

DISEÑO DE SISTEMA DE CLASIFICACIÓN DE COMUNICACIONES RECIBIDAS VÍA CORREO ELECTRÓNICO POR EL CONSEJO FEDERAL DE GOBIERNO

Postgrado en Ciencias de la Computación

Facultad de Ciencias de la Universidad Central de Venezuela

Caracas, Venezuela

Eliezer Peñaloza

eliezer.oniros@gmail.com

Luis Acevedo

laar@protonmail.com

Resumen — A menudo, el Consejo Federal de Gobierno recibe correos electrónicos o comunicaciones de las distintas Organizaciones de Base del Poder Popular (O.B.P.P.) que hay en el país, muchas de éstas, requieren de cierta atención inmediata, debido a que suelen estar relacionadas con dificultades de registros de miembros, cambios de voceros, solicitudes de aperturas de cuentas, entre otras. Sin embargo, dada la cantidad de comunicaciones recibidas, no se puede dar respuesta pronta debido a la falta de personal con el que presenta el departamento, esto causa que algunos usuarios reenvíen sus correos electrónicos con sus peticiones en función de sus necesidades o problemática. Con ayuda de la experiencia del personal del departamento y la base de datos que poseen, han logrado crear y también organizar categorías que les permiten saber qué tipo de respuesta darle a cada correo recibido. A través de la minería de datos se busca detectar patrones claves en los cuerpos de texto de los correos electrónicos recibidos y se intentará construir una

aplicación que automatice este proceso de clasificación.

I. INTRODUCCIÓN

El Consejo Federal de Gobierno es el órgano encargado de la coordinación y la planificación de políticas y acciones para el desarrollo del proceso de descentralización y transferencia de competencias del Poder Nacional a los Estados, Municipios, Consejos Comunales y Asociaciones Vecinales, entre otras competencias; para este caso en específico, evaluación, aprobación y financiamiento de proyectos socio-productivos.

Con alrededor de cincuenta mil consejos comunales a nivel nacional, el proceso de clasificación de comunicaciones recibidas se torna cuesta arriba porque es llevado a cabo manualmente. Por medio de la minería de datos, se puede analizar y descubrir patrones que

ayuden a comprender mejor la relación que guardan entre sí la composición de los textos en las comunicaciones recibidas, para así, lograr establecer una categoría de forma automatizada a los nuevos correos recibidos.

II. COMPRENSIÓN DEL NEGOCIO

Objetivos del negocio.

El C.F.G. es una institución con mucha relevancia dentro del ámbito político y económico nacional, y por lo tanto muy solicitada. En vista de la masiva cantidad de comunicaciones que llegan a la institución, y el poco personal, se quiere automatizar el proceso de categorización de correos electrónicos entrantes, con ello acortar el tiempo de respuesta, de esta manera los usuarios podrán solventar sus dificultades para el desarrollo pleno de sus proyectos.

El personal del departamento ha hecho un trabajo previo de clasificación de forma manual, logrando además, crear y organizar los datos, junto con esto han creado un documento de respuestas para las distintas categorías que puede llegar a tener un correo. A partir de lo anterior se traza el siguiente objetivo:

Crear una aplicación que logre asignar de forma automática las categorías a los comunicados nuevos entrantes.

De acuerdo a lo mencionado anteriormente, se busca que al saber a qué categoría pertenece una determinada comunicación, dar respuesta a ésta de forma inmediata, sin necesidad de leer su contenido.

Desde el punto de vista de la minería de datos, se plantean las siguientes metas:

Construir un modelo, usando la minería de texto, que clasifique de forma automática los correos nuevos entrantes, todo esto, acorde a la información y clases existentes detalladas por los expertos.

1. Se quiere que el modelo tenga al menos un 70% de precisión.

En este sentido, se determinó la ruta a seguir para alcanzar los objetivos:

1. Obtención de los datos.
2. Limpieza de los datos.
3. Creación del corpus.
4. Generación de la vista minable.
5. Elección y aplicación de algoritmos.
6. Evaluación de los resultados.
7. Desarrollo y aplicación del producto dentro de la institución.

III. COMPRENSIÓN DE LOS DATOS

Recolección inicial de los datos

Los datos son los correos electrónicos recibidos en el C.F.G. almacenados

manualmente, es decir, son copiados y pegados en una hoja de cálculo, todo esto a partir del año 2017. Estos correos, en su mayoría, contienen errores ortográficos, y debido al proceso manual de almacenamiento son comunes los errores de formato, la información innecesaria y hasta elementos duplicados. Es por ello que se editó cada celda correspondiente al asunto interno (clase) para tener una mejor claridad de los datos y resultados.

Descripción de los datos

Los datos están compuestos por 6000 registros aproximadamente, y seis (6) atributos, los cuales son:

- *codigo_situr_obpp* (código asociado a la O.B.P.P.).
- *fecha_comunicacion* (fecha en la que llegó la comunicación).
- *correlativo_comunicacion* (identificador de la comunicación).
- *Asunto* (categoría asignada por la O.B.P.P.).
- *asunto_interno* (categoría asignada por la institución).
- *texto_comunicacion* (contenido del mensaje).

Para los fines del proyecto, sólo son relevantes los dos últimos atributos: *asunto_interno* y *texto_comunicacion*. Pues, de esta forma, el corpus se construyó con estas dos variables en formato csv.

Exploración de los datos

Existe una clase mayoritaria, correspondiente a “Solicitud de activación de usuario Sinco”, que representa más del 51% de los datos (ver figura 1). Este des-balance puede influir en el rendimiento del modelo causando sobre-entrenamiento.

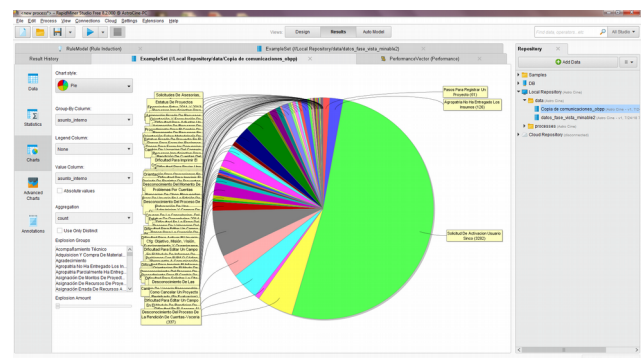


Figura 1.

Calidad de los datos

Debido a errores humanos, cuando se copian y pegan los correos a la hoja de cálculo, el contenido de éstos tiene las etiquetas de los formatos de tipo XML y JSON. Por otro lado, algunas instancias carecían de la clase (*asunto_interno*). En total, 338 instancias duplicadas fueron encontradas. Los errores ortográficos representaron una amenaza para la construcción del modelo, al igual que la alusión a otras direcciones de correos electrónicos inmersas en el cuerpo del mensaje (*texto_comunicacion*).

IV. PREPARACIÓN DE LOS DATOS

Selección de los datos

La hoja de cálculo con todos los correos posee seis (6) atributos, de los cuales sólo son relevantes dos de ellos: *asunto_interno* y *texto_comunicacion*. Se ignoró el atributo *asunto* por recomendación de los expertos, el resto de los atributos no aporta información relevante para este trabajo.

Limpieza del conjunto de datos

Las 338 instancias duplicadas fueron eliminadas, al igual que otras 44 que no poseían *asunto_interno* (clase). Los caracteres especiales en general fueron reemplazados o removidos, haciendo hincapié en las tildes, y las ñ (eñes) fueron reemplazadas por n (enes).

Construcción de los datos

A partir del atributo *texto_comunicacion* se generaron nuevos atributos, provenientes de las palabras contenidas en el cuerpo del mensaje. Para ello, con la ayuda del software Rapid Miner, se empleó la “tokenización” y, posteriormente, eliminación de palabras vacías (“stopwords”), mediante el uso de un diccionario que contiene artículos, conectores, nombres propios, direcciones de correo electrónico, entre otros. También, se lematizó usando el algoritmo “Snowball” adaptado al español.

Posterior al proceso anterior, se realizó una segunda limpieza al conjunto de datos, puesto que

se obtuvo 17847 nuevos atributos. Usando Rapid Miner, se aplicó el operador de ganancia de información, tomando en cuenta sólo aquellos datos que poseían un peso mayor a 0,03, reduciéndose a 92 el número de atributos, ver figura 2.

Row No.	asunto_interno	abon	acta	activ	activacion	adjunt	administr	agropatr	aparec
1	Pasos Para...	0	0	0	0	0	0	0	0
2	Agropatria No...	0	0	0	0	0	0.022	0.038	0
3	Solicitud De...	0	0	0	0.040	0	0	0	0
4	Solicitud De...	0	0	0	0	0	0	0	0
5	Solicitud De...	0	0	0	0.142	0	0	0	0
6	Solicitud De...	0	0	0	0	0	0	0	0
7	Solicitud De...	0	0	0	0	0	0	0	0
8	Solicitud De...	0	0	0	0	0	0	0	0
9	Solicitud De...	0	0	0	0	0	0	0	0
10	Desconocim...	0	0	0	0.135	0	0	0	0
11	Dificultad En...	0	0	0	0.088	0	0	0	0
12	Solicitud De...	0	0	0	0	0	0	0	0
13	Solicitud De...	0	0	0	0	0	0	0	0
14	Solicitud De...	0	0	0	0	0	0	0	0
15	Solicitud De...	0	0	0	0	0	0	0	0
16	Solicitud De...	0	0	0	0	0	0	0	0
17	Solicitud De...	0	0	0	0	0	0	0	0
18	Solicitud De...	0	0	0	0	0	0	0	0
19	Solicitud De...	0	0	0	0	0	0	0	0
20	Solicitud De...	0	0	0	0	0	0	0	0

Figura 2.

V. MODELADO

Selección de técnicas de modelado

Se usaron distintos algoritmos: Decision Tree, Rules Induction, Random Forest y Deep Learning. Esto, con intención de cotejar los resultados y construir el modelo con el que mejor se adapte a la necesidad.

Generación de diseños de pruebas

El rendimiento de los modelos fue evaluado aplicándose la técnica de validación cruzada, tomándose como criterio el valor de precisión (*accuracy*). El resultado de estas pruebas se puede apreciar en la Figura 3.

Construcción del modelo

Como se refleja en la Figura 2, el algoritmo con mejor rendimiento fue *Deep Learning*. Sin embargo, cabe destacar que el algoritmo *Decision Tree* sólo pudo predecir las instancias con la categoría predominante.

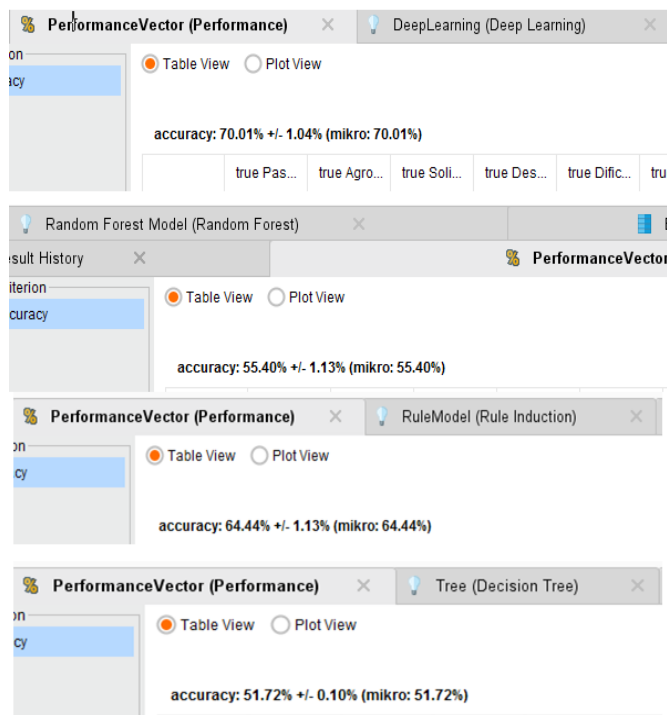


Figura 3.

VI. EVALUACIÓN

Evaluación de los resultados

De acuerdo con el criterio de éxito establecido, se logró la meta de precisión de al menos un 70% (*Deep Learning*). La precisión de los demás algoritmos puede ser mejorada con la incorporación de más datos, en especial las categorías con menos registros.

Con pruebas hechas a menor escala, y en aquellas categorías con número de instancias superior a 30, los algoritmos alcanzan una precisión (*accuracy*) entre 87% y 92%.

Determinación de pasos a seguir

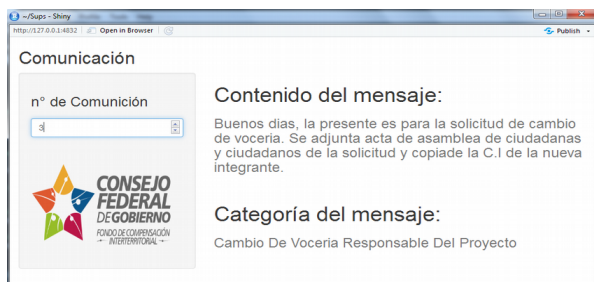
Uno de los factores clave en este problema está en el des-balance de las clases. Por lo que una opción es utilizar datos sintéticos para poder ayudar a mejorar el rendimiento, o esperar a que el departamento suministre más datos.

Otra posibilidad es crear varios modelos para que cada uno identifique una categoría en específico.

VII. PUESTA EN PRODUCCIÓN

Una vez obtenidos los resultados, se pasó inmediatamente a la elaboración de una aplicación, que facilitará la apreciación del contenido del mensaje y, a su vez, su clasificación en tiempo real, sólo se preprocesaron los datos nuevos para transformarlos, de modo que tengan el mismo formato que la vista minable, además del número de atributos. Se realizó una función para ello, además de calcular el TF-IDF usando las palabras del total registrado en los datos. Todo esto fue hecho con la herramienta R, mediante la biblioteca *Shiny* se creó la aplicación, ver figura 4. En donde el usuario puede desplegar el

mensaje, a medida que mueve el número al que está asociado.



The screenshot shows a web browser window with a Shiny application. The title bar says "Shiny". The address bar shows "http://127.0.0.1:4832". The page has a "Publish" button in the top right. The main content area is titled "Comunicación". On the left, there is a section labeled "n° de Comunicación" with a text input field containing the number "4". Below this is the logo of the "CONSEJO FEDERAL DE GOBIERNO FONDO DE COMPENSACIÓN INTERTERITORIAL". On the right, there are two sections: "Contenido del mensaje:" with the text "Buenos días, la presente es para la solicitud de cambio de vocería. Se adjunta acta de asamblea de ciudadanas y ciudadanos de la solicitud y copiado la C.I de la nueva integrante." and "Categoría del mensaje:" with the text "Cambio De Vocería Responsable Del Proyecto".

Figura 4.

VIII. CONCLUSIONES

El modelo puede estar sujeto a los errores ortográficos cometidos por el usuario que envía el correo, y esto afectaría su desempeño. Las desproporciones tan marcadas entre categorías, en especial “Solicitud de Activación de Usuario Sinco”, afecta al modelo dándole un sobre entrenamiento por este sesgo. El modelo funciona mejor con el algoritmo Deep Learning, dado que supo encontrar mejor los patrones intrínsecos de los mensajes. El modelo puede mejorar el rendimiento agregándole más registros.