UNIVERSIDAD TORCUATO DI TELLA

MASTER IN MANAGEMENT + ANALYTICS

EXTRACCIÓN DE PATRONES EN LAS RESEÑAS SOBRE CELULARES MEDIANTE EL MODELADO DE TEMAS Y EL ANÁLISIS DE SENTIMIENTOS

TESIS

Consuelo Nazar Anchorena .

Mayo 2022.

Tutor: Luca Rabbione.

Resumen

En la era digital, las redes sociales han cambiado la forma de comunicarnos: las mismas se convirtieron en una fuente de información e intercambio fundamental. El contenido que se genera en ellas requiere ser analizado mediante la aplicación de diversas técnicas de procesamiento del lenguaje natural, con el propósito de encontrar tendencias o patrones en las opiniones y comportamientos de las personas. Dicho análisis, le permite a las distintas áreas de las organizaciones enfocar sus esfuerzos en desarrollar estrategias que busquen la satisfacción de los consumidores, así como también que les permita posicionar sus propuestas y productos.

Este estudio se centra en la identificación de las dimensiones claves relacionadas con la compra de teléfonos móviles a través de internet. Específicamente nos basamos en información recolectada de Mercado Libre, ya que es un comercio electrónico que contiene un gran volumen de datos. En primer lugar, extrajimos los datos de las reseñas de la categoría "Celulares y Teléfonos" y realizamos un preprocesamiento de los mismos, que incluyó la eliminación de palabras vacías, la normalización y tokenización de los datos. Luego, para comenzar a comprender las razones en las cuáles los consumidores se basan para realizar sus elecciones, aplicamos métodos de aprendizaje no supervisado, que incluyeron la extracción de los cinco tópicos principales, utilizando la transformación del texto a una bolsa de palabras (en inglés, *bag of words*) y el método de Asignación latente de Dirichlet (LDA). También lo complementamos con técnicas de análisis de sentimiento, que están enfocadas en comprender las diversas palabras y expresiones que los seres humanos utilizamos para expresar nuestro grado de aceptación hacia un tema o producto, de manera de poder convertir las emociones en información objetiva.

Adicionamos a lo mencionado anteriormente, métodos de aprendizaje supervisado para aprovechar la información contenida en las etiquetas, es decir, en los puntajes de las reseñas. Para ello utilizamos una combinación de dos tipos de enfoques para extraer características: el enfoque de la bolsa de palabras previamente mencionado y *TF-IDF* (del inglés *Term frequency – Inverse document frequency*, frecuencia de término – frecuencia inversa de documento). Luego, entrenamos y evaluamos algoritmos de clasificación capaces de predecir los puntajes, de manera tal que puedan darnos una valoración social lo más acertada posible. Nos enfocamos en cuatro modelos de clasificación: *Random Forest* (en español, Bosque Aleatorio), *Support Vector Machine* (en español, Máquinas de Vector Soporte), *Naive Bayes* (en español, Bayes Ingenuo) y *Logistic Regression* (en español, Regresión Logística). Los resultados del estudio encuentran implicaciones prácticas para el desarrollo de los celulares, ya que permiten hacer foco en los tópicos y aspectos clave en los que los consumidores se basan para hacer sus elecciones.

Abstract

In the digital age, social networks have changed the way we communicate: they are a fundamental source of information and a place to exchange ideas. The content generated in them needs to be analyzed through the application of natural language processing techniques, in order to find trends and patterns in people's opinions and behaviors. This allows the different areas of organizations to focus their efforts on developing strategies that seek consumer satisfaction and also that help them position their proposals and products.

This study focuses on the identification of the key dimensions related to the purchase of mobile phones through the Internet, specifically we rely on information collected from Mercado Libre, since it is an electronic commerce that contains a large volume of data. First, we extracted the data from the reviews of the "Cell Phones" category and pre-processed them, which included the elimination of stop words, normalization and tokenization of the data. Then, to begin to understand the reasons consumers base their choices on, we applied unsupervised learning methods, which include extracting the top five topics using bag-of-words transformation and the Latent Dirichlet Assignment (LDA) method. We also complemented it by sentiment analysis techniques, which are focused on understanding the various words and expressions that human beings use to express our degree of acceptance towards a topic or product, in order to convert emotions into objective information.

In addition to the aforementioned, supervised learning methods were used to take advantage of the information contained in the labels, that is, in the review scores. To do this, we used a combination of two types of approaches to extract features: the bag of words approach and TF-IDF (Term Frequency – Inverse document frequency). Then, classification algorithms capable of predicting the scores were trained and evaluated, in such a way that they can give us a social evaluation that is as accurate as possible. We focus on four classification models that are considered the most appropriate for our problem, a Random Forest, a Support Vector Machine, a Naive Bayes and a Logistic Regression. The results of the study find practical implications for the development of cell phones, because they allow to focus on the topics and key aspects on which consumers base their choices.

Índice

Índice de Tablas

Índice de Figuras

1. Introducción

1.1. Contexto

Comercio electrónico

Las empresas de comercio electrónico, o también conocido como *e-commerce*, se encargan de la compra y venta de productos o servicios a través de medios digitales, como páginas web, aplicaciones móviles y redes sociales. Esta manera de comercializar ha ido cambiando la forma en la que, tanto los vendedores como compradores, interaccionan y se comunican, ya que ha estimulado el pago electrónico, el marketing digital, las distintas maneras de administrar los procesos involucrados en las necesidades de suministro y el intercambio de datos electrónicos.

El comercio electrónico se encuentra en constante crecimiento como una consecuencia de la revolución digital y se ha visto enormemente acelerado por la pandemia de COVID-19, que provocó que entren personas que quizás no lo habían probado nunca o que adquirían productos en el comercio electrónico pero con poca frecuencia (Yujnovsky&Asoc., 2020).

La popularidad de estas nuevas plataformas y tendencias han contribuido ampliamente a desarrollar un nuevo paradigma en los hábitos de compra: la desaparición del clásico proceso de compra lineal y la aparición de un proceso de decisión de compra más circular, iterativo e influido por nuestro círculo social (Bear, D., & Szabo, M., 2012). Bajo este paradigma, en lugar de desarrollarse un proceso de compra lineal desde el conocimiento hasta la compra, en el que los consumidores seleccionan entre un conjunto de opciones para decidirse entre un pequeño grupo y comprar una de ellas, el proceso de añadir marcas y productos a las primeras posiciones de consideración es mucho más extenso y se modifica continuamente. Influidos por la familia, amigos, vecinos y por lo dicho en redes sociales, el proceso de decisión de compra es mucho más circular y en él cobra importancia la experiencia tras la compra y la prescripción de "colegas compradores" (Barrio, J., 2017).

Las personas intercambian sus pensamientos a través de diferentes redes sociales, blogs, y foros web. En las compras en línea brindan opiniones y reseñas que sirven no solo como decisión de compra para otros usuarios sino que también ayudan a mejorar la calidad de los productos o servicios y permiten descubrir diversos problemas o perspectivas para las empresas, como por ejemplo, qué les gusta a los consumidores y qué no, por qué compran, qué características prefieren, cuál ha sido su experiencia de compra.

Por todo lo mencionado anteriormente, mapear y explicar la experiencia de los clientes o los problemas con una marca, producto o servicio se volvió vital, es materia prima fundamental ya que tiene una gran amplitud de información que impulsa las ventas en línea y las tasas de conversión más altas.

Mercado Libre

Mercado Libre es la plataforma líder en comercio electrónico de América Latina. Fue fundada en 1999, y tiene sus orígenes en Argentina. Hoy en día cuenta con operaciones diarias en Argentina, Bolivia, Brasil, Chile, Colombia, Costa Rica, Ecuador, Guatemala, México, Panamá, Perú, Portugal, República Dominicana, Uruguay y Venezuela. La misma está compuesta por MercadoPago, para transacciones de dinero, MercadoShops que es la plataforma de E-commerce para PyMES y MercadoEnvíos como servicio de mensajería puerta a puerta en convenio con empresas de correo locales.

Es la firma más valiosa en Argentina. En el tercer trimestre del año 2021 informó que tuvo ingresos netos por US$1900 millones, con un incremento del 72,9% en comparación con igual período de 2020, para lo que en ese momento, según su cotización en Wall Street, la empresa tenía una valuación de US$77.485 millones. La acelerada adopción de plataformas digitales producto de la pandemia y el cambio de hábitos en los consumidores, le dieron una oportunidad a Mercado Libre de potenciar su negocio. Según datos de la propia empresa, entre Julio y Septiembre del 2021, su base de usuarios activos creció un 3,4% interanual, y llegó a los 78,7 millones de personas en la región. A su vez, en esos tres meses se vendieron 259,8 millones de artículos (un 26,3% más que en el tercer trimestre de 2020), y se procesaron 865,7 millones de transacciones (una suba del 54,7%) (Herrera, E., 2021).

Reseñas en línea

Las reseñas u opiniones de consumidores en línea, forma digital de comunicarse de “boca a boca”, se han convertido en una fuente de información insustituible en la toma de decisiones de otros consumidores (Filieri, R., & Mariani, M., 2021). Los clientes que han experimentado los productos o servicios describen o eligen puntuaciones. Dichas puntuaciones, contienen información que se percibe como más confiable y objetiva que la información proporcionada por las propias empresas, lo que les permite a los usuarios crear una imagen de las mismas y de lo que éstas ofrecen previa a su compra.

A diferencia de las revisiones generales, las revisiones en línea son accesibles las 24 horas del día, lo que permite el almacenamiento continuo de información en forma de texto o imágenes (Lee, B. C., & Byun, H. J., 2014). Por lo que se ha ampliado la cobertura y la difusión de la información que se da de forma rápida, lo que significa que los clientes no tienen limitaciones de tiempo ni espacio para escribirlas y brindar su opinión.

El mercado de la telefonía celular

El mundo contemporáneo o era informacional exige circulación permanente de mercaderías, deseos, valores, ideas, tecnologías, culturas, modas, mitos: la sociedad postmodernista ha arrasado el mito de lo permanente, de lo estático e inmutable. Las nuevas tecnologías han triturado los viejos esquemas de la vieja comunicación sedentaria, y, consecuentemente, han impuesto a la humanidad nuevos ritmos en las relaciones sociales y productivas, se ha desplazado de lo estático a lo dinámico (Flórez Calderón, M., & Ramos, Z., 1993). Este marco es clave para comprender la revolución que existe en las comunicaciones móviles en general y más específicamente en la telefonía celular, que exige que los vendedores estén buscando formas de diferenciar sus productos de los de la competencia constantemente.

El negocio celular ha tenido un crecimiento exponencial en todo el mundo y las tecnologías móviles e inalámbricas se están generalizando cada vez más. Los teléfonos móviles que alguna vez se consideraron un lujo, ahora están reemplazando a los teléfonos convencionales en el uso residencial. Las redes inalámbricas liberan a los usuarios de las ataduras que los tienen atados a su escritorio, lo que les permite vivir y trabajar de manera más flexible y conveniente (Kumar, S., & Zahn, C., 2003).

Particularmente este mercado en Argentina ha crecido a pasos agigantados durante los últimos años y se espera que así lo siga haciendo. En el 2021, se estimó que alrededor de 34,8 millones de argentinos eran usuarios de algún tipo de teléfono móvil inteligente (considerando individuos de todas las edades que tenían un smartphone y que lo usaban mínimo una vez al mes), lo que representaría un incremento de 1,6 millones en comparación con la cifra para 2020. Se estima que para 2025, el número de usuarios de smartphones en Argentina supere los 40 millones (Statista Research Department., 2021). *Ver Anexo A.*

Procesamiento del lenguaje natural: modelado de tópicos y análisis de sentimiento

El procesamiento del lenguaje natural es un subcampo del aprendizaje automático que ha ganado mucha popularidad; es una técnica para que las computadoras aprendan cómo funciona el lenguaje humano con el uso de la lingüística computacional y el dominio de la informática.

El modelado de tópicos, en inglés *topic modeling*, es una de las técnicas más poderosas en la minería de textos. Se basa en el descubrimiento de datos latentes y la búsqueda de relaciones entre datos y documentos de texto. Existen varios métodos para el modelado de temas: la asignación de Dirichlet latente (LDA) es una de las más populares en este campo (Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L., 2019). La suposición básica de LDA es que todos los documentos están representados por distribución de temas y todos los temas por distribución de palabras.

Otra herramienta muy importante en el procesamiento del lenguaje natural es el análisis de sentimientos. Consiste en una metodología para determinar los sentimientos, reacciones emocionales y perspectivas de los clientes hacia los bienes, marcas o servicios (Khan, M. T., Durrani, M., Ali, A., Inayat, I., Khalid, S., & Khan, K. H., 2016). El análisis de sentimiento detecta cambios en la opinión pública de una marca que pueden apoyar o contradecir la estrategia de la empresa. Cuando los equipos deben reformarse o establecerse nuevas tácticas inventivas, las calificaciones de sentimiento lo indican.

Hoy en día las empresas se enfrentan a la mercantilización de productos, así como a clientes que están más informados y son más exigentes que nunca antes. Tratar de competir administrando los costos ya no es suficiente. Se requiere utilizar estrategias para poder cumplir con la gran demanda de los consumidores de los comercios electrónicos, por eso la evaluación de los productos por parte de los usuarios es un buen criterio para definir la base de una estrategia de management.

1.2. Problema

En los últimos años, se ha vuelto muy importante que todas las empresas que ofrecen, tanto productos como servicios, tengan presencia en el mundo virtual; ya sea publicando lo que ofrecen en un comercio electrónico o a través de su propia página web. Los sitios web comenzaron a ser el principal medio de compra y uno de los motivos centrales reside en el acceso a la información que tienen los consumidores antes de realizar la compra (Barrio, J., 2017). Una parte central de esa información viene dada por los comentarios, reseñas, valoraciones o *reviews*, que dejan los distintos compradores narrando sus experiencias con el producto, servicio, negocio o experiencia de compra en general.

En la sociedad del intercambio de la era digital actual, a cada segundo se genera una gran cantidad de reseñas en línea. La velocidad de producción y el volumen de nuevas reseñas están incitando a compradores y vendedores constantemente. Cualquier producto publicado en línea con muchas críticas positivas le aporta gran legitimidad al mismo. Por el contrario, cualquier artículo que no

contenga reseñas genera desconfianza en los potenciales consumidores, ya que las personas confían en la experiencia de los demás.

Tanto las reseñas positivas como las negativas son de vital importancia e influyen en la empresa y los productos que ofrecen. En el caso de que la reseña sea positiva, fomentará que otros clientes interesados en este tipo de productos se decidan finalmente a comprar en esta empresa, sirviendo de escaparate a los potenciales consumidores. Por otro lado, si estos comentarios son negativos, sirven para mejorar los productos y para identificar deficiencias o necesidades que se demandan (Macario, A., 2018). En consecuencia, todas las empresas que buscan intensificar y mejorar sus negocios necesitan recolectar los datos recibidos en las redes y analizarlos, ya que les permite encontrar respuestas para impulsar el negocio y mejorar la toma de decisiones estratégicas para su crecimiento.

Se ha producido una sobrecarga informativa que ha permitido multiplicar las fuentes del saber: la información que circula en las redes es cada vez mayor y como consecuencia también evolucionaron los métodos, técnicas y aplicaciones para interpretar y procesar mejor los datos.

1.3. Objetivo

El enfoque principal de este trabajo es identificar patrones en las opiniones sobre los productos en línea del comercio electrónico, específicamente los productos de la sección "Celulares y Teléfonos" de la plataforma Mercado Libre. De manera que, los vendedores y fabricantes de celulares móviles, puedan tomar mejores decisiones utilizando las experiencias reales de otros clientes que han experimentado previamente sus productos. Las reseñas brindan la oportunidad de conocer cuáles son los problemas o aspectos del producto que hay que corregir, como también permiten saber cuáles son los puntos favorables o con los cuáles los clientes están satisfechos para poder potenciarlos. Por este motivo, se llevará adelante la construcción de modelos que utilizarán técnicas de procesamiento de lenguaje natural y de aprendizaje automático para el resumen y análisis de un gran volumen de opiniones.

Su implementación para resumir y analizar comentarios supondrá un importante aporte al sector de la venta online, ya que los resultados sobre la opinión y experiencia de usuario de los productos pueden ser de utilidad para potenciales compradores y vendedores. Dicho análisis permitirá entender la conversación desde el punto de vista emocional de los consumidores, permitirá conocer a la comunidad en profundidad, ayudará a detectar cuáles son los puntos que los deja conformes o disconformes a los seguidores para emprender campañas y servicios segmentados hacia ellos o

modificar y mejorar aquellos aspectos más débiles que pueden generar un impacto importante en las ventas.

Adicionalmente si se consigue que las críticas positivas sean mayores que las negativas, permitiría aumentar las conversiones ya que si el producto tiene buena calidad y los clientes expresan conformidad con los mismos cada vez más personas se animarán a comprar ese producto. Además el contenido generado por el usuario impulsa el *SEO* (en inglés, *Search Engine Optimization* lo que significa Optimización para motores de búsqueda), lo que puede generar mayor tráfico ya que las reseñas pueden generar que el negocio en línea tenga una clasificación más alta en los resultados de búsqueda.

2. Datos

Para este trabajo utilizamos datos que descargamos mediante el acceso a la interfaz de programación de aplicaciones de Mercado Libre[1] mediante una técnica para extraer información de sitios webs denominada *web scrapping*. La *API* se trata de un conjunto de definiciones y protocolos que se utiliza para desarrollar e integrar el software de las aplicaciones, permitiendo la comunicación entre dos aplicaciones de software a través de un conjunto de reglas (Rojas, A., 2021).

Nos centramos en la categoría de "Celulares y Teléfonos"[2], ya que es una de las categorías más deseadas y con más reseñas en sus publicaciones dentro de la web de Mercado Libre *(Ver Anexo B).*

Las subcategorías que se encuentran contenidas dentro son: "Accesorios para Celulares" ,"Celulares y Smartphones", "Handies y Radiofrecuencia", "Lentes de Realidad Virtual", "Repuestos de Celulares", "Smartwatches y Accesorios", "Tarifadores y Cabinas", "Telefonía Fija e Inalámbrica", "Telefonía IP" y "Otros". De todas las subcategorías, centramos nuestro análisis únicamente en la de "Celulares y Smartphones" que es la de nuestro interés y que además es la que mayor cantidad de reseñas contiene.

En total filtramos 34.076 reseñas sobre "Celulares y Smartphones" con sus respectivas variables asociadas que son: "item_key", "item_title", "review_key", "review_title", "review_content" y "review_rate".

[1] API Mercadolibre
[2] Codificada como "MLA 1051".

A continuación explicamos brevemente el contenido de las mismas.

- **item_key:** clave de identificación de cada ítem dentro de la categoría.
- **item_title:** título del ítem.
- **review_key**: clave de identificación de la reseña.
- **review_title:** título de la reseña.
- **review _content**: contenido de la reseña.
- **review_rate**: valoración en estrellas (de 1 a 5). Estos valores se pueden ver como etiquetas de opinión : valor 1 para "Muy negativo", 2 para "Algo negativo", 3 para "Neutral", 4 para "Algo positivo" y 5 para "Muy positivo".

A continuación podemos ver una tabla ejemplificativa de los datos.

Tabla 1. Muestra del dataset de Mercado Libre.

item_key	item_title	review_key	review_title	review_content	review_rate
MLA1111890165	Moto G60s 128 Gb Azul 6 Gb Ram	138048401	Bueno	En todo es un buen celular, lo malo que tiene es su tamaño es muy grande usaba uno de 5,5 pulgadas y se nota mucho la diferencia, ah la pantalla no es muy buena, se ve bien pero no tiene contraste, cualquier video con mucha iluminación se ve pálido. En todo lo demás es bueno, tiene 120hz es rápido, para juego corre todo.	3
MLA1111890165	Moto G60s 128 Gb Azul 6 Gb Ram	140419952	Recomendable, excelente relación precio/ calidad	Excelente, buen equipo, rápido y con muy buenas prestaciones de audio, vídeo y fotografía, nada que envidiarle a equipos que valen más del doble. Mejoraron la ubicación del lector de huella digital está puesto en un lugar muy accesible, respecto a versiones anteriores de motorola.	5
MLA1111890165	Moto G60s 128 Gb Azul 6 Gb Ram	140031283	Regular	Falla a veces, es muy incómodo el sistema de atajos tactil.	2
MLA915377519	Samsung Galaxy A52 128 Gb Awesome Black 6 Gb Ram	139231425	Malo	Este teléfono tiene un serio problema con las notificaciones. Aparentemente "mata" a las aplicaciones para "ahorrar" batería y resulta que no te llegan las notificaciones de apps de mensajería como whatsapp, por ejemplo. Perdí llamadas, mensajes, etc. Porque viene esa configuración por defecto. Es realmente inentendible cómo pueden sacar al mercado un dispositivo así, se supone que el celular está para comunicarse justamente. Encima sale casi un sueldo entero. Inaceptable. Primera y última vez que compro samsung.	2
MLA915377519	Samsung Galaxy A52 128 Gb Awesome Black 6 Gb Ram	107880328	Excelente	Excelente, la cámara genial, es muy lindo y el precio es adecuado.	5
MLA915377519	Samsung Galaxy A52 128 Gb Awesome Black 6 Gb Ram	139778700	No es lo que esperaba	Anda muy mal la conexión wifi, puede demorar horas.	3
MLA923362790	Samsung Galaxy S20 Fe 128 Gb Cloud Navy 6 Gb Ram	123752598	Excelente relación precio calidad y prestaciones	El celular es muy bueno. A nivel poder y velocidad va de 10. La cámara triple funciona muy bien y la frontal es excelente!. La batería me dura el día completo con uso intensivo. Por el precio es muy recomendable!.	4
MLA923362790	Samsung Galaxy S20 Fe 128 Gb Cloud Navy 6 Gb Ram	135991852	Malo, no fue lo esperado	No fue lo que esperaba, tuve que desactivar muchas opciones para que la bateria dure un poco más, porque sino tenía que cargarlo hasta 2 veces por día. Una decepción. No lo devolví porque no puedo quedarme sin celular, pero no lo recomiendo para nada.Además, si usas wifi, tiene tan poca señal que por defecto te usa datos. Y en algunas redes sociales como instagram, al ingresar se queda pensando por un rato largo hasta que carga. No lo compren, hay mejores modelos u otras marcas. Yo vengo comprando siempre samsung, pero despues de esto en el próximo cambiare de marca.	1
MLA923362790	Samsung Galaxy S20 Fe 128 Gb Cloud Navy 6 Gb Ram	130734961	Excelente	Una de las mejores camaras del mercado incluso en generacion actual, y por muchisimo la mejor en relacion precio calidad, te lo dice un fotografo.La pantalla es excelente en tamaño, peso, luminosidad y representación de colores.	5
MLA1111361645	Motorola Edge 20 Pro 256 Gb Blanco Optic 12 Gb Ram	132567460	Excelente	Muy bueno. Linda imgen buena resolución. Buen sonido sistema operativo muy bueno.	5
MLA1111361645	Motorola Edge 20 Pro 256 Gb Blanco Optic 12 Gb Ram	127112347	Buen equipo.	Buen aparato, la duración de la batería no es lo que esperaba.	4
MLA1111361645	Motorola Edge 20 Pro 256 Gb Blanco Optic 12 Gb Ram	140242374	Malo	Malo para el precio , yo tengo mil problema con el mio.	1
MLA1111361645	Motorola Edge 20 Pro 256 Gb Blanco Optic 12 Gb Ram	135606806	Excelente	Excelente, rápido, simple, luviano, muchas aplicaciones para recomendar.	5

3. Metodología

Comenzamos por extraer los datos de la categoría de nuestro interés mediante *web scraping* que es un proceso que permite recopilar información de forma automática de la web, en este caso, utilizamos la *API* del sitio web. Luego realizamos el preprocesamiento de los datos que obtuvimos, un paso preliminar durante el proceso de minería de datos. Incluye cualquier tipo de procesamiento que se le realice a los datos brutos para transformarlos en el formato que nos resulte más práctico para el análisis, en nuestro caso aplicamos la eliminación de palabras vacías, la normalización, la tokenización y la vectorización de palabras. Ésta última técnica se basa en la transformación de palabras en vectores para poder ingresar dichos vectores en nuestros modelos, y eso lo hicimos a través del método de bolsa de palabras (en inglés, *bag of words*) y de la transformación *TF-IDF* (en inglés, *term frequency – inverse document frequency*).

Luego, continuamos con la aplicación de algoritmos de aprendizaje no supervisados, que infieren patrones de un conjunto de datos sin etiquetas o sin referencia a resultados conocidos. Se utilizan para descubrir la estructura subyacente de los datos y agrupar datos no estructurados según sus similitudes y diferencias. El término "no supervisado" se refiere a la falta de una variable dependiente (y) en un modelo. Comenzamos con un análisis de tópicos, utilizando la técnica *Latent Dirichlet Allocation* (*LDA*) que busca construir temas en base a las distribuciones de palabras en un conjunto de documentos. Nos permitió comprender cuáles son los tópicos principales sobre los que suelen comentar los usuarios y nos ayudó a descubrir las estructuras semánticas latentes en un cuerpo de texto extenso. Adicionalmente, realizamos un análisis de sentimiento para comprender qué emociones o actitudes manifiestan los consumidores específicamente cuando mencionan los distintos temas identificados anteriormente y así identificamos el tono emocional detrás de una serie de palabras, si son comentarios positivos, negativos, neutros o cuál es la reacción de los clientes respecto a un producto, servicio o aspecto específico de los mismos.

En último lugar, complementamos las técnicas desarrolladas anteriormente con el entrenamiento de algoritmos de aprendizaje supervisado, que trabajan con los datos etiquetados y aprenden iterativamente de ellos. Para eso utilizamos el identificador numérico que representa el sentimiento de cada reseña, que puede ir desde el 1 que es el menor puntaje al 5 que es el mayor puntaje *(Ver Anexo C).* Por lo tanto, el problema se veía como un proceso de clasificación donde buscamos predecir la polaridad existente en los comentarios, es decir la valoración positiva o negativa en función de la intención del autor al momento de redactarlo, y además obtuvimos una valoración social de los mismos. Entrenamos cuatro algoritmos: *Random Forest* (en español, Bosque Aleatorio),

Support Vector Machine (en español, Máquinas de Vector Soporte), *Naive Bayes* (en español, Bayes Ingenuo) y *Logistic Regression* (en español, Regresión Logística). Luego predijimos el puntaje y evaluamos esas predicciones con distintas métricas, como son el *Accuracy, Precision, Recall, F1-score* y *AUC*, para poder identificar cuál de los cuatro algoritmos se adaptaba mejor a nuestros datos.

3.1. Recopilación de datos

Se recopilaron los datos de las reseñas en línea publicadas en Mercado Libre, para eso realizamos un llamado a su respectiva *API*[3]. De esa manera, descargamos la información necesaria para el análisis y la exploración del significado latente a través de los resultados de varias técnicas de minería de texto. Nos centramos en las reseñas de una sola categoría de las treinta y dos que posee la página web de Mercado Libre, que es la categoría de "Celulares y Teléfonos". Dentro de esa categoría, recuperamos hasta 100 comentarios por cada publicación para abarcar una variedad importante de ítems dentro de ella. A su vez, dicha categoría contiene distintas subcategorías pero filtramos los ítems de la subcategoría "Celulares y Smartphones" abarcando un total de 34.076 reseñas que fueron guardadas en un archivo CSV.

3.2. Preprocesamiento de datos

Los datos descargados los analizamos utilizando Python, pero previo a eso comenzamos con este paso de limpieza y preprocesamiento de los datos que es fundamental en la minería de texto. Se trata de todo el procesamiento de los datos brutos para transformarlos en datos que tengan la estructura o el formato necesario para poder aplicar luego los métodos. A los mismos los desglosamos en las siguientes etapas:

A) Concatenación del título y el contenido de la reseña

Dado que el título y el cuerpo de la reseña están divididos en dos columnas diferentes ("review_title" y "review_content") los unimos en una sola columna para que se transformen en el cuerpo de nuestro análisis.

[3] API Mercadolibre

B) Tokenización

Transformamos el texto de las reseñas concatenadas con su correspondiente título y tokenizamos cada palabra. Para lograr eso:

- Separamos el texto en oraciones.
- Separamos las oraciones en tokens (cada palabra es un token).

C) Lematización

La lematización es un proceso lingüístico que consiste en convertir cada *token* en su lema o significado raíz previsto. El lema es la forma que está aceptada como representante de todas las formas flexionadas de una misma palabra (Wikipedia, 2022).

D) Eliminación de palabras vacías

Las palabras vacías o en inglés *stopwords* son, básicamente, conjunciones, artículos, preposiciones y adverbios (Blackbeast, 2022) que se quitan de los documentos ya que no aportan información útil para nuestro análisis.

E) Normalización

La normalización es el proceso de organización de la base de datos para evitar la redundancia, optimizar el espacio de almacenamiento, proteger la integridad de los datos y reducir el tiempo y complejidad de la base de datos, para eso por ejemplo eliminamos las filas y campos duplicados, así como también corroboramos que los datos de las columnas sean del mismo tipo.

F) Vectorización: Bolsa de palabras

Transformamos todos los lemas en una representación de bolsa de palabras (*Bag of Words*, o *BoW*) para que nuestros modelos puedan procesar los datos. Para hacerlo, mapeamos las palabras de todo el cuerpo del texto en un diccionario con el que podamos contar las ocurrencias de cada palabra.

Una bolsa de palabras es una representación vectorial de texto donde cada elemento denota el número normalizado de ocurrencia de un término base en el documento. Para contar el número de ocurrencias de un término base, *BoW* realiza una coincidencia exacta de palabras, que puede considerarse como un mapeo duro de las palabras al término base (Zhao, R., & Mao, K., 2017).

Se le llama “bolsa” de palabras, porque se descarta cualquier información sobre el orden o la estructura de las palabras en el documento. Al modelo solo le interesa entender si las palabras

conocidas aparecen en el documento, no en qué parte del documento. La intuición es que los documentos son similares si tienen un contenido similar. Además, que solo del contenido podemos aprender algo sobre el significado del documento.

En nuestro caso armamos una matriz Documentos-términos, ya que tenemos en las filas los documentos y en las columnas los tokens. Los documentos se pueden pensar como vectores de 0 y 1 que indican la ausencia o presencia de las palabras.

3.3. Modelado de tópicos

El modelo de aprendizaje automático que usamos se llama asignación latente de Dirichlet (LDA), que es una de las técnicas más populares para el modelado de temas. Es un modelo probabilístico generativo que asume que cada tema es una combinación de un conjunto subyacente de palabras, y cada documento es una combinación de un conjunto de probabilidades de temas (Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L., 2019).

¿Cómo funciona LDA?

Hay 2 partes en LDA:

- Las palabras que pertenecen a un documento que ya conocemos.
- Las palabras que pertenecen a un tema o la probabilidad de que las palabras pertenezcan a un tema, que necesitamos calcular.

El algoritmo para encontrar este último:

- Revisa cada documento y asigna aleatoriamente cada palabra del documento a uno de los k temas (k se elige de antemano)(Kulshrestha, R., 2019). Para elegir el k se calculan modelos diferentes con un número creciente de temas y se escoge el mejor en términos de coherencia de los temas encontrados. La coherencia es una medida de cuán similares semánticamente son las palabras en el tema y nos proporciona un indicador de cuán interpretable es para un ser humano (Morstatter, F., & Liu, H., 2018).
- Para cada documento d , revisa cada palabra w y calcula:
    1) $p(tema\ t \mid documento\ d)$***:*** la proporción de palabras en el documento d que están asignadas al tema t. Intenta capturar cuántas palabras pertenecen al tema t para un documento dado d. Excluyendo la palabra actual, es decir que toma todas las palabras que pertenecen a un tema dado un documento sin contar la palabra en

cuestión. Si muchas palabras de d pertenecen a t , es más probable que la palabra w pertenezca a t.

$$\left(\frac{\#palabras\ en\ d\ con\ t + alpha}{\#palabras\ en\ d\ con\ cualquier\ t + k \times alpha}\right) \quad (1)$$

(Kulshrestha, R., 2019).

alpha determina cómo se distribuyen los temas en los documentos. Un valor más alto para alfa significa que un tema se distribuirá más ampliamente en los documentos, mientras que un valor más bajo para alfa significa que un tema se distribuirá de manera más estrecha en los documentos. En nuestro caso utilizamos el valor predeterminado para alfa que es "simétrico". Esto significa que es uniforme para cada tema. La fórmula que se utiliza para calcular el valor simétrico de alfa es dividir 1.0 por el número de temas del modelo. (Saxton, M., 2018).

2) $p(palabra\ w \mid tema\ t)$: la proporción de asignaciones al tema t sobre todos los documentos que provienen de esta palabra w. Intenta capturar cuántos documentos hay en el tema t debido a la palabra w .

LDA representa los documentos como una mezcla de temas. Del mismo modo, un tema es una mezcla de palabras. Si una palabra tiene una alta probabilidad de estar en un tema, todos los documentos que tengan w estarán más fuertemente asociados con t también. De manera similar, si no es muy probable que w esté en t , los documentos que contienen el w tendrán una probabilidad muy baja de estar en t, porque el resto de las palabras en d pertenecerán a algún otro tema y, por lo tanto, d tendrá una probabilidad más alta para esos temas. Entonces, incluso si se agrega w a t, no traerá muchos documentos de este tipo a t .

Para actualizar la probabilidad de que la palabra w pertenezca al tema t , usamos:

$$p(palabra\ w\ con\ tema\ t) = p(tema\ t \mid documento\ d) \times p(palabra\ w \mid tema\ t) \quad (2)$$

(Kulshrestha, R., 2019).

3.4. Análisis de sentimiento con métodos no supervisados

El Análisis de Sentimientos (AS), también conocido como Minería de Sentimientos o Análisis Subjetivo se define como la clasificación o cuantificación computacional de sentimientos y emociones expresadas en textos para convertirlos en información objetiva (Pang, B., & Lee, L., 2008).

Por lo que si el modelado de temas presentado anteriormente era una técnica de minería de texto que identifica el "objetivo cubierto por el texto", el análisis de sentimiento es una técnica de minería de texto que estima la "actitud contenida en el texto". Así como el modelado de temas extrae palabras que encarnan temas que se supone que son inherentes al texto y estima los temas, el análisis de sentimientos también estima los sentimientos inherentes al texto (Liu, B., 2012).

Los métodos no supervisados son algoritmos que basan su proceso de entrenamiento en datos sin clases previamente definidas o sin etiquetas. Es decir, que en principio no se conoce ningún valor objetivo o de clase, ya sea categórico o numérico. El aprendizaje no supervisado está dedicado a las tareas de agrupamiento o segmentación, donde su objetivo es encontrar grupos similares en el conjunto de datos.

Para realizar este estudio utilizamos una librería de análisis de sentimiento en español de Python que se llama “pysentimiento” que puede clasificar las emociones como positivas, negativas y neutrales. Dicha librería emplea modelos transformadores pre entrenados, redes neuronales que aprenden el contexto y, por lo tanto, el significado mediante el seguimiento de las relaciones en datos secuenciales (como las palabras de esta oración). En 2018 Google publicó un modelo transformador llamado *BERT (Bidirectional Encoders Representations for Transformers*) (Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., 2018). En este caso se usa la versión traducida al español de *BERT* llamada BETO (Canete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., & Pérez, J., 2020). Esta librería no solo realiza análisis de sentimiento, detectando comentarios positivos, negativos o neutros, sino que también puede hacer análisis de emociones (alegría, ira, miedo, tristeza, etc) y detectar lenguajes de odio. Estos dos últimos temas, si bien se mencionan, escapan al alcance del presente trabajo.

3.5. Modelos de clasificación de sentimiento con aprendizaje supervisado

Existen diferentes recursos para recopilar información que contiene opiniones, como correos electrónicos, cuestionarios en línea, publicaciones en redes sociales y publicaciones en blogs. Pero la cantidad de datos recopilados de estos recursos suele ser demasiado grande para el análisis manual por parte de humanos. Por otro lado, son demasiado valiosos para ser ignorados. Entonces el análisis de sentimiento intenta resolver este problema entrenando el modelo en opiniones ya identificadas o etiquetadas con diferentes polaridades y usa este modelo para predecir opiniones no identificadas o no etiquetadas (Ay Karakuş, B., Talo, M., Hallaç, İ. R., & Aydin, G., 2018).

Por lo útil que resultan estos modelos complementamos el análisis con métodos no supervisados realizado previamente con diversos algoritmos de aprendizaje supervisado. Los resultados de estos

últimos reflejan la capacidad para generalizar el conocimiento de los datos disponibles con casos objetivo o etiquetados, de modo que los mismos puedan usarse para predecir casos nuevos (no etiquetados)(Berry, M. W., Mohamed, A., & Yap, B. W., 2019). Los algoritmos que utilizamos nos permitieron clasificar las reseñas en positivas o negativas, utilizando las etiquetas de los datos para que los algoritmos "aprendan". El objetivo de utilizar los mismos era poder tener una "valoración social", es decir queríamos predecir cuál era el puntaje de los consumidores frente a diferentes temas o comentarios, ya que esto permite a los tomadores de decisiones planificar tareas relevantes de una manera más consciente de la opinión de sus consumidores.

Nos referimos a "valoración social" ya que estos modelos nos permiten evaluar el corpus de un conjunto de reseñas, en nuestro caso de miles, lo que nos brinda un panorama mucho más amplio que el caso en el que evaluemos las reseñas y puntajes de manera individual. Estos modelos de predicción entrenados con tantas opiniones distintas son capaces de predecir o conocer cómo evaluarán los consumidores en sus reseñas los distintos aspectos relacionados con el producto o el servicio que la empresa brinda. De cierta manera las valoraciones y ponderaciones cobran más sentido y pesan más porque dejan de ser opiniones aisladas o personales y el puntaje (*output*) del modelo se convierte en una "opinión social".

A) Preprocesamiento

Adicionamos al preprocesamiento realizado previamente a la vectorización (*BoW*), algunos pasos más para poder ingresar los datos en los modelos de aprendizaje supervisado. En primer lugar balanceamos nuestro dataset, ya que para realizar de forma correcta la clasificación de las clases debíamos tener datos con un número de observaciones similar para cada una. De no ser así, los algoritmos tienden a favorecer la clase con mayor proporción de observaciones pudiendo obtener un modelo que no predice de forma correcta la clase minoritaria (es decir, aquella clase que tiene menor cantidad de observaciones)(Dataequity, 2020). Al balancear el dataset quedaron igual cantidad de reseñas positivas (que son aquellas con puntaje "4" o "5") que negativas (que son aquellas con puntaje "1" o "2"),y descartamos las neutrales (que son aquellas con puntaje de "3") por lo que en total nos quedaron 22.442 reseñas (11.221 positivas y 11.221 negativas), todas tokenizadas. Luego convertimos todas las palabras de las reseñas en minúscula y filtramos los caracteres alfanuméricos.

B) Frecuencia de los tokens

Como segundo paso generamos un diccionario en donde por cada token tendremos su frecuencia. Para eso creamos una función cuyo output es una lista de diccionarios compuesta por:

- Id de la reseña.
- Calificación de la reseña (aún no convertida en binaria).
- Reseña tokenizada (cada palabra es un token).
- Un contador por cada token (indicando la cantidad de veces que aparece cada palabra en el review). Esto nos permite quedarnos con aquellos tokens que aparezcan mínimo 5 veces.

C) Binarización

Convertimos el problema en binario para que el clasificador (que es la columna de "review_rate") sea únicamente "Negativo" cuando la calificación sea 1 o 2 y "Positivo" cuando la calificación sea igual a 4 o 5.

D) Vectorización de palabras

Aplicamos los siguientes tipos de vectorización (*word embeddings*) para transformar las palabras en valores numéricos, *Bag of Words* y *TF-IDF*.

- **Bolsa de palabras (*Bag of Words*)**

Volvemos a aplicar el mismo método de bolsa de palabras mencionado en el preprocesamiento pero sobre la base de datos modificada para transformar todas las palabras en vectores y para que nuestros modelos de clasificación puedan interpretarlos .

- **Similitud coseno**

La similitud coseno es una métrica que nos servirá para poder comparar el contenido de las reseñas. Se basa en proyectar cada elemento a comparar en un espacio N-dimensional para comparar el coseno del ángulo que forman los elementos proyectados entre sí. De esta forma, a menor ángulo entre las proyecciones de los elementos mayor similitud entre los mismos.

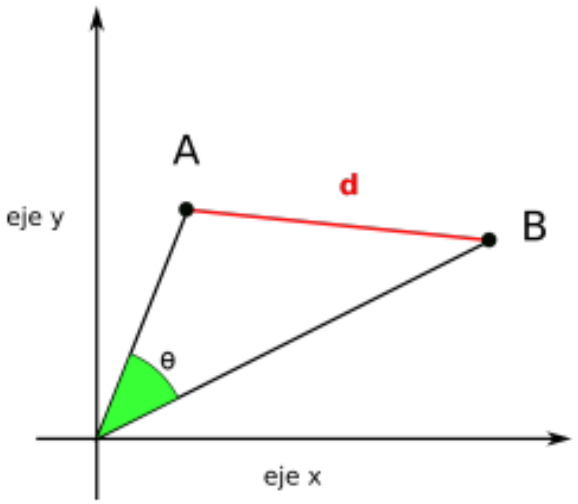
La fórmula matemática de ésta similitud es:

$$cos\,(\theta) = \frac{A \times B}{|A| \times |B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\ \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (3)$$

donde A_i y B_i son componentes de los vectores de las proyecciones de los elementos A y B , y donde $|A|$ y $|B|$ son los módulos de dichos vectores respectivamente. Al ser una métrica basada en el coseno, puede tomar valores en el rango $[-1, 1]$, siendo el valor -1 el peor valor de similitud posible y el 1 el mejor valor de similitud posible.

Como se puede observar en la Figura 1, dicha métrica de similitud se encarga de medir la orientación de dichos vectores de proyección y no la magnitud de los mismos (al tratar con el ángulo entre ellos), por lo que resulta muy útil para encontrar elementos que tengan relaciones parecidas en sus coordenadas, sea cual fueren sus magnitudes.

Figura 1. Se presentan las distintas reseñas (documentos) a comparar como vectores (A y B) y la similitud coseno es inversamente proporcional al ángulo formado por ambos vectores.



(Anguiano Batanero, E., 2019).

Dentro de la similitud de coseno dos vectores con la misma orientación tienen una similitud de coseno de 1, dos vectores orientados a 90° entre sí tienen una similitud de 0, y dos vectores diametralmente opuestos tienen una similitud de -1, independientemente de su magnitud.

- **Normalización TF-IDF**

El segundo método de vectorización que utilizamos es *TF-IDF* que es el acrónimo de *"Term Frequency times Inverse Document Frequency"*, que podemos traducir como "frecuencia del término por frecuencia inversa de documento".

Se utiliza para disminuir la frecuencia de palabras que se repiten mucho, ya que mide con qué frecuencia aparece un término dentro de un documento determinado, y lo compara con el número de documentos que mencionan ese término dentro de una colección entera de documentos.

Es la unión de dos métricas, la primera de ellas es la de frecuencia del término, *term frequency* o *TF* y la segunda que es la frecuencia inversa de los documentos, *inverse document frequency* o *IDF.* El valor *TF-IDF* aumenta proporcionalmente al número de veces que una palabra aparece en el documento, pero es compensada por la frecuencia de la palabra en la colección de documentos, lo que permite manejar el hecho de que algunas palabras son generalmente más comunes que otras. A continuación analizamos su fórmula matemática.

- *Term frequency (TF):* frecuencia de la palabra.

$$tf_{t,d=} \begin{cases} 1 + log_{10}\, count(t,d) & if\ count(t,d) > 0 \\ 0 & otherwise \end{cases} \quad (4)$$

Como podemos observar en la fórmula, la función crece logarítmicamente y con base diez, es decir que lo que aumenta por pasar de 10 a 100 es lo mismo que lo que aumenta por pasar de 100 a 1000. Como consecuencia, cada vez que aparece mucho más una palabra empieza a pesar menos, de manera que le quita peso a las palabras excesivamente frecuentes.

- *Inverse document frequency (IDF):*

  - N es el número total de documentos.

  - df_t = frecuencia del término t en el documento.

$$idf_t = log\left(\frac{N}{df_t}\right) \quad (5)$$

Una palabra que aparece en todos los documentos es poco informativa.

  - $w_{t,d}$ = Término t en el documento d:

$$w_{t,d} = tf_{t,d} * idf_t \quad (6)$$ (Jurafsky, D., & Martin, J. H., 2021).

De esta manera tendrán mucha importancia las palabras que aparecen mucho y en pocos documentos.

E) Entrenamiento de clasificadores

Después del preprocesamiento de datos, encontramos grupos de características similares usando la medida de distancia del coseno, es decir, encontramos palabras similares o palabras más cercanas en representación vectorial utilizando la distancia del coseno. Una distancia de coseno más alta indica una orientación semántica más cercana de las palabras (Bansal, B., & Srivastava, S., 2018).

Después de encontrar las características más similares, y de tener los datos divididos en positivos y negativos, podemos entrenar nuestros cuatro clasificadores para comparar luego su perfomance.

- **Bosque aleatorio** (en inglés, *Random Forest*)

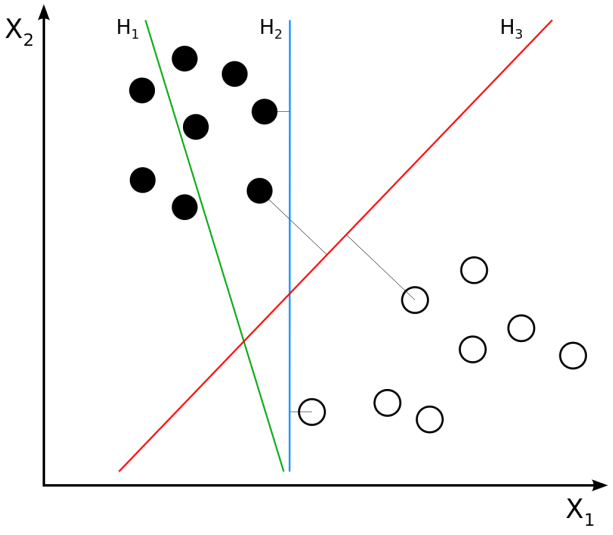
Un modelo *Random Forest* está formado por un conjunto o ensamble de árboles de decisión individuales, cada uno entrenado con una muestra aleatoria extraída de los datos de entrenamiento originales mediante *bootstrapping*. Esto implica que cada árbol se entrena con unos datos ligeramente distintos. En cada árbol individual, las observaciones se van distribuyendo por bifurcaciones (nodos) generando la estructura del árbol hasta alcanzar un nodo terminal. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo (Rodrigo, J., 2020).

En lo que respecta a los hiperparámetros, es decir las configuraciones utilizadas para el entrenamiento, usamos todos los que la librería "Sklearn" trae por defecto, excepto el número de árboles incluidos en el modelo que lo elevamos a 400 y el número de cores (núcleos del procesador) empleados para el entrenamiento que lo fijamos en -1 para utilizar todos los disponibles, ya que en el *Random Forest l*os árboles se ajustan de forma independiente, por lo que la paralelización (es decir, la distribución de la ejecución de esos ajustes) reduce notablemente el tiempo de entrenamiento.

- **Máquinas de vector soporte** (en inglés, *Support Vector Machine*)

El objetivo de una máquina de vector soporte es encontrar el hiperplano con el máximo margen que puede separar a las clases linealmente. A dicho hiperplano de separación definido como el vector entre los 2 puntos de las 2 clases más cercanas se lo llama vector soporte. Cuando las nuevas muestras se agregan en dicho modelo, en función de los espacios a los que pertenezcan, pueden ser clasificadas a una o la otra clase (Abd Elkarim, I. S., & Agbinya, J., 2019).

Figura 2. Ejemplos de hiperplanos. H1 no separa las clases. H2 las separa, pero solo con un margen pequeño. H3 las separa con el margen máximo (vector soporte).



(Weinberg, Z., 2020).

En nuestro modelo utilizamos los hiperparámetros que la librería trae por defecto que incluyen el kernel de función de base radial (RBF), pero le establecimos que el "random_state" que controla la generación de números pseudoaleatorios para mezclar los datos para estimaciones de probabilidad sea igual a 1, y que "probabilty = True" para habilitar las estimaciones de probabilidad.

- **Bayes ingenuo** (en inglés, *Naive Bayes*)

El clasificador *NB* es un clasificador probabilístico básico basado en el teorema de Bayes. Puede predecir las probabilidades de pertenencia a una clase, como la probabilidad de que una muestra dada pertenezca a una clase en particular (Leung, K. M., 2007). Este modelo es denominado "*Naive*" ya que asume que las variables predictoras son independientes entre sí. En otras palabras, que la presencia de una cierta característica en un conjunto de datos no está en absoluto relacionada con la presencia de cualquier otra característica.

En nuestro trabajo utilizamos un modelo de Bayes Ingenuo Multinomial ya que es adecuado para la clasificación con características discretas (por ejemplo, el recuento de palabras para la clasificación de texto). Elegimos este modelo ya que el clasificador de Bayes Ingenuo Simple considera la probabilidad de aparición de cada término dada la clase de forma binaria, es decir el término aparece o no y entonces su probabilidad condicional dada la clase es o no considerada. En cambio, el clasificador de Bayes Ingenuo Multinomial suele mejorar el desempeño pues considera el número de apariciones del término para evaluar la contribución de la probabilidad condicional dada la clase con lo que el modelado de cada documento se ajusta mejor a la clase a la que pertenece. Puede pensarse que si se modifica la representación de los documentos, de manera que el conteo de los

términos que aparecen en él se cambia por el número de apariciones del término en la clase cuya probabilidad de pertenencia del documento se está evaluando, se está proporcionando información adicional al clasificador para que la asignación de clase mejore (Anguiano-Hernández, E. 2009).

- **Regresión logística** (en inglés, *Logistic Regression*)

La regresión logística es un instrumento estadístico de análisis multivariado, que se usa para predecir el resultado de una variable categórica (que puede adoptar un número limitado de categorías) en función de las variables predictoras (Chitarroni, H., 2002). Sirve para modelar la probabilidad de un evento ocurriendo en función de otros factores y las probabilidades que describen el posible resultado se modelan como una función de variables explicativas, utilizando una función logística.

En nuestro trabajo el algoritmo que utilizamos para la optimización fue "liblinear". Se implementa utilizando la biblioteca liblinear de código abierto y utiliza internamente el método de descenso del eje de coordenadas para optimizar iterativamente la función de pérdida (Katastros, 2022).

F) Evaluación de resultados

Utilizamos diversas métricas de evaluación de resultados para los algoritmos de clasificación como *Accuracy* (en español, Exactitud), *Precision* (en español, Precisión), *Recall* (en español, exhaustividad) y *F1-score* (en español, Valor-F).

Para evaluar los resultados tenemos que considerar la matriz de confusión de cada algoritmo. En dicha matriz cada columna representa el número de predicciones de cada clase y la fila representa las instancias en la clase real. Es decir que nos permite comprender que tipos de aciertos y errores está teniendo el modelo al clasificar los datos. Los resultados pueden ser cuatro, ya que surgen dos posibles valores reales y dos posibles valores predichos que se resumen en la siguiente tabla.

Tabla 2. Matriz de confusión.

	Predicción	
Realidad	Verdaderos Negativos (VN)	Falsos Positivos (FP)
	Falsos Negativos (FN)	Veraderos Positivos (VP)

- Verdadero positivo: El valor real es positivo y el algoritmo predijo también que era positivo.
- Verdadero negativo: El valor real es negativo y el algoritmo predijo también que el resultado era negativo.
- Falso negativo: El valor real es positivo, y el algoritmo predijo que el resultado es negativo.

- ❖ Falso positivo: El valor real es negativo, y el algoritmo predijo que el resultado es positivo.

Las ecuaciones (A)-(E) muestran los parámetros numéricos para evaluar el desempeño de cada clasificador.

A) Accuracy = $\frac{VP+VN}{VP+VN+FP+FN}$ (7)

El accuracy o *exactitud* mide el porcentaje de casos que el modelo ha acertado.

B) Precisión = $\frac{VP}{VP+FP}$ (8)

Con la métrica de precisión podemos medir la calidad del modelo de machine learning en tareas de clasificación, ya que se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión. En forma práctica, es el porcentaje de predicciones positivas correctas.

C) Recall = $\frac{VP}{VP+FN}$ (9)

La métrica de *recall* o exhaustividad nos va a informar sobre la cantidad que el modelo de *machine learning* es capaz de identificar, ya que es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo.

D) F1-score = $2 \times \frac{precision \times recall}{precision + recall}$ (10)

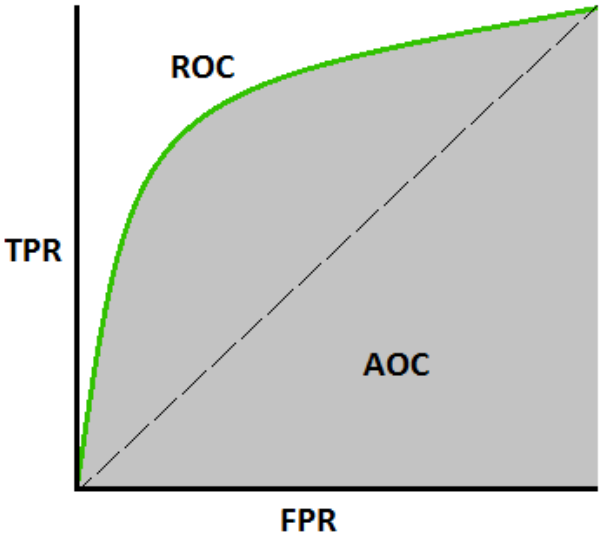
El valor F1 se utiliza para combinar las medidas de *precision* y *recall* en un sólo valor. Esto es práctico porque hace más fácil el poder comparar el rendimiento combinado de la precisión y la exhaustividad entre varias soluciones.

E) Curva AUC - ROC

Es una medida de rendimiento para problemas de clasificación en varias configuraciones de umbrales. ROC es una curva de probabilidad y AUC representa el grado o medida de separabilidad. Indica cuánto es capaz el modelo de distinguir entre clases. Cuanto mayor sea el AUC, mejor será el modelo para predecir 0 como 0 y 1 como 1.

La curva ROC se traza con TPR (Tasa de Verdaderos Positivos) contra FPR (Tasa de Falsos Positivos) donde TPR está en el eje y y FPR está en el eje x.

Figura 3. AUC - Curva ROC.



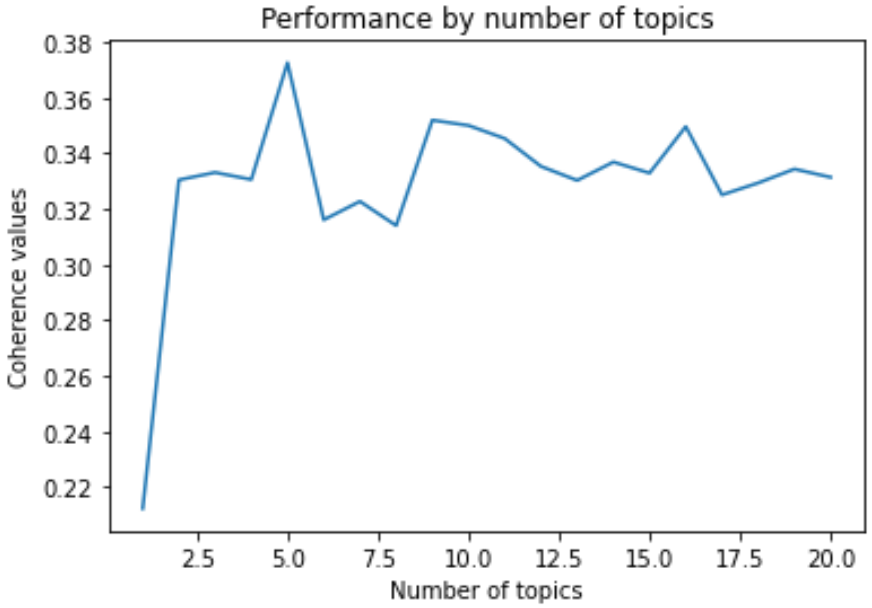
(Narkhede, S., 2018).

4. Resultados

4.1. Modelado de tópicos

Para definir el número adecuado de tópicos para entrenar el modelo, trazamos el rendimiento de cada modelo a medida que aumentamos el número de temas y observamos que un modelo con 5 tópicos era el que mejor funcionaba, es decir que era aquel modelo con mayor coherencia (el más interpretable para un ser humano ya que tiene palabras semánticamente parecidas en cada tema).

Figura 4. Rendimiento del modelo (valor de coherencia) frente a la cantidad de temas.



Los resultados del análisis de modelado de tópicos se muestran en la Figura 5. Como podemos observar, filtramos únicamente aquellas palabras que sean adjetivos y sustantivos para formar las nubes de palabras de cada uno de los cinco tópicos encontrados por el modelo LDA.

Figura 5. Resultados del análisis de tópicos.

Entre las palabras derivadas de cada gráfico, el tamaño de la palabra está relacionado con la importancia, lo que indica que aquellas palabras más grandes tienen el significado más importante en ese tema. Luego, establecimos un nombre de tema para describir las palabras contenidas en cada tópico.

El tópico 0 era un tema representativo de 10 palabras y se lo denominó “Relación precio-calidad”, que puede explicarse con palabras como “producto”, “relación”, “precio”, “calidad”, “descripción”, “expectativa”. El tópico 1 se denominó “Sistema operativo” con palabras clave como “android”,

“sistema”, “aplicación”, “función”, “modelo”. El tópico 2 se denominó “Diseño” con palabras clave como “color”, “tamaño”, “original”, “modelo”, “fino”. El tópico 3 se denominó “Características básicas” ya que incluye palabras generales como “memoria”, “cámara”, “sonido”, “pantalla”. Por último, el tópico 4 se denominó “Duración de la batería” debido a la diferencia en los valores de importancia en palabras como “batería”, “carga”, “dura”, “día”, “vida”.

Como podemos ver en los resultados mencionados, los temas latentes en las reseñas en línea de Mercado Libre sobre los “Celulares y Teléfonos”, se agruparon en cinco temas en total. Los clientes con alguna experiencia de uso en esa categoría de productos pueden juzgar que las razones para realizar sus elecciones se basan en factores claves como “Relación precio-calidad”, “Sistema operativo”, “Diseño”, “Características básicas” y “Duración de la batería”.

4.2. Análisis de sentimiento con métodos no supervisados

En la siguiente tabla podemos observar algunos de los resultados del análisis de sentimiento realizado con métodos no supervisados, donde tenemos el sentimiento asignado por el modelo a cada *review* en la última columna denominada “sentiment”.

Tabla 3. Ejemplos de los resultados del análisis de sentimiento.

item_key	item_title	review_key	review_title	review_content	review_rate	review_title+review_content	sentiment
MLA1111890165	Moto G60s 128 Gb Azul 6 Gb Ram	139249923	Bueno	Buena relacion precio calidad, anda muy bien para el precio.	3	Bueno.Buena relacion precio calidad, anda muy bien para el precio.	POS
MLA1111890165	Moto G60s 128 Gb Azul 6 Gb Ram	138048401	Bueno	En todo es un buen celular, lo malo que tiene es su tamaño es muy grande usaba uno de 5,5 pulgadas y se nota mucho la diferencia, ah la pantalla no es muy buena, se ve bien pero no tiene contraste, cualquier video con mucha iluminación se ve pálido. En todo lo demás es bueno, tiene 120hz es rápido, para juego corre todo.	3	Bueno.En todo es un buen celular, lo malo que tiene es su tamaño es muy grande usaba uno de 5,5 pulgadas y se nota mucho la diferencia, ah la pantalla no es muy buena, se ve bien pero no tiene contraste, cualquier video con mucha iluminación se ve pálido. En todo lo demás es bueno, tiene 120hz es rápido, para juego corre todo.	NEU
MLA915377519	Samsung Galaxy A52 128 Gb Awesome Black 6 Gb Ram	110097807	Miserables	El teléfono funciona, nada espectacular dentro de las prestaciones mínimas que uno pretende por lo pagado pero es malísimo. Sin auriculares y un cargador justo con un cable de pésima calidad, la tapa posterior es muy blanda y si o si necesitas protegerlo con una funda. Recomendable si, decepcionante es su presentación también.	3	Miserables.El teléfono funciona, nada espectacular dentro de las prestaciones mínimas que uno pretende por lo pagado pero es malísimo. Sin auriculares y un cargador justo con un cable de pésima calidad, la tapa posterior es muy blanda y si o si necesitas protegerlo con una funda. Recomendable si, decepcionante es su presentación también.	NEG

Es interesante resaltar estos ejemplos donde el "review_rate" de las reseñas es 3, pero sin embargo el modelo de análisis de sentimiento no supervisado, que fue entrenado sin considerar las etiquetas de los datos, puede detectar distintas emociones en los distintos comentarios que no necesariamente se corresponden entre ellas o son las mismas. Podemos observar que a la primera reseña el modelo de análisis de sentimiento la clasifica como positiva, y cuando es leída e interpretada por un humano también da la sensación de estar expresando un aspecto positivo del producto como es la buena relación precio-calidad. Pero a la segunda reseña el modelo la clasificó como neutral y también parece lógico ya que el consumidor en la misma está expresando algunas ventajas y también desventajas del producto. Y a la última reseña la clasifica como negativa y cuándo la interpretamos vemos que aparecen palabras que expresan descontento como "miserables", "malísimo", "pésima", "decepcionante". Este análisis, en el que dejamos sin consideración las etiquetas, nos permite ir un poco más en profundidad con cada comentario e interpretar esas emociones y aspectos que quizás ni el propio consumidor tenía en claro al momento de publicar su reseña. También deja a la vista la discordancia lógica que existe entre las percepciones de los usuarios, ya que para algunos un 3 puede ser un puntaje neutral, pero quizás para otro es positivo o negativo.

Aprovechamos dicha clasificación de comentarios para derivar aquellas palabras que expresan emociones tanto positivas como negativas y graficarlas en una nube de palabras para observar cuáles son las más importantes.

Figura 6. Palabras negativas del análisis de sentimiento.

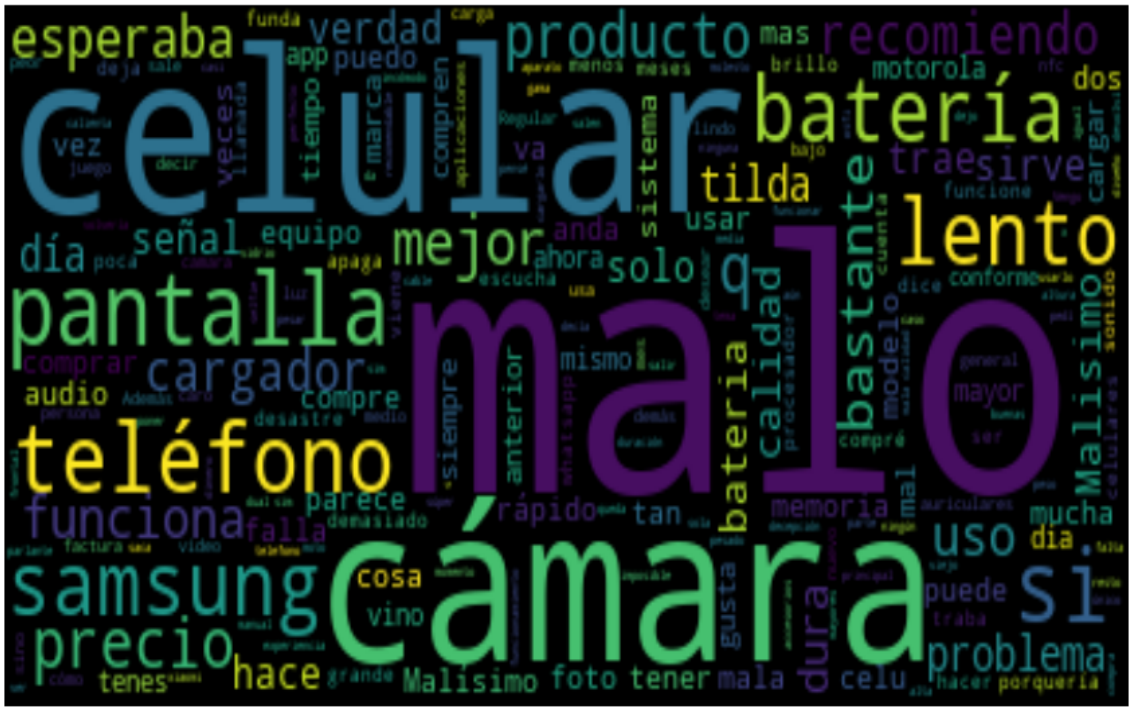


Figura 7. Palabras positivas del análisis de sentimiento.

Los resultados mostraron que las palabras que expresan emociones negativas como muestra la Figura 6, suelen estar relacionadas con "la cámara", "precio", "batería", "se tilda", "lento", "falla", "mala calidad", "pantalla", "cargador", "señal". Las palabras que expresaban sentimientos positivos

como muestra la Figura 7 incluían "buena calidad", "excelente producto", "lo recomiendo", "muy lindo", "precio calidad".

4.3. Modelos de clasificación de sentimiento con aprendizaje supervisado

En esta sección presentamos los resultados de los cuatro algoritmos de clasificación utilizados sobre nuestro dataset en la detección de sentimientos a través de los modelos de aprendizaje supervisado.

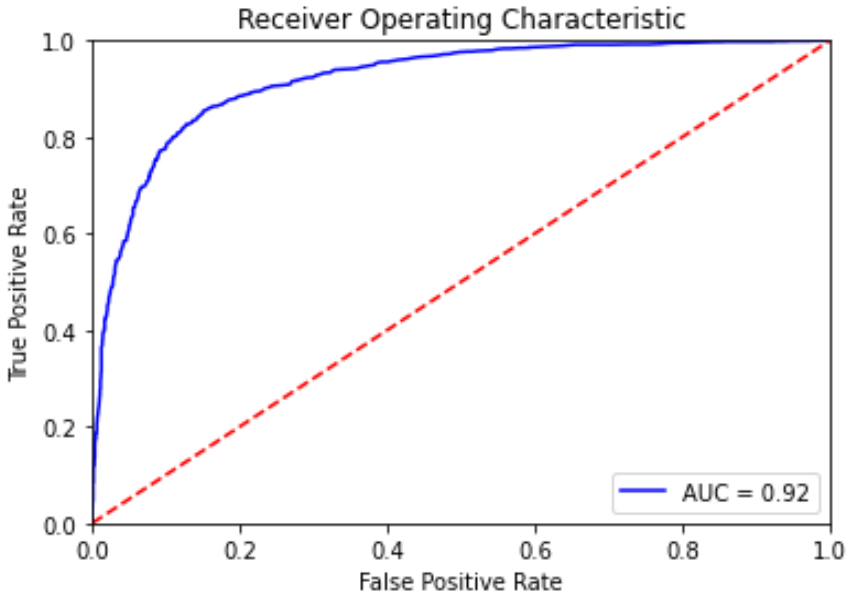
Tabla 4. Resultados de los modelos de clasificación.

Algoritmo	Accuracy	Precision	Recall	F1-score
RF	0.85	0.88	0.81	0.84
SVM	0.78	0.81	0.73	0.77
NB	0.78	0.76	0.83	0.79
LR	0.72	0.72	0.73	0.73

Como podemos observar en la tabla el modelo que obtuvo mayor *Accuracy* fue el *Random Forest* clasificando el 85% de los datos correctamente. Ver *Apéndice A* para visualizar las figuras de comparación de los resultados de todos los algoritmos.

Para complementar el análisis de resultados en la siguiente figura mostramos el resultado del Área Debajo de la Curva o *Area Under the Curve* del algoritmo que mejor performo que es el *Random Forest*. Ver *Apéndice B* para observar las figuras del resto de los algoritmos.

Figura 8. AUC del modelo Random Forest.



5. Conclusiones y trabajos futuros

5.1. Conclusiones

En este estudio utilizamos el modelado de tópicos y el análisis de sentimiento para identificar las necesidades y opiniones de los clientes que compran teléfonos móviles a través de internet, específicamente a través de la web de Mercado Libre. Mediante el análisis de las reseñas en línea escritas por clientes que han realizado alguna compra en este sitio, exploramos los factores que influyen en la intención de compra. Las relaciones duraderas con los clientes requieren de una comunicación constante entre ellos y las empresas, de manera que tengan libertad de transmitir su visión sobre el producto, sus ideas y sugerencias.

El uso de los métodos que aplicamos muestran resultados prometedores para los negocios, ya que permiten analizar las reseñas y en base a ellas identificar con qué se sienten satisfechos e insatisfechos los consumidores. Para ayudar a las organizaciones en sus actividades de gestión y en el proceso de toma de decisiones, proporcionando datos claves y métodos que sean útiles para retener a los clientes y también para ampliar el mercado. Poder conocer lo que se menciona sobre la marca en las redes y los sentimientos, emociones u opiniones que se expresan sobre la misma, es una de las maneras más directas de tener feedback de los consumidores que deben ser el principal foco de atención en las empresas de cualquier rubro, ya que son quienes tienen "el poder" de decidir si volverán a elegir los productos o no.

En los últimos años hemos sido testigos del crecimiento exponencial de los datos. El modelado de temas utilizando LDA asignación de Dirichlet latente es un método útil para extraer conocimiento y relaciones en estos datos sin tener que dedicarle tiempo excesivo, ya que permite reducir la cantidad de los mismos, facilita su interpretación arrojando resultados comprensibles y garantiza que no se pierda información importante. La esencia de LDA está en su exploración de distribuciones de temas dentro de documentos y de distribuciones de palabras dentro de temas, lo que permite identificar temas coherentes a través de un proceso iterativo.

En lo que respecta al análisis de sentimiento con métodos no supervisados, en el dominio empresarial, podemos concluir que sirve principalmente para detectar la voz del consumidor, conocer la reputación de la marca, la tendencia de la publicidad y del comercio en línea. Ya que permite entender los aspectos clave del producto y del servicio que una marca le brinda a sus clientes, además deja en claro cuáles son los elementos que deben mejorarse y cuáles hacen felices a los clientes. También pueden utilizarse para realizar una investigación de mercado y de competencia,

ya que permiten observar qué marcas reciben más comentarios positivos o negativos y también se puede saber de qué tratan los mismos, eso puede guiar los esfuerzos de marketing para posicionar a la marca como se desea.

Adicionalmente el análisis de sentimiento con los métodos supervisados que implementamos para predecir la polaridad de los sentimientos en función de los algoritmos que entrenamos, son útiles para tener una valoración social ya que sus resultados nos demuestran su buena capacidad para poder predecir cuál será la opinión de miles de consumidores respecto a diversos aspectos de los productos o servicios. Es decir que dejamos de considerar las reseñas de manera individual para analizar el corpus de un conjunto de reseñas que nos dan una visión mucho más amplia de la realidad.

Los resultados obtenidos del rubro analizado específicamente nos permiten mencionar las siguientes implicaciones. En primer lugar, derivado del análisis de temas y de las nubes de palabras, observamos que es clave que los productos presenten una relación acorde entre el precio y la calidad; éste punto está muy relacionado con la imagen que da la marca, con el cumplimiento de expectativas y con la transparencia que cualquier entidad empresarial debe transmitir para retener a sus clientes. La transmisión de información precisa y verídica lleva a poder cumplir con las medidas de satisfacción de los clientes y lograr así la fidelización de los mismos. En segundo lugar, es muy valorado por los usuarios que los celulares tengan un buen sistema operativo (IOS, Android), ya que es es el software principal del dispositivo, que administra los recursos de hardware y software interactuando con el usuario a través de interfaces gráficas y conectividad, y sobre el que se instala el resto de aplicaciones que emplea el usuario final. En tercer lugar, el diseño de los celulares aparece como uno de los puntos claves para colocarse en el gusto y el corazón de los consumidores, ya que actualmente es difícil imaginar nuestro dia a dia sin nuestro celular, por lo que es algo que miramos repetidas veces al día y es necesaria nuestra aceptación estética ya que no es algo que pase desapercibido. En cuarto lugar, podríamos mencionar como importante todas las características básicas de los celulares que deben cumplir con nuestras expectativas y necesidades, como lo es la memoria, la cámara, el sonido, la pantalla y no menos importante la vida útil y duración de la batería. Ya que como lo mencionamos anteriormente los celulares se han convertido en una necesidad para la mayoría de las personas en todo el mundo, y cumplen múltiples funciones como mantenernos conectados con la familia, colegas de trabajo, poder acceder al correo electrónico, redes sociales, almacenamiento de datos, toma de fotografías, entre otras de las tantas funciones disponibles.

En lo que respecta al análisis de sentimiento con métodos no supervisados, podemos referirnos a las expresiones negativas, que son aquellos factores que afectan negativamente a la futura compra o

que hace que los clientes se sientan disconformes. Se centran básicamente en los problemas técnicos que los celulares pueden presentar como por ejemplo que contengan alguna falla, que se tilden, que sea lentos y además surgen algunos aspectos que son los que suelen estar relacionado con dichas cualidades que son una mala duración de la batería, baja señal, que el cargador sea lento o se rompa y la mala calidad de la cámara. Es importante que los fabricantes o vendedores se centren en estos factores, ya que quizás con una mejora que no suponga mucho costo pueden mejorar la percepción del cliente y evitar que sea un obstáculo para el desarrollo de su marca. También podemos referirnos a las expresiones positivas que surgieron del análisis que permiten conocer cuáles son los aspectos que los clientes valoran, y estos están relacionados principalmente con la calidad y su adecuada relación con el precio del producto. Es importante resaltar en este punto que aquellas personas que expresan sentirse conformes o que expresan un sentimiento positivo hacia su compra, suelen recomendar a la marca y el producto, y eso es información muy confiable para los futuros compradores, ya que estos últimos suelen depositar su confianza en personas que han probado previamente la marca. Además está demostrado que incluir recomendaciones o valoraciones positivas sobre las empresas en tu web o redes sociales, logra diferenciarte del resto de los competidores. No hay mejor marketing para una empresa que conseguir que sus propios clientes hablen bien de ella.

Finalmente, los resultados arrojados por los algoritmos de clasificación automáticos son capaces de predecir el sentimiento (positivo o negativo) de los mensajes publicados en Mercado Libre, es decir que nos permiten hablar de *data-feeling*, de manera de que se pueda pasar de tomar decisiones basadas en datos a decidir con base en lo que indican los sentimientos de los comentarios y así percibir a través de los datos digitales lo que los consumidores piensan y opinan en el mundo real. Para encontrar el mejor clasificador entrenamos cuatro modelos distintos: *Random Forest*, *Support Vector Machine*, *Naive Bayes* y *Logistic Regression.* El que mejor resultados obtuvo fue el *Random Forest* con un *accuracy* de 85% y un *AUC* de 0.92, es decir que es un modelo capaz de predecir correctamente el 85% de los comentarios y el que peor resultados logro fue el de *Logistic Regression* con un *accuracy* del 72% y un *AUC* de 0.78, lo que significa que el modelo predice correctamente el 72% de los comentarios lo que es un resultado aceptable pero puede mejorarse.

5.2. Trabajos futuros

Durante el desarrollo de la presente tesis fueron surgiendo algunas ideas o líneas que serían interesantes retomar en trabajos de investigación futuros, ya que no pudieron abordarse en profundidad en este trabajo, que son las que mencionamos a continuación.

- Implementar la metodología y los modelos propuestos en otras categorías de productos en Mercado libre u otro comercio eléctronico para evaluar la eficacia y replicabilidad de los mismos.

- En lo que respecta al preprocesamiento realizado para ingresar los datos a los modelos de aprendizaje supervisado, podría evitarse descartar reseñas y pasarlas por el algoritmo con su puntaje original, sin binarizar. O probar con otros métodos de preprocesamiento y analizar si los resultados obtenidos por los modelos mejoran, por ejemplo, no balancear el dataset, o en vez de una bolsa de palabras utilizar N-gramas o el algoritmo Word2vec que usa un modelo de red neuronal para aprender las asociaciones.

- Eliminar de la base de datos preposiciones o letras sueltas poco informativas como las que se pueden observar en las nubes de palabras formadas a partir del análisis de sentimiento con métodos no supervisados, como por ejemplo “de”, “n”, “y”, “que”, “q”, “por”, “el”, entre otras. Ya que son palabras que mostraron relevancia en los resultados pero podrían estar ensuciando el dataset sin sumar información con poder predictivo o de clasificación.

- Dada la buena performance del modelo *Random Forest* podría utilizarse como modelo alternativo un Gradient Boosting Machine o algún derivado, ya que pueden ser más precisos que el *Random Forest* debido a que los árboles se entrenan para corregir los errores de los demás y no de manera independiente, por lo que son capaces de capturar patrones más complejos en los datos.

- Realizar el análisis de sentimiento con métodos no supervisados y supervisados considerando las diferentes marcas de celulares, por ejemplo Samsung, Nokia, Motorola, entre otras y poder llegar a realizar una comparación en profundidad de las ventajas y desventajas de cada una de ellas. También otra alternativa podría ser aplicar el análisis de sentimiento sobre los diferentes temas encontrados por el modelo de análisis de tópicos y no aplicar el análisis directamente sobre la base de datos completa.

Referencias

- Abd Elkarim, I. S., & Agbinya, J. (2019). A Review of Parallel Support Vector Machines (PSVMs) for Big Data classification. Australian Journal of Basic and Applied Sciences, 13(12), 61-71.

- Anguiano Batanero, E. (2019). Aprendizaje por refuerzo y técnicas profundas aplicadas a un sistema de recomendación de venta al por menor (Master's thesis).

- Anguiano-Hernández, E. (2009). Naive Bayes Multinomial para clasificación de texto usando un esquema de pesado por clases.

- Ay Karakuş, B., Talo, M., Hallaç, İ. R., & Aydin, G. (2018). Evaluating deep learning models for sentiment classification. Concurrency and Computation: Practice and Experience, 30(21), e4783.

- Bansal, B., & Srivastava, S. (2018). Sentiment classification of online consumer reviews using word vector representations. Procedia computer science, 132, 1147-1153.

- Barrio, J. (2017). La influencia de los medios sociales digitales en el consumo. La función prescriptiva de los medios sociales en la decisión de compra de bebidas refrescantes en España.

- Bear, D., & Szabo, M. (2012). Social Shopping. El impacto del social media en nuestras decisiones de compra. CP Proximity.

- Berry, M. W., Mohamed, A., & Yap, B. W. (Eds.). (2019). Supervised and unsupervised learning for data science. Springer Nature.

- Blackbeast, Agencia de Marketing Digital. (2022). Concepto de Stop Words.

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

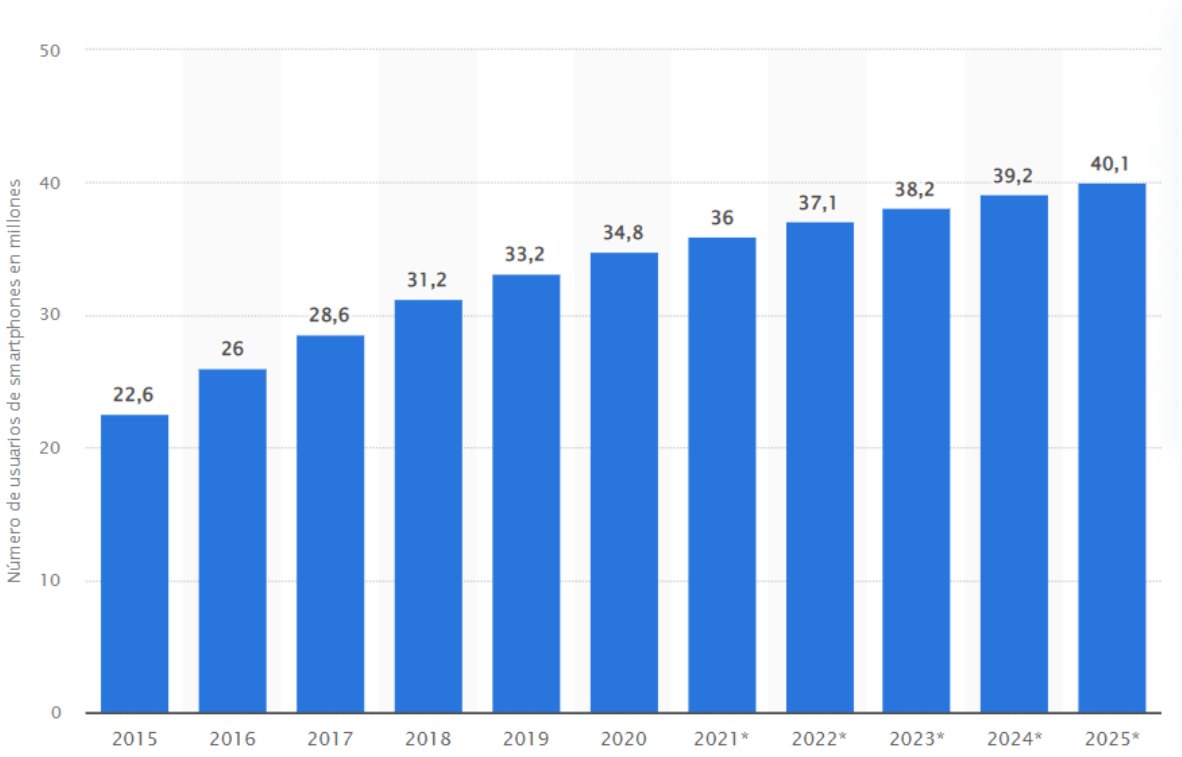
- Canete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., & Pérez, J. (2020). Spanish pre-trained bert model and evaluation data.

- Chitarroni, H. (2002). La regresión logística.

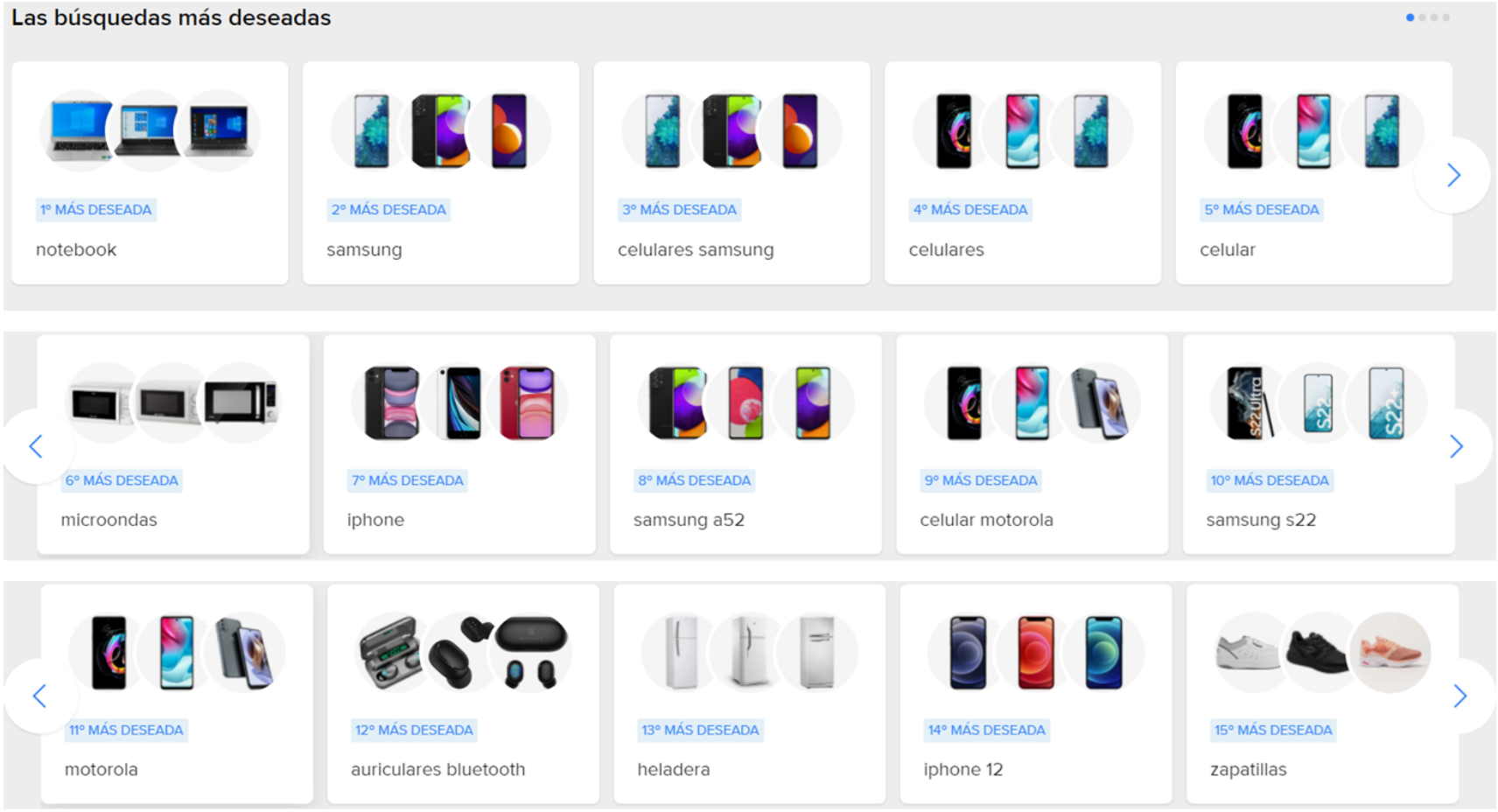
- Dataequity (2020). Dataset desbalanceado: Prediciendo eventos muy poco frecuentes.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.
- Filieri, R., & Mariani, M. (2021). The role of cultural values in consumers" evaluation of online review helpfulness: a big data approach. International Marketing Review.
- Flórez Calderón, M., & Ramos, Z. (1993). El mercado de la telefonía celular. Ingeniería e Investigación.
- Herrera, E. (2021). Mercado Libre tuvo ganancias por US$95,2 millones en el tercer trimestre del año. La Nación.
- Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78(11), 15169-15211.
- Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78(11), 15169-15211.
- Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
- Katastros (2022). LogisticRegressionCV parameter usage and meaning notes. https://blog.katastros.com/a?ID=00900-c2332bbf-6987-4fef-b3fa-9f7eb87edf3b.
- Khan, M. T., Durrani, M., Ali, A., Inayat, I., Khalid, S., & Khan, K. H. (2016). Sentiment analysis and the complex natural language. Complex Adaptive Systems Modeling, 4(1), 1-19.
- Kulshrestha,R. (2019). A Beginner's Guide to Latent Dirichlet Allocation(LDA).
- Kumar, S., & Zahn, C. (2003). Mobile communications: evolution and impact on business operations. technovation, 23(6), 515-520.
- Lee, B. C., & Byun, H. J. (2014). The impact of online review on purchasing behavior: A case of hotel and resort. Journal of Korean Tourism & Leisure, 26(7), 59-79.

- Leung, K. M. (2007). Naive bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering, 2007, 123-156.
- Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.
- Macario, A. (2018). La importancia de las reseñas online. Puro Marketing.
- Morstatter, F., & Liu, H. (2018). In search of coherence and consensus: measuring the interpretability of statistical topics. Journal of Machine Learning Research, 18(169), 1-32.
- Narkhede, S. (2018). Understanding auc-roc curve. Towards Data Science, 26(1), 220-227.
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in information retrieval, 2(1–2), 1-135.
- Rodrigo, J (2020). Random Forest con Python available under a Attribution 4.0 International.
- Rojas, A. (2021). ¿Qué es una API? ¿Para qué sirven las API?.
- Saxton, M. (2018). Analysis: Studying the Properties of Topic Models with Different Alpha Values.
- Statista Research Department. (2021). Número de usuarios de teléfonos móviles inteligentes en Argentina de 2015 a 2025 (en millones). https://es.statista.com/estadisticas/598527/numero-de-usuarios-de-moviles-en-argentina/.
- Weinberg, Z. (2020). Svm separating hyperplanes (SVG). Wikimedia Commons, the free media repository.
- Wikipedia, La enciclopedia libre. (2022). Lematización. https://es.wikipedia.org/w/index.php?title=Lematizaci%C3%B3n&oldid=140877099.
- Yujnovsky&Asoc. (2020). La pandemia y el boom del comercio electrónico en América Latina.
- Zhao, R., & Mao, K. (2017). Fuzzy bag-of-words model for document representation. IEEE transactions on fuzzy systems, 26(2), 794-804.

Anexo A. Número de usuarios de teléfonos móviles inteligentes en Argentina de 2015 a 2025 *(en millones).*



Anexo B. Búsquedas más deseadas de Mercado Libre.

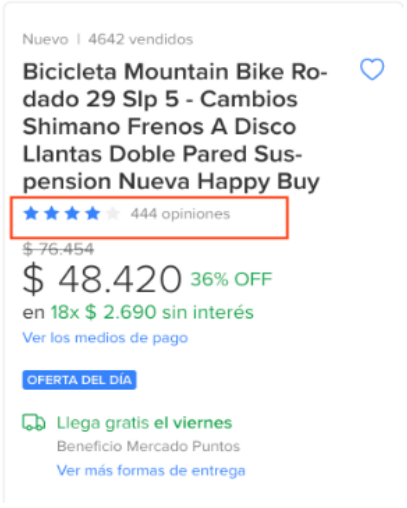


Fuente: *https://tendencias.mercadolibre.com.ar/*

Anexo C. Opiniones de productos.

Opiniones de productos

Una vez entregado el producto al comprador, este podrá opinar de acuerdo a su experiencia indicando cuántas estrellas le daría al producto, realizar un comentario y, en caso de ser un producto de Moda, indicar si le quedó como esperaba o no. De esta manera, el vendedor podrá saber el promedio de estrellas sobre sus productos. En Mercado Libre, verán las estrellas y cantidad de opiniones debajo del título de la publicación de la siguiente manera:



Fuente: *https://developers.mercadolibre.com.ar/es_ar/opiniones-sobre-producto*

Apéndice A. Figuras con los resultados de los algoritmos.

Figura 8. Comparación del Accuracy de los cuatro modelos.

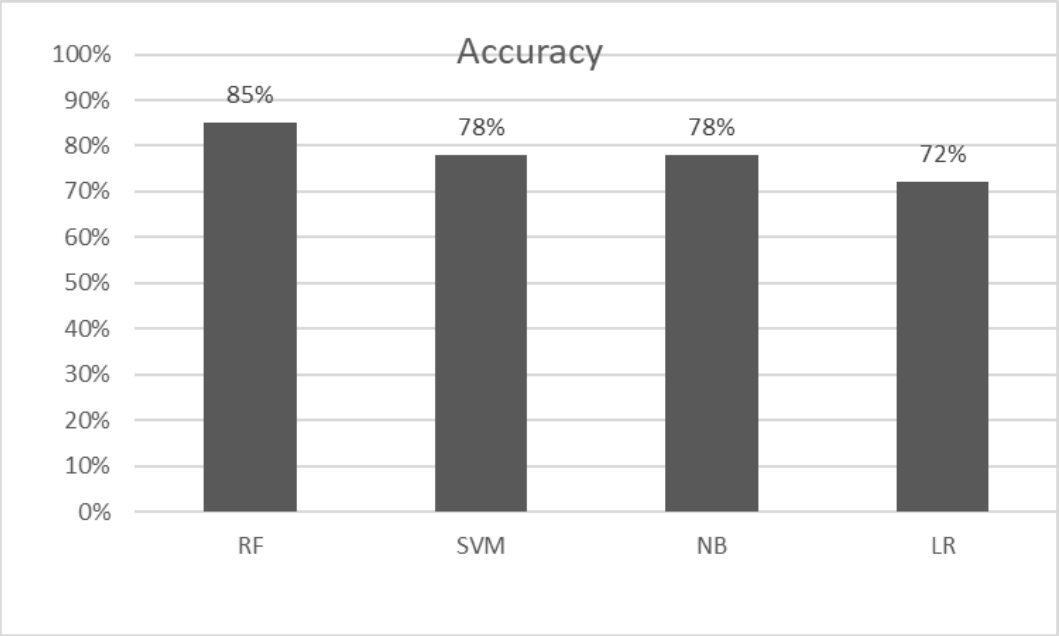
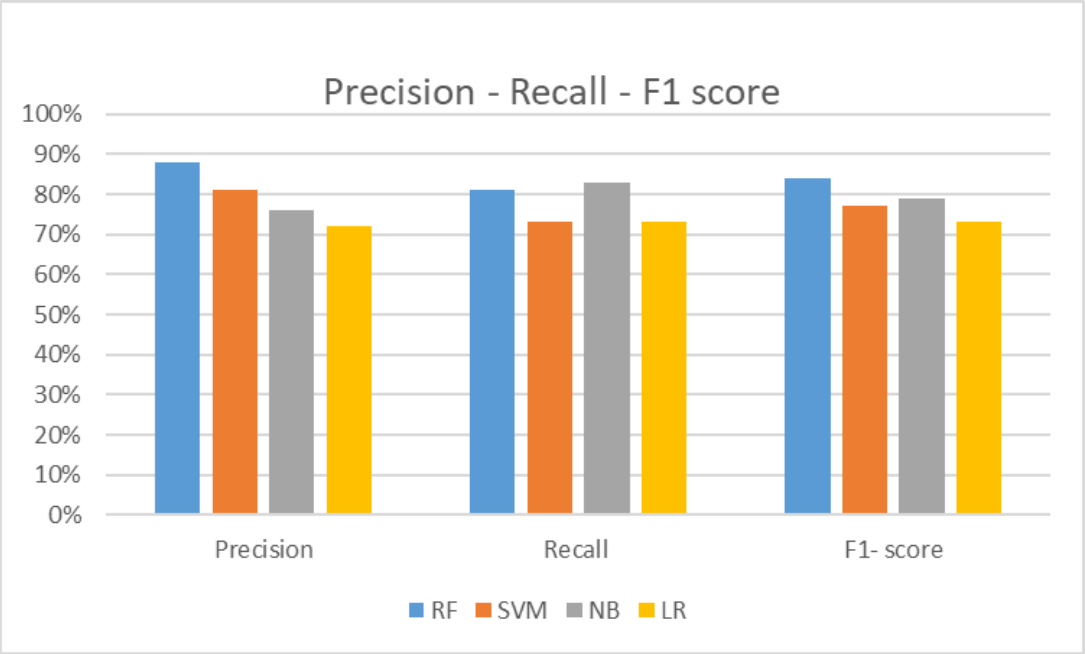
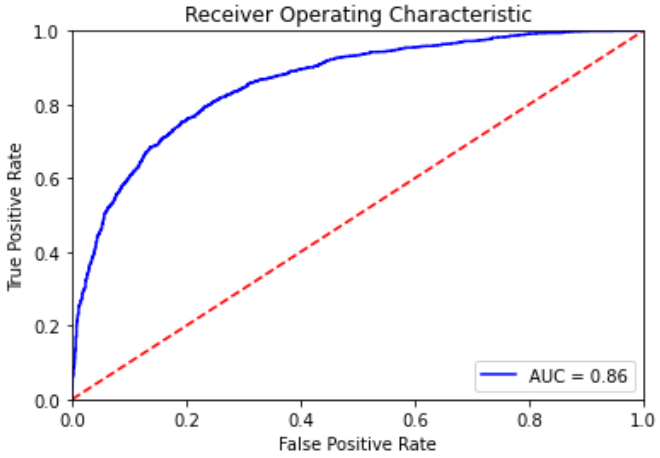


Figura 9. Precision, Recall y F1 score para los cuatro modelos.

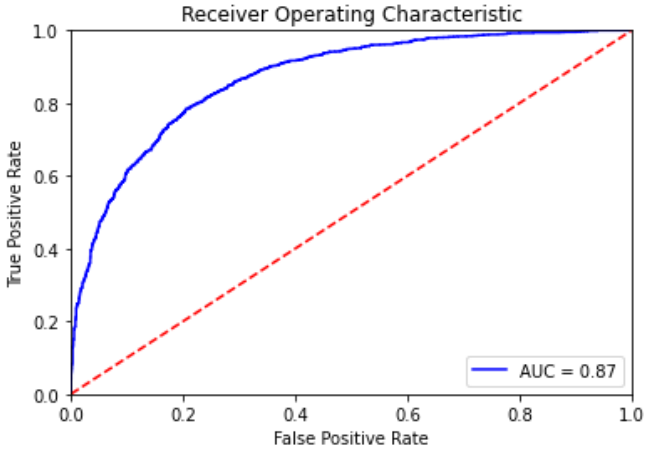


Apéndice B. AUC del modelo SVM, NB y LR.

AUC del modelo Support Vector Machine.



AUC del modelo Naive Bayes.



AUC del modelo Logistic Regression.