Machine Learning para la predicción de alertas en la ejecución de proyectos de Regalías

Unidad de Científicos de Datos

2017

7. Modelo para la clasificación de proyectos de regalías

Participantes: Edwin Torres y José Zea. Bajo la coordinación de Mario Morales

Problema

Diseñar un sistema que permita evaluar, en la etapa de formulación, el éxito o el fracaso de un proyecto de inversión presentado al Sistema General de Regalías. Este modelo debe realizar un análisis de la información de texto que se incluye en los estudios técnicos que describen a cada proyecto.

Insumos

Documentos formulación proyectos de regalías (pdfs), MGA, SUIFP

Objetivo

Apoyar en el acopio y depuración de la información para construir un modelo que permita predecir si un proyecto de regalías presentará una alerta:

1. Lectura de base de datos de regalías del servidor *vbidev* de la base de datos MAPAI_Fuente1.

2. Lectura de las siguientes variables de texto:

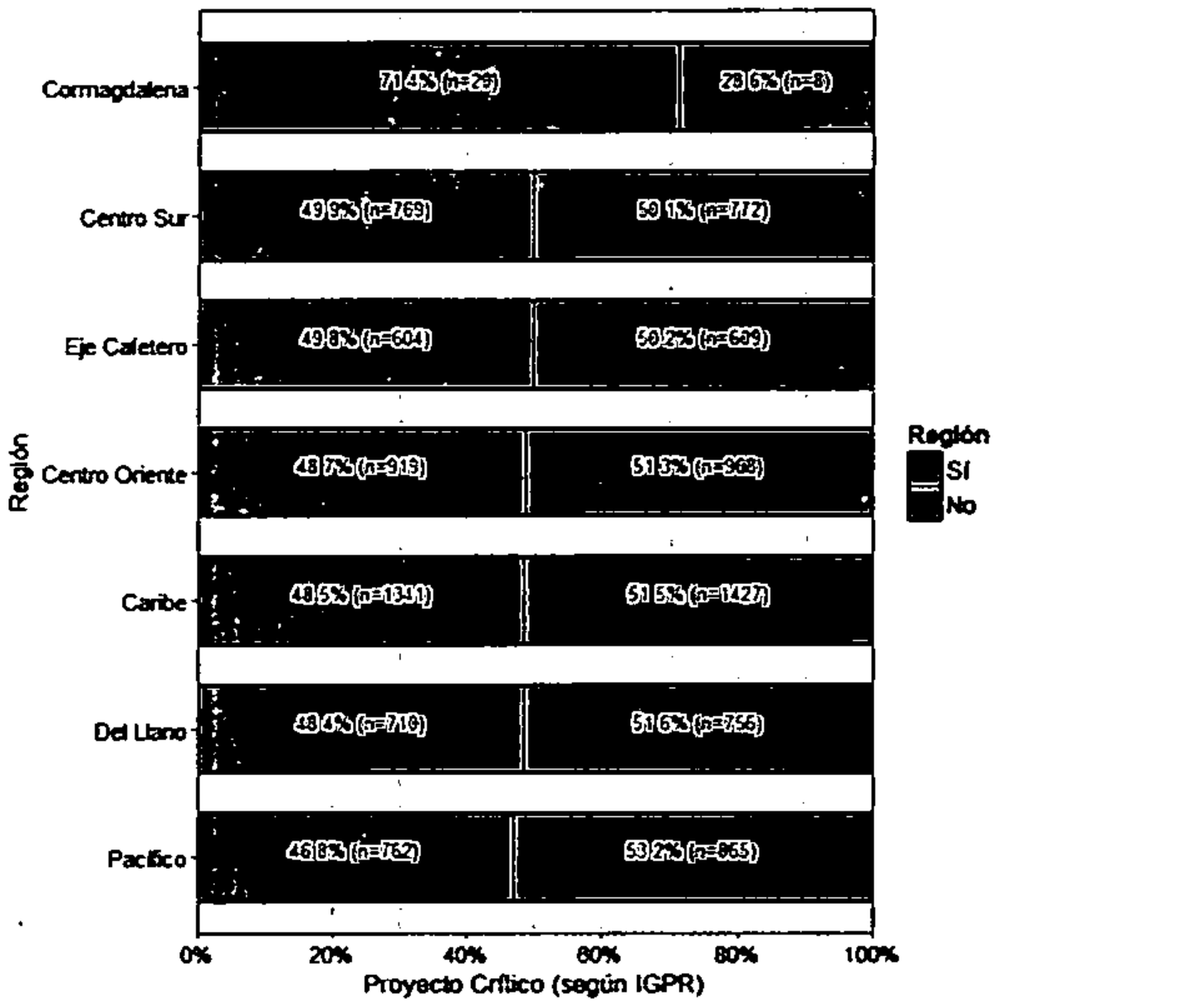
    - Identificador del texto: BPIN,
    - Nombre del proyecto
    - Problema central
    - Descripción
    - Causas directas e indirectas
    - Efectos directos e indirectos
    - Objetivo general
    - Objetivos específicos.
    - Actividades
    - Productos

    a. Limpieza básica del texto: eliminación de símbolos de puntuaciones, números, espacios
    b. Aplicación de correctores ortográficos
    c. Lematización con *Treetagger*
    d. Remover nombres geográficos del texto.
    e. Corregir las palabras mal lematizadas por *Treetagger*

3. Lectura y procesamiento de índice de Gestión de Regalías

4. Análisis exploratorio de datos.

    Se analizó el índice de regalías por



Descripción

No hay un mecanismo que permita identificar anticipadamente (en la etapa de formulación) si un proyecto entrará en estado crítico o no durante su etapa de ejecución.

El siguiente documento describe el proceso realizado para el diseño y construcción de la solución al problema formulado. El proyecto se desarrolló en tres fases: la primera en la cual se contextualiza el proceso que genera la información a analizar, en concreto a la metodología empleada en la formulación, aprobación y seguimiento que adelanta el Sistema General de Regalías (SGR) a los proyectos que solicitan su financiación. En la segunda fase se realiza un análisis exploratorio de los datos y un preprocesamiento de la información de texto recopilada y su transformación en un tipo de dato apto para ser suministrado al modelo. En la fase tres se implementan algoritmos de clasificación que permitan determinar la probabilidad de ocurrencia de una alerta.

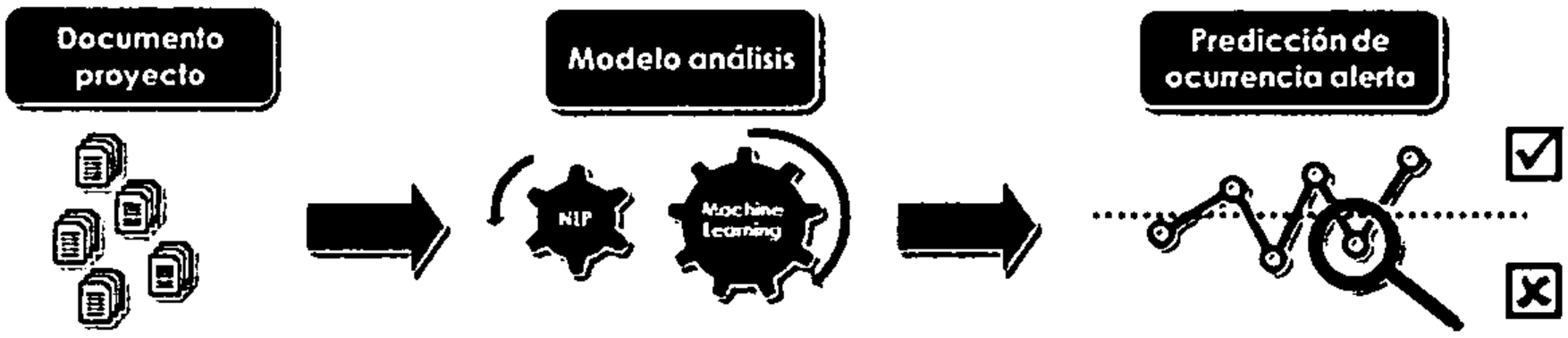


Figura 5. Flujo para la predicción de alertas.

Datos

La consolidación de la base de datos utilizada se realizó mediante la captura de la información de los proyectos almacenada en los sistemas de información usados para

tal fin. Estos son la MGA, SUIFP y Gesproy. Se cuenta con un total de 10688 proyectos desde el año 2012.

A partir de este conjunto se hace una revisión de las variables disponibles, tomando en cuenta que estas deben estar presentes para todos los proyectos. De esta revisión se define que se trabajará con las variables contenidas en los campos de la las fichas EBI (Estadísticas básicas de inversión), fichas en las cuales se encuentra la información básica de los proyectos.

La ficha EBI se encuentra estructurada por los siguientes módulos:

1. Módulo De identificación del problema o necesidad
2. Módulo de preparación de la alternativa de solución
3. Módulo de evaluación de la alternativa de solución
4. Módulo de programación
5. Módulo de decisión

La extracción de la información de texto relevante se hace del módulo 1, en el que se encuentran los siguientes campos:

- Problema central
- Descripción de la situación existente
- Magnitud actual
- Causas que generan el problema (Dir e ind)
- Efectos generados por el problema (Dir e ind)
- Objetivo general
- Objetivos específicos
- Alternativa de solución

Cada campo de estos corresponde a un texto de longitud variable. Debido a la forma en la que la información se ingresó a las bases de datos fue necesario pasar todos los textos por un corrector ortográfico que incluía terminología propia de las diferentes regiones del país. Posteriormente se procesó el texto a través de un lematizador con el fin de reducir el vocabulario sin afectar el significado de las palabras dentro del contexto de cada campo. El resultado de este procesamiento se resume en lo siguiente:

- 10.529 Proyectos
- 1'591.595 palabras (sin stop words y lematizado)
- 68.539 palabras únicas
- Longitud promedio 151 palabras por documento
- Promedio 91 palabras únicas por documento

Para las etiquetas se trabajó con el índice de gestión de regalías IGRP, que es un indicador para medir y valorar la gestión de las entidades ejecutoras que visibiliza la gestión de los proyectos. Esta herramienta hace un análisis contemplando dos dimensiones:

- Administrativa
    - Transparencia y medidas del SMSC
- Desempeño
    - Eficiencia y eficacia

En resumen, el IGRP es una sumatoria de los factores ponderados:

IGPR = f (Transparencia, SMSCE, Eficiencia, Eficacia | θ)
Se definió un umbral ρ con el objetivo de segmentar el conjunto de datos en dos categorías.

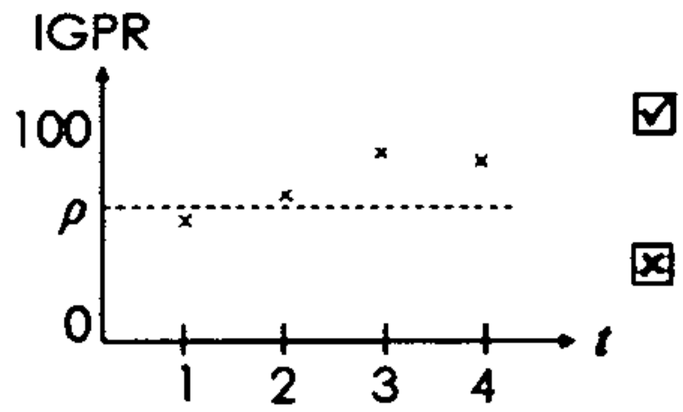


Figura 6. Umbral sobre el IGPR para la generación de etiquetas.

$$y(IGPR) = \begin{cases} 1, & IGPR \leq \rho\ ,\ \forall\ t \\ 0, & IGPR > \rho\ ,\ \forall\ t \end{cases}$$

Transformación del texto

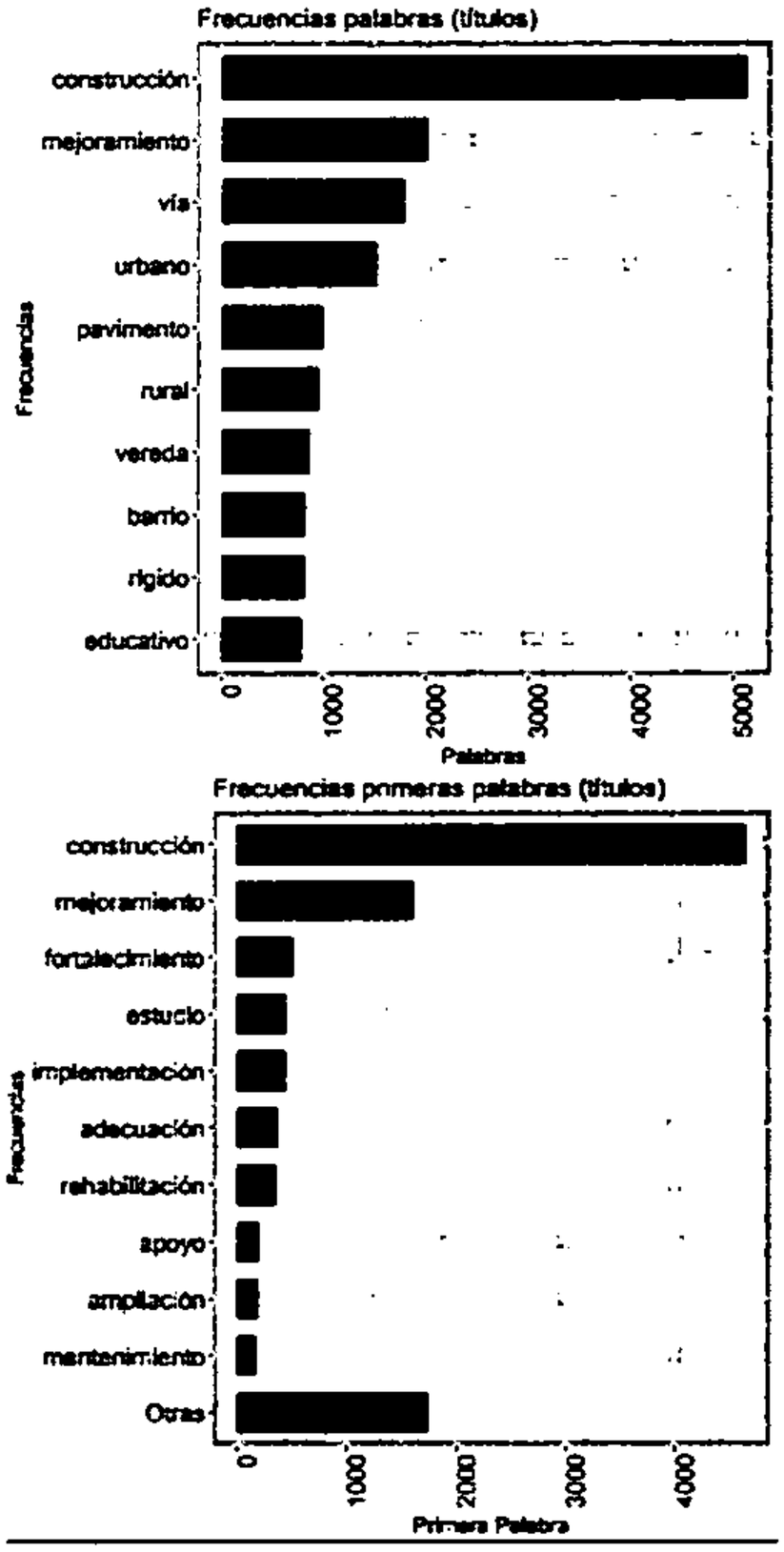
Se inicia con un pre-procesamiento estándar del texto para reducir los posibles errores que se hayan generado en la etapa de ingreso y captura de la información por parte de los proponentes de los proyectos. Este preprocesamiento consiste en los siguientes pasos:

- Segmentación por palaras para todo el texto
- Eliminar información no textual (caracteres especiales, puntuación, símbolos, palabras con longitud menor a dos caracteres)
- Lematizar todo el texto
- Corrección ortográfica.

Se continuo con una etapa de procesamiento específico, es decir, en función del campo a analizar (SGR), fuertemente influenciado por suposiciones que debemos hacer para segmentar eficientemente el análisis.

1. Títulos: Hipótesis - Primera palabra es el indicador más importante para determinar en que consiste el proyecto.
2. Cadena de valor
    a. Objetivos (General, específico): Primeras palabras
    b. Producto
    c. Actividades

Los resultados de esta etapa se pueden apreciar con mayor claridad en las siguientes graficas:


Frecuencias palabras (títulos)
construcción
mejoramiento
vía
urbano
pavimento
rural
vereda
barrio
rígido
educativo
Frecuencias
0
1000
2000
3000
4000
5000
Palabras
Frecuencias primeras palabras (títulos)
construcción
mejoramiento
fortalecimiento
estudio
implementación
adecuación
rehabilitación
apoyo
ampliación
mantenimiento
Otras
Frecuencias
0
1000
2000
3000
4000
Primera Palabra

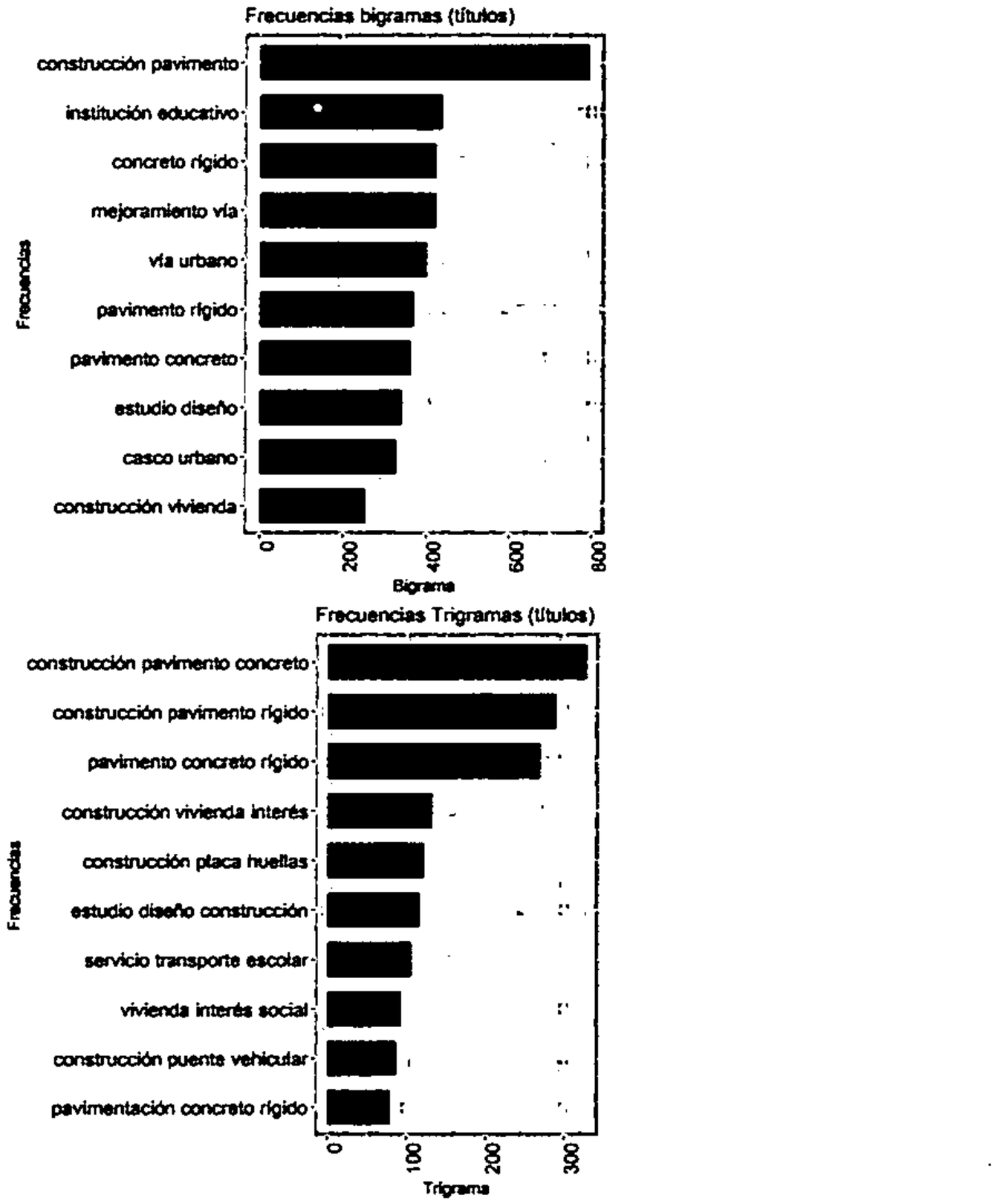


Figura 7. Resultados del análisis de frecuencias de palabras y n-gramas

Vectorización

Para convertir el texto en información cuantificable que pueda ser ingresada al modelo clasificador, fue necesario realizar un proceso de vectorización sobre el texto completo de cada proyecto. De esta forma cada proyecto es representado por un vector M-dimensional.

Se utilizo la vectorización a través de redes neuronales implementada a través del algoritmo Doc2Vec. El resultado de esta vectorización se representa en la siguiente figura a través de una conversión a un espacio bidimensional haciendo uso del algoritmo t-SNE.

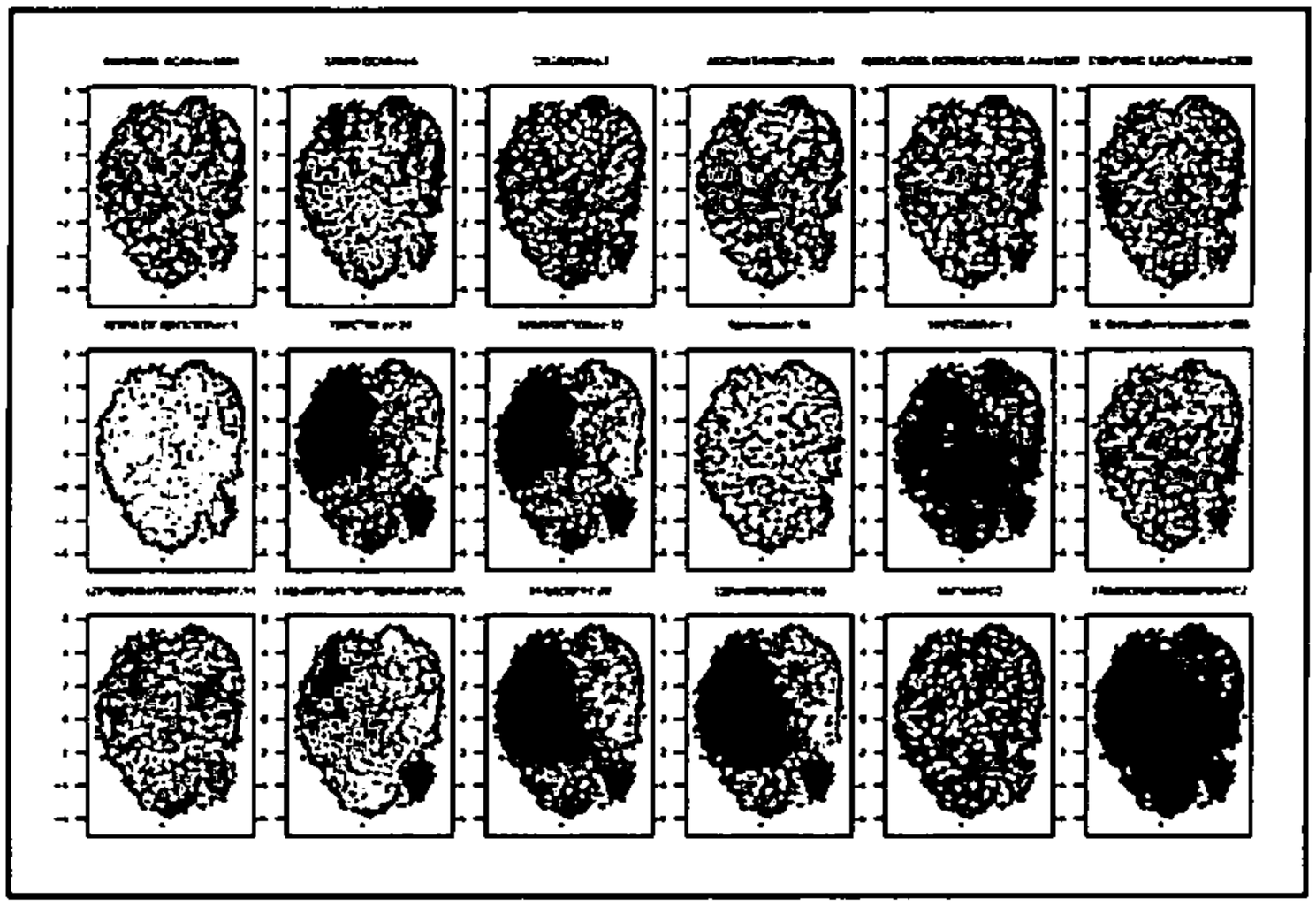


Figura 8. Visualización de los vectores en una representación 2D mediante el uso de t-SNE.

En la figura cada punto representa un proyecto y su color corresponde al valor de la variable categórica con la que se esté graficando.

Modelo clasificador

Para la construcción del modelo se hacen múltiples particiones del dataset, de tal forma que se divida en dos: un conjunto de entrenamiento con el 70% de los datos y un conjunto de prueba con el 30% de los datos.

Se realizaron pruebas con tres modelos: regresión logística, random forest y support vector machine (SVM). Se tomo el clasificador que en promedio entrega menor porcentaje de error en la clasificación del conjunto de prueba. El modelo final es un SVM con lo siguiente configuración:

- Kernel: RBF
- C = 1
- F1 score: 0,65
- MSE-training: 0,19
- MSE-test: 0,33

Conclusiones

- El análisis del texto complementa la información cuantitativa.
- Descubrir contexto a través de la vectorización.
- Evidenciar fallas "tempranas" en la formulación (correctivos anticipados).
- Priorización del seguimiento
    - Clasificar proyectos o entidades ejecutoras
    - Optimización de los recursos
- Clasificador adaptable a otras métricas de desempeño SGR