Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

El futuro es de todos

DNP
Departamento
Nacional de Planeación

ANÁLISIS DE PROPUESTAS PARA LAS MESAS DE CONVERSACIÓN NACIONAL

Entidad
Departamento Nacional de Planeación
- Dirección de Desarrollo Digital.
- Dirección General

Sector	**Lenguaje**	**Fuente de datos**
Planeación	R, Python.	Plataforma Web de la gran conversación nacional

Presentación
En noviembre de 2019, el Gobierno Nacional abrió la Conversación Nacional, un espacio de participación para escuchar las voces de los colombianos, profundizar en propuestas que construyen, diseñar rutas de transformación y acelerar la ejecución de acciones para cerrar las brechas sociales históricas que existen en el país. Con este propósito, se establecieron 6 grandes temáticas para poner sobre la mesa: Educación, Ambiente, Crecimiento con Equidad, Juventud, Paz con Legalidad y Lucha contra la Corrupción y se dispuso de una plataforma digital para facilitar la participación de los colombianos en todos los rincones del país, que para el mes de marzo de 2020 había registrado más de 12.000 propuestas. El proyecto aquí descrito permitió agrupar y clasificar estas propuestas en categorías de interés, utilizando minería de texto. Esta clasificación permitió asignar con mayor facilidad las propuestas a las distintas direcciones técnicas del DNP para su revisión y análisis, al igual que permitió identificar las preocupaciones más representativas que expresaban los colombianos a través de la plataforma.

In November 2019, the National Government opened the National Conversation, a space for participation to listen to the voices of Colombians, deepen in proposals that build, design routes of transformation and accelerate the implementation of actions to close the historical social gaps that exist in Colombia. To this end, six major themes were established to be put on the table: Education, Environment, Growth with Equity, Youth, Peace with Legality and the Fight against Corruption. A digital platform was also set up to facilitate the participation of Colombians within all the territory, which by March 2020 had registered more than 12,000 proposals. The project described here made it possible to group and classify these proposals into categories of interest, using text mining. This classification made it easier to assign the proposals to the various technical directorates of the DNP for review and analysis, as well as to identify the most representative concerns expressed by Colombians through the platform.

Objetivo general
Facilitar el análisis y respuesta de las propuestas registradas en la plataforma web de las mesas de conversación nacional a través del uso de técnicas de minería de texto.

Objetivos específicos
1. Identificar grupos característicos de propuestas a través de la agrupación automática de los textos, a nivel de mesa y pregunta.
2. Identificar categorías específicas de interés mediante la identificación de palabras clave en las propuestas.
3. Caracterizar las propuestas mediante la visualización de sus términos más representativos con distintos niveles de agregación.

Metodología

El trabajo realizado para analizar y explorar el texto de las respuestas se dividió en varias etapas, que se presentan a continuación. Los pasos 1, 2 y 3 (preprocesamiento, cálculo de sentimiento y etiquetado por palabras clave) fueron aplicados a toda la base de datos (es decir, a todas las propuestas), mientras que los pasos 4 y 5 (exploración y agrupación) se aplicaron varias veces, unas veces filtrando la base por mesa de conversación (Juventud, Crecimiento con Equidad, etc.) y otras por pregunta específica, según los resultados fueran mejores con uno u otro enfoque. Finalmente, los pasos 6 y 7 (análisis demográfico y análisis temporal) se realizaron para toda la base y a diferentes niveles de desagregación (tema y pregunta), tomando como insumo los resultados de los pasos anteriores. Toda la metodología aquí descrita se realizó inicialmente con un corte de la base de datos que contenía 10625 propuestas, pero luego se actualizó con corte al 18 de febrero de 2020, fecha en que contenía 12015 registros.

Preprocesamiento de texto y limpieza de base

Para el preprocesamiento de texto se realizaron las siguientes acciones sobre el campo de las respuestas de los ciudadanos:

- Pasar el texto a minúsculas
- Remover signos de puntuación, números y caracteres extraños
- Remover stopwords estándar del idioma español
- Remover de la respuesta las palabras contenidas en la pregunta a la que se responde.
- Remover palabras adicionales por mesa de conversación, que pudieran generar ruido en los análisis. Se definió con el equipo asesor del DNP una lista de palabras y expresiones por eliminar para cada mesa.
- Remover palabras que en general no aportaban al análisis (por ejemplo, groserías)
- Simplificar el texto, utilizando técnicas de lematización, que ayudaron a agrupar textos similares entre sí.

Una vez se tuvo el texto preprocesado, o "limpio", lo siguiente que se realizó fue remover filas de la base de datos que pudieran entorpecer el análisis. En particular, se eliminaron:

- Filas en las que los campos de "nombre" (de quien llena el formulario) y "respuesta" estuvieran duplicadas.
- Filas en las que las respuestas tuvieran solamente 3 palabras o menos.

Cálculo de sentimiento de las propuestas ciudadanas

El siguiente paso consistió en utilizar el texto limpio para estimar el sentimiento asociado a cada propuesta. Se utilizó una red neuronal para asignar un puntaje de "sentimiento positivo" a las propuestas, con una escala de 0 (nada positivo) a 100 (totalmente positivo). Esta red neuronal fue utilizada en 2019 para el seguimiento a comentarios sobre el PND en Twitter. Se entrenó con Tweets en español etiquetados como "positivos" y "negativos" a partir de palabras clave y emoticones. Cabe mencionar que este modelo identifica relaciones entre términos, por lo que la puntuación se basa en todos los términos del Tweet y no solo en los utilizados para el etiquetado inicial.

Cabe mencionar que los resultados están sujetos a error, especialmente en algunas mesas o preguntas (por ejemplo, en la mesa de transparencia es posible que una respuesta sea calificada como negativa sin serlo, solo por usar palabras como "corrupción" o "delito"). Sin embargo, la calificación de sentimiento de las respuestas puede servir como un proxy para medir el sentir de la ciudadanía alrededor de un tema o una pregunta en particular.

Etiquetado por palabras clave

Tomando como insumo el texto limpio, se realizó un etiquetado de las propuestas a partir de una lista de palabras clave que definían subcategorías dentro de cada mesa. Por ejemplo, si en una propuesta de la mesa de "Paz con legalidad" se encontraba la expresión "líderes sociales", la propuesta se clasificaba en "Condiciones y garantías de seguridad y humanitarias para la construcción de paz", o si se encontraba el término "tala" en una propuesta de ambiente, esta se clasificaba en la categoría de "Deforestación". Este procedimiento se aplicó para las más de 12 mil propuestas utilizando 1253 términos clave brindados por el equipo asesor del DNP para etiquetar 43 categorías diferentes, previamente definidas.

Exploración de frecuencia de n-gramas

A partir del texto limpio, se identificaron los n-gramas (grupos de 1, 2 o 3 palabras consecutivas) y sus respectivas frecuencias. Esta información fue presentada en un aplicativo interactivo de dos maneras: por medio de una nube de palabras y por medio de una tabla que presentaba los 20 n-gramas más frecuentes en las propuestas, junto con sus respectivas frecuencias.

División de los textos en temas

A partir del texto limpio, se segmentaron las respuestas en diferentes grupos, intentando que los textos fueran muy similares dentro de cada grupo y muy diferentes de los textos de los otros grupos. Para hacer esta segmentación primero se generó una representación vectorial de cada respuesta con el modelo de Bolsa de Palabras con y sin TF-IDF (es decir, se representó cada texto de forma numérica), y luego se consideraron distintas alternativas de agrupamiento:

- **Latent Dirichlet Allocation (LDA) para modelamiento de temas:** En este caso, se integró en el aplicativo la posibilidad de dividir las respuestas en 2, 3, 4 o hasta 30 temas mediante un enfoque probabilístico (LDA), presentando para cada tema las palabras más representativas.
- **Método de agrupamiento de K-medias:** Este algoritmo también permitió dividir los textos en 2, 3, 4 y hasta 30 grupos. En este caso se utilizaron proyecciones entre los espacios vectoriales de los grupos para conocer las palabras exclusivas y más relevantes que caracterizaban a cada grupo resultante.
- **Método de K-medias recursivo:** Esta alternativa funciona de manera equivalente al algoritmo de K-medias, pero se realizó la agrupación de manera iterativa: se dividieron las propuestas en 5 grupos, cada uno de los 5 grupos se dividió en 5 subgrupos, luego los 25 subgrupos resultantes se dividieron también en 5 subgrupos cada uno y así sucesivamente hasta que cada subgrupo tuviera un número de propuestas menor o igual que 5.
- **Método de agrupamiento jerárquico aglomerativo:** Este método permitió agrupar las propuestas que fueran muy similares, aunque dejando el resto de las propuestas sin agrupar. Como medida de similitud entre cada par de propuestas se tomó la similitud coseno y se utilizó el método de enlace completo para ejecutar el algoritmo.
- **Método de agrupamiento basado en redes:** En este método se construye un grafo de coocurrencias para identificar las palabras que más aparecen juntas en las propuestas y sus relaciones. Un agrupamiento sobre esta red utilizando el algoritmo "*chinese whispers*" permite identificar temas latentes en las propuestas y con base en estas palabras se pueden clasificar las propuestas en categorías utilizando la misma técnica de "etiquetado por palabras clave" explicada anteriormente.

Cabe resaltar que se crearon herramientas para aplicar uno u otro enfoque de agrupamiento a las propuestas a nivel de mesa o a nivel de pregunta. El equipo asesor del DNP utilizó el enfoque que se consideró que arrojaba mejores resultados para cada mesa o pregunta y con base en ello se definieron las agrupaciones resultantes.

Estadísticas demográficas descriptivas

Posteriormente, el equipo asesor del DNP realizó una caracterización de las propuestas y de sus proponentes cruzando las agrupaciones y categorías resultantes con las variables demográficas de las personas que diligenciaron las respuestas, lo que permitió identificar temas de mayor interés para distintos grupos poblacionales. Las variables utilizadas fueron:

- Rango de edad
- Nivel de escolaridad
- Sexo

Análisis temporal

Finalmente, se agruparon las propuestas por semana de recepción (es decir, cada 7 días desde el lanzamiento de la plataforma) y se realizaron gráficos descriptivos, integrando los resultados de los puntos anteriores, para analizar:

- El número de propuestas recibidas en cada mesa, cada semana.
- Los temas de mayor interés (utilizando las etiquetas obtenidas en el paso 3) por cada mesa y por cada semana.
- La evolución del sentimiento promedio (nivel de sentimiento positivo) en las propuestas de cada mesa, por semana.

Análisis de propuestas en mesas regionales

De manera adicional, se realizaron nubes de palabras para identificar temáticas principales en las mesas regionales, con base en las propuestas registradas en ellas.

Resultados

El desarrollo metodológico aquí presentado brindó numerosos insumos de análisis a la Dirección General del DNP, cuyo equipo consolidó los principales resultados para realizar los análisis de cuáles son las principales preocupaciones que se observan en cada región por grupo de edad, sexo y nivel de escolaridad. Presentar estos resultados con el mismo nivel de detalle va más allá del objeto de este informe técnico, por lo que acá los solo se presentan resultados que permiten ilustrar los tipos de gráficos obtenidos y las herramientas desarrolladas para facilitar el análisis.

Comenzando por los gráficos descriptivos (nubes de palabras), estos se automatizaron y se generaron para cada una de las mesas y para cada una de las preguntas. Por ejemplo, para la pregunta "¿Cómo evolucionar hacia el modelo de educación que responda a las necesidades y desafíos propios del siglo XXI?", se obtuvo la nube presentada en la figura 1.

En cuanto al análisis de sentimiento, se etiquetaron las propuestas en la base de datos en 5 niveles (positivo, ligeramente positivo, neutro, ligeramente negativo y negativo). Adicionalmente, se graficó la evolución del nivel de sentimiento promedio por mesa, obteniendo el gráfico presentado en la figura 2. Allí se observa que mesas como la de educación y la de juventud tienen propuestas con mayor sentimiento positivo estimado, aunque las tendencias no son concluyentes en ninguna de las mesas.

Figura 1: Nube de palabras construida para las respuestas a la pregunta "¿Cómo evolucionar hacia el modelo de educación que responda a las necesidades y desafíos propios del siglo XXI?"

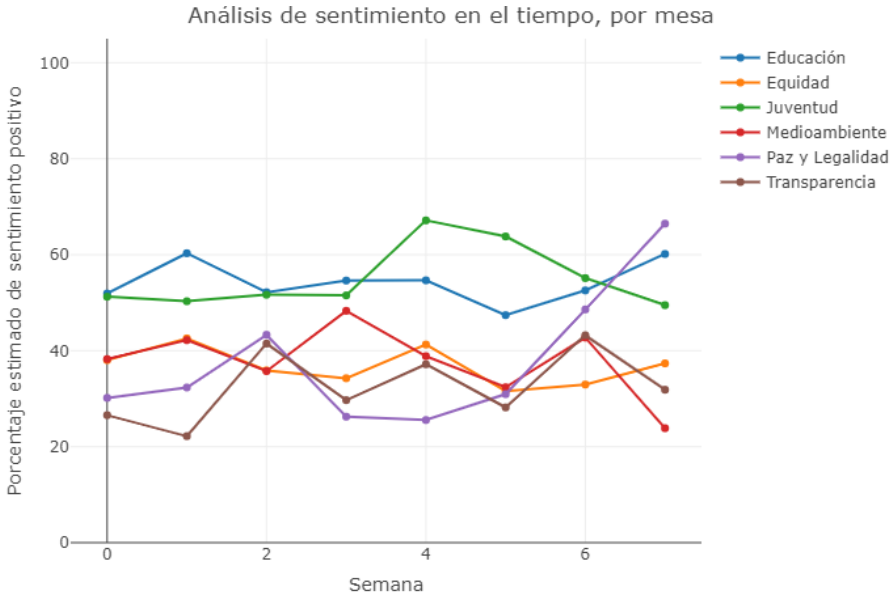

Figura 2: Evolución del nivel de sentimiento promedio, por mesa y semana.

En cuanto al etiquetado por palabras clave, esta fue la técnica que presentó mejores resultados para la identificación de temáticas relevantes e incluso como proxy de la agrupación de las propuestas, con la única desventaja de que permitió clasificar un poco menos de la mitad de las propuestas. De las 10625 propuestas, 4498 se lograron clasificar con las palabras clave en las categorías presentadas en la figura 3.

Respecto a las otras metodologías de agrupación, el agrupamiento jerárquico aglomerativo permitió agrupar (por lo menos emparejar) 2498 propuestas con un umbral de similitud coseno de 0.7, mientras que con un umbral de 0.5 se lograron agrupar (por lo menos emparejar) 5750 propuestas. La clasificación de palabras realizada con agrupamiento basado en redes permitió obtener la red ilustrada en la figura 4, pero no se realizó la clasificación basada en palabras ya que las palabras y categorías brindadas por la Dirección General se consideraron más pertinentes para el análisis.

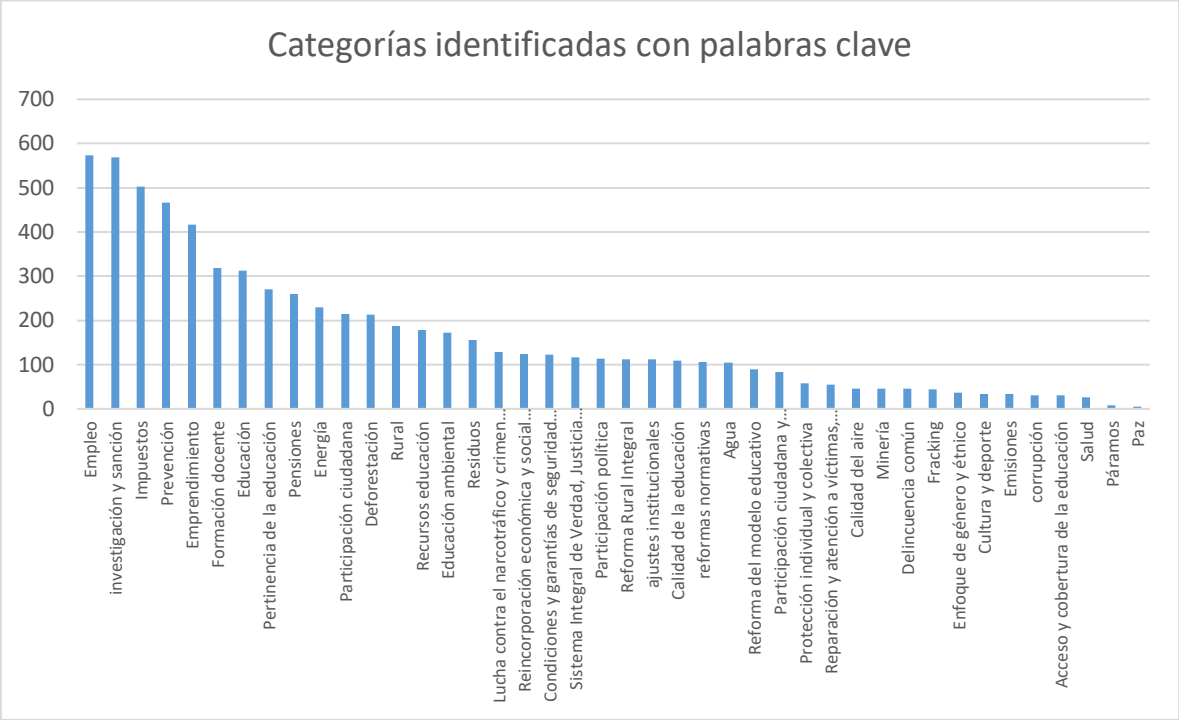

Figura 3: Propuestas clasificadas en las categorías identificadas por palabras clave.

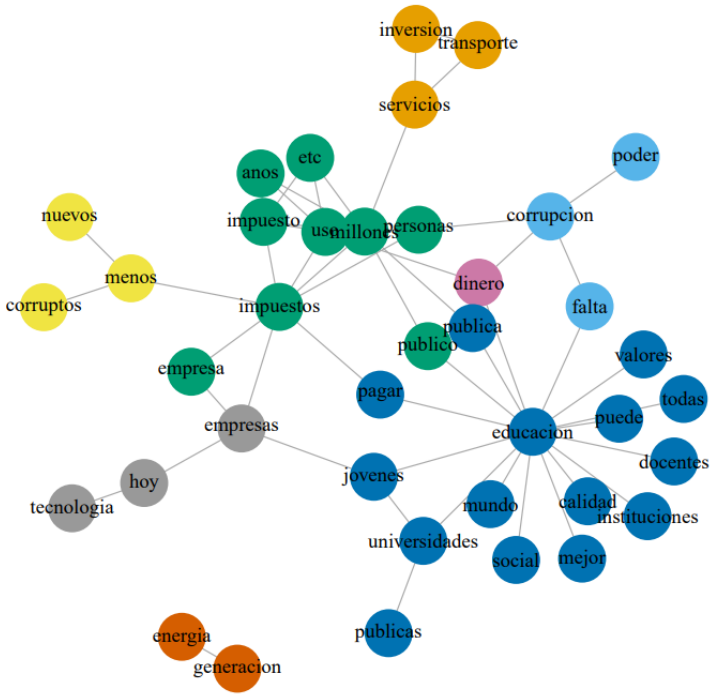

Figura 4: Agrupación de palabras sobre la red de coocurrencias.

Por otro lado, las técnicas de LDA, k-medias y k-medias recursivo permitieron agrupar las propuestas en su totalidad y sus resultados se presentaron en los aplicativos desarrollados. Uno de ellos, desarrollado en Python, permite filtrar las propuestas por mesa, pregunta, fecha y sentimiento, tras lo cual se realizan los gráficos descriptivos y se reporta el agrupamiento con LDA y k-medias. La interfaz de este aplicativo se ilustra en la figura 5. El otro aplicativo, desarrollado en R, permite filtrar también por mesa y pregunta y contiene únicamente los resultados del agrupamiento con k-medias recursivo, los cuales se pueden visualizar de forma interactiva (haciendo clic sobre un nodo para desagregar el grupo correspondiente en 5 nuevas categorías), acompañados de una nube de palabras y de una tabla que se actualizan de acuerdo con las propuestas del grupo sobre el cual el usuario hace clic. La figura 6 ilustra su funcionamiento para el caso en que el usuario hace clic sobre el nodo “contratos corrupción congreso”, con 337 propuestas.

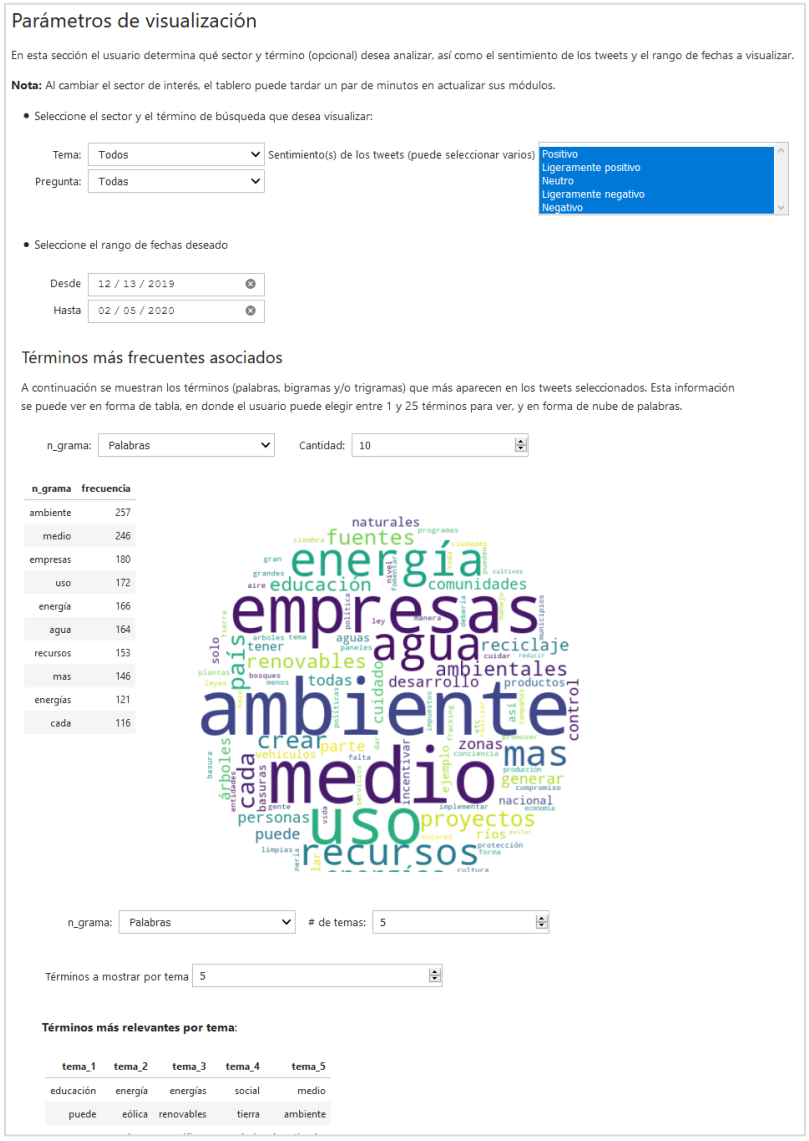

Figura 5: Interfaz del aplicativo desarrollado en Python. Los gráficos descriptivos y los grupos (clusters) se actualizan con base en los filtros escogidos por el usuario.

Figura 6: Interfaz del aplicativo desarrollado en R. La nube de palabras y la tabla corresponden a la interacción con el nodo "contratos corrupción congreso".

Conclusiones y recomendaciones

1. El uso de técnicas de minería de texto mostró ser una alternativa pertinente para facilitar la clasificación y análisis de las más de 12.000 propuestas recibidas a través de la plataforma de la Conversación Nacional.
2. El proyecto permitió identificar las principales preocupaciones expresadas por la población en las diferentes mesas, lo que constituyó un insumo valioso para la Dirección General, quien cruzó esta información con características demográficas y en el marco del análisis integral de las propuestas.
3. Las técnicas y herramientas desarrolladas permitieron a los usuarios visualizar, de manera agregada, la información más relevante, con la posibilidad adicional de interactuar con ella para realizar análisis de interés a diferentes niveles de agregación.

Socialización

Este proyecto se socializó con la Dirección General del DNP y las categorías identificadas con palabras clave se incluyeron en la base de datos que se presentó y compartió a las direcciones técnicas del DNP para su análisis.