Comparación entre artículos de Mercado Libre por medio del análisis, clasificación y ranqueo de comentarios

Agustín Facundo Jobson

Universidad Católica de Santiago del Estero, Departamento Académico Rafaela

Resumen

Hoy en día la compra y venta por la plataforma web es uno de los principales medios por los cuales las personas adquieren los productos que desean, sin embargo, en muchas ocasiones se dificulta en gran medida el hecho de elegir un producto en lugar de otro de similares características, por lo que las opiniones y calificaciones que otras personas le dan a dichos productos juega un papel fundamental a la hora de tomar una decisión. El gran problema ahí es que las personas somos subjetivas, por lo que en muchas ocasiones los comentarios escritos y la calificación otorgada no suelen estar muy correlacionados entre sí. El objetivo de este estudio es describir y enunciar una serie de técnicas de análisis de sentimiento y Machine Learning que se centren en analizar los comentarios de un determinado producto y otorgarle una calificación objetiva al mismo (eliminando la subjetividad mencionada), con el fin de que se le simplifique a las personas el poder elegir entre productos muy similares. Para ello debemos obtener los comentarios de los productos que queremos evaluar con una API que el propio Mercado Libre provee, luego esos datos sufren un proceso de Tokenización y se le remueven las "stop words", y luego se le aplica uno de los diferentes métodos expuestos en este informe. De esta manera, al comparar los resultados con distintos productos, se podrá saber cuál de todos es el más factible de adquirir.

Palabras Clave

Sentiment Analysis, Comentarios de productos, Mercado Libre, Comparación de productos.

Problema de Investigación

El mundo está en constante cambio, ese es un hecho irrefutable, hace años la idea de tener computadoras interconectadas entre sí por medio de Internet y de compartir información en un instante entre distintas partes del mundo no era más que un simple deseo, para muchos, imposible. Pero hoy en día, eso que parecía inalcanzable, es nuestra realidad y nos resulta sumamente común, incluso indispensable en la vida cotidiana. Es porque el mundo cambia continuamente, que también lo hacen las personas, su forma de pensar, socializar, y actuar. A lo largo de los años, la forma en la que las empresas y las personas han comercializado sus productos ha cambiado rotundamente. En este contexto dinámico, pasamos de una comercialización cara a cara donde la información sobre los productos solía difundirse de boca en boca y a través de la publicidad, a un contexto en el que, si bien estos factores siguen siendo importante, también existe internet, que es sin dudas uno de los factores (por no decir el que más) más influyente en las decisiones de compra de las personas. Solo contando con una computadora y una conexión a internet, tenemos a disposición una amplia gama de información sobre cualquier producto, descripciones, experiencias de otros usuarios, puntos positivos y negativos, entre muchas otras cosas, y es por este "overflow" de información que en la mayoría de las ocasiones se nos dificulta entre elegir entre una serie de productos de características homogéneas, sin mencionar que también entra en juego la subjetividad de cada persona que realiza un comentario. Centrarse en la totalidad del Internet para comparar productos de similares características es por demás imposible, es por este motivo, que, a lo largo de este informe, se va a abordar una parte muy pequeña de lo que es el problema en general.

Mercado Libre es la comunidad de compra y venta online más grande de América Latina. [1] Desde su surgimiento por el año 1999 no ha parado de evolucionar, y pronto

permitió que cualquier usuario pueda publicar una opinión y ranquear cualquiera de los millones de productos que la página ofrece. La idea de ver las opiniones y experiencias de otros usuarios sobre un producto que me interesa es una característica fundamental a la hora de realizar una compra, sin embargo, a pesar de ser una funcionalidad muy útil que la pagina provee, es también un problema en sí mismo, ya que las personas en la gran mayoría de las ocasiones no somos objetivas, sino que nos dejamos llevar por nuestros pensamientos y sentimientos provocando que el comentario y el ranking que posteamos no refleje la realidad del producto en cuestión.

A continuación, se presentan una serie de imágenes extraídas directamente de los comentarios de algunos productos, con el fin de ilustrar la problemática planteada:

Puede suceder, además, que el comentario y posterior ranqueo, no se limite solo al producto en sí, sino además a cuestiones externas e independientes del mismo, como, por ejemplo, el servicio propuesto por el vendedor, de la propia página, de envío, etc.

La propuesta que presento en este proyecto y la cual voy a desarrollar a lo largo del mismo, viene a dar resolución a las cuestiones y problemáticas antes mencionadas. La misma se basa en realizar un "tracking" de todos los comentarios que se encuentran a disposición de un determinado producto. Posteriormente utilizar "Sentiment Analysis" para realizar un "re-scaling" objetivo, es decir, otorgarle una calificación evitando la subjetividad propia de las personas, la cual provoca en muchas ocasiones que la calificación final sea injusta. De esta manera se le puede aplicar una calificación general objetiva al producto, lo cual hace más fácil saber, entre productos similares, cual es mejor opción o no.

Preguntas de Investigación

¿Cómo se pueden obtener los comentarios de un producto? ¿Qué técnicas de análisis de sentimientos existen? ¿Cuál se puede utilizar para el problema planteado? ¿Qué métodos son los más eficientes para afrontar el problema en cuestión?

Objetivos de Investigación

- Exponer la/las maneras de obtención de los comentarios de un producto.
- Identificar que técnicas de análisis de sentimientos existen.
- Enunciar la técnica que se pueda utilizar para solventar la problemática planteada.
- Detallar el o los métodos que son los más eficientes para afrontar el problema en cuestión

Revisión de Antecedentes

Desde hace décadas el estudio del análisis de sentimiento o del "opinion mining" es algo que ha despertado el interés de los investigadores. A partir de ese momento, el análisis de sentimientos es una de las áreas de investigación de más rápido crecimiento en ciencias de la computación, es por este motivo que resulta complicado hacer un seguimiento exhaustivo de este fenómeno.

A pesar de que un primer enfoque del tema se remonta a trabajos realizados en el siglo XX, en el análisis subjetivo de texto realizado por la comunidad de lingüística computacional en la década de 1990, fue aproximadamente en la década del 2000 cuando se dio el estallido del análisis de opiniones basados en computadoras, con la disponibilidad de textos subjetivos en la Web [2].

Entre los años 2003 y 2005, según indica Sonai en su artículo, surge la idea de tratar a las palabras individualmente e identificarlas como una de las partes del habla, por ejemplo, verbo, adjetivo, sustantivo, adverbio, etc., lo que se conoce como bolsa de palabras, mientras que el clasificador procesa la palabra independientemente de su estructura gramatical. En esta fase, se presentan muchos documentos para indicar la importancia de subjetividad y objetividad en la etapa de extracción.[3]

A partir del año 2006 y en adelante, la minería de opinión comenzó a utilizarse en una gran cantidad de portales web, donde lo que interesaba era sobre todo la opinión de las personas acerca de candidatos presidenciales, sobre evento y entretenimiento, y las empresas comienzan a interesarse sobre lo que la gente opinaba de los productos que comercializaban. [3] Tal es la evolución de esta práctica que para finales del año 2016, ya se encontraban publicados aproximadamente 7000 trabajos de investigación, lo cual demuestra una alta tendencia en la utilización de estas técnicas, principalmente motivada por la posibilidad de recopilar y analizar cantidades masivas de opiniones y comentarios con la ayuda de herramientas de minería de texto.[2]

Una vez desarrollada brevemente la evolución que sufrió el análisis de sentimiento en los últimos años, es posible centrarse más específicamente en el problema planteado en este informe.

En lo que concierne al mismo, varios autores se han planteado problemáticas similares. Por ejemplo, en el año 2013, Rain y Callen exponen en su artículo una manera de detectar actitudes positivas o negativas hacia los productos a través de la utilización de Machine Learning, una "bolsa de palabras", un clasificador y una "lista de decisión" entendible por los humanos. Donde el número de estrellas que las personas le daban al producto servía de entrenamiento para realizar el aprendizaje automático. Al analizar los resultados, llegaron a la conclusión que el proyecto fue exitoso puesto que el clasificador logró etiquetar más del 50% de los comentarios de una manera satisfactoria, y concluyeron que el éxito se debía principalmente a la utilización de la bolsa de palabras. [4]

Unos años después, en el 2015, Bhatt, en conjunto con otros autores, idearon la premisa de que los comentarios de los

usuarios son en realidad una “mezcla de opiniones” entre opiniones sobre el producto, y sobre el servicio. Por lo que se propusieron desarrollar un modelo cuya función sea la de separar exitosamente dicha mezcla de opiniones, además de hacer hincapié en la característica que el usuario describió en su comentario. Por ejemplo: si el usuario escribe que “la cámara es buena”, el modelo clasificará la cámara como positiva. Básicamente el “workflow” de trabajo consistía en lo siguiente: obtenían la URL de los cometarios de Amazon para obtener toda la información, “limpiaban” la información, es decir, removían caracteres especiales y las denominadas “stop words” y almacenaban toda esa información en una Base de Datos para su posterior visualización y análisis, donde cada frase era enviada a un polarizador que retornaba un 1 si el comentario era positivo y un -1 si era negativo. Realizaron una serie de pruebas dentro de las cuales se destaca la prueba de análisis de un IPhone 5 donde concluyeron que el sistema retornaba resultados satisfactorios.[5]

Un trabajo realizado recientemente fue llevado a cabo en 2018 por unos investigadores de la Universidad de Ciencia y Tecnología de Bangladesh. Tanjim Ul Haque y su equipo de trabajo pensaron en “categorizar los comentarios positivos y negativos de los clientes sobre diferentes productos y construir un modelo de aprendizaje supervisado para polarizar una gran cantidad de opiniones”. [6] La metodología de trabajo es muy parecida a las anteriores, utilizaron archivos JSON para guardar la información (almacenaban, por ejemplo, ID del producto, ID del cliente, texto del comentario, etc.), la pre-procesaban, es decir, separaban el string (la sentencia) completo en muchas palabras, eliminaban las palabras “vacías” y utilizaron una bolsa de palabras y la distribución Chi-Cuadrado para determinar la diferencia entre los datos observados y los datos esperados, por supuesto esperando la menor diferencia posible. Empleando estas técnicas lograron unos resultados de aproximadamente el 90% de precisión.

Marco Teórico

Dentro del marco teórico se encuentran conceptos que se pueden considerar como generales, tales como “Sentiment Analysis” o Machine Learning, o conceptos más específicos, los cuales serán detallados más adelante.

Como el tópico principal del trabajo es sobre el análisis de sentimiento, se debe comprender correctamente cuál es su significado. Análisis de sentimiento hace referencia a la “minería contextual de texto que identifica y extrae información subjetiva en el material de origen”[7], y ayuda a quien la está utilizando a entender el sentimiento social del texto el cual está evaluando. Es un concepto que va de la mano con el concepto de “opinion mining” puesto que son extremadamente parecidos. “Opinion mining” es una técnica que, “dado un conjunto de documentos de texto que contienen opiniones sobre un objeto, la minería de opinión tiene como objetivo extraer atributos y componentes del objeto que se han comentado en cada documento y determinar si los comentarios son positivos, negativos o neutrales. [8]

La diferencia, aunque sutil, radica en que en el “opinion mining” solo se extrae la opinión de la persona, mientras que el análisis de sentimiento va más profundo y busca encontrar las emociones ocultas en los mensajes utilizando herramientas o algoritmos. Sin embargo, son conceptos sumamente parecidos. Ambos combinan el lenguaje natural con algoritmos de machine Learning para analizar un determinado texto. Lo cual, da la pauta para definir el próximo punto.[9]

El Machine Learning (o aprendizaje automático) fue definido en el 1959 por Arthur Samuel como “campo de estudio que da a las computadoras la capacidad de aprender sin ser programado explícitamente”. Posteriormente, Tom M. Mitchell, proporcionó una definición más formal: "Se dice que un programa informático aprende de la experiencia E con respecto a alguna clase de tareas T y la medida de rendimiento P, si su rendimiento en las tareas en T, medida por P, mejora con la experiencia E". [10] En síntesis es una rama de la inteligencia artificial que permite que las máquinas aprendan sin ser expresamente programadas para ello.

Sin embargo, ambas técnicas se deben utilizar en conjunto con otras técnicas o conceptos para maximizar los resultados positivos y minimizar los puntos negativos.

Uno de esos conceptos es el de Tokenización, lo cual se refiere al “proceso de separar una secuencia de cadenas en individuos como palabras, palabras clave, frases, símbolos y otros elementos conocidos como tokens. Las fichas pueden ser palabras individuales, frases o incluso frases enteras. En este proceso se descartan algunos caracteres como signos de puntuación. Los tokens funcionan como entrada para diferentes procesos como el análisis y la minería de texto”. [6]

Otro de los conceptos que complementan las técnicas expuestas anteriormente es el de “stop words” o palabras vacías, que son aquellos “objetos de una oración que no son necesarios en ningún sector de la minería de texto. Por lo tanto, generalmente estas palabras son ignoradas para mejorar la precisión del análisis” (por ejemplo: artículos, pronombres, preposiciones, etc.).[6]

Otro concepto importante que también es imperativo definir es el de “Bag of Words” (bolsa de palabras en español), que se trata de una lista de palabras útiles que ayudan a hacer más preciso el análisis del texto. [6]

Para concluir el marco teórico, se definirá un concepto que puede parecer redundante pero que, por su importancia en el trabajo, es necesario conocer, y es el concepto de Mercado Libre. Según información brindada por la propia empresa, se trata de “un ecosistema completo compuesto por Mercado Pago, Mercado Shops, Mercado Libre Publicidad y Mercado Envíos”, que busca ofrecer soluciones para que tanto individuos como empresas puedan comprar, vender, anunciar, enviar y pagar por bienes y servicios por Internet. [1]

Metodología

Enfoque Metodológico

El marco metodológico ha sido descripto por varios autores a lo largo de los años, sin embargo, aproximadamente en el año 2006, Balestrini lo define como “la instancia referida a los métodos, las diversas reglas, registros, técnicas y protocolos con los cuales una teoría y sus métodos calculan las magnitudes de lo real”.[11] De esta definición se obtiene que el marco metodológico es la estructura sistemática para la recolección y posterior análisis de información, que permite una interpretación de los resultados obtenidos a lo largo del proyecto en función del problema planteado con anterioridad.

Una vez realizado un acercamiento a lo que marco o enfoque metodológico se refiere, es posible continuar con el mismo. Recordando que el objetivo del presente informe es el estudio y el análisis de las diferentes técnicas o métodos del análisis de sentimiento orientados al procesamiento y clasificación de comentarios de productos de la plataforma de venta online Mercado Libre, se optó por emplear un diseño experimental, ya que, aunque el tema de

investigación tiene un sustento teórico bastante amplio, se desea ver que métodos son los que brindan mejores resultados, por lo que se debe realizar un proyecto de experimentación. Además, se trata de una investigación de tipo descriptiva en la cual se dará a conocer cuáles son los métodos de análisis de sentimiento más eficientes en relación con el problema planteado.

Para finalizar este apartado, hay una última cuestión que se debe dejar en claro, la cual se trata del enfoque que se le dará al presente trabajo. Con vistas en el problema presentado, el informe será diseñado bajo el enfoque cuantitativo, puesto que es el que mejor se adapta a las características del problema en cuestión, pues los resultados obtenidos serán analizados para la posterior formulación de conclusiones.

Como establecen Fernández y Baptista en la página 37 de su trabajo, "el enfoque cuantitativo utiliza la recolección de datos para probar hipótesis con base en la medición numérica y el análisis estadístico, con el fin establecer pautas de comportamiento y probar teorías".[12]

Técnicas, instrumentos/procedimientos

Antes de siquiera plantearnos comenzar a analizar los comentarios de un producto, debemos tener en claro cómo se pueden obtener dichos comentarios. Mercado Libre desarrolló una API la cual permite obtener todos los comentarios de un producto de una forma simple y rápida, de manera que solo debemos pasarle el ID del producto y la API retornará un archivo JSON con todos los datos de este. A modo de ejemplo se utilizó la herramienta Postman para acceder a la API y obtener los comentarios del producto, y la página web codebeautify.org para presentarlo de una manera más limpia y legible

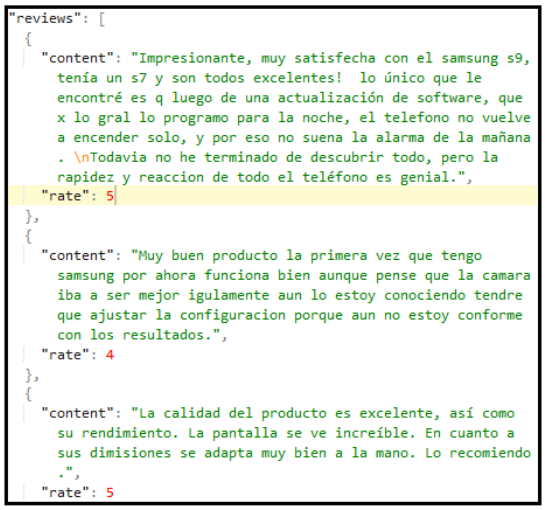
Accediendo a esta API podemos obtener los distintos comentarios de dicho producto.

https://api.mercadolibre.com/reviews/item/MLA723647586


```
  "title": "Excelente",
  "content": "La verdad un muy equipo pero si van a comprar
    algo de este precio les recomiendo pagar un poquito mas y
    obtener s9plus.",
  "rate": 5,
  "valorization": -5,
  "likes": 1,
  "dislikes": 6,
  "reviewer_id": 150947709,
  "buying_date": "2019-02-02T04:00:00Z",
  "relevance": 23,
  "forbidden_words": 0
},
{
  "id": 31680980,
  "reviewable_object": {
    "id": "MLA723647586",
    "type": "product"
  },
  "date_created": "2018-08-03T12:44:50Z",
  "status": "published",
  "title": "Muy bueno",
  "content": "Muy bueno,es rapido ,tiene una camara excelente
    , la bateria es un poco floja, pero estoy conforme muy
    buen producto.",
```


Una API se trata de "un conjunto de definiciones y protocolos que se utiliza para desarrollar e integrar el software de las aplicaciones. API significa interfaz de programación de aplicaciones".[13] Postman "es una plataforma de colaboración para el desarrollo de APIs, simplificando cada paso de la creación de la misma y optimizando la colaboración para que las mismas sean más rápidas y mejores".[14] Para finalizar, un archivo JSON es un "formato de texto sencillo para el intercambio de datos. Es un acrónimo de JavaScript Object Notation"[15]

Como se puede observar en la imagen, la API desarrollada por Mercado Libre nos brinda una amplia gama de información la cual no toda es necesaria para resolver el problema planteado, por lo que luego de obtener todos los datos, las reseñas del producto deben ser guardadas en otro archivo JSON, el cual consta del texto de la reseña y la puntuación otorgada por el revisor. En la siguiente imagen se puede apreciar la figura que tomarán los datos.



```
"reviews": [
  {
    "content": "Impresionante, muy satisfecha con el samsung s9,
    tenía un s7 y son todos excelentes!  lo único que le
    encontré es q luego de una actualización de software, que
    x lo gral lo programo para la noche, el telefono no vuelve
    a encender solo, y por eso no suena la alarma de la mañana
    . \nTodavia no he terminado de descubrir todo, pero la
    rapidez y reaccion de todo el teléfono es genial.",
    "rate": 5
  },
  {
    "content": "Muy buen producto la primera vez que tengo
    samsung por ahora funciona bien aunque pense que la camara
    iba a ser mejor igulamente aun lo estoy conociendo tendre
    que ajustar la configuracion porque aun no estoy conforme
    con los resultados.",
    "rate": 4
  },
  {
    "content": "La calidad del producto es excelente, así como
    su rendimiento. La pantalla se ve increíble. En cuanto a
    sus dimisiones se adapta muy bien a la mano. Lo recomiendo
    .",
    "rate": 5
```

Para finalizar con el proceso de obtención de los datos, a ese nuevo archivo generado, se lo debe someter a una serie de técnicas de Tokenización, en la cual, como se explicó anteriormente en el marco teórico, el comentario es separado en palabras clave, frases u otros elementos conocidos como tokens, y se les elimina las "stop words" como los signos de puntuación, artículos, preposiciones, etc.

Una vez que tenemos a nuestro alcance aquellas "reviews" de productos que queremos evaluar, toca continuar con la siguiente etapa del análisis de sentimiento. En esta próxima etapa debemos extraer y seleccionar las denominadas "text features" o características del texto, las cuales pueden involucrar las siguientes partes del texto:

- **Partes del habla** ("parts of speech o POS"): que se basa en encontrar los adjetivos que contiene el texto evaluado los cuales son indicadores importantes de opiniones.[16]
- **Palabras y frases de opinión**: estas son palabras comúnmente utilizadas para expresar opiniones, tales como bueno, malo, amo u odio. También se deben considerar expresiones o frases que expresan una opinión, pero que no emplean palabras de opinión, como, por ejemplo, "me costó un brazo".[16]
- **Negaciones**: la aparición de palabras negativas puede cambiar el sentido de un texto. Por ejemplo, si un texto dice que "no es bueno", sería lo mismo que escribir "malo".[16]
- **Presencia de términos**: se trata de palabras individuales que nos interesa saber si el texto posee o no. Se le otorga a la palabra un peso binario (un uno si la palabra aparece en el texto o viceversa)[16]

De estos ejemplos presentados se puede concluir que las "text features" se tratan de determinadas palabras o frases de corta longitud que se encuentran presentes en el texto que queremos analizar y que se deben identificar para llevar a cabo el procedimiento de análisis de sentimiento con éxito.

Para seleccionar las características del texto se pueden utilizar una serie de métodos muy diversos. Abarcar la totalidad de los métodos en un trabajo de corta duración resulta por demás imposible, por lo que se hará hincapié en solo algunas técnicas. Técnicamente hablando, podemos encontrar diferentes enfoques que se utilizan para dicha selección, dentro de los cuales podemos encontrar:

- Enfoques basados en el **Machine Learning**, los cuales utilizan una serie de algoritmos de aprendizaje para determinar el sentimiento de un texto, por medio del entrenamiento empleando un "dataset" de datos conocido. [17]
- Enfoques basados en el **léxico** ("Lexicon-based"), los cuales involucran calcular la polaridad de los sentimientos utilizando la orientación semántica de las palabras u oraciones.[17] Este tipo de enfoque por lo general requiere de anotación humana.[16]

- El enfoque basado **en reglas** busca palabras de opinión en un texto y luego las clasifica basado en el número de palabras positivas y negativas. Considera diferentes reglas para clasificación como polaridad de diccionario, palabras de negación, palabras de refuerzo, modismos, emoticonos, opiniones mixtas, etc.[17]
- El último enfoque se centra en **modelos estadísticos**, los cuales representan cada revisión como una mezcla de aspectos latentes y calificaciones, los cuales se asume que pueden ser representados por multinomiales distribuciones y tratar de agrupar términos principales en aspectos y sentimientos en calificaciones.[17] Este tipo de enfoques es completamente automático. [16]

Todas estas técnicas de selección de características tratan a los documentos como grupo de palabras ("bag of words") o como una cadena que conserva la secuencia de palabras en el documento, lo cual provee una gran simplicidad para el proceso de clasificación.

Una vez definidos los diferentes enfoques, es posible proseguir. Sin embargo, debido a la gran cantidad de información y a la amplia gama de métodos que cada enfoque provee, con el objetivo de sintetizar este informe, solo me centraré en los métodos del enfoque de modelos estadísticos y los enfoques basados en Machine Learning, ya que se trata de los enfoques que generalmente son más utilizados, sin mencionar que los modelos estadísticos, al estar basados en la teoría matemática, provoca que sus resultados posean una gran exactitud.[17]

Dentro del enfoque de modelos estadísticos, los enfoques más frecuentes son:

- **Chi-cuadrado (x^2):** ya he hablado de esta distribución en los antecedentes del proyecto. Básicamente Chi-cuadrado es un cálculo que se utiliza para determinar qué tan pequeña es la diferencia entre los datos observados y los datos esperados.[6] Para dicho cálculo, emplea la siguiente fórmula:

$$\chi_i^2 = \frac{n \cdot F(w)^2 \cdot (p_i(w) - P_i)^2}{F(w) \cdot (1 - F(w)) \cdot P_i \cdot (1 - P_i)}$$

  Donde n es el número total de documentos en la colección, pi(w) es la probabilidad condicional de clase i para los documentos que contienen la palabra w, Pi es la fracción de documentos de clase i, y F(w) es la fracción de documentos que contienen la palabra w. [16]
- **Latent Semantic Indexing (LSI):** se trata de un método de transformación de características, los cuales se basan en crear un conjunto más pequeño de características en función del conjunto original de características. Básicamente LSI transforma el espacio de texto en un nuevo sistema de ejes que son una combinación lineal de las características de la palabra original.[16]
- Hay otros enfoques estadísticos que podrían utilizarse, como "**Hidden Markov Model" (HMM) o "Latent Dirichlet Allocation" (LDA)**, los cuales fueron utilizados por Duric y Song para separar las entidades en un documento de opinión de las expresiones subjetivas que describen esas entidades.[16]

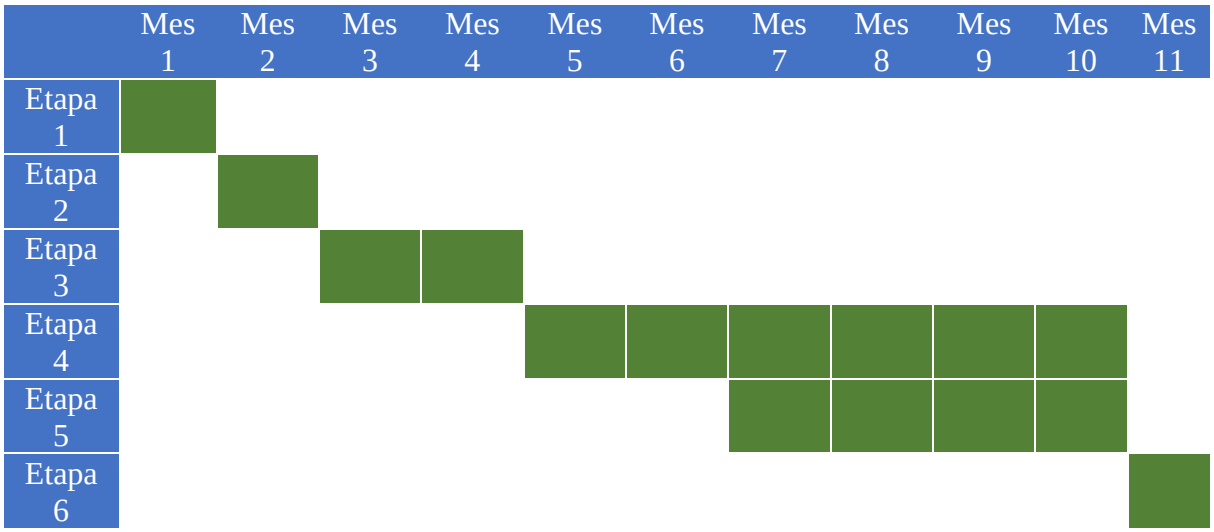
Dentro de los algoritmos de Machine Learning, **Support Vector Machine (SVM), Naïve Bayes** y **N-Gram.**

- **SVM** es un clasificador discriminativo formalmente definido por un separador de hiperplano.[18] En otras palabras, este algoritmo requiere para comenzar a funcionar que se la hayan proporcionado previamente unos datos de entrenamiento (por lo que se considera aprendizaje supervisado), y posteriormente genera un hiperplano óptimo que categoriza nuevos ejemplos.
- El clasificador **Naïve Bayesiano** se basa en el teorema de Bayes con supuestos de independencia entre predictores. En otras palabras, en este clasificador se asume que las variables predictoras son independientes entre sí, es decir, que la presencia de una cierta característica en un conjunto de datos no está en absoluto relacionada con la presencia de cualquier otra característica.[19] Este clasificador es particularmente útil para conjuntos de datos muy grandes, y a pesar de su simplicidad, es frecuentemente utilizado porque generalmente supera a otros métodos de clasificación más sofisticados.[18]
- **N-gram** es un modelo simple que asigna probabilidades a frases de palabras o de secuencias enteras. [18] Este tipo de modelo también puede ser considerado como un modelo estadístico, puesto que predicen la siguiente palabra en la oración usando las palabras n-x anteriores. [20] Por ejemplo, si tenemos la frase: "No puedo leer sin tener …", se puede inferir que la siguiente palabra será "anteojos" o “lentes”. N-gram predice la siguiente palabra en la secuencia usando la probabilidad condicional de la siguiente palabra.

Cabe destacar que no se entrará en detalle en lo respectivo a los últimos dos enfoques mencionados, debido a que, generalmente, no se utilizan para resolver este tipo de problemas. Sin embargo, es importante destacar el hecho de que, en el enfoque basado en el léxico, se utilizan palabras de opinión positivas para expresar algunos estados, mientras que las palabras de opinión negativas se utilizan para expresar estados no deseados. También hay frases de opinión y modismos que juntos se llaman “léxico de opinión”. Hay tres enfoques principales:

- El **enfoque manual** consume mucho tiempo y no se utiliza solo. Por lo general se combina con los otros dos enfoques automatizados.[16]
- Enfoque **basado en diccionarios**, en el cual un pequeño conjunto de palabras de opinión se recopila manualmente con orientaciones conocidas, y se busca en el texto que se desea evaluar sinónimos o antónimos a dichas palabras.[16]
- Enfoque **basado en el cuerpo**, el cual ayuda a resolver el problema de encontrar palabras de opinión con orientaciones específicas de contexto. Sus métodos dependen de patrones sintácticos o patrones que ocurren junto con una “lista semilla” de palabras de opinión para encontrar otras palabras de opinión en un cuerpo de una gran extensión.[16]

Plan de Trabajo



Se decidió desarrollar un plan de trabajo basado en una serie de etapas definidas:

- Etapa 1: enfocada principalmente a la revisión bibliográfica sobre el tema y a la realización de un análisis de antecedentes más exhaustivo.
- Etapa 2: en esta etapa se deben identificar y detectar los diferentes requerimientos que el sistema deberá contar.
- Etapa 3: etapa de análisis y diseño del sistema
- Etapa 4: etapa de codificación del sistema. Esta es la etapa del desarrollo propiamente dicha del sistema.
- Etapa 5: etapa de pruebas y experimentación.
  Cabe destacar que las etapas 4 y 5 se realizan en conjunto ya que se trata de un momento en el que se experimentará con los diferentes métodos de análisis de sentimiento desarrollados para la posterior observación y obtener conclusiones al respecto.
- Etapa 6: análisis de resultados y conclusiones sobre los mismos

Resultados Esperados

Como se indicó al inicio de este informe, uno de los objetivos de la investigación era identificar y describir que técnicas de análisis de sentimientos existen hoy en día y enunciar cual de todas puede ser utilizada para resolver el problema en cuestión.

Una vez que el plan de trabajo haya finalizado se espera poder analizar claramente los resultados obtenidos, siendo capaces de determinar claramente que método de análisis de sentimiento es el que arroja mejores resultados para solucionar el problema en cuestión.

Un resultado que se puede esperar, que luego será validado o no por el propio sistema, es que el uso de la distribución Chi-Cuadrado es un enfoque por demás correcto, y que además se sabe que ya se ha utilizado para resolver problemáticas de esta índole, por lo que sabemos que su influencia en el proyecto es positiva. Es por este motivo que se esperan buenos resultados con el empleo de esta técnica.

Referencias

[1] MercadoLibre, “Historia de Mercado Libre: conocé todo sobre la compañía | IDEAS Mercado Libre Argentina,” *Ideas Mercado Libre*. 2020, [Online]. Available: https://ideas.mercadolibre.com/ar/noticias/historia-de-mercado-libre/.

[2] M. V Mäntylä, D. Graziotin, and M. Kuutila, “The Evolution of Sentiment Analysis,” *Comput. Rev.*, vol. 27, no. February, pp. 16–32, 2018, [Online]. Available: https://doi.org/10.1016/j.cosrev.2017.10.002.

[3] K. Sonai, M. Anbananthen, and A. M. H. Elyasir, “Evolution of Opinion Mining,” *Aust. J. Basic Appl. Sci.*, vol. 7, no. 6, pp. 359–370, 2013.

[4] C. Rain, “Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning,” *Swart. Coll.*, 2013, doi: 10.1097/IOP.0b013e318213f5d9.

[5] A. Bhatt, A. Patel, H. Chheda, and K. Gawande, “Amazon Review Classification and Sentiment Analysis,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 6, pp. 5107–5110, 2015.

[6] T. U. Haque, N. N. Saber, and F. M. Shah, “Sentiment analysis on large scale Amazon product reviews,” *2018 IEEE Int. Conf. Innov. Res. Dev. ICIRD 2018*, no. June 2019, pp. 1–6, 2018, doi: 10.1109/ICIRD.2018.8376299.

[7] “Sentiment Analysis: Concept, Analysis and Applications | by Shashank Gupta | Towards Data Science.” [Online]. Available: https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17.

[8] B. Liu and S. M. Street, “Opinion mining,” no. 1, pp. 1–7.

[9] R. Baldania, “What is the Difference between opinion mining and sentiment analysis?” 2017, [Online]. Available: https://www.quora.com/What-is-the-Difference-between-opinion-mining-and-sentiment-analysis.

[10] P. Dönmez, “Introduction to Machine Learning, 2nd ed., by Ethem Alpaydın. Cambridge, MA: The MIT Press2010. ISBN: 978-0-262-01243-0. $54/£ 39.95 + 584 pages.,” *Nat. Lang. Eng.*, vol. 19, no. 2, pp. 285–288, 2013, doi: 10.1017/s1351324912000290.

[11] Varios, “Capitulo III: Marco Analìtico,” pp. 83–101, 2008, [Online]. Available: http://virtual.urbe.edu/tesispub/0094671/cap03.pdf.

[12] *R. Hernandez Sampieri, Carlos Fernandez, Pilar Baptista, "Metodología de la investigación", 6ta edición. 2014*

[13] Red Hat, “¿Qué es una API?,” *Documentación proporcionada por Red Hat sobre el Concepto de API (application programming interface)* . 2019, [Online]. Available: https://www.redhat.com/es/topics/api/what-are-application-programming-interfaces.

[14] o.V., “Introduction | Postman Learning Center.” [Online]. Available: https://learning.getpostman.com/docs/postman/launching-postman/introduction/.

[15] “JSON - Wikipedia, la enciclopedia libre.” . [online]. Available: https://es.wikipedia.org/wiki/JSON

[16] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.

[17] S. W. Reilly and I. Catton, “Utilization of pore-size distributions to predict thermophysical properties and performance of biporous wick evaporators,” *J. Heat Transfer*, vol. 136, no. 6, 2014, doi: 10.1115/1.4026624.

[18] M. Ridzwan Yaakub, M. Iqbal Abu Latiffi, and L. Safra Zaabar, “A Review on Sentiment Analysis Techniques and Applications,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 551, no. 1, 2019, doi: 10.1088/1757-899X/551/1/012070.

[19] Victor Roman, “Algoritmos Naive Bayes: Fundamentos e Implementación | by Victor Roman | Ciencia y Datos | Medium.” 2019, [Online]. Available: https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementación-4bcb24b307f.

[20] “nlp - N-GRAMS nlp Tutorial.”. [online]. Available: https://riptutorial.com/es/nlp/topic/8851/n-grams