Dirección de Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento Nacional de Planeación

CLASIFICADOR DE PROYECTOS DE INVERSIÓN EN EL MARCO DE LA IMPLEMENTACIÓN DEL ACUERDO DE PAZ

Entidad
Departamento Nacional de Planeación
- Dirección de Desarrollo Digital.
- Dirección General – Grupo de Proyectos Especiales.

Sector	**Lenguaje**	**Fuente de datos**
Planeación	R.	SUIFP, Matriz PMI.

Presentación
Para el cumplimiento del "Acuerdo Final para la terminación del conflicto y la construcción de una paz estable y duradera", suscrito entre el Gobierno de Colombia y las FARC-EP el 24 de noviembre de 2016, el Grupo de Proyectos Especiales del DNP elaboró el Plan Marco de Implementación del Acuerdo Final (PMI), que es el principal referente para la inclusión de los componentes de paz en los Planes Nacionales de Desarrollo y que contiene el conjunto de pilares, estrategias, productos, metas trazadoras e indicadores necesarios para la implementación del Acuerdo Final. Ya que esta implementación requiere de un gran nivel de inversión de recursos públicos, es importante realizar una cuantificación de los montos destinados al desarrollo de los productos descritos en el PMI, algo que se posibilita con este trabajo, utilizando algoritmos de minería de texto, agrupación y clasificación.

To ensure compliance with the "Final Agreement for the termination of the conflict and the construction of a stable and lasting peace", signed between the Government of Colombia and the FARC-EP on November 24, 2016, the Special Projects Group of the DNP elaborated the Framework Plan for the Implementation of the Final Agreement (PMI), which is the main reference for the inclusion of peace components in the National Development Plans and which contains the set of pillars, strategies, products, tracing goals and indicators necessary for implementing the Final Agreement. Since this implementation requires a high level of investment of public resources, it is important to quantify the amounts allocated for the development of the products described in the PMI, what is facilitated with this work using text mining, clustering and classification algorithms.

Objetivo general
Analizar la inversión del Gobierno de Colombia para el cumplimiento de los Acuerdos de Paz, en función de los productos e indicadores definidos en el Plan Marco de Implementación (PMI).

Objetivos específicos
1. Analizar los pilares, estrategias, productos, metas trazadoras e indicadores del PMI mediante algoritmos de minería de texto y de agrupamiento, identificando relaciones entre ellos y obteniendo grupos con textos similares.
2. Clasificar los proyectos de inversión de acuerdo a su contenido textual entre proyectos correspondientes a paz, postconflicto y víctimas.
3. Desarrollar un algoritmo basado en texto para medir la similitud existente entre los proyectos de inversión (del PGN, del SGR y de inversión territorial) y los distintos elementos del PMI.
4. Clasificar los proyectos de inversión dentro de las categorías que puedan definirse a partir de los elementos definidos en el PMI.

Metodología

La metodología desarrollada para la clasificación de los proyectos de inversión articula distintos procedimientos y algoritmos de análisis de texto que pueden resumirse en una vectorización de los textos, una reducción de dimensionalidad, un agrupamiento (*clustering*) y una clasificación.

Vectorización de textos

En primer lugar, se analizaron los textos que describen cada estrategia, producto, meta trazadora (si aplica) e indicador definido en el PMI, para lo cual se utilizó la matriz del PMI para el PND, brindada por el Grupo de Proyectos Especiales (GPE). Estos ítems se agruparon, obteniendo una cadena de texto por cada fila de la matriz. La limpieza de las cadenas de texto obtenidas consistió en la transformación del texto a minúsculas y en la remoción de números, signos de puntuación y demás caracteres distintos a las letras que conforman las palabras, también se removieron conectores, preposiciones y palabras que no agregan significado al texto, entre las cuales se incluyen zonas geográficas y palabras como “nación”, “proyecto” y “Colombia”, que son transversales a todos los textos y no ayudan a diferenciarlos entre sí. Si bien en algunos casos no es necesario eliminar tildes, en este se optó por hacerlo pues, aunque puede generar ambigüedades, estas no se utilizan en las descripciones de un gran número de proyectos. Para ejemplificar este procedimiento, puede pensarse en una cadena de texto como “4. Sustitución de cultivos de Uso Ilícito” que quedaría convertida en “sustitucion cultivos ilicito".

Habiendo obtenido las cadenas de texto, se utilizó el modelo de bolsa de palabras (*bag of words*), que consiste en construir una matriz con palabras en las columnas y textos en las filas para representar el número de veces que aparece cada palabra en cada documento. Por ejemplo, si el texto 12 equivale a “sustitucion cultivos ilicito", las columnas “sustitucion”, “cultivos” e “ilicito” tendrán un valor de 1 en la fila 12 de la matriz. Adicionalmente, se convirtieron estas frecuencias dividiendo sobre el número de palabras contenidas en el texto, lo que puede interpretarse como una estimación de la probabilidad de que una palabra se encuentre en un texto. Para el caso del texto “sustitucion cultivos ilicito", de 3 palabras, los valores en la matriz en la fila 12 pasarían de ser 1 a 1/3.

Reducción de dimensionalidad y agrupación

A partir de la matriz obtenida, se puede pensar en los textos como vectores de "n" dimensiones, es decir, vectores de "n" posiciones, donde "n" corresponde al número de palabras distintas que hay en el conjunto de todos los textos. Visualizar estos resultados, sin embargo, puede ser complicado puesto que, al haber muchas palabras diferentes, se debe manejar un gran número de dimensiones. Por ello, se optó por utilizar una técnica de reducción de dimensionalidad no lineal que permitiera visualizar los productos como puntos en un espacio de pocas dimensiones, el cual conservara la mayor cantidad de información posible. La técnica utilizada fue el algoritmo t-SNE (*T-distributed Stochastic Neighbor Embedding*), cuya aplicación permitió obtener una representación de los textos en un espacio de 2 dimensiones. Al contar con los textos representados en este espacio, se ejecutó un algoritmo de agrupación (*clustering*) para conformar 12 conjuntos de textos con características similares.

Clasificación de proyectos de inversión

Posteriormente, se construyó un algoritmo de clasificación para asignar los textos de los proyectos a las 12 categorías definidas por los grupos. Para ello, se entrenó una red neuronal utilizando la representación vectorial

de los textos (la obtenida del modelo de bolsa de palabras), aprovechando que se conocía la clasificación de cada texto del PMI en las 12 categorías, equivalente a identificar a qué grupo pertenece cada texto. Tras entrenar la red neuronal, se sometieron los textos de los proyectos al mismo proceso de limpieza y vectorización, únicamente con palabras contenidas en el PMI, consideradas alusivas a paz, postconflicto y víctimas.

Obtenidos los textos de los proyectos, se realiza una clasificación previa para descartar los proyectos que no hacen referencia a los temas de paz. Este paso permite mejorar la clasificación final y se genera a partir de una integración entre indicadores de la diferencia de frecuencias, la proporción de frecuencias y una proyección en el espacio nulo de los vectores de pablaras referentes a paz (obtenidas de los documentos del acuerdo final y el PMI) y los vectores referentes a 33 CONPES de 11 sectores de la economía. Estos indicadores permiten obtener las palabras exclusivas y las más distintivas de los textos de paz. Tras identificar aquellas palabras, se asigna un puntaje de similitud a los textos de cada proyecto en función de las palabras exclusivas y representativas que contiene, las cuales otorgan un puntaje mayor o menor según el valor calculado de sus indicadores. Así, los textos vectorizados de los proyectos se hicieron pasar a través de la red neuronal, obteniendo para cada uno su probabilidad de pertenecer a cada categoría y clasificando cada texto a la categoría a la que pertenece con mayor probabilidad y estimando la clasificación en puntos del acuerdo según el porcentaje de elementos del PMI de cada grupo (*cluster*) alusivos a cada punto.

Resultados

A partir del procedimiento realizado sobre la matriz del PMI, se obtuvieron 502 cadenas de texto para analizar. Tras depurarlas e identificar palabras únicas, se obtuvieron 1.012 palabras distintas, por lo que la matriz resultante del modelo de bolsa de palabras contenía 502.964 elementos (497 x 1.012). De la aplicación de los algoritmos t-SNE y de agrupación sobre textos vectorizados (es decir, sobre las filas de la matriz), se obtuvo la representación bidimensional de los elementos del PMI que se presenta en la figura 1, donde el nombre asignado a cada grupo corresponde a las dos palabras que más veces aparecen en los textos que lo conforman.

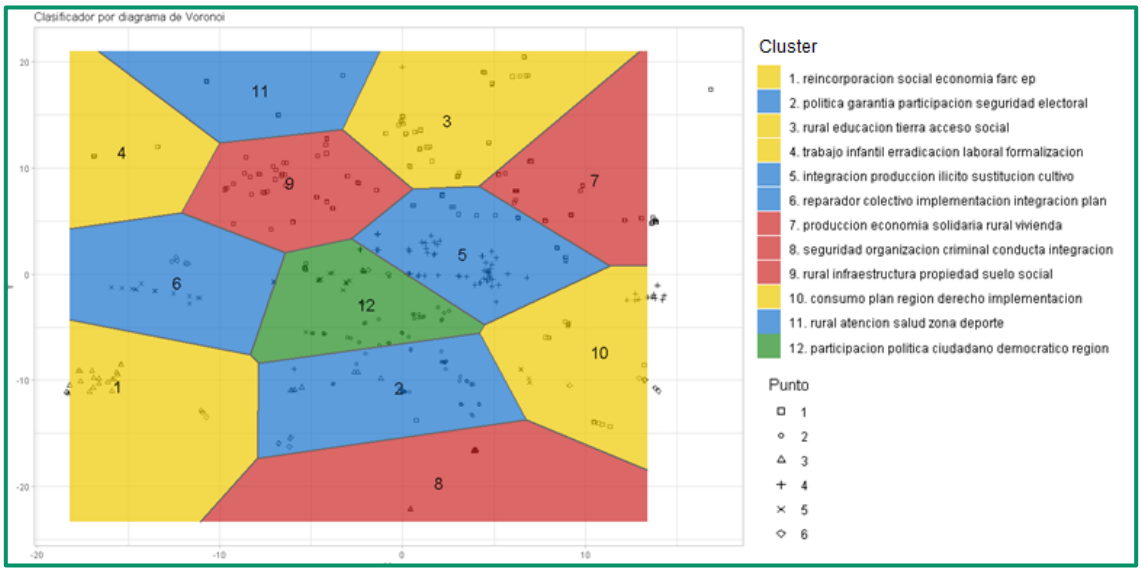

Figura 1: Cadenas de texto construidas a partir de los elementos del PMI representadas en los clusters.

En el caso de los proyectos, se analizaron 6.141 proyectos de inversión del Presupuesto General de la Nación (PGN), 13.122 proyectos de inversión del Sistema General de Regalías (SGR) y 13.123 proyectos de inversión territorial, para un total de 32.386 textos. Al asignar puntajes a los proyectos para la clasificación intermedia, se filtraron los proyectos no referentes al tema de paz, tomando como propuesta un umbral que clasifica los primeros 510 proyectos, descartando 31.876. Estos 510 fueron clasificados en las 12 categorías utilizando la red neuronal entrenada con los 502 textos, ejercicio que permitió también identificar los proyectos con mayor puntaje, algunos de los cuales se muestran en la tabla 1.

N°	Nombre del proyecto
1	Implementación del sistema de alertas tempranas para la prevención de las violaciones masivas de derechos humanos en Colombia.
2	Implementación de medidas de reparación integral para las víctimas del conflicto armado en el departamento del Cauca.
3	Mejoramiento de los canales de atención y comunicación para las víctimas para facilitar su acceso a la oferta institucional.
4	Implantación de acciones y medidas orientadas a generar condiciones de seguridad y convivencia ciudadana a nivel nacional, a través del fondo nacional de seguridad y convivencia ciudadana FONSECON.
5	Asistencia a la niñez y apoyo a la familia para posibilitar el ejercicio de los derechos.
6	Apoyo al fortalecimiento del sector justicia para la reducción de la impunidad en Colombia.

Tabla 1: Proyectos de inversión con mayor puntaje de relevancia en el tema de paz.

Aplicación en Shiny

La clasificación intermedia tiene el objetivo de separar los proyectos de inversión en aquellos que hablan de paz (cualquier tema) y aquellos que no hablan de paz, para lo que se define un puntaje de similitud a partir del cual un proyecto será clasificado como proyecto de paz. Como este puntaje puede arrojar resultados poco robustos y depende principalmente del juicio experto, se desarrolló -utilizando el paquete *Shiny* de R- una aplicación que permite interactuar con el puntaje y ajustar el número correcto de proyectos de paz, que automáticamente son clasificados dentro de los 12 clústeres generados anteriormente. La interfaz de usuario de la aplicación desarrollada se muestra en la figura 2.

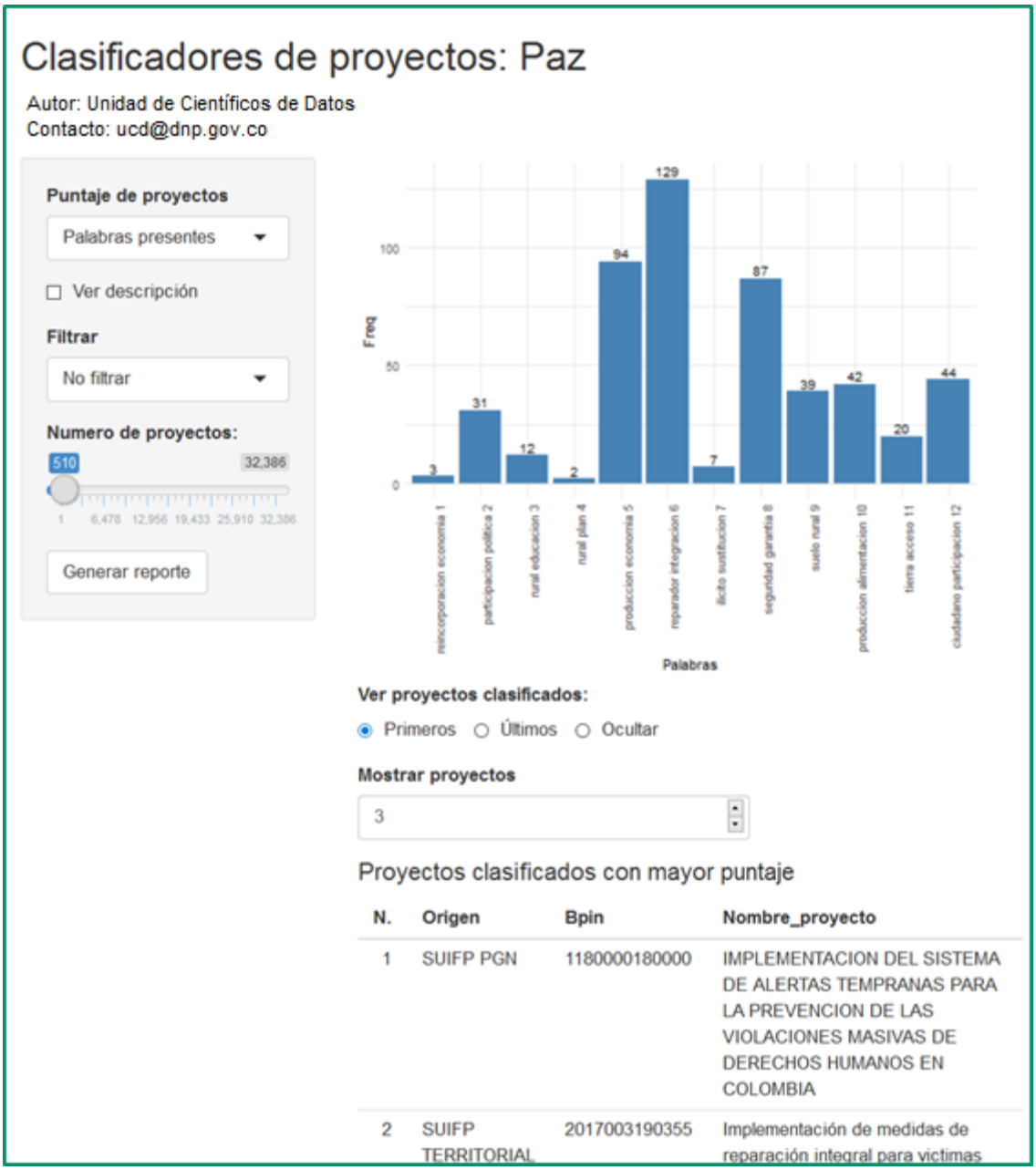

Figura 2: Interfaz de usuario de la aplicación desarrollada.

Conclusiones

1. Es posible agrupar los productos del PMI en relación con los ejes temáticos a los cuales hace referencia cada producto.
2. Los algoritmos utilizados permiten visualizar el avance en la implementación de los puntos del acuerdo en función de los proyectos de inversión analizados.
3. Mediante la inclusión de los recursos destinados a cada proyecto, se podría cuantificar la inversión total en cada punto del acuerdo y en cada uno de los ejes temáticos identificados.
4. La aplicación permite revisar detalladamente cada proyecto clasificado pues genera un reporte de aquellos proyectos clasificados y no clasificados dentro del tema de paz.

Socialización

Este proyecto se socializó con el Grupo de Proyectos Especiales de la Dirección General.