Análisis de registros de mantenimiento de centrales de generación de energía con técnicas de procesamiento de lenguaje natural

Andrés Alonso Ocampo Dávila
aaocampod@eafit.edu.co

Director
Carlos Andrés Salazar Martínez
csalaz22@eafit.edu.co

Escuela de Ciencias Aplicadas e Ingeniería
Universidad EAFIT, Medellín, Colombia

Índice

Índice de figuras

Resumen

Las actividades de mantenimiento se documentan en los sistemas de información de las organizaciones, incluyendo detalles sobre las tareas ejecutadas, los equipos intervenidos y su condición. Una parte de la información registrada corresponde a textos libres, no estructurados, ingresados por los técnicos de mantenimiento. Estos textos se caracterizan por combinar lenguaje técnico, abreviaciones y jerga específica con una redacción informal y presentar numerosos errores gramaticales y ortográficos. Dado el volumen y características de estos textos, para extraer de ellos información relevante para la evaluación de la condición de los activos se requiere de su revisión en forma manual. Este procedimiento de revisión, aunque puede resultar efectivo, es extremadamente costoso en términos del tiempo que demanda por parte de personal técnico capacitado. Es por esta razón que esta potencial fuente de conocimiento comúnmente se desaprovecha y no contribuye como debería a mejorar el proceso de mantenimiento y el desempeño de los activos.

Este trabajo presenta el uso de diversas técnicas de procesamiento de lenguaje natural y modelos de aprendizaje automático para obtener datos estructurados a partir de los textos libres no estructurados, correspondientes a los registros de mantenimiento de un conjunto de centrales de generación de energía. De esta forma se busca validar e identificar aquellas técnicas que permitan la extracción de información estructurada a partir de los textos libres, que sirva como insumo para posteriores análisis de confiabilidad, mantenibilidad y disponibilidad.

Los resultados demuestran que al aplicar técnicas de procesamiento de lenguaje natural combinadas con modelos de aprendizaje automático es posible etiquetar los datos no estructurados de los registros de mantenimiento, identificando el defecto en el equipo, asociado con el componente con falla, y la causa de la falla, proporcionando información acerca del historial de condición de los activos, con lo cual en última instancia es posible soportar la gestión de la condición de los equipos y la toma de decisiones de mantenimiento.

1. Planteamiento del problema

Entre los procesos de negocio de una empresa prestadora de servicios públicos se incluye el mantenimiento de la infraestructura requerida para la generación de energía eléctrica.

Como parte de la gestión administrativa de las actividades de mantenimiento de las centrales de generación de energía eléctrica, se realiza la documentación correspondiente en los sistemas de información de la empresa. Esta documentación incluye el registro de las tareas ejecutadas, los equipos intervenidos y su condición antes y después de la intervención. Aunque en su mayoría la información registrada corresponde a datos estructurados, también se incluyen textos libres no estructurados, ingresados por los técnicos de mantenimiento describiendo los trabajos realizados.

A diferencia de otros campos de datos, las entradas de texto permiten a los usuarios escribir libremente lo que consideren necesario para especificar el trabajo ejecutado, con cualquier conjunto de palabras y sin límites en su extensión. Por lo tanto, estos textos se caracterizan por combinar lenguaje técnico, abreviaciones y jerga específica, con una redacción informal y por presentar numerosos errores gramaticales y ortográficos.

Como ocurre en cualquier organización intensiva en activos[1], el monitoreo a la gestión y desempeño del mantenimiento es una actividad de especial importancia, que se basa principalmente en los históricos de mantenimiento. Por las características que ya se han mencionado, para extraer información relevante de los textos libres en los registros de mantenimiento es necesario un proceso de revisión "manual", es decir, su lectura uno a uno. Este proceso, aunque puede ser efectivo, resulta extremadamente costoso en términos del tiempo que demanda por parte de personal técnico capacitado, con la experiencia y conocimiento requeridos para ejecutarla, si se considera además la cantidad de información generada, que supera los 4000 registros de trabajo por año para las actividades correctivas (destinadas a atender averías) solamente.

Es por las razones expuestas que la información contenida en los textos libres no estructurados de los registros de mantenimiento, aunque constituye una fuente potencial de conocimiento para los equipos responsables de la gestión de los activos, comúnmente se desaprovecha, y no contribuye como debería a evaluar y mejorar el proceso de mantenimiento y el desempeño de los equipos.

El caso analizado en el marco de este proyecto corresponde a una empresa prestadora de servicios públicos en Colombia, la cual se ha anonimizada y renombrada como La Empresa por razones de protección y confidencialidad de los datos.

[1]Una industria intensiva en activos es aquella que requiere una inversión significativa en activos físicos, como maquinaria, equipos, instalaciones y propiedades, para operar y generar ingresos. Estos activos suelen tener un ciclo de vida prolongado y están diseñados para ser utilizados a largo plazo en la producción de bienes o servicios

2. Justificación

Como ocurre en el caso de las industrias intensivas en activos, para los activos que hacen parte de la infraestructura para la generación de energía, una gran cantidad de información se encuentra almacenada en sus registros históricos de mantenimiento. Parte de esta información la constituyen textos no estructurados, que redactan los usuarios de forma libre, describiendo los trabajos de mantenimiento realizados, la condición del activo y otros aspectos que pueden considerarse relevantes para evaluar el desempeño de los equipos a lo largo del tiempo (Stenström et al., 2015).

Dadas las características y volumen de estos textos libres, para extraer de ellos información relevante se requiere de un proceso de revisión "manual", que resulta considerablemente costoso por las horas de personal capacitado necesarias para su procesamiento. Lo anterior hace que esta información, potencialmente valiosa, en la mayoría de los casos se desaproveche y no sea usada, como se supone, para apoyar la toma de decisiones en el proceso de mantenimiento y la gestión de los activos.

Es así como se hace necesario implementar procedimientos y herramientas que, haciendo uso de técnicas de procesamiento de lenguaje natural y modelos de aprendizaje automático, permitan obtener datos estructurados a partir de los textos libres no estructurados incluidos en los registros de mantenimiento. De esta forma se logrará facilitar el proceso de revisión de dichos textos y la extracción de información que servirá como apoyo para posteriores análisis de confiabilidad, mantenibilidad y disponibilidad y como soporte a la toma de decisiones en mantenimiento y la evaluación de la condición de los equipos.

Específicamente, en el contexto descrito, resulta de especial interés obtener un flujo de trabajo automatizado y eficiente que permita extraer datos estructurados a partir los textos libres de los registros de mantenimiento, es decir, identificar elementos específicos, como el defecto en el equipo y la causa de la falla. Esto permitirá aprovechar de mejor manera la información contenida en los textos, reduciendo el tiempo de procesamiento, facilitando la identificación de patrones de fallas recurrentes o de alto impacto y apoyando la definición de la estrategia de mantenimiento predictivo y la gestión de los indicadores de mantenimiento.

3. Objetivos

3.1. Objetivo General

Analizar el uso de diversas técnicas de procesamiento de lenguaje natural y modelos de aprendizaje automático para obtener datos estructurados a partir de los textos libres, no estructurados, correspondientes a los registros de mantenimiento, que sirvan como insumo para posteriores análisis de confiabilidad, mantenibilidad y disponibilidad de un conjunto de centrales de generación de energía.

3.2. Objetivos Específicos

- Definir un conjunto de procedimientos y tareas a ejecutar sobre los textos crudos extraídos de la base de datos de registros de mantenimiento para su preprocesamiento y preparación para el posterior análisis.
- Aplicar herramientas de procesamiento de lenguaje natural y modelos de aprendizaje automático sobre los textos preparados, para identificar en el conjunto de datos no estructurados el defecto en el equipo y la causa inmediata de la falla documentadas en los registros de mantenimiento.
- Desarrollar un *pipeline* de datos, que incluyendo las etapas de preprocesamiento y la aplicación de los modelos, permita extraer de los textos libres de los registros de mantenimiento elementos como el defecto en el equipo y la causa de falla.

4. Marco Teórico

En esta sección se presenta una revisión de los temas relacionados con modelos de aprendizaje automático usados en tareas de procesamiento de lenguaje natural de textos, con aplicación al análisis de datos no estructurados que hacen parte registros de mantenimiento.

El mantenimiento es el conjunto de acciones técnicas, administrativas y de gestión realizadas a lo largo del ciclo de vida de un activo, con el fin de conservarlo o devolverlo, a un estado en el cual pueda desarrollar la función para la que es requerido (Asociación Española de Normalización y Estandarización, 2002). Como parte de las actividades de gestión, se realiza el seguimiento y la evaluación de los resultados del mantenimiento. Para esto son utilizadas tecnologías de la información, incluyendo sistemas de planificación de recursos empresariales (ERP), sistemas de gestión de mantenimiento (MMS) y sistemas de gestión de activos empresariales (EAM). Haciendo uso de estos sistemas, se almacenan los datos de los trabajos de mantenimiento, denominados como registros de mantenimiento, reportes de mantenimiento u órdenes de trabajo. Estos datos permiten recopilar la historia de la condición de un sistema de activos, incluyendo detalles sobre inspecciones, diagnósticos y correctivos ejecutados (Stenström et al., 2015; Brundage et al., 2021).

En organizaciones intensivas en activos, el monitoreo del desempeño y los costos del mantenimiento es una actividad de especial importancia, que se apoya principalmente en las órdenes de trabajo (Stenström et al., 2015). Estas contienen información detallada sobre los activos, sus componentes, condición y mecanismos de falla, que puede ser aprovechada por los equipos de mantenimiento para el pronóstico de fallas, el análisis de causa raíz y la toma de decisiones basada en datos (Bhardwaj et al., 2022). Además estos datos son esenciales para la evaluación y gestión de los indicadores de confiabilidad, disponibilidad y mantenibilidad de los activos.

En los sistemas de información, los registros de mantenimiento son gestionados a través de una interfaz gráfica de usuario. Normalmente, cada registro contiene datos del activo, la descripción y causa de falla y la actividad de mantenimiento ejecutada. La interfaz de usuario dispone de diferentes tipos de campos para el ingreso de esta información, incluyendo listas desplegables, casillas de verificación, cuadros de lista y campos de entrada de texto. A diferencia de los otros campos de datos, el campo de entrada de texto permite que el usuario ingrese libremente una respuesta con aquello que considere necesario para detallar las situaciones encontradas y el trabajo realizado, con cualquier conjunto de palabras, en cualquier cantidad, lo cual se traduce en un texto "crudo" relativamente desestructurado (Brundage et al., 2021).

La extracción de información a partir de los textos en los registros de mantenimiento es una tarea compleja debido a que estos datos se presentan en formato no estructurado, lo que implica que su análisis requiere de un esfuerzo importante. Igualmente, estos textos contienen lenguaje técnico y jerga específica y presentan errores de digitación, ortográficos o gramaticales, lo que genera una reducción de la calidad de los datos de mantenimiento y una restricción significativa para su análisis. Así, con el fin de disponer herramientas para la extracción de información y generar conocimiento valioso a partir de los registros de mantenimiento, se han desarrollado diferentes enfoques, incluyendo de minería de texto basadas en técnicas de procesamiento de lenguaje natural (Sala et al., 2022).

De manera general, el Procesamiento de Lenguaje Natural o NLP (*Natural Language Processing*) hace referencia a cualquier tipo de manipulación computarizada del lenguaje natural, es decir, del lenguaje que es usado por los humanos para su comunicación cotidiana (Bird, 2009). Más específicamente, el NLP corresponde al campo de la lingüística y las ciencias de la computación que se ocupa de darle a las computadoras capacidades y comprensión del lenguaje humano. Esto incluye tareas como la extracción de información y la clasificación o agrupación de textos, como los que se encuentran en los registros de mantenimiento (Öztürk et al., 2022).

El avance actual en el área del NLP, incluyendo el auge de las técnicas de aprendizaje profundo, permite que el análisis automático de textos libres sea viable y que se haya mejorado considerablemente su precisión. En consecuencia, aplicaciones convencionales del NLP, como el análisis de sentimientos y la clasificación de opiniones, se han convertido en algo relativamente trivial. Sin embargo, las aplicaciones de NLP para el procesamiento de datos provenientes de sistemas de gestión de mantenimiento son todavía poco comunes (Deloose et al., 2023).

No obstante las dificultades que reviste el procesamiento de los textos en los registros de mantenimiento, se pueden encontrar aplicaciones de NLP en este dominio. Entre ellas, la predicción de los códigos de falla asociados a los reportes de averías a partir de los textos libres (Zhang et al., 2020; Deloose et al., 2023) o la mejora en el etiquetado (metadatos para categorizar el tipo de fallas, sus causas y las acciones correctiva) de los registros de mantenimiento, identificando las notificaciones mal etiquetadas o con etiquetas ambiguas

(Arif-Uz-Zaman et al., 2017; Deloose et al., 2023). En ambos casos, se utilizaron modelos de espacio vectorial para la representación de los textos y modelos supervisados como máquinas de soporte vectorial, *naive Bayes* y *Random Forest* para su clasificación.

Los modelos de aprendizaje profundo también han sido usados en aplicaciones de NLP. Las redes neuronales recurrentes (RNN) y las redes neuronales convolucionales (CNN) son las opciones normalmente elegidas para este tipo de problemas (Deloose et al., 2023). Más recientemente, los *transformers* se están ubicando como el estado del arte en las aplicaciones de NLP y modelos como BERT *(Bidirectional Encoder Representations from Transformers)* han sido utilizados exitosamente para la clasificación de registros de mantenimiento y obtener datos estructurados y conocimiento a partir de textos libres, incluso en problemas con conjuntos de datos con clases desbalanceadas (Usuga-Cadavid et al., 2020, 2022).

Otros enfoques han abordado el problema de extraer información descriptiva de los textos libres de los registros de mantenimiento, buscando identificar las partes o componentes afectados, el mecanismo de falla o las condiciones operativas. El objetivo propuesto en estos casos es identificar grupos de trabajos de mantenimiento similares, utilizando dos fuentes de datos: los propios registros de mantenimiento y la taxonomía de los equipos. Haciendo uso de modelos de aprendizaje no supervisado y al combinar información semántica y taxonómica para la segmentación de los datos no estructurados de los textos, se logra la extracción precisa de conocimiento basado en el contexto documentado en los registros de mantenimiento (Bhardwaj et al., 2022).

Por otra parte, el Reconocimiento de Entidades Nombradas, NER *(Named Entity Recognition)*, ha sido utilizado para abordar problemas en dominios que presentan similitudes con el análisis de los registros de mantenimiento. El NER es una técnica de extracción de información orientada a identificar y clasificar las palabras de un documento en unas categorías predeterminadas llamadas "Entidades Nombras", NEs *(Named Entities)*, lo que lo convierte en una herramienta importante en muchas de las áreas de aplicación del NLP (Raja et al., 2019).

Entre los enfoques clásicos del NER se encuentra el basado en reglas. Los sistemas de NER que se desarrollan bajo este paradigma identifican entidades aplicando extensos conjuntos de reglas que, en general, se componen de patrones basados en características gramaticales, sintácticas y ortográficas, en combinación con diccionarios especializados. Este tipo de modelos demuestran un desempeño superior en dominios específicos y destaca por su capacidad de detectar entidades complejas, que frecuentemente representan un desafío para los modelos de aprendizaje. No obstante, se ven limitados en su portabilidad y robustez y conllevan costos significativos en el mantenimiento de las reglas, incluso ante modificaciones ligeras en los datos, lo que hace que no se adaptan necesariamente bien a nuevos dominios y lenguajes (Raja et al., 2019).

Entre las aplicaciones de NER basado en reglas se incluye la extracción de elementos

específicos, pertenecientes a categorías preestablecidas, a partir de textos libres no estructurados, incluidos en registros de dominios diversos, pero que resultan similares en el tipo de problema que se quiere resolver para los registros de mantenimiento. Estos sistemas han sido empleados con éxito en múltiples contextos, tales como la identificación de factores desencadenantes y consecuencias en informes de accidentes de trabajo (Tixier et al., 2016), la obtención de indicadores de inestabilidad residencial a partir de datos no estructurados en historias clínicas electrónicas (Hatef et al., 2022) y la extracción de causas, consecuencias y riesgos de los textos libres de los reportes de ocurrencia de eventos en sistemas aeronáuticos (Ricketts et al., 2022). En todos estos casos un sistema de NER fue utilizado para aplicar un conjunto de reglas o patrones definidos en función del conocimiento experto en el dominio, logrando identificar con éxito entre el documento los elementos particulares de interés para el problema específico, lo que resultó en la disminución significativa del esfuerzo involucrado en la revisión manual de los documentos.

5. Metodología

Para el desarrollo del proyecto se ha utilizado como base la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), aquella que se usa de forma más generalizada, en su estructura original o con algunas adaptaciones, en aplicaciones de minería de datos y proyectos ciencia de datos de diferentes dominios (Martínez-Plumed et al., 2021). Esta metodología propone seis pasos, que se detallarán más adelante: Entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue (Chapman et al., 2000).

5.1. Entendimiento del negocio

Como se ha mencionado, el contexto en el cual se desarrolla este proyecto es el proceso de mantenimiento de la infraestructura de generación de energía de una empresa de servicios públicos.

Como resultado de la gestión administrativa de las tareas de mantenimiento, se documentan las actividades realizadas en un EAM (*Enterprise Asset Manager*), el sistema de información dispuesto en la organización para este fin. Este registro histórico posee una relevancia significativa como fuente de información primaria para el seguimiento y control a la gestión del mantenimiento y para el monitoreo al desempeño y la condición de los activos que hacen parte del proceso de generación de energía. En este sentido, las actividades correctivas, aquellas destinadas a la atención de las averías en los equipos, resultan de especial interés. La documentación de estas actividad puede incluir información acerca de elementos específicos como el defecto identificado en el equipo, el componente fallado y la causa inmediata de la falla, como se ilustra en la figura 1. Estos datos, adecuadamente tabulados, pueden ser usados como insumo para el desarrollo de análisis orientados a estimar la condición de los equipos, identificar patrones de falla recurrentes o de alto impacto (en costo o tiempo) y a soportar las estrategias de mantenimiento preventivo y predictivo.

Figura 1: Ejemplo de elementos de información en los textos libres

Texto libre	Elementos de información		
	Defecto en el equipo	Componente	Causa
SE REALIZA INSPECCIÓN AL VENTILADOR EXTRACTOR E2 EL CUAL SE REPORTÓ CON RUIDO. ESTE EQUIPO SOLO PRESENTA PROBLEMA EN LAS DOS BANDAS VA-75 QUE SE ENCUENTRAN MUY DESGASTADAS Y A PUNTO DE REVENTARSE. EL EQUIPO SE DEJA EN FUNCIONAMIENTO.	RUIDO	BANDAS	DESGASTE
SE INSPECCIONA FUGA. SE ENCUENTRA ROTURA EN LA TUBERIA POR CORROSION.	FUGA	TUBERIA	ROTURA
SE REVISA EL REGULADOR; YA QUE SE DISPARÓ LA UNIDAD POR BAJO NIVEL DE ACEITE. SE ENCUENTRA LA BOMBA PRINCIPAL PARADA, DURANTE LA REVISIÓN SE ENCONTRÓ QUE EL MICROSUICHE DE ACCIONAMIENTO DE LA BOMBA ESTABA PEGADO, POR ESTE MOTIVO NO ARRANCÓ LA BOMBA PRINCIPAL	BAJO NIVEL ACEITE	BOMBA	MICROSUICHE PEGADO

De acuerdo con lo anterior, resulta de especial interés para las áreas responsables del mantenimiento de la infraestructura para la generación de energía, contar con un mecanismo que permita extraer en forma automática y eficiente los datos estructurados a partir los textos libres de los registros de mantenimiento, es decir, identificar elementos específicos de interés, como los antes citados.

5.2. Entendimiento de los datos

Para el desarrollo del proyecto se ha dispuesto de un conjunto de datos conformado por aproximadamente 13.400 registros de mantenimiento, donde se han documentado trabajos de tipo correctivo únicamente (atención de fallas). Estos registros se encuentran almacenados en la base de datos transaccional (RDBMS Oracle) del sistema de información (EAM) que soporta el proceso de mantenimiento de los activos de la infraestructura de generación de energía.

Los datos corresponden a 21 centrales de generación de energía de la Empresa y han sido recolectados en el periodo comprendido entre agosto de 2018 y febrero de 2023. Cada registro se compone de un conjunto de variables que incluye un identificador del registro y la fecha de creación del mismo, una descripción corta de la actividad, datos relativos al elemento mantenido (sistema, subsistema, activo, componente) y el campo de texto donde, de manera libre, el técnico ha descrito las situaciones encontradas y el trabajo realizado (figura 2). Como se ha establecido previamente, es sobre este último campo, identificado como “LDTEXT”, sobre el cual se ha enfocado el desarrollo de este proyecto.

5.3. Preparación de los datos

Durante esta etapa se han ejecutado las tareas requeridas para construir y ajustar el conjunto de datos final sobre el cual se realizaron los análisis y se desarrollaron los modelos. La preparación de los datos, específicamente los textos de mantenimiento que fueron analizados, ha incluido las etapas que se describen a continuación.

Figura 2: Campos del registro de mantenimiento

Variable	Descripción de la variable	Tipo de dato
REPORTDATE	Fecha de creación de la orden de trabajo	DATE
WONUM	Número de la orden de trabajo	VARCHAR (10)
WORKTYPE	Tipo de trabajo (correctivo, preventivo)	VARCHAR (5)
DESCRIPTION	Descripción de la orden de trabajo	VARCHAR (100)
LOCATION	Sistema (sección, subsistema) de la orden de trabajo	VARCHAR (35)
ASSETNUM	Activo de la orden de trabajo	VARCHAR (20)
STATUS	Estado de la orden de trabajo	VARCHAR (16)
FAILURECODE	Código de anomalía	VARCHAR (15)
PROBLEMCODE	Código de problema	VARCHAR (15)
WORKLOGID	Identificador único del registro de texto	NUMBER
LDDESCRIPTION	Resumen del registro de texto	VARCHAR (100)
LDTEXT	Detalle del registro de texto	CLOB
ANCESTOR	Instalación (planta, circuito)	VARCHAR (35)

I **Extracción e integración de los datos**

Mediante consultas SQL a la base de datos Oracle del sistema de transaccional, EAM, se ha extraído el conjunto de datos conformado por 13425 registros de trabajo de mantenimiento, los cuales, como se mencionó, incluyen los textos "crudos" objeto de análisis.

Los datos extraídos se almacenaron como archivos planos, en formato *.csv*, para ser procesados.

II **Selección de variables**

El conjunto de datos de los registros de trabajo, conformado por 13 variables (figura 2), fue segmentado en dos partes. La primera parte incluye el campo "LDTEXT", que contiene el texto sin procesar sobre el cual se aplicaron las técnicas NLP, junto con el identificar único del registro de texto "WORKLOGID". La segunda parte comprende las variables restantes, que contienen los datos estructurados que caracterizan cada una de las actividades de mantenimiento realizadas y que posteriormente permiten relacionar los resultados obtenidos de la minería de texto con dichos trabajos y los activos sobre los que fueron ejecutados.

III **Limpieza de los datos**

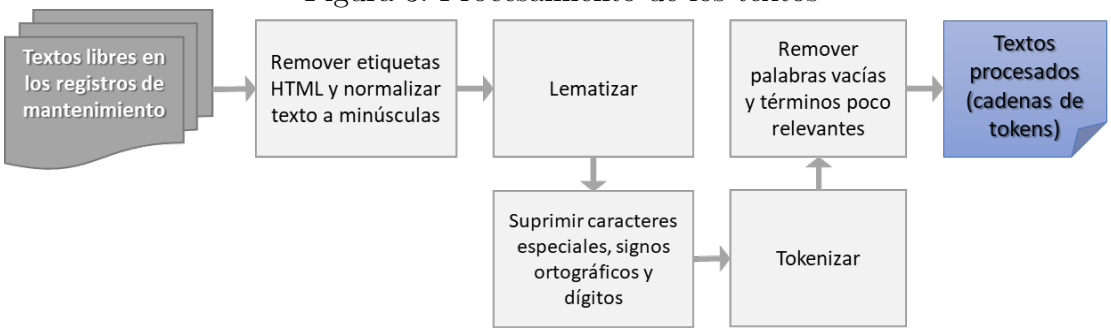
Con el fin de llevar los datos no estructurados (texto) a un dominio numérico que permitiera un modelado adecuado se definieron y aplicaron algunas funciones en Python para realizar el procesamiento y limpieza de los textos "crudos". También fueron utilizadas funciones disponibles en los paquetes *NLTK* (Bird, 2009) y *spaCy* Honnibal et al. (2020).

En primer lugar, se excluyeron todos los registros en los cuales el campo de interés, "LDTEXT", se encontrara vacío. A continuación, se ejecutaron diversos procedimientos de NLP para la limpieza de los textos "crudos", haciendo uso de expresiones regulares (*RegEx*) y funciones disponibles en los paquetes NLTK y spaCy. Las acciones realizadas incluyeron, en su orden (figura 3),

- remoción de etiquetas HTML,
- normalización a minúsculas,
- "lematización"[2],
- supresión de caracteres especiales, tildes, diéresis, eñes, virgulillas y dígitos,
- "tokenización"[3]
- remoción de "palabras vacías" o que agregan poco valor (*stop words*) y de términos poco relevantes para el análisis (como nombres de personas o lugares)

No se aplicó el procedimiento de *stemming* por el pobre desempeño observado al aplicarse sobre los textos analizados, en el dominio específico.

Figura 3: Procesamiento de los textos



De esta manera, el texto libre en cada uno de los registros de mantenimiento se ha representado como una lista de *tokens* o palabras procesadas (figura 4).

Con el conjunto de *tokens* resultantes del procesamiento se conformó la bolsa de palabras (BoW, *Bag of Words*), que fue usado en la representación de documentos y la construcción de los modelos.

Figura 4: Resultado del procesamiento de los textos

Texto libre (incluye etiquetas HTML)	Tokens procesados
Se realiza adecuación de los múltiples de distribución de mangueras de los porta escobillas, lo que consiste en instalarlos 20 cm más abajo de la ubicación original para que la succión tenga un mejor direccionamiento, se acondicionan todas las mangueras y se cambian la #1, #5 y #11 por deterioro, se realizan pruebas de funcionamiento. Se entrega a operación en condiciones normales. <!-- RICH TEXT -->	['realizar', 'adecuacion', 'multiple', 'distribucion', 'manguera', 'portar', 'escobilla', 'consistir', 'instalar', 'mas', 'abajo', 'ubicacion', 'original', 'succion', 'tener', 'mejor', 'direccionamiento', 'acondicionar', 'manguera', 'cambiar', 'deterioro', 'realizar', 'prueba', 'funcionamiento', 'entregar', 'operacion', 'condicion', 'normal']

[2]Proceso que asigna las diversas formas de una palabra a la forma canónica o forma de cita de la palabra, también conocida como lexema o lema. (Bird, 2009)

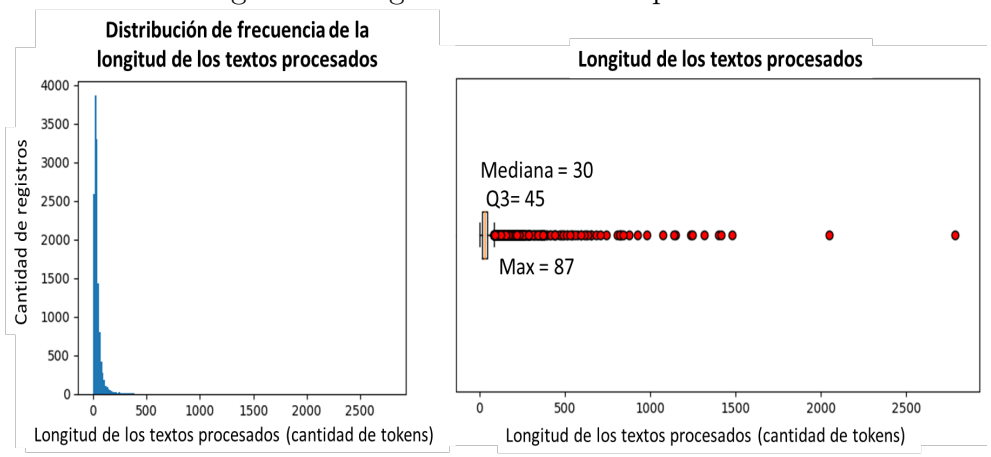
[3]Segmentación de un texto en unidades lingüísticas básicas e identificables o *tokens* (como palabras) que constituyen un fragmento de datos de lenguaje (Bird, 2009)

IV Análisis exploratorio de los datos

En esta etapa se realizó un análisis descriptivo de los textos procesados (cadenas de *tokens*), con el fin de identificar características relevantes en el conjunto de datos que pudieran orientar el modelado.

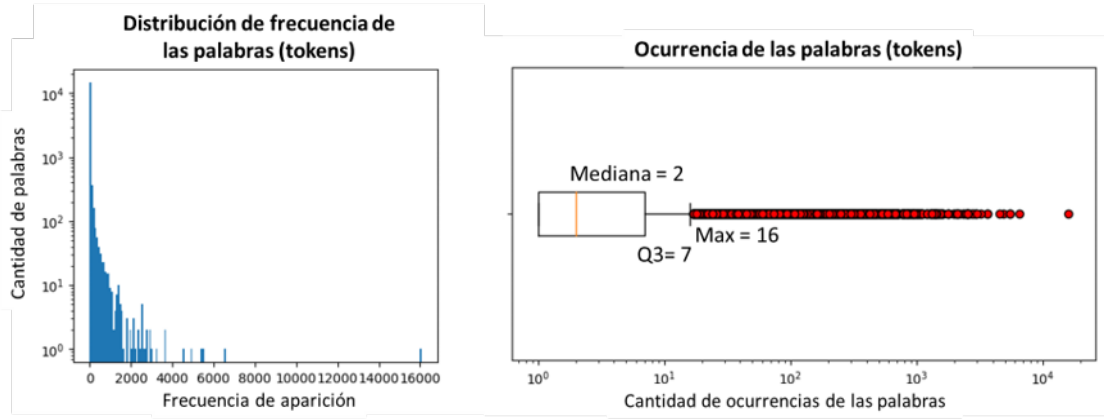
Respecto a la longitud de los textos procesados, se ha podido determinar que las cadenas de *tokens* resultantes de la transformación de los textos sin procesar varían desde un mínimo de un token hasta en ocasiones superar los 2000 *tokens*. No obstante, se destaca que aproximadamente el 75 % de los registros cuentan con 45 *tokens* o menos, es decir que, en general, los textos en los registros de trabajo tienden a ser relativamente cortos (figura 5).

Figura 5: Longitud de los textos procesados



El conjunto de textos procesados se compone de un total de 49159 *tokens*, lo que resulta en una BoW con un tamaño de 15563 *tokens*. Al analizar la distribución de frecuencia de las palabras, se destaca que de los 15563 *tokens* presentes en la BoW, aproximadamente el 50 % de ellos aparece en los textos en no más de dos ocasiones, mientras que el 75 % se repite en ocho ocasiones o menos. Así, el análisis las cadenas de *tokens* permite establecer que la mayoría de las palabras en el conjunto de datos tienen una frecuencia de aparición bastante baja, lo que sugiere una diversidad léxica considerable en los textos procesados (figura 6).

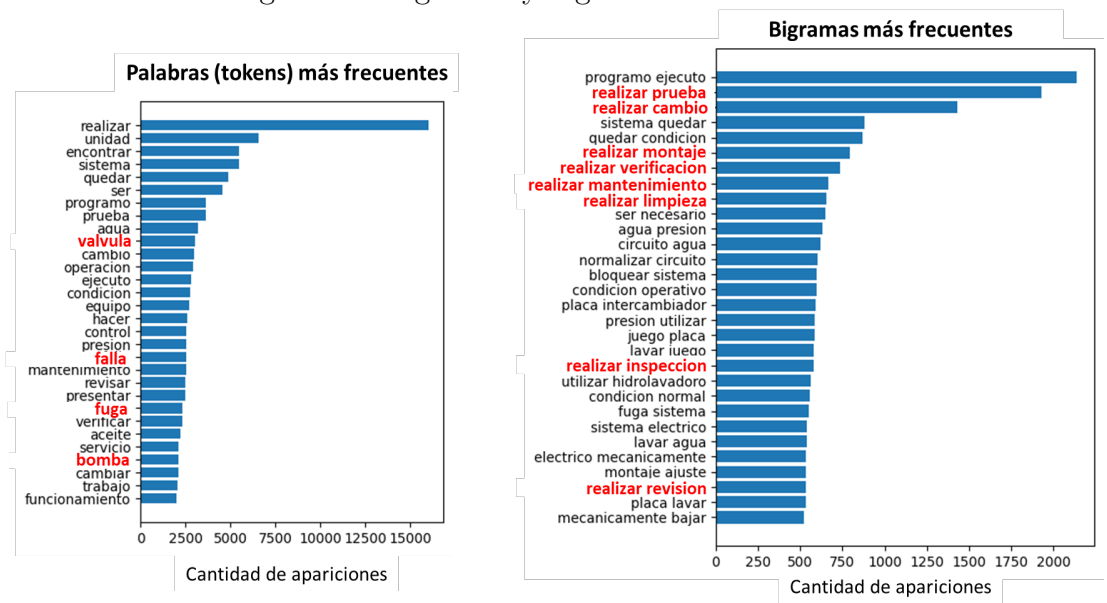
Figura 6: Distribución de frecuencia de las palabras en los textos



Realizando una inspección más detallada a la BoW, se procedió a identificar los primeros 30 unigramas (palabras o *tokens*) más frecuentes presentes en los textos procesados. Se llevó a cabo un análisis similar para el conjunto de bigramas obtenidos de la BoW, obteniendo una visión de las combinaciones de palabras que aparecen con mayor frecuencia (también las primeras 30) en los registros de trabajo.

En el análisis de los unigramas se encontraron pocos indicios de componentes críticos o fallas recurrentes que se hayan documentado en los registros de mantenimiento. Por otra parte, al examinar los bigramas, como se ha resaltado en la figura 7, se identifica de manera repetida la palabra "realizar" acompañada de otras palabras que indicarían acciones correctivas ejecutadas durante los trabajos de mantenimiento y que fueron registradas en los textos de los registros de trabajo.

Figura 7: Unigramas y bigramas más frecuentes



V **Transformación de los datos**

Después de procesar los textos, se procedió a crear una representación vectorial para cada una de las cadenas de *tokens* correspondientes a los registros en el conjunto de datos. Esto permitió que los datos fueran aptos para su modelado utilizando de algoritmos de aprendizaje automático. Así, se generó una matriz de frecuencias para los *tokens* aplicando la representación *tf-idf (term frequency × inverse document frequency)*. Para llevar a cabo esta tarea se utilizó la clase *TfidfVectorizer* del módulo *feature_extraction.text* del paquete *scikit-learn*[4](Pedregosa et al., 2011).

Considerando las características previamente identificadas para el conjunto de datos durante el análisis exploratorio (pág. 14) en lo referente a la distribución de frecuencia de las palabras y siguiendo una estrategia recomendada en estudios previos (Turney and Pantel, 2010), se determinó aplicar filtros sobre la matriz de representación. Esto implicó eliminar componentes que aparecían en muy pocos documentos ($df < 0{,}5\,\%$) o en demasiados documento ($df > 95\,\%$).Esta estrategia se implementó con el objetivo de reducir el número de componentes de los vectores resultantes, al mismo tiempo que se preservaba la información esencial de los registros y se buscaba mejorar el desempeño de los modelos y reducir el costo de procesamiento.

Un procedimiento idéntico al descrito para el conjunto de *tokens* (unigramas) se aplicó también para lograr la representación vectorial de los bigramas.

5.4. Modelado

Como se ha mencionado, en el desarrollo de este proyecto se exploraron diversas técnicas de NLP y modelos de aprendizaje automático para etiquetar los textos (datos no estructurados) correspondientes a los registros de mantenimiento, con el propósito de extraer información en forma de datos estructurados mediante la asignación de categorías relevantes.

Así, durante esta etapa sobre los textos procesados, ya transformados a sus formas de unigramas y bigramas y representados en una matriz de frecuencias, se aplicaron diferentes modelos, con el fin de etiquetar los datos no estructurados y, en particular, identificar posibles elementos de información relevante, como defectos en los equipos y causas de fallas.

I **Extracción de características**

Con el objetivo de identificar términos clave, relacionados con los elementos de interés, se emplearon técnicas de extracción de características y reducción de dimensionalidad.

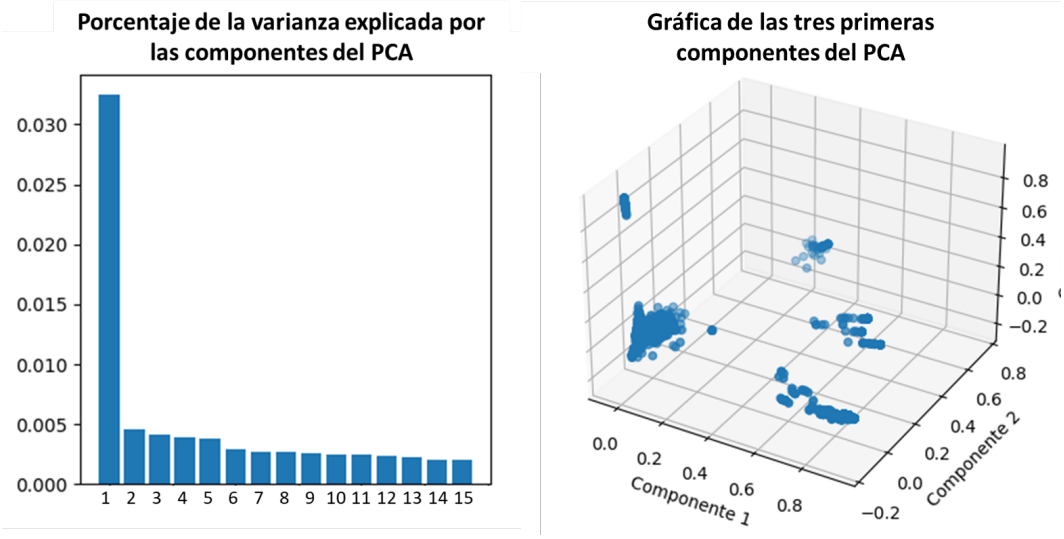
[4]El valor de tf-idf para un termino "t" de un documento "d" en un conjunto de documentos es calculado como:
tf-idf(t, d) $= tf(t,d) * idf(t)$, donde idf(t) $= log[n/df(t)] + 1$, n es el número total de documentos, $df(t)$ es la frecuencia de aparición de "t" en el conjunto de documentos.

Este análisis se utilizó como punto de partida para la aplicación de técnicas no supervisadas de segmentación (*clustering*) que permitieran obtener grupos de registros en los cuales identificar, a través de un análisis exploratorio, las "etiquetas" o categorías relevantes correspondientes a las fallas en los activos y sus causas.

Las técnicas de análisis utilizadas fueron PCA (Análisis de componentes principales) y t-SNE (*T-distributed Stochastic Neighbor Embedding*). Estas técnicas se aplicaron a las representaciones vectoriales de los unigramas y bigramas. No obstante, durante el modelado por agrupamiento (*clustering*) pudo establecerse que fue el análisis de los bigramas el que presentó resultados más consistentes. Por lo tanto será esta la linea de análisis que se detallará en los apartes subsiguientes.

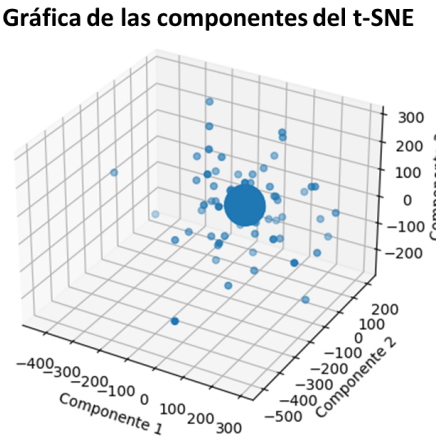
Al aplicar PCA a la matriz de frecuencia de los bigramas se evidencia que la varianza explicada por las primeras componentes del PCA es baja (figura 8). Este resultado sugiere que las componentes no logran capturar adecuadamente la variabilidad presente en los datos originales. Así, es probable que las características (bigramas) no estén altamente correlacionadas y que la simplificación del conjunto de datos podría conllevar la pérdida de información relevante. No obstante, al graficar las tres primeras componentes, buscando comprender la estructura de los datos, se logró identificar un patrón evidente de agrupamiento en los datos (figura 8).

Figura 8: Resultado PCA



Como se mencionó anteriormente, se utilizó la técnica de t-SNE de manera alternativa al PCA, con el objetivo de explorar posibles relaciones y estructuras no lineales en el conjunto de datos. Sin embargo, a diferencia de lo observado con el PCA, al evaluar la visualización de los resultados (figura 9), estos no proporcionan indicios de patrones de agrupación claramente diferenciados en los datos.

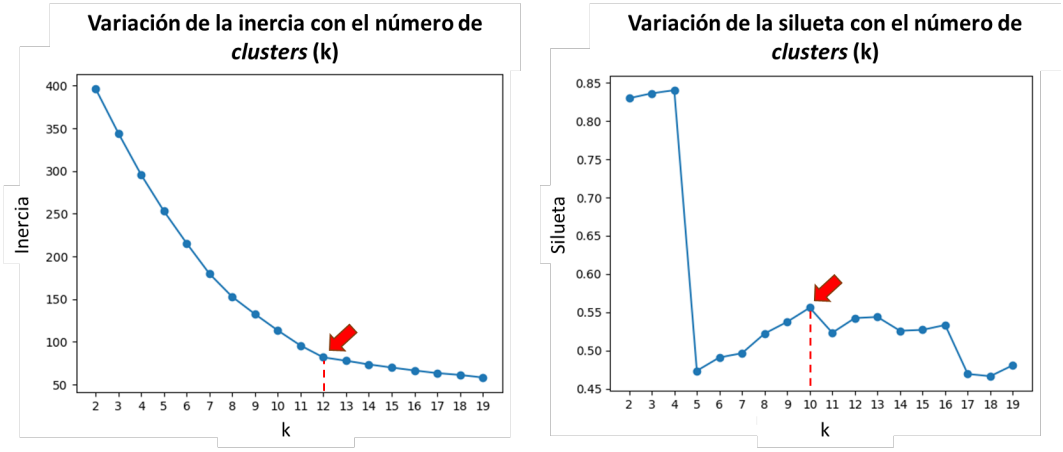
Figura 9: Resultado t-SNE



II Segmentación (*clustering*)

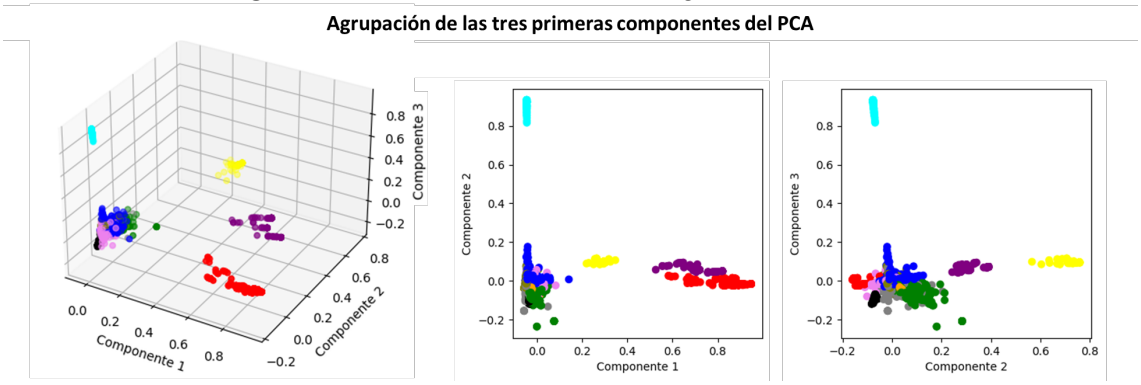
En línea con los resultados obtenidos del análisis PCA, se procedió a la implementación de un modelo no supervisado mediante el empleo del algoritmo de *k-Means*. Esto con el propósito de caracterizar las agrupaciones observadas en los textos y buscando identificar en ellas algunos de los elementos de interés (defectos en los activos y causas de fallas). Al usar las métricas de inercia y silueta para determinar el número k óptimo de *clusters* (método del codo) se obtuvo que este se ubicaría entre 10 y 12 (figura 10).

Figura 10: Identificación del k óptimo



Así el algoritmo de *k-Means* fue ejecutado con la configuración de 12 *clusters*, lo que resultó en un modelo que segmentó los datos de manera que parecía corresponder a los patrones previamente identificados a través del PCA, aunque con alguna superposición en los grupos obtenidos (figura 11).

Figura 11: Resultado del *clustering* usando *k-Means*



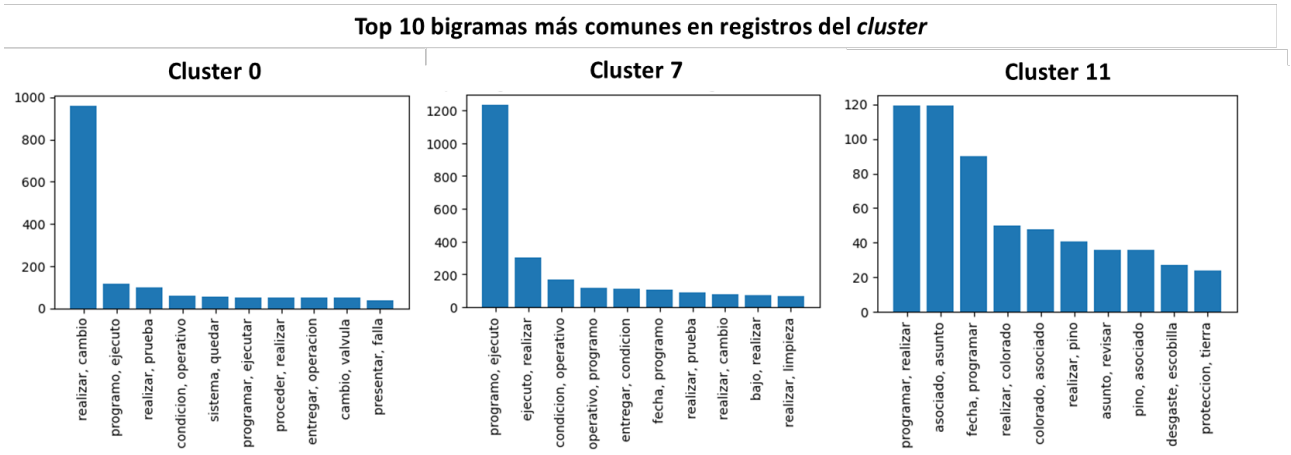
Sin embargo, el número de *clusters* se consideró insuficiente, dado el amplio espectro de posibles categorías asociadas con defectos en activos y causas de falla que se esperaría encontrar en los textos. Lo anterior sugirió que la segmentación de los datos no se ajustaba a los criterios previstos. Esta evaluación se confirmó al explorar los registros más cercanos a los centroides de cada *cluster* (como un medio para perfilar cada grupo), así como los bigramas más frecuentes, lo que reveló que no es posible observar de forma clara las categorías de interés. Para ilustrar esta situación se incluye a continuación (figura 12) el análisis descriptivo de algunos de los *clusters*.

Con base en los resultados obtenidos hasta este punto, podemos evidenciar que el *clustering* proporcionó una comprensión más profunda del conjunto de datos, complementando el análisis descriptivo de los documentos. Sin embargo, dada la necesidad de identificar y categorizar una amplia variedad de elementos en los textos y teniendo en cuenta que los resultados obtenidos con esta técnica no se ajustan a las categorías predefinidas, esta no se ha considerado una herramienta adecuada para extraer la información específica requerida de los textos no estructurados. Por lo tanto se determinó explorar enfoques alternativos para lograr el resultado propuesto.

III Reconocimiento de Entidades Nombradas (NER) basado en reglas

Como se mencionó anteriormente, una de las aplicaciones principales del NER basado en reglas es la extracción de elementos específicos en un documento, pertenecientes a categorías preestablecidas. Por esta razón se definió utilizar un modelo de NER en la identificación de los elementos particulares correspondientes a las fallas en los activos y sus causas, dentro de los textos no estructurados de los registros de mantenimiento. Este modelo permitió aplicar un conjunto de reglas, conformado por patrones previamente definidos a partir del conocimiento especializado y relevante del personal técnico encargado del análisis de los informes de mantenimiento de la infraestructura de generación.

Figura 12: Análisis de frecuencia de bigramas por *cluster* (muestra)



El modelo de NER fue implementado mediante un *pipeline* de procesamiento de lenguaje del paquete spaCy, versión 3.6.1., usando como base el modelo "es_core_news_lg". En este *pipeline* se incorporó el componente *EntityRuler*, que permite realizar reconocimiento de entidades basado en reglas, aplicadas sobre patrones de *tokens*, operando de manera similar a las expresiones regulares. De esta forma se buscó ajustar las predicciones del componente de NER que ya viene incluido por defecto en el *pipeline* y mejorar así su desempeño[5][6][7].

En primer lugar, se definió el conjunto de entidades utilizando como referencia información tabulada sobre fallas recurrentes o de alto impacto en sistemas o equipos críticos de la infraestructura de generación de energía. Esta información había sido recopilada previamente por el personal técnico de la Empresa, como parte de los análisis de confiabilidad, mantenibilidad y disponibilidad de los activos. Con base en dicha información se extrajeron los defectos en los equipos, asociados a los componente con fallas (entidad "COMPONENTE") y las causas de las fallas (entidad "CAUSA") (figura 13).

Figura 13: Ejemplos definición de entidades

Información tabulada sobre fallas

Sistema	Equipo/Item Mantenible	Modo de Falla
Circulación Aceite	Tubería De Aceite	Fuga De Aceite
Comunicaciones	Relé Sel 2505	Conector Contaminado
Generador	Estator Generador	Falla Tierra

Conjunto de entidades

Causa	Componente
Fuga de aceite	Tubería
Conector Contaminado	Relé
Falla Tierra	Estator

[5] https://spacy.io/usage/rule-based-matching#entityruler
[6] https://spacy.io/api/entityruler
[7] https://spacy.io/usage/rule-based-matching#matcher

Tal como es requerido por el modelo de NER, cada conjunto de entidades fue transformado en un listado de patrones de entidades. En este caso, cada patrón buscaba definir la coincidencia con una entidad que podía estar conformada por una o varias palabras. Luego, estos patrones fueron estructurados como datos tipo "diccionario" con dos claves: *"label"*, especificando la etiqueta que se le asignará a la entidad que coincide con el patrón (en este caso "CAUSA" o "COMPONENTE") y *"pattern"*, con el *token* o listado de *tokens* y las diversas características que conforman el patrón a buscar (figura 14).

En este punto, el listado completo de patrones de entidades definido y que conforma el conjunto de reglas que aplica el NER, se cargó al modelo como un archivo JSONL (JSON delimitado por saltos de línea) que contienen un objeto (patrón de entidad) por línea.

Figura 14: Ejemplos de patrones de entidades

Entidades (Coincidencias buscadas)	Patrones de entidades
contaminacion contaminado contaminada contaminados contaminadas	{"label": "CAUSA", "pattern": [{"LOWER":{"REGEX":"^contamina(cion|do|da)s?$"}}]}
fuga fugas fuga aceite fuga agua fuga aire	{"label": "CAUSA", "pattern": [{"LOWER":{"REGEX":"^fugas?$"}},{"LOWER": {"IN" : ["aceite", "agua", "aire"]},"OP":"?"}]}
variador variador frecuencia variador velocidad	{"label": "COMPONENTE", "pattern": [{"LOWER":"variador"},{"LOWER": {"IN" : ["frecuencia", "velocidad"]},"OP":"?"}]}
suiche comunicaciones suiches comunicacion switch comunicación switches comunicaciones	{"label": "COMPONENTE", "pattern": [{"LOWER":{"REGEX":"^s[uw]it?che?s?$"}},{"LOWER":{"REGEX":"^comunicacion(es)?$"}}]}

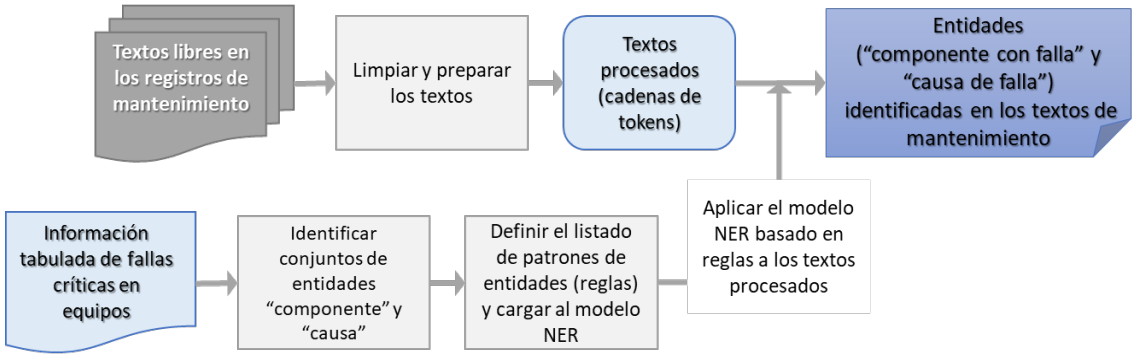
Finalmente, el modelo de NER basado en reglas fue aplicado sobre cada una de las cadenas de *tokens* previamente procesadas, correspondientes a los registros en el conjunto de datos. Esto permitió la identificación y extracción en forma precisa, a partir de los textos de mantenimiento, de los elementos que previamente se habían incluido en los listados de causas de falla -entidad "CAUSA"-, y defectos de los equipos -entidad "COMPONENTE"- (figura 15).

Figura 15: Ejemplos de identificación de entidades “CAUSA” y “COMPONENTE”

Texto libre (incluye etiquetas HTML)	Causas	Componentes
SEP 10/2018 OT: 132072 PROGRAMÓ: ESTEFAN G. EJECUTO: R.ROJAS AL INSPECCIONAR LA TRAMPA DE CONDENSADO DEL COJINETE GUIA SUPERIOR DE LA UNIDAD 2 SE ENCUENTRA QUE LA VALVULA DE DRENAJE DE LA MISMA TIENE PASO DE ACEITE, SE PROCEDE ENTONCES A REALIZAR EL RESPECTIVO CAMBIO DE LA VALVULA, ESTA ES UNA VALVULA DE BOLA DE 3/8". SE APLICA SILICONA EN LAS UNIONES Y SE ENTREGA EN CONDICIONES OPERATIVAS. <!-- RICH TEXT -->	['paso aceite']	['valvula drenaje’, 'valvula’, 'cojinete guia']
SE REVISO CONTROL ELECTRICO DE LAS UNIDADES DE CONDENSACION. SE ENCONTRO TRANSFORMADOR DE CONTROL 440/120 VAC 250 VA QUEMADO. SE CAMBIO. EL EQUIPO ARRANCO BIEN. <!-- RICH TEXT -->	['quemado']	['transformador control’, 'control']
referente a la falla de TPU U2 Transferencia bloqueada, se revisa y se encuentra fuente de 125vdc / 24vdc en corto siendo necesario reemplazarla por fuente de repuesto nueva existente en almacén, se realiza el cambio, se toman medidas y se calibra quedando un voltaje de salida de 24vdc, se normaliza el sistema de transferencia y se hace entrega del equipo al personal de operación en condiciones normales.<!-- RICH TEXT -->	['corto’, ‘falla’, ‘bloqueada']	['transferencia', 'fuente’, 'tpu']

Se presenta a continuación un esquema que incluye las etapas generales para la definición y aplicación del modelo NER basado en reglas(figura 16).

Figura 16: *Pipeline* para el procesamiento de los textos de los registros de mantenimiento



5.5. Evaluación

Como se estableció previamente, la aplicación del modelo de NER basado en reglas sobre el texto no estructurado en los registros de mantenimiento, previamente procesados, posibilitó la extracción a partir de estos de uno o más elementos (entidades) correspondientes a dos categorías, causas de falla y componentes con falla. De un total de 13425 registros analizados, fue posible extraer al menos una entidad “CAUSA” para 6489 y al menos una entidad “COMPONENTE” para 10690, lo que equivale al 48 % y 80 % respectivamente. Por otra parte, no fue posible identificar ninguna entidad para 2202 registros, lo que corresponde al 16 % de los datos analizados.

Estos resultados sería un indicador de que no se fue suficientemente exhaustivo en la construcción de los conjuntos de entidades y los listados de patrones correspondientes que fueron cargados al modelo NER, es decir que algunas “CAUSAS” y “COMPONENTES” no fueron identificados a partir de la información tabulada sobre fallas recurrentes que se tomó como insumo, o simplemente no se encontraban documentados en la dicha fuente. Se

muestran a continuación (figura 17), a modo de referencia, algunos ejemplos de casos en donde existiendo información correspondiente a causas de falla y componentes con falla en los registro, esta no fue extraída.

Figura 17: Entidades no identificadas

Texto libre (incluye etiquetas HTML)	Elementos de información no identificados	
	Causas	Componentes
ENERO 04/2021 OT: 4703023 PROGRAMÓ: JANER G. EJECUTÓ: ERICK E. SE INTERVIENE EL EXTRACTOR E1 DE LOS TRANSFORMADORES DE POTENCIA. SE DETECTA JUEGO RADIAL EN CHUMACERAS, SE SOLICITA COMPRA POR CAJA MENOR AL NO HABER EXISTENCIA EN ALMACÉN. ENERO 05/2021 OT: 4703023 PROGRAMÓ: JANER G. EJECUTÓ: ERICK E. SE INSTALAN CHUMACERAS DE 1" REFERENCIA (P205). SE ACOPLA BANDA DE TRANSMISIÓN Y SE ALINEA. SE REALIZAN PRUEBAS Y SE ENTREGA EN CONDICIONES OPERATIVAS. <!-- RICH TEXT -->	JUEGO CHUMACERAS	EXTRACTOR, TRANSFORMADOR DE POTENCIA
Se realiza ajuste de los Mw totales en el sistema scada en el programa de los controladores de las unidades 3 y 4 que se encontraban desfasados. Cesar Tulio / Luis Eduardo <!-- RICH TEXT -->	DESFASADO	SCADA, CONTROLADOR DE UNIDAD
ES NECESARIO EL CAMBIO DE LOS DOS FLTROS POR ESTAR OBTRUIDOS <!-- RICH TEXT -->	OBSTRUIDO	FILTRO

Igualmente, los resultados podría explicarse por el hecho de que en algunos de los textos efectivamente no se encontraba el tipo de elementos de información buscados, es decir no se hacía referencia a causas de falla o a componentes con falla (figura 18).

Figura 18: Textos sin información

ID del registro	Texto libre (incluye etiquetas HTML)	Texto procesado
76199	SE ATIENDE EN LA OT 92084. SE HIZO DUPLICADA. <!-- RICH TEXT -->	atiende hizo duplicada
139723	22/10/2018. NO SE GENERA ACCION EN ESTE MOMENTO. <!-- RICH TEXT -->	genera accion momento
1099681	PRUEBAS <!-- RICH TEXT -->	pruebas

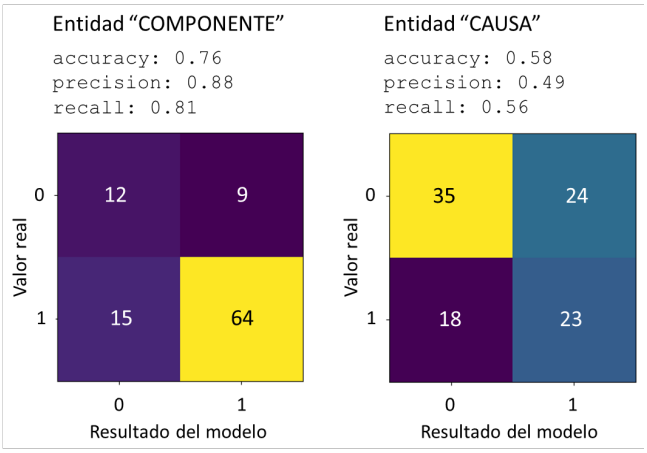
Como se ha mencionado anteriormente, el problema planteado y la solución propuesta usando un modelo NER, se enfocan en una tarea de extracción de información. Sin embargo, con el fin de evaluar el desempeño del modelo y su capacidad para identificar los elementos de interés (entidades) presentes en los textos, se asimiló el resultado obtenido a una clasificación binaria. Así, se tomó una muestra aleatoria de 100 registros y a cada texto se le asignó una etiqueta, para cada categoría ("CAUSAS" y "COMPONENTES") de acuerdo con una evaluación manual basada criterio de personal técnico. Estas etiquetas correspondían a "1" si existía al menos una entidad en el registro, o "0" en caso contrario, considerándose este como el "valor real" (*ground truth*). El mismo procedimiento se aplicó a los resultados generados por el modelo, asignando, en cada categoría, etiquetas "1"si el modelo había identificado al menos una entidad y "0"si no se habían extraído elementos (figura 19).

Figura 19: Etiquetado de los registros

ID del registro	Texto libre (incluye etiquetas HTML)	Etiquetas valores reales		Resultado del modelo		Etiquetas valores modelo	
		Causa_true	Componente _true	Causas	Componentes	Causa_pred	Componentes _pred
652187	Se revisó controlador del Canal # 1, se encontró protección en alimentación de 24VDC disparadas a causa de cortocircuito en alimentación de respaldo del CCM1, se desconectó alimentación de respaldo del CCM1, queda trabajando por alimentación principal (X2 1-2), se descarga nuevamente backup a CCM1 y controlador #1. Se procede a normalizar regulador de tensión, quedando los canales 1, 2 y backup habilitados. Realizo: Zoraida Monsalve Informo: Luis Adrián Cardona <!-- RICH TEXT -->	1	1	disparadas, cortocircuito	alimentacion, regulador tension, controlador	1	1
3021830	SE GENERÓ LA OR, CUANDO SE RECIBA EL ELEMENTO SE PROGRAMARA LA INSTLACION.<!-- RICH TEXT -->	0	0			0	0
2667486	SE REALIZA INSPECCIÓN DE LA VENTOSA DESTAPANDOLA Y EVIDENCIANDO GRAN CANTIDAD DE MATERIAL ANTIESPUMANTE PROVENIENTE DE LA PTAR (SE ANEXA FOTOGRAFIA DE LO ENCONTRADO), EL CUAL OBSTRUIA EL CELLADO HERMETICO DEL DIAFRAGMA. SE EXTRAE DICHO MATERIAL Y SE REALIZAN PRUEBAS OBTENIENDO CESE DE LA FUGA.<!-- RICH TEXT -->	1	1	fuga		1	0

Con este subconjunto de datos, etiquetado según lo descrito, fue evaluada la capacidad del modelo para la extracción de entidades para las dos categorías de interés, "CAUSAS" y "COMPONENTES" usando las métricas *accuracy*, *precision* y *recall*. Para llevar a cabo esta tarea se utilizaron las clases disponible en el módulo *Metrics*[8] del paquete *scikit-learn* (Pedregosa et al., 2011). El valor de estas métricas para cada categoría, acompañado de la correspondiente matriz de confusión, se muestran a continuación (figura 20).

Figura 20: Métricas calculadas para la extracción de entidades



Estos resultados indican que el modelo NER tiene un mejor desempeño en la identificación de entidades de la categoría "COMPONENTES" que de la categoría "CAUSAS". Con valores de 88 % y 81 % para el *precision* y el *recall* respectivamente, el modelo se muestra altamente efectivo para extraer al menos una entidad "COMPONENTE" de los textos, en forma correcta y omitiendo pocas ocurrencias. Por otra parte, en el caso de las entidades "CAUSAS" el modelo presentó un desempeño considerablemente menor, tendiendo a entregar un número mayor de falsos positivos

[8] `https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics`

(*precision*=49 %), a la vez omite una cantidad importante de entidades de este tipo en los registros (*recall*=56 %).

Esta diferencia en el desempeño del modelo puede atribuirse a la complejidad y variabilidad inherente a los textos no estructurados, en relación a la descripción de las causas de las fallas. Así, el hecho de que se presenten dificultades para extraer entidades de la categoría "CAUSAS", usando un modelo como el implementado, es esperable dado que las causas de falla pueden estar registradas de maneras muy diversas en los textos o pueden no haber sido establecidas de forma explícita y tienen una mayor posibilidad de presentar ambigüedad, lo que dificulta su identificación precisa mediante reglas predefinidas.

Así, en términos generales, la aplicación del modelo de NER basado en reglas sobre el texto no estructurado de los registros de mantenimiento demostró ser efectiva en la identificación de elementos correspondientes a los componentes con falla (entidades "COMPONENTES"). Sin embargo, su capacidad para identificar los elementos asociados a las causas de falla (entidades "CAUSAS") presenta margen para ser mejorada, lo cual podría lograrse mediante una definición más exhaustiva del conjunto de entidades y los patrones a través de los cuales se capturan sus múltiples variaciones y condiciones. A pesar de estos desafíos, los resultados obtenidos señalan que el modelo usado y el flujo de trabajo (*pipeline*) en su conjunto entregan una respuesta satisfactoria al problema planteado.

6. Discusión de los resultados

El flujo de trabajo propuesto, incluyendo la aplicación del modelo de NER basado en reglas, proporciona información relevante sobre el historial de condición de los activos y la infraestructura de generación de energía. Esta información resulta de utilidad y puede ser usada por los profesionales encargados del análisis de datos para la toma de decisiones y la definición de estrategias de mantenimiento.

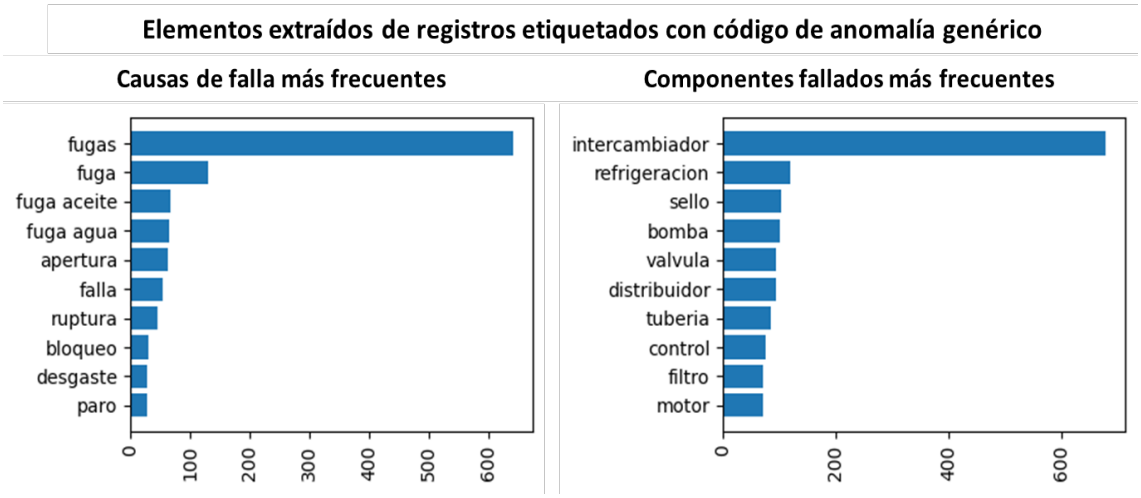
Una de las aplicaciones potenciales y, de hecho, la más directa, es la mejora en la clasificación de los registros de mantenimiento. Como se indicó anteriormente, estos registro cuentan con un campo identificado como "FAILURECODE", donde se diligencia un "código de anomalía" (figura 2), buscando especificar el tipo de falla atendida durante la intervención de mantenimiento. No obstante, en numerosas ocasiones este campo se diligencia con códigos o etiquetas genéricas, como "MECGEN" (mecánico genérico) o "ELECTGEN" (eléctrico genérico), que no ofrecen detalles sobre la naturaleza de la falla, lo que limita significativamente las posibilidades de análisis. Utilizando las entidades extraídas a través del modelo de NER, es posible mejorar el etiquetado de estos registros asignando una "causa de falla" específica, identificada a partir de las descripciones incluidas en los textos no estructurados.

Figura 21: Refinamiento del etiquetado de los datos a partir de las entidades extraídas

N° orden de trabajo	Descripción de la orden de trabajo	Texto procesado	Failurecode (Código de anomalía)	Causa_1	Causa_2
1247765	REVISAR ILUMINACION EXTERIOR PLANTA	mayo de se revisa el alumbrado exterior con la colaboracion de una grua que nos fue prestada por el personal de distribucion se corrige un cortocircuito en una de las luminarias se cambia y queda todo el sistema de iluminacion exterior en optimas condiciones realizo grupo electrico informo william lopera	ELECTGEN	cortocircuito	
3131514	VERIFICAR RUIDO ANORMAL VENTILADOR GABINETE 208 VAC 1NJ3	se verifica ventilador frontal de los servicios auxiliares miscelaneos de g se encuentra desajustado y contaminado se desmonta se lipia y se vuelve a ajustar en su base se aprovecha y se limpian los ventiladores de la parte de atras que se encontraban muy contaminados juan a caicedo henao	ELECTGEN	desajustado	contaminado
3865008	Modulo PL 810 Fallado G34	por presentar perdida de las senales de temperatura en intercambiador y del cojinete superior de la unidad se hace una revision de las senales de temperatura y se encuentra el modulo pl de la link del piso de turbina quemado se cambia dicho modulo por uno de repuesto y se normalizan las senales de temperatura realizo grupo de electronica	MECGEN	perdida senales	quemado
6410544	REVISAR VALVULA DE 4 VIAS DE LA UNIDAD 6 DE G3	ot para la correccion de la fuga de aceite de la valvula de apertura y cierre de la valvula esferica g fue necesario retirar la tapa frontal para el cambio del orring en eje central debido a que este ya presentaba desgaste este fue cambiado por uno de mm se instala tapa se coloca tornilleria se realizan pruebas y se hace entrega a operacion	MECGEN	desgaste	fuga aceite

En esta misma línea de análisis, al disponer de una identificación específica de los elementos "causa de falla" y "componente con falla", obtenidos a partir de la entidades en los registros de trabajo donde se documentan las averías, se posibilita la identificación de condiciones de fallas recurrentes. A modo de ejemplo, se presenta un gráfico donde se muestran, para el conjunto de registros etiquetados con el código de anomalía genérico "MECGEN", las "causas de falla" y "componentes con falla" más frecuentes.

Figura 22: Identificación de condiciones de falla recurrentes (causas y componentes)



Además de lo mencionado, la identificación precisa de las condiciones de falla a partir de entidades extraídas de los textos no estructurados permite el análisis de los indicadores de desempeño de los activos. Para el caso de los "componentes con falla", por ejemplo, es posible calcular métricas clave como el tiempo medio entre fallas (MTBF) para sistemas y componentes específicos. Para ilustrar este caso de uso, se presenta a continuación un conjunto de eventos tabulados correspondiente a un sistema particular, codificado como "G3_CNV_AUXELEC_C/M", en el cual se ha identificado la entidad "cargador baterias" como uno de los "componentes con falla". Con información estructurada de esta

manera, incluyendo la fecha de ocurrencia de cada evento, es posible calcular, de forma prácticamente directa, el tiempo promedio en el que se produce una falla asociada con este componente en el sistema mencionado.

Figura 23: Cálculo de indicadores para condiciones de falla específicas identificadas

N° orden de trabajo	Fecha de la orden de trabajo	Código del sistema	Descripción de la orden de trabajo	Componentes	Tiempo entre eventos (horas)
1905021	2019/11/9 10:22	G3_CNV_AUXELEC_C/M	CAMBIAR BATERÍA #41 BANCO 1 125VDC GIII	[bateria, cargador, banco, baterias, cargador baterías]	
4702443	2021/01/4 8:12	G3_CNV_AUXELEC_C/M	CORREGIR FALLA EN CARGADOR DE BATERIAS NRO. 2 - MODULO 3 - SOBREVOLTAJE.	[cargador baterias, modulos, modulo, cargador]	10126
9711128	2022/05/4 7:53	G3_CNV_AUXELEC_C/M	REVISAR MODULOS CARGADOR No2	[cargador baterias, modulos, barra, cargador]	11640
9913389	2022/05/18 13:58	G3_CNV_AUXELEC_C/M	VERIFICAR/AJUSTAR VOLTAJE MODULOS CARGADORES SA 125 VDC G3	[cargador baterias, barra, modulo, cargador]	342
12731620	2022/09/16 11:22	G3_CNV_AUXELEC_C/M	REALIZAR MANTENIMIENTO MODULO 4 CARGADOR 1	[cargador baterias, modulo]	2901
				Tiempo promedio entre eventos (MTBF) = 5002 horas	

Se observa entonces que al extraer información estructurada de los textos no estructurados de los registros de mantenimiento, esta puede ser usada como apoyo para posteriores análisis de confiabilidad, mantenibilidad y disponibilidad y se convierte en insumo para la evaluación de la condición de los equipos y soporte para la toma de decisiones en mantenimiento.

7. Productos Obtenidos

Como resultado de este proyecto se entrega un flujo de trabajo (*pipeline*), según fue descrito previamente (figura 16), incluyendo documentación y código fuente. Este permite identificar los elementos correspondientes a las causas de falla y los componentes con falla a partir de los datos no estructurados de los textos libres de los registros históricos de mantenimiento de un conjunto de centrales de generación de energía.

8. Plan de Gestión de Datos

Para el desarrollo del proyecto se usaron datos extraídos de la base de datos histórica del sistema transaccional que soporta el proceso de mantenimiento. El conjunto de registros fue obtenido a través de consultas específicas y, como se describió anteriormente, contiene tanto datos estructurados como no estructurados correspondientes a la documentación de

los trabajos de mantenimiento ejecutados.

Los datos crudos y en las diferentes etapas de procesamiento, los resultados obtenidos de la aplicación de los modelos y los propios modelos se encuentran almacenados en sistemas de archivos en la red interna de la Empresa. De esta manera todas las etapas de gestión de datos se han realizado siguiendo los protocolos y políticas establecidas por la Empresa para garantizar la seguridad y confidencialidad de los datos.

Los datos proporcionados por la Empresa y los obtenidos durante del proyecto fueron usados exclusivamente por el autor. Igualmente podrán tener acceso a ellos el director y otras personas que por su relación académica con el proyecto así lo requieran para el correcto cumplimiento de sus obligaciones o gestiones. A estas personas se les advierte sobre su obligación de abstenerse de reproducir, modificar o divulgar a terceros los datos y resultados del proyecto, sin previa autorización escrita y expresa de la Empresa.

9. Aspectos Éticos

En el marco del presente proyecto, los datos proporcionados por la Empresa se han utilizado exclusivamente para las actividades descritas en este documento, requeridas para el logro de los objetivos propuestos y no se han empleado o emplearán para beneficio propio o de terceros. De igual manera, los modelos y resultados obtenidos durante el desarrollo del proyecto se utilizarán únicamente para los fines aquí establecidos.

El producto resultante de este proyecto permitirá procesar en forma automática los textos de los registros de mantenimiento para obtener datos estructurados, con lo cual se optimiza el tiempo de procesamiento y análisis de estos datos. En última instancia, esto significa que la Empresa podrá aprovechar de una mejor manera la información contenida en los textos para apoyar la definición de la estrategia de mantenimiento y la gestión de sus activos.

10. Conclusiones

A pesar de que el uso de un modelo no supervisado permitió una mejor comprensión del conjunto de datos y complementó el análisis descriptivo de los documentos, no se considera un enfoque adecuado para extraer la información específica requerida de los textos no estructurados. Esto se debe a su limitación para identificar y categorizar la amplia variedad de elementos relacionados con defectos en activos y causas de falla que se esperaría encontrar en los textos de los registros de mantenimiento.

El modelo de Reconocimiento de Entidades Nombradas (NER) basado en reglas demostró ser efectivo en la identificación de entidades en los textos no estructurados de los registros de mantenimiento, destacándose particularmente en la extracción de entidades

de la categoría “componentes con falla”. Un desempeño menos satisfactorio se observó en la identificación de “causas de falla”, debido a la variabilidad y ambigüedad en la forma en que estas se registran. Sin embargo, en este aspecto puede lograrse una mejora en la capacidad del modelo mediante una construcción más exhaustiva del conjunto de entidades y el listado de patrones correspondiente, de manera que capture en forma más adecuada sus múltiples variaciones y condiciones.

En general, los resultados obtenidos señalan que el modelo y el flujo de trabajo (*pipeline*) propuestos son válidos en el contexto del problema planteado. Este enfoque mejora significativamente el etiquetado y la extracción de información a partir de lo textos no estructurados de los registros de mantenimiento, permitiendo la identificación de elementos relacionados con las condiciones de falla (“causas” y “componentes”). Por lo tanto, es una aproximación útil y pertinente al proporcionar información valiosa sobre el historial de condición de los activos y habilitar la realización de análisis relacionados con la confiabilidad, mantenibilidad y disponibilidad de los equipos.

Se identifica como una línea de trabajo a desarrollar la implementación de un proceso iterativo donde los resultados obtenidos del propio modelo de NER basado en reglas, en combinación con otras fuentes de información técnica relevante, servirán para refinar continuamente los patrones de entidades (reglas), con el objetivo de mejorar en forma progresiva el rendimiento del modelo.

Además, se propone explorar el desarrollo de un modelo NER basado en técnicas de aprendizaje profundo, en línea con el estado del arte. Este modelo, siendo entrenado con un conjunto de datos etiquetado específico para el dominio del problema, permitiría identificar las entidades no solo mediante palabras clave o patrones, sino también aprovechando el contexto en el que aparecen en los documentos. Se esperaría entonces que esto resulte en un aumento significativo en la precisión de la extracción de información.

11. Anexos

Repositorio en GitHub del proyecto. **Proyecto de Grado MCDA**

Referencias

Arif-Uz-Zaman, K., Cholette, M. E., Ma, L., and Karim, A. (2017), "Extracting failure time data from industrial maintenance records using text mining," *Advanced Engineering Informatics*, 33.

Asociación Española de Normalización y Estandarización (2002), "UNE-EN 13306:2002 Terminología de Mantenimiento," Norma.

Bhardwaj, A. S., Deep, A., Veeramani, D., and Zhou, S. (2022), "A Custom Word Embedding Model for Clustering of Maintenance Records," *IEEE Transactions on Industrial Informatics*, 18, 816–826.

Bird, Steven; Klein, E. L. E. (2009), *Natural Language Processing with Python*, O'Reilly.

Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., and Lukens, S. (2021), "Technical language processing: Unlocking maintenance knowledge," *Manufacturing Letters*, 27.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T. P., Shearer, C., and Wirth, R. (2000), "CRISP-DM 1.0: Step-by-step data mining guide," .

Deloose, A., Gysels, G., Baets, B. D., and Verwaeren, J. (2023), "Combining natural language processing and multidimensional classifiers to predict and correct CMMS metadata," *Computers in Industry*, 145.

Hatef, E., Rouhizadeh, M., Nau, C., Xie, F., Rouillard, C., Abu-Nasser, M., Padilla, A., Lyons, L. J., Kharrazi, H., Weiner, J. P., and Roblin, D. (2022), "Development and assessment of a natural language processing model to identify residential instability in electronic health records' unstructured data: A comparison of 3 integrated healthcare delivery systems," *JAMIA Open*, 5.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020), "spaCy: Industrial-strength Natural Language Processing in Python," .

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., and Flach, P. (2021), "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Transactions on Knowledge and Data Engineering*, 33, 3048–3061.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011), "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 12, 2825–2830.

Raja, N. K., Bakala, N., and Suresh, S. (2019), "NLP: Rule based name entity recognition," *International Journal of Innovative Technology and Exploring Engineering*, 8.

Ricketts, J., Pelham, J., Barry, D., and Guo, W. (2022), "An NLP framework for extracting causes, consequences, and hazards from occurrence reports to validate a HAZOP study," volume 2022-September.

Sala, R., Pirola, F., Pezzotta, G., and Cavalieri, S. (2022), "NLP-based insights discovery for industrial asset and service improvement: An analysis of maintenance reports," *IFAC-PapersOnLine*, 55.

Stenström, C., Aljumaili, M., and Parida, A. (2015), "Natural language processing of maintenance records data," *International Journal of COMADEM*, 18.

Tixier, A. J., Hallowell, M. R., Rajagopalan, B., and Bowman, D. (2016), "Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports," *Automation in Construction*, 62.

Turney, P. D. and Pantel, P. (2010), "From frequency to meaning: Vector space models of semantics," *Journal of Artificial Intelligence Research*, 37, 141–188.

Usuga-Cadavid, J. P., Grabot, B., Lamouri, S., Pellerin, R., and Fortin, A. (2020), "Valuing free-form text data from maintenance logs through transfer learning with CamemBERT," *Enterprise Information Systems*.

Usuga-Cadavid, J. P., Lamouri, S., Grabot, B., and Fortin, A. (2022), "Using deep learning to value free-form text data for predictive maintenance," *International Journal of Production Research*, 60.

Zhang, T., Bhatia, A., Pandya, D., Sahinidis, N. V., Cao, Y., and Flores-Cerrillo, J. (2020), "Industrial text analytics for reliability with derivative-free optimization," *Computers and Chemical Engineering*, 135.

Öztürk, E., Solak, A., Bäcker, D., Weiss, L., and Wegener, K. (2022), "Analysis and relevance of service reports to extend predictive maintenance of large-scale plants," *Procedia CIRP*, 107.