Ingeniería y
Competitividad
Vol 25 (3) 2023; doi: 10.25100/iyc.v25i3.13220

Vol 25 (3)

ISSN
0123-3033
e- 2027-8284

Artículo de investigación

Categorización e integración de columnas de opinión y contenido de páginas web aplicando técnicas de Procesamiento de Lenguaje Natural

Categorization and Integration of Opinion Columns Content in Web Pages Applying Natural Language Processing Techniques

Jorge-Alexander Acevedo-Castiblanco[1] Marco-Javier Suárez-Barón[1] Juan-Sebastian Gonzalez-Sanabriaz[1]

[1]Universidad Pedagógica y Tecnológica de Colombia, Escuela de Ingeniería de Sistemas y Computación, Sogamoso, Colombia.

Resumen

Se presenta la aplicación de técnicas de Procesamiento de Lenguaje Natural para el análisis de textos, describiendo el proceso realizado desde la extracción de datos hasta la identificación y detección de opiniones de manera automática. Los textos analizados fueron columnas de opinión que reflejan los criterios de las personas sobre temas de actualidad. Lo anterior con el fin de proporcionar una manera ágil de identificar los temas de interés en la comunidad para proporcionar a los interesados de forma resumida lo que se expresa sobre estos temas. Para tal fin, se implementó un algoritmo que permite extraer información de manera precisa y limpia desde páginas web y posteriormente otro algoritmo que se encarga de efectuar la categorización automática de la información extraída, generando un resumen preciso de los principales temas en cada escrito.

Abstract

The application of Natural Language Processing techniques for text analysis is presented, describing the process carried out from data extraction to the identification and detection of opinions automatically. The texts analyzed were opinion columns that reflect the criteria of people on current issues. The foregoing to provide an agile way to identify topics of interest in the community to provide those interested in a summary of what is expressed on these topics. For this purpose, an algorithm was implemented that allows information to be extracted accurately and cleanly from web pages and later another algorithm that oversees carrying out the automatic categorization of the extracted information, generating an accurate summary of the main topics in each writing.

Keywords:
Text Classification, Opinion Columns, Natural Language Processing, Web Scrapping

Palabras clave:
Clasificación de texto, Columnas de opinión, Procesamiento de Lenguaje Natural, Web Scrapping.

Cómo citar:
Acevedo-Castiblanco, J.A., Suárez-Barón, M.J., Gonzalez-Sanabriaz, J.G. Categorización e integración de columnas de opinión contenido en páginas web aplicando técnicas de Procesamiento de Lenguaje Natural. Ingeniería y Competitividad,2023, 25(3); e-22313220. doi: https://doi.org/ 10.25100/iyc.v25i3.13220

Recibido: 09-05-23
Aceptado 11-13-23

Correspondencia:
jorge.acevedo01@uptc.edu.co, marco.suarez@uptc.edu.co, § juansebastian.gonzalez@uptc.edu.co

Conflicto de intereses:
Ninguno declarado

Contribución

¿Por qué se realizó?

Se realiza el trabajo dado que la mayoría de los métodos de caracterización de textos actuales que requieren numero considerable de ejemplos para incluir un nuevo tema o un riguroso entrenamiento en columnas de opinión; así pues, para nuestro caso solo es necesario crear un registro del nuevo tema en la base de datos, y el algoritmo automáticamente lo detectará y validará mediante el modelo Bloom. Esto representa un avance significativo que simplifica y agiliza el proceso de ampliar el alcance del análisis y caracterización de las columnas de opinión lo que repercute en menores costos.

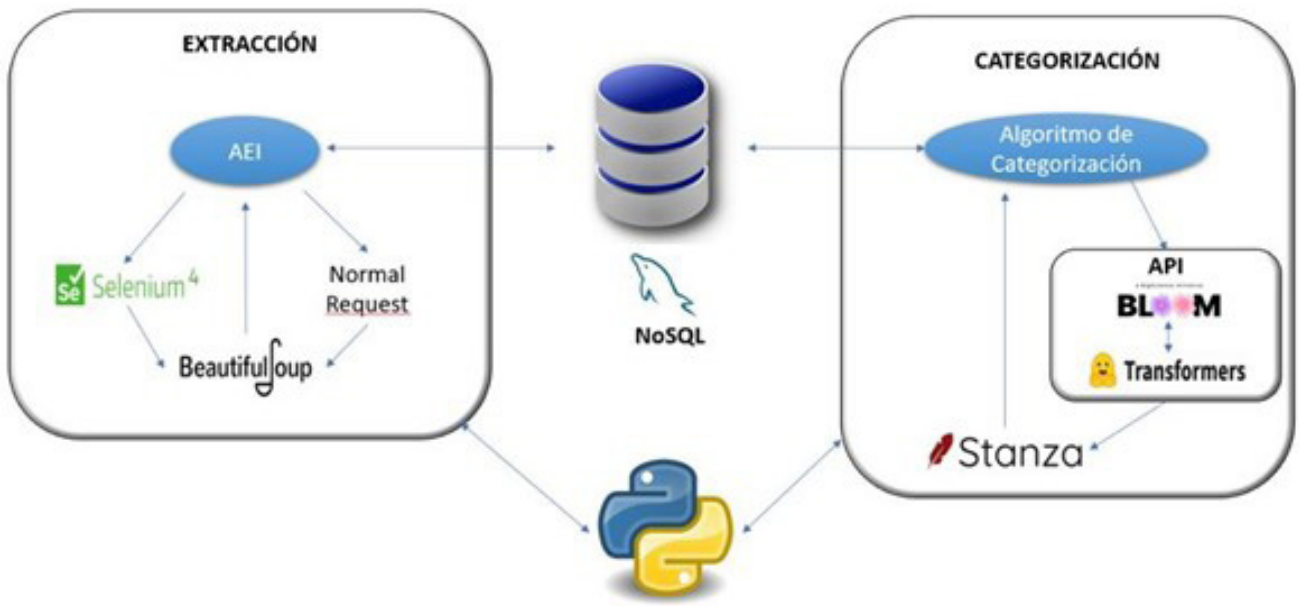
¿Cuáles fueron los resultados más relevantes?

Este trabajo desarrollo un algoritmo parta extraer datos automáticamente de Internet a partir de un alto volumen de información no estructurada que se encuentra disponible. Mediante el desarrollo se recupera e integra la información de manera organizada, permitiendo la aplicación de filtros y la posibilidad de realizar limpieza durante o al final del proceso de extracción. Se ha logrado crear una herramienta escalable y útil que puede ser aplicada en diversos campos de extracción e integración de información.

¿Qué aportan estos resultados?

El aporte se evidencia en la capacidad de proporcionar la información relevante de una columna de opinión y de esta manera identificar la apreciación de la sociedad respecto a una organización o persona en particular, identificar los temas de actualidad o comprender en qué temas se especializa un determinado autor.

Graphical Abstract



Introducción

El Procesamiento del Lenguaje Natural (PLN) presenta un gran conjunto de técnicas que ofrecen una amplia gama de ventajas como puede ser la capacidad de resumir y obtener la idea principal de textos complejos (1). Por esta razón, aplicar PLN para la automatización en el análisis de columnas de opinión puede convertirse en una herramienta de gran utilidad para la optimización de procesos y el análisis de textos, en una sociedad donde día a día se incrementa el número de publicaciones de diferentes temas en la web (2).

Ahora bien, para dicho análisis se requiere realizar un proceso de extracción de información de diversas fuentes, proceso para el que existen varias herramientas que suelen funcionar mediante un modelo de negocio, es decir, son provistas con una suscripción mensual con limitaciones en el número de peticiones o con un tamaño máximo en la cantidad de datos que se puede manipular (3). Adicionalmente, la información extraída por estas herramientas suele contener caracteres o datos incoherentes debido al análisis y decodificación genérico de etiquetas, ocasionando que se añadan, quiten o alteren los símbolos originales y, por ende, la información queda inservible e incoherente, desencadenando en la categorización incorrecta de la información (4).

Por otra parte, la escasez de modelos entrenados en español para la categorización de columnas de opinión dificulta el proceso, obligando a utilizar técnicas con un alto margen de error (5). Esta última dificultad puede deberse a la gran cantidad de terminologías en el idioma español (6), sumada a su profunda semántica y la capacidad de que cada columna aborde múltiples temas simultáneamente, lo cual representa un desafío para que un modelo pueda determinar un resultado preciso. Sumado a lo anterior, se encuentra el desafío de los acentos, conjugaciones y el uso del sarcasmo en el español, que añaden un nivel adicional de complejidad al proceso de categorización.

Por lo anterior, el objetivo principal de este trabajo se centró en obtener las columnas de opinión encontradas en páginas web e identificar los temas relevantes abordados en su contenido en donde mediante el uso de algoritmos entrenados junto a técnicas de PLN. Inicialmente, se procedió a categorizar las columnas según su contenido, permitiendo una comprensión de los temas referenciados con la finalidad de dar un entendimiento de las perspectivas u opiniones de los individuos/grupos que las redactan; las cuales son la principal fuente de información para proporcionar una posible idea de lo que se trata en la actualidad.

La investigación buscó simplificar la comprensión de las conversaciones que mantienen ciertas personas sobre temas específicos mediante el resumen de las opiniones de individuos con una amplia base de seguidores a través del análisis de los puntos destacados en sus columnas de opinión. Además, brinda la oportunidad de conocer mejor a un autor, incluyendo sus áreas de interés, plataformas de publicación habituales entre otros detalles relevantes.

Estado de la cuestión

Una contribución significativa al campo del PLN es la introducción de manera innovadora las Redes Neuronales Convolucionales en tres dimensiones (CNNs 3D) en tareas de clasificación de texto. Aprovechando las ventajas de la extracción de características locales y el uso de información jerárquica de modelos preentrenados como BERT, Text3D demuestra un rendimiento superior en la clasificación de sentimientos y categorización de temas en comparación con las metodologías convencionales. Esto resalta la versatilidad de las CNNs 3D en PLN y la importancia de explorar nuevas dimensiones de enfoque en el procesamiento de texto para avanzar en la precisión y eficacia de las aplicaciones de procesamiento del lenguaje natural (7).

Por otra parte, el análisis de sentimientos permite extraer opiniones de datos no estructurados, especialmente para el análisis de datos de Twitter, que refleja las opiniones en

doi: 10.25100/iyc.v25i3.13150

tiempo real sobre una variedad de temas. La introducción del modelo de redes neuronales BERT y el optimizador *Ranger AdaBelief* muestra avances significativos en la precisión de la clasificación de análisis de sentimientos. Estos resultados refuerzan la importancia de la innovación en el PLN y la optimización de algoritmos para obtener una comprensión precisa de las opiniones públicas, lo que puede ser de utilidad en la toma de decisiones empresariales e investigaciones de mercado (8).

Teniendo en cuenta lo anterior, la combinación de PLN y análisis de sentimientos para comprender a los usuarios, en un caso de estudio en Italia, radicó en la categorización de los usuarios en cuatro clases distintas, lo que ofreció una perspectiva completa de las opiniones en función de su perfil. Además, la incorporación de un léxico específico en el idioma nativo (italiano) enriquece la capacidad de discernir el "tono de voz" de cada categoría. Los resultados revelan la capacidad de la metodología para identificar cambios en la actitud de los usuarios en respuesta a eventos importantes tras algún suceso. Por tanto, la estrategia de combinación puede proporcionar información valiosa sobre la percepción pública y ayudar en la toma de decisiones (9-10).

Asimismo, el uso de PLN y reconocimiento de patrones para interpretación de la información textual de manera que se asemeje al procesamiento cognitivo humano representa un gran desafío. Por lo cual, es necesario enriquecer la representación semántica del texto y evaluar y aplicar las particularidades de los idiomas, propendiendo porque la inteligencia artificial se catapulte hacia la comprensión de textos de manera cercana a la mente humana (11-12).

Estudios preliminares, reflejan la necesidad de abordar la brecha de recursos en el procesamiento del lenguaje natural para cada idioma de forma particular, pues esto proporciona un recurso curado para la clasificación de oraciones en el respectivo idioma, promoviendo la diversidad lingüística y modelos de lenguaje inclusivos. Proceso que debe contar con el acompañamiento de hablantes nativos que evalúen las etiquetas de Transformación de Oraciones, mejorando la confiabilidad de los conjuntos de datos, empoderando a investigadores, profesionales y desarrolladores para construir modelos de PLN precisos y sólidos adaptados a un idioma (13-16).

Materiales y métodos

En el trabajo se realizó una adaptación de la metodología CRISP-ML aplicando los algoritmos de PLN reduciendo el esfuerzo humano necesario para tareas de análisis repetitivas. Con base en esta metodología se aplicaron una serie de pasos para dar solución al problema representado en la figura 1.

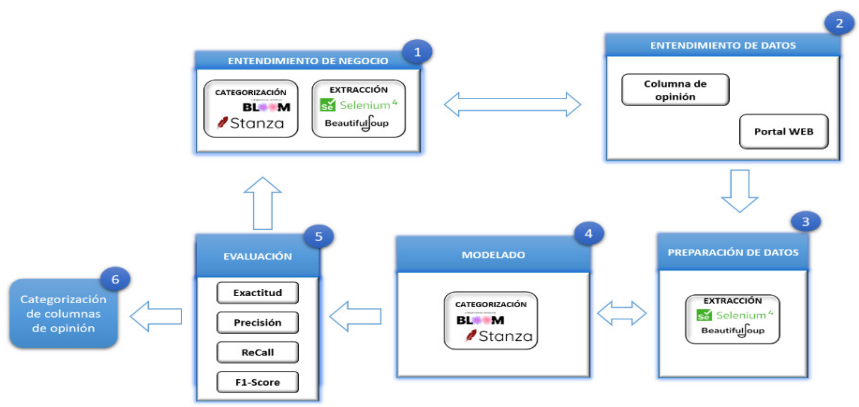


Figura . 1. Metodología CRISP-ML

doi: 10.25100/iyc.v25i3.13150

Recolección de datos: se desarrolló un algoritmo eficiente para recopilar una cantidad considerable de columnas de opinión provenientes de diversas páginas web. Este algoritmo se adapta de manera personalizada a cada portal web, empleando web *scraping* para extraer la información de manera automatizada con el objetivo de proporcionar estos datos al modelo encargado de llevar a cabo la categorización.
Análisis y limpieza: una vez obtenido el contenido de las columnas de opinión, se realizó una validación rigurosa para asegurar que la información extraída esté lo más limpia y precisa posible. En caso de detectar algún tipo de error se aplicaban las correcciones necesarias.
Validación del modelo de categorización: utilizando el contenido extraído de las columnas de opinión, se desarrolló un algoritmo basado en un modelo PLN usando el modelo "Bloom" y otras herramientas, con el fin de categorizar el contenido de cada columna.
Evaluación del desempeño: se emplearon métricas de calidad para evaluar el rendimiento del algoritmo de categorización al determinar los temas de cada columna de opinión. La exactitud, precisión, el recall y la F1-score fueron las métricas aplicadas en una muestra de registros analizados manualmente para evaluar el rendimiento en la clasificación de las columnas en sus respectivas categorías.
Despliegue: al ejecutar el aplicativo se determinaron los temas de cada columna obtenida automáticamente provenientes de los portales previamente almacenados con su respectiva configuración. La automatización de este proceso ahorro tiempo y esfuerzo, y contribuyó a reducir errores humanos en la extracción y clasificación de la información. Además, al ser una ejecución automática, se garantiza una mayor coherencia y consistencia en los resultados obtenidos en comparación con un proceso manual.

Para el proyecto, la población constó de 10327 columnas de opinión que fueron almacenadas en la base de datos y recopiladas a lo largo de casi un año desde 18 portales configurados en páginas web. Además, para la categorización de cada una de estas columnas, se contó con una base de alrededor de 2600 temas, lo que proporcionó un contexto completo para el análisis. Una vez que el algoritmo de categorización fue implementado y ajustado con éxito, se procedió a tomar una muestra de 150 columnas con el objetivo de validar los resultados mediante métricas que permitieron evaluar la calidad de estos.

Asimismo, se propuso la métrica (1) para la exactitud debido a que se debe validar la cantidad de predicciones correctas sobre la cantidad total de predicciones, pero al haber múltiples predicciones posibles se debe aplicar otra metodología, ya que la exactitud de la ecuación (2) simplemente verifica si el resultado de predicciones correctas es igual al total de predicciones.

Como el algoritmo arroja una lista de temas, se considera que lo mejor es que se valide el número de temas correctos entre todos los propuestos:

$$Exactitud\ propuesta = \frac{\sum_1^n t_p}{n} \quad (1)$$

En donde, *n* es el número de columnas analizadas manualmente, *t* es el número de temas verdaderos por columna y *p* el número de temas predichos por el algoritmo de categorización por columna.

La ecuación de exactitud (2) se divide en predicciones correctas sobre todas las predicciones.

$$Exactitud = \frac{predicciones\ correctas}{todas\ las\ predicciones} \quad (2)$$

El resultado de la métrica de precisión es la división entre los verdaderos positivos y la suma entre los verdaderos positivos y los falsos positivos (3) lo que permite conocer cuántos de los temas predichos realmente se debieron incluir.

$$Precisión = \frac{VP}{VP+FP} \quad (3)$$

doi: 10.25100/iyc.v25i3.13150

La métrica ReCall es la proporción de verdaderos positivos (VP) entre todos los casos positivos reales (falsos negativos y verdaderos positivos) (4).

$$Recall = \frac{VP}{VP+FN} \quad (4)$$

Para calcular la métrica F1-score, se utilizan los valores de las métricas recall y precisión. Esto proporciona una forma de evaluar el equilibrio de los resultados generados por el algoritmo de categorización (5).

$$F1 - score = 2 * \frac{Precisión*recall}{Precisión+recall} \quad (5)$$

Resultados

La figura 2 presenta la arquitectura propuesta donde se identifican las conexiones entre los algoritmos de extracción y recuperación de información y el algoritmo de categorización automática de columnas de opinión desde páginas web: inicialmente se utiliza una configuración manual específica para cada portal web donde se indican las etiquetas HTML que deben ser analizadas. El proceso inicia con la creación de configuraciones individuales para cada portal del cual se desee extraer y recuperar la información.

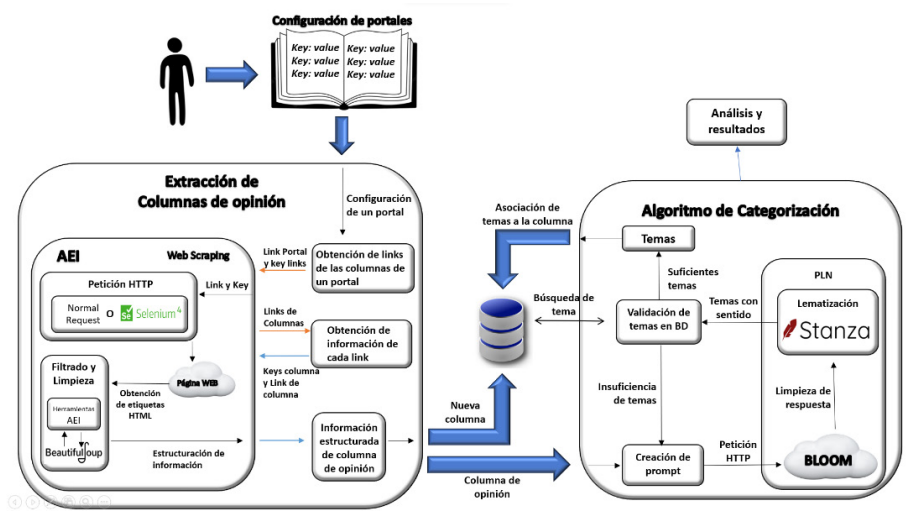


Figura. 2. Arquitectura del modelo propuesto.

Una vez se cuenta con la configuración específica, se procede a aplicar la técnica de web scraping a través del algoritmo AEI. El algoritmo realiza una petición normal o utiliza Selenium para simular la interacción de un usuario real en una página web, según sea el caso, para acceder a cada portal web. Luego, se realiza una nueva petición para cada enlace de la lista, ubicando automáticamente la información y filtrando hasta obtener una estructura adecuada de la columna de opinión. La información extraída se almacena en la base de datos.

Finalmente, se valida la existencia de cada tema en la base de datos y en caso de encontrar coincidencias, se procede a almacenar la relación entre el tema y la columna de opinión en la base de datos.

Extracción de columnas de opinión

En primer lugar, se extrae la información de páginas web mediante herramientas de *web scraping* y la creación de un algoritmo que logra analizar, limpiar y almacenar la información obtenida de

doi: 10.25100/iyc.v25i3.13150

las etiquetas HTML que se quiere obtener para luego, con ayuda de técnicas de PLN categorizar la información obtenida. Para obtener la información requerida de las columnas de opinión se creó un algoritmo de detección y limpieza de la información apoyándose de *BeautifulSoup*, una librería que permite determinar en dónde se abren, se cierran y algunas características de las etiquetas HTML como sus atributos o identificadores, simulando al lenguaje XPATH, pero con algunos filtros adicionales como expresiones regulares, validación de tamaño cadenas de caracteres, verificación de etiquetas hijo, entre otras que logran identificar patrones en todos los links dentro de una misma página web.

Para extraer la información de una página web se debe configurar un portal web para ese se diseño un tipo de plantilla que le permite al algoritmo de extracción conocer la ubicación de la información y cómo debe limpiar o filtrar la información que obtendrá. La figura 3 ilustra la estructura de una configuración para un portal web. Las características se refieren a los atributos que singularizan a un portal web; estos se almacenarán en la base de datos. Además, se incluyen opciones que guiarán al algoritmo en la ubicación y extracción de la información, así como en su posterior proceso de limpieza.

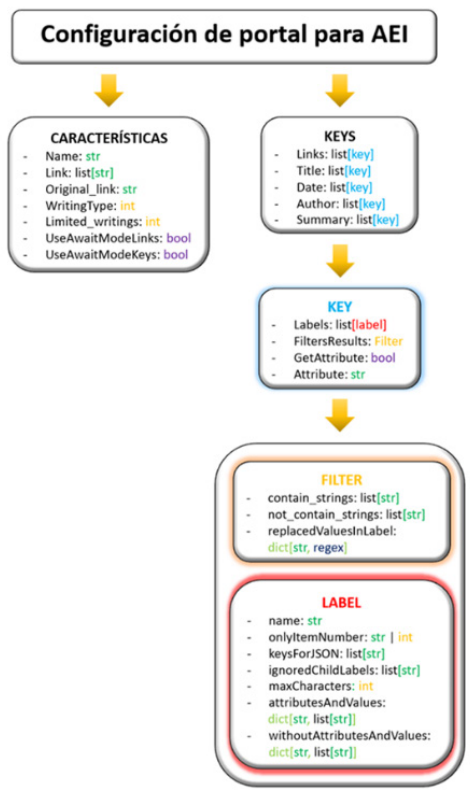


Fig. 3. Configuración de un portal.

Luego de extractar la información se debe realizar un control de calidad manual para verificar la limpieza, coherencia de la información y se establece que la información no cumple con lo que se esperaba. La información que se extrajo y recuperó consiste en el nombre del autor, fecha de publicación, título, contenido, fecha de extracción y link de la columna de opinión. En la figura 4 se evidencia el proceso que se siguió para extraer las columnas de opinión y validar las configuraciones respectivas para cada portal.

doi: 10.25100/iyc.v25i3.13150

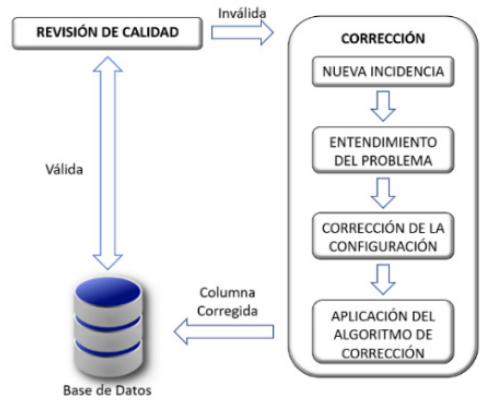


Figura. 4. Proceso de limpieza y corrección de columnas de opinión

Algoritmo de categorización

Se elaboró un algoritmo que recibe información extraída y rectificada del paso anterior y la convierte en una lista de temas, utilizando el API de *Hugging Face*, además del manejo de *prompts* que dependen del tamaño y los símbolos de puntuación de su contenido y Stanza que aplica la lematización. La API de *Hugging Face* limita el número de peticiones que recibe y la respuesta dependiendo del tamaño del *prompt* o de la cantidad de tokens que se le envíe, por eso se diseñó un algoritmo orientado a solucionar los conflictos identificados en donde para cada petición se añade un intervalo de tiempo que varía dependiendo de la respuesta que se obtiene.

Se debe limpiar mediante el uso de expresiones regulares y algoritmo de cadenas de caracteres hasta dejar únicamente una lista de palabras, con el fin de evitar resultados incoherentes. Finalmente, se aplica la técnica de lematización mediante la librería Stanza para conservar las palabras que verdaderamente son de interés y que tienen sentido, para luego validar si existen en la base de datos. Si la respuesta es afirmativa se crea un nuevo registro que asocia y almacena la columna con el tema obtenido como se representa en la figura 5.

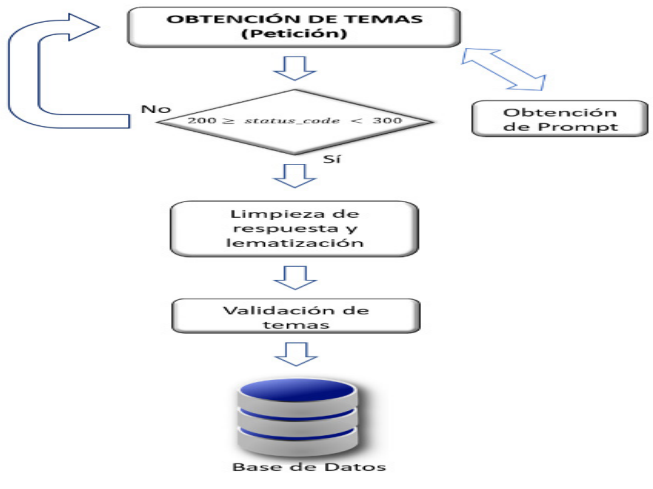


Figura. 5. Proceso de Categorización

Para validar los resultados del algoritmo de categorización y asegurar la precisión en la predicción de los temas relacionados con cada columna de opinión, se aplicaron métricas de evaluación. Estas métricas permitieron medir y analizar la efectividad del algoritmo, asegurando así la confiabilidad de los resultados obtenidos. Se recolectaron 10484 registros

doi: 10.25100/iyc.v25i3.13150

de publicaciones y 10424 registros de columnas de opinión, que incluyeron información relevante como el título, contenido, fecha de publicación, autor, entre otros. Las columnas extraídas provienen de 18 portales con su respectiva configuración y para categorizar las columnas se creó una lista con 2602 temas que permiten dar contexto a las columnas. En la figura 6 el algoritmo extrae 5 columnas de opinión del portal web elheraldo.co con ayuda de Selenium, simulando la interacción de una persona con el sitio. Se muestran los temas obtenidos del análisis de las cuatro primeras columnas, junto con el título, el link, el nombre del autor y la fecha de publicación de cada columna de opinión obtenida.

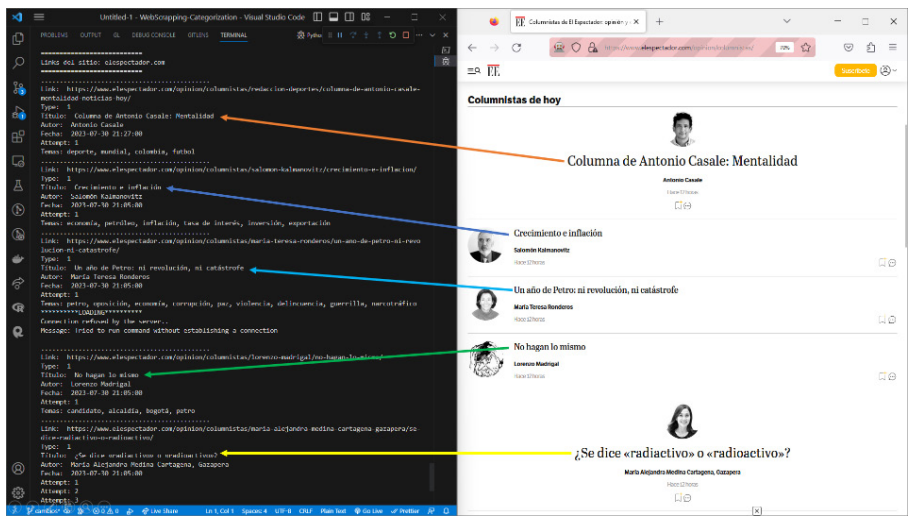


Fig. 6. Extracción con Selenium

Se obtuvieron algunos ejemplos de cómo utilizar la información recolectada entre las columnas y los temas que se identificaron por cada una. A continuación, se muestran algunas gráficas que representan la información resumida en algunos casos específicos: en la figura 7a “Autores más activos en Junio” se muestran los 5 autores que más columnas redactaron durante el mes de junio del 2023 de 18 portales web. Los autores que más columnas redactaron en el mes de junio fueron: Gonzalo Gallo con un total de 29, seguido de Rafael Nieto Loaiza con 16, Luis Alonso Colmenares con 14, José Lafaurie con 13 y Mario Fernando Prado con 8. En la figura 7b “Temas más populares el 6 de diciembre” se identificaron cuáles fueron los temas que más se abordaron en las columnas de opinión durante el 6 de diciembre de 2022 de todos los portales web que tenían una configuración predefinida hasta ese momento. El tema más popular en ese día fue economía.

En la figura 7c “Temas más populares de junio” se hizo un análisis para el mes de junio del 2023 en donde se obtuvieron los 10 temas más tratados durante este mes sin importar el autor o el portal. El tema más popular de junio fue la corrupción en donde la mayoría de los autores que redactaron columnas en este lapso hablaron sobre este tema.

En la Figura 7d se observa como los temas más relevantes fueron corrupción y economía para los dos meses, pero con una pequeña disminución en la cantidad de columnas. Lo que indicaría que en estos meses pudo haber escándalos de corrupción o discusiones que incluían al gobierno y que afectaban al Gobierno y, por ende, al presidente actual. El tema paz es un tema en común para los dos meses y se evidencia que en junio se redactó más sobre ese tema tal vez por los acuerdos de cese al fue con los grupos armados del país.

doi: 10.25100/iyc.v25i3.13150

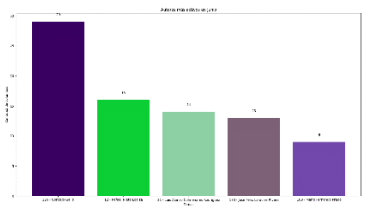


Fig. 7a. Autores más activos en Junio.

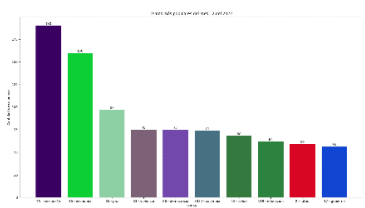


Fig. 7b. Temas más populares en diciembre 2022

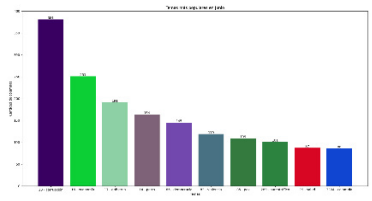


Fig. 7c. Temas más populares en junio 2023

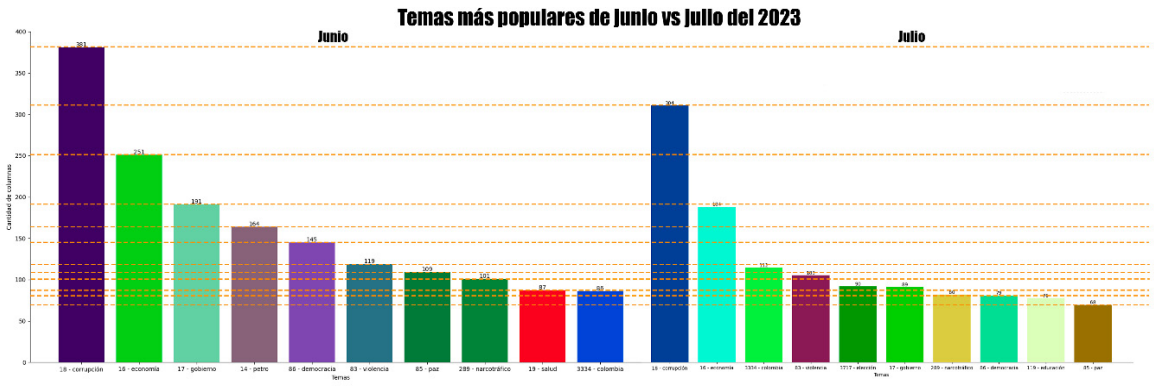


Fig. 7d. Temas más populares de junio vs julio del 2023

Validación

Se evaluó la relación que existe entre los autores, las publicaciones, los escritos y los portales de un dataset de 9263 registros de la información de las columnas de opinión. Se evidenció que en el escrito y la publicación es donde mayor correlación se presenta debido a que a medida que se creaban registros de los escritos, al mismo tiempo se creaba uno de publicación. Y no es perfecta debido a que el software detecta cuando un escrito ya existe y está presente en otro portal, por lo tanto, se crea únicamente un registro de la publicación.

La figura 8 muestra la relación entre el portal y el autor donde se evidencia una mayor correlación comparándola con el resto de resultados y puede deberse a que cada portal tiene sus propios escritores o que estos escriben únicamente para un portal.

Se toma como umbral 0.4 como criterio para tomar una correlación como positiva debido a que se plantea conocer con qué frecuencia un autor escribe para un portal. Del resultado obtenido se puede concluir que las correlaciones entre autores y portales, y publicación y escrito, son los únicos verdaderos positivos debido a que normalmente un autor tiene una mayor inclinación por escribir para un mismo portal y una columna de opinión en la mayoría de casos es publicada en un solo sitio web.

doi: 10.25100/iyc.v25i3.13150

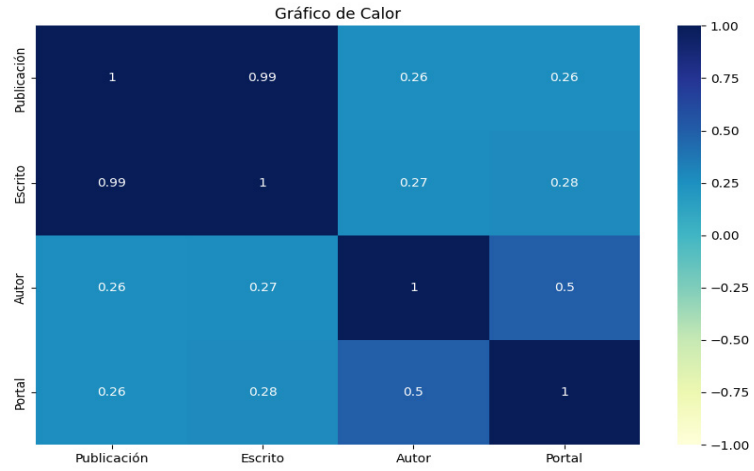


Fig. 8. Matriz de correlación de los registros almacenados

Aplicación de métricas

Para evaluar los resultados del algoritmo de categorización se aplicaron métricas para validar la exactitud del algoritmo cuando se utiliza el modelo para determinar los temas de una columna de opinión tomando como muestra 150 columnas de opinión previamente extraídas, almacenadas y analizadas manualmente en donde se determinaron cuáles verdaderamente son los temas más relevantes de una columna dando como resultado 0.56 para la métrica de exactitud de la gráfica (2) y 0.885 para la métrica de exactitud propuesta (1). Al ser una lista de respuestas proporcionadas por el algoritmo de categorización se debe validar cuántos de los valores de esa lista son correctos y verdaderamente tienen que ver con el resultado total.

En la tabla 1 se presentan los resultados obtenidos al aplicar las métricas de calidad. En la métrica de precisión el algoritmo determinó cuántos de los temas predichos están verdaderamente relacionados con la columna. En otras palabras, se evaluaron los falsos y los verdaderos positivos en cada columna de opinión dando como resultado 0.931. El resultado obtenido indica que de las 150 columnas de opinión el 92% de los temas que el algoritmo arrojó verdaderamente están relacionados a la columna.

En cuanto a la métrica recall, se evaluó la capacidad del algoritmo para no pasar por alto temas que estuvieran verdaderamente relacionados con el contenido de la columna. Se analizaron los falsos negativos en contraste con los verdaderos positivos dando como resultado 0.975. En las 150 columna de opinión que se analizaron, se determinó que el algoritmo omitió un poco más del 2% de los temas que podrían haber sido incluidos en el análisis.

Finalmente, con la métrica f1-*score* que evalúa el equilibrio entre las métricas de precisión y recall. Se demuestra que el algoritmo ha arrojado resultados sobresalientes en la clasificación de las columnas en donde se refleja que ha logrado clasificar de manera precisa y acertada la gran mayoría de las columnas, al mismo tiempo ha minimizado los falsos positivos y los falsos negativos con un resultado de 0.948.

doi: 10.25100/iyc.v25i3.13150

Tabla 1. Resultados métricas

Métrica	Resultado
Exactitud	0,56
Exactitud estimada	0,885
Precisión	0,931
ReCall	0,975
F1-Score	0,948

Conclusiones

El análisis de la métrica de exactitud propuesta permite concluir que el algoritmo logra predecir los temas de cada columna con un alto grado de acierto. Además, de los temas de las 150 columnas de opinión analizadas, el algoritmo logró identificar o caracterizar correctamente el 88% de los temas. Esto proporciona un resultado mucho más acorde con la calidad de los resultados obtenidos y demuestra que el algoritmo arroja muy buenos resultados. El algoritmo AEI implementado tiene la versatilidad de ser utilizado en la extracción de información de diversos sitios web. Para emplearlo, simplemente es necesario proporcionarle la ubicación de la etiqueta y especificar las características particulares que se desean extraer.

A diferencia de la mayoría de los métodos de caracterización de textos que requieren una gran cantidad de ejemplos para incluir un nuevo tema o un riguroso entrenamiento, en este caso solo es necesario crear un registro del nuevo tema en la base de datos, y el algoritmo automáticamente lo detectará y validará mediante el modelo *Bloom*. Esto representa un avance significativo que simplifica y agiliza el proceso de ampliar el alcance del análisis y caracterizacion de las columnas de opinión lo que repercute en menores costos.

El modelo *Bloom* resulta altamente útil, siempre y cuando el patrón indexado en el prompt sea preciso y específico en relación a lo que se desea obtener. La efectividad del modelo depende en gran medida de la adecuada elección del patrón, ya que este guiará la búsqueda y determinará la calidad de los resultados obtenidos. Es esencial comprender la importancia de definir un patrón adecuado, que abarque de manera precisa los elementos clave de la información que se busca extraer. Mediante el uso de técnicas avanzadas de PLN, es posible extraer información relevante, identificar patrones y categorizar el contenido de manera altamente eficiente, como se ha evidenciado a través de los resultados obtenidos en las métricas. Estos resultados demuestran que se ha logrado analizar una gran cantidad de información en un corto período de tiempo, obteniendo resultados coherentes y de alta calidad.

Extraer datos automáticamente de Internet puede resultar un desafío debido a la gran cantidad de información no estructurada que se encuentra disponible. Sin embargo, mediante el desarrollo de un algoritmo capaz de recopilar esta información de manera organizada, permitiendo la aplicación de filtros y la posibilidad de realizar limpieza durante o al final del proceso de extracción, se ha logrado crear una herramienta sumamente útil que puede ser aplicada en diversos campos.

La capacidad de proporcionar la información relevante de una columna de opinión brinda la posibilidad de conocer la opinión de la sociedad sobre una empresa o persona en particular, identificar los temas de actualidad o comprender en qué temas se especializa un determinado autor, entre muchas otras posibilidades.

doi: 10.25100/iyc.v25i3.13150

Referencias

1. Moreno A. [Internet] Procesamiento del lenguaje natural ¿qué es?, 2023. Disponible en: https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/
2. Kaur G, Sharma A. A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. Journal of Big Data. 2023; 10(1):10-18. https://doi.org/10.1186/s40537-022-00680-6
3. Haque R, Islam N, Tasneem M, Das AK. Multi-class sentiment classification on Bengali social media comments using machine learning. International Journal of Cognitive Computing in Engineering. 2023; 4: 21-35. https://doi.org/10.1016/j.ijcce.2023.01.001
4. Martínez N, Téllez J, Barrero J, Chaves L. Automatic method for the prediction of the commercial appraisal of a property in Bogota city. 7th Congreso Internacional de Innovación y Tendencias En Ingeniería. 2021. https://doi.org/10.1109/CONIITI53815.2021.9619685
5. Báez P, Arancibia AP, Chaparro MI, Bucarey T, Núñez F, Dunstan J. Natural language processing for clinical text in Spanish: The case of waiting lists in Chile. Revista Médica Clínica Las Condes. 2022; 33(6): 576-582. https://doi.org/10.1016/j.rmclc.2022.10.002
6. Garrido-Muñoz I, Montejo-Ráez A, Martínez-Santiago F. Exploring gender bias in Spanish deep learning models. CEUR Workshop Proceedings. 2022; 3224: 44-47
7. Wang J, Li J, Zhang Y. Text3D: 3D Convolutional Neural Networks for Text Classification. Electronics (Switzerland). 2023; 12(14):e87. https://doi.org/10.3390/electronics12143087
8. Gouthami S, Hegde NP. An improved sentiment classification model using BERT classification with ranger Adabelief Optimizer. Journal of Theoretical and Applied Information Technology. 2023; 101(12): 5102-5113.
9. Catelli R, Pelosi S, Comito C, Pizzuti C, Esposito M. Lexicon-based sentiment analysis to detect opinions and attitude towards COVID-19 vaccines on Twitter in Italy. Computers in Biology and Medicine, 2023; 158:e106876. https://doi.org/10.1016/j.compbiomed.2023.106876
10. Yang Z, Zhang L, Wang X, Mai Y. ESG Text Classification: An Application of the Prompt-Based Learning Approach. Journal of Financial Data Science. 2023; 5(1): 47-57. https://doi.org/10.3905/jfds.2022.1.115
11. De Santis E, Rizzi A. Prototype Theory Meets Word Embedding: A Novel Approach for Text Categorization via Granular Computing. Cognitive Computation. 2023; 15(3): 976-997. https://doi.org/10.1007/s12559-023-10132-9
12. Siddiqui T, Amer, A. A comprehensive review on text classification and text mining techniques using spam dataset detection. Mathematics and Computer Science. 2024; 2: 1-18. https://doi.org/10.1002/9781119896715.ch1
13. Das RK, Islam M, Khushbu SA. BTSD: A curated transformation of sentence dataset for text classification in Bangla language. Data in Brief. 2023; 50:e109445. https://doi.org/10.1016/j.dib.2023.109445
14. Bi H, Li B, Qiu Y, Change M. EnvText: A Chinese text mining tool for environmental domain with advanced BERT model. Software Impacts. 2023; 17:e100559. https://doi.org/10.1016/j.simpa.2023.100559
15. Palai P, Agrawal K, Mishra DP, Salkuti SR. Text grouping: a comprehensive guide. IAES International Journal of Artificial Intelligence. 2023; 12(3): 1476-1483. https://doi.org/10.11591/ijai.v12.i3.pp1476-1483

doi: 10.25100/iyc.v25i3.13150

16. Fonseca CA, de Souza Netto RS, Bodolay AN, Carvalho Guelpeli MV. AnoTex: Structured data filtering routine of the scientific article genre as contribution to PLN. Texto Livre. 2018; 11(3): 40-64. https://doi.org/10.17851/1983-3652.11.3.40-64

doi: 10.25100/iyc.v25i3.13150