Revista Latinoamericana de Metodología de la Investigación Social.
Nº29. Año 15. Abril - Septiembre 2025. Argentina. ISSN 1853-6190. Pp. 51-67.

Uso de Procesamiento de Lenguaje Natural para procesar respuestas abiertas de una encuesta de Opinión Pública

Use of Natural Language Processing to process open responses of a Public Opinion survey

Esteban Martínez Porras, Adrián Ramírez Fernández, Laura Solís Bastos y José André Díaz-González

Resumen

Este artículo revisa la utilidad de utilizar Procesamiento de Lenguaje Natural (PLN) para el análisis de las respuestas abiertas brindas a una encuesta de opinión pública. Se utilizan datos de una encuesta sobre percepción de la población costarricense respecto a diversos grupos migrantes y, a partir de ellos, se prueban diversos algoritmos, con el propósito de identificar cuál de ellos realiza una mejor clasificación de los datos. Se concluye que el algoritmo *random forest* es el que realiza una mejor clasificación automática de las respuestas, ayudando a disminuir la ambigüedad de la clasificación. El trabajo busca resaltar la utilidad que tiene el PLN para las investigaciones en Ciencias Sociales y, en especial, para el análisis de datos de preguntas abiertas aplicadas en estudios de opinión pública, ya que facilita la clasificación y análisis de gran cantidad de información no estructurada.

Palabras clave: Procesamiento de Lenguaje Natural; Minería de Texto; Opinión Pública; métodos de encuestas; *Machine Learning.*

Abstract

This article reviews the usefulness of using Natural Language Processing (NLP) for the analysis of open responses given to a public opinion survey. Data from a survey about the perception of the Costa Rican population regarding various migrant groups are used to test algorithms to identify which performs a better classification of the data. It is concluded that the random forest algorithm performs a better automatic classification of the answers, helping to reduce the ambiguity of the classification. The work seeks to highlight the usefulness of the PNL for research in Social Sciences and, especially, for the analysis of data from open questions applied in studies of public opinion, since it facilitates the classification and analysis of a large amount of unstructured information.

Keywords: Natural Language Processing; Text Mining; Public Opinion; survey methods, *Machine. Learning.*

Introducción

Este artículo explora la pertinencia de utilizar técnicas de Procesamiento de Lenguaje Natural (PLN) para analizar las respuestas a preguntas abiertas obtenidas mediante la aplicación de una encuesta de opinión pública. Dado esto, el trabajo aquí presentado es en cierta forma exploratorio, ya que busca determinar la pertinencia de aplicar un método de análisis de uso común en el *Big Data* en una encuesta de opinión pública, con lo que se pretende generar un aporte metodológico que facilite el procesamiento y análisis de información recolectada en preguntas abiertas.

Para ello, el presente artículo se divide en seis secciones: en la primera se realiza una revisión de literatura académica sobre el uso de algoritmos en las ciencias sociales; en la segunda sección se realiza una presentación de la encuesta de la cual se toman los datos para realizar el presente análisis; en la tercera sección se explican los alcances y limitaciones del procesamiento de lenguaje natural para el análisis de textos; en la cuarta sección se expone el procedimiento metodológico seguido para el análisis de los datos; en la quinta sección se indican los resultados del análisis y, en la última sección, se encuentran las conclusiones.

Revisión de literatura académica sobre Algoritmos y Ciencias Sociales

Desde las Ciencias Sociales se han realizado diversos acercamientos hacia el uso de los algoritmos, sin embargo, prioritariamente la literatura académica se ha concentrado en los algoritmos y sus usos como fenómeno social, más que en identificar sus aportes en el desarrollo de procesos de investigación y análisis.

En el área de la comunicación social es una en la que se han desarrollado más investigaciones relacionadas al impacto de los algoritmos en el acceso y consumo de la información, especialmente en las redes sociales. Estos trabajos han mostrado que el uso de algoritmos que ayudan a promover información a partir de las reacciones emotivas de las personas usuarias tiende a facilitar la polarización de las opiniones (Fondevila Gascón, 2017). Otras investigaciones establecen que dada la implementación de algoritmos para seleccionar la información que difunden entre los usuarios, las redes sociales se han alejado de ser las plataformas de difusión o distribución de información que pretendían ser, y más bien han asumido un rol de editor de información, por las que decisiones que adoptan en el diseño de sus algoritmos deben ser discutidas a partir de la responsabilidad ética de dicho papel (Cetina Presuel & Martínez Sierra, 2019).

Otro grupo de trabajos se ha concentrado en estudiar la utilidad de los algoritmos para la toma de decisiones. Así, se ha determinado que es posible diseñar algoritmos que faciliten tomar decisiones en temas sociales complejos, o bien, que simplemente ayudan a simplificar las tareas cotidianas, lo cual permita a las personas concentrarse en tareas y problemas prioritarios (Zoran, 2016). También se ha investigado la utilidad de los algoritmos en los procesos de toma de decisiones públicas, así como en el diseño de políticas públicas; estos trabajos han mostrado que, si bien el uso de algoritmos puede ser útil para generar modelos de toma de decisiones ante situaciones complejas, al mismo tiempo reconocen que un diseño inadecuado de estos puede generar que ciertos grupos sean excluidos o no beneficiados de las decisiones adoptadas (Beer, 2017; Maxwell & Tomlinson, 2020).

Un tercer grupo de trabajos se han concentrado en identificar el valor de los algoritmos para el desarrollo de simulaciones, y han recalcado cómo estos han contribuido en el Modelamiento y Simulación Basado en Agentes (ABMS) para la comprensión y explicación de los fenómenos sociales, su emergencia, evolución y adaptación más que la predicción típica de los demás tipos de simulación (Díaz & Domínguez, 2013). También se ha destacado la utilidad que tiene el *Big Data* y el uso de algoritmos en los estudios de opinión pública, destacándose su aporte tanto en el proceso de recolección de información en redes o plataformas digitales, hasta sus aportes en los análisis de información y en la identificación de patrones de opinión (Cabrera-Álvarez, 2022; González, 2019; Gualda, 2022; Mamaqi *et al*., 2021; Porter *et al*., 2020; Tu *et al*., 2021).

Así, el uso de algoritmos en los análisis sociales y, especialmente, en los estudios de opinión pública dista de ser campo nuevo o emergente; sin embargo, posiblemente no puede considerarse aún que sea una herramienta de análisis consolidado en las disciplinas sociales. Desde las ciencias de la información se ha demostrado la utilidad de los algoritmos para clasificar y procesar información, y esto ha sido aprovechado por las ciencias sociales para generar modelos para la toma de decisiones y procesar datos de alta complejidad; no obstante, aún es necesario profundizar en los alcances y limitaciones que tienen los algoritmos como herramienta de investigación y, principalmente, promocionar y fomentar sus uso entre las personas profesionales del área de ciencias sociales.

Debido a lo anterior, este trabajo busca realizar un aporte respecto al uso de algoritmos de aprendizaje automático (*Machine Learnig*) para procesar y analizar los datos recopilados en las preguntas abiertas aplicadas en una encuesta de opinión pública. Con esto, se espera mostrar como este tipo de algoritmo puede ser tanto una herramienta metodológica que facilita el trabajo de procesamiento de datos, como una herramienta analítica que facilite a las personas investigadoras la labor de comprender e interpretar la información recopilada.

Revisión de literatura académica referente a estudios de percepción sobre la población migrante en Costa Rica

El Instituto de Estudios Sociales en Población (IDESPO), en sus más de cuatro décadas de existencia ha abordado el fenómeno migratorio en el país, como un tema de relevancia para conocer el acontecer nacional y las percepciones de la población al respecto.

Cabe destacar que la labor que realiza este instituto en el marco del análisis de las percepciones de la población costarricense tiene un amplio reconocimiento a lo largo de su trayectoria, este tipo de estudios por lo regular se han desarrollado en la mayoría de los casos a través de encuestas telefónicas dirigidas a la población nacional y residente en el país.

A continuación, se detallarán algunos de los principales resultados de estudios sobre percepción acerca de la población migrante en Costa Rica, los que ejemplifican el desarrollo de material académico en esta línea.

- *Agosto, 2000.* La población costarricense de la gran área metropolitana frente a percepción hacia la población migrante, la política y los impuestos. Esta encuesta telefónica fue realizada a 400 personas, en este caso se aborda la percepción hacia la población migrante y en particular sobre la población nicaragüense.
- *Agosto, 2005.* Percepciones de la población costarricense sobre la inmigración de nicaragüenses. En esta encuesta telefónica el tamaño de la muestra fue de 600 y objetivo fue aportar a la opinión pública información referida a las percepciones de los costarricenses acerca de los inmigrantes.
- *Junio, 2006.* Se publica el informe de encuesta Identidades nacionales, integración y ciudadanía: percepciones hacia la inmigración. Presenta resultados de una encuesta nacional telefónica a hogares, la cual obtuvo una muestra de 600 personas mayores de edad. En esta oportunidad se abordan temas como los efectos de la migración sobre las instituciones sociales, políticas y económicas de nuestro país. Se plantea en específico como la inmigración es percibida como un "problema" de orden público, el cual necesita ser controlado mediante la restricción del acceso de inmigrantes a nuestro país.
- *Octubre, 2008.* Se publica el informe de encuesta Percepciones y actitudes de la población costarricense hacia la inmigración nicaragüense y la emigración de costarricenses al exterior. Corresponde a datos de una encuesta nacional telefónica a hogares, la cual obtuvo una muestra de 600 personas mayores de edad. En esta oportunidad se abordan temas como las percepciones acerca de Costa Rica como país receptor de población migrante (nicaragüenses, colombianos, indígenas), así como la construcción de esa visión de las otredades.

- *Agosto, 2014.* Se publica el informe de encuesta Construcción de opiniones públicas sobre la migración en Costa Rica. En este estudio, que empleó como medio de contacto la telefonía fija, se encuestaron 1000 personas y se valoraron las percepciones respecto a nacionalidades específicas de la población migrante residente en el país o en tránsito.
- *Abril, 2016.* Se publica el informe de encuesta Percepciones acerca de las relaciones entre Costa Rica y Nicaragua. Se realizó vía telefónica, a una muestra de 800 personas, mayores de edad y costarricenses. Tuvo por objetivo determinar las percepciones y opiniones de la población costarricense respecto a las relaciones entre Costa Rica y Nicaragua, a partir de su opinión sobre coyunturas o acontecimientos acaecidos en los meses anteriores en los que intervienen o se ven afectados ambos países.
- *Setiembre, 2019.* El informe percepciones de la población nacional sobre las migraciones, convivencia e integración en Costa Rica, muestra los resultados de una encuesta telefónica que empleó como método de contacto líneas celulares activas en el país, dirigida a personas costarricenses por nacimiento, de 18 años o más, la cual tuvo un alcance de 1002 personas encuestadas. Esta encuesta tuvo como objetivo determinar las percepciones de la población nacional acerca de las dinámicas de integración y convivencia de la población inmigrante en Costa Rica.

Es así como resulta interesante comprender de qué forma los estudios de percepción mediante encuestas son un insumo de gran relevancia para la investigación social, lo cual ha sido visible a través del tiempo para abordar temáticas específicas tales como las migraciones en Costa Rica.

Los estudios de percepción acerca de las migraciones en Costa Rica han permitido evidenciar como se construyen las opiniones en torno a la población migrante, en este sentido se ha observado que hay múltiples factores que pueden influir en la manera en la que forman estas percepciones, tales como la experiencia personal, la información publicada en medios de comunicación, los discursos políticos y políticas migratorias, estereotipos culturales, entre otros.

En el caso de la experiencia personal se observa que las interacciones individuales con personas migrantes podrían llegar a influir en la forma en que se percibe a esta población. Si alguien ha tenido experiencias positivas con migrantes, es más probable que tenga una percepción favorable hacia ellos. Por el contrario, las experiencias negativas pueden generar percepciones negativas.

Por otra parte, los medios de comunicación desempeñan un papel importante en la construcción de las percepciones sobre la migración. La forma en que se presentan las noticias y los reportajes sobre migrantes podría llegar a influir en cómo se percibe a esta población. Si los medios enfatizan los aspectos negativos o sensacionalistas exaltando el fenómeno migratorio o la nacionalidad de la persona migrante, esto puede contribuir a la formación de percepciones negativas.

Además, en el caso de discursos políticos y las políticas migratorias pueden tener un impacto significativo en la formación de percepciones sobre la población migrante. Dependiendo de cómo los líderes políticos aborden el tema de la migración, se podría generar percepciones positivas o negativas. Si se utiliza un lenguaje que estigmatiza a los migrantes o se promueven políticas restrictivas, esto puede influir en la forma en que se percibe a esta población.

Finalmente, los estereotipos culturales también podrían influir en la forma en que se percibe a los migrantes. Los estereotipos negativos o la falta de conocimiento sobre las culturas de origen de los migrantes pueden llevar a construir percepciones negativas. Por otro lado, si existe una mayor familiaridad y comprensión de las diferentes culturas, esto podría contribuir a la integración y la convivencia.

Es importante tener en cuenta que las percepciones sobre la población migrante pueden ser diversas. Además, estas percepciones pueden llegar a cambiar con el tiempo a medida que se

desarrollan condiciones sociopolíticas que favorezcan nuevas visiones en referencia a este fenómeno o la población migrante.

Las percepciones de la población costarricense sobre la población migrante

En el año 2012, desde el Programa Migraciones, Cambio Social e Identidades adscrito al IDESPO-UNA, se llevó a cabo la encuesta "Construcción de opiniones públicas sobre la migración en Costa Rica", los datos fueron recolectados en el mes de mayo. Esta encuesta fue realizada a través de telefonía fija a 1000 personas costarricenses, mayores de 18 años y residentes del hogar contactado, contó con un error de muestreo del 3,1% y un nivel de confianza del 95%. Estos datos han sido retomados en la actualidad con el propósito de analizarlos desde una lectura distinta, empleando algoritmos de aprendizaje automático para clasificar y procesar información, de forma tal que pueda compararse los resultados obtenidos en el año 2012, y así valorar como el uso de un algoritmo puede llegar a facilitar su procesamiento.

En el marco de esta encuesta se aplicó un módulo de preguntas referentes a la percepción sobre distintas poblaciones migrantes en Costa Rica; primero se consultó: A continuación, me gustaría que me indique qué piensa de algunos grupos y de forma posterior se preguntaba ¿Por qué creé eso? Dentro de los grupos en consulta se encontraban: haitianos, indígenas panameños, colombianos, nicaragüenses, dominicanos, españoles, africanos, estadounidenses y chinos. El objetivo de esta pregunta se asociaba a valorar las posibles caracterizaciones acerca de la población migrante en Costa Rica, a partir de los imaginarios sociales que se construyen vinculados a las nacionalidades de las personas migrantes.

La migración es un fenómeno de gran relevancia para Costa Rica debido a la complejidad de su contexto migratorio. El país se caracteriza principalmente por ser receptor de migrantes, con la comunidad nicaragüense como el grupo mayoritario dentro de su territorio. En años recientes, además, ha adquirido un papel significativo como país de tránsito para personas migrantes que se desplazan, en su mayoría, hacia los Estados Unidos. Finalmente, aunque en menor medida, Costa Rica también es un país de origen de migrantes, siendo precisamente los Estados Unidos el principal destino elegido por los costarricenses que deciden emigrar.

Comprender las percepciones sobre las migraciones es un reto de complejidades, esto debido a que el fenómeno se encuentra presente dentro de la cotidianidad de la población, enmarcado en las condiciones en cuales se da esta movilidad poblacional, tanto regulares como irregulares. Además, esta construcción de las percepciones puede tender en algunos casos a estar marcada por mitos y estereotipos en torno a estos grupos poblacionales, tal como lo manifiesta Delgado Montaldo (2008) en referencia a Goffman:

> De acuerdo con el estigma se manifiesta como una actitud negativa, discriminatoria, que se dirige siempre contra un individuo o un grupo al que se le considera inferior. En este sentido, aunque en ocasiones la víctima pueda ser una persona aislada, la principal razón por la cual se le estigmatiza, se le discrimina o se le evalúa negativamente, es porque pertenece a un determinado grupo (los 'nicas' o nicaragüenses, los 'paisas' o colombianos, los 'nochis' o chinos, en el caso de poblaciones inmigrantes, o bien, los afrocaribeños y los indígenas, quienes también son estigmatizados) (Delgado Montaldo, 2008: 86-87).

Dentro de las características que se mencionan de forma más frecuente, es común identificar algunas asociadas a elementos positivos, otras a elementos negativos y algunas poseen carácter ambivalente. También se observa que, según la nacionalidad, algunas surgen únicamente para grupos específicos, mientras que otras características se asocian a condiciones que fueron relacionadas al país de origen, incluso a la percepción sobre la construcción del imaginario de la nacionalidad, lo que plasma una diversidad de respuestas sobre la población migrante y muestra la complejidad que se requiere para el procesamiento de este tipo de datos (ver Figura 1).

Figura 1. Nubes de palabras asociadas a las caracterizaciones de población migrante en Costa Rica. 2012.

Fuente: Construcción propia, a partir de datos de IDESPO-UNA (2012).

De acuerdo con lo anterior, destaca esa visión diferenciadora que se refleja en el marco de las percepciones mencionadas por la población encuestada, y que trazan líneas para quién es denominado como migrante según su país de procedencia, tal como lo destaca Sandoval García (2004):

> "Inmigración" se ha convertido en un concepto de "sentido común" que requiere ser discutido críticamente. Por ejemplo, europeos o norteamericanos que invierten en actividades turísticas en Costa Rica podrían ser considerados "inmigrantes", pues han abandonado su país y residen en una nueva nación. Sin embargo, se les conoce como "inversionistas", "pensionados" o "turistas". Así, "inmigrante" es un término cuyo empleo es altamente selectivo, reservado para aquellos grupos considerados, en uno u otro sentido, como "conflictivos" (Sandoval García, 2004: 157).

Por otra parte, cabe destacar que las percepciones se construyen desde elementos subjetivos que corresponden a un momento determinado en el tiempo y que estas pueden ir variando según elementos que puedan incidir de forma significativa. Esto se refleja en las respuestas brindadas a la pregunta *¿Por qué creé eso?*, que procuran obtener esa explicación en torno a los calificativos otorgados a la población migrante según su nacionalidad.

Es así como se identifica que de las respuestas de mayor recurrencia destacan los medios de comunicación como una fuente importante de información sobre población migrante, asimismo, la experiencia personal basada en la convivencia con personas migrantes en diferentes espacios sociales, y rasgos estereotipados asociados por la población encuestada a la nacionalidad de la persona migrante o al país de origen.

Dado lo anterior, es que los datos recolectados en la pregunta abierta *¿Por qué cree eso?* Se someten a análisis utilizando PNL ya que, debido a la gran cantidad y diversidad de respuestas recibidas, aplicar esta metodología resulta valiosa para valorar tanto la existencia de

correspondencia entre las razones expresadas por la población encuestada y su percepción sobre los diversos grupos de población migrante sobre los cuales se consulta; asimismo, utilizar PNL aporta elementos analíticos adicionales que permitan a las personas investigadoras tener una mejor comprensión de las razones e imaginarios de la población costarricense que sustentan sus opiniones y percepciones en relación a estos grupos.

Análisis de texto usando el Procesamiento de Lenguaje Natural

Para el análisis de datos de estudios de opinión resulta apropiado utilizar un conjunto de técnicas que hacen uso de algoritmos de aprendizaje automático o *machine learning*; el cual trabaja sobre documentos de texto, y considera tanto su estructura interna y la distribución de las palabras para la codificación y categorización, las cuales se conoce como Procesamiento de Lenguaje Natural (PNL por sus siglas en inglés) (Bonaccorso, 2017). Así, el PNL permite analizar, comprender y entender el significado de las palabras a través de una computadora que realiza el proceso de manera automática. Como señalan Tintinago *et al*. (2018), el PNL es un campo de estudio que se enfoca en las interacciones entre el lenguaje humano y los ordenadores. Para esto utiliza como herramienta la inteligencia artificial.

Así, se denomina como NLP a un conjunto de técnicas de "*machine learning*" que trabaja sobre documentos de texto, considerando su estructura interna y la distribución de las palabras para la codificación y categorización (Bonaccorso, 2017). Es decir, el PNL permite analizar, comprender y entender el significado de las palabras a través de una computadora realizando el proceso de manera automática.

Una de las aplicaciones que tiene el NLP es la extracción de información de textos, la cual, según Moreira *et al*. (2021) "consiste en la obtención de partes que son importantes en el contenido para pasarlos a una base de datos llenos sobre un tema específico" (p. 130); lo cual facilita a grupo de investigadores en diversas disciplinas sistematizar y clasificar datos relevantes de información textual, que posteriormente puede servir para la interpretación, comprensión o búsqueda de significados para un mejor entendimiento del pensamiento humano.

Además, el NLP busca precisión y comprensión de los significados de acuerdo con consensos sociales de determinada lengua, por medio de algoritmos que hacen una valoración adecuada de las palabras. Según Fernández (2012: 3) el PNL "ocupa de la formulación e investigación de mecanismos eficaces para la comunicación entre personas, o entre personas y máquinas por medio de lenguajes de comunicación humana".

Una de las ventajas de utilizar NLP es que minimiza las interpretaciones erróneas en el lenguaje. Es así como los modelos de inteligencia artificial aplicados a los lenguajes naturales, no solo se centran en la comprensión de los lenguajes sino en aspectos sobre el pensamiento humano y la organización de la información. En el caso de información presente en un documento de texto, o en un mensaje textual, el NLP sirve para facilitar la comprensión de este, dado que los datos en texto tienen mucha información, pero esta no está estructurada. Mediante un algoritmo de inteligencia artificial se puede preprocesar y convertir toda esa información en datos numéricos, además se puede extraer la semántica y significado del contenido.

Para el análisis de texto utilizando NLP, se aplica un método de *tokenización* de documentos, que consiste en separar palabra por palabra, en la cual se crean una serie de parámetros que se obtienen de un vocabulario común para crear su propia "bolsa de palabras". Para ello se utilizan unidades semánticas que son palabras o grupos de palabras que se conocen como tokens. Con estos tokens se forman vectores y luego una matriz que finalmente se usa como entrada para los algoritmos de clasificación como por ejemplo *Naybe Bayes* el más utilizado para esto (Jansen, 2018).

De acuerdo con Fernández (2004), este modelo de disgregación del texto establece un procedimiento analítico que no se centra en unidades textuales aisladas, sino en el valor significativo social del texto en su totalidad; esto se logra a partir del establecimiento del corpus y los datos de entrenamiento y test que se ejecutan en el algoritmo.

La relevancia de la dimensión social del texto se origina desde el sentido práctico de las palabras, "ya que el procesamiento del significado dependerá de las condiciones no lingüísticas en las que se produce la instancia no comunicativa" (Fernández, 2004: 4). Es decir, el procesamiento del lenguaje a través del análisis textual deberá responder a la forma que se produce la convencionalización del significado del texto y esto se logra mediante la transformación del vocabulario en vectores que pueden ser fácilmente usados para clasificar y agrupar las palabras.

Como parte del proceso del algoritmo de PNL se remueven las *stopword* que son palabras de uso común en el habla como: artículos, conjunciones, preposiciones, adverbios, que no proveen una información semántica relevante para el análisis. Además, se realiza una simplificación de las palabras, al hacer una transformación de palabras a verbos o raíces, por ejemplo: El *ama* se transforma al verbo *amar.*

Este conjunto de reglas se aplica varias veces para transformar el texto etiquetado en oraciones que definen la asociación entre una palabra y una parte del habla con un sentimiento calificado. Para la implementación se usan herramientas para etiquetar y una base de datos con claves / frases con evaluaciones de polaridad de emociones" (Hernández y Gómez, 2013: 90).

Por consiguiente, el uso de algoritmos de PNL no es solo para clasificar palabras, sino que permite realizar un análisis más profundo del texto, como señalan Almela *et al*. (2012), la gramática en el lenguaje tiene una relevancia en la lingüística computacional sino también en la psicología, donde "las variaciones en su uso pueden aportar información valiosa sobre el estado mental de la persona, su edad, su sexo, su estatus social o la condición de verdad de su discurso" (p. 70). Es decir, un algoritmo de NLP puede facilitar a los investigadores en el área de lenguaje y psicología la clasificación e identificación de patrones de frases o palabras que permiten caracterizar a los individuos que las escriben.

Aunque actualmente la *minería de opiniones* (término que hace referencia al análisis de sentimientos por medio de textos de manera digital), está orientada más en áreas de mercadeo y manejo de imagen (Hernández y Gómez, 2013), estas técnicas de PNL pueden favorecer diversas áreas sociales para la comprensión del pensamiento humano.

De acuerdo con Bonaccorso (2017) el proceso NLP se puede resumir en las siguientes 3 fases:

1) Limpiar textos y prepararlos para aplicar los algoritmos de ML o solo quedarse con la palabra más relevante.
2) Crear el modelo de *Bag of words* (saco de palabras), que serán las palabras claves a clasificar por medio del algoritmo más apropiado.
3) Aplicar el modelo de ML, más apropiado al *Bag of words*.

Tras completar los pasos anteriores, se realiza el análisis del modelo aplicado usando la matriz de confusión, la cual muestra cómo es la precisión del algoritmo de clasificación que se utiliza. Esta matriz se compone de las siguientes entradas:

- Verdadero positivo (TP): Una muestra positiva correctamente clasificada.
- Falso positivo (FP): Una muestra negativa clasificada como positiva.
- Verdadero negativo (TN): Una muestra negativa correctamente clasificada.
- Falso negativo (FN): Una muestra positiva clasificada como negativa.

Los datos obtenidos de esta matriz se pueden mejorar realizando algunos cálculos o modificaciones. De acuerdo con Raschka y Mirjalili (2017) tenemos dos medidas generales que son el *error de predicción (*ERR) y la *exactitud* (ACC) que proporcionan información general acerca de cuántas muestras se clasifican erróneamente. Así que el ERR se puede entender como el ponderado de la suma de predicciones falsas con todas las predicciones. La fórmula es:

$$ERR = \frac{FP + FN}{TN + TP + FN + FP}$$

Por su parte, la exactitud es un cociente que se calcula como la suma de las predicciones correctas divididas por el número total de predicciones, como se muestra en la fórmula:

$$ACC = \frac{TN + TP}{TN + TP + FN + FP} = 1 - ERR$$

La exactitud (o *accuracy*) muestra el porcentaje de predicciones correctas (verdaderos positivos y verdaderos negativos) frente al total de predicciones. Es útil para datos equilibrados. Además, se cuenta con otras métricas como son: la tasa de verdadero positivo (TPR) y la tasa de falsos positivos (FPR) métricas de rendimiento que son especialmente útiles para problemas de clase desequilibrado (Raschka y Mirjalili, 2017). Dichas métricas se calculan de la siguiente forma:

$$TRP = \frac{TP}{FN + TP}, \quad FPR = \frac{FP}{TN + FP}$$

La tasa de verdaderos positivos (TPR) se conoce también como *recall* (sensibilidad), esta métrica tiene la capacidad de detectar muestras positivas verdaderas entre todos los positivos potenciales. Por su parte el FPR se conoce como la especificidad, que de acuerdo con Bonaccorso (2017) es la proporción entre los casos negativos bien clasificados por el modelo, respecto al total de negativos. Está métrica permite discriminar los casos negativos, es decir valorar que tan bueno el modelo para no obtener falsos positivos.

Otra de las métricas utilizadas es la precisión, la cual refiere a que tan cerca está el resultado de una predicción del valor verdadero; para lo cual se calcula el cociente entre los casos positivos bien clasificados por el modelo y el total de predicciones positivas:

$$Precision = \frac{TP}{FP + TP}$$

Se puede emplear una métrica más para ver el comportamiento de los valores, en este caso se aplica una media armónica ponderada entre la precisión y la tasa de verdaderos positivos (*recall*), la cual suele denominarse como *F1-score* y el valor de beta está entre 1 y 2, y se calcula de la siguiente manera:

$$F_{Beta} = (\beta^2 + 1)\frac{Precision * Recall}{\beta^2 * Prescision + Recall}$$

Esta métrica nos proporciona información sobre la precisión de la predicción, dado que la puntuación más alta se logra dando más importancia a la precisión (que es mayor), mientras que el menor corresponde a un predominio de *recall*. Por lo tanto, *F-Beta (F1-score)* es útil para tener una imagen compacta de la exactitud como compensación entre alta precisión y un número limitado de falsos negativos.

Método de análisis

Para el desarrollo del análisis, se aplicaron técnicas de procesamiento de lenguaje natural a datos de una encuesta de opinión sobre migraciones con el objetivo de hacer una clasificación de las opiniones en varias categorías. En síntesis, el NLP consiste en tomar un texto en lenguas naturales (español, inglés, etc) y dividirlo en palabras, analizarlas, buscar características comunes (patrones) que luego pueden ser usadas como entrada a diversos algoritmos de *machine learning* (regresión lineal, *Naive Bayes*, árboles de decisión, redes neuronales) y hacer clasificaciones. El proceso realizado se describe a continuación:

Fase 1: Limpieza de archivo. Se recibe un archivo Excel con datos preprocesados de una encuesta sobre migraciones de población. En el archivo vienen pares de columnas una con el texto de la opción y otra con la categoría asociada al tipo de población migrante. Se procesa el archivo y se unen todos los pares de columnas: Queda finalmente un archivo con 8919 registros con dos columnas que contiene todos los datos de los diferentes grupos de población (etnia) y sus diferentes categorías por grupo. Como cada grupo poblacional viene con diferentes categorías realiza un proceso de homologación del cual resultan 11 categorías como se muestra la figura 2.

Figura 2. Homologación de las categorías para su análisis

	HOMOLOGACIÓN								
Grupo	POBREZA (Cat 1)	BUSQ TRABA (Cat 2)	NECES APOYO (Cat 3)	REPRES OTRA CULTURA (Cat 4)	RASGOS NEGATIVOS (Cat 5)	NO CONOZCO (Cat 6)	RASGOS POSITIVO (Cat 7)	LUCHA DERE (Cat 8)	AMBIVALENTE (Cat 9)
ETNIA	1	2	3	4	5	6	7	8	9
HAITI	1	X	X	4	5	6	2,3	X	X
Indígenas panameños	1	X	X	7	4	6	5,2,3	8	X
Colombianos	X	X	X	X	1,3	X	2,5	X	4
Nicaragüenses	X	X	X	X	2,5	X	1,4	X	3
Dominicanos	X	X	X	X	1,3	X	2	X	4
Españoles	X	X	X	X	4	6	1,2,3,5	X	X
Africanos	X	X	X	X	1	3	2,4	X	X
Estadounidenses	X	X	X	X	3,6	X	1,2,4,5	X	X
Chinos	x	x	x	7	5,6		1,2,3,4	X	x

Fuente: Construcción propia.

En la Figura 2 se muestra la tabla de homologación de las categorías, la columna "Grupo" representa cada uno de los grupos de población analizados. En el resto de las columnas representan cada una de las categorías nuevas. El número de cada columna indica cuál o cuáles categorías de un grupo coinciden con la categoría nueva. Las "X" son las categorías que fueron movidas para formar parte de la nueva categorización. Por ejemplo, en Haití, categorías 2 (búsqueda de trabajo) y 3 (necesidad de apoyo para mejor calidad de vida) pasan ser parte de la categoría 7 (rasgos positivos).

Fase 2: Tokenizar. Se toma cada una de las filas del archivo que contienen una muestra del texto o párrafo ingresada por cada usuario. Cada uno de estos textos se divide en palabras

Fase 3: Eliminar caracteres. Se eliminan tildes, comas, caracteres especiales y se dejan solo caracteres de la entre "a" y "z".

Fase 4: Solo minúsculas. Se pasa todo el texto a minúscula.

Fase 5: Quitar palabras irrelevantes. Dejar solo la raíz de las palabras. De cada verbo se elimina la conjugación presente, pasado y futuro y se deja solo la raíz. Por ejemplo: de la palabra "ayudarles" se obtiene "ayudar"

Fase 6: Crear la bolsa de palabras. Es una matriz donde se toman todas las pablaras que quedaron del proceso anterior (Fase 5), y se coloca cada una en una columna. La última columna es el código de categoría. Las filas son cada una de las observaciones de los usuarios, y el numero representa cuantas veces aparece la palabra en cada observación. La tabla 1 ilustra una parte de la bolsa de palabras.

Tabla 1. Ejemplo de matriz dispersa de la bolsa de palabras

	Problemas	Oportunidades	Bueno	Malo	Clasificación CÓDIGO
observación usuario 1	1	0	1	0	1
observación usuario 2	0	0	1	0	1
observación usuario 3	1	1	0	1	2
observación usuario n	0	0	1	0	7

Fuente: Construcción propia.

Fase 7: Seleccionar el algoritmo para el modelo de clasificación. En esta etapa, utilizando los datos de la etapa anterior, se probaron varios algoritmos y se determinó cuál de ellos daba la mejor clasificación. Los algoritmos que fueron probados fueron: algoritmo de *K-NN*, algoritmo de SVM, algoritmo de Bayes, algoritmo de *Random Forest* y algoritmo de Regresión Lineal

El modelo de clasificación consiste en tres etapas: 1) Entrenamiento y prueba, en el cual se toman una parte de los datos de encuesta para entrenar el modelo y otra parte para probar su funcionamiento. Estos datos corresponden con la bolsa de palabras del paso anterior. Los datos se dividieron de la siguiente manera: 80% datos de entrenamiento y 20% datos de prueba. 2) Crear el modelo, se usan los datos de entrenamiento para encontrar patrones aplicando los algoritmos de clasificación antes mencionados. 3) Probar el modelo, se comparan los datos de prueba con los datos de entrenamiento para medir la precisión de la clasificación realizada por el modelo. Para cada algoritmo se calcula la matriz de confusión y se obtienen las siguientes métricas: precisión, *recall, F1-score y Accuracy*. Finalmente, con el algoritmo con mejores resultados se aplican las técnicas de *K-fold cross y Grid Search* para optimizar los resultados.

Resultados del análisis

Dadas las características del proceso de recolección de datos, en las cuales las personas entrevistadas podían responder libremente sobre qué opinan respecto a las poblaciones migrantes en Costa Rica, las categorías 5 (rasgos negativos) y 7 (rasgos positivos) permiten recoger cómo dichas opiniones tienden a identificar rasgos positivos o negativos sobre dichos grupos; lo cual es un elemento valioso para comprender la construcción que la valoración costarricense realiza sobre las poblaciones migrantes y su disposición a aceptarlas o rechazarlas. Aunado a esto, en el proceso de homologación de categorías (ver figura 2), algunas categorías movieron todos o la mayoría de sus datos a otra, dejando estas sin datos (categorías 2 y 3) o con muy pocos datos (categorías 1,4,8 y 9). Por esta razón, los algoritmos de *Machine Learning* utilizados dan mejores clasificaciones con las categorías que tienen más datos, es el caso de las categorías 5 y 7 (relevantes para el estudio) y las categorías 88 y 99 (NSR y otros, no relevantes para el estudio).

Se puede observar en la figura 4 el análisis de todas las categorías para los datos de prueba (20%). Las categorías 2 y 3 ni siquiera son considerados para la prueba pues no cuentan con datos para ello. Las categorías 1,4 y 8 cuentan con muy pocos de entrenamiento y por ello las bajas predicciones. La categoría 9 cuenta con buenas predicciones, pero al tratarse de la categoría "Ambivalente" son sirve para el análisis.

Una vez obtenidas todas las matrices de confusión se calculó el *Accuracy* de cada algoritmo y la precisión, así como el *recall* y *f1-score* para las categorías más relevantes para la investigación que son la categoría 5 (rasgos negativos) y la categoría 7 (rasgos positivos) de cada algoritmo (ver tabla 2).

Tabla 2. Resultados de las métricas aplicadas

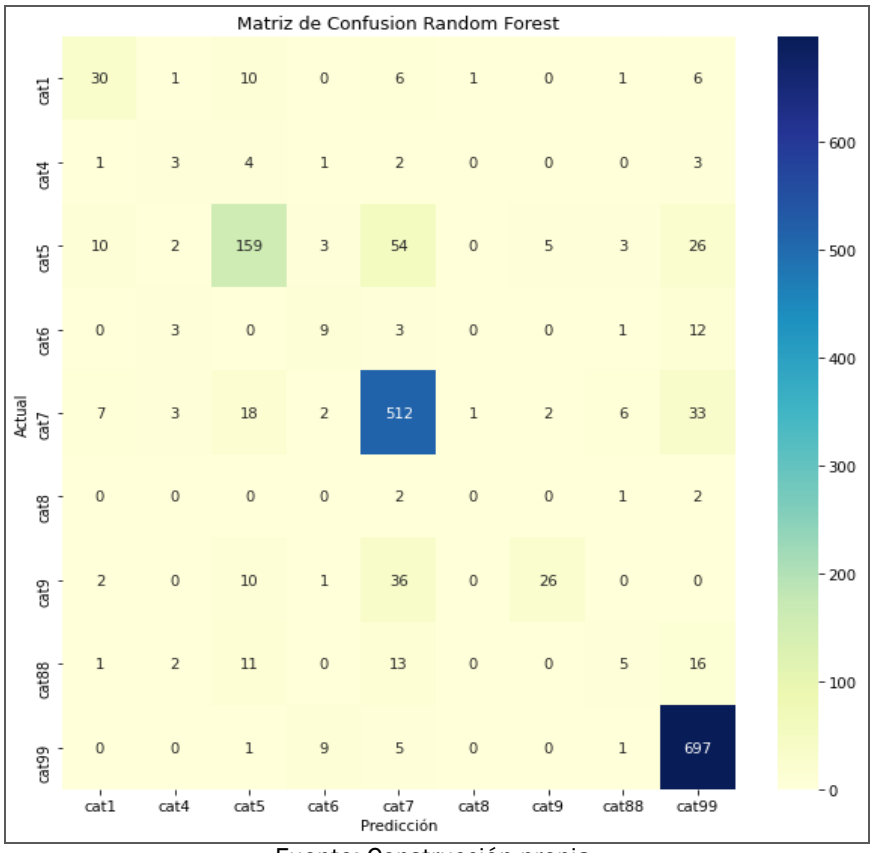
Algoritmo	Accuracy	Precisión (Cat 5)	Precisión (Cat 7)	Recall (Cat 5)	Recall (Cat 7)	F1 (Cat 5)	F1 (Cat 7)
algoritmo de K-NN	0.76	0.66	0.78	0.49	0.83	0.56	0.80
algoritmo de SVM	0.79	0.69	0.84	0.61	0.79	0.75	0.81
algoritmo de Bayes	0.56	0.34	0.75	0.31	0.29	0.33	0.42
algoritmo de Random Forest	0.81	0.74	0.81	1.58	0.88	0.65	0.84
algoritmo de Regresión Lineal	0.80	0.67	0.84	0.61	0.83	0.64	0.84

Fuente: Construcción propia.

Por los resultados obtenidos se determina que el algoritmo que realiza mejor la clasificación es el *Random Forest.* Donde se obtiene una precisión de un 74% y un 81% respectivamente. Y además todo el algoritmo tiene una exactitud de un 81%

Por otra parte, el gráfico 1 muestra el número de predicciones buenas (aciertos) y predicciones malas (desaciertos) por cada categoría. Así, el gráfico muestra que en las categorías 5 y 7 las predicciones buenas superan a las predicciones malas a diferencia de las otras categorías. La categoría 99 es un caso especial pues corresponde con la categoría "no se/no respondo".

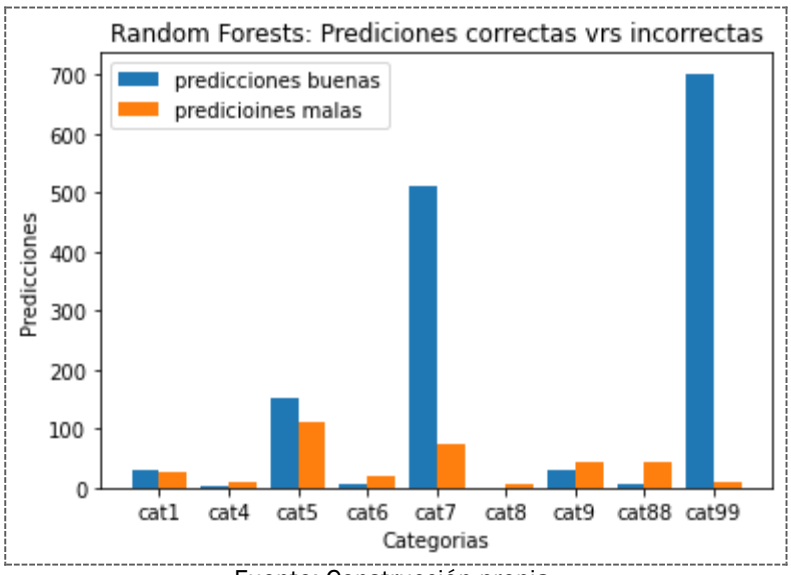
Figura 3. Matriz de confusión del *random forest*



Fuente: Construcción propia.

La figura 4 muestra la matriz de confusión, en la cual cada fila representa la predicción realizada por el modelo para cada categoría. La diagonal principal muestra los *verdaderos positivos.* Es decir, el dato estaba clasificado en una categoría y el modelo lo clasificó en esa misma categoría. Por ejemplo, de toda la fila de la categoría 5 (rasgos negativos), con 262 muestras, en 151 hubo acierto. El resto de los datos de la fila muestra en qué otra categoría cayó el dato en el que no hubo acierto. Además, la barra lateral corresponde con un mapa de calor, donde el color más intenso representa más aciertos.

Gráfico 1. Cantidad de predicciones buenas vs. malas



Fuente: Construcción propia.

El gráfico 1 muestra las predicciones correctas e incorrectas de cada categoría, se puede resaltar la predicción correcta de las categorías 5 (rasgos negativos) y categoría 7 (rasgos positivos). En el caso de la predicción de categoría 99 (no se no responde) el modelo acierta con mayor precisión porque todas las observaciones son los caracteres NS, ns, N/S (no respondo).

Una vez determinado que el *random forest* es el algoritmo con mejores resultados, se aplicó el *Grid search* para comprobar que los parámetros usados por el algoritmo fueron los mejores. Luego se calculó la precisión del algoritmo usando la técnica de *K-Fold Cross* y se obtuvo una media global de 0.7751586188472583 que es un 77,6% de precisión y una desviación estándar 0.013976548078564405, es decir, 1.4%. Con de *K-Fold Cross* se puede determinar que la calidad de la precisión está balanceada y que se mantendrá en el rango de 72.2% a 78.8%. Por lo cual el modelo tiene una probabilidad baja de tener sobreajuste.

Conclusiones

El uso de NLP aplicado a encuestas con preguntas abiertas, demuestra que se puede automatizar el proceso clasificación con una precisión buena respecto al proceso que hace manualmente un grupo de clasificadores. Además, mejora el tiempo de clasificación, con lo cual los investigadores pueden enfocarse más rápido en el análisis de los datos.

Una vez construido el modelo este puede continuar entrenando y aplicando en diferentes encuestas que contengan preguntas similares. Entre más datos se tengan de entrenamiento el algoritmo mejora su precisión. Considerando que se debe ir analizando un posible sobreajuste a medida que se van aumentando los datos.

Se pudo observar que al aplicar el NLP con el algoritmo *random forest* a encuestas con preguntas abiertas permite que el sistema automáticamente clasifique las opiniones en diferentes categorías ayuda a disminuir la ambigüedad en la clasificación. Esto es un aporte importante cuando se está trabajando con datos de percepción sobre temas polémicos, o bien, donde el tipo, cantidad y calidad de la información recolectada es amplia, tal y como sucede con los datos analizados en el presente artículo. Asimismo, el proceso investigativo permitió crear una metodología para preprocesado de los datos. En este caso se pudo unificar datos de encuestas de diferentes grupos de población para generar un base más grande para entrenar el modelo.

Así, al aplicar el algoritmo *random forest* a datos sobre precepción de las poblaciones migrantes en Costa Rica, se pudo corroborar que este pudo clasificar de manera eficiente y adecuada los datos recolectados, tal y como otros trabajos han demostrado (Bonaccorso, 2017; Tintinago *et al.,* 2018), demostrando que con suficiente entrenamiento puedo replicar los resultados de clasificación generados por las personas investigadoras, pero disminuyendo los problemas de ambigüedad y sesgos de clasificación, permitiendo tener una mejor calidad de datos para comprender el fenómeno de estudio.

Por último, lo expuesto en el artículo muestra lo importante que es ampliar el abordaje interdisciplinario en los estudios de opinión. Tradicionalmente los estudios de opinión han sido trabajado por disciplinas como la ciencia política, la sociología, la psicología, el marketing, entre otras; las cuales han permitido construir instrumentos y propuestas para mejorar la calidad y validez de dichos estudios; no obstante, el incorporar disciplinas no tradicionales, como en este caso la ciencia de datos, pueden generar nuevas herramientas y procedimientos para procesar y analizar la información recolectada en los estudios de opinión, permitiendo llegar a obtener un análisis más precisos de las actitudes, valoraciones y opiniones expresadas por las personas a través de las encuestas.

Referencias

ALMELA SÁNCHEZ-LAFUENTE, Á., VALENCIA GARCÍA, R. & CANTOS GÓMEZ, P. (2012). Detectando la mentira en lenguaje escrito. *Procesamiento Del Lenguaje Natural,* (48), 65-72. Recuperado de http://rua.ua.es/dspace/handle/10045/22032

BEER, D. (2017). The social power of algorithms. *Information, Communication & Society*, *20*(1), 1-13. https://doi.org/10.1080/1369118X.2016.1216147

BONACCORSO, G. (2017). *Machine learning Algorithms*. Birminghan, Mumbai: Packt Publishing Ltd.

CABRERA-ÁLVAREZ, P. (2022). Survey Research in Times of Big Data: Investigación con encuestas en los tiempos del big data. *EMPIRIA: Revista de Metodología de Ciencias Sociales*, *53*, 31-51. https://doi.org/10.5944/empiria.53.2022.32611

CETINA PRESUEL, R. & MARTÍNEZ SIERRA, J. M. (2019). Algorithms and the News: Social Media Platforms as News Publishers and Distributors. *Revista de Comunicación*, *18*(2), 261-285. https://doi.org/10.26441/rc18.2-2019-a13

DELGADO MONTALDO, D. (2008). Percepciones de la inmigración e integración en Costa Rica. *Papeles de población*, 14(57), 65-91.

DÍAZ, R. M. & DOMÍNGUEZ, Á. Z. (2013). Las ciencias sociales y los dispositivos de la complejidad. *Cuadernos de Administración*, *29*(50), 123-131.

FERNÁNDEZ, A. M. (2004). El procesamiento del texto como lenguaje natural. *Hermeneus: Revista de la Facultad de Traducción e Interpretación De Soria,* (6), 75-98.

FERNÁNDEZ, G. M. (2012). Adquisición y representación del conocimiento mediante procesamiento del lenguaje natural. Universidad de la Coruña [Tesis Doctoral]. Recuperado de https://ruc.udc.es/dspace/bitstream/handle/2183/10057/FernandezGavilanes_Milagros_TD_2012.pdf?sequence=5&isAllowed=y

FONDEVILA GASCÓN, J. F. (2017). Algoritmos sobre el impacto de los medios de comunicación en medios sociales: Estado de la cuestión. *Icono 14*, *15*(1), 21-41. https://doi.org/10.7195/ri14.v15i1.948

GONZÁLEZ, F. (2019). Big data, algoritmos y política: Las ciencias sociales en la era de las redes digitales. *Cinta de moebio*, *65*, 267-280. https://doi.org/10.4067/s0717-554x2019000200267

GUALDA, E. (2022). Social big data y sociología y ciencias sociales computacionales. *EMPIRIA: Revista de Metodología de Ciencias Sociales*, *53*, 147-177. https://doi.org/10.5944/empiria.53.2022.32631

HERNÁNDEZ, M., Y GÓMEZ, J. (2013). Aplicaciones de Procesamiento de Lenguaje Natural. *Revista Politécnica*, *32* (1), 87-96. https://doi.org/10.33333/rp.vol32n0.32

IDESPO-UNA (2012). Encuesta: Construcción de opiniones públicas sobre la migración en Costa Rica.

MAMAQI, X., BANDRES GOLDARAZ, E., & PEREZ CALLE, R. D. (2021). Analisis Big Data de la opinión pública sobre inmigración en redes sociales.En Pérez Calle, R.D., Trincado Aznar, E. y Gallego Abaroa, E. (Coord.) *Economia, empresa y justicia. Nuevos retos para el futuro* (pp. 1451-1467). Madrid : Dikson

MAXWELL, J. & TOMLINSON, J. (2020). Proving algorithmic discrimination in government decision-making. *Oxford University Commonwealth Law Journal*, *20*(2), 352-360. https://doi.org/10.1080/14729342.2020.1833604

MOREIRA, D., CRUZ, I., GONZÁLEZ, K., QUIRUMBAY, A., MAGALLAN, C., GUARDA, T., ANDRADE, A. & CASTILLO, C. (2021). Análisis del Estado Actual de Procesamiento de Lenguaje Natural. *Revista Ibérica De Sistemas e Tecnologias De Informação*, 126-136.

PORTER, N. D., VERDERY, A. M. & GADDIS, S. M. (2020). Enhancing big data in the social sciences with crowdsourcing: Data augmentation practices, techniques, and opportunities. *PLoS ONE*, *15*(6), 1-21. https://doi.org/10.1371/journal.pone.0233154

RASCHKA, S. y MIRJALILI, V. (2017). *Python Machine learning. Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow.* Mumbai: Packt Publishing.

SANDOVAL GARCÍA, C. (2004). El "otro" nicaraguense en el imaginario colectivo costarricense. Algunos retos analíticos y políticos. *Nómadas* (Col), (20),152-159.

TINTINAGO, A., MIÑOZ, Y., URIBE, G. A. & ÁLVAREZ SÁNCHEZ, P. H. (2018). Etiquetado asistido de documentos de investigación mediante procesamiento de lenguaje natural y tecnologías de la web semántica. *Scientia et Technica*, 23(4), 528–537.

TU, S.-T., LU, L., HSIEH, C.-H. & WU, C.-Y. (2021). A New Internet Public Opinion Evaluation Model: A Case Study of Public Opinions on COVID-19 in Taiwan. *International Journal of Big Data and Analytics in Healthcare*, *6*, 1-17. https://doi.org/10.4018/IJBDAH.287603

ZORAN, U. (2016). Las Ciencias Computacionales como recurso para la toma de decisiones: Los algoritmos. *Revista Antioqueña de las Ciencias Computacionales*, *6*(2), 55-59.

Autores.

Esteban Martínez Porras

Escuela de Informática, Universidad Nacional (Costa Rica) y Escuela de Matemáticas, Universidad de Costa Rica.

Máster en Planificación curricular y Licenciado en Enseñanza de la Matemática ambos grados por la Universidad de Costa Rica. Docente en la Sede Interuniversitaria de la Universidad Nacional. Investigador del proyecto Evaluación de las tecnologías digitales que refuerzan el desarrollo del pensamiento lógico matemático en estudiantes de ingeniería industrial: propuesta de indicadores de calidad de la Universidad de Costa Rica.

E-mail: esteban.martinez.porras@una.ac.cr

Adrián Ramírez Fernández

Escuela de informática, Universidad Nacional, Costa Rica.

Máster en Telemática por el Instituto Tecnológico de Costa Rica y Bachiller en Ciencias de la computación, Universidad de Costa Rica. Docente en la Sede Interuniversitaria de la Universidad Nacional.

E-mail: adrian.ramirez.fernandez@una.ac.cr

Laura Solís Bastos

Instituto de Estudios Sociales en Población (IDESPO), Universidad Nacional, Costa Rica.

Doctora en Demografía por la Universidad Nacional de Córdoba, Argentina, Máster en Estudios Latinoamericanos con énfasis en cultura y desarrollo, Licenciada en Sociología, ambos grados por la Universidad Nacional, Costa Rica. Investigadora del programa Umbral Político, adscrito al Idespo, Universidad Nacional, Costa Rica

E-mail: laura.solis.bastos@una.ac.cr

José André Díaz-González

Instituto de Estudios Sociales en Población (IDESPO), Universidad Nacional, Costa Rica.

Doctor en Gobierno y Políticas Públicas, Magíster en Historia y Licenciado en Ciencias Políticas por la Universidad de Costa Rica. Investigador del Programa Umbral Político del Instituto de Estudios Sociales en Población de la Universidad Nacional, y docente de la Escuela de Ciencias Políticas de la Universidad de Costa Rica.

E-mail: jose.diaz.gonzalez@una.ac.cr

Citado.

MARTÍNEZ PORRAS, Esteban; RAMÍREZ FERNÁNDEZ, Adrián; SOLÍS BASTOS, Laura y DÍAZ-GONZÁLEZ, José André (2025). Uso de Procesamiento de Lenguaje Natural para procesar respuestas abiertas de una encuesta de Opinión Pública. *Revista Latinoamericana de Metodología de la Investigación Social - ReLMIS*. Nº29, Año 15, pp. 51-67.

Plazos.

Recibido: 28/10/2022. Aceptado: 15/06/2023.