Desarrollo de un algoritmo para el análisis de sentimientos de comentarios de clientes sobre productos de tecnología de Alkosto/Ktronix

Gabriela Lara
Facultad de Ingeniería de Sistemas e Informática
UPB Bucaramanga
Bucaramanga, Colombia
gabriela.lara.2019@upb.edu.co

Luisa Monsalve
Facultad de Ingeniería de Sistemas e Informática
UPB Bucaramanga
Bucaramanga, Colombia
luisa.monsalve.2019@upb.edu.co

Silvia Coy
Facultad de Ingeniería de Sistemas e Informática
UPB Bucaramanga
Bucaramanga, Colombia
silvia.coy.2019@upb.edu.co

Yireth Aldana
Facultad de Ingeniería de Sistemas e Informática
UPB Bucaramanga
Bucaramanga, Colombia
yireth.aldana.2017@upb.edu.co

***Resumen*— Debido a la influencia que tienen los comentarios en las decisiones de compra online, se analizó el comportamiento de los clientes con los productos que pertenecen al sector de tecnología, por medio de un análisis de sentimientos dentro del área de estudio del Procesamiento de Lenguaje Natural (NLP). Se llevó a cabo utilizando la técnica de extracción de datos llamada Web Scraping con la que extraen los datos necesarios de la página web de Alkosto/Ktronix en el lenguaje HTML luego, se aplicó el procesamiento de lenguaje natural, tokenización, filtro de stopwords y regresión logística con el fin de encontrar patrones que permitan definir la polaridad de los comentarios.**

Palabras Clave*— *Polaridad, Lenguaje Natural, Machine Learning, web scraping.

***Abstract*— Due to the influence that feedback has on online purchase decisions, customer behavior with products belonging to the technology sector was analyzed by means of a sentiment analysis within the Natural Language Processing (NLP) study area. It was carried out using the technique of data extraction called Web Scraping with which the necessary data is extracted from the Alkosto/Ktronix web page in the HTML language. Then, natural language processing, tokenization, stopword filtering and logistic regression were applied in order to find patterns that allow the definition of the polarity of the comments.**

Keywords*— *Polarity, Natural Language Processing, Machine Learning, Web Scraping.

I. INTRODUCCIÓN

Alkosto líder en tecnología lanza en 1998 Ktronix como la única tienda multimarca especializada en electrónica y tecnología, ellos velan por la capacitación completa al personal, tanto en conocimiento de producto como en parámetros de atención y servicio. En la actualidad tiene clientes en varias partes de Colombia con sus tiendas en Bogotá (11), Medellín (3), Bucaramanga (1), Cali (1), Villavicencio (1), Tunja (1), Manizales (1). Por otra parte, en 2010 deciden abrir una tienda virtual con más de mil seiscientas (1.600) referencias de electrodomésticos, tecnología, accesorios y consumibles.

Debido a la influencia que tienen los comentarios en las decisiones de compra online, Alkosto/Ktronix busca analizar el comportamiento de los clientes con los productos que se encuentran en el sector de tecnología, por medio de un análisis de sentimientos dentro del área de estudio del procesamiento de Lenguaje Natural (NLP).

Desarrollar un algoritmo que por medio del Procesamiento de Lenguaje Natural (NLP) realice un análisis de sentimientos a partir de comentarios de clientes sobre productos de tecnología de Alkosto/Ktronix, con el fin de clasificar de manera automática los comentarios con una connotación positiva o negativa.

II. FUNDAMENTACIÓN TEÓRICA

Esta sección se divide en cuatro temáticas de estudio, consideradas relevantes para el desarrollo de la investigación: Análisis de Sentimientos, Procesamiento del Lenguaje Natural, Machine Learning, Web Scrapping.

A. Análisis de Sentimientos

El Análisis de Sentimientos es un campo de investigación dentro del PLN que trata de extraer de manera automática información subjetiva expresada en el texto de un documento, y de esta forma, por medio de la polaridad identificar si la información extraída se clasifica como una connotación positiva o negativa [1]. Cabe resaltar que este tipo de procesamiento de la información generalmente se basa en información estadística y no en un análisis lingüístico.

Para dejar en claro el término “polaridad”, se explicará a continuación:

La polaridad es un proceso de clasificación utilizado para etiquetar fragmentos de un texto como positivo, negativo o neutro [2]. Actualmente la polaridad se presenta por medio de las reacciones que ofrecen diferentes redes sociales (Facebook, Twitter, YouTube, etc.). Por ejemplo:

Fig. 1. Reacciones que ofrece la Red Social Facebook [3]

B. Procesamiento del Lenguaje Natural

El Procesamiento del Lenguaje Natural o Natural Language Processing (NLP) es un campo de estudio dentro de la inteligencia artificial que mediante la comprensión y generación de lenguaje natural intenta que las máquinas entiendan, interpreten y manipulen el lenguaje humano [4]. En la mayoría de los casos se asocia NLP con la creación de chatbots o asistentes virtuales, sin embargo, existen más utilidades como: resumen de textos, donde busca encontrar la idea principal del texto e ignorar lo irrelevante; traducción automática, por medio de la implementación de NLP en un sistema logra traducir textos en diferentes idiomas; reconocimientos de entidades, el cual pretende encontrar personas, entidades comerciales, ciudades o marcas [5].

C. Machine Learning

Es una disciplina del campo de la inteligencia artificial que crea sistemas de aprendizaje automático y por medio de algoritmos que identifican patrones complejos en datos masivos es capaz de predecir comportamientos futuros

[6]. Existen tres tipos de aprendizaje automático:

- Aprendizaje supervisado: Se entrena al sistema proporcionándole datos con etiquetas detalladas que le permiten asociar y clasificar datos sin etiquetas.

- Aprendizaje no supervisado: Se basa en la compresión y abstracción de patrones de información de manera directa.

- Aprendizaje por refuerzo: Esta técnica se basa en prueba y error donde se optimiza el comportamiento del sistema a partir de la experiencia [7].

D. Web Scrapping

Es una técnica que permite extraer información de páginas web de manera automática a través de algoritmos de búsqueda con los que se puede rastrear la información que se desea extraer [8]. Algunas de sus aplicaciones son [9]:

- Marketing de contenidos
- Monitorización de la competencia (contexto de negocio)
- Reputación de paginas

III. ESTADO DEL ARTE

D. J. Calatrava [10] en 2019, afirmó en su trabajo de grado que a partir de los resultados en su chiclet slicer (visualizador en el que se categorizaron los tweets), gráfico de columnas agrupado, funnel (visualizador utilizado para analizar los tweets y cantidad de usuarios que generaron tweets en un determinado día), indicador (promedio de sentimientos que se han generado con los tweets) y gráfico circular, se logró deducir y comprender el comportamiento de los usuarios ante una marca o producto.

A. I. Yañez [11] en 2019, una de las funciones utilizadas para el tratamiento del lenguaje natural fue la eliminación de stopwords(palabras que no tienen un significado por sí solas, suelen ser: artículos, pronombres, preposiciones, adverbios y a veces verbos), aseguró que es un paso común en un proceso de tratamiento para reducir la dimensionalidad del problema eliminando palabras que no aportan valor a la polaridad (valor que adopta la opinión, ya sea de forma positiva, negativa o neutral) del documento. Para ello definió un diccionario de palabras comunes compuesto por preposiciones, conjunciones, pronombres y ciertas conjunciones de verbos comunes.

F. N. Machado [12] en 2018, utilizó Ixa Pipes un conjunto de módulos para el procesamiento del lenguaje natural creada por el grupo IXA NLP de la Universidad del País Vasco. Se puso en práctica esta herramienta por sus características, las cuales incluyen la tokenización del texto en palabras, segmentación del texto en frases, análisis morfológico, reconocimiento de nombres de entidades (NER). Los resultados después del uso de la herramienta fueron aceptables, por ende, hubo un manejo correcto de las herramientas.

V. A. Hernández [13] en 2015, describió el Corpus para la detección de ironía como un detalle importante que debe poseer los datos suficientes de entrenamiento para así poder realizar un aprendizaje del algoritmo supervisado de clasificación. Debido a que resulta difícil etiquetar lo datos manualmente, se aprovechó la existencia de hashtags en Twitter para realizar una recolección rápida de datos que además tenga relación con la ironía verbal.

G. A. Molina [14] en 2017, hizo uso de Clasificadores Bayesianos, ya que estos permiten realizar la clasificación de

las expresiones idiomáticas con las que sean entrenados. El autor mencionó que los clasificadores que realizan la categorización entre positivo y negativo, son binarios ya que poseen solamente dos clases para realizar la clasificación. Por otro lado, los que tienen la capacidad de clasificar entre diferentes clases son clasificadores multiclase, los cuales también fueron aplicados en el proyecto.

P. P. Agustin [15] en 2019, realizo un análisis comparativo de algoritmos predictivos y métodos utilizando un lexicón español, con el fin de evaluar con cuál de ellos se obtiene un mejor resultado en términos de precisión y fiabilidad. Por medio de los resultados del análisis se lograron obtener los siguientes datos:

	Exactitud	Precisión positivos	Precisión negativos
Naive Bayes	70,64%	70,6%	79,6%
Regresión Logística	81.55%	81%	74,3%
SVM	72,3%	75,5%	71%

Fig. 2. Tabla de resultados finales de los algoritmos utilizados [15]

El autor concluyó que utilizando el método de Regresión Logística se obtuvo un mejor porcentaje de exactitud comparado a los demás, que, aunque obtuvieron menor valor tampoco dejan de ser buenos resultados. Además, logro observar que el método de Naive Bayes obtuvo un menor puntaje en términos de exactitud, pero al mismo tiempo fue más preciso en la clasificación de comentarios negativos.

IV. METODOLOGIA

A. *Investigación preliminar*

Se identifica el mercado de Alkosto/Ktronix, y por toma de datos a través de la página web extrayendo las características de cada uno de sus productos y distribución.

B. *Especificación de requerimientos y definición de clases*

Se busca el límite del proyecto con respecto a la información requerida anteriormente y con lo que quiere el cliente.

C. *Extraer los datos requeridos por medio del método Web Scrapping*

Se extraen las características de cada producto como lo es el nombre del producto, el precio, nombre del usuario (revisar que esté disponible), calificación incluida en cada comentario y el cuerpo del comentario. Para poder extraer las estrellas, el cuerpo y el autor se crea un diccionario ya que estos no se permitían ver a simple vista.

D. *Preparación de datos y modelamiento del algoritmo*

Se limpiaron los datos extraídos en el paso anterior, se hicieron las conversiones necesarias, se elaboraron los diagramas correspondientes y los datos se suben a una base de datos que va a contener los comentarios de los clientes.

E. *Análisis de datos*

En este punto se define la polaridad de los comentarios con la clasificación que cada uno tiene a partir de las estrellas.

F. *Validación y visualización de datos*

Se revisa la exactitud y precisión del algoritmo para al final permitir la visualización de la gráfica porcentual de los comentarios.

V. RESULTADOS

En la siguiente sección se presentan los resultados obtenidos al aplicar los procesos metodológicos detallados en la sección anterior, siguiendo el mismo orden presentado en la sección:

A. *Diagrama de barras del porcentaje de las estrellas*

Esta gráfica tiene como finalidad dar a conocer la clasificación de los comentarios en un intervalo de 1 a 5, permitiendo identificar que gran parte de los comentarios se relacionaban con una puntuación mayor a 4. Sin embargo, los comentarios con puntuación de 3 fueron descartados en la aplicación del modelo debido a que hacían referencia a sentimientos neutros.

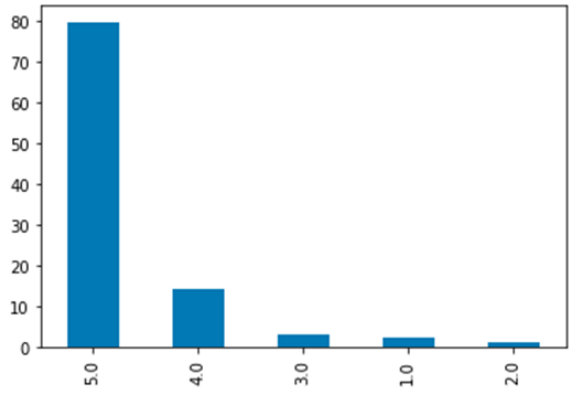

Fig. 3. Gráfica porcentual de calificación de estrellas [16]

B. Diagrama de barras de la polaridad del comentario

Esta gráfica representa el porcentaje de comentarios con connotación positiva y negativa, los cuales son indicados con clasificación binaria, donde el 0 hace referencia a los comentarios relacionados con la puntación de 1 y 2 estrellas relacionadas con el sentimiento negativo, y el 1 que corresponde a comentarios positivos clasificados con más de 4 estrellas.

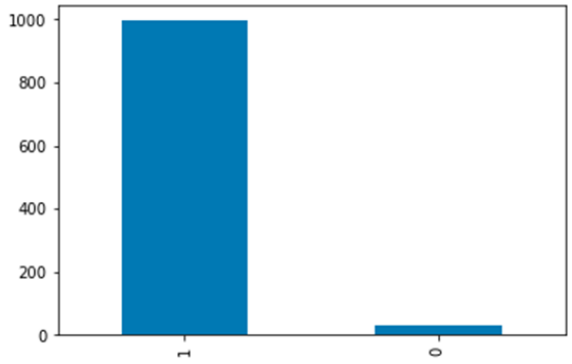

Fig. 4. Gráfica de polaridad de los comentarios [16]

C. Nube de palabras positiva y negativa

Las nubes de palabras de las figuras 5 y 6 hacen referencia a las 10 palabras más usadas en comentarios positivos y negativos, respectivamente. Se observa que la nube de palabras negativas resalta las palabras "malo" y "dañado", es decir, que el producto no fue de su agrado. Por otro lado, en la nube de palabras positivas las palabras que más resaltan son "excelente" y "buena", es decir, que el producto ha gustado y sus comentarios son positivos.

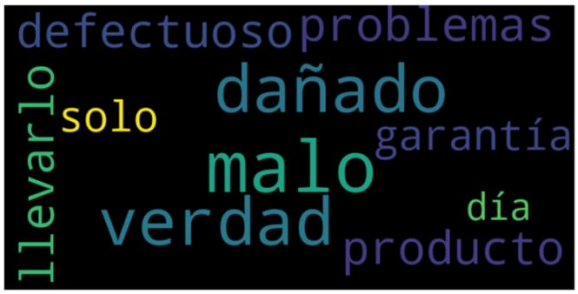

Fig. 5. Nube de palabras positivas [16]

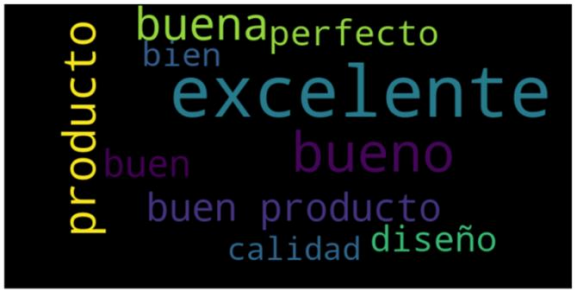

Fig. 6. Nube de palabras negativas [16]

D. Matriz de correlación existente entre producto y precio

Por medio de la técnica de visualización de datos llamada matriz de correlación, se representa la correlación existente entre el producto y su respectivo precio. La matriz de correlación está dada en un rango de [-1,1] y permite observar que el resultado obtenido fue 0.083 lo que refleja que la correlación existente es nula a partir de la tabla de la figura 7. A continuación, se muestra la matriz mencionada anteriormente, figura 8.

±0.00	±0.09	Correlación nula
±0.10	±0.19	Correlación muy débil
±0.20	±0.49	Correlación débil
±0.50	±0.69	Correlación moderada
±0.70	±0.84	Correlación significativa
±0.85	±0.95	Correlación fuerte
±0.96	±1.0	Correlación perfecta

Fig. 7. Tabla de valores de matriz de correlación [16]

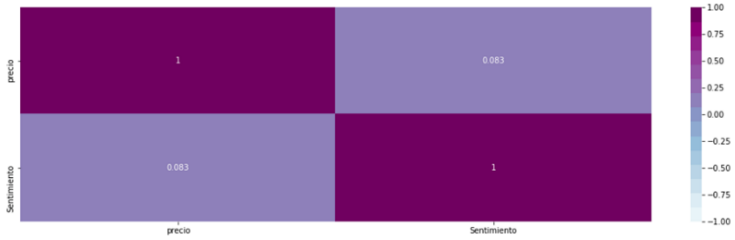

Fig. 8. Matriz de correlación existente entre precio y sentimiento [16]

E. Vectorización

La bolsa de palabras es una técnica conocida como representación basada en el conteo la cual busca analizar documentos señalando la frecuencia con la que se producen ciertas estructuras. Esta técnica se realiza a través de una matriz de coocurrencia donde automáticamente tiene en cuenta el preprocesamiento. A continuación, se muestra la implementación en el algoritmo, figura 9.

	02	10	100	1100	15	18	1a	20	2018	2020	...	ya	zenfone	zoom	ágil	él	óptima
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0

Fig. 9. Bolsa de palabras [16]

F. Tabla de top 10 de palabras con más peso en la connotación positiva y negativa

Por medio de las tablas de las figuras 1 y 2 se refleja el peso de las 10 palabras que más influyen en la clasificación de sentimientos. En la figura 10 se encuentra la tabla con las

palabras positivas y en la figura 11 las tablas con las palabras negativas.

	palabra	peso
2182	excelente	1.259067
599	buen	0.774890
726	buena	0.691694
3587	perfecto	0.595172
498	bien	0.568011
3811	producto	0.486684
1707	diseño	0.412745
661	buen producto	0.271591
4468	super	0.252209
842	bueno	0.246948

Fig. 10. Tabla de palabras más frecuentes en comentarios positivos [16]

	palabra	peso
2977	lento dañado	-0.464809
3278	microfono	-0.464809
2976	lento	-0.464809
1584	defectuoso	-0.508203
1869	día	-0.570670
3088	mala calidad	-0.841810
3087	mala	-0.891249
1565	dañado	-1.332061
4103	regular	-1.716049
3092	malo	-1.950353

Fig. 11. Tabla de palabras más frecuentes en comentarios negativos [16]

G. Matriz de confusión

La matriz de confusión es una herramienta que nos permite visualizar el desempeño del algoritmo de aprendizaje supervisado. Esta matriz se divide en cuatro partes que son:

- Verdaderos positivos: Son aquellos comentarios positivos que se analizaron correctamente.
- Falsos positivos: Son aquellos comentarios que se analizaron como positivos, pero eran negativos.
- Falsos negativos: Son aquellos comentarios que eran positivos, pero fueron clasificados como negativos.
- Verdaderos Negativos: Se refiere a los comentarios negativos clasificados como tal.

Según los datos obtenidos de la matriz de confusión (figura 12) y la métrica F1 score, el algoritmo tiene un desempeño del 93%.

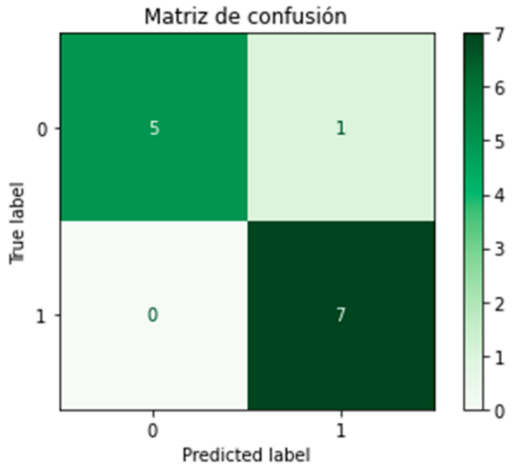

Fig. 12. Matriz de confusión [16]

H. Gráfico de barras porcentuales de métricas de exactitud

En la gráfica presentada en la figura 13 se refleja el porcentaje de las métricas del modelo las cuales son: Precisión, exactitud y sensibilidad

- Precisión: Son los casos clasificados de manera correcta por el modelo. En este caso, el modelo tiene una precisión del 55%.
- Exactitud: Es el porcentaje de casos en los que el modelo ha acertado. Según los resultados la exactitud del modelo es del 62%.
- Sensibilidad: Es la proporción de casos positivos clasificados correctamente respecto al total de casos positivos. A partir de los resultados obtenidos se puede concluir que el modelo es en su totalidad sensible debido a que tiene un porcentaje del 100%.

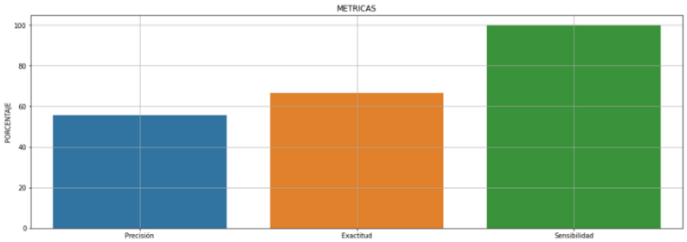

Fig. 13. Gráfica porcentual de las métricas del modelo [16]

VI. DISCUSIÓN

Hoy en día las empresas están preocupadas del impacto que tienen sus productos en los clientes, entonces, lo que buscamos es que la empresa Alkosto/Ktronix pueda visualizar de manera más sencilla el impacto que tienen sus productos en la sociedad, es decir, que no deban revisar uno por uno si no que puedan revisar todos a partir de un análisis estadístico. Como limitaciones tendríamos que es un aplicativo de escritorio y está en el lenguaje de Python, entonces la persona tendría que conocer este lenguaje de lo contrario no entendería el código, otras restricciones es que el sistema debe ser capaz de determinar la polaridad de los comentarios y debe generar una gráfica con una escala de calificación.

VII. CONCLUSIONES

El algoritmo para el análisis de sentimientos expresados en comentarios de la plataforma de Alkosto/Ktronix que fue propuesto permitió reflejar la polaridad mediante la extracción de datos por medio de la técnica de Web Scrapping, preprocesamiento, entrenamiento del modelo de regresión logística, validación y pruebas de este. El modelo mencionado anteriormente se desarrolló a partir de las técnicas de TF-IDF y bolsa de palabras las cuales permitieron la clasificación. Se consideraron tres métricas para medir el desempeño del modelo obteniendo como resultado:

una precisión del 55%, una exactitud del 62% y una sensibilidad del 100%.

A través de los resultados anteriores se concluyó que el modelo presenta predicciones más sensibles a clasificar comentarios con connotación positiva que negativa, Sin embargo, demostró tener un porcentaje de asertividad aceptable con respecto a la polaridad de comentarios en general.

REFERENCIAS

[1] J. Sobrino, «Universitat Oberta de Catalunya,» Junio 2018. [En línea]. Available: http://openaccess.uoc.edu/webapps/o2/bitstream/10609/81435/6/jsobrinosTFM0618memoria.pdf. [Último acceso: 22 Octubre 2020].

[2] M. A. A. Fernández, «DSpace@UCLV,» 12 Junio 2016. [En línea]. Available: http://dspace.uclv.edu.cu:8089/handle/123456789/9214. [Último acceso: 20 Octubre 2020].

[3]«Wikiwand,» [En línea]. Available: https://www.wikiwand.com/es/An%C3%A1lisis_de_sentimiento. [Último acceso: 20 Octubre 2020].

[4] «Decide Soluciones,» 12 Septiembre 2019. [En línea]. Available: https://decidesoluciones.es/procesamiento-del-lenguaje-natural-pln-o-nlp-que-es-y-para-que-se-utiliza/. [Último acceso: 20 Octubre 2020].

[5] «Aprende Machine Learning,» 22 Octubre 2020. [En línea]. Available: https://www.aprendemachinelearning.com/. [Último acceso: 27 Octubre 2020].

[6] «Iberdrola,» 2020. [En línea]. Available: https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico. [Último acceso: 27 Octubre 2020].

[7] Redacción APD, «apd,» 4 Marzo 2019. [En línea]. Available: https://www.apd.es/que-es-machine-learning/. [Último acceso: 20 Octubre 2020].

[8] M. Martí, «sitelabs,» 8 Abril 2016. [En línea]. Available: https://sitelabs.es/web-scraping-introduccion-y-herramientas/. [Último acceso: 20 Octubre 2020].

[9] A. Lafuente, «aukera,» 10 Enero 2019. [En línea]. Available: https://aukera.es/blog/web-scraping/. [Último acceso: 20 Octubre 2020].

[10] J. D. Tenorio, «Repositorio Digital Universidad de las Américas,» 2019. [En línea]. Available: http://dspace.udla.edu.ec/handle/33000/10600. [Último acceso: 09 2020].

[11] A. I. Yañez, «Repositorio Universidad de Coruña,» 2019. [En línea]. Available: http://hdl.handle.net/2183/25152. [Último acceso: 09 2020].

[12] F. Nantes, «Repositorio Institucional Universidad de la Laguna,» 03 09 2018. [En línea]. Available: https://riull.ull.es/xmlui/bitstream/handle/915/10412/Analisis%20de%20sentimientos%20basado%20en%20opiniones%20turisticas..pdf?sequence=1&isAllowed=y. [Último acceso: 09 2020].

[13] V. A. Hernandez, «Repositorio Academico de la Universidad de Chile,» 2015. [En línea]. Available: http://repositorio.uchile.cl/handle/2250/134793. [Último acceso: 10 2020].

[14] R. D. USCG, «Julio Oswaldo Vasconez,» 21 03 2017. [En línea]. Available: http://repositorio.ucsg.edu.ec/handle/3317/7649. [Último acceso: 10 2020].

[15] P. P. Agustin, «Repositorio ITBA Principal,» 29 Julio 2019. [En línea]. Available: https://ri.itba.edu.ar/bitstream/handle/123456789/1782/Proyecto%20Final.pdf?sequence=1&isAllowed=y. [Último acceso: 1 Noviembre 2020].

[16] G. Lara, L. Monsalve, Y. Aldana y S. Coy, Gylus Comment, Bucaramanga, 2020.