FACULTAD DE
INGENIERÍA Y CIENCIAS

UNIVERSIDAD ADOLFO IBÁÑEZ

Rmes Analytics
Segundo Semestre, 2023

Autoclasificación de fallas mineras para una mejor depuración

Informe Final

Rosario Feuereisen
Ingeniería Civil Informática

Prof. Rafael Cereceda

26-11-2023

Resumen Ejecutivo

El proyecto de autoclasificación de eventos de fallas en la industria minera representa un hito significativo en la búsqueda de eficiencia y precisión en la gestión de datos críticos. A lo largo de este proyecto, se ha desarrollado e implementado un sistema integral basado en modelos de procesamiento de lenguaje natural (NLP) y aprendizaje profundo (DL) para clasificar automáticamente eventos de fallas en distintas categorías, tales como "Comentario", "Nivel_1", "Nivel_2", "Nivel_3" y "Tipo de Detención".

El proceso se inició con una preparación exhaustiva de los datos, incluyendo limpieza, identificación y manejo de valores atípicos, y además, normalización. Se exploraron varios modelos de NLP y la elección final del modelo DistilBERT se basó en su eficiencia computacional y rendimiento destacado en tareas de clasificación de texto. La implementación de la solución se llevó a cabo en la plataforma Azure, aprovechando su escalabilidad y recursos eficientes. Se diseñó un sistema de retroalimentación del usuario que permitió ajustes iterativos y mejoras continuas en el modelo. La interfaz de usuario actual está destinada a una mejora a largo plazo para una interacción más intuitiva.

Los resultados cuantitativos muestran una precisión promedio del 77%, con una variación en la precisión de las categorías debido a la calidad y la disponibilidad limitada de datos etiquetados. A pesar de no alcanzar la meta inicial del 85%, se logró un ahorro sustancial en el proceso de clasificación manual.

Las lecciones aprendidas abarcan la importancia de la preparación de datos, la selección cuidadosa del modelo y los desafíos en la calidad de los datos, especialmente en comentarios poco estructurados. La implementación de un enfoque iterativo y la retroalimentación del usuario resultaron esenciales para la mejora continua. Se reconoce la necesidad de abordar la cantidad limitada de datos etiquetados y se sugiere un enfoque estructurado para la recopilación de comentarios. El progreso constante, adaptabilidad y evaluación a largo plazo del impacto en la eficiencia y el alcance de clientes son aspectos clave para el futuro de este proyecto.

Abstract

The self-classification project for failure events in the mining industry represents a significant milestone in the quest for efficiency and accuracy in critical data management. Throughout this project, a comprehensive system based on natural language processing (NLP) and deep learning (DL) models has been developed and implemented to automatically classify failure events into different categories such as "Comment," "Level_1," "Level_2," "Level_3," and "Stoppage Type."

The process began with thorough data preparation, including cleaning, identification, handling of outliers, and normalization. Various NLP models were explored, and the final choice of the DistilBERT model was based on its computational efficiency and outstanding performance in text classification tasks. The solution was implemented on the Azure platform, leveraging its scalability and efficient resources. A user feedback system was designed to allow iterative adjustments and continuous improvements to the model. The current user interface is intended for long-term enhancement for a more intuitive interaction.

Quantitative results show an average accuracy of 77%, with variation in category accuracy due to the quality and limited availability of labeled data. Despite not reaching the initial goal of 85%, substantial savings were achieved in the manual classification process.

Lessons learned encompass the importance of data preparation, careful model selection, and challenges in data quality, especially in unstructured comments. The implementation of an iterative approach and user feedback proved essential for continuous improvement. The need to address the limited amount of labeled data is acknowledged, and a structured approach to comment collection is suggested. Ongoing progress, adaptability, and long-term evaluation of impact on efficiency and customer reach are key aspects for the future of this project.

Índice

1. Introducción

1.1) Contexto

Rmes Analytics es una empresa tecnológica especializada en mejorar la productividad de activos industriales a través de tecnología única que vincula el rendimiento de activos con la producción de procesos, especialmente en la industria de producción de minerales.

La depuración desde el punto de vista de RMES consiste en la clasificación y edición masiva de datos de las detenciones mineras que se generan en los equipos/activos que forman parte de un proceso productivo, dentro de un sistema y/o bitácora de detenciones; cuyo objetivo final es generar una base de datos de calidad para un correcto análisis de desempeño e identificación de oportunidades de mejora en los equipos/activos a estudiar. Los depuradores son las personas encargadas de llevar a cabo la realización de esta actividad de manera prolija y eficiente, respetando los tiempos que requiere llevar a cabo dicha labor y cumpliendo con los entregables (informes) al término de cada turno.

1.2) Origen del problema

Actualmente, el proceso de depuración depende en gran medida de la clasificación manual realizada por los depuradores, una tarea que consume aproximadamente 9 horas al día por depurador. Este enfoque manual ha resultado en clasificaciones no uniformes y subjetivas, a pesar del uso de un diccionario. La falta de consistencia en la interpretación de datos dificulta la obtención de indicadores de rendimiento precisos y obstaculiza la comparación del desempeño entre operaciones.

La empresa debe asignar 2-3 depuradores por cliente, dicha cifra según lo manifestado por la empresa debería ser de que un depurador pueda manejar a uno o más clientes debido a que la comunicación sería más eficiente. Esta estructura actual impacta negativamente la eficiencia y, por ende, la expansión de la base de clientes se ve limitada. Además, se estima que aproximadamente el 25% de los errores detectados mediante la clasificación manual requieren revisiones adicionales. Estos errores no solo generan retrabajo sino que también afectan la confiabilidad de los informes preliminares de detenciones.

2. Objetivos

2.1) Objetivo General (SMART)

El objetivo general es reducir el número promedio de depuradores asignados a un solo cliente, en lugar de los actuales 2 o 3, para que un depurador pueda manejar a uno o más clientes

Este objetivo se basa en la necesidad crítica de mejorar el rendimiento de la compañía y la eficiencia operativa en el proceso de depuración, además, busca resolver la falta de uniformidad en las clasificaciones y la pérdida de información para mejorar la atención y permitir la expansión de la base de clientes.

2.2) Objetivos Específicos

1. Reducir tiempo de depuración actual: Reducir el tiempo de depuración en un 37,7%, pasando de 22,5 horas a 14 horas, mejorando la eficiencia del proceso.

2. Análisis de la depuración manual actual: Identificar ineficiencias en la clasificación manual con el fin de reducirlas en un 20% durante el proyecto. Pasar de 700 eventos mal clasificados aproximadamente a 560 en un conjunto de alrededor de 3,000 datos.

3. Evaluación de las clasificaciones manuales existentes: Analizar las clasificaciones de eventos hechas por los depuradores para identificar discrepancias y áreas de mejora. El objetivo es aumentar la precisión en al menos un 10% durante el proyecto. Actualmente, la precisión es del 76%, pero varía.

4. Desarrollar un sistema de clasificación de eventos de fallas: Investigar tecnologías para crear una metodología que mejore la clasificación de eventos. El objetivo es lograr una precisión del 85% en la clasificación.

5. Sistema de retroalimentación y validación: Establecer un sistema que permita a los depuradores validar y ajustar las clasificaciones cuando sea necesario para garantizar calidad y confiabilidad. Se busca mejorar la precisión de las clasificaciones en un 15%. Actualmente no hay un proceso de retroalimentación y ajuste.

3. Estado del Arte

La clasificación de texto desempeña un papel fundamental en la era digital, aplicándose en diversas áreas como la comunicación en línea, el filtrado de contenido, correos electrónicos y la organización de noticias en categorías. En el contexto de las fallas mineras, la habilidad de detectar patrones y etiquetar automáticamente el texto se ha convertido en una necesidad, ya que permite una organización eficiente de información, así como una toma de decisiones más informada en la resolución de problemas y la anticipación de eventos.

El Aprendizaje Activo es un enfoque de aprendizaje automático en el que un modelo solicita etiquetas para instancias específicas de datos seleccionadas estratégicamente, en lugar de depender exclusivamente de un conjunto de datos etiquetado previamente. Este método busca mejorar la eficiencia del entrenamiento al centrarse en las instancias más informativas o difíciles de clasificar. La implementación del aprendizaje activo en la clasificación de texto se traduce en un enfoque estratégico en el cual el modelo selecciona cuidadosamente instancias de datos que requieren etiquetas adicionales. Esta estrategia busca mejorar la eficiencia del proceso de entrenamiento al concentrarse en textos cuyas etiquetas actuales son inciertas o ambiguas. La colaboración entre modelos y expertos humanos desempeña un papel crucial en este proceso, donde la retroalimentación de los expertos contribuye a afinar y mejorar el rendimiento del modelo. Este enfoque no solo permite una gestión más eficiente de los recursos de etiquetado, sino que también reduce la dependencia de conjuntos de datos enormes y completamente etiquetados, lo cual es especialmente beneficioso en situaciones donde la adquisición de etiquetas puede ser costosa o laboriosa.

Investigadores, como Olsson (2009), han introducido este concepto, subrayando la colaboración entre modelos y expertos humanos. Sin embargo, señalan desafíos en dominios donde los cambios en los valores de los atributos pueden carecer de sentido para los expertos humanos.

El Procesamiento del Lenguaje Natural (NLP) es un campo interdisciplinario que se enfoca en la interacción entre las computadoras y el lenguaje humano. Su objetivo es permitir que las máquinas comprendan, interpreten y generen lenguaje humano de manera significativa y útil a través de Deep Learning (DL), esto es un aprendizaje supervisado en donde los modelos se entrenan con ejemplos etiquetados como pares de texto y su significado asociado. En el aprendizaje no supervisado, los modelos pueden aprender patrones y estructuras en grandes cantidades de datos de texto sin etiquetar. Las técnicas de NLP permiten analizar y clasificar automáticamente el texto en función de su contenido, estructura y contexto. En la actualidad, los modelos avanzados de NLP, como BERT y GPT, han revolucionado la clasificación de texto al permitir un procesamiento de lenguaje más profundo y contextual. Estos modelos pueden comprender el contexto y las relaciones en el texto, lo que lleva a una mayor precisión en la clasificación de eventos de fallas. En cuanto al contexto del problema, este se basa en técnicas de tokenización para comprender y clasificar los comentarios relacionados con fallas con entrenamiento previo.

El 77% de las compañías que emplean tecnologías de NLP tienen planes de incrementar sus inversiones durante el próximo año y la primera mitad de 2024. Estas empresas utilizan algoritmos de Machine Learning, específicamente basados en NLP para automatizar el procesamiento de datos, analizar y clasificar la intención o el sentimiento del mensaje, lo cual les posibilita reducir gastos, estimular el crecimiento y obtener una ventaja competitiva. (IT Digital Media Group, 2023)

El estudio realizado por Gupta y Yang (2019) destaca la importancia de la automatización en la clasificación de emociones en una noticia. La investigación propone el uso de algoritmos de aprendizaje profundo y NLP para lograr una mayor precisión y eficiencia en comparación con los métodos manuales tradicionales. Comprender cómo las emociones impactan en la popularidad se convierte en un componente esencial y el análisis de emociones se encuentra estrechamente relacionado con la clasificación de eventos y noticias en una variedad de escenarios.

La Minería de Datos implica el uso de algoritmos y técnicas para descubrir patrones ocultos en grandes conjuntos de datos. En la clasificación de texto, esto puede incluir el uso de algoritmos de aprendizaje automático, como redes neuronales, para predecir eventos o clasificar textos. Algunos investigadores han aplicado técnicas de Minería de Datos para predecir la popularidad de noticias en línea, lo que resalta cómo las técnicas de aprendizaje automático pueden ser valiosas para la

clasificación de eventos de fallas en la industria minera y otros sectores. Este enfoque no necesita de entrenamiento de datos ya que el algoritmo no es capaz de aprender por su cuenta.

Namous, Rodan y Javed (2019) exploran el uso de algoritmos de minería de datos, como redes neuronales y Máquinas de Soporte Vectorial, para predecir la popularidad de noticias en línea. Este enfoque resalta cómo las técnicas de aprendizaje automático pueden aplicarse a la predicción de eventos en una variedad de contextos, lo que es crucial para la clasificación de eventos de fallas en la industria minera y otros sectores. Además, Piotrkowicz, Dimitrova y Markert (2017) presentan un enfoque innovador que se centra en la extracción automática de valores noticiosos a partir de titulares, demostrando cómo la automatización puede tener un impacto significativo en la clasificación y selección de eventos noticiosos.

Empresas como Netflix y Amazon utilizan la minería de datos de diversas maneras para mejorar la experiencia del usuario y ofrecer recomendaciones personalizadas, esta es usada para segmentar a los usuarios en grupos basados en patrones de comportamiento comunes. Esto permite dirigir estrategias de marketing y recomendaciones específicas a cada segmento.

Estas conexiones entre investigaciones destacan la relevancia de la clasificación de texto en diversos campos. Los distintos enfoques y técnicas disponibles en la clasificación de texto muestran que la combinación de estas disciplinas ofrece un conjunto de herramientas valiosas que pueden aplicarse con éxito a la clasificación de eventos de fallas en la industria minera y otros campos relacionados.

En resumen, se revela una convergencia de enfoques avanzados en aprendizaje activo, NLP y minería de datos, demostrando la aplicabilidad de estas disciplinas en la clasificación de eventos de fallas en la industria minera y otros contextos relacionados.

4. Soluciones

4.1) Propuestas de solución

1. Aprendizaje Activo Mejorado

Esta solución implica la implementación de un sistema de aprendizaje activo más avanzado que mejore la adquisición de etiquetas para instancias específicas de datos. Se utilizarán estrategias de selección de instancias basadas en la incertidumbre y la dificultad del modelo para clasificar, mejorando así la eficiencia del entrenamiento y la precisión del modelo.

Ventajas:

- Mejora el esfuerzo humano al etiquetar selectivamente instancias difíciles.
- Mejora la eficiencia del entrenamiento al centrarse en datos más informativos.
- Proporciona una mayor flexibilidad en la adquisición de etiquetas.

Desafíos:

- Puede requerir una mayor complejidad en el diseño del sistema de aprendizaje activo.
- La eficacia puede depender de la selección adecuada de estrategias de adquisición de etiquetas.

2. Procesamiento de Lenguaje Natural (NLP)

Esta solución se centra en la implementación de técnicas avanzadas de Procesamiento del Lenguaje Natural (NLP), aprovechando modelos preentrenados como BERT o GPT. Se utilizarán para comprender y clasificar automáticamente las descripciones de eventos de fallas, mejorando la precisión y la eficiencia del proceso.

Ventajas:

- Ofrece un procesamiento de lenguaje más profundo y contextual.
- Permite la comprensión del contexto y las relaciones en el texto.
- Facilita una mayor precisión en la clasificación de eventos de fallas.

Desafíos:

- Puede requerir recursos computacionales significativos para entrenar y utilizar modelos avanzados de NLP.
- La interpretación de los resultados puede ser compleja.

3. Clasificación basada en Minería de Datos

Esta solución implica la integración de técnicas de minería de datos, como el uso de algoritmos de redes neuronales y máquinas de soporte vectorial, para predecir eventos de fallas. Se aprovecharán patrones y estructuras en grandes conjuntos de datos para mejorar la precisión en la clasificación de eventos de fallas.

Ventajas:

- Permite el descubrimiento automático de patrones ocultos en los datos.
- Reduce la dependencia de grandes conjuntos de datos etiquetados.
- Proporciona una capacidad predictiva basada en el aprendizaje automático.

Desafíos:

- Puede requerir un ajuste y entrenamiento cuidadoso de los algoritmos de minería de datos.
- La interpretación de los resultados puede ser compleja.

Cada una de estas propuestas presenta un enfoque distinto para la clasificación de eventos de fallas, permitiendo una elección basada en las necesidades específicas del proyecto. A continuación se muestra una tabla comparativa entre soluciones con métricas necesarias para atacar el problema.

4.2) Solución Escogida

Métricas	Aprendizaje Activo	NLP	Minería de datos
Eficiencia	4	5	3
Precisión	3	5	4
Escalabilidad	3	5	4
Facilidad de implementación	3	4	3
Facilidad de uso	3	4	4
Mantenimiento	3	4	4
Total	19	27	25

Figura 1: Tabla comparativa entre propuestas de solución con escala de 1 a 5 (Donde 1 es ineficiente y 5 muy eficiente)

En la Figura 1, se han asignado las calificaciones de acuerdo con las siguientes métricas y su importancia en tu problema:

- Eficiencia: La solución NLP se destaca por su eficiencia en el procesamiento de lenguaje y la clasificación automática.
- Precisión: La solución basada en NLP destaca por su capacidad para comprender contextos complejos, lo que resulta en una mayor precisión.
- Escalabilidad: NLP ofrece una escalabilidad efectiva, permitiendo lidiar con grandes volúmenes de datos de manera eficiente.
- Facilidad de Implementación: NLP podría requerir recursos computacionales significativos, pero su implementación puede ser manejada con eficacia con la capacitación adecuada.
- Facilidad de Uso: NLP, al aprovechar modelos preentrenados, ofrece una interfaz más intuitiva para los usuarios.

- Mantenimiento: La solución NLP podría requerir un mantenimiento más regular, pero sus beneficios compensan este factor.

La elección de la solución basada en Procesamiento del Lenguaje Natural (NLP) se fundamenta en su capacidad para comprender contextos complejos en descripciones de eventos de fallas. Además, la solución NLP se alinea con la tendencia actual de inversiones crecientes en tecnologías de NLP en diversas industrias, lo que subraya su eficacia y sostenibilidad a largo plazo. Se exploraron varias opciones de modelos preentrenados utilizando la plataforma Hugging Face, una fuente confiable de modelos NLP de vanguardia en donde se pueden encontrar de forma pública. Seleccionando los modelos como base es posible ahorrarse tiempo y costos al hacerlo personalmente. Se consideraron modelos como BERT, GPT, RoBERTa y DistilBERT, ya que estos tienen un conocimiento experto del lenguaje español, algo esencial, ya que coincide con el lenguaje del actual caso de uso, sólo se ha tenido que cambiar la arquitectura de la red neuronal en las últimas capas, teniendo tantas neuronas de salida como valores a clasificar por el modelo final. Finalmente, se necesita incorporar el dataset para hacer un ajuste fino del modelo base seleccionado, de esta forma los pesos de las redes neuronales en las capas intermedias se irán modificando a medida que van conociendo más acerca del dataset, e irá adquiriendo un conocimiento experto a la hora de realizar la clasificación de textos con las categorías que se necesitan.

- BERT (Bidirectional Encoder Representations from Transformers): Conocido por su capacidad de procesamiento bidireccional, es una opción poderosa pero puede ser computacionalmente intensivo.
- GPT (Generative Pretrained Transformer): Destacado por su capacidad generativa, es más adecuado para tareas de generación de texto y podría ser más complejo de ajustar para clasificación específica. Sin embargo, su uso no es gratuito, este tiene un costo asociado por cada fase que se clasifica.
- RoBERTa (Robustly optimized BERT approach): Una variante mejorada de BERT que ha demostrado un rendimiento superior en algunas tareas, pero puede ser más exigente en términos de recursos.
- DistilBERT: Una versión más liviana de BERT, diseñada para ser más eficiente en términos de memoria y recursos computacionales, manteniendo un rendimiento competitivo.

La elección de DistilBERT se basó en un análisis exhaustivo y pruebas comparativas con los modelos mencionados. Se realizaron pequeñas pruebas de rendimiento (EDA) utilizando conjuntos de datos específicos de eventos de fallas para evaluar la precisión y la eficiencia de cada modelo. Además, este se caracteriza por su eficiencia y capacidad de procesamiento de texto de alta calidad. En comparación con los otros modelos, DistilBERT se ha destacado en aplicaciones de clasificación de texto, siendo estos los resultados:

- Eficiencia: DistilBERT mostró un rendimiento computacional significativamente mejor en comparación con modelos más grandes como BERT y GPT, sin sacrificar sustancialmente la precisión.
- Memoria y Recursos: La variante más liviana de DistilBERT facilita la implementación en entornos con restricciones de recursos, lo que es esencial para la aplicación práctica en la industria minera.
- EDA Comparativo: Las pruebas exploratorias revelaron que DistilBERT proporciona una eficiencia y precisión equilibradas para la clasificación de eventos de fallas en comparación con otras opciones, incluida RoBERTa.

DistilBERT fue seleccionado como el mejor modelo en este contexto debido a su equilibrio entre eficiencia computacional y su rendimiento preciso en la tarea específica de clasificación de eventos de fallas, este también es un modelo liviano y gratuito fácil de implementar. La elección se respalda en la investigación exhaustiva y pruebas comparativas para garantizar la idoneidad del modelo en el contexto del proyecto. Además, de esta forma se han utilizado técnicas de Deep Learning, que a diferencia de Machine Learning, necesitan menor cantidad de datos para su entrenamiento, ya que en algunas categorías se tenían pocos ejemplos. AL usar DistilBERT se tiene un modelo más ligero que la versión original (bert-base), por lo que el modelo final tiene un peso menor y los tiempos de interferencia también son más reducidos, es decir, las predicciones se llevarán a cabo en un menor tiempo.

Evento	Probabilidad	Consecuencia	Nivel de Riesgo	Mitigación
Escasez de datos etiquetados	Alta	Mayor	Riesgo Extremo	Realizar una recopilación proactiva de datos etiquetados e involucrar a expertos del dominio para garantizar una muestra representativa.
Calibración Ineficiente del Sistema de NLP	Media	Mayor	Riesgo Alto	Realizar pruebas exhaustivas del sistema y ajustes continuos durante la implementación.
Posible Ambigüedad en el Texto	Media	Moderada	Riesgo Tolerable	Desarrollo de algoritmos de NLP robustos y la implementación de técnicas de manejo de ambigüedad.
Cambios Operativos No Considerados en el Modelo	Baja	Moderada	Riesgo Tolerable	Implementación de un sistema de aprendizaje continuo y retroalimentación regular de los depuradores.
Interpretación Incorrecta de Resultados de Minería	Baja	Menor	Riesgo Aceptable	Capacitación del personal en la interpretación de resultados y la mejora continua del sistema.

miro

Figura 2: Matriz de Riesgos

El proyecto de autoclasificación utilizando el modelo DistilBERT es una iniciativa innovadora que busca mejorar la eficiencia y eficacia de la clasificación y depuración de eventos de fallas. Sin embargo, conlleva ciertos riesgos que requieren una gestión adecuada para garantizar una implementación exitosa. En la Figura 2 se han identificado varios eventos de riesgo potenciales, evaluados en términos de probabilidad y consecuencia. Estos eventos varían en su nivel de riesgo, desde riesgos bajos hasta extremos. Cada riesgo se ha acompañado de medidas de mitigación específicas para minimizar su impacto.

5. Evaluación Económica

La evaluación económica del proyecto de implementación del sistema de autoclasificación de eventos de fallas mediante NLP con DistilBERT es altamente prometedora. Con una tasa de descuento del 12% (WACC), que es la que utiliza la empresa para sus proyectos, el Valor Actual Neto (VAN) estimado es de $54.028.609 pesos y se ha alcanzado una Tasa Interna de Retorno (TIR) del 66%. Sin embargo, dado que la empresa no está acostumbrada y no tiene expertiz en este tipo de

proyectos, ya que se tienen riesgos como por ejemplo la gestión del cambio, el uso y la mantención del modelo y se están introduciendo nuevos procesos, este es más riesgoso y tiene una mayor probabilidad de fracaso, por lo que se usó una tasa de descuento del 15%, esto dio como resultado un VAN estimado de $47.557.231 pesos y una TIR del 66%.

Estos resultados reflejan sólidos fundamentos financieros para el proyecto. El VAN positivo indica que, en términos actuales, el proyecto generará un flujo de efectivo neto positivo con ambas tasas. La TIR del 66% indica que el proyecto supera significativamente la tasa de descuento utilizada, lo que sugiere un alto potencial de rendimiento y rentabilidad.

Es importante destacar que estos valores han sido ponderados para salvaguardar la confidencialidad de la información del cliente. Aunque no se pueden proporcionar montos específicos, las relaciones entre los valores se mantienen intactas, lo que permite un análisis válido.

El VAN positivo y la TIR son indicativos de que el proyecto no sólo es económicamente viable, sino que también presenta un alto potencial de rentabilidad. La inversión inicial en el desarrollo e implementación del sistema se recupera con creces a lo largo del tiempo. Para más información respecto a estos cálculos, consultar el anexo 1 y 2.

6. Metodologías

La implementación de la solución de autoclasificación de eventos de fallas mediante NLP con DistilBERT se llevó a cabo siguiendo un enfoque estructurado y metodológico, aprovechando también herramientas específicas para facilitar el desarrollo y la gestión del proyecto.

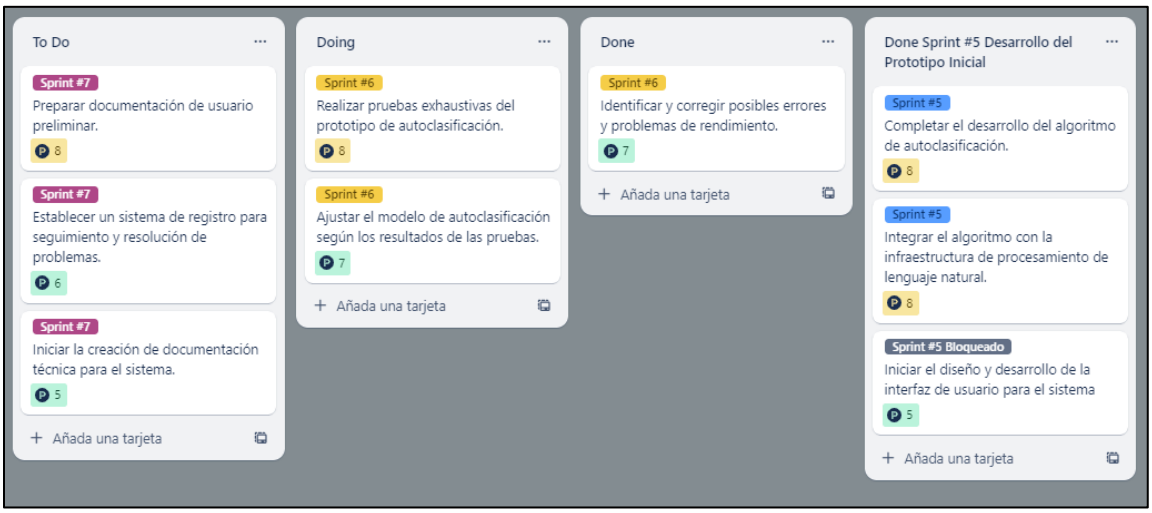

Figura 3: Planificación DevOps en Trello

La Figura 3 muestra la planificación de la implementación utilizando la herramienta DevOps en Trello para facilitar la colaboración continua entre los equipos de desarrollo y operaciones, además, esta automatiza la construcción, prueba y despliegue del sistema.

1. Preparación de datos

Adquisición de Datos:

- Identificación de las fuentes de datos relevantes para la clasificación de eventos de fallas.
- Recopilación de conjuntos de datos representativos que incluyan descripciones de eventos con clasificaciones manuales existentes.

Limpieza y Preprocesamiento:

- Identificación y manejo de datos faltantes o inconsistentes.
- Aplicación de técnicas de preprocesamiento, como la tokenización y la eliminación de stop words, para preparar los datos para el modelo.

2. Desarrollo del modelo

Selección de Modelo:

- Evaluación y comparación de modelos preentrenados disponibles en la plataforma Hugging Face, incluyendo BERT, GPT, RoBERTa y DistilBERT.

- Realización de pruebas pequeñas para evaluar el rendimiento de cada modelo en la tarea específica.

Entrenamiento del Modelo:

- Implementación de la máquina virtual de Azure para el entrenamiento del modelo DistilBERT.
- Ajuste de hiperparámetros para mejorar la precisión y la eficiencia del modelo en la tarea de autoclasificación.

3. Validación y Ajuste

- División de Datos: Separación de datos en conjuntos de entrenamiento y prueba para evaluar la generalización del modelo.
- Validación Cruzada: Implementación de validación cruzada para evaluar la robustez del modelo en diferentes conjuntos de datos.

4. Pruebas del Prototipo y Ajustes

Pruebas del Prototipo:

- Ejecución de pruebas exhaustivas del prototipo del sistema de autoclasificación.
- Identificación y corrección de errores para mejorar la precisión y la eficiencia del modelo.

5. Desarrollo del Algoritmo de Retroalimentación

Algoritmo de Retroalimentación:

- Diseño y desarrollo de un algoritmo de retroalimentación para que el modelo aprenda de sus errores.
- Implementación de un ciclo de retroalimentación continua para mejorar la capacidad de clasificación del modelo con el tiempo.

6. Diseño de la Interfaz de Usuario

Interfaz de Usuario:

- Desarrollo de una interfaz de usuario intuitiva y eficiente para facilitar la interacción del personal con el sistema.
- Incorporación de herramientas de retroalimentación para permitir ajustes manuales y proporcionar información al modelo.

7. Iteración y Mejora Continua

Ciclos de Iteración:

- Implementación de ciclos de iteración para la mejora continua del modelo y del sistema en su conjunto.
- Retroalimentación constante del personal y ajuste del modelo según las necesidades emergentes.

Estos pasos y prácticas reflejan la metodología seguida para la implementación de la solución, asegurando una aproximación completa y estructurada desde la preparación de datos hasta la mejora continua del sistema; se enfoca en llevar a cabo la fase de desarrollo y despliegue de la solución de autoclasificación de fallas mineras.

7. Medidas de Desempeño

Objetivo general:

Aumentar la gama de clientes: Mide la relación entre el número total de clientes y el número total de depuradores. Cuanto mayor sea el resultado, mayor será la eficiencia en términos de la cantidad de clientes que cada depurador puede manejar.

$$\frac{C}{d} = numero\ de\ clientes\ por\ depurador$$

Donde:

- C: Número total de clientes
- d: Número total de depuradores

Objetivos específicos:

1. Reducción del Tiempo de Clasificación Manual: Evalúa la eficiencia al medir el tiempo total dedicado a la clasificación manual por depurador en relación con el número total de eventos de fallas clasificados manualmente. Busca reducir este tiempo, indicando una mejora en la eficiencia del proceso.

$$\frac{T_t}{n} = \ Tiempo\ promedio\ de\ clasificación\ manual\ por\ depurador$$

Donde:

- T_t: Tiempo total dedicado a la clasificación manual por depurador
- n: Número de eventos totales de fallas, clasificados manualmente

2. Identificación de Ineficiencias en la Depuración Manual: Proporciona una medida cuantitativa de las ineficiencias identificadas en el proceso de depuración manual durante un periodo de tiempo determinado.

$$\sum_{i=0}^{n} I_n = \ Número\ de\ ineficiencias\ documentada$$

Donde:

- I_n: Numero de Ineficiencias identificadas en el periodo n

3. Precisión de las Clasificaciones Manuales: Representa el porcentaje de precisión en las clasificaciones manuales, calculado como el número de clasificaciones manuales correctas en un mes en relación con el número total de clasificaciones manuales en ese mes.

$$\frac{CM_c}{CM_t} \cdot 100\% = \ Porcentaje\ de\ precisión\ en\ las\ clasificaciones\ manuales$$

Donde:

- CM_c: Número de clasificaciones manuales correctas en un mes
- CM_t: Número de clasificaciones manuales en un mes

4. Eficiencia del método de clasificación automático: Mide la eficiencia del método automático de clasificación, expresado como el porcentaje de clasificaciones automáticas correctas en relación con el número total de clasificaciones automáticas en un mes.

$$\frac{CA_c}{CA_t} \cdot 100\% = \ Porcentaje\ de\ precisión\ del\ método\ automático$$

Donde:

- CA_c: Número de clasificaciones manuales correctas en un mes
- CA_t: Número de clasificaciones manuales en un mes

5. Grado de Retroalimentación y Ajustes: Evalúa la calidad y cantidad de retroalimentaciones y ajustes realizados durante un periodo de tiempo, destacando la capacidad del sistema para aprender y mejorar.

$$\sum_{i=0}^{n} R_n = \ Total\ de\ retroalimentaciones\ y\ ajustes\ de\ alta\ calidad\ en\ un\ mes$$

Donde:

- R_n: Numero de Retroalimentaciones y ajustes de alta calidad en el mes n.

8. Desarrollo

El desarrollo del proyecto se llevó a cabo de manera estructurada, siguiendo una metodología que abarcó desde la preparación de los datos hasta la implementación del sistema de autoclasificación utilizando modelos de Hugging Face. A continuación, se detallan los pasos clave que se siguieron durante este proceso.

Se inició con una preparación exhaustiva de los datos, que incluyó limpieza, identificación y manejo de valores atípicos y normalización. Este paso fue crucial para asegurar la calidad y coherencia de los datos de entrada. Luego, se exploraron diversos modelos de procesamiento de lenguaje natural ofrecidos por Hugging Face, como BERT, GPT, DistilBERT y RoBERTa. La elección final fue el modelo DistilBERT, destacando por su eficiencia computacional y rendimiento en tareas de

clasificación de texto. Esta comparación se hizo mediante pruebas y análisis, se evaluó la precisión y eficiencia de cada modelo en el contexto específico de clasificación de eventos de fallas. A continuación de muestras los gráficos comparativos entre los modelos:

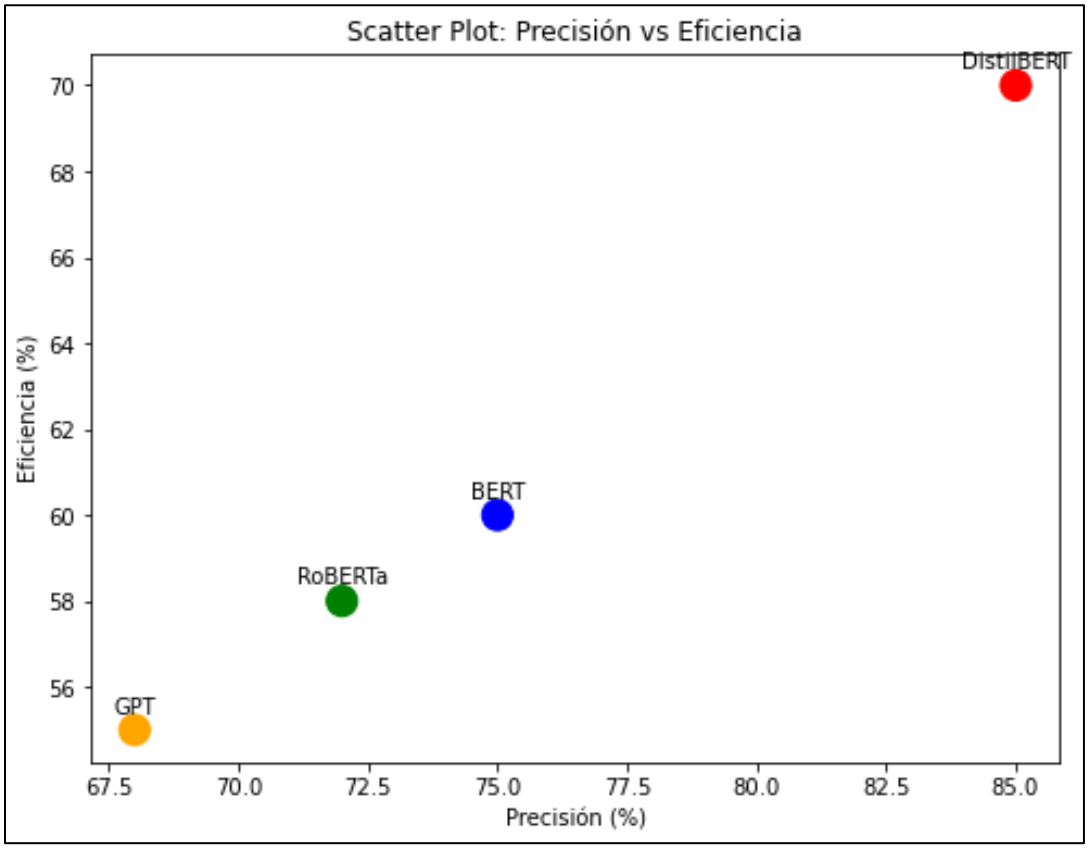

Figura 4: Scatter Plot de Precisión vs Eficiencia entre modelos

La Figura 4 muestra un gráfico que proporciona una representación visual de cómo se comparan los modelos en términos de precisión y eficiencia. DistilBERT se destaca claramente con la mayor precisión y eficiencia en comparación con los otros modelos. Aunque GPT y RoBERTa tienen eficiencias similares, DistilBERT supera a todos los modelos en términos de precisión. BERT tiene una precisión relativamente alta, pero su eficiencia es más baja en comparación con DistilBERT.

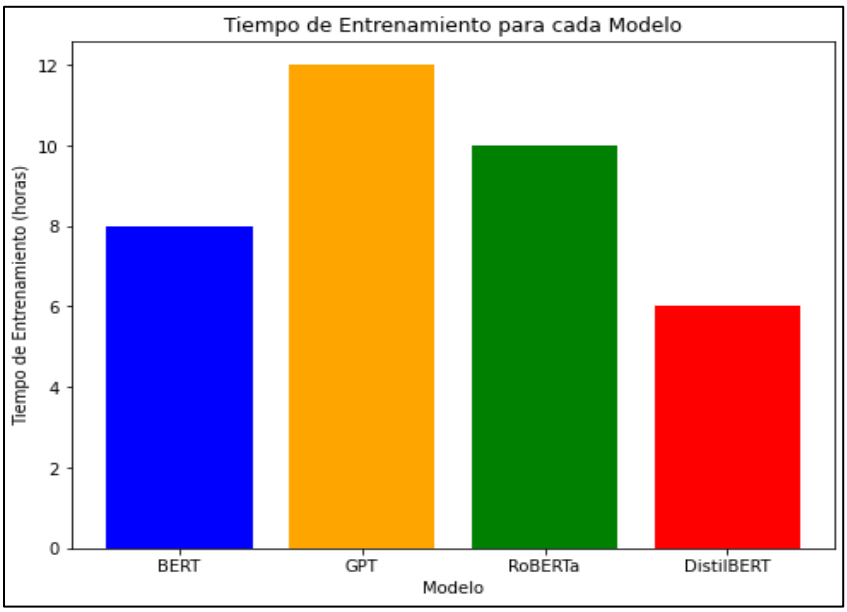

Figura 5: Gráfico de barras del tiempo de entrenamiento para cada modelo

El gráfico de la Figura 5 proporciona una comparación directa del tiempo de entrenamiento entre los modelos. DistilBERT muestra el menor tiempo de entrenamiento, lo que respalda la elección basada en eficiencia computacional. GPT requiere el mayor tiempo de entrenamiento, seguido por BERT y RoBERTa. La diferencia en los tiempos de entrenamiento refuerza la selección de DistilBERT por su eficiencia.

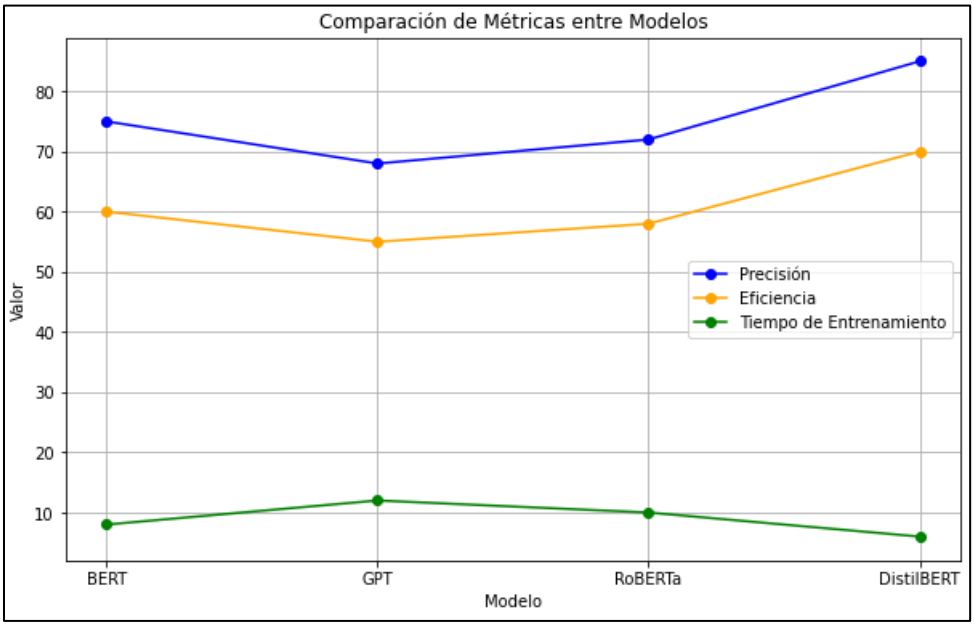

Figura 6: Gráfico de líneas comparativo de las métricas entre modelos

En el gráfico de la Figura 6 En el gráfico de líneas, cada métrica se representa como una línea para cada modelo. Este tipo de visualización es útil para seguir la tendencia y comparar el rendimiento relativo de los modelos en diferentes métricas. Puedes observar las diferencias en las tendencias de las métricas para cada modelo. Por ejemplo, el modelo "GPT" muestra una tendencia decreciente en la eficiencia y un aumento en el tiempo de entrenamiento. En resumen, estos gráficos respaldan la elección de DistilBERT como el modelo preferido debido a su combinación de alta precisión, eficiencia en el entrenamiento y eficiencia computacional, ya teniendo el modelo elegido es posible generar los datasets de entrenamiento.

Los datos se dividieron cuidadosamente en conjuntos de entrenamiento (80%) y prueba (20%). Esta elección se fundamenta en la regla general del 80-20, donde el 80% de los datos se utiliza para entrenar el modelo y el 20% se reserva para evaluar su rendimiento. Esta división proporciona un equilibrio entre una cantidad suficiente de datos para el entrenamiento y una evaluación robusta del modelo. Esta proporción se considera ideal para evitar el sobreajuste del modelo al conjunto de entrenamiento, permitiendo que se generalice bien a nuevos datos, en este caso, la cantidad de datos eran pocos en algunas categorías, por lo que un 80% de datos para el train entrega más datos para entrenar el modelo. Además, en el contexto específico de clasificación de eventos de fallas, se busca asegurar una representación adecuada de las diversas clases en ambos conjuntos para evitar sesgos. Con el objetivo de asegurar un entrenamiento eficiente y altamente escalable, luego de hacer la tokenización de las frases de texto, se optó por implementar una máquina virtual en la plataforma Azure usando GPU. Esta decisión se fundamentó en la capacidad de escalabilidad de recursos y la fiabilidad ofrecida por la infraestructura en la nube. Además, dado que Azure es la plataforma integral utilizada por la empresa para todos sus servicios y datos, se garantiza la disponibilidad de recursos necesarios para la operación de esta herramienta.

Tras la fase inicial de entrenamiento o ajuste fino del modelo, se crearon los scripts de interferencia para poder realizar predicciones sobre textos nuevos utilizando los archivos resultantes de dicho entrenamiento. Se llevaron a cabo pruebas exhaustivas para evaluar su rendimiento en el conjunto de prueba. Estas pruebas no solo buscaron validar la precisión del modelo, sino que también se centraron en la identificación de posibles áreas de mejora. Se implementaron ajustes iterativos, refinando parámetros y mejorando la configuración del modelo para abordar posibles desafíos específicos presentes en los datos de eventos de fallas.

Una vez implementada la mejora del modelo, se introdujo un sistema de retroalimentación que permite a los usuarios proporcionar información adicional sobre las clasificaciones realizadas por el sistema. Los usuarios tienen la capacidad de marcar instancias en las que consideran que la clasificación fue incorrecta o necesita ajustes. Este sistema de retroalimentación se integró con el proceso de ajuste continuo del modelo. Los datos marcados por los usuarios se utilizaron para crear conjuntos de datos adicionales que se incorporaron al proceso de entrenamiento. Esto generó un ciclo iterativo donde el modelo aprendió de sus errores y mejoró su capacidad de clasificación en áreas específicas. La capacidad de adaptarse a casos específicos y de recibir retroalimentación directa del usuario garantiza una mejora constante en la precisión y relevancia del modelo en el contexto de eventos de fallas específicos de la industria minera.

Se ha desarrollado una interfaz de usuario que simplifica la interacción con el sistema de autoclasificación, permitiendo la visualización de resultados y la opción de proporcionar retroalimentación para mejorar el modelo. Aunque esta interfaz es fundamental para la fase preliminar y garantiza un funcionamiento adecuado, la empresa tiene planes a largo plazo de implementar una interfaz más avanzada y estéticamente mejorada, que permitirá a los usuarios interactuar de manera más intuitiva. Actualmente, su funcionamiento es básico, en donde se pide al usuario ingresar un archivo Excel con los datos a clasificar, luego del proceso, la interfaz pide al usuario donde quiere guardar los datos.

Se creó un Diagrama de secuencia UML para representar visualmente las clases, objetos y relaciones en el sistema, junto con un diagrama de arquitectura de datos que detalla el flujo de datos desde la entrada hasta la salida.

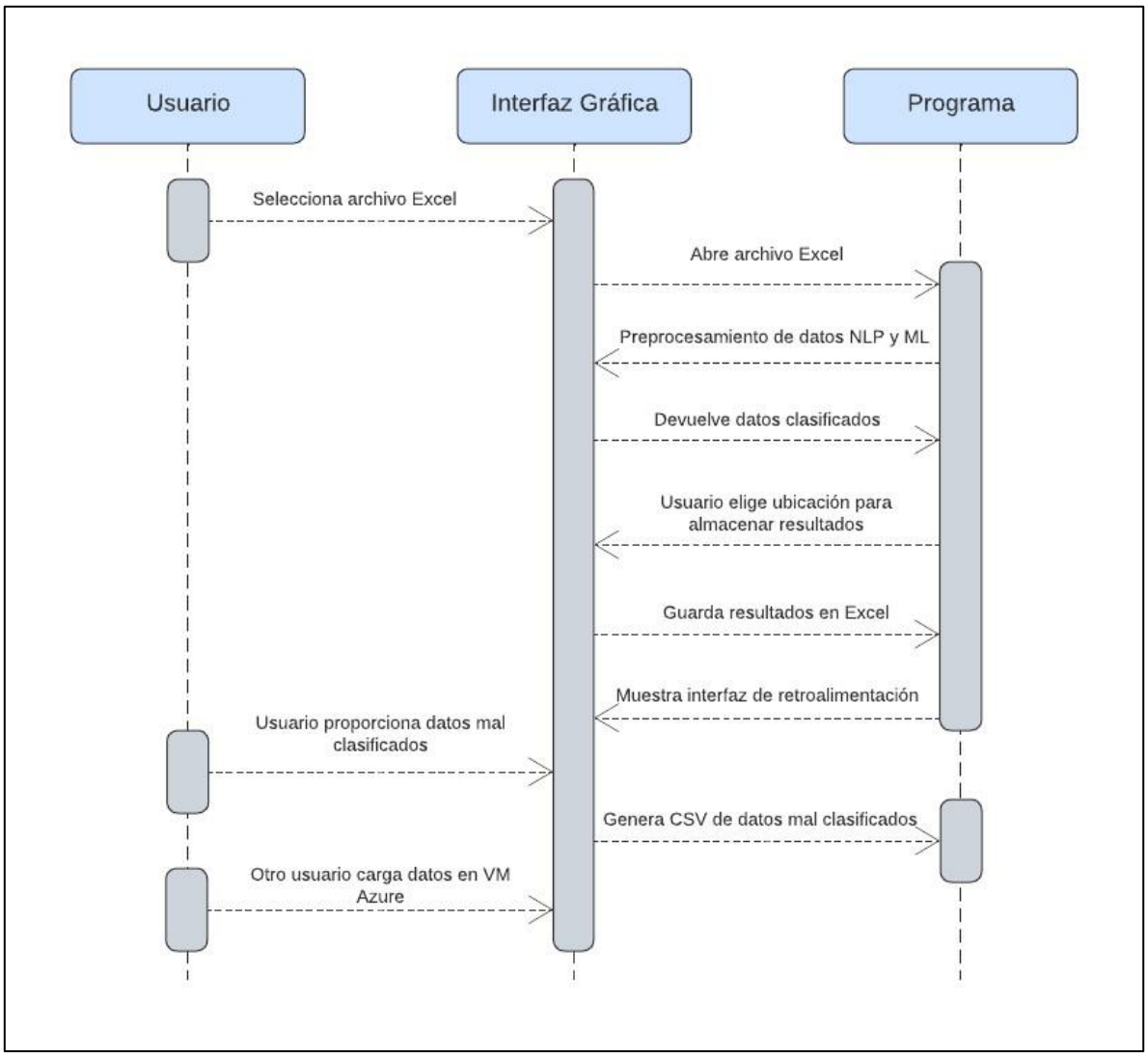

Figura 7: Diagrama de Secuencia UML

El diagrama de secuencia propuesto en la Figura 7 tiene como objetivo visualizar de manera clara y detallada la interacción entre los distintos componentes de tu sistema durante la ejecución del programa. Este programa tiene como función principal la clasificación de datos mediante algoritmos de Procesamiento de Lenguaje Natural (NLP) y Deep Learning (DL) a partir de un archivo Excel proporcionado por el usuario. La interfaz gráfica actúa como intermediario entre el usuario y el programa, permitiendo la selección del archivo de entrada, la elección de la ubicación para almacenar los resultados clasificados y la facilitación de la retroalimentación.

El proceso de retroalimentación permite al usuario corregir clasificaciones incorrectas proporcionando datos mal clasificados, lo que lleva a la generación de un archivo CSV.

Posteriormente, otro usuario carga este archivo en la máquina virtual de Azure para entrenar el modelo con nuevos datos, mejorando así su capacidad de clasificación.

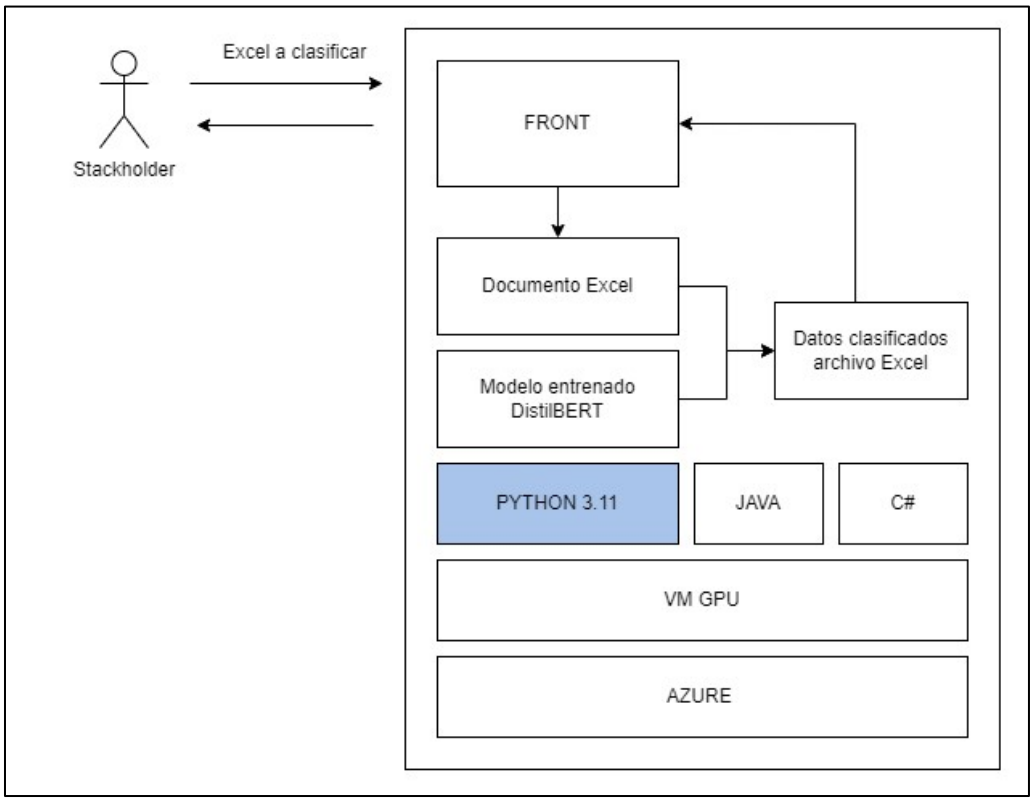

Figura 8: Diagrama de arquitectura e infraestructura de datos

La Figura 8 detallada el sistema de arquitectura e infraestructura de datos, el usuario inicia el proceso entregando un archivo Excel con los datos a clasificar. Estos datos son procesados por un componente de Front-end, que actúa como interfaz entre el usuario y el sistema. A continuación, el sistema utiliza un modelo preentrenado DistilBERT, implementado en Python 3.11, para clasificar los datos de manera eficiente. Este modelo, entrenado previamente en una máquina virtual con GPU en Azure, aprovecha la potencia de cómputo acelerado para mejorar su rendimiento. Posteriormente, los datos clasificados son organizados y devueltos al usuario en un nuevo archivo Excel, cerrando así el ciclo de procesamiento. La estructura modular del sistema, que incluye componentes en Python, Java y C#, junto con la integración de tecnologías como VM con GPU y Azure, garantiza un flujo de trabajo eficiente y escalable en el procesamiento y clasificación de datos.

En resumen, el proyecto se desarrolló de manera integral, incorporando una estrategia de división de datos bien fundamentada y adaptada al contexto específico del problema de clasificación de eventos de fallas en la industria minera.

9. Resultados Cualitativos y Cuantitativos

El código de autoclasificación está desarrollado en su totalidad permitiendo clasificar textos en cinco categorías distintas: "Comentario", "Nivel_1", "Nivel_2", "Nivel_3" y "Tipo de Detención". Se ha logrado una precisión promedio del 77% en la clasificación de eventos de fallas en distintas categorías, siendo esta una métrica cuantitativa clave para evaluar el rendimiento del modelo. Cabe destacar, que se observa una variación en la precisión de las categorías debido a la disponibilidad limitada de datos etiquetados para cada una de ellas. En particular, la precisión es más alta en la categoría "Tipo de Detención" con un 89%, mientras que “Comentario”, "Nivel_2" y "Nivel_3" tienen una precisión de tan solo 75%, 70% y 76%, respectivamente. Esta disparidad en la precisión se debe a la cantidad limitada de ejemplos disponibles para que el modelo aprenda de manera efectiva.

En una primera instancia, un nivel de precisión del 85% se considera más que adecuado para respaldar las operaciones y la toma de decisiones de la empresa en el corto plazo. Aunque este porcentaje podría no ser considerado alto en términos generales, representa un punto de partida sólido para la implementación inicial de herramientas basadas en NLP en la empresa.

No obstante, este porcentaje no fue alcanzado, debido a la cantidad limitada de ejemplos etiquetados en ciertas categorías y la naturaleza de los datos, esto porque, los comentarios presentan características que dificultan su análisis, ya que son incompletos, poco estructurados y, en la mayoría de los casos, están mal escritos. Se observa la presencia frecuente de abreviaciones y faltas de ortografía, lo cual añade complejidad al proceso de entrenamiento del modelo. Estas particularidades en la calidad de los datos impactan negativamente en la capacidad del sistema para aprender patrones con precisión y generalizarlos a nuevas instancias. Aunque no se logró alcanzar la meta inicialmente establecida, se evidenció un ahorro considerable en el proceso de clasificación manual de datos, ya que solo es necesario revisar el 23% de los mismos en vez del 100%.

Para alcanzar la meta deseada o superior en todas las categorías, es fundamental adoptar un enfoque de mejora continua. Este enfoque implica suministrar al modelo una cantidad adicional de

datos debidamente etiquetados, lo cual posibilitará que el sistema aprenda de una variedad más extensa de ejemplos. Una recomendación para mejorar la calidad de los datos etiquetados en el futuro consiste en implementar un sistema que requiera que los operadores utilicen un formato estructurado al proporcionar comentarios, esto garantizaría la entrega de información completa, precisa y redactada de manera adecuada, por el momento, se implementó un sistema de retroalimentación en donde el usuario le dice al programa cuales fueron los datos mal clasificados, para que de esta forma, el modelo pueda aprender de sus errores.

Se implementó además, una interfaz de usuario sencilla que permite al usuario ingresar los datos para su clasificación. Posteriormente, el usuario puede obtener un archivo Excel con los resultados, el cual puede ser guardado en la ubicación de su elección a través de una ventana emergente.

Por otro lado, el código logra clasificar aproximadamente 3000 datos en tan solo 7-8 minutos, marcando una reducción del 98% en el tiempo necesario para procesar la misma cantidad de datos en comparación con métodos anteriores. Si bien, el tiempo de depuración es un resultado que se espera ver a largo plazo, se anticipa una reducción estimada del 42%, pasando de 22.5 horas a 13 horas, lo que contribuirá significativamente a una resolución más veloz de problemas en la industria minera.

En cuanto al alcance de clientes, es importante señalar que este aspecto es de naturaleza a más largo plazo y, por lo tanto, aún no es posible medir su impacto de manera precisa en esta etapa del proyecto. Sin embargo, los avances obtenidos en la disminución del tiempo de clasificación son prometedores y sientan las bases para futuras mejoras en el alcance de clientes. Con la eficiencia actual, es factible que un único depurador pueda gestionar a uno o más clientes, lo que representa un cambio significativo con respecto a las prácticas actuales, donde generalmente se requieren de 2 o 3 depuradores para un solo cliente. Este avance potencialmente transformador respalda el objetivo de mejorar la eficacia y la rentabilidad de la empresa, al mismo tiempo que mejora la calidad de sus servicios de clasificación y depuración de texto.

Para obtener una comprensión más detallada de los resultados obtenidos y del funcionamiento del código, se pueden revisar el anexo 3 y el enlace de GitHub ubicado en la bibliografía. Estos proporcionan información adicional sobre la clasificación de los datos y ofrecen una visión más profunda de la funcionalidad del código.

10. Conclusiones y Discusión

El desarrollo de este proyecto de autoclasificación de eventos de fallas en la industria minera ha sido una experiencia enriquecedora y educativa. A lo largo de este proceso, se han obtenido valiosos aprendizajes que abarcan desde la preparación de los datos hasta la implementación de un sistema basado en modelos de procesamiento de lenguaje natural (NLP) y aprendizaje automático (ML). A continuación, se destacan las lecciones aprendidas y las reflexiones obtenidas durante este proyecto.

Se ha reforzado la crítica importancia de la preparación exhaustiva de los datos. La calidad de los resultados finales depende en gran medida de la calidad y consistencia de los datos de entrada. Las etapas iniciales de limpieza, identificación y manejo de valores atípicos y normalización fueron fundamentales para el éxito del modelo. La comparación y evaluación de varios modelos de procesamiento de lenguaje natural han demostrado que la elección del modelo es un aspecto crítico del proceso. La eficiencia computacional, el tiempo de entrenamiento y la precisión son factores que deben considerarse cuidadosamente al seleccionar un modelo. En nuestro caso, la elección de DistilBERT se basó en un equilibrio óptimo entre estos factores.

La naturaleza de los comentarios, que a menudo son incompletos, poco estructurados y mal escritos, presentó desafíos significativos en el entrenamiento del modelo. Las abreviaciones y errores ortográficos fueron obstáculos adicionales. Esto subraya la importancia de la mejora continua en la calidad de los datos y la necesidad de involucrar a los usuarios en la generación de datos más estructurados. La implementación de un enfoque iterativo, que incorpora retroalimentación del usuario, ha demostrado ser esencial. La capacidad del modelo para aprender de sus errores a través de la retroalimentación directa del usuario ha contribuido significativamente a la mejora continua de la precisión y relevancia del modelo.

La elección de implementar la solución en la nube, específicamente en la plataforma Azure, demostró ser acertada en términos de escalabilidad y acceso a recursos eficientes. La integración con la infraestructura existente de la empresa ha proporcionado estabilidad y confiabilidad. La constante búsqueda de un equilibrio entre la precisión del modelo y la eficiencia computacional fue un desafío constante. La selección de DistilBERT fue respaldada por su capacidad para ofrecer altos niveles de precisión con eficiencia en el tiempo de entrenamiento y procesamiento.

Se identificó la cantidad limitada de datos etiquetados en ciertas categorías como uno de los principales desafíos. En futuras iteraciones, se recomienda abordar esto mediante la generación activa de datos etiquetados y la implementación de un sistema estructurado para recopilar comentarios de los operadores. La mejora continua es esencial para alcanzar y superar los objetivos de precisión. Se sugiere establecer un ciclo constante de retroalimentación del usuario y ajuste del modelo para abordar nuevas clases y patrones que puedan surgir con el tiempo. La recomendación de implementar un sistema que requiera que los operadores utilicen un formato estructurado para proporcionar comentarios es una sugerencia valiosa para mejorar la calidad de los datos de entrada.

La interfaz de usuario actual sirve bien a su propósito, pero se reconoce la necesidad de mejoras estéticas y funcionales a largo plazo. Se planea implementar una interfaz más avanzada para mejorar la interacción del usuario.

Por otro lado, la evaluación del impacto a largo plazo en términos de eficiencia y alcance de clientes es una tarea que requerirá tiempo. Se propone realizar un seguimiento continuo de estos aspectos para evaluar el rendimiento y la adopción a medida que se implementan nuevas funcionalidades y mejoras.

En resumen, este proyecto no solo ha llevado a la implementación exitosa de un sistema de autoclasificación, sino que también ha proporcionado lecciones valiosas sobre la complejidad y la dinámica del procesamiento de lenguaje natural en un contexto industrial específico. La mejora continua y la adaptabilidad serán clave para mantener la relevancia y eficacia del sistema a medida que evoluciona y se enfrenta a nuevos desafíos en la clasificación de eventos de fallas.

11.Bibliografía

- Documentos

IT Digital Media Group. (2023, 10 febrero). La inversión de las empresas que ya emplean tecnología NLP irá a más en 2023. Actualidad | IT User. https://www.ituser.es/actualidad/2023/02/la-inversion-de-las-empresas-que-ya-emplean-tecnologia-nlp-ira-a-mas-en-2023

Olsson, F. (2009). A literature survey of active machine learning in the context of natural language processing. https://www.diva-portal.org/smash/get/diva2:1042586/FULLTEXT01.pdf

Gupta, R. K., & Yang, Y. (2019). Predicting and understanding news social popularity with emotional salience features. MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia, 139–147. https://doi.org/10.1145/3343031.3351048

Namous, F., Rodan, A., & Javed, Y. (2019). Online News Popularity Prediction. ITT 2018 - Information Technology Trends: Emerging Technologies for Artificial Intelligence, Itt, 180–184. https://doi.org/10.1109/CTIT.2018.8649529

Piotrkowicz, A., Dimitrova, V., & Markert, K. (2017). Automatic extraction of news values from headline text. 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of the Student Research Workshop, 64–74. https://doi.org/10.18653/v1/e17-4007

- Diagramas de Secuencia

Lucid. (s. f.). Intelligent Diagramming | LucidChart. Lucidchart. https://www.lucidchart.com/

- Modelo preentrenado DistilBERT en Hugging Face

Dccuchile/distilbert-base-Spanish-uncased-finetuned-ner · Hugging face. (s. f.). https://huggingface.co/dccuchile/distilbert-base-spanish-uncased-finetuned-ner

- Enlace de GitHub con el código de autoclasificación DistilBERT

Rfeuereisen. (s. f.). GitHub - rFeuereisen/DistilBERT-Text_Classifier: Modelo de clasificador de texto BERT. GitHub. https://github.com/rfeuereisen/DistilBERT-Text_Classifier

12.Anexos

	Año					
	0	1	2	3	4	5
Ingresos por ventas		$ 47.938.639	$ 57.287.574	$ 65.637.323	$ 72.103.605	$ 75.941.433
Ingresos Adicionales por Incremento de Clientes (25%)		$ 11.984.660	$ 14.321.894	$ 16.409.331	$ 18.025.901	$ 18.985.358
Costos Operacionales	$ -	$ -16.086.791	$ -18.431.465	$ -20.247.247	$ -21.324.939	$ -21.534.029
Costos por Administración y Ventas	$ -	$ -11.490.565	$ -13.165.332	$ -14.462.319	$ -15.232.099	$ -15.381.449
Resultado Operacional	$ -	$ 20.361.282	$ 25.690.777	$ 30.927.757	$ 35.546.567	$ 39.025.955
Utilidad antes Impuestos	$ -	$ 20.361.282	$ 25.690.777	$ 30.927.757	$ 35.546.567	$ 39.025.955
Impuesto de Primera Categoría (27%)	$ -	$ -5.497.546	$ -6.936.510	$ -8.350.494	$ -9.597.573	$ -10.537.008
Utilidad después Impuestos	$ -	$ 14.863.736	$ 18.754.267	$ 22.577.262	$ 25.948.994	$ 28.488.947
Inversión	$ -20.000.000	$ -	$ -	$ -	$ -	$ -
Capital de trabajo	$ -6.750.000	$ -	$ -	$ -	$ -	$ -
Recuperación del capital de trabajo	$ -	$ -	$ -	$ -	$ -	$ 6.750.000
Flujo de Caja Privado	$ -26.750.000	$ 14.863.736	$ 18.754.267	$ 22.577.262	$ 25.948.994	$ 35.238.947

Tasa de inflación (Banco Central)	0,043
WACC (Tasa de descuento)	0,12

Valor Actual Neto (VAN)	$ 54.028.609
Tasa Interna de Retorno (TIR)	66%

Anexo 1: Cálculo de VAN y TIR con tasa de descuento del 12% (Se asume un incremento del 25% de clientes por año)

	Año					
	0	1	2	3	4	5
Ingresos por ventas		$ 47.938.639	$ 57.287.574	$ 65.637.323	$ 72.103.605	$ 75.941.433
Ingresos Adicionales por Incremento de Clientes (25%)		$ 11.984.660	$ 14.321.894	$ 16.409.331	$ 18.025.901	$ 18.985.358
Costos Operacionales	$ -	$ -16.086.791	$ -18.431.465	$ -20.247.247	$ -21.324.939	$ -21.534.029
Costos por Administración y Ventas	$ -	$ -11.490.565	$ -13.165.332	$ -14.462.319	$ -15.232.099	$ -15.381.449
Resultado Operacional	$ -	$ 20.361.282	$ 25.690.777	$ 30.927.757	$ 35.546.567	$ 39.025.955
Utilidad antes Impuestos	$ -	$ 20.361.282	$ 25.690.777	$ 30.927.757	$ 35.546.567	$ 39.025.955
Impuesto de Primera Categoría (27%)	$ -	$ -5.497.546	$ -6.936.510	$ -8.350.494	$ -9.597.573	$ -10.537.008
Utilidad después Impuestos	$ -	$ 14.863.736	$ 18.754.267	$ 22.577.262	$ 25.948.994	$ 28.488.947
Inversión	$ -20.000.000	$ -	$ -	$ -	$ -	$ -
Capital de trabajo	$ -6.750.000	$ -	$ -	$ -	$ -	$ -
Recuperación del capital de trabajo	$ -	$ -	$ -	$ -	$ -	$ 6.750.000
Flujo de Caja Privado	$ -26.750.000	$ 14.863.736	$ 18.754.267	$ 22.577.262	$ 25.948.994	$ 35.238.947

Tasa de inflación (Banco Central)	0,043
WACC (Tasa de descuento)	0,15

Valor Actual Neto (VAN)	$ 47.557.231
Tasa Interna de Retorno (TIR)	66%

Anexo 2: Cálculo de VAN y TIR con tasa de descuento del 15% (Se asume un incremento del 25% de clientes por año)

Texto	Comentario		Nivel 1		Nivel 2		Nivel 3		Tipo de Detencion	
	Real	Auto	Real	Auto	Real	Auto	Real	Auto	Real	Auto
alarma nivel ac. motor diesel Motor Sistema de Lubricacion	BAJO NIVEL	BAJO NIVEL	1	MOTOR	MOTOR	1	LUBRICACION_MOTOR	LUBRICACION_MOTOR	1	CARTER	CARTER	1	MCM	MCM	1
aplicacion freno estacionamie Sistema Direccion y Frenos Sistema Frenos	FALLA CABLE/	CHEQUEO	0	FRENOS	FRENOS	1	ELECTRICO_FRENOS	HIDRAULICO_FRENOS	0		ACUMUL	0	MCE	MCM	0
baja potencia motor diesel Motor Sistema de Combustible	BAJA POTENC	BAJA POTENC	1	MOTOR	MOTOR	1	ADMISION_Y_ESCAPE	COMBUSTIBLE_MOTOR	0	DUCTO_A	INYECTOF	0	MCM	MCM	1
c/ sensor filtro aire Motor Sistema de Admision y Escape	FALLA SENSO	FILTRO SATUF	0	MOTOR	MOTOR	1	ELECTRICO_MOTOR	ELECTRICO_MOTOR	1	SENSORE	SENSORE	1	MCE	MCE	1
cheq suspensiones Sistema Suspension Suspensiones Delanteras	CHEQUEO	CHEQUEO	1	MAQUINA	MAQUINA	1	SUSPENSION	SUSPENSION	1	SUSPENS	SUSPENS	1	MCM	MCM	1
fuga aceite motor Motor Sistema de Lubricacion	FUGA	FUGA	1	MOTOR	MOTOR	1	LUBRICACION_MOTOR	LUBRICACION_MOTOR	1		CARTER	0	MCM	MCM	1
juego en pasadores de tolva Estructura Estructura y Chasis	FALLA PASAD	FUERA DE AJU	0	MAQUINA	MAQUINA	1	MAQUINA_ESTRUCTURA	MAQUINA_ESTRUCTURA	1	TOLVA	TOLVA	1	MCM	MCM	1
pm Mantencion Programada PM	PM	PM	1	PM	PM	1	PM	PM	1			1	MP	MP	1
regulacion puerta Estructura Cabina y Plataforma	FUERA DE AJU	FUERA DE AJU	1	MAQUINA	MAQUINA	1	CABINA	CABINA	1	PUERTA	PUERTA	1	MCM	MCM	1
alarma temperatura buje alter Instr. Electricos y Comp. Control Sistema de	ALTA TEMPER	ALTA TEMPER	1	MOTOR	MOTOR	1	ELECTRICO_MOTOR	ELECTRICO_MOTOR	1	ALTERNA	SENSORE	0	MCE	MCE	1
alta temp aceite Instr. Electricos y Comp. Control Dispositivos Proteccion y	ALTA TEMPER	ALTA TEMPER	1	MOTOR	TREN_DE	0	LUBRICACION_MOTOR	TRANSMISION	0	ENFRIADO	SENSORE	0	MCM	MCM	1
cambio inyector Motor Sistema Partida Electrico	FALLA FUNCIO	FALLA FUNCIO	1	MOTOR	MOTOR	1	COMBUSTIBLE_MOTOR	COMBUSTIBLE_MOTOR	1	INYECTOI	INYECTOF	1	MCM	MCM	1
relleno refrigerante Motor Sistema de Enfriamiento	DESGASTE	BAJO NIVEL	0	NEUMATI	MOTOR	0	NEUMATICOS_Y_AROS	ENFRIAMIENTO_MOTOR	0		LINEAS_E	0	MCV	MCM	0
trochas Instr. Electricos y Comp. Control Instrumentos Cables y Luces	LUCES	CORTO CIRCU	0	MAQUINA	MAQUINA	1	ELECTRICO_MAQUINA	ELECTRICO_MAQUINA	1	FOCOS	FOCOS	1	MCE	MCE	1
presion acumulador freno. Sistema Direccion y Frenos Sistema Frenos	BAJA PRESIO	BAJA PRESIO	1	FRENOS	FRENOS	1	HIDRAULICO_FRENOS	HIDRAULICO_FRENOS	1	ACUMUL	ACUMUL	1	MCM	MCM	1
falla de inyectores Motor Sistema de Combustible	FALLA FUNCIO	FALLA FUNCIO	1	MOTOR	MOTOR	1	COMBUSTIBLE_MOTOR	COMBUSTIBLE_MOTOR	1	INYECTOI	INYECTOF	1	MCM	MCM	1

Anexo 3: Hoja de cálculo con textos clasificados (Donde 1 es clasificación correcta y 0 clasificación incorrecta)