proceso de validación datos.

Luego de realizar dicho proceso de validación, es posible evidenciar gráficamente, por medio de un diagrama de barras (Bar plot), que los datos efectivamente se encuentran balanceados.

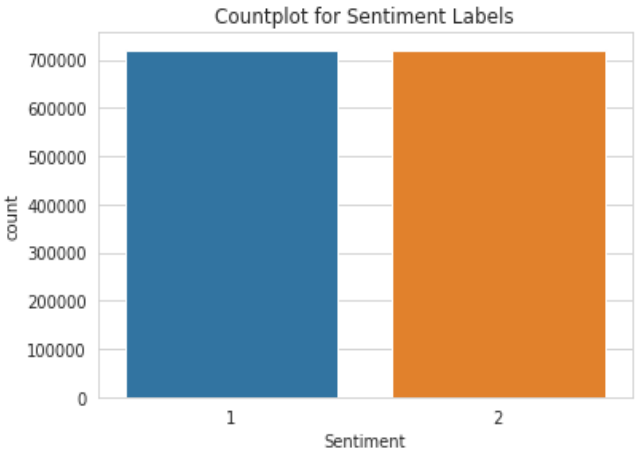

Fig 4. Diagrama de barras resultado del Balanceo

Posteriormente, el dataset es limpiado aplicando distintos procedimientos y filtros como darles un formato a los datos, remover signos de puntuación, sintaxis HTML, URLS, Emoticones e incluso eliminar las *Stopwords*, o palabras vacías del texto, palabras que no agregan valor a la polarización de un texto, con el fin de luego poder lematizar los datos, es decir, llevar las palabras a su forma de origen.

```python
# #Removes Punctuations
def remove_punctuations(data):
    punct_tag=re.compile(r'[^\w\s]')
    data=punct_tag.sub(r'',data)
    return data

#Removes HTML syntaxes
def remove_html(data):
    html_tag=re.compile(r'<.*?>')
    data=html_tag.sub(r'',data)
    return data

#Removes URL data
def remove_url(data):
    url_clean= re.compile(r"https://\S+|www\.\S+")
    data=url_clean.sub(r'',data)
    return data

#Removes Emojis
def remove_emoji(data):
    emoji_clean= re.compile("["
                            u"\U0001F600-\U0001F64F"
                            u"\U0001F300-\U0001F5FF"
                            u"\U0001F680-\U0001F6FF"
                            u"\U0001F1E0-\U0001F1FF"
                            u"\U00002702-\U000027B0"
                            u"\U000024C2-\U0001F251"
                            "]+", flags=re.UNICODE)
    data=emoji_clean.sub(r'',data)
    url_clean= re.compile(r"https://\S+|www\.\S+")
    data=url_clean.sub(r'',data)
    return data
```

Fig 5. Código para eliminar puntuaciones, html, url y emoticones.

```python
def remove_stopwords(data):
    stop_words = set(stopwords.words('english'))
    word_tokens = word_tokenize(data)
    filtered_sentence = [w for w in word_tokens if not w in stop_wor
    filtered_sentence = []
    for w in word_tokens:
        if w not in stop_words:
            filtered_sentence.append(w)
    return " ".join(filtered_sentence)
```

Fig 6. Código para eliminar las palabras vacías.

Luego se realizar este proceso, es posible identificar las palabras más utilizadas dentro de cada tipo de reseñas, positiva o negativa, para el dataset de Kaggle. Es posible visualizar esta información por medio de un Word Cloud, un gráfico donde a mayor tamaño más se ha repetido la palabra dentro de un conjunto de estas. Para este caso, los conjuntos de palabras serán aquellas que hacen parte de las reseñas positivas y aquellas que hacen parte de las reseñas negativas para el dataset de Kaggle.

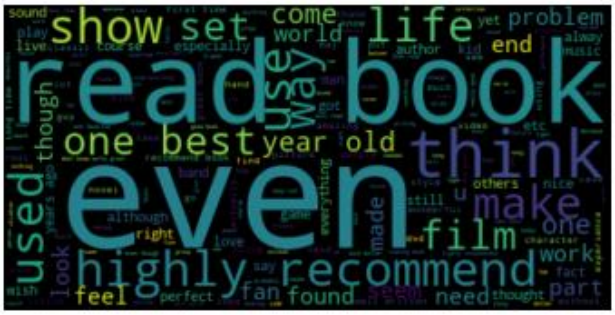

Fig 7. Word cloud para reseñas positivas.

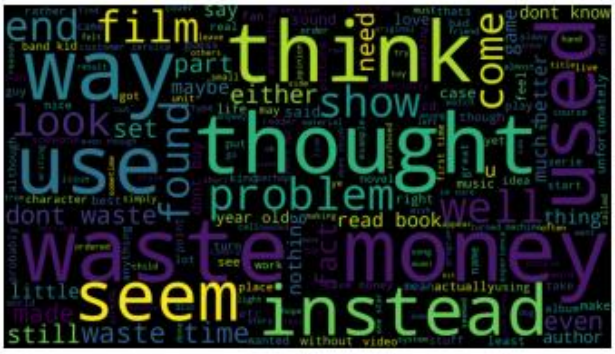

Fig 8. Word cloud para reseñas negativas.

A partir de lo anterior, luego de lematizar las palabras, es posible definir un modelo. Para esto se dividen los datos entre un set de entrenamiento (Train) y un set de prueba (Test) con ayuda de la librería Sklearn. Esta misma librería será la que se encargará de los distintos pasos para la implementación del modelo, como la vectorización, regresión logistica, escalamiento, *fitting*, predicciones y resultado preliminares.

```python
vectorizer = CountVectorizer(token_pattern=r'\b\w+\b')
train_matrix = vectorizer.fit_transform(Xt['Review Text'])
test_matrix = vectorizer.transform(Xv['Review Text'])
```

Fig 9. Código del proceso de vectorización

```python
lr = LogisticRegression(solver='lbfgs', max_iter=6000)
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler(with_mean=False)
X_train = scaler.fit_transform(train_matrix)
X_test = scaler.transform(test_matrix)
```

Fig 10. Código regresión logística y escalamiento.

Es importante mencionar, que todo el proceso anteriormente mencionado se seccionó para limitar el uso de recursos debido a que la plataforma de Google, Colab, no permite el uso de máquinas fuertes en las versiones gratuitas.

Con el modelo finalizado, se pueden observar distintas métricas del modelo implementado por medio Sklearn Metrics donde se le enviará como entrada las predicciones

realizadas con nuestro set de pruebas. Por medio de estos será posible generar un reporte donde se observe la precisión, exhautividad y puntaje f1 de nuestro modelo (Fig 11) y evaluar el porcentaje de verdadero positivo y falsos positivos. Así mismo, es posible generar una matriz de confusión que nos permita observar con claridad los valores bien clasificados, para el dataset de Kaggle, en verdadero positivo, verdadero negativo, falso positivo y falso negativo (Fig 12).

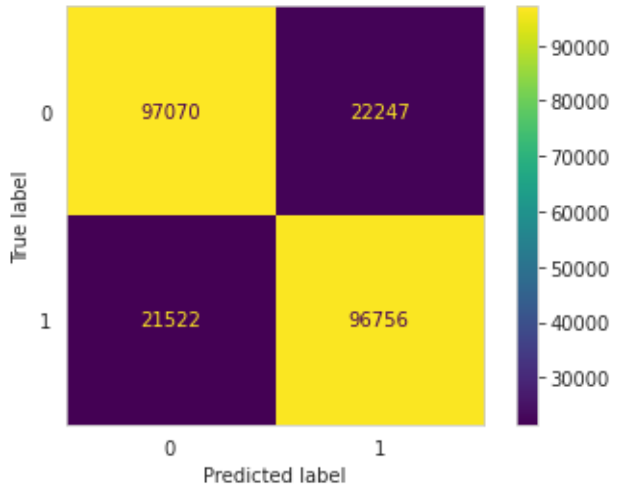

Fig 12. Matriz de confusión.

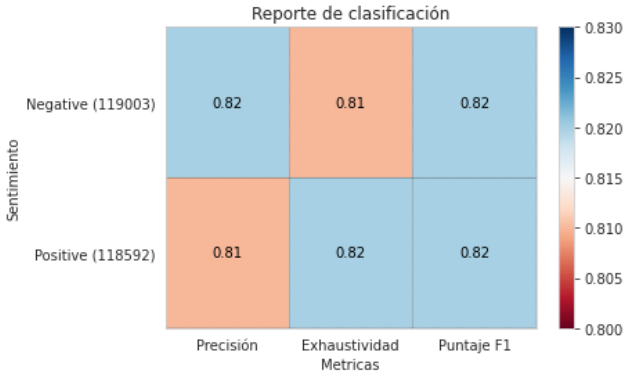

Fig 13. Reporte de clasificación.

Posteriormente, se procede al scrapping, este se realiza con ayuda de la librería BeautifulSoup, el prototipo recibe un término de búsqueda, que usa el buscador de Amazon para obtener los resultados de esa búsqueda. El usuario debería entonces poder seleccionar de entre los productos seleccionados el que se desea scrappear.

```
print("Select a product to scrape: ")
for idx, (url, prod) in enumerate(prods_with_urls, start=1):
    print(f"\t{idx}. {prod}")

selection: int = int(input("Selection: "))

selected_url: str = f"{BASE_AMZN_URL}{prods_with_urls[selection-1][0]}"
selected_url = selected_url.replace("/dp/", "/product-reviews/")

print(f"The URL to be scrapped is: {selected_url}")
```

Fig 14. Código selección de producto

A partir de lo anterior, se extrae el texto de los comentarios en la página de un E-commerce (Amazon), y se utiliza el algoritmo para clasificarlos en positivos y negativos. A partir de esto es posible observar un gráfico que resume la polaridad de los comentarios de un artículo.

Fig 15. Resultados de la polaridad de los artículos.

XI. VALIDACION PROTOTIPO.

Es necesario aplicar una evaluación competente del prototipo, y del proyecto desarrollado en general, con el fin de validar que se cumplen los objetivos planteados y que el propósito de la aplicación se está llevando a cabo exitosamente. Debido a lo anterior, en el ámbito de la asignatura Proyecto Final, se realizó la validación del prototipo en grupos de pares haciendo uso del estándar ISO 15504 por medio del Instrumento de evaluación del prototipo. Teniendo en cuenta que 5 es totalmente de acuerdo, 1 totalmente en desacuerdo y N es no aplica, los resultados obtenidos se pueden observar en la *Tabla 1*.

Caracteristica	Definición o descirpción	1	2	3	4	5
Understandability	¿Facil de comprender?					5
Documentation	¿Documentación de usuario completa,					5
Buildability	¿Fácil de construir en un sistema compatible?			3		
Installability	¿Fácil de instalar en un				4	
Learnability	¿Fácil de aprender a	N				
Identity	¿La identidad del proyecto / software es				4	
Copyright	¿Es fácil ver quién posee el proyecto /					
Licencing	Adopción de la licencia	N				
Governance	¿Fácil de entender cómo se ejecuta el proyecto y cómo se gestiona el					5
Community	¿Evidencia de					
Accessibility	¿Evidencia de capacidad de descarga				4	
Testability	¿Fácil de probar la corrección del					5
Portability	¿Utilizable en múltiples				4	
Supportability	¿Evidencia de soporte para desarrolladores		2			
Analysability	¿Fácil de entender a				4	
Changeability	¿Fácil de modificar y aportar cambios a los					5
Evolvability	¿Evidencia de					5
Interoperability	software requerido / relaciona				4	
Modelo de evaluación basado en el estandar ISO 9126 ISO 15504 + ESCALA DE LIKERT						

Tabla 3. Instrumento de evaluación prototipo

XII. Conclusiones.

Después de trabajar en la investigación y desarrollo del proyecto es posible afirmar que se cumplió los objetivos propuestos. Se comenzó realizando una revisión sistémica de la literatura que tratara temas relacionados con el análisis de sentimientos y otras formas de determinar la polaridad de un texto, procesamiento de lenguaje natural, web scrapping y otras técnicas para extraer reseñas de un comercio en línea, entre otros. Esto se realizó con el fin de obtener las bases para elaborar el informe y el prototipo de la solución.

Al obtener y organizar la información relevante encontrada, fue posible idear distintas rutas de acción que nos permitirían proponer una solución al problema identificado. Posterior a esto, tomando como referencia trabajos realizados anteriormente sobre la misma temática fue posible identificar las herramientas que podrían ser útiles a la hora de implementar dicha la solución del proyecto, logrando establecer la arquitectura del prototipo a realizar.

Después de realizar lo anteriormente mencionado, se definió la metodología de desarrollo que permitiría implementar el prototipo y validar su funcionamiento. Para el desarrollo del prototipo, se utilizaron las herramientas previamente definidas logrando así recolectar y visualizar información, procesar los datos, y definir un modelo que se utilizaría con los datos obtenidos para dar respuesta la problemática.

Luego de realizar el prototipo del proyecto, se realizó la validación del mismo por medio del instrumento de evaluación del prototipo. Lo anterior, se realizó con el fin de conocer la opinión de un grupo par acerca del cumplimiento de algunas características que se plantean en el estándar ISO 15504. Por medio de dicha evaluación, fue posible observar que es un prototipo fácil de entender en general y se encuentra bien documentado, sin embargo, futuros desarrolladores no contarán con soporte, así como se les dificultará construir un sistema compatible.

Para finalizar, gracias a cada una de las etapas anteriormente descritas fue posible alcanzar cada uno de los objetivos específicos planteados al inicio del proyecto. Es necesario destacar que el desarrollo del proyecto busca brindar una herramienta que ayude a la toma de decisiones a la hora de realizar compras en línea para los posibles clientes de un e-commerce, o bien, una forma de evaluar los productos que un vendedor tiene dentro de su catálogo por lo que el proyecto. Por consiguiente, este estudio se podría tener en cuenta para futuras investigaciones de temática similar, utilizando distintos comercios on-line o incluso una mayor cantidad de reseñas, por lo que se extiende la invitación a otras personas o grupos de investigación a tomar como base este estudio para realizar uno que incluya otros factores que puedan ayudar a mejorar la polarización de una reseña teniendo en cuenta una mayor cantidad de factores y conjunto de datos.

Referencias.

Baid, P., Gupta, A., & Chaplot, N. (2017). Sentiment analysis of movie reviews using machine learning techniques. International Journal of Computer Applications, 179(7), 45-49.

Galán, V. (2016). Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario (Bachelor's thesis).

Hu, S., Kumar A., Al-Turjman, F., Gupta, Seth, S. y Shubman, S. (2020). Reviewer Credibility and Sentiment Analysis Based User Profile Modelling for Online Product Recommendation. IEEE Acesss, 8, 26172-26189. 10.1109/ACCESS.2020.2971087

Huang, L., Dou, Z., Hu, Y., y Huang, R. (2019). Textual Analysis for Online Reviews: A Polymerization Topic Sentiment Model.IEEE Access, 7, 91940-91945. https://doi.org/10.1109/ACCESS.2019.2920091.

Kanwal, S., Nawaz, S., Malik, K. and Nawaz, Z. (2021). A Review of Text-Based Recommendation systems.

IEEE Access, 9, 31638-31661. https://doi.org/10.1109/ACCESS.2021.3059312.

Kausar, S., Huahu, X., Ahmad, W., Shabir, Y. y Ahmad, W. (2020). A Sentiment Polarity Categorization Technique for Online Product Reviews. IEEE Access, 8, 3594-3605. https://doi.org/10.1109/ACCESS.2019.2963020.

Martinis, M., Zucco, C. y Cannataro, M. (2022). An Italian lexicon-based sentiment analysis approach for medical applications. In Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '22). Association for Computing Machinery, 70, 1–4. https://doi.org/10.1145/3535508.3545594

López, J. (2018). Web scraping. Tomado de: https://www.academia.edu/download/55775125/web-scraping.pdf

Oracle. (2018). What Is Data Mining. (Oracle) Tomado de: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#

Tchalakova, M., Gerdemann, D. y Meurers, D. (2011). Automatic sentiment classification of product reviews using maximal phrases-based analysis. In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA '11).

The Royal Society. (2017). Machine learning: the power and promise of computers that learn by example. Report by the Royal Society (Vol. 66). https://doi.org/10.1126/scitranslmed.3002564

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. Knowledge discovery and data mining: towards a unifying framework. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, 82–88.

Yang, L., Li, Y., Wang, J. y Sherratt. S. (2020), Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning. IEEE Access, vol. 8, pp. 23522-23530. https://doi.org/10.1109/ACCESS.2020.2969854.

XIII. ANEXOS

A. Anexo 1: Revisión sistemática de la literatura

Title	Authors	Abstract	Keywords	Cite (IEEE)
A Sentiment Polarity Categorization Technique for Online Product Reviews	S. Kausar, X. Huahu, W. Ahmad, M. Y. Shabir and W. Ahmad	Sentiment analysis is also known as opinion mining which shows the people's opinions and emotions about certain products or services. The main problem in sentiment analysis is the sentiment polarity categorization that determines whether a review is positive, negative or neutral. Previous studies proposed different techniques, but still there are some research gaps, i) some studies include only 3 sentiment classes: positive, neutral and negative, but none of them considered more than 3 classes ii) sentiment polarity features were considered on individual basis but none of them considered on both individual and on combined basis iii) No previous technique considered five sentiment classes with 3 sentiment polarity features such as a verb, adverb, adjective and their combinations. In this study, we propose a sentiment polarity categorization technique for a large data set of online reviews of Instant Videos. A comprehensive data set of five hundred thousand online reviews is used in our research. There are five classes (Strongly Negative, Negative, Neutral, Positive and Strongly Positive). We also consider three polarity features Verb, Adverb, Adjective and their combinations with their different senses in review-level categorization. Our experiments for review-level categorization show promising outcomes as the accuracy of our results is 81 percent which is 3 percent better than many previous techniques whose average accuracy is 78 percent.	Review sentiment analysis Opinion mining Categorization	S. Kausar, X. Huahu, W. Ahmad, M. Y. Shabir and W. Ahmad, "A Sentiment Polarity Categorization Technique for Online Product Reviews," in IEEE Access, vol. 8, pp. 3594-3605, 2020, doi: 10.1109/ACCESS.2019.2963020.

Reviewer Credibility and Sentiment Analysis Based User Profile Modelling for Online Product Recommendation	S. Hu, A. Kumar, F. Al-Turjman, S. Gupta, S. Seth and Shubham	Deciphering user purchase preferences, their likes and dislikes are a very tricky task even for humans, making its automation a very complex job. This research work augments heuristic-driven user interest profiling with reviewer credibility analysis and fine-grained feature sentiment analysis to devise a robust recommendation methodology. The proposed credibility, interest and sentiment enhanced recommendation (CISER) model has five modules namely candidate feature extraction, reviewer credibility analysis, user interest mining, candidate feature sentiment assignment and recommendation module. Review corpus is given as an input to the CISER model. Candidate feature extraction module uses context and sentiment confidence to extract features of importance. To make our model robust to fake and unworthy reviews and reviewers, reviewer credibility analysis proffers an approach of associating expertise, trust and influence scores with reviewers to weigh their opinion according to their credibility. The user interest mining module uses aesthetics of review writing as heuristics for interest-pattern mining. The candidate feature sentiment assignment module scores candidate features present in review based on their fastText sentiment polarity. Finally, the recommendation module uses credibility weighted sentiment scoring of user preferred features for purchase recommendations. The proposed recommendation methodology harnesses not only numeric ratings, but also sentiment expressions associated with features, customer preference profile and reviewer credibility for quantitative analysis of various alternative products. The mean average precision (MAP@1) for CISER is 93% and MAP@3 is 49%, which is better than current state-of-the-art systems.	Review sentiment analysis Recommendation system Feature extraction	S. Hu, A. Kumar, F. Al-Turjman, S. Gupta, S. Seth and Shubham, "Reviewer Credibility and Sentiment Analysis Based User Profile Modelling for Online Product Recommendation," in IEEE Access, vol. 8, pp. 26172-26189, 2020, doi: 10.1109/ACCESS.2020.2971087.
Textual Analysis for Online Reviews: A Polymerization on Topic Sentiment Model	L. Huang, Z. Dou, Y. Hu and R. Huang	More and more e-commerce companies realize the importance of analyzing the online reviews of their products. It is believed that online review has a significant impact on the shaping product brand and sales promotion. In this paper, we proposed a polymerization topic sentiment model (PTSM) to conduct textual analysis for online reviews. We applied this model	Review sentiment analysis	L. Huang, Z. Dou, Y. Hu and R. Huang, "Textual Analysis for Online Reviews: A Polymerization Topic Sentiment Model," in IEEE Access, vol. 7, pp. 91940-91945, 2019, doi: 10.1109/ACCESS.2019.2920091.

		to extract and filter the sentiment information from online reviews. Through integrating this model with machine learning methods, the results showed that the prediction accuracy had improved. Also, the experimental results showed that filtering sentiment topics hidden in the reviews are more important in influencing sales prediction, and the PTSM is more precise than other methods. The findings of this paper contribute to the knowledge that filtering the sentiment topics of online reviews could improve the prediction accuracy. Also, it could be applied by e-commerce practitioners as a new technique to conduct analyses of online reviews.	Textual analysis	
Chinese Text Sentiment Analysis Based on Extended Sentiment Dictionary	G. Xu, Z. Yu, H. Yao, F. Li, Y. Meng and X. Wu	The method of text sentiment analysis based on sentiment dictionary often has the problems that the sentiment dictionary doesn't contain enough sentiment words or omits some field sentiment words. In addition, due to the existence of some polysemic sentiment words with positivity, negativity, and neutrality, the words' polarity cannot be accurately expressed, so the accuracy of text sentiment analysis is reduced to some extent. In this paper, an extended sentiment dictionary is constructed. The extended sentiment dictionary contains the basic sentiment words, the field sentiment words, and the polysemic sentiment words, which improves the accuracy of sentiment analysis. The naive Bayesian classifier is used to determine the field of the text in which the polysemic sentiment word is. Thus, the sentiment value of the polysemic sentiment word in the field is obtained. By utilizing the extended sentiment dictionary and the designed sentiment score rules, the sentiment of the text is achieved. The experimental results prove that the proposed sentiment analysis method based on extended sentiment dictionary has certain feasibility and accuracy. The research is meaningful for the sentiment recognition of the comment texts.	Sentiment analysis Sentiment dictionary Classification	G. Xu, Z. Yu, H. Yao, F. Li, Y. Meng and X. Wu, "Chinese Text Sentiment Analysis Based on Extended Sentiment Dictionary," in IEEE Access, vol. 7, pp. 43749-43762, 2019, doi: 10.1109/ACCESS.2019.2907772.

Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning	L. Yang, Y. Li, J. Wang and R. S. Sherratt	In recent years, with the rapid development of Internet technology, online shopping has become a mainstream way for users to purchase and consume. Sentiment analysis of many user reviews on e-commerce platforms can effectively improve user satisfaction. This paper proposes a new sentiment analysis model-SLCABG, which is based on the sentiment lexicon and combines Convolutional Neural Network (CNN) and attention-based Bidirectional Gated Recurrent Unit (BiGRU). In terms of methods, the SLCABG model combines the advantages of sentiment lexicon and deep learning technology and overcomes the shortcomings of existing sentiment analysis model of product reviews. The SLCABG model combines the advantages of the sentiment lexicon and deep learning techniques. First, the sentiment lexicon is used to enhance the sentiment features in the reviews. Then the CNN and the Gated Recurrent Unit (GRU) network are used to extract the main sentiment features and context features in the reviews and use the attention mechanism to weight. And finally classify the weighted sentiment features. In terms of data, this paper crawls and cleans the real book evaluation of dangdang.com, a famous Chinese e-commerce website, for training and testing, all of which are based on Chinese. The scale of the data has reached 100000 orders of magnitude, which can be widely used in the field of Chinese sentiment analysis. The experimental results show that the model can effectively improve the performance of text sentiment analysis	Sentiment analysis Sentiment lexicon Classification	L. Yang, Y. Li, J. Wang and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," in IEEE Access, vol. 8, pp. 23522-23530, 2020, doi: 10.1109/ACCESS.2020.2969854.
Automatic Expressive Opinion Sentence Generation for Enjoyable Conversational Systems	Y. Matsuyama, A. Saito, S. Fujie and T. Kobayashi	In terms of functional conversations, Grice's Maxim of Quantity suggests that responses should contain no more information than was explicitly asked for. However, in our daily conversations, more informative response skills are usually employed in order to hold enjoyable conversations with interlocutors. These responses are usually produced as forms of one's additional opinions, which usually contain their original viewpoints as well as novel means of expression, rather than simple and common responses characteristic of the general public. In this paper, we propose automatic expressive opinion sentence	Sentence generation Paraphrase generation Generated opinions	Y. Matsuyama, A. Saito, S. Fujie and T. Kobayashi, "Automatic Expressive Opinion Sentence Generation for Enjoyable Conversational Systems," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 2, pp. 313-326, Feb. 2015, doi: 10.1109/TASLP.2014.2363589.

		generation mechanisms for enjoyable conversational systems. The generated opinions are extracted from many reviews on the web, and ranked in terms of contextual relevance, length of sentences, and amount of information represented by the frequency of adjectives. The sentence generator also has an additional phrasing skill. Three controlled lab experiments were conducted, where subjects were requested to read generated sentences and watch videos filmed about conversations between the robot and a person. The results implied that mechanisms effectively promote users' enjoyment and interests.		
A Review of Text-Based Recommendation Systems	S. Kanwal, S. Nawaz, M. K. Malik and Z. Nawaz	Many websites over the Internet are producing a variety of textual data; such as news, research articles, ebooks, personal blogs, and user reviews. In these websites, the textual data is so large that the process of finding pertinent information by a user often becomes cumbersome. To overcome this issue, “Text-based Recommendation Systems (RS)” are being developed. They are the systems with the capability to find the relevant information in a minimal time using text as the primary feature. There exist several techniques to build and evaluate such systems. And though a good number of surveys compile the general attributes of recommendation systems, there is still a lack of comprehensive literature review about the text-based recommendation systems. In this paper, we present a review of the latest studies on text-based RS. We have conducted this survey by collecting literature from preeminent digital repositories, that was published during the period 2010-2020. This survey mainly covers the four major aspects of the textual based recommendation systems used in the reviewed literature. The aspects are datasets, feature extraction techniques, computational approaches, and evaluation metrics. As benchmark datasets carry a vital role in any research, publicly available datasets are extensively reviewed in this paper. Moreover, for text-based RS many proprietary datasets are also used, which are not available in the public. But we have consolidated all the attributes of these publically available and proprietary datasets to familiarize these attributes to new researchers. Furthermore, the feature extraction	Recommendation system Feature extraction	S. Kanwal, S. Nawaz, M. K. Malik and Z. Nawaz, "A Review of Text-Based Recommendation Systems," in IEEE Access, vol. 9, pp. 31638-31661, 2021, doi: 10.1109/ACCESS.2021.3059312.

		methods from the text are briefed and their usage in the construction of text-based RS are discussed. Later, various computational approaches that use these features are also discussed. To evaluate these systems, some evaluation metrics are adopted. We have presented an overview of these evaluation metrics and diagramed them according to their popularity. The survey concludes that Word Embedding is the widely used feature selection technique in the latest research. The survey also deduces that hybridization of text features with other features enhance the recommendation accuracy. The study highlights the fact that most of the work is on English textual data, and News recommendation is the most popular domain.		
Recommendation System Based on Heterogeneous Feature: A Survey	H. Wang, Z. Le and X. Gong	Recommendation systems have become an important field of research in computer science and physics. In recent years, breakthroughs have been achieved in social, biological, and research cooperation networks. With the popularization of big data and deep learning technology development, graph structures are increasingly being used to represent large-scale and complex data in the real world. In this paper, we reviewed the progress made in recommendation systems research in the past 20 years and comprehensively classified recommender systems based on the heterogeneous input features. We introduced layering in the classification of recommendation systems. Furthermore, we proposed a new hierarchical classification model of recommendation systems divided into three layers: feature input, feature learning, and output layers. In the feature learning layer, existing recommendation systems were divided into graph-based, text-based, behavior-based, spatiotemporal-based, and hybrid recommendation systems. Additionally, we provided evaluation index, open-source implementation, experimental comparison and the relative merits for each recommendation method. Subsequently, future development directions of recommendation systems are discussed.	Text-based recommendation system Classification	H. Wang, Z. Le and X. Gong, "Recommendation System Based on Heterogeneous Feature: A Survey," in IEEE Access, vol. 8, pp. 170779-170793, 2020, doi: 10.1109/ACCESS.2020.3024154.

User-Oriented Paraphrase Generation with Keywords Controlled Network	D. Zeng, H. Zhang, L. Xiang, J. Wang and G. Ji	Paraphrase generation can help with both downstream tasks in natural language processing (NLP) and human writing in our daily life. Most of the prevalent neural models focus on the former usages and generate uncontrolled paraphrase while they ignore the subtleties of users' requirement. In addition, the existing tools for users are usually rule-based which is unnatural due to the complexity of the paraphrase nature. To this end, we propose a keyword-controlled network (KCN) which can be used as an assistant paraphrase generation tool. The KCN works in an interactive manner and generates different paraphrases given different keywords. The model is based on a Sequence-to-Sequence (Seq2Seq) framework integrated with copy mechanism. Given the source sentence and the keywords, two encoders transform them into vector representations. Then, the representations are fused together and used for the decoder. The decoder with attention mechanism either copies the words from the keywords or generates words from the whole dictionary. In the training stage, as the source sentence and the target sentence are all valid paraphrases, the model is trained to generate each given different keywords, which simulates the behaviors of users. The extensive experiments on three datasets show that our method outperforms baselines in the automatic evaluation (0.06 absolute improvement in BLEU) and the generated paraphrases meet user expectation in the human evaluation.	Paraphrase generation Natural language processing	D. Zeng, H. Zhang, L. Xiang, J. Wang and G. Ji, "User-Oriented Paraphrase Generation with Keywords Controlled Network," in IEEE Access, vol. 7, pp. 80542-80551, 2019, doi: 10.1109/ACCESS.2019.2923057.
A System for Automatic English Text Expansion	S. García-Méndez, M. Fernández-Gavilanes, E. Costa-Montenegro, J. Juncal-Martínez, F. J. González-Castaño and E. Reiter	We present an automatic text expansion system to generate English sentences, which performs automatic Natural Language Generation (NLG) by combining linguistic rules with statistical approaches. Here, “automatic” means that the system can generate coherent and correct sentences from a minimum set of words. From its inception, the design is modular and adaptable to other languages. This adaptability is one of its greatest advantages. For English, we have created the highly precise aLexiE lexicon with wide coverage, which represents a contribution on its own. We have evaluated the resulting NLG library in an Augmentative and Alternative Communication (AAC)	Sentence generation Natural language generation Lexicon	S. García-Méndez, M. Fernández-Gavilanes, E. Costa-Montenegro, J. Juncal-Martínez, F. J. González-Castaño and E. Reiter, "A System for Automatic English Text Expansion," in IEEE Access, vol. 7, pp. 123320-123333, 2019, doi: 10.1109/ACCESS.2019.2937505.

		proof of concept, both directly (by regenerating corpus sentences) and manually (from annotations) using a popular corpus in the NLG field. We performed a second analysis by comparing the quality of text expansion in English to Spanish, using an ad-hoc Spanish-English parallel corpus. The system might also be applied to other domains such as report and news generation.		
Sentiment Analysis by Capsules	Yequan Wang, Aixin Sun, Jialong Han, Ying Liu, Xiaoyan Zhu	In this paper, we propose RNN-Capsule, a capsule model based on Recurrent Neural Network (RNN) for sentiment analysis. For a given problem, one capsule is built for each sentiment category e.g., 'positive' and 'negative'. Each capsule has an attribute, a state, and three modules: representation module, probability module, and reconstruction module. The attribute of a capsule is the assigned sentiment category. Given an instance encoded in hidden vectors by a typical RNN, the representation module builds capsule representation by the attention mechanism. Based on capsule representation, the probability module computes the capsule's state probability. A capsule's state is active if its state probability is the largest among all capsules for the given instance, and inactive otherwise. On two benchmark datasets (i.e., Movie Review and Stanford Sentiment Treebank) and one proprietary dataset (i.e., Hospital Feedback), we show that RNN-Capsule achieves state-of-the-art performance on sentiment classification. More importantly, without using any linguistic knowledge, RNN-Capsule is capable of outputting words with sentiment tendencies reflecting capsules' attributes. The words well reflect the domain specificity of the dataset	Sentimental Analysis Classification	Yequan Wang, Aixin Sun, Jialong Han, Ying Liu, and Xiaoyan Zhu. 2018. Sentiment Analysis by Capsules. In Proceedings of the 2018 World Wide Web Conference (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1165–1174. https://doi.org/10.1145/3178876.3186015
Customer Sentiment in Web-Based Service Interactions: Automated Analyses and New Insights	Galit B. Yom-Tov, Shelly Ashtar, Daniel Altman, Michael Natapov, Neta Barkay, Monika Westphal, Anat Rafaeli, Authors Info & Claims	We adjust sentiment analysis techniques to automatically detect customer emotion in on-line service interactions of multiple business domains. Then we use the adjusted sentiment analysis tool to report insights about the dynamics of emotion in on-line service chats, using a large data set of Telecommunication customer service interactions. Our analyses show customer emotions starting out negative and evolving into positive as the interaction ends. Also, we identify a close relationship between customer emotion dynamics	Sentiment Lexicon Sentiment analysis	Galit B. Yom-Tov, Shelly Ashtar, Daniel Altman, Michael Natapov, Neta Barkay, Monika Westphal, and Anat Rafaeli. 2018. Customer Sentiment in Web-Based Service Interactions: Automated Analyses and New Insights. In Companion Proceedings of the The Web Conference 2018 (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1689–1697.

		during the service interaction and the concepts of service failure and recovery. This connection manifests in customer service quality evaluationsafter the interaction ends. Our study shows the connection between customer emotion and service quality as service interactions unfold and suggests the use of sentiment analysis tools for real-time monitoring and control of web-based service quality.		https://doi.org/10.1145/3184558.3191628
An Italian lexicon-based sentiment analysis approach for medical applications	Maria Chiara Martinis, Chiara Zucco, Mario Cannatar, Authors Info & Claims	Sentiment analysis aims at extracting opinions and or emotions mainly from written text. The most popular problem in sentiment analysis certainly is polarity detection, which falls into the broader class of Natural Language Processing (NLP) problems of text classification. To date, state-of-the-art approaches to text classification use neural language models built on popular architectures such as Transformers. However, these approaches are difficult to apply in low-resource languages and domains, as for instance the Italian language or small clinical trials. Motivated by this, this paper presents VADER-IT, a lexicon-based algorithm for polarity prediction in written text, that is an adaptation to the Italian language of the popular VADER. Unlike VADER, our system also predicts a polarity class (i.e. positive, negative or neutral). The system was tested on a dataset of 5495 healthcare related reviews from QSalute https://www.qsalute.it/, reaching a micro averaged F1--score = 81% and a micro averaged Jaccard -score = 73%.	Sentiment Lexicon Natural language processing	Maria Chiara Martinis, Chiara Zucco, and Mario Cannataro. 2022. An Italian lexicon-based sentiment analysis approach for medical applications. In Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '22). Association for Computing Machinery, New York, NY, USA, Article 70, 1–4. https://doi.org/10.1145/3535508.3545594
Generating focused topic-specific sentiment lexicons	Valentin Jijkoun, Maarten de Rijke, Wouter Weerkamp	We present a method for automatically generating focused and accurate topic-specific subjectivity lexicons from a general-purpose polarity lexicon that allow users to pin-point subjective on-topic information in a set of relevant documents. We motivate the need for such lexicons in the field of media analysis, describe a bootstrapping method for generating a topic-specific lexicon from a general-purpose polarity lexicon, and evaluate the quality of the generated lexicons both manually and using a TREC Blog track test set for opinionated blog post retrieval. Although the generated lexicons can be an order of magnitude more selective than the general-purpose lexicon, they maintain, or even improve, the	Building lexicons Opinion retrieval system	Valentin Jijkoun, Maarten de Rijke, and Wouter Weerkamp. 2010. Generating focused topic-specific sentiment lexicons. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10). Association for Computational Linguistics, USA, 585–594.

		performance of an opinion retrieval system.		
Automatic sentiment classification of product reviews using maximal phrases-based analysis	Maria Tchalakova, Dale Gerfrmann, Detnas Meurers	In this paper we explore the use of phrases occurring maximally in text as features for sentiment classification of product reviews. The goal is to find in a statistical way representative words and phrases used typically in positive and negative reviews. The approach does not rely on predefined sentiment lexicons, and the motivation for this is that potentially every word could be considered as expressing something positive and/or negative in different situations, and that the context and the personal attitude of the opinion holder should be considered when determining the polarity of the phrase, instead of doing this out of context.	Sentiment lexicons Classification	Maria Tchalakova, Dale Gerdemann, and Detmar Meurers. 2011. Automatic sentiment classification of product reviews using maximal phrases-based analysis. In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA '11). Association for Computational Linguistics, USA, 111–117.
Generating a Context-Aware Sentiment Lexicon for Aspect-Based Product Review Mining	Jurgen Bross, Heiko Ehrig, Authors Info & Claims	A great share of current sentiment analysis techniques is based on special purpose lexicons providing information about the semantic orientation (e.g., positive, negative, neutral) of its entries. Due to the high labor costs of manually assembling such resources, recent work has focused on automatically inducing the polarity of given terms. We follow this line of work while focusing on the domain of user-generated product reviews, a main field of application for sentiment analysis. In this domain, a major observation is that the semantic orientation of terms is often context-dependent which poses an additional challenge to the automatic construction of such lexicons (e.g. positive: “longbattery life” vs. negative: “long shutter lag time”). We propose a novel unsupervised method to induce a context-aware sentiment lexicon by utilizing the semi-structuredness of user-generated product reviews.	Generating a Context Sentiment Lexicon	Jurgen Bross and Heiko Ehrig. 2010. Generating a Context-Aware Sentiment Lexicon for Aspect-Based Product Review Mining. In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01 (WI-IAT '10). IEEE Computer Society, USA, 435–439. https://doi.org/10.1109/WI-IAT.2010.56

Building Sentiment Lexicons for All Major Languages	Yanqing Chen and Steven Skiena.	Sentiment analysis in a multilingual world remains a challenging problem, because developing language-specific sentiment lexicons is an extremely resourceintensive process. Such lexicons remain a scarce resource for most languages. In this paper, we address this lexicon gap by building high-quality sentiment lexicons for 136 major languages. We integrate a variety of linguistic resources to produce an immense knowledge graph. By appropriately propagating from seed words, we construct sentiment lexicons for each component language of our graph. Our lexicons have a polarity agreement of 95.7% with published lexicons, while achieving an overall coverage of 45.2%. We demonstrate the performance of our lexicons in an extrinsic analysis of 2,000 distinct historical figures' Wikipedia articles on 30 languages. Despite cultural difference and the intended neutrality of Wikipedia articles, our lexicons show an average sentiment correlation of 0.28 across all language pairs.	Building lexicons Sentiment analysis	Yanqing Chen and Steven Skiena. 2014. Building Sentiment Lexicons for All Major Languages. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 383–389, Baltimore, Maryland. Association for Computational Linguistics.