© Revista de Matemática: Teoría y Aplicaciones 2022 **29**(2) : 261–287
cimpa – ucr issn: 1409-2433 (Print), 2215-3373 (Online)
doi: https://doi.org/10.15517/rmta.v29i2.46379

MINERÍA DE TEXTO EN LA ENCUESTA NACIONAL DE TRANSPARENCIA 2019

TEXT MINIG IN THE NATIONAL TRANSPARENCY SURVEY 2019

Felipe González-Évora* Óscar Centeno-Mora†

Received: 20/Apr/2021; Revised: 18/Sep/2021;
Accepted: 30/May/2022

---

*Universidad de Costa Rica, Escuela de Estadística, San José, Costa Rica. E-Mail: juan.gonzalezevora@ucr.ac.cr

†Universidad de Costa Rica, Escuela de Tecnologías en Salud, San José, Costa Rica. E-Mail: oscar.centenomora@ucr.ac.cr

Resumen

Codificar y analizar preguntas abiertas provenientes de encuestas de opinión suele ser laborioso. La minería de texto ofrece una alternativa para ese tipo de problemática. Se utilizaron los datos de preguntas abiertas provenientes de la Encuesta Nacional de Percepción sobre la Transparencia 2019. Se aplica la minería de texto desde un enfoque descriptivo como predictivo: este último posee un interés predominante al realizar la codificación automática de respuestas o categorías a partir del aprendizaje automático supervisado. Se emplean algoritmos de máquinas de soporte vectorial, clasificador ingenuo de Bayes, bosques aleatorios, XGBoost y vecinos más cercanos. Los resultados del análisis descriptivo permiten apreciar las descripciones, visualizaciones y relaciones en el análisis de las preguntas abiertas. El análisis predictivo reseña que los algoritmos seleccionados con mayor ocurrencia para las preguntas abiertas fueron el clasificador ingenuo de Bayes y los bosques aleatorios, mostrando precisiones de entre 48% y 76%. Se obtuvieron resultados similares en comparación con las categorías que fueron codificadas manualmente.
Se aprecian resultados satisfactorios en el análisis integral de las 12 preguntas de la encuesta.

Palabras clave: encuesta de opinión; preguntas abiertas; minería de texto; aprendizaje automático supervisado.

Abstract

Coding and analyzing open-ended questions from opinion survey is often time consuming. Text mining offers an alternative for this type of problem. Data comes from the 2019 National Survey of Perception on Transparency open-ended questions. Text mining is applied from a descriptive and predictive approach: the latter has a predominant interest in performing the automatic coding of responses or categories using supervised machine learning. Support vector machine algorithms, naive Bayes classifier, random forests, XGBoost, and closest neighbors are used. The results of the descriptive analysis improve the descriptions, visualizations and relationships in the analysis of the open-ended questions. The predictive analysis reports that the algorithms with the highest selection occurrence for the open-ended questions were the naive Bayes classifier and the random forests, showing accuracies between 48% and 76%. Similar results were obtained compared with the pre-established categories. Satisfactory results are seen in the comprehensive analysis of the 12 survey questions.

Keywords: opinion surveys; open questions; text mining; supervised machine learning.

Mathematics Subject Classification: 68T45, 68T50.

1 Introduction

La cantidad de texto contenida en repositorios, bibliotecas digitales, blogs, informes, redes de medios sociales y correos electrónicos produce la necesidad de recurrir a una forma evolutiva de tratar la data textual: métodos automáticos, tanto en el procesamiento como en el análisis de esta información. Se han desarrollado técnicas de minería de texto que emplean un conjunto de algoritmos para convertir información textual en datos estructurados a fin de poder utilizar métodos analíticos en el procesamiento de la información (Maheswari y Sathiaseelan, [10]). La minería de texto busca aplicar técnicas de análisis de información para brindar significado a los datos no estructurados (Gurusamy y Kannan, [9]).

El presente trabajo utiliza técnicas de minería de texto aplicadas a preguntas abiertas incluidas en la Encuesta Nacional de Percepción sobre la Transparencia 2019 (ENPT - 2019). El objetivo de esta encuesta es indagar la percepción del grado de transparencia en la gestión pública. Por medio de la ENPT–2019 se intentan conocer las principales dificultades de acceso a la información y participación ciudadana en los temas del manejo de fondos públicos (Contraloría General de la República, [3]). Además, se buscó determinar la percepción de los ciudadanos y funcionarios públicos respecto del estado del acceso de la información sobre la Hacienda Pública y su relación con la rendición de cuentas y la participación ciudadana sobre su gestión.

El tratamiento que se le da a las preguntas abiertas realizadas en la ENPT–2019 es subjetivo, laborioso y costoso. Las aplicaciones realizadas en este trabajo muestran una alternativa de análisis distinto para preguntas abiertas en el análisis de texto. Las técnicas utilizadas pretenden mostrar la exploración y la visualización de las respuestas textuales, así como la aplicación de algoritmos que permitan codificar las preguntas abiertas de manera automática.

2 Minería de texto

La minería de datos podría presentarse como la extracción no trivial de información implícita, previamente desconocida y potencialmente útil que se encuentra en grandes bases de datos. La búsqueda de patrones se lleva a cabo utilizando métodos estadísticos, matemáticos y algorítmicos (Zaïane, [19]). La minería de datos descubre el conocimiento de datos estructurados, mientras que la minería de textos descubre y extrae el conocimiento de datos que no están estructurados, provenientes del lenguaje natural (Ananiadou, Keek y Tsujii, [2]). El análisis de minería de texto requiere una mezcla entre los saberes y técnicas de la Lingüística y la Estadística. La minería de datos representa la fusión de una serie de otras disciplinas, especialmente la Estadística y el Aprendizaje automático. Muchos de los algoritmos de la minería de datos se basan en estadística y métodos de probabilidad (Tufféry, [17]).

Los datos de texto pueden analizarse en diferentes niveles de representación: se pueden tratar como una bolsa de palabras o como una cadena de palabras. En la mayoría de las aplicaciones sería deseable representar la información del texto de forma semántica, con el propósito de realizar un análisis y una minería de datos que contenga mayor significado. Los métodos modernos de procesamiento del lenguaje natural aún no son lo suficientemente robustos para funcionar adecuadamente en dominios textuales sin el uso de restricciones para la generación de una semántica precisa del texto. La mayoría de los enfoques de minería de texto en la actualidad todavía dependen de la representación más superficial basada en palabras, especialmente el enfoque de bolsa de palabras (Solka, [16]).

La minería de texto ha resultado de mucha utilidad en una gran cantidad de aplicaciones; sin embargo, se suele ignorar que el texto es un formalismo de representación con una capacidad expresiva mucho mayor que las estructuras de datos. Al reducir el contenido del texto a una forma intermedia, se pierde una gran cantidad de información valiosa para la obtención de nuevo conocimiento. El texto contiene conocimiento, expresado mediante lenguaje natural, mucho más rico que una estructura de datos; por ejemplo, los textos incluyen de manera habitual expresiones condicionales, disyunciones, negaciones, entre otros muchos recursos expresivos (Consuelo, [8]).

3 Metodología

A continuación, se describen tanto los materiales y métodos utilizados tanto en el procesamiento de los datos textuales como en el análisis de las 12 preguntas abiertas de la ENPT–2019.

3.1 Material

La población de estudio está conformada por los ciudadanos mayores de 18 años residentes en Costa Rica, y por los funcionarios públicos activos que laboran para instituciones del sector público. La recolección de los datos se llevó a cabo del lunes 11 al viernes 22 de febrero del año 2019. Los datos se obtuvieron en las instalaciones de la Contraloría General de la República (CGR), utilizando diversas herramientas de la plataforma de Google, como el Site, formularios de Google y hojas de cálculo de Google. Estas herramientas permitieron crear cuestionarios en línea para ser aplicados; y estos, a su vez, posibilitaron registrar los resultados en un archivo de datos.

Los datos fueron recolectados mediante una encuesta telefónica. Para la selección de la muestra, se contó con un marco muestral en la consulta a la ciudadanía (números telefónicos celulares)[1] y con un marco muestral en la consulta a los funcionarios públicos[2]. Se utilizó un muestreo simple al azar para seleccionar a los ciudadanos y funcionarios públicos. La estimación del tamaño de la muestra en la ciudadanía se calculó con un nivel de confianza del 95% y con un margen error de 3 puntos porcentuales, lo que determinó un tamaño de muestra mínimo de 1.068 personas por entrevistar. En cuanto a la estimación del tamaño de la muestra de los funcionarios públicos, esta se calculó con un nivel de confianza del 95%, y con un margen de error de 4 puntos porcentuales, lo que determinó que se debía consultar a un mínimo de 600 funcionarios.

Se diseñaron tres cuestionarios para la consulta a la ciudadanía y un cuestionario para la consulta a los funcionarios públicos. Cada cuestionario se entendió como un módulo o un componente referido al tema de transparencia. En estos módulos se preguntó acerca de la percepción del acceso a los sitios web de las instituciones públicas, el nivel de participación, y la evaluación de las municipalidades y de las instituciones públicas. Dado que el presente trabajo encuentra su fundamento en el análisis de las preguntas abiertas, estas se presentan en la Tabla 1. Las preguntas se agrupan según el cuestionario en el que aparecen: Los cuestionarios fueron elaborados internamente en la CGR, mediante el trabajo conjunto del personal de la División de Fiscalización Operativa y Evaluativa, el Despacho Contralor y la División de Contratación Administrativa. El programa de análisis de datos empleado fue R Studio. Se crearon un total de 37 funciones con el fin de preparar los datos, procesarlos y visualizarlos. Para esto fue necesario utilizar 42 librerías.

3.2 Métodos

El análisis de la información de la presente encuesta posee tanto una vertiente exploratoria como predictiva en las posibles respuestas de las preguntas abiertas.

[1] A partir de un marco muestral de números provenientes de los registros del Instituto Costarricense de Electricidad (ICE).

[2] A partir de la base de datos SICERE, administrada por la Caja Costarricense del Seguro Social (CCSS).

3.2.1 Análisis exploratorio

El análisis exploratorio de los datos textuales utiliza técnicas de correlaciones, nubes de palabras, análisis de redes y cluster jerárquico.

La nube de palabras se puede generar particionando el texto en una sola palabra, en dos o en más de dos. La tokenización se puede realizar mediante N-gramas. Esta es una subsecuencia de n elementos de una secuencia de texto dada (Milios et al., [12]). Si $N = 2$, las subsecuencias se denominan bigramas; si $N = 3$, trigramas; si $N \geq 4$, N-gramas. Estos permiten estudiar los textos con un mejor entendimiento, debido a que es un método en el que se pueden apreciar mejor las respuestas, según su estructura semántica.

El análisis de redes, de forma descriptiva, consiste en determinar el grado de asociación entre los elementos, acá las palabras. Para esto, se calcula el coeficiente de correlación de Mathews, también conocido como el coeficiente phi. Este devuelve un valor entre 0 y 1. Una correlación de 1 indica que las dos palabras aparecen juntas en todas las respuestas, mientras que un valor de 0 significa que las palabras nunca aparecen en la misma respuesta. El coeficiente Φ para la palabra X y la palabra Y está dado por la siguiente ecuación:

$$\Phi(X, Y) = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_1 n_0 n_{0*} n_{1*}}} \tag{1}$$

en donde: n_{11}, el número de respuestas donde aparecen tanto la palabra X como la palabra Y, n_{10} y n_{01}, los casos en que una palabra aparece sin la otra, n_0, el número donde no aparece la palabra X, n_1, el número donde aparece X, n_{0*}, el número donde no aparece la palabra Y, y n_{1*}, el número donde aparece Y (Silge y Robinson, [15]).

Utilizando el mismo esquema del análisis exploratorio, se analizan las relaciones jerárquicas entre las palabras mediante un análisis de cluster jerárquico. En este análisis, las palabras son representadas por medio de un dendrograma. El método utilizado para unir los grupos fue el de Ward. Este método indica la distancia entre dos grupos, A y B, y está determinado por la siguiente ecuación (Murtagh y Legendre, [13]):

$$\Delta(A, B) = \frac{n_A n_B}{n_A + n_B} \|m_A - m_B\|^2, \tag{2}$$

donde n_A (resp. n_B) es el número de elementos de A (resp. de B), y m_A (resp. m_B) es el centroide de A (resp. de B). Para finalizar, otra forma de realizar el análisis de conglomerados es mediante un árbol filogenético. Un árbol filogenético es un esquema arborescente que muestra las relaciones entre varias entidades que se cree que poseen características en común. Esta estructura de presentar la información proviene de la biología; sin embargo, se puede aplicar para presentar datos de texto (Vani, Appa, Sridhar, Chakravarthy, Nageshwararo y Rao, [6]).

3.2.2 Análisis predictivo

Algoritmos de predicción Los modelos que se comparan con el propósito de obtener el mejor clasificador para cada pregunta son el clasificador ingenuo de Bayes, bosques aleatorios, XGBoost, máquinas de soporte vectorial y vecinos más cercanos.

El clasificador ingenuo de Bayes se construye utilizando un conjunto de datos de entrenamiento, esto con el fin de estimar la probabilidad de cada clase dados los valores de las palabras de los documentos (Allahyari et al., [1]). Para esto se usa el Teorema de Bayes, el cual estima las probabilidades:

$$p(c_j|d) = \frac{P(c_j)P(d|c_j)}{P(d)}, \tag{3}$$

donde c_j clases y d es un documento. El denominador de la ecuación (3) indica la probabilidad de obtener el documento, y no distingue entre las clases. Por esta razón, el denominador va a ser igual para cada clase y no necesita ser maximizado; en consecuencia, se puede eliminar de la ecuación. Este método asume que las palabras son condicionalmente independientes, dada la clase. Esto simplifica los cálculos. La ecuación del clasificador de Bayes es dada por:

$$P(c_j|d) = \text{argmax}_{c_j \in C} P(c_j) \prod_{i=1}^{M} P(d_i|c_j). \tag{4}$$

A pesar de que el supuesto de independencia condicional es generalmente falso para la aparición de una palabra en documentos, el clasificador ingenuo de Bayes es efectivo (Araujon, [3]).

Los bosques aleatorios utilizan la agregación de Bootstrap, en los que se construyen una gran cantidad de árboles y luego se promedian sus resultados. Si se tienen K árboles, la clase estimada estaría dada por:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{5}$$

donde x_i es un documento, f_k son los árboles de clasificación individuales y F es la función que contiene todos los árboles. El hiperparámetro por calibrar en este algoritmo es el número de árboles de clasificación por entrenar.

El método de XGBoost (*Extreme Gradient Boosting*) está conformado por árboles de clasificación, al igual que los bosques aleatorios; sin embargo, funciona de manera distinta. Este modelo, en vez de implementar la agregación de Bootstrap, utiliza el método de potenciación que consiste en combinar varios árboles de clasificación débiles para producir un clasificador robusto (Hastie et al., [7]). Las predicciones de todos los clasificadores $f_k(x)$ se combinan según la cantidad mayoritaria de votos ponderados para producir la predicción final:

$$f(x) = \sum_{k=1}^{K} a_k f_k(x) \tag{6}$$

donde $a_1, a_2, \ldots, a_K$ se calculan mediante el algoritmo de potenciación, el cual se encarga de ponderar la contribución de cada $f_k(x)$. Su efecto es incrementar la influencia de los clasificadores con una mayor precisión (Hastie et al., [14]). Dado los árboles de clasificación f_k, se define la función objetivo por optimizar para un conjunto de parámetros θ como (Mateo, [11]):

$$\mathrm{Obj}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{7}$$

donde $l(y_i, \hat{y}_i)$ representa la función de pérdida que compara el valor verdadero y_i con el predicho $\hat{y}_i$; en este caso se utilizó el error de clasificación. El segundo término $\Omega(f_k)$ es el término de regularización que penaliza la complejidad del modelo utilizado para evitar sobreajuste[3]. En XGBoost la fórmula de regularización está dada por:

$$\Omega(f) = \Gamma^T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2, \tag{8}$$

donde T es el número de nodos, mientras que el segundo término representa la regulación $L2$ para los puntajes w_j en cada nodo, Γ representa la penalización definida por el usuario con la finalidad de podar los árboles, mientras que λ es el término de regularización $L2$ sobre los puntajes. Los k vecinos más cercanos clasifica cada observación según la clase mayoritaria presente entre los vecinos más cercanos que se encuentran en el conjunto de entrenamiento. Esto se realiza mediante una función de distancia o similitud. La calidad de la clasificación del método depende de la manera en que se calculan las distancias entre las

[3]El sobreajuste es el efecto de sobreentrenar un algoritmo de aprendizaje con unos ciertos datos para los que se conoce el resultado deseado.

diferentes observaciones (Nguyen, Rivero y Morell, [14]). La distancia más común, y la que se utilizó en este trabajo, para entrenar este algoritmo es la Euclidiana (Hastie et al., [7]).

En este método se debe elegir el valor de k que optimice el error de clasificación. El valor máximo asignado que puede tomar el valor de k es de $\sqrt{m}$. El algoritmo toma el valor de k que minimice el error de clasificación.

Finalmente, las máquinas de soporte vectorial se fundamentan en modalidades de hiperplanos para la separación de las observaciones, la máquina de soporte vectorial para el problema multiclase debe de optimizar la siguiente función para encontrar el hiperplano con margen máximo (Xu et al., [18]):

$$\min_{w_m b_m \xi_m} \frac{1}{2}(w_m^2)^T + C\sum_{i=1}^{n} \xi_{m_i} \tag{9}$$

donde w es el vector de pesos que representa el hiperplano de separación entre las dos clases. El peso asociado a cada variable predictora proporciona información sobre la relevancia para la discriminación entre las clases. C es un hiperparámetro de regularización, el cual puede ayudar a reducir el sobreajuste y, por ende, el error de clasificación; ξ_i es una variable que funciona para minimizar el error por medio de la distancia entre la observación y el hiperplano, ξ_i es igual a 0 para los casos que fueron separados correctamente por un margen suficiente, ξ_i se encuentra entre 0 y 1 en los casos que fueron separados correctamente, pero por un margen menor que el deseado por el modelo; mientras tanto, ξ_i es mayor a 1 para los casos que fueron clasificados incorrectamente. Finalmente, m representa cada par de modelos binarios (Ben-Hur y Weston [4]). La función anterior se encuentra sujeta a:

$$\begin{array}{rcl} w_m^T x_i + b_m & \geq & 1 - \xi_i, \text{ cuando } y_i = 1 \\ w_m^T x_i + b_m & \leq & 1 - \xi_i, \text{ cuando } y_i = -1 \\ \xi_{m_i} & \geq & 0, \end{array}$$

donde x_i representan el vector que contiene el conjunto de n observaciones en el conjunto de datos de entrenamiento, y_i es la variable binaria por predecir, que puede tomar valores de 1 y -1, m representa cada uno de los modelos binarios de máquinas de soporte vectorial. Por último, b es el intercepto o sesgo del hiperplano.

Validación En la validación de los métodos de predicción se utiliza una validación cruzada y medidas de ajuste para datos categóricos para así determinar la precisión utilizadas en la determinación del rendimiento de los modelos.

La validación cruzada tomó K grupos, en donde una parte de los datos es destinada para entrenar el modelo y otra parte para validarlo, dividiendo el conjunto de datos en K grupos del mismo tamaño, realizando el proceso K veces. Este método, al utilizar datos nuevos o de prueba, evita obtener resultados que podrían ser engañosos debido al sobreajuste de los modelos (Hastie et al., [7]):

$$\mathrm{CV}_K = \frac{1}{K}\sum_{k=1}^{K} \mathrm{error}_k. \tag{10}$$

El presente análisis utilizó un $K = 10$. El procedimiento de validación cruzada puede realizarse solo una vez para medir el error obtenido en la clasificación, o puede efectuarse repetidas veces para obtener una mayor certeza y confianza del valor del error. El procedimiento de validación cruzada se realizó un total de cinco veces.

Se utilizó la medida de ajuste global con la finalidad de seleccionar el mejor modelo predictivo. Con el modelo elegido, se calculó la precisión para cada una de las categorías j. Con el fin de evitar que las métricas mencionadas se vieran afectadas por el sobreajuste de los modelos, estas se estimaron a partir de la validación cruzada y sus repeticiones. La precisión del modelo se encuentra dada por la ecuación:

$$\text{Precisión} = \frac{\text{Clasificados correctamente}}{\text{Clasificados correctamente} + \text{Clasificados incorrectamente}}. \tag{11}$$

La medida utilizada para seleccionar el mejor modelo para cada pregunta es la precisión estimada, a partir de la validación cruzada con diez grupos (K) y de las cinco repeticiones (i). Esta está es dada por la ecuación:

$$\text{Precisión} = \frac{\sum_{i=1}^{5} \frac{\sum_{k=1}^{10} \text{Precisión}_{ik}}{10}}{5}. \tag{12}$$

Por otra parte, se estimó la matriz de confusión a partir de las predicciones obtenidas en la validación cruzada y las repeticiones. En cada validación cruzada se obtuvo una predicción para cada respuesta. Como el proceso se repite cinco veces, la clase predicha se calculó con la moda de las clases obtenidas en cada repetición. Además, con el fin de comparar las clases predichas con las que fueron codificadas manualmente, se estimó la distribución porcentual y las precisiones para las clases predichas, por medio de los resultados obtenidos en la matriz de confusión. La precisión para cada clase se calculó a partir de las categorías predichas. Estas se obtuvieron en la matriz de confusión mediante el cociente de los casos clasificados correctamente en la clase j entre el total de los casos que fueron clasificados en la clase j por el modelo:

$$\text{Precisión}(j) = \frac{\text{Clasificados correctamente}(j)}{\text{Predichos}(j)}.$$

Asimismo, por medio de un gráfico de dispersión, se estudió si existe relación entre la precisión de los modelos seccionados, el tamaño de muestra y el número de categorías de las preguntas.

4 Resultados

Integralmente se analizaron las respuestas de las preguntas abiertas de la ENTP–2019. El análisis de las doce preguntas genera una gran cantidad de resultados; por lo tanto, en esta sección se muestran, para los análisis exploratorios y predictivos, únicamente los resultados para la pregunta que hace referencia a los impedimentos que han surgido a la hora de participar en asuntos del sector público, la cual se incluye en el cuestionario de participación ciudadana.

4.1 Análisis exploratorio

El Cuadro 1 indica el tamaño de muestra inicial, la cantidad de muestra efectiva y el porcentaje que no sabe o no responde (NS/NR) para cada cuestionario y pregunta. En cada cuestionario, las personas que no respondieron la pregunta sobre barreras o impedimentos, no se les indagó sobre soluciones o qué harían desde su propio ámbito. Se aprecia que el porcentaje de observaciones denominadas como no sabe o no responde, varía entre 5% y 18%. El primer análisis del texto muestra las palabras que más se repiten. Para ello, se estimó la frecuencia de cada una de las palabras para la pregunta que se refiere a los impedimentos a la hora de participar en asuntos del sector público. Los resultados se muestran en forma de nube de palabras (Figura 1). La frecuencia de los bigramas que más aparecen en las respuestas se muestra en la Figura 2. Los que más se repiten son: ‘falta de información’ (72), ‘falta de tiempo’ (44), ‘falta de interés’ (31) y ‘falta de conocimiento’ (22). Asimismo, entre los bigramas más frecuentes se encuentran: ‘tomar en cuenta’, ‘tiempo horario’ y ‘falta de comunicación’, entre otros.

Otra alternativa para visualizar los bigramas es mediante un análisis de redes, en donde se puede apreciar la centralidad que presenta cada nodo o palabra por medio del indicador de grado total. Mediante este procedimiento, se logra evidenciar que la palabra 'tiempo' es la que presenta una mayor centralidad en la red; a esta le siguen las palabras: 'faltar', 'información', 'participar' y 'horario' (Figura 3). De forma complementaria al análisis de redes en palabras, se analizaron las correlaciones entre estas por medio del coeficiente phi. En la red, el gradiente de color y el grosor de la línea de conexión representan la magnitud de la correlación. En la Figura 4, se aprecia que dos palabras que se encuentran altamente correlacionadas son 'adulto' y 'mayor', al igual que 'sector' y 'público'. Además, se logra observar la conexión por medio de la magnitud de las correlaciones de otras palabras como: 'toman' y 'cuenta', 'ninguna' y 'limitación', y 'ser' y 'escuchado'.

Cuadro 1: Tamaño de muestra, muestra efectiva y porcentaje de no sabe/no responde para cada pregunta.

Pregunta	Muestra inicial[1]	Muestra efectiva[2]	Porcentaje de NS/NR
Acceso a la información			
Impedimentos	1105	995	9.95
Soluciones	825	768	5.17
¿Qué haría la persona?	825	622	18.38
Rendición de cuentas			
Impedimentos	1095	960	12.33
Soluciones	764	664	9.16
¿Qué haría la persona?	763	599	14.99
Participación ciudadana			
Impedimentos	1090	919	15.69
Soluciones	871	713	14.50
¿Qué haría la persona?	882	688	17.80
Funcionarios públicos			
Impedimentos	607	557	8.24
¿Qué haría si fuera jerarca?	529	458	11.70
¿Qué haría la persona?	520	437	13.71

Nota[1]: la muestra inicial representa la cantidad de personas a las que se les realizó la respectiva pregunta; eliminando los no aplica.
Nota[2]: la muestra efectiva resulta de eliminar los NS/NR.

Figura 1: Nube de palabras y de bigrama que hace referencia a los impedimentos para participar en asuntos del sector público.

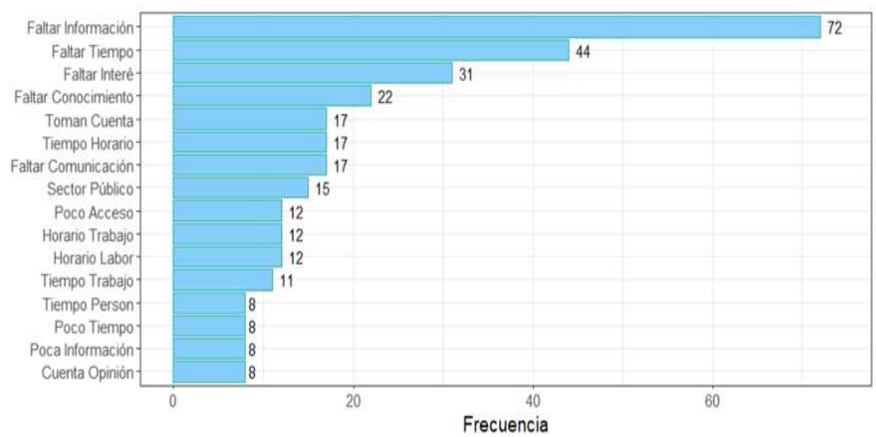

Figura 2: Frecuencia de bigramas que más ocurren en referencia a los impedimentos para participar en asuntos del sector público.

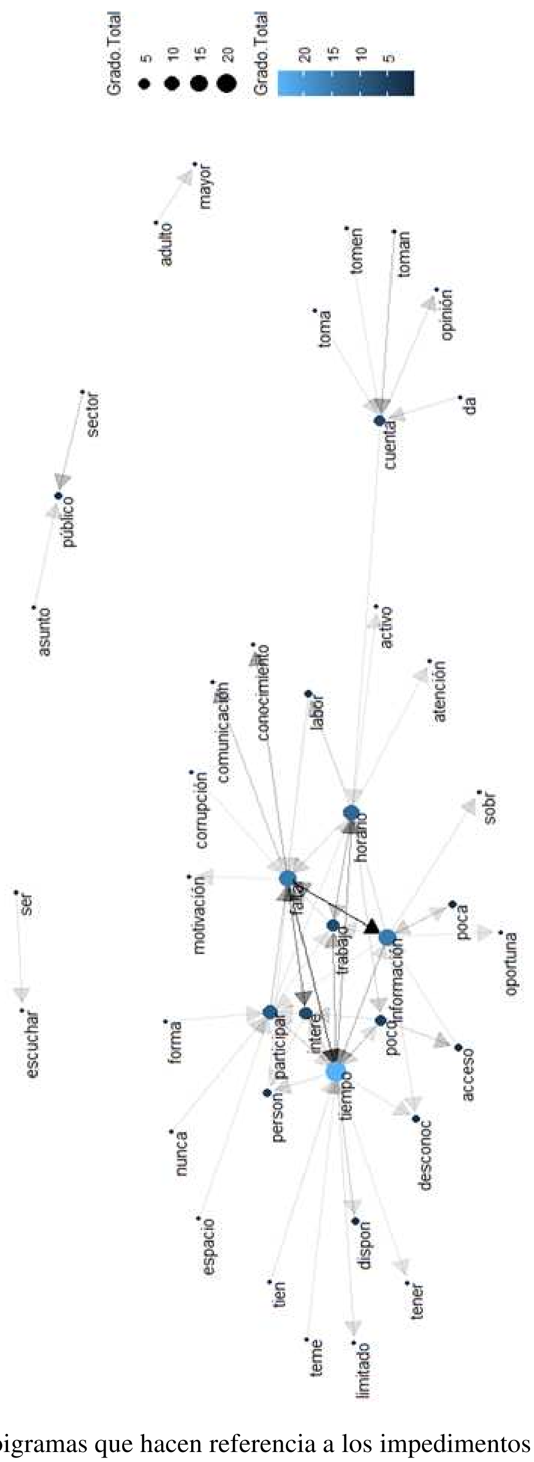

Figura 3: Red de bigramas que hacen referencia a los impedimentos para participar en asuntos del sector público.

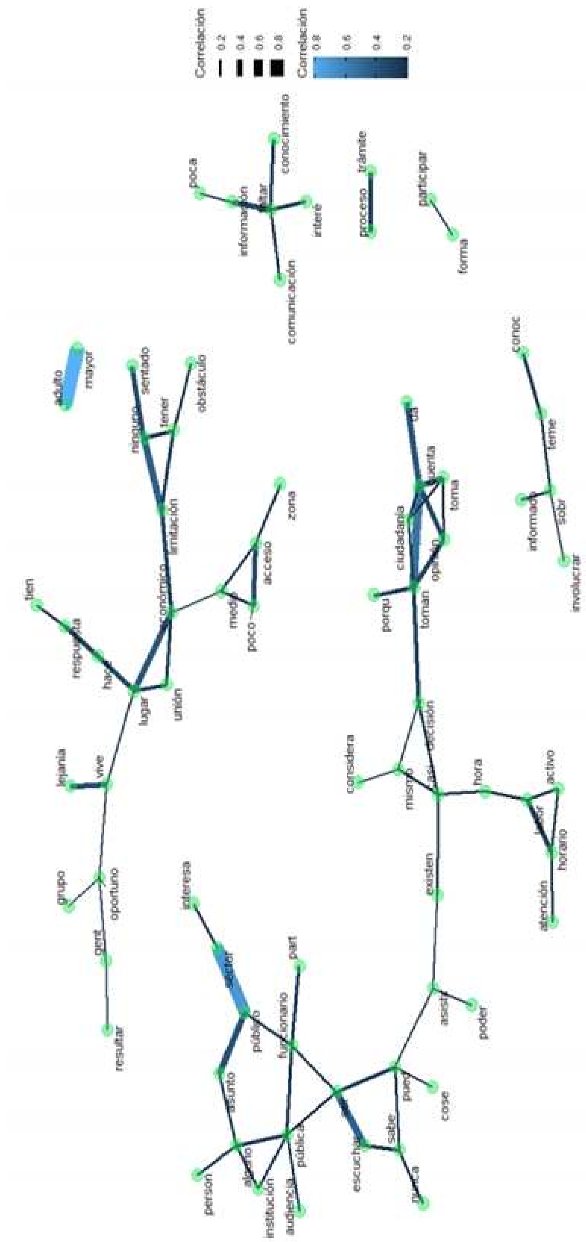

Figura 4: Red de correlaciones entre palabras que hacen referencia a los impedimentos para participar en asuntos del sector público.

Se presentan los resultados a partir del análisis de clusters jerárquico, este análisis se muestra la agrupación de las palabras en conglomerados, según sus distancias o similitudes. La Figura 5 representa el análisis de conglomerados mediante un árbol filogenético, en donde se evidencia que se correlacionan palabras como 'adulto' y 'mayor'. Luego, al agregar la palabra 'porque' en el grupo, se puede interpretar que el impedimento al que se hace referencia es 'porque se es adulto mayor'. Adicionalmente, se forman otros conglomerados: 'vive lejos', 'estudio limitado', 'poco acceso', 'funcionarios públicos', entre otros.

A grandes rasgos, el análisis de las 12 preguntas describe para las barreras el mejorar los sitios web y las redes sociales, actualizar la información, aumentar la cantidad de información que ofrecen y mejorar el acceso. Respecto a la pregunta de qué podría hacer la persona desde su propio ámbito, las personas respondieron lo siguiente: participar, buscar información, solicitar información, obtener provecho de los sitios web o no hacer nada. Cuando se les preguntó a los funcionarios públicos qué harían, si fueran jerarcas, para eliminar o solucionar las barreras, mencionaron: dar más información, mejorar la rendición de cuentas, y brindar información por medios de comunicación y páginas web. Por otra parte, al preguntarles qué podrían hacer desde su propio ámbito, señalaron: brindar la información, nada, ser transparentes, actualizar la información, ser honestos y ofrecer un mejor servicio.

4.2 Análisis predictivos

El Cuadro 2 muestra los algoritmos seleccionados para cada pregunta, esto según su precisión al realizar las validaciones cruzadas. El clasificador ingenuo de Bayes fue el que se seleccionó en más ocasiones, seguido de los bosques aleatorios. Las máquinas de soporte vectorial fueron elegidas en dos ocasiones; al considerar la pregunta sobre cómo solucionaría la barrera, incluida en el cuestionario de acceso a la información, y la pregunta realizada a los funcionarios públicos sobre qué harían desde su propio ámbito para solucionar la barrera. La pregunta en la que se consiguió una mayor precisión, a partir del uso de la validación cruzada, fue la que hace referencia a los impedimentos al participar en asuntos del sector público, pregunta incluida en el cuestionario de participación ciudadana. Dicha pregunta obtuvo un 76,3%, y fue seguida por la pregunta sobre barreras para obtener resultados de lo que hace el sector público, con un 74,4%. Esta última pregunta está incluida en el cuestionario de rendición de cuentas. La pregunta referente a qué haría si fuera jerarca para solucionar las barreras, correspondiente al cuestionario de funcionarios públicos, fue en la que se obtuvo una precisión menor a la hora de predecir las categorías (48%).

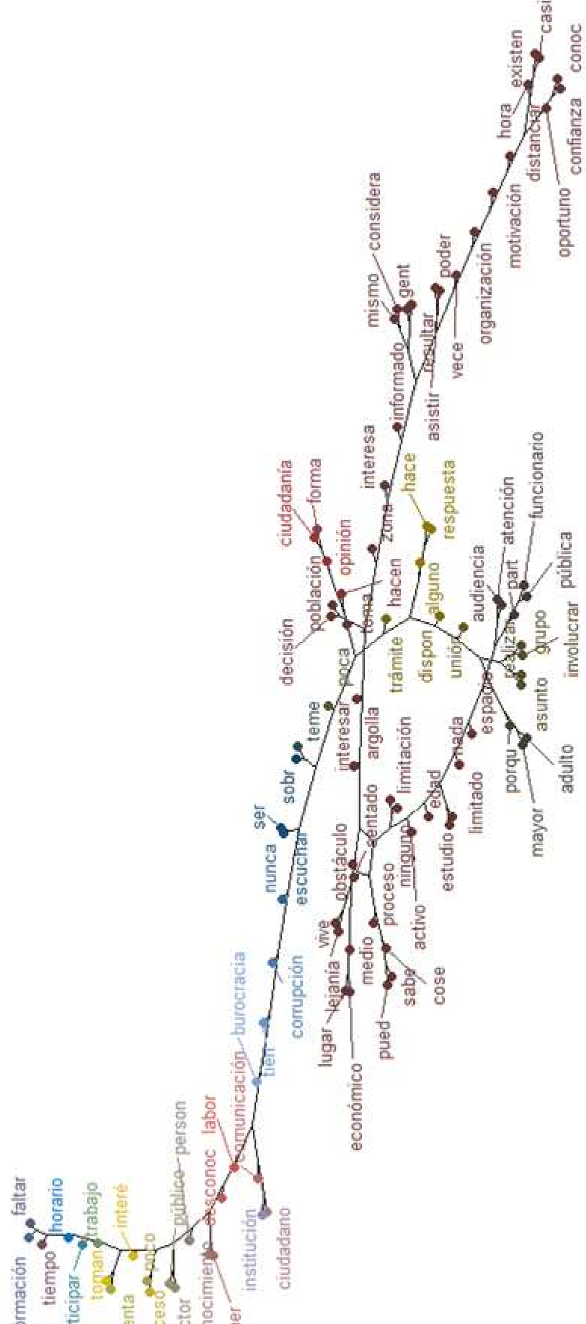

Figura 5: Dendrograma filogenético de palabras1 que hacen referencia a los impedimentos para participar en asuntos del sector público.

Cuadro 2: Modelos de predicción seleccionados para cada pregunta, número de predictores iniciales y finales y su respectiva precisión

Pregunta	Modelo elegido	Número de categorías	Total de palabras iniciales	Total de palabras finales[1]	Precisión[2]
Acceso a la información					
Impedimentos	Bayes	22	590	108	71,37
Soluciones	SVM–Lineal	35	708	100	58,56
¿Qué haría la persona?	Bosques aleatorios	36	602	160	62,25
Rendición de cuentas					
Impedimentos	Bayes	21	578	139	74,41
Soluciones	Bosques aleatorios	26	725	113	54,78
¿Qué haría la persona?	Bosques aleatorios	19	590	137	64,17
Participación ciudadana					
Impedimentos	Bosques aleatorios	21	599	241	76,34
Soluciones	Bayes	28	806	215	61,59
¿Qué haría la persona?	Bayes	24	691	196	64,58
Funcionarios públicos					
Impedimentos	Bayes	19	582	227	62,20
¿Qué haría si fuera jerarca?	Bayes	23	767	218	47,50
¿Qué haría la persona?	SVM–Sigmoideo	26	622	157	51,79

Nota[1]: las palabras finales son las que se obtienen al eliminar las que se repiten una sola vez, y con la utilización de la prueba Chi-cuadrado.
Nota[2]: la precisión se obtuvo a partir del promedio de las precisiones en la validación cruzada, considerando diez grupos y cinco repeticiones.

Se estudió la relación entre la precisión obtenida, el tamaño de muestra y el número de categorías. La Figura 6 destaca que, entre las preguntas, existe una relación lineal positiva entre el tamaño de muestra utilizado para entrenar los modelos y la precisión obtenida. Lo anterior indica que, al aumentar el tamaño de muestra, la precisión aumenta. Se observa, además, una relación lineal negativa entre el número de categorías de las preguntas y la precisión obtenida. También, se realizó un gráfico de dispersión con el propósito de apreciar la relación existente entre la distribución porcentual de las categorías obtenidas mediante la codificación manual y la distribución de las categorías predichas para cada una de las preguntas de los cuatro cuestionarios. El gráfico resultante muestra una relación lineal positiva entre ambas distribuciones, inclusive en la pregunta del cuestionario de funcionarios públicos que hace referencia a qué harían estos si fueran jerarcas, en la que se obtuvo una precisión del 48% (Figura 7). El que se hayan encontrado correlaciones relativamente fuertes entre ambas distribuciones no indica que se obtienen conclusiones sustantivas similares entre las categorías codificadas manualmente y las predichas: para determinar esto se deben analizar y comparar los valores de la distribución porcentual, como lo muestra más adelante el Cuadro 3.

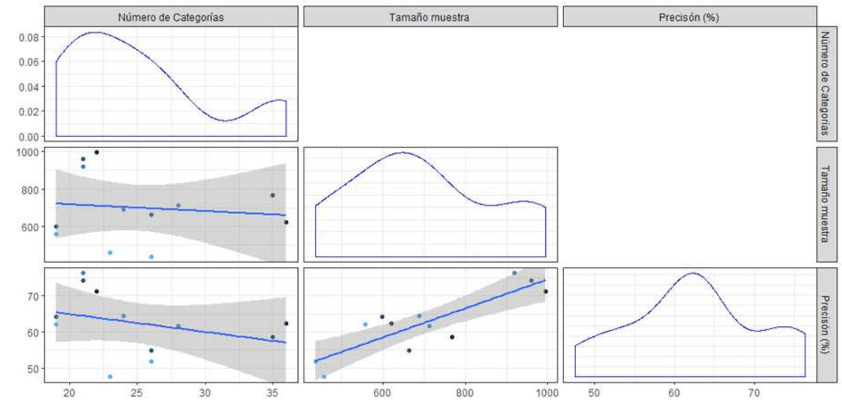

Figura 6: Relación entre la precisión obtenida, el tamaño de muestra y el número de categorías de las doce preguntas.

Se presenta cómo se realizó la selección, validación e interpretación de los resultados del modelo seleccionado; esto para la pregunta que se refiere a los impedimentos para participar en asuntos del sector público, incluida en el cuestionario de participación ciudadana. Se debe señalar que el mismo procedimiento de selección y validación se realizó para las restantes preguntas; sin embargo, a manera de ejemplo, se muestra para una única pregunta. Para elegir el mejor modelo para esta pregunta, se efectuó una validación cruzada que consideró diez grupos y cinco repeticiones. Los resultados que comparan los modelos entrenados se muestran en el Figura 8. En este, se aprecia que el modelo que tiene una mayor precisión en las cinco validaciones cruzadas son los bosques aleatorios; por ende, este algoritmo es el que se escogió para estimar las predicciones de las respuestas. En el Figura 9 se aprecia la matriz de confusión obtenida con el modelo de bosques aleatorios. En dicha matriz, se observa cómo clasificó el algoritmo las respuestas y la gradiente de color que representa la precisión de las categorías predichas. El modelo clasificó 90 respuestas en la categoría “otros”, de las cuales únicamente 29 pertenecen a dicha categoría. Las restantes 61 respuestas pertenecen a otras categorías. Por otro lado, el Cuadro 3 muestra la comparación entre la distribución porcentual de las categorías codificadas manualmente y las predichas por los bosques aleatorios. Para las categorías preestablecidas, las dos clases con mayor frecuencia son las que indican que su mayor impedimento son el ‘tiempo’ (23,7%) y la ‘falta de información’ (21,2%). El resultado obtenido fue el mismo para las clases predichas: las dos clases con mayor frecuencia son el tiempo (24,0%) y la falta de información (21,6%). Las categorías de ‘horarios’, ‘falta de interés’ y ‘falta de comunicación’ representan 7,4%, 7,3% y 6.3%, respectivamente. Por otra parte, estas mismas categorías para las clases predichas representan 7,9%, 7,1% y 3,3%, respectivamente. Estos resultados indican una similitud entre la distribución porcentual de las categorías codificadas manualmente que presentan porcentajes menores y sus respectivas categorías predichas. Las precisiones de las categorías predichas para las clases de 'edad', 'trámites' y 'corrupción' son mayores a 91%, mientras que para las categorías 'tiempo' y 'falta de información', las precisiones respectivas son 88,7% y 81,9%. Las categorías que muestran precisiones menores son las de 'funcionarios públicos', 'otro' y 'falta de espacios'.

Cuadro 3: Comparación de clases codificadas manualmente con las clases predichas1 por los bosques aleatorios para la pregunta que hace referencia a los impedimentos para obtener información de instituciones públicas.

Codificación manual		Predicho		
Categoría	Distribución porcentual (%)	Categoría	Distribución porcentual (%)	Precisión (%)
Tiempo	23,72	Tiempo	24,05	88,69
Falta de información	21,22	Falta de información	21,65	81,91
Horarios	7,40	Otro	9,79	32,22
Falta de interés	7,29	No ha participado	8,05	66,22
Falta de comunicación	6,31	Horarios	7,94	87,67
No ha participado	6,20	Falta de interés	7,07	81,54
Otro	4,90	No es tomado en cuenta	4,46	63,41
No es tomado en cuenta	4,79	Falta de comunicación	3,26	86,67
Trámites	3,70	Trámites	2,94	96,30
Corrupción	2,94	Corrupción	2,50	91,30
Lejanía a institución	2,61	No ha tenido obstáculos	1,74	68,75
No ha tenido obstáculos	1,52	Lejanía a institución	1,41	92,31
Accesibilidad	1,31	Accesibilidad	1,31	58,33
Edad	1,31	Edad	0,98	100,00
Desmotivación	0,98	Falta de espacios	0,54	40,00
Falta de espacios	0,98	Desconfianza	0,54	60,00
No genera cambios	0,87	Desmotivación	0,44	75,00
Funcionarios públicos	0,65	No genera cambios	0,44	75,00
Desconfianza	0,44	Limitaciones económicas	0,44	50,00
Limitaciones económicas	0,44	No recibe respuesta	0,44	75,00
No recibe respuesta	0,44	Funcionarios públicos	0,00	0,00

Nota[1]:la distribución porcentual y la precisión para las clases predichas se obtuvieron por medio de los datos de la matriz de confusión, a partir de los resultados de la validación cruzada y sus repeticiones.
Fuente: Contraloría General de la República. Encuesta Nacional de Percepción sobre la Transparencia 2019.

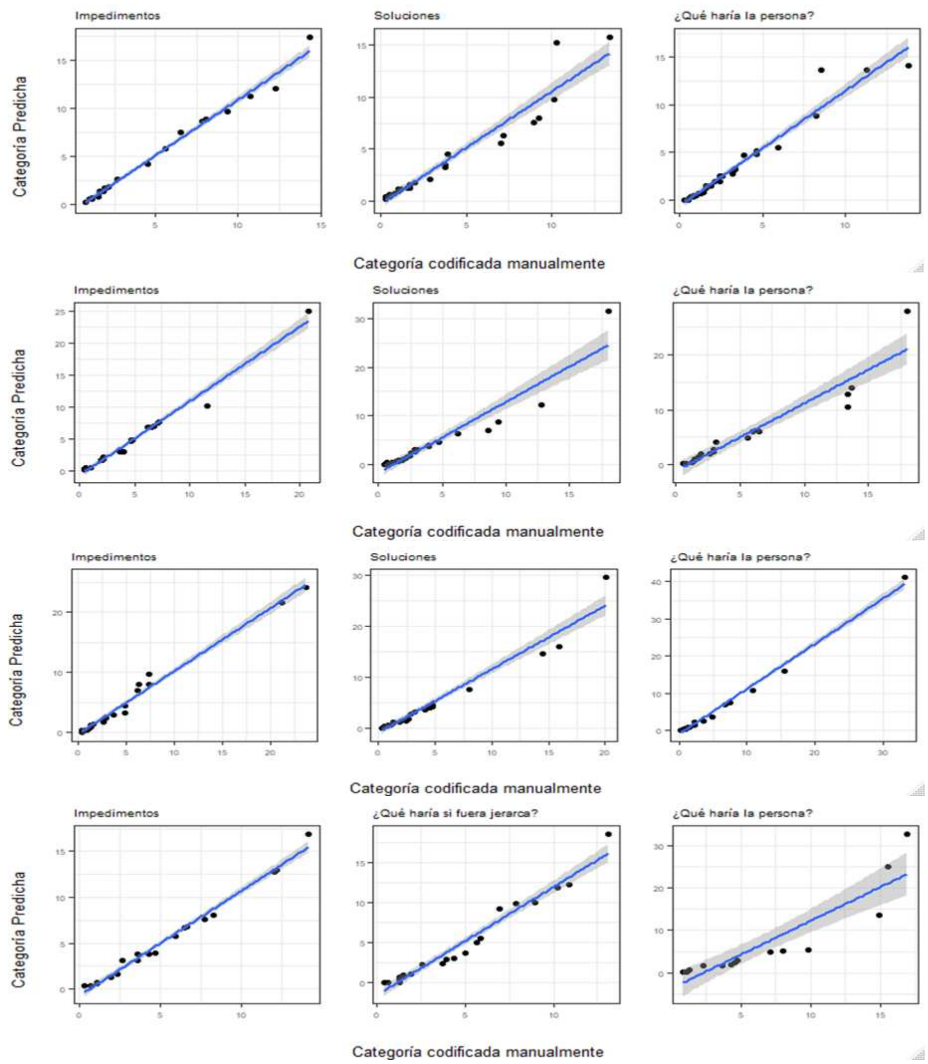

Figura 7: Relación entre distribución porcentual de las categorías codificadas manualmente y distribución de las categorías predichas.

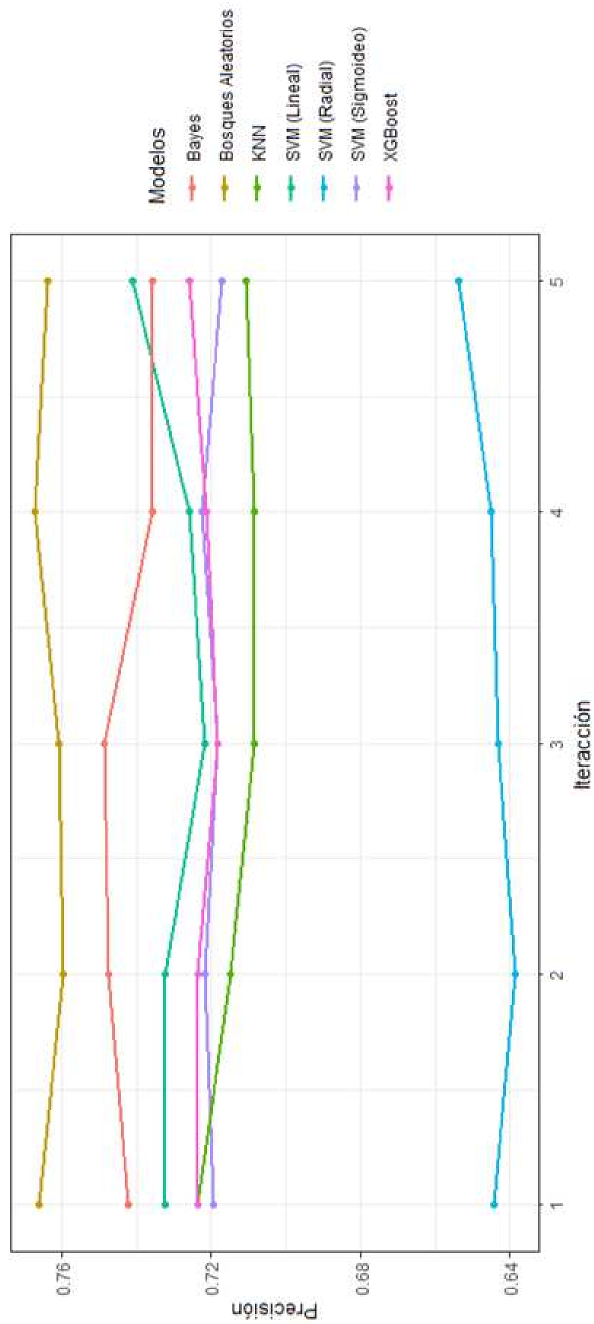

Figura 8: Repeticiones de validación cruzada para los modelos propuestos para la pregunta que hace referencia a los impedimentos para participar en asuntos del sector público.

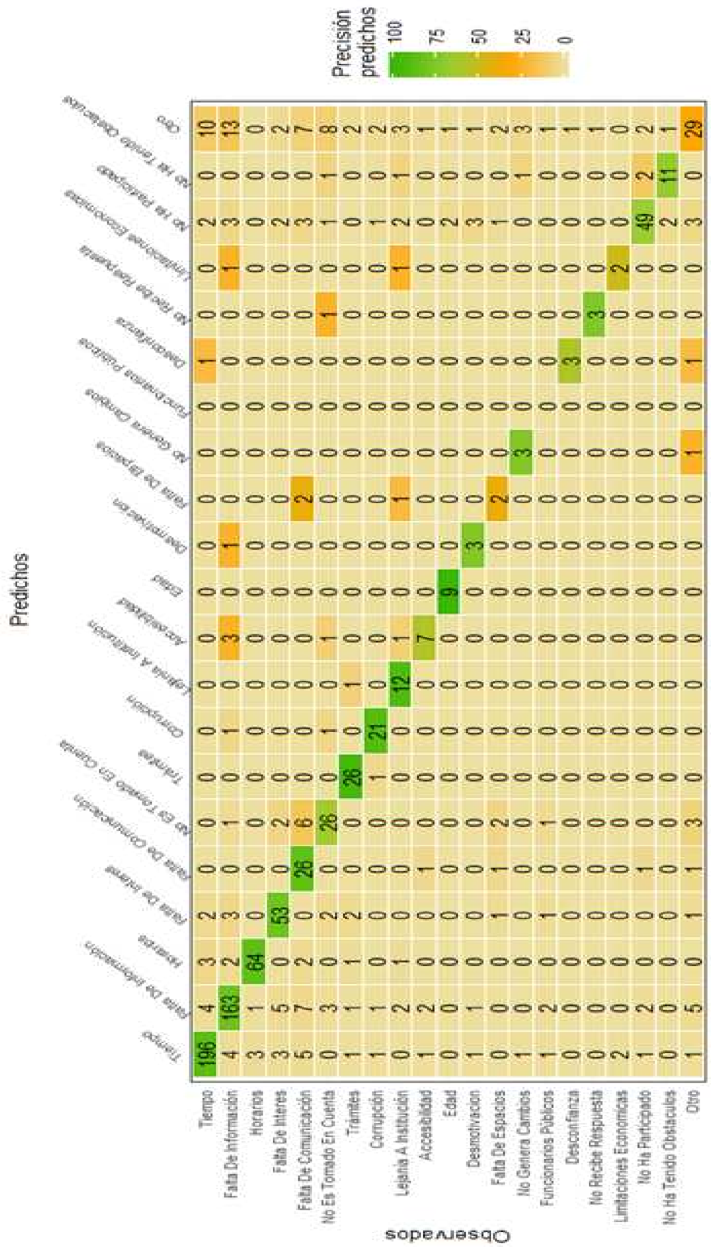

Figura 9: Matriz de confusión[1] obtenida a partir de la predicción de los bosques aleatorios para la pregunta que hace referencia a los impedimentos para participar en asuntos del sector público.

Nota[1]: los datos de la matriz de confusión fueron estimados a partir de los resultados de la validación cruzada y las repeticiones.

5 Conclusión y discusión

El presente trabajo mostró la aplicación de minería de texto en preguntas abiertas de la ENPT - 2019. La primera etapa en el análisis texto efectuado fue la limpieza y el preprocesamiento de la información. Durante el proceso de análisis, se mostró la utilización de técnicas de análisis descriptivas, como lo son las frecuencias de palabras (nubes), redes de texto, análisis de correlación, clusters y análisis de sentimientos; asimismo, se mostró la utilización de técnicas predictivas de aprendizaje automático supervisado. Se demostró que es posible realizar un análisis del texto de manera automatizada, a diferencia de lo que ocurre en la codificación manual de las preguntas abiertas, en la que el proceso es subjetivo y conlleva una gran cantidad de tiempo. También las técnicas utilizadas en este trabajo permiten visualizar e identificar rápidamente los principales temas o conceptos presentes en las respuestas y las relaciones entre las palabras de una forma explícita. Los resultados más importantes producto del análisis descriptivo destacan las principales barreras o impedimentos que las personas encuentran a la hora de obtener información y participar en asuntos del sector público son, entre otros, la gran cantidad de trámites, el tiempo, el desconocimiento y la complicación al usar las páginas web. Al preguntarles a las personas cómo se podrían eliminar esas barreras, mencionaron soluciones como las siguientes: simplificando trámites, hacer eficiente el proceso, y por medio de mejorar y actualizar los sitios web. La predicción de respuestas mostró que la tarea de codificación automática puede plantearse como un problema de categorización multi-clase por medio del aprendizaje automático supervisado. Se presentó una alternativa para seleccionar el número de palabras elegidas para entrenar los modelos, lo que disminuye considerablemente la dimensionalidad. Además, se mostró cómo seleccionar el modelo más adecuado para cada una de las preguntas mediante validación cruzada. Los algoritmos que fueron seleccionados con mayor ocurrencia para las doce preguntas fueron el clasificador ingenuo de Bayes y los bosques aleatorios. La precisión resultante para cada uno de los modelos seleccionados para cada pregunta varió entre 48% y 76%. A partir de los resultados obtenidos de los modelos predictivos, se encontró que existe una relación lineal positiva entre el tamaño de muestra y la precisión: a mayor tamaño de muestra, mayor es la precisión. Estos resultados permiten concluir que, en algunas preguntas, al incluirse tantas categorías, se debe aumentar el tamaño de muestra para incrementar las precisiones de los modelos, o se deben crear menos categorías, pues a veces no es fácil aumentar el tamaño de muestra. Por otro lado, al observar las matrices de confusión, se observó el ruido que añade la categoría “otros” en los modelos, debido a que las respuestas en esa categoría no siguen ningún patrón. Los resultados obtenidos a partir de técnicas exploratorias como la frecuencia de palabras, análisis de redes y clusters son similares a los resultados que fueron codificados manualmente. Por otra parte, las categorías predichas por los modelos elegidos para cada pregunta permiten establecer resultados similares en comparación a las categorías preestablecidas. Esto permite concluir que los resultados obtenidos con la codificación manual y por medio de la utilización de técnicas de la minería de texto, tanto exploratorias como predictivas, son semejantes.

El archivo de datos utilizado para realizar los análisis contiene tres variables de ponderación para ajustar los resultados a nivel nacional. Habría sido deseable ponderar los datos; sin embargo, esto no se hizo debido a la falta de recursos computacionales y la complejidad de las técnicas aplicadas en el documento. El presente trabajo constituye una primera aproximación del uso de técnicas provenientes de la minería de texto para analizar la información contenida en preguntas abiertas en encuestas de opinión. Esta modalidad de análisis se debe seguir investigando y aplicando al contexto de Costa Rica, debido a que produce ventajas considerables. El análisis de este tipo de preguntas tradicionalmente se ha realizado mediante el método de la codificación manual; no obstante, las capacidades computacionales con las que se cuenta en la actualidad permiten realizar un análisis más automatizado de las preguntas abiertas.

Agradecimientos y financiamiento

Agradecemos a la Constraloría Genera de la República por proporcionarnos la información, la cual fue el insumo para el análisis antes visto. Además, agradecemos a la Universidad de Costa Rica por patrocinar y difundir la presente investigación.

Referencias

[1] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E.D. Trippe, J.B. Gutiérrez, K. Kochut, *A brief survey of text mining: Classification, clustering and extraction techniques*, arXiv, 2017. Doi: https://arxiv.org/abs/1707.02919

[2] S. Ananiadou, D.B. Kell, J.i. Tsujii, *Text mining and its potential applications in systems biology*, Trends in Biotechnology **24** (2006), no. 12, 571–579. Doi: 10.1016/j.tibtech.2006.10.002

[3] N.P. Araujo, *Método semisupervisado para la clasificación automática de textos de opinión*. Masters Thesis in Computer Science, Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, México, 2009. Available from: Link

[4] A. Ben-Hur, J. Weston, *A User's Guide to Support Vector Machines*, in: O. Carugo & F. Eisenhaber (Eds) Data Mining Techniques for the Life Sciences. Methods in Molecular Biology 609, Humana Press, Springer, New York, 2009, pp. 223–239. Doi: 10.1007/978-1-60327-241-4_13

[5] Contraloría General de la República, *Memoria Anual 2018*, San José. Costa Rica, 2019. Available from: Link

[6] S.V. Guttula, A.A. Rao, G.R. Sridhar, M.S. Chakravarthy, K. Nageshwararo, P.V. Rao, *Cluster analysis and phylogenetic relationship in biomarker indentification of type 2 diabetes and nephropathy*, International Journal of Diabetes in Developing Countries **30** (2010), 52–56. Doi: 10.4103/0973-3930.60003

[7] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, 2nd Edition, Springer, New York, 2009. Doi: 10.1007/978-0-387-84858-7

[8] M.C. Justicia de la Torre, *Nuevas técnicas de minería de textos: Aplicaciones*. Doctorate Thesis in Communication Sciences and Artificial Intelligence, University of Granada, Spain, 2017. https://digibug.ugr.es/handle/10481/46975

[9] S. Kannan, V. Gurusamy, Preprocessing Techniques for Text Mining. Preprint, Madurai Kamaraj University, India, 2014. Available from: Link

[10] M. Maheswari, J.G.R. Sathiaseelan, *Text mining: Survey on techniques and applications*, International Journal of Science and Research **6** (2017), no. 6, 1660–1664. Link

[11] J.D. Mateo Vázquez, *Competición de Kaggle.com: Santander Customer Satisfaction* Master Thesis, Universidad Internacional de Andalucía, Huelva, España, 2014. Available from: Link

[12] E.E. Milios, M.M. Shafiei, S. Wang, R. Zhang, B. Tang, J. Tougas, *A Systematic Study on Document Representation and Dimensionality Reduction for Text Clustering*, Preprint, Faculty of Computer Science, Dalhouse University, 2007. Available from: Link

[13] F. Murtagh, P. Legendre, *Hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion?*, Journal of Classification **31** (2014), 274–295. Doi: https://doi.org/10.1007/s00357-014-9161-z

[14] B. Nguyen Cong, J. Rivero Pérez, C. Morell, *Aprendizaje supervisado de funciones de distancia: estado del arte* Revista Cubana de Ciencias Informáticas **9**(2015), no. 2, 14–28. Available from: Link

[15] J. Silge, D. Robinson, *Text Mining with R. A Tidy Approach*. O'Reilly, Sebastopol CA, 2019. https://www.tidytextmining.com/

[16] J.L Solka, *Text data mining: Theory and methods*, Statistics Surveys **2** (2008), 94–112. Doi: 10.1214/07-SS016

[17] S. Tufféry, *Data Mining and Statistics for Decision Making*, John Wiley & Sons, New York, 2011. Doi: 10.1002/9780470979174

[18] J. Xu, X. Liu, Z. Huo, C. Deng, F. Nie, H. Huang, *Multi-class support vector machine via maximizing multi-class margins*, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017 pp. 3154–3160. Doi: 10.24963/ijcai.2017/440

[19] O.R. Zaïane, *Introduction to Data Mining*, Chapter 1 in: Principles of Knowledge Discovery in Databases, Departament of Computer Science, University of Alberta. Canada. Available from: Link