Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

CLASIFICACIÓN Y GENERACIÓN DE RESPUESTAS AUTOMÁTICAS A PETICIONES, QUEJAS, RECLAMOS, SUGERENCIAS Y DENUNCIAS (PQRSD) RADICADAS EN EL DNP

Entidad
Departamento Nacional de Planeación
- Dirección de Desarrollo Digital.
- Centro de Servicio al Ciudadano.

Sector	**Lenguaje**	**Fuente de datos**
Planeación.	R.	ORFEO - Base de datos de PQRSD del Centro de Servicio al Ciudadano del DNP.

Presentación
El sistema de peticiones, quejas, reclamos, sugerencias y denuncias (PQRSD) es una herramienta que permite a las entidades públicas conocer las distintas necesidades por parte de los ciudadanos con el objetivo de fortalecer los programas ofrecidos y focalizar la inversión social. Dado que los ciudadanos buscan soluciones prontas a sus inquietudes, es relevante el desarrollo de un mecanismo que permita clasificar automáticamente las PQRSD enviadas por los ciudadanos y de esta manera brindarles respuestas de forma automática para así ser un canal de atención más eficiente y reducir la cantidad de solicitudes tramitadas en el DNP. Por tanto, se presenta una metodología de clasificación automática de las PQRSD mediante un modelo de aprendizaje de máquina que permite identificar peticiones sobre inconformidad con el puntaje de SISBÉN para poder así brindar una respuesta automática.

The system of petitions, complaints, claims, suggestions and denunciations (PQRSD) is a tool that allows public entities to know the citizens needs to strengthen the programs offered and focus social investment. Given that citizens seek prompt solutions to their concerns, it is important to develop a mechanism to automatically classify the PQRSD sent by citizens and thus provide them with automatic responses to be a more efficient channel of attention and reduce the number of requests serviced by DNP. Therefore, an automatic PQRSD classification methodology is presented through a machine learning model that allows the identification of petitions on non-conformity with SISBEN score so that an automatic response can be provided.

Objetivo general
Brindar respuestas automáticas a las PQRSD sobre el tema de inconformidad con el puntaje de SISBÉN a través de la aplicación de técnicas de minería de texto y aprendizaje de máquina.

Objetivos específicos
1. Identificar palabras clave en las PQRSD sobre la temática de inconformidad con el puntaje de SISBÉN a través de un análisis descriptivo de las PQRSD.
2. Etiquetar las PQRSD de manera automática mediante la detección de las palabras clave identificadas.
3. Entrenar un modelo de aprendizaje de máquina que identifique las PQRSD sobre inconformidad con el puntaje de SISBÉN.
4. Diseñar un módulo de respuestas automáticas para implementar la solución dentro de la infraestructura tecnológica de DNP.

Antecedentes

En el marco del proyecto "Análisis y clasificación de las PQRSD radicadas en DNP", realizado por la Unidad de Científicos de Datos en primer semestre del 2019, se identificó que uno de los temas más comunes en las PQRSD que llegan al DNP correspondía a inconformidades con el puntaje de SISBÉN, un tipo de solicitud al que deben dar respuesta las entidades territoriales pero que en numerosas ocasiones se radican en DNP a través de distintos medios, entre los cuales se destaca el formulario web. Con el objetivo de disminuir la carga que representa para el DNP la atención de estas PQRSD, se planteó un proyecto que permitiera dar respuesta más rápidamente a las PQRSD sobre este tipo, como se presenta a lo largo del presente documento.

Metodología

La metodología utilizada para analizar las PQRSD puede dividirse en cinco grandes fases: (1) el preprocesamiento de los textos, (2) la identificación de palabras latentes, (3) clasificación inicial de las PQRSD a partir de palabras claves (4) la representación vectorial (matemática) de los textos, (5) el desarrollo de un algoritmo de clasificación supervisado para identificar el tema correspondiente y brindar la respuesta al usuario.

Preprocesamiento o limpieza del texto

Al realizar la lectura de los asuntos contenidos en la base correspondientes a cada PQRSD se obtuvo una cadena de texto correspondiente a cada asunto. Con el fin de limpiar el texto contenido en cada cadena, se realizó una serie de transformaciones al mismo, la cual consistía en cambiar mayúsculas por minúsculas, remoción de números, signos de puntuación y demás caracteres que no fueran letras así como de los correos electrónicos y palabras que estuvieran compuestas de más de 19 caracteres y de menos de 2; adicionalmente se removieron conectores, preposiciones y palabras que no eran relevantes en el contexto, entre estas se incluyeron nombres, apellidos y zonas geográficas.

El texto obtenido anteriormente fue corregido ortográficamente, para lo cual se toma la palabra que se quiere corregir, se calcula la distancia de Levenshtein entre ella y un listado de palabras ordenado por probabilidad de ocurrencia, y se toma aquella más probable entre las que tienen la menor distancia. En caso de que todas las palabras tengan una distancia de Levenshtein mayor a 2, la palabra no se modifica. El listado de referencia para realizar la corrección está compuesto por las 10.000 formas (palabras) más frecuentes del español según la RAE, junto con palabras propias del problema como "SISBÉN", "RUV" y "GESPROY", a las cuales se les asignó la mayor probabilidad de ocurrencia.

Tomando como base el texto corregido, se efectuó también un proceso de lematización, que consiste en obtener el lema de cada palabra, siendo esta una forma inflexiva del término y permitiendo así la eliminación de formas verbales conjugadas, plurales, entre otras (por ejemplo, "jugar, jugamos y jugaríamos" se convierten en "jugar"). Al texto resultante se le denominó "texto lematizado" y este fue la base para procedimientos posteriores.

Inferencia de palabras latentes

Una vez lematizado el texto, se buscó implementar una metodología de inferencia de palabras latentes haciendo uso de dos metodologías principales: teoría de grafos y probabilidades condicionales.

Teoría de grafos

Para construir el grafo, como primera medida se obtuvo la matriz de coocurrencias de las palabras, esta es una matriz simétrica y cuadrada cuyas filas y columnas son todas las palabras que aparecen en el texto y da cuenta de la cantidad de veces que dos palabras aparecen juntas en una misma oración, por ejemplo, para las oraciones: “Quiero un auto rojo”, “Quiero un auto azul”, “Quiero un auto gris”, la matriz de coocurrencia es la mostrada en la Tabla 1.

Palabra	Quiero	un	auto	rojo	gris	verde
Quiero	0	3	3	1	1	1
Un	0	0	3	1	1	1
Auto	0	0	0	1	1	1
Rojo	0	0	0	0	0	0
Gris	0	0	0	0	0	0
Verde	0	0	0	0	0	0

Tabla 1: Ejemplo matriz de coocurrencias

A partir de la matriz obtenida se construye un grafo no dirigido cuyos nodos son las palabras, las conexiones entre dos palabras están dadas si estas aparecen juntas en la misma oración y el peso del enlace es la cantidad de veces que esto sucede.

Posterior a esto, se implementaron dos metodologías para inferir las palabras latentes. La primera consistió en calcular la probabilidad de conexión entre dos palabras haciendo uso del índice de similitud de Jaccard, este índice, dadas dos palabras A y B se calcula como $J(A,B) = |A \cap B| / |A \cup B|$ que representa la relación de los vecinos comunes de A y B. Para implementar esta metodología, se hizo un conteo de las apariciones de cada palabra en todas las PQRSD, posteriormente se tomó cada PQRSD individualmente y de acuerdo con las frecuencias calculadas en el paso anterior, se extrajeron las tres palabras con mayor cantidad de apariciones y a cada una le fue calculado el índice de Jaccard con todas las demás palabras presentes en todas las PQRSD, obteniendo la probabilidad de conexión de cada una de las tres palabras con todas las restantes. Una vez realizado este proceso y con el fin de obtener una probabilidad conjunta de la aparición de todas las palabras dadas las 3 de interés, se sumaron las probabilidades individuales y se dividió por la probabilidad máxima con el fin del obtener valores entre cero y uno, posteriormente se seleccionaron aquellas palabras que superaran el 90% de probabilidad de aparición dadas las tres palabras obtenidas a partir de la PQRSD.

Para la segunda metodología se tiene que para una PQRSD determinada, se toman en consideración todos los posibles subconjuntos de nodos (subgrafos) donde se encuentren presentes las palabras contenidas en esta y se les calcula la medida de coherencia que se define como la razón entre la media geométrica y la media aritmética de los pesos. Un ejemplo de esta metodología se muestra en la Figura 1.

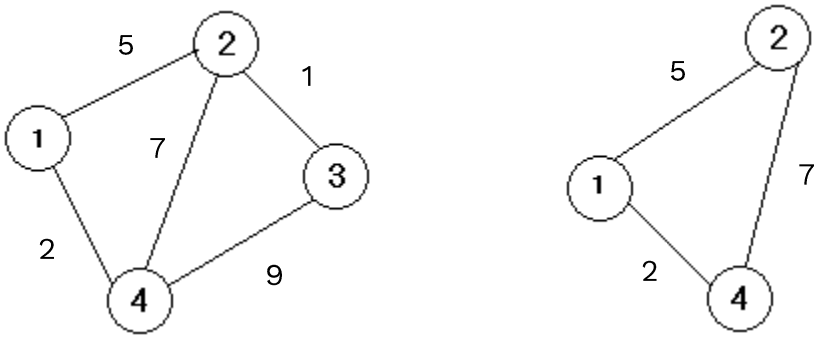

Figura 1: Ejemplo de grafo y subgrafo con pesos

En la parte izquierda de la Figura 1 se ilustra un grafo no dirigido con cuatro nodos, cuyos pesos se encuentran a lado de las conexiones, en la parte derecha de la misma figura se presenta un subgrafo del grafo inicial cuyos pesos de enlace son 5, 7 y 2, con base en esto se tiene que la coherencia del subgrafo es:

$$Q(g) = \sqrt[3]{14/4{,}6}$$

Una vez calculada la coherencia para todos los subgrafos, se seleccionaron las cinco palabras que mayor valor presentaron de todos los subgrafos considerados, siendo estas las palabras latentes de la PQRSD y este proceso se realizó para todos los asuntos.

Probabilidades condicionales

Haciendo uso de los conteos de aparición de las palabras mencionados previamente, se extrajeron las dos palabras con mayor frecuencia en cada PQRSD y se calculó la probabilidad de aparición de cada palabra dadas las dos existentes mediante la siguiente fórmula:

$$P(A|\ B \cap C) = P(A \cap B \cap C)/P(B \cap C)$$

Donde la probabilidad del numerador se calcula como la razón entre la cantidad de veces que aparecen las palabras A, B, C en la misma PQRSD y todos los trigramas presentes en los asuntos, mientras que el denominador se calcula como la proporción de aparición de las palabras B y C juntas con respecto al total de bigramas presentes en los textos.

Clasificación inicial de las PQRSD

Dada la información contenida en los asuntos de las PQRSD, se identifica que uno de los temas principales temas a tratar por parte de los ciudadanos es la inconformidad con el puntaje del Sisbén que les fue asignado, así como la solicitud de una nueva encuesta, por tanto, se exploraron las PQRSD referentes a este tema con el fin de determinar algunas palabras características de este tipo de solicitudes, permitiendo así una clasificación inicial de los asuntos.

A partir de esto se lograron identificar palabras tales como: puntaje, inconformidad, alto, Sisbén, madre, cabeza hogar, bajar, bajo, entre otras, con base en esto se conformaron combinaciones de dichas palabras que solían

aparecer juntas en los diferentes requerimientos, para así determinar si una PQRSD dada era referente a inconformidad con el puntaje o no, y así efectuar el posterior entrenamiento del modelo.

Cabe destacar que con el fin de validar esta clasificación inicial se tomaron 10 muestras de 30 PQRSD cada una y se realizó la verificación del etiquetado de forma manual, cambiando las etiquetas en caso de ser necesario, para así garantizar el aprendizaje adecuado del modelo.

Representación vectorial del texto

Una vez obtenida la clasificación inicial de las PQRSD, se procedió a representar los textos de forma vectorial haciendo uso de un método conocido como *hashing,* este consiste en asignar un vector único de longitud fija predefinida a cada texto mediante una función de hash, en este caso, la longitud del vector fue de 100.

Modelo de clasificación

Para realizar la clasificación de las PQRSD correspondientes a inconformidad con el puntaje del Sisbén, se entrenó un modelo de respuesta binaria, dividiendo el conjunto de datos en 80% para entrenamiento (para ajustar el modelo) y 20% para prueba (para estimar y validar su capacidad de predicción). Cabe destacar que la cantidad PQRSD etiquetadas como inconformidad es considerablemente menor a las que no son identificadas como pertenecientes de esta temática, por tanto, previo al entrenamiento del modelo se realizó un muestreo de las PQRSD que no pertenecen a inconformidad con el fin de balancear las clases 60% a 40% de no inconformidad e inconformidad respectivamente. Por otra parte, el objetivo principal en la evaluación del desempeño del modelo fue reducir la cantidad de falsos positivos, dado que es de especial relevancia no responder a una PQRSD que no es de inconformidad como si lo fuese, para lograr este objetivo, se evaluaron tres métricas de desempeño: precisión, recall y el puntaje F1 (Tabla 2).

$$\text{Precisión} = \frac{VP}{VP+FP} \qquad \text{Recall} = \frac{VP}{VP+FN} \qquad \text{F1} = \frac{2\,Recall*Precisión}{Recall+Precisión}$$

		Clase predicha	
		Sí	No
Clase real	Sí	Verdadero Positivo (VP)	Falso Negativo (FN)
	No	Falso Positivo (FP)	Verdadero Negativo (VN)

Tabla 2: Estructura general de una matriz de confusión

Al obtener las predicciones del modelo, los resultados están dados como probabilidades de pertenencia de las PQRSD a la temática de inconformidad, por tanto se decidió variar el umbral a partir del cual una PQRSD puede ser considerada de inconformidad, la decisión del valor óptimo del umbral se tomó con base en las métricas previamente mencionadas, dado que la precisión permite garantizar que la tasa de falsos positivos sea baja, mientras que el recall muestra la pérdida de verdaderos positivos que se obtiene conforme se aumenta la precisión y finalmente el puntaje F1 condensa la información contenida en las dos métricas previamente mencionadas.

Resultados

Clasificación inicial

Se eligieron algunas palabras que permitían identificar las PQRSD referentes a la inconformidad de un usuario con respecto a su puntaje del Sisbén, algunas de ellas fueron: inconformidad, alto, puntaje, Sisbén, madre, cabeza, hogar, entre otras. Con base en ello, se realizó la clasificación inicial de los asuntos en dos clases: "Sí" y "No", permitiendo así la identificación de 1565 PQRS referentes a inconformidad con el puntaje.

Inferencia de palabras latentes

Como primera medida se realizó la inferencia de palabras latentes a partir de probabilidades condicionales, sin embargo, esto fue computacionalmente costoso dado el tiempo que tarda el algoritmo en ejecutarse, además no es posible realizar la inferencia con más de dos palabras, pues las intersecciones reducen su probabilidad de ocurrencia, tendiendo a cero al aumentar el número de palabras con las cuales se infiere, por lo tanto no es posible hacer uso de toda la información disponible de la PQRSD.

Por otra parte, a partir de la matriz de coocurrencias se obtuvo el grafo presentado en la Figura 2, una vez construido, se realizó la inferencia de conexiones entre palabras haciendo uso del índice de Jaccard, sin embargo, dada la metodología planteada para su cálculo, no es posible incluir más de tres palabras para realizar la inferencia de términos, excluyendo información relevante del análisis.

Figura 2: Grafo obtenido de la matriz de coocurrencias

Como consecuencia de lo anterior, se decide implementar la metodología donde se examina la estructura del grafo, obteniendo resultados más consistentes que fueron revisados detalladamente. Por ejemplo, se encontró que al ingresar la PQRSD *puntaje Sisbén alto,* texto obtenido después del tratamiento y preprocesamiento de la información, el algoritmo infiere las palabras *solicitud, encuesta, nueva, cambiar, inconforme*, lo cual es consistente con la información previa que se tiene acerca de este tipo de solicitudes, es relevante aclarar que el grafo fue construido a partir del texto lematizado, pues de esta manera se evitaba redundancia en la

información y por ende en las conexiones que se pueden dar a causa de plurales o conjugaciones verbales. Cabe destacar que al aplicar el algoritmo de clasificación nuevamente sobre las PQRSD con las palabras latentes se logró identificar algunas adicionales, aumentando de 1565 sin palabras latentes a 1904 si estas son tomadas en consideración.

Modelo de aprendizaje supervisado

Se ajustaron cuatro tipos de modelo diferentes: Regresión logística, Random Forest, Regresión logística con boosting y máquinas de soporte vectorial lineales y radiales (SVM), obteniendo los mejores resultados con la regresión logística al comparar las matrices de confusión de los respectivos modelos.

Inicialmente, con la regresión logística se obtuvieron 87 falsos positivos, dando como resultado una precisión de 0.74, un recall de 0.69 y por tanto un puntaje F1 de 0.71. Con el objetivo de reducir la cantidad de falsos positivos sin perder verdaderos positivos en el proceso, se decidió variar el umbral de probabilidad a partir del cual una PQRSD puede ser considerada como referente a inconformidad, para ello se realizaron variaciones de dicho umbral en un intervalo de 0.5 a 1 para determinar el valor óptimo. Los resultados obtenidos se muestran en las Figuras 3 – 5.

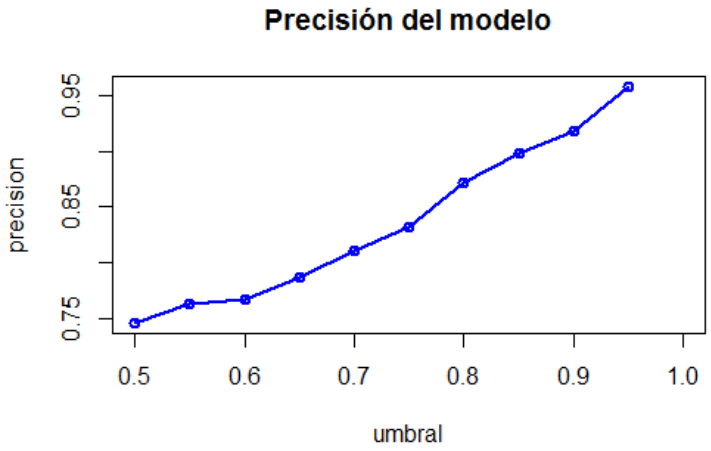

Figura 3: Precisión del modelo logístico

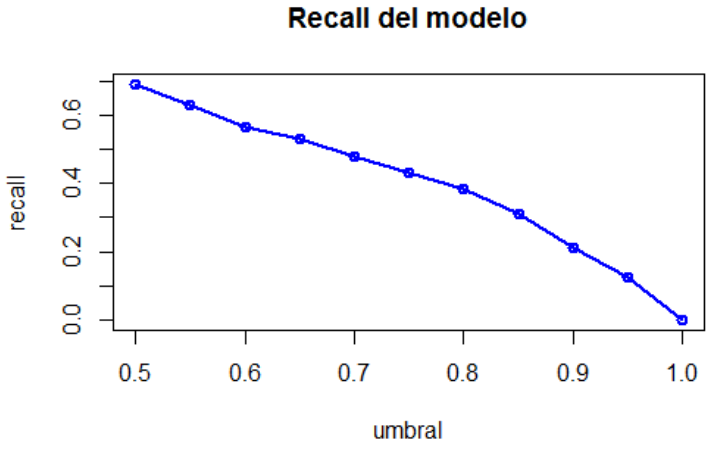

Figura 4: Recall del modelo logístico

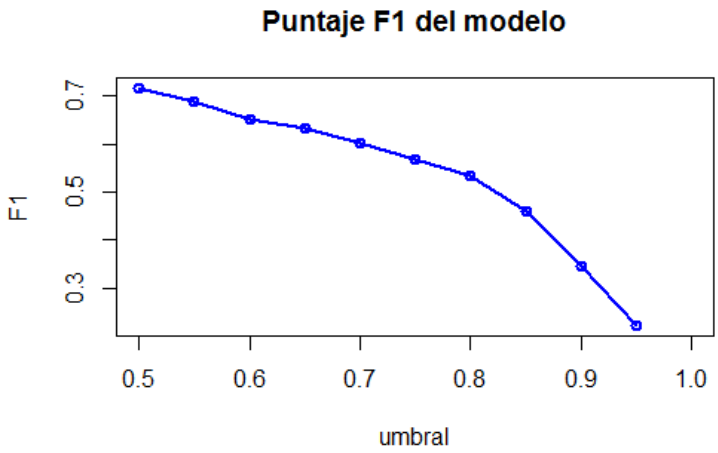

Figura 5: Puntaje F1 del modelo logístico

A partir de lo presentado en las Figuras 3, 4 y 5 se decidió tomar un valor de 0.8 para el umbral, considerando que este es el punto óptimo al tomar como referencia las 3 medidas halladas. Al realizar el cambio del valor se encontró que la cantidad de falsos positivos se redujeron a 21, siendo este el mejor desempeño presentado por el modelo.

Conclusiones y recomendaciones

1. El desarrollo propuesto muestra que el uso de minería de texto y modelos de aprendizaje de máquina constituye una alternativa viable a nivel técnico para dar respuesta a las PQRSD sobre inconformidad con el puntaje de SISBÉN.
2. Se logró entrenar un modelo que identifica las PQRSD de interés con una precisión cercana al 90%, sin embargo, no es viable a nivel técnico reducir a cero (0%) la tasa de falsos positivos, lo que implica que el modelo podría generar respuestas automáticas a algunas solicitudes que no fueran sobre inconformidad con el puntaje de SISBÉN, elemento que debe ser considerado al momento de definir si se cuenta con la viabilidad jurídica necesaria para implementarlo.
3. Se debe evaluar la viabilidad de implementar la solución dentro de la infraestructura tecnológica de DNP con base en los recursos tecnológicos existentes y en los lenguajes de programación en ellos utilizados.

Socialización

El presente proyecto se socializó con el Centro de Servicio al Ciudadano de la Secretaría General y con el Programa Nacional de Servicio al Ciudadano.