MACHINE LEARNING

para la categorización de respuestas de preguntas abiertas

BICENTENARIO DEL PERÚ 2021 - 2024

OFICINA DE SEGUIMIENTO Y EVALUACIÓN ESTRATÉGICA (OSEE)

Síntesis

Diciembre 2022

Jefe de la Oficina de Seguimiento y Evaluación Estratégica

Juan Manuel García Carpio

Elaboración de contenidos:

- Erik Carl Candela Rojas
- Elsa Gabriela Cañari Huerta

MINISTERIO DE EDUCACIÓN
Oficina de Seguimiento y Evaluación Estratégico
Diciembre - 2022
Sede Central: Calle Del Comercio N° 193 Lima - Lima - San Borja - 15021 Perú
Teléfono: (01) 615-5800
https://www.gob.pe/minedu

MACHINE LEARNING
para la categorización de respuesta de preguntas abiertas

Presentación:

La presente propuesta de innovación metodológica consiste en categorizar automáticamente las "respuestas de preguntas abiertas" provenientes de los instrumentos de recojo de información con el fin de reducir los costos que requiere una categorización manual.

Las preguntas abiertas son utilizadas en los instrumentos[1] de recojo de información que elabora las unidades de la Oficina de Seguimiento y Evaluación estratégica (OSEE) de la Secretaria de Planificación Estrategia (SPE) del Ministerio de Educación. Este tipo de preguntas son empleadas en los instrumentos cuando se busca explorar ciertos temas de interés a mayor profundidad,[2] y no se usan categorías preestablecidas de respuestas.

Las respuestas a estas preguntas corresponden a textos (datos no estructurados) que requieren de un tratamiento previo para transformarlas en datos estructurados, facilitando así su análisis. Actualmente, un tratamiento que viene realizando la OSEE es agrupar o categorizar las respuestas con contenido similar de forma manual. Sin embargo, este proceso suele ser muy costoso en términos de tiempo y dinero, especialmente cuando existen miles de respuestas que requieren categorización.

Por tal motivo, se buscó alternativas para poder reducir el costo de categorizar las respuestas de preguntas abiertas provenientes de instrumentos de recojo de información. Específicamente, se busca categorizar las respuestas abiertas automáticamente, de forma total o parcial. En ese sentido, el equipo de la OSEE diseñó la innovación "Machine Learning para la categorización de respuesta de preguntas abiertas". La innovación empleó técnicas analíticas de *Machine Learning*[3] para poder procesar y categorizar todas las respuestas provenientes de una pregunta abierta, reduciendo así el costo que implica categorizarlas de forma manual.

Palabras claves: Pregunta abierta, *topic labeling*, *open-ended question*, *Machine Learning*, aprendizaje automático, *clustering*.

[1] El instrumento empleado para la innovación corresponde al módulo de "materiales educativos" del Sistema de Información de Apoyo a la Gestión de la Institución Educativa (SIAGIE).
[2] Bradburn et al. (2004) identifica múltiples ventajas y desventajas de las preguntas abiertas.
[3] Machine Learning (ML) o aprendizaje automático es la ciencia (y arte) de programar las computadoras para que puedan aprender de los datos (Géron, 2019, pág. 8).

1. LA INNOVACIÓN

1.1 Problemática

Las preguntas abiertas constituyen una parte fundamental de los instrumentos de recojo de información, ya que permiten recoger la opinión de los encuestados en todos sus matices, permitiendo así explorar o profundizar el tema de interés, así como obtener respuestas que no han sido anticipadas previamente (Bradburn et al., 2004).

Sin embargo, Bradburn et al. (2004) también señala que, para analizar las respuestas de preguntas abiertas, estas deben ser previamente codificadas o categorizadas manualmente, generando un costo en términos de tiempo y dinero. Incluso los autores manifiestan que esta categorización puede introducir un error al momento de codificar cada respuesta por ser un procedimiento que se realiza manualmente.

El costo elevado que implica categorizar las respuestas de preguntas abiertas afecta también a la labor que pueden estar realizando las diferentes Direcciones y Oficinas del Ministerio de Educación, como es el caso de la Oficina de Seguimiento y Evaluación estratégica (OSEE). Así, se pudo identificar que en un análisis realizado por la OSEE sobre las "Observaciones de directores sobre recepción y asignación de Tabletas" provenientes del módulo de "materiales educativos" del Sistema de Información de Apoyo a la Gestión de la Institución Educativa (SIAGIE), el proceso de categorizar las 81,488 observaciones en 20 categorías tomo un alrededor de aproximadamente 7 días.

1.2 Propuesta de solución

Por esta razón, se impulsó la búsqueda de alternativas de bajo costo que permitan organizar y facilitar el análisis de textos. En ese sentido, se empleó el proceso de *Text Data Mining* el cual permite aplicar algoritmos y métodos del campo de *Machine Learnig* y estadística para encontrar patrones y así estructurar o categorizar los datos no estructurados de los textos (Hotho et al., 2005).

Entre los métodos disponibles de *Text Data Mining*, se optó por emplear el *Clustering* ya que, en base a lo señalado por Hotho et al. (2005), permite ordenar y agrupar textos con similar contenido. De este modo, se revisaron algunos estudios que emplean este método a través de diversos algoritmos.

En primer lugar, Roberts et al. (2014) empleó el "Modelo de Temas Estructurales" (STM) en la encuesta nacional de elección americana para categorizar las opiniones de los encuestados sobre las políticas más importantes que enfrenta los Estados Unidos. Asimismo, Buenaño et al. (2020) empleó el algoritmo de "Asignación de Dirichlets Latentes" (LDA) para agrupar respuestas de una encuesta de opinión realizada a profesores universitarios. Por su parte, Vidal et al. (2022) empleó el "Modelo de Temas Bi-término" (BTM) para conocer el interés de consumidores a partir de una encuesta de opinión sobre granjas verticales. Por último, Chen y Beaver (2022) emplearon el algoritmo "Louvain" para categorizar las noticias de la sección de "gobierno y sociales" de la base de datos de Reuters.

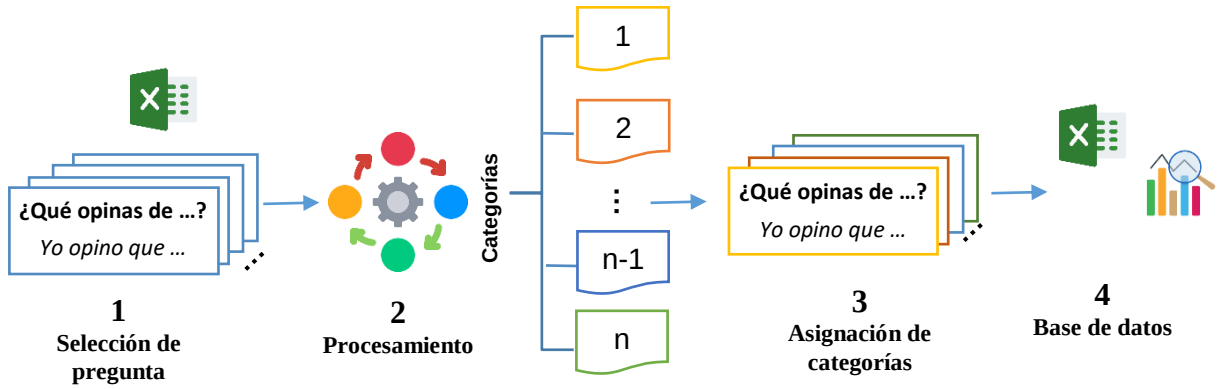
Por otro lado, existen servicios de pago que permiten asignar una categoría a textos de forma automática y sin necesidad de realizar alguna configuración, como es el caso de Google que ofrece el producto de API de *Natural Language*[4] el cual se enfoca en la clasificación de contenidos en base a categorías predefinidas. Asimismo, Bytesview ofrece un servicio similar llamado *Topic Labeling*[5], que permite asignar una categoría estándar a un texto determinado. Por último, OpenAI pone a disposición el modelo *text-davinci-003*[6] mediante el API *Completions* el cual puede ser empleado para la clasificación de textos[7].

[4] La documentación oficial del API (Interfaz de Programación de Aplicaciones) se encuentra disponible en el siguiente link: https://cloud.google.com/natural-language?hl=es
[5] Mayor información del servicio *Topic Labeling* se encuentra disponible en: https://www.bytesview.com/topic-labeling
[6] Modelo de la familia GPT-3 desarrollado por OpenAI, que es empleado mediante el API *Completions* cuya documentación se encuentra en el siguiente link: https://beta.openai.com/docs/api-reference/completions
[7] Ver ejemplo en el siguiente link: https://beta.openai.com/examples/default-classification

Con base a lo señalado hasta el momento, la solución propuesta buscó emplear algoritmos de *Text Data Mining* para realizar el proceso de categorización automática, sin incurrir en costos adicionales de servicios en línea. En la figura 1 se describe los 4 pasos que emplea la innovación desarrollada para categorizar textos.

1. En el paso 1 se selecciona la pregunta abierta, del que se recopilara todas las "m" respuestas que realizaron los encuestados.
2. En el paso 2 se procesa las "m" respuestas con el objetivo de encontrar "n" categorías que las resuman.
3. En el paso 3 se asigna a cada respuesta su respectiva categoría.
4. Por último, en el paso 4 se genera la base de datos (BD) de repuestas de preguntas abiertas con sus respectivas categorías, convirtiendo los textos de las respuestas a datos estructurados que faciliten su análisis.

Figura 1. Pasos para la categorización de textos



2. METODOLOGIA DE CATEGORIZACIÓN

2.1 Datos

Los métodos y algoritmos de *Text Data Mining* se aplicaron a los datos provenientes del módulo de "materiales educativos" del SIAGIE. Este módulo permite registrar y asignar material educativo de acuerdo a los estudiantes matriculados en las instituciones educativas. Su objetivo es contar con información confiable para la identificación del material faltante y sobrante de las IIEE (MINEDU, 2022).

Para este proyecto, se empleó la información del submódulo de Asignación[8] del módulo de "materiales". Específicamente, se utilizaron los datos de observaciones de directores sobre la asignación de tabletas a los estudiantes y docentes, con corte al 18 de febrero del 2022. De esta manera se conformó la base de datos que se utilizó para el desarrollo del presente documento.

Asimismo, esta base de datos se compone de 81,488 observaciones que se realizaron en 7,772 instituciones educativas. Esta base de datos contiene información sobre el código modular, la observación realizada y la categorización manual.

Cabe señalar que la categorización manual se realizó de forma posterior al registro de las observaciones, identificando así 20 categorías codificadas numéricamente. En la Figura 2 se muestra las frecuencias de cada categoría, en donde se observa que la categoría 97 comprende más de 50,000 observaciones y está referida a que no existe ninguna observación con la asignación de la Tableta a los estudiantes y docentes.

[8] El submódulo de "Asignación" permite registrar la asignación de tabletas y cargadores solares a los estudiantes y docentes de la IE.

Figura 2 Frecuencia de observaciones según las 20 etiquetas manuales- "Asignación de tabletas"

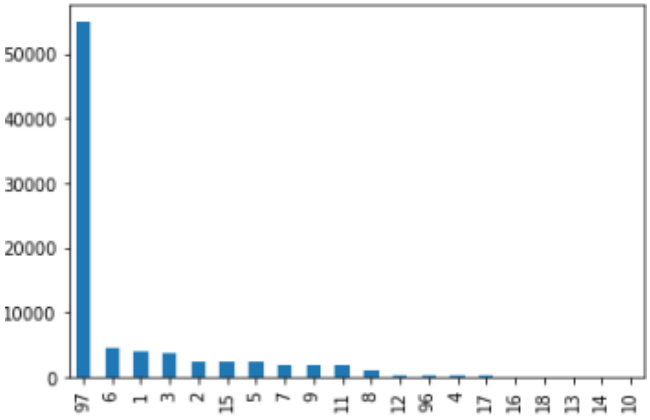
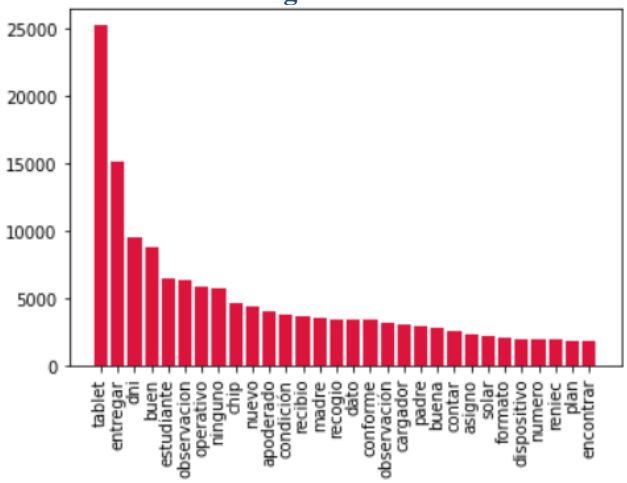


Figura 3 Frecuencia de las 30 palabras más comunes - "Asignación de tabletas"



Adicionalmente, se hizo un conteo de las 30 palabras más comunes y se observó que la palabra "tablet", "entregar" y "dni" son las tres más referidas, como se muestra en la Figura 3.

2.2 Modelo de categorización

Para la categorización de respuestas, solo se consideraron el 15% (12,915) de las respuestas que contienen observaciones del director sobre la asignación de tabletas.

Se aplicaron diversos algoritmos que permiten realizar *Clustering* para agrupar textos con similar contenido, los cuales fueron empleados en los diversos estudios descritos en el punto 1.2 del documento. En la tabla 1 se describen los algoritmos empleados:

Tabla 1: Propuestas de Algoritmos para categorización de textos

N°	Algoritmo propuesto	Descripción
1.	Asignación de Dirichlets Latentes (LDA)	Al ser unos de los modelos más frecuentemente usados para modelamiento de temas, se implementó el LDA. Este trabaja bajo la premisa de que cada documento contiene una mezcla de temas, y cada tema se compone de una mezcla de palabras (Blei et al., 2003).
2.	Modelo de temas estructurales (STM)	El modelo de temas estructurales se destaca del resto de modelos al permitir incluir covariables (variables propias del autor del texto) que afectan la prevalencia o frecuencia de temas (Roberts et al., 2014).
3.	Modelo de temas bitermino (BTM)	Con base a lo señalado por Yan et al. (2013), el algoritmo se basa en la coocurrencia de palabras que identifica temas mediante el modelado de bitérminos, donde un bitérmino consta de dos palabras que se encuentran en el mismo contexto. Asimismo, el autor señala que el BTM es apropiado para textos cortos.
4.	Redes complejas (Grafos) y Louvain	Permite representar los datos (nodos) en un grafo que posee propiedades estadísticas y topológicas para entender la conexión entre estos (Blondel et al., 2008).

2.2.1 Propuesta 1: Asignación de Dirichlets Latentes (LDA)

La tabla 2 muestra las 8 agrupaciones de temas identificadas mediante el algoritmo LDA.

Tabla 2: Identificación de grupos de textos mediante LDA

Tema 1	Tema 2	Tema 3	Tema 4	Tema 5	Tema 6	Tema 7	Tema 8
serie	cargador	formato	dni	ente	tablet	chip	estudiante
tableta	chip	control	madre	cobertura	estudiante	tableta	tableta
estudiante	dato	entregar	sistema	chip	asignado	internet	dni
grado	firmar	tableta	padre	plan	interno	conectividad	entrega
asignar	solar	apoderado	valido	datos	tener	operador	padre
año	memoria	firmar	apoderado	tablet	alumno	celular	tableta
numero	jurado	estudiante	recoger	lugar	alumna	registro	motivo
tablet	declaracion	padre	estudiante	estudiante	tableta	entregar	recoger
docente	entregar	numero	mama	rosario	asigno	señal	madre
coincidir	validado	cambiar	validar	poner	entregar	problema	encontrar

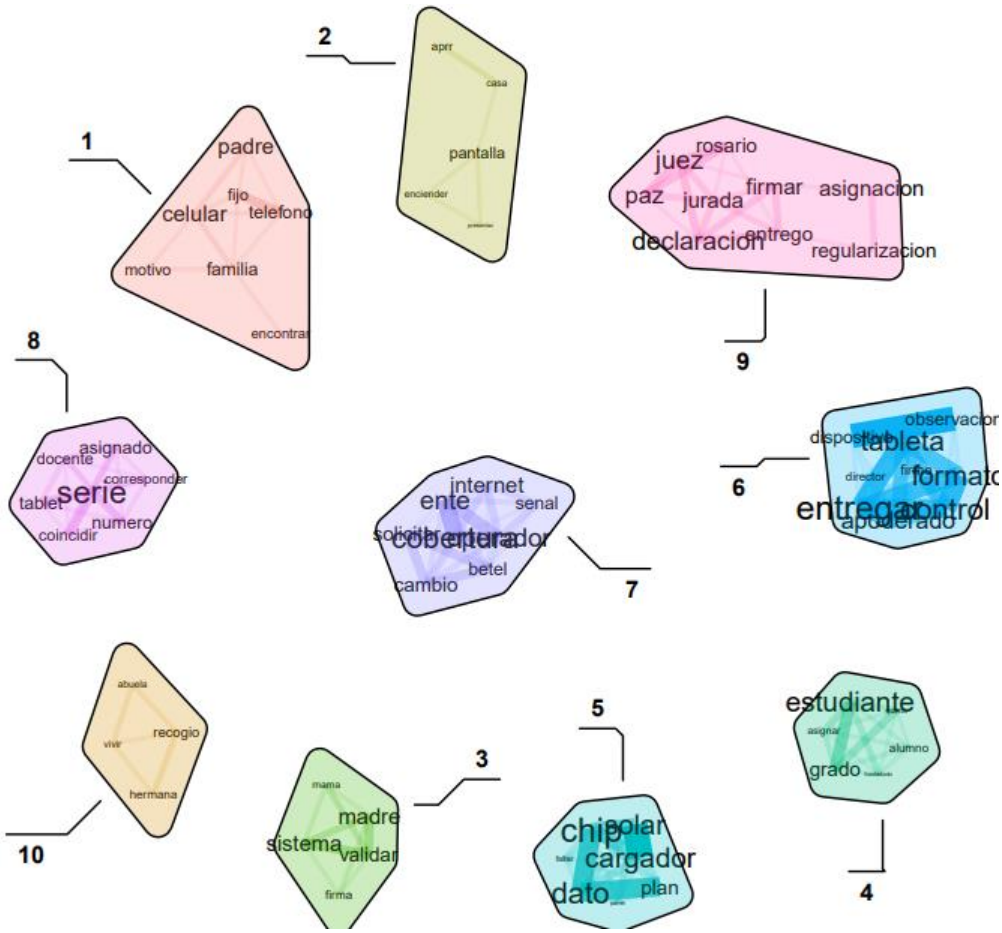
A partir de las 8 agrupaciones se pudo inferir las siguientes categorías de observaciones:

- Tema 1: Serie/número de tableta asignada.
- Tema 2: Accesorios de la tableta.
- Tema 3: Formato de control. La tableta fue entregada al apoderado.
- Tema 4: La validez del DNI del padre/madre en el sistema.
- Tema 5: Plan de datos, operador de internet.
- Tema 6: Asignación de tableta al alumno
- Tema 7: Chip y operador de internet.
- Tema 8: Entrega de la tableta al padre o madre del estudiante.

2.2.2 Propuesta 2: Modelo de temas bitérmino (BTM)

Se evaluó el modelo BTM con 5, 10 y 15 temas. De acuerdo a los resultados obtenidos, la interpretación resulto más sencilla en el modelo con 10 temas. La figura 4 muestra los 10 temas identificados mediante el algoritmo BTM.

Figura 4: Identificación de grupos de textos mediante BTM



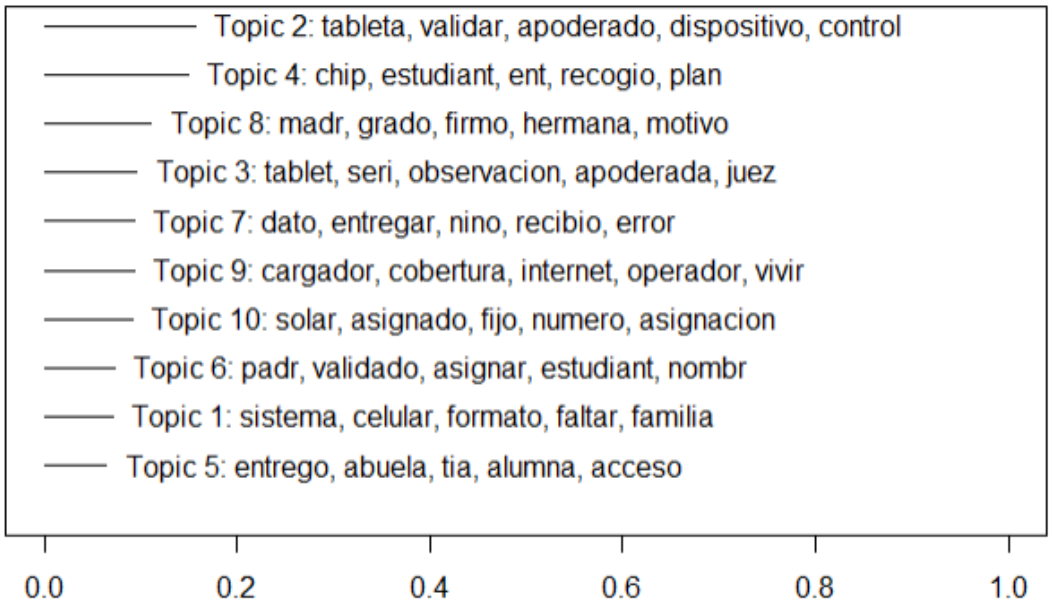
A partir de las 10 agrupaciones se pudo inferir las siguientes categorías de observaciones:

- Tema 1: Teléfono fijo y celular del padre
- Tema 2: Pantalla y encendido del dispositivo
- Tema 3: Sistema y validación de registros de familiares.
- Tema 4: Asignación al estudiante y su grado de estudio.
- Tema 5: Cargador solar, chip, plan de datos.
- Tema 6: Formato de control y observaciones con la firma.
- Tema 7: Operador, de internet, señal de internet y cobertura.
- Tema 8: Nro. de serie, problemas con la coincidencia y correspondencia.
- Tema 9: Firma y declaración jurada.
- Tema 10: recogió hermana o abuela que vivía con el alumno.

2.2.3 Propuesta 3: Modelo de temas estructurales (STM)

La figura 5 muestra las 10 agrupaciones identificadas mediante el algoritmo de STM.

Figura 5: Identificación de grupos de textos mediante STM



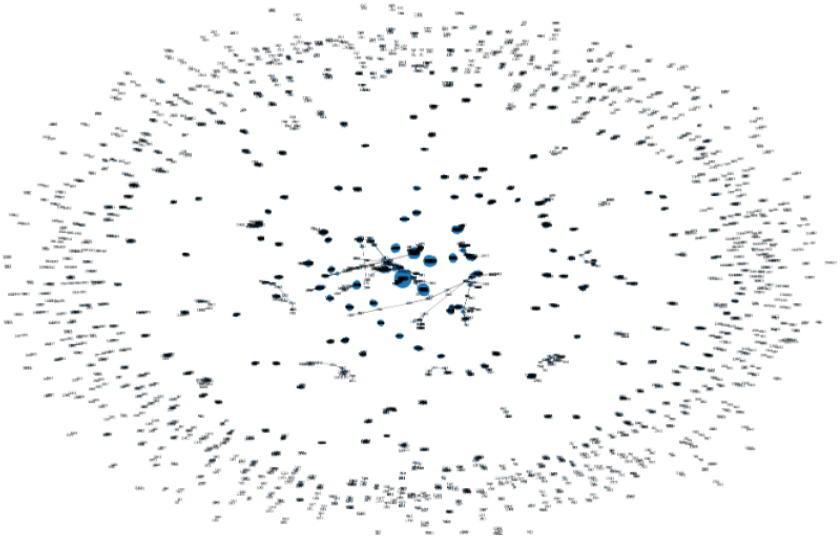
A partir de las 10 agrupaciones se pudo inferir las siguientes categorías de observaciones:

- Tema 1: Sistema de registro.
- Tema 2: Validez del apoderado del alumno.
- Tema 3: Número de serie.
- Tema 4: Chip y operador de internet.
- Tema 5: Entrega y firma de un familiar del alumno.
- Tema 6: Validez del padre y asignación.
- Tema 7: Algún error en la entrega y recepción.
- Tema 8: Firma de algún familiar del alumno.
- Tema 9: Cargador, así como la cobertura del operador de internet.
- Tema 10: Número (de teléfono) fijo.

2.2.4 Propuesta 4: Modelo Redes complejas y Louvain

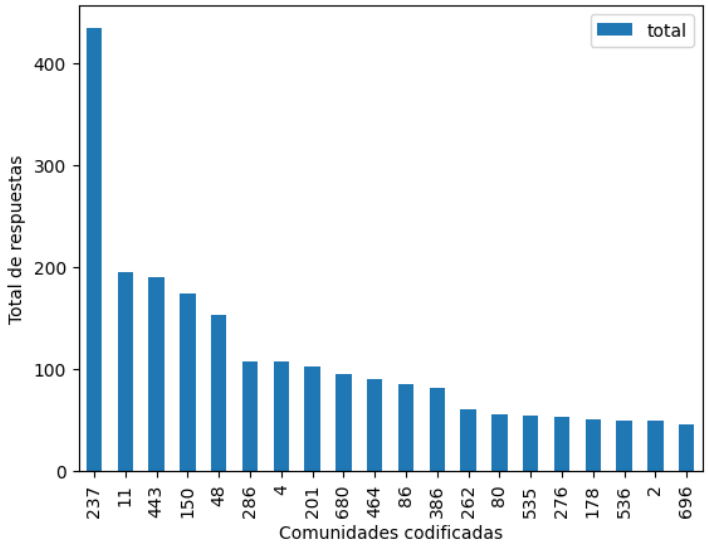
Esta propuesta está compuesta en dos etapas. En primer lugar, se transforman todos los textos para que puedan ser representados a través de redes complejas (grafo). A partir de la red se identifican comunidades mediante el algoritmo Louvain. Cuando se aplicó el procedimiento en las 12,915 observaciones se pudo identificar 813 comunidades (categorías), donde cada comunidad agrupa textos con similar contenido. La figura 6 muestra la red compleja y las múltiples agrupaciones que el algoritmo Louvain pudo identificar.

Figura 6: Identificación de grupos de textos mediante Louvain



Cada una de estas 813 comunidades agrupan distintas cantidades de observaciones. La figura 7 resalta las 20 comunidades que agrupan la mayor cantidad de observaciones, donde se destaca la comunidad con código 237 que agrupa más de 434 observaciones (referida a la observación de “formato de control de entrega de tabletas”), seguido de la comunidad con código 11 que agrupa 195 observaciones.

Figura 7: Top 20 de comunidades con mayor respuestas agrupadas



Asimismo, a diferencia de la propuesta 1, 2 y 3, el nombre de la categoría de cada comunidad se establece a partir del texto de la observación más representativa de cada comunidad. Para dicho propósito se empleó el algoritmo de *PageRank* desarrollado por Page et al. (1999).

3. EVALUACIÓN DE RESULTADOS

Para evaluar la calidad de los resultados obtenidos de cada una de las propuestas, se utilizaron las siguientes métricas de rendimiento:

- Coherencia de los temas: Indicador que mide qué tan bien un tema está respaldado por un conjunto de textos (corpus). Este indicador utiliza las estadísticas y probabilidades extraídas del corpus, enfocadas en el contexto de la palabra, para dar un puntaje de coherencia a un tema que va desde 0 a 1. Mientras más cercano a 1 mejor es la similitud de los textos dentro de los grupos.
- Tiempo de ejecución del modelo: Las técnicas de modelamiento se compara en función a su respectivo tiempo de ejecución.

Los resultados obtenidos se encuentran en la Tabla 3.

Tabla 3. Resultado de las métricas de rendimiento

Algoritmo	Categorías	Tiempo de procesamiento	Coherencia
LDA	8	10 min	0.64
STM	10	20 min	0.59
BTM	10	20 min	0.31
Louvain	813	15 min	0.72

A continuación, se describen los resultados obtenidos:

- Como se puede apreciar en la tabla 3, el algoritmo Louvain pudo identificar 813 categorías, cuyo valor es muy diferente al total de categorías detectadas por LDA, STM y BTM. Un resultado similar se encuentra en el trabajo de Chen y Beaver (2022), quienes señalan que los textos cuentan con categorías o temáticas jerarquizadas y el algoritmo Louvain detecta, de forma más fina, subcategorías dentro de cada nivel jerárquico.
- Esto podría explicar los valores de coherencia de LDA, STM y BTM, los cuales tratan de agrupar los distintos tipos de observaciones en 10 o menos categorías, provocando así que cada una de las categorías estén agrupando observaciones con poca similitud en sus contenidos, generando valores de coherencia inferiores al de Louvain.
- Para el caso del algoritmo Louvain, como la detección es más granular, los textos que en las agrupaciones identificadas son más similares, como lo evidencia su valor de 0.72 de coherencia.
- Por último, se puede evidenciar que todos los algoritmos empleados tienen un tiempo de ejecución menor o igual a 20 minutos, el cual es una mejora significativa si se compara con el tiempo que tomaría una categorización manual.

4. CONCLUSIONES

- Se aplicó la metodología de LDA, STM, BTM y Louvain a 12,915 respuestas, las cuales contenían textos sobre observacióones a la asignación de tabletas, y se pudo determinar que el algoritmo de Louvain cuenta con el mejor valor de "Coherencia" (0.72) y permitió identificar 812 categorías. Con relación a las 68,573 respuestas restantes, no fue necesario aplicar algoritmos para identificar categorías ya que estas respuestas no contenían textos de observaciones de asignación de tabletas.
- Ningún algoritmo coincidió en encontrar las 20 categorías que se determinaron manualmente, principalmente Louvain que detecto 812 categorías. Este podría deberse a que los textos cuentan con temas jerárquicos y Louvain está detectando más subcategorías en las distintas jerarquías (Chen & Beaver, 2022)
- Como se evidencia en los resultados obtenidos, la categorización automática mediante la técnica de Machine Learning no es suficiente para realizar una correcta categorización de las respuestas abiertas. Esto implica que se requiera realizar un trabajo manual posterior para obtener una correcta categorización de las respuestas abiertas. A este mismo razonamiento llegaron Schonlau y Couper (2016), quienes realizaron un estudio similar.
- En ese sentido, la metodología descrita en este documento representa una línea de partida para el proceso de categorización de respuestas de preguntas abiertas. Es decir, en lugar de analizar 12,915 repuestas para agruparlas manualmente, solo se analizarían 812 respuestas si se emplean esta metodología utilizando el algoritmo Louvain.
- Adoptar esta metodología permitirá mejorar la eficiencia del proceso de etiquetado manual de las respuestas a preguntas abiertas utilizadas en los instrumentos de recojo de información. En ese sentido, se recomienda su uso para poder facilitar el proceso de categorización de datos en formato de texto que se encuentran en las respuestas de preguntas abiertas.
- Específicamente, como línea de mejora se podría implementar una técnica para agrupar las categorías identificadas por el algoritmo de Louvain.
- Cabe indicar que la presente metodología se encuentra en un proceso de mejora continua con el objetivo reducir la tasa de error al momento de agrupar textos similares.

.

BIBLIOGRAFÍA

Buenaño-Fernández, D., Gonzalez, M., Gil, D., & Luján-Mora, S. (2020). *Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach.* IEEE.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 993-1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment, 2008(10), P10008.*

Bradburn, N., Sudman, S., & Wansink, B. (2004). *Asking Questions: The Definitive Guide to Questionnaire Design -- For Market Research, Political Polls, and Social and Health Questionnaires, 2nd, Revised Edition.* Hoboken: John Wiley & Sons.

Chen, X., & Beaver, I. (2022). *An Adaptive Deep Clustering Pipeline to Inform Text Labeling at Scale.* Baltimore, Maryland, USA: Proceedings of the 39 th International Conference on Machine Learning.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow.* O'REILLY.

Hotho, A., Nurnberger, A., & Paaß, G. (2005). *A Brief Survey of Text Mining.* UFPE.

MINEDU. (2022). *Módulo Materiales Educativos - Tablets.* SIAGIE: https://siagie.minedu.gob.pe/archivos/Instructivo_ME_SIAGIE_Tablets.pdf

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web.* Stanford InfoLab.

Roberts, M., Stewart, B., Tingley, D., Lucas, C., Leder-Luis, J., Kushner Gadarian, S., . . . Rand, D. (2014). *Structural Topic Models for Open-Ended Survey Responses.* Wiley.

Robinson, I., Webber, J., & Eifrem, E. (2015). *Graph databases: new opportunities for connected data.* O'Reilly Media, Inc.

Schonlau, M., & Couper, M. (2016). Semi-automated categorization of open-ended questions. *Survey Research Methods*. https://doi.org/https://doi.org/10.18148/srm/2016.v10i2.6213

Vidal, L., Ares, G., & Jaeger, S. (2022). *Biterm topic modelling of responses to open-ended questions: A study with US consumers about vertical farming.* Food Quality and Preference.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A Biterm Topic Model for Short Texts. *Institute of Computing Technology*, 1445–1456.