UNIVERSDIDAD DE CONCEPCIÓN

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

Entrenamiento de un Modelo de Lenguaje Natural para clasificar proyectos evaluados en el programa Capital Semilla Emprende de Sercotec

POR

Edgardo Benjamín Cea Salamanca

Memoria de Título presentada a la Facultad de Ingeniería de la Universidad de Concepción para optar al título profesional de Ingeniero Civil Industrial

Profesores Guía

Marcela Parada Contzen
Juan Carlos Caro Seguel

Agosto 2024
Concepción, Chile

Dedicatoria

A mi madre, Marisol Salamanca Hernández.

A mi prima, Ana María Berrios.

A mi tío, Ivor Molina Salamanca (Q.E.P.D).

Agradecimientos

A Dios, por permitirme cumplir mi sueño.

A mi madre, Marisol, quien, con su sacrificio y amor incondicional, me ha apoyado en cada paso de este camino.

A mi prima, Ana María, por ser una fuente de inspiración, por instarme siempre a perseverar, y por ayudarme en todo momento.

A toda mi familia, especialmente a mis tíos y primos, por su sincero cariño.

A los amigos que hice en la universidad, en especial a los miembros de Desastres Fútbol Club.

A mis profesores, por su dedicación y paciencia.

A todas las personas que, de alguna manera, contribuyeron en mi vida a lo largo de la carrera.

Índice General

Índice de Tablas

Índice de Figuras

Resumen

El Servicio de Cooperación Técnica (Sercotec) es un organismo dependiente del Ministerio de Economía que apoya a pequeños emprendedores a través de programas concursables. Actualmente, Sercotec enfrenta una gran carga al evaluar las postulaciones a sus programas, especialmente en el Programa Capital Semilla Emprende (PCSE). Esta evaluación, que está a cargo de funcionarios, se ve ralentizada por la gran cantidad de postulaciones.

Para abordar la problemática de Sercotec, se propone un modelo basado en aprendizaje automático para la revisión y evaluación de postulaciones. Este modelo emplea un algoritmo que recopila los datos de las postulaciones, evalúa las respuestas correspondientes al Modelo Canvas en el formulario, y clasifica el texto asignando una calificación acorde a la calidad de cada respuesta.

El modelo es entrenado con datos reales de Sercotec sobre la base de BETO bajo hiperparámetros ajustados por el optimizador Optuna, donde BETO es una versión en español del modelo de lenguaje pre-entrenado BERT. Dicho modelo se valida con el método de Matriz de Confusión.

El modelo está restringido a las nueve secciones clásicas del Canvas, excluyendo Sustentabilidad, Coherencia global de respuestas y otros elementos del formulario de postulación. Además, enfrenta desafíos relacionados con el procesamiento de texto debido a errores en los datos reales y está limitado por la capacidad computacional disponible, lo que restringe los hiperparámetros y el volumen de datos utilizados en el entrenamiento.

Una futura integración de este modelo debería agilizar la evaluación y, en consecuencia, la selección de solicitudes para PCSE, haciéndolo un proceso más eficiente en términos de tiempo y esfuerzo, al mismo tiempo que se elimina el factor de error humano en esta tarea.

Abstract

The Technical Cooperation Service (Sercotec) is an agency dependent on the Ministry of Economy that supports small entrepreneurs through competitive programs. Currently, Sercotec faces a great burden when evaluating applications for its programs, especially in the Capital Semilla Emprende Program (PCSE). This evaluation, which is carried out by officials, is slowed down by the large number of applications.

To address Sercotec's problem, a model based on machine learning is proposed for the review and evaluation of applications. This model uses an algorithm that collects data from applications, evaluates the responses corresponding to the Canvas Model in the form, and classifies the text by assigning a rating according to the quality of each response.

The model is trained with real data from Sercotec based on BETO under hyperparameters adjusted by the Optuna optimizer, where BETO is a Spanish version of the BERT pre-trained language model. This model is validated with the Confusion Matrix method.

The model is restricted to the nine classic Canvas sections, excluding Sustainability, Global Response Consistency, and other elements of the application form. In addition, it faces challenges related to text processing due to errors in the real data and is limited by the available computational capacity, which restricts the hyperparameters and volume of data used in training.

A future integration of this model should streamline the evaluation and, consequently, the selection of applications for PCSE, making it a more efficient process in terms of time and effort, while eliminating the human error factor in this task.

1. Introducción

El Servicio de Cooperación Técnica (Sercotec) es una corporación de derecho privado dependiente del Ministerio de Economía. Su misión consiste en brindar apoyo a emprendedores, así como a micro y pequeñas empresas, con el objetivo de fomentar su desarrollo y contribuir al crecimiento del país (Sercotec, 2023a). Para lograrlo, Sercotec ofrece asesorías, capacitaciones y financia proyectos a través de fondos concursables, permitiendo a los usuarios postular a programas destinados a las áreas Crecer y Fortalecer. Dentro de estos programas se encuentra el Programa Capital Semilla Emprende (PCSE), que es donde se centra esta investigación.

El Programa Capital Semilla Emprende tiene las barreras de postulación más bajas de Sercotec, ya que solo requiere siete requisitos base (Sercotec, 2022a), a diferencia de otros programas que exigen requisitos adicionales, como Capital Abeja Emprende donde solo pueden postular mujeres (Sercotec, 2022b) o el programa Crece que exige una formalización de la empresa (Sercotec, 2023b). Debido a esto, en cada convocatoria se registra una cantidad masiva de postulaciones. Por ejemplo, para la Región del Biobío, y solo considerando el año 2022, ingresaron más de 1400 postulaciones (Sercotec, 2023c). El número de postulaciones plantea un desafío para la institución, ya que la cantidad de solicitudes a evaluar supera la capacidad del servicio pues la calificación de cada postulación se realiza de manera manual por un funcionario de Sercotec.

La revisión manual de postulaciones puede generar complicaciones, algunas relacionadas con el bienestar del trabajador, como el estrés y la desmotivación, y otras relacionadas con el tiempo invertido y los posibles errores debido a la monotonía del trabajo. Un error en la evaluación de formularios puede tener consecuencias perjudiciales, ya que podría resultar en la exclusión de una persona con una sólida solicitud de la convocatoria (El Empleo, 2023).

Ante ello, y para el desarrollo de esta investigación, se propone la implementación de un modelo que califique las nueve secciones del Modelo Canvas utilizado en el formulario de postulación de Sercotec, que es el ítem más relevante en los formularios de postulación al programa. Este es un prototipo predictivo de evaluación en base a BETO, que es una versión en español de BERT, un modelo procesamiento de lenguaje natural (PLN) de tipo *Transformer* pre-entrenado (Conejeros, 2023). BETO es entrenado con respuestas reales dadas por los usuarios en antiguas postulaciones junto a su respectiva calificación, y luego es capaz de generar una calificación (1, 3, 5 o 7 según corresponda) de manera automática para cada sección del Canvas.

El modelo propuesto permite mejorar la evaluación de formularios, ya que el proceso es mucho más eficiente al calificar un mayor número de formularios en menos tiempo. Además, liberaría horas del personal para otras tareas disponibles. Importante es notar que los modelos PLN Transformers, como BETO, son los que mejores resultados entregan, ya que tienen la cualidad analizar todos los *tokens* ingresados (Donoso, 2021). Los buenos resultados hacen que sean modelos fáciles de replicar en diferentes procesos de evaluación y/o selección. En Chile ya se han aplicado este tipo de soluciones, como el trabajo de Álvarez (2021), donde lo utilizó en encuestas de satisfacción en la industria del *retail*.

Antes de llevar a cabo el entrenamiento del modelo BETO, es fundamental comprender tanto el tipo de respuestas que se espera que ingrese, como las respuestas que el modelo debería generar. Por esta razón, se realiza un pequeño análisis exploratorio de datos (AED) utilizando los datos proporcionados por Sercotec. Luego, se deben generar las calificaciones mediante la tokenización de las palabras ingresadas. Además, se trabaja con Optuna para mejorar la calidad de las notas puestas mediante la configuración de hiperparámetros. Posteriormente, se valida el modelo con la toma de datos fuera de muestras a través de una matriz de confusión.

1.1 Objetivos de la memoria

1.1.1 Objetivo General

- Entrenar un modelo basado en procesamiento de lenguaje natural para clasificar las respuestas al Canvas de los postulantes al Programa Capital Semilla Emprende de Sercotec.

1.1.2 Objetivos Específicos

- Revisar y analizar la literatura relacionada al modelo BETO, su aplicación en clasificación de texto, y estudiar el código propuesto con datos simulados.
- Analizar y depurar la data con las respuestas reales que dieron usuarios del programa Capital Semilla Emprende, y sus utilizarlos en el modelo propuesto.
- Extender el código base desde la sección clientes hacia las ocho secciones restantes del Canvas de las postulaciones.
- Iterar la búsqueda de resultados del modelo con una configuración de hiperparámetros con el optimizador Optuna.
- Validar el modelo con datos fuera de muestra a través de una matriz de confusión.

1.2 Alcances y limitaciones

1.2.1 Alcance

El alcance de esta memoria de título implica la formulación y entrenamiento de un modelo computacional con base en BETO. Su finalidad es la evaluación y calificación automática de las nueve secciones del Canvas, para los formularios que postulan al programa Capital Semilla Emprende de Sercotec. El entrenamiento es realizado a partir de datos reales, y son recogidos mediante Python. Para mejorar la calidad de los resultados, se utiliza Optuna, que busca

respuestas con buena precisión y eficiencia según la configuración de hiperparámetros dada.

1.2.2 Limitaciones

Para esta Memoria de Título de consideran las siguientes tres limitaciones:

- El modelo es limitado a evaluar las respuestas del modelo Canvas del formulario de postulación al programa CSE, en particular a la evaluación de las nueve secciones clásicas. No se aborda la sección 10 "sustentabilidad", debido a que existen pocos datos respecto al resto de secciones, ni la sección 11 "coherencia global de la idea de negocio", puesto que es una nota extra que no depende de una respuesta puntual (ver en Anexo A, Tabla 25). Tampoco se evalúan otros ítems del formulario de postulación como el video pitch.
- Al trabajar con datos reales, se tiene como limitante la capacidad de BETO, en cuanto al procesamiento de texto, debido al ingreso de caracteres extraños, palabras mal escritas, mala redacción, o alguna otra anomalía no considerada por BETO durante el pre-entrenamiento, que pueda repercutir en malas predicciones.
- Se tiene como un limitante crítico la capacidad computacional que requiere el entrenamiento y testeo del modelo. Para evaluar respuestas reales con textos extensos, se necesita una potencia mínima en la generación de redes neuronales de *deep learning*. Con lo que la capacidad del *hardware* (CPU, GPU y memoria RAM), limita la configuración entre el tamaño máximo del paquete de datos que se ingresa, y los hiperparámetros con los que se realiza el entrenamiento.

1.3 Organización del Documento

La presente memoria de título es organizada de la siguiente manera:

- Capítulo 1: Introducción. Se expone el tema central de la memoria de título en conjunto a sus respectivos objetivos (general y específicos), además del alcance y limitaciones que la memoria considera.
- Capítulo 2: Revisión bibliográfica. Se exploran los conceptos teóricos y esenciales para contextualizar bien esta investigación. Esto incluye datos sobre Sercotec y Capital Semilla, y estudios sobre *machine learning*, procesamiento de lenguaje natural, BERT y BETO.
- Capítulo 3: Metodología. Se establece el modelo seleccionado para el estudio, y las diferentes metodologías para el entrenamiento de este. Se abarca el tratamiento de los datos reales, la implementación del código de BETO y del optimizador Optuna, la validación del modelo, y como se extiende a todas las secciones del Canvas.
- Capítulo 4: Datos. Se concentra en seleccionar, depurar y estudiar los datos que se utilizan durante el entrenamiento. Los datos no seleccionados quedan como fuera de muestra para realizar la validación del modelo.
- Capítulo 5: Resultados. Se presentan resultados obtenidos con la aplicación de BETO, junto a un análisis de calidad de las calificaciones generadas.
- Capítulos 6: Conclusiones. Se presentan las conclusiones del estudio respecto a generalidades de las inteligencias artificiales y herramientas tecnológicas, y del modelo presentado en particular.

2. Revisión bibliográfica

Para abordar de manera integral el contexto del estudio, es fundamental explorar tanto los aspectos relacionados con el proceso de postulación a programas de apoyo al emprendimiento, como el uso de tecnologías avanzadas que pueden ayudar en la evaluación de dichas propuestas.

Este capítulo analiza dos áreas clave: el proceso de postulación al PCSE de Sercotec, y la aplicación de la inteligencia artificial en la clasificación de texto. Primero, se explora cómo Sercotec emplea el Modelo de Negocios Canvas para evaluar propuestas emprendedoras. Luego, se explora la evolución de la IA y su implementación en Chile, incluyendo el procesamiento de lenguaje natural y el uso de modelos avanzados como BETO. Finalmente, se introduce Optuna, una herramienta diseñada para optimizar hiperparámetros y mejorar la eficiencia en el entrenamiento de modelos de *machine learning* y *deep learning*.

2.1 Sercotec y Canvas

2.1.1 Programa Capital Semilla Emprende de Sercotec

Sercotec, es una corporación dependiente del Ministerio de Economía. Se dedica a apoyar a las micro y pequeñas empresas (Mipymes) a través de financiamiento y asesorías. Trabajan con pequeños emprendedores que buscan concretar sus proyectos de negocio a lo largo del país (Sercotec, 2023a).

Por cumplir su misión, Sercotec ha creado diferentes programas, otorgando apoyo a quién cumplan ciertos requisitos. Aquí nace el Programa Capital Semilla Emprende (PCSE), cuyo enfoque es dar el impulso inicial a las ideas de negocios de los beneficiarios. Los negocios seleccionados pueden acceder a un monto de entre $3 y $3.5 millones de pesos para inversión en bienes y acciones de gestión empresarial. Estos pueden ser utilizados en mejoras de infraestructura, compra de materiales, acciones de marketing, capacitación, entre otros (Sercotec, 2022a). Desde el año 2012, ha beneficiado a más de 13

mil emprendedores y emprendedoras, por lo demás el número de postulaciones de cada año es inmenso.

Para ello, los postulantes deben cumplir ciertos requisitos, entre los que destacan los siguientes tres componentes: admisibilidad, evaluación técnica, y fase de desarrollo y formalización.

2.1.2 Modelo de Negocios Canvas

El Modelo de Negocio Canvas es una herramienta visual ampliamente utilizada en el mundo empresarial para describir, analizar y planificar un modelo de negocio de manera clara y concisa. Fue creado por Osterwalder & Pigneur (2009), desde entonces se ha convertido en un estándar en la estrategia empresarial. Canvas se utiliza para representar de manera gráfica los componentes clave de un negocio y cómo interactúan entre sí (Clavijo, 2024).

El Modelo Canvas consta de nueve bloques: segmentos de clientes, propuesta de valor, canales de distribución, relaciones con los clientes, fuentes de ingresos, recursos clave, actividades clave, socios clave y estructura de costos. Cada bloque se completa con información específica relacionada con el negocio. Permite a los emprendedores y empresarios comprender mejor su modelo de negocio, identificar áreas de mejora y comunicar eficazmente su estrategia a los demás.

A pesar de su amplia aceptación y utilidad, existen otros modelos que ayudan con el modelo de negocio. Una de las alternativas más conocidas es el Mapa de Empatía, que se centra en comprender profundamente las necesidades, deseos y emociones de los clientes (Miró, 2023).

Otra alternativa, es el Modelo Canvas Lean, o Canvas Inclinado, creado por Maurya (2012) en su libro *Running Lean*, que es una variante del modelo original de Osterwalder & Pigneur (2009). Se enfoca en aspectos clave como los problemas a resolver, las soluciones propuestas, las métricas clave, los

canales de adquisición de clientes, los ingresos esperados, los costos y los factores de riesgo (Van Zandt, 2023). El Lean Canvas enfatiza que una empresa nace de la identificación de un problema que puede resolverse de manera factible y donde alguien estaría dispuesto a pagar. Ambos modelos relacionan el concepto de problema con necesidades, aunque se mencionen indistintamente sin discutir sus implicaciones para la creación de nuevos productos y servicios (Mejía-Giraldo, 2019).

El Modelo de Negocio Canvas es una herramienta ágil y efectiva para analizar y comunicar modelos de negocio (Clavijo, 2024). Es por ello que Sercotec, y otras entidades similares, lo utilizan en sus convocatorias a fondos concursables, y así evalúan que tan bien estructurados están los proyectos de los postulantes, tal como lo hacen con el PCSE.

Zumarán & Cortés (2021), funcionarios del CFT PUCV, indican que el Modelo Canvas facilita la simplificación y proyección eficiente de ideas de negocios. Además, este modelo mejora la comprensión, amplía la perspectiva y permite un análisis estratégico del negocio. Instituciones como FOSIS, Sercotec y CORFO, que forman parte de la Red de Fomento del Estado, lo emplean en la elaboración de formularios para postular a sus fondos concursables.

Sercotec trabaja con una rúbrica propia para evaluar cada proyecto, esta se centra en si responde completa, parcial, superficialmente o no responde a cada pregunta respectiva al Modelo Canvas del proyecto puntual.

2.2 Inteligencia artificial y aprendizaje automático

2.2.1 Inteligencia artificial y la cuarta revolución industrial

En diversas disciplinas se hace referencia a la cuarta revolución industrial, este término se utiliza debido a la transformación digital impulsada por las nuevas tecnologías de procesamiento de datos y la automatización total de procesos (Schwab, 2016b). La Cuarta Revolución Industrial, es un fenómeno a nivel

mundial, que trae avances muy significativos, similar a lo que ocurrió con la industrialización a fines del siglo XVIII, tras la llegada de la máquina a vapor (Gayubas, 2017).

A diferencia de las revoluciones anteriores, esta vez la importancia no radica tanto en el desarrollo tecnológico en sí, sino en la gran interconexión de estas tecnologías entre sí, y con los usuarios. La cuarta revolución hace uso del internet, la robótica, el *big data*, los metadatos, la inteligencia artificial (IA), entre otros, que otorgan una capacidad de interacción que fusiona el mundo físico con el digital, lo que comúnmente se denomina como IoT, o el Internet de las Cosas (Porcelli, 2020).

Si bien la idea de la IA nació con el artículo de McCulloch & Pitts (1943), no fue hasta que McCarthy (1955) ocupó dicha terminología de forma oficial. La IA se planteó casi como una premisa de ciencia ficción, que con el paso de las décadas se asentó hasta los grandes desarrollos de hoy en día (Maguregui, 2023). Este fenómeno se ha destacado debido a la velocidad, amplitud y profundidad de impacto en los sistemas de información, lo que lo convierte en más que simplemente un avance tecnológico (E. Caro, 2017). Este avance tecnológico permite la automatización de proceso, como por ejemplo: el almacenamiento, clasificación, análisis y conexión de datos (Martineau, 2023).

La evolución de la IA inició con orientaciones distintos a los de hoy en día. El cambio de en enfoque en la IA, se da al pensar en ella como una máquina con la capacidad de procesar información similar a un cerebro humano. Esta premisa permitió un avance importante para la conceptualización de la IA actual, que al llegar al siglo XXI se empezó a emplear en problemas y soluciones más específicos (Mendoza, 2024). Al observar resultados evidenciables, se empezó a aplicar en diversas disciplinas de la informática, en conjunto al *machine learning* y otros métodos de aprendizajes (Arenas et al., 2023).

En el libro La cuarta revolución industrial de Schwab (2016), se nos plantean los diversos problemas éticos que conlleva el desarrollo de estas tecnologías, como en el área laboral, donde se perderán muchos puestos de trabajo. No obstante, el desarrollo de las IAs también propicia el nacimiento de otras profesiones relacionadas a estas áreas. Por otro lado, el uso desmedido de la IA, o su utilización con fines maliciosos, pueden acarrear consecuencias graves, como el incremento de una dependencia tecnológica, problemas con la privacidad de la información de las personas, o un traspaso de límites éticos que puedan afectar a alguna persona o comunidad en particular (ISDI, 2023). Es por lo anterior, que se ve una necesidad de discusión respecto al fenómeno de las IAs, a fin de establecer el uso de estas tecnologías de forma responsable para aprovechar de buena forma su potencial.

2.2.2 Aplicación de las IAs en Chile

Desde su popularización en los últimos años alrededor del mundo, el uso de la IAs se disparó. Hoy en día se están aplicando en distintos sectores industriales, con el fin de automatizar y optimizar procesos, bajando costos y tiempos (Baena, 2023).

En Chile, el escenario no es distinto. La introducción e implementación de estas nuevas tecnologías ha permitido que su uso se diversifique, y no solo eso, también se plantea reforzar la infraestructura informática actual y potenciar sus capacidades. A modo de ejemplo, el Ministerio de Ciencias en el año 2021 formuló una política de IAs, donde el foco de la discusión va desde "el impacto real que puede tener la automatización en el sistema laboral chileno" hasta "cómo la IA puede cambiar el concepto de realidad que existe en la sociedad, relativizando lo verdadero y lo falso" (El Mostrador, 2023).

Asimismo, el proyecto de ley N°15869-19 presentado el lunes 24 de abril de 2023 ante el Parlamento propone un marco regulatorio a los sistemas de IA y la robótica, además de otras tecnologías conexas en sus distintos ámbitos de aplicación (Cámara de Diputados, 2023). Este proyecto sigue en trámite en la

Cámara de Diputados, pero ejemplifica la relevancia que ha adquirido la IA en nuestro país (Cámara de Senadores, 2024).

Hace poco, en España se presentó la Nueva Estrategia de IA, donde el uso activo de la IAs es el eje, y cuyas aplicaciones buscan potenciar Pymes y nuevos proyectos tecnológicos (Moya, 2024). Chile está algo más avanzado, ya que en el 2021 se presentó el Plan de Acción de IA dispuesto por el Ministerio de Ciencia. Plan que promueve la inversión educacional con becas de estudio en el extranjero, la colaboración de I+D entre universidades y el sector productivo, y la generación de incentivos y de acuerdos público-privados. El plan cuenta con un presupuesto de $26 mil millones de pesos (Ministerio de Ciencia, 2021).

En la actualidad, el uso de las IAs, sobre todo las generativas, se ha incrementado, especialmente luego de la salida de herramientas como Chat GPT, Dall-E o Midjourney (Mittal, 2023). Debido al fácil uso de dichas IAs, se hace una tecnología accesible a un gran número de usuarios. Dentro del país un gran número de empresas han declarado usar IAs para generar valor, y es en el sector minero donde esta automatización de procesos se puede evidenciar con más ejemplos (Vera-Cruz, 2023).

Desde la ingeniería, el objetivo de trabajar con las IA es poder diseñar y crear sistemas inteligentes para utilizarlas en diversas aplicaciones, como lo son: Ventas y marketing, Manufactura, Recursos Humanos, Gestión, Estrategias y Finanzas, etc. (B. Caro, 2021). De este modo, el diseño y estrategias de posibles aplicaciones de las IA nace de cómo utilizarlas para "facilitar el trabajo más mecánico o repetitivo, con el fin de dedicar más tiempo y esfuerzos en aquello que es más importante y esencial" (Sanhueza, 2023).

En el artículo de Venegas (2021) sobre la IA en el área de la ingeniería, se muestra su utilización en el análisis de patrones discursivos a través de un sistema de Procesamiento de Lenguaje Natural. En particular, se utilizó un modelo de aprendizaje profundo en español BETO, que fue efectiva al trabajar

con grandes cantidades de texto, gracias a la auto capacidad adaptarse a modelos de clasificación avanzados (Cañete et al., 2023).

2.2.3 Procesamiento de Lenguaje Natural

El Procesamiento de Lenguaje Natural (PLN), o NLP en inglés, dentro de la IA se refiere al campo de conocimiento e investigación referentes al estudio y desarrollo de comunicación entre una IA y las personas mediante lenguajes naturales, es decir, el inglés, el español, el francés, etc. (Moreno, 2018). El PLN está relacionado directamente con la IA aplicada en *software* de uso cotidiano, como los traductores, *chatbot* o clasificadores de archivos (Holdsworth, 2024).

Según Álvarez (2021), podemos clasificar los conceptos fundamentales del PLN según la

Tabla 1. Como se observa en la tabla, cada proceso se ejecuta en conjunto a otro, es decir, es una cadena de tratamientos que se les da a un texto.

Tabla 1: Conceptos fundamentales del PLN

Concepto	Definición
Corpus	Se refiere a el conjunto de textos que se utilizan como set de datos para trabajar.
Preprocesamiento de texto	Para poder realizar un análisis del *corpus*, es necesario que éste sea limpiado y transformado. Lo que incluye la normalización del texto, es decir, eliminar caracteres especiales, corregir los errores ortográficos, esto para que el texto ingresado esté en un formato coherente.
Tokenización	En este proceso, el texto se divide en unidades más pequeñas llamadas tokens, pueden ser palabras individuales, sub-palabras y caracteres gramaticales como el uso del punto o la coma.
Vocabulario	Es el producto de los procesos anteriores, son todos los tokens resultados del preprocesamiento.
n-gram	Consiste en una secuencia de caracteres de largo n.
Stop worlds	Son las palabras que se remueven porque carecen de contenido informativo. Palabras como La o Que. Dichas palabras se remueven porque se repiten muchas veces dentro del *corpus* y generan ruido.
One-shot	Consiste en la asociación de relación uno a uno de los tokens del vocabulario de largo V, ordenados en un índice, a vectores en un espacio {0, 1} V con un 1 en la posición correspondiente al índice de cada token y ceros en el resto de los componentes del vector.

Fuente: Elaboración propia a partir de Álvarez (2021).

Con el tiempo se han creado modelos y estructuras que trabajan sobre esa base de cadena de tratamientos y los suplementan. Por ello, su evolución ha sido exponencial en los últimos años, debido a grandes avances y aportes al desarrollo de la IA (Giraldo & Orozco, 2023). Según Giraldo y Orozco (2023), los principales avances en PLN corresponden a los siguientes 3 desarrollos: *word embeddings*, la arquitectura *transformer*, y el modelo generativo multipropósito. Y es la arquitectura transformers, la base de los modelos como el que se emplea en este estudio.

Para desarrollar modelos avanzados, se utilizan redes neuronales que imitan el aprendizaje humano para clasificar y predecir datos (Venegas, 2021). Estas redes, organizadas en capas y basadas en *machine learning*, son capaces de aprender de la experiencia, generalizar casos previos y abstraer características clave de datos relevantes (Matich, 2001).

2.2.4 Machine learning

El *machine learning* (ML), según Mamani (2022), es una rama de la IA que usa computación científica, matemáticas y estadística para resolver problemas de clasificación, regresión y agrupamiento mediante técnicas automatizadas. En esencia, permite a las computadoras aprender de datos y cálculos anteriores para mejorar sus capacidades.

El ML trabaja mediante algoritmos para solucionar problemas o brindar una respuesta a través de una serie de instrucciones. Hoy es muy usada en distintas empresas a nivel mundial. Dos gigantes de la industria tecnológica, como lo son Meta (ex Facebook) y Google ocupan el ML para automatizar procesos a todo nivel. Por ejemplo, en sus soluciones publicitarias, que fueron desarrollando y orientando cada vez a mecanismos más inteligentes de selección de audiencias. En estos procesos, el operador sólo indica unos pocos datos, y la campaña se ejecuta con métodos de aprendizaje automatizado (Zelcer, 2022).

2.2.5 Deep learning

El *deep learning* (DL), o aprendizaje profundo, es una rama del ML, y por tanto de la IA, que se centra en la construcción y entrenamiento de redes neuronales artificiales profundas para resolver tareas complejas de procesamiento de datos (Alarcón, 2020). Según Burns (2021), el DL se considera como una forma de automatizar el análisis predictivo. En general, los algoritmos desarrollados en el ML son lineales, mientras que los algoritmos de DL se apilan en una jerarquía de complejidad y abstracción mayor, con múltiples capas (Burns, 2021).

El DL surgió como una evolución del ML tradicional en la década de 2010, y su popularidad creció gracias a los avances en hardware, grandes conjuntos de datos y técnicas de optimización (Dallas, 2024). Una de las principales diferencias entre el DL y el ML tradicional radica en la complejidad de las redes neuronales utilizadas en el aprendizaje (Gorini, 2024). Mientras que el ML suele emplear modelos más simples como regresión lineal o máquinas de soporte vectorial, el DL utiliza redes neuronales profundas de muchas capas, lo que permite capturar relaciones más complejas entre los datos (Sotaquirá, 2018). Es por esto que requiere grandes cantidades de datos para un entrenamiento efectivo (Alonso, 2024).

El uso de modelos entrenados por DL permite aprender características específicas, de manera similar a como lo hace el cerebro humano (UNIR, 2021). Estas peculiaridades, hace que dichos modelos sean particularmente adecuados extracción y reducción de características, por lo que es usado en aplicaciones de reconocimiento visual computarizado, análisis de texto y habla, o en el modelado de relaciones probabilísticas (Sarmiento-Ramos, 2020).

Los modelos en base a DL han mostrado ser exitosos en varias áreas. Una de estas áreas es el PLN, dónde los métodos de redes neuronales logran resultados de vanguardia en varias aplicaciones, como la clasificación de texto, reconocimiento de voz, generación de subtítulos, o resúmenes, entre otros (Brownlee, 2019).

2.2.6 Modelo BERT y BETO

El modelo BERT toma su nombre de: *Bidirectional Encoder Representations from Transformers* (Representación de Codificador Bidireccional de Transformadores) es un modelo basado en el aprendizaje profundo y redes neuronales dedicado al lenguaje, fue diseñado por Google y está pensado para trabajar en el idioma inglés (Devlin et al., 2019). BERT es un método de representaciones lingüísticas previas al entrenamiento. Este método presenta una mejora frente métodos anteriores por ser un sistema no supervisado, contextual y profundamente bidireccional usado de manera previa al PNL (López & Gonzales, 2021).

Este modelo ha tenido adaptaciones a distintos idiomas, entre ellos el español, como BETO que es una composición de redes neurales desarrollada en Chile. BETO es un modelo de PLN, está compuesto de redes de gran tamaño con la capacidad de analizar grandes volúmenes de texto (*Corpus*) (Cañete et al., 2023). Gracias a un pre-entrenamiento puede aprender la estructura de las palabras en un idioma (IIC, 2023).

Las diferencias internas entre BERT y BETO son pocas ya que siguen la misma lógica. Como principales diferencias está la gramática propia el idioma, y trabaja con caracteres presentes en español, como lo son el uso de las tildes y la letra Ñ.

En la investigación de Venegas (2021) se muestra a BETO como un modelo robusto, ya que destaca que este modelo cuenta con 110 millones de parámetros y fue entrenado con un corpus de aproximadamente 3.000 millones de palabras en español.

Hoy en día, están surgiendo nuevas variantes de BERT en español, se tiene registro de MarIA y RigoBERTa, entre otras.

MarIA: Según la página del Ministerio para la Transformación Digital y de la Función Pública, MarIA el primer sistema de inteligencia artificial experto en comprender y escribir en lengua española (Gobierno de España, 2022).

RigoBERTa: Según el Instituto de Ingeniería del Conocimiento de la Universidad Autónoma de Madrid, RigoBERTa es un modelo de lenguaje entrenado para la comprensión general de nuestro idioma español. Además, cuenta con la posibilidad de adaptarse a diferentes contextos del lenguaje, y se especializa en tareas de análisis y comprensión. (IIC, 2023).

Se prefiere el uso de BETO, ya que desde su origen fue pre-entrenado con textos cargados al dialecto chileno, por lo que BETO debería ser mejor al usarse en *corpus* de generados por habitantes de Chile (Cañete et al., 2023). En cambio, MarIA (Gutiérrez-Fandiño et al., 2022) y RigoBERTa (Vaca et al., 2022) se centran en su dialecto de origen, que es el de España.

2.2.7 Optuna

Optuna es una biblioteca de optimización automática de hiperparámetros de código abierto para ML y otros problemas de optimización. Esta biblioteca se utiliza comúnmente para encontrar los mejores conjuntos de hiperparámetros para modelos de ML de manera eficiente (Optuna, 2018).

Optuna utiliza distintos algoritmos de búsqueda como TPE (*Tree-structured Parzen Estimator*), Procesos Gaussianos, u otros, dependiendo del marco de aprendizaje. Así, busca entre distintas configuraciones de hiperparámetros de manera eficiente, para encontrar una buena combinación que maximice o minimice una función objetivo (Akiba et al., 2019).

Esta biblioteca es muy útil cuando se trabaja con algoritmos de ML y DL, utiliza rutas (conexiones entre distintos puntos de una red) existentes para determinar el área prometedora. Así, buscar la optimización del hiperparámetro, resultando en el encuentro del hiperparámetro buenos valores en un tiempo mínimo (Es & Bajaj, 2023). Además, es compatible con varios

frameworks (o plantillas de *software*) de ML, como TensorFlow, PyTorch (Akiba et al., 2019), haciéndolo un optimizador bastante versátil.

Existen varias alternativas a Optuna para la optimización automática de hiperparámetros, por ejemplo, Hyperopt y Skopt.

Hyperopt: utiliza un enfoque de optimización bayesiana similar al de Optuna. Hyperopt ha estado disponible por más tiempo y es ampliamente adoptado en la comunidad de aprendizaje automático. Sin embargo, algunos usuarios encuentran que la interfaz de Hyperopt puede ser un poco más compleja (Czakon, 2023).

Skopt: también conocido como Scikit-Optimize, es una biblioteca que se integra directamente con scikit-learn (*software* de ML), y se enfoca en la optimización en modelos específicos del programa. Skopt tiene una integración perfecta con scikit-learn. Por contraparte, puede ser menos versátil si se está buscando optimizar hiperparámetros para una variedad de modelos de aprendizaje distintos a scikit-learn (Amat, 2020).

3. Metodología

Para comprender el enfoque y las técnicas empleadas en el desarrollo de esta memoria de título. En esta sección, es necesario conocer los procesos y herramientas utilizadas para el tratamiento de los datos, la implementación del modelo predictor y las etapas de evaluación y validación. Cada uno de estos pasos es necesario para obtener un modelo capaz de entregar resultados válidos.

Primeramente, se explica el tratamiento de datos, que abarca la limpieza y preparación de los conjuntos de información, es el primer paso para asegurar que el modelo funcione de manera eficiente y con el menor número de errores. A continuación, se presenta el proceso de implementación del modelo, incluyendo su entrenamiento, testeo y validación. Estas fases permiten refinar y ajustar el modelo para que pueda realizar predicciones precisas y relevantes en las respuestas al Canvas de Sercotec.

3.1 Tratamiento de datos

Antes de la implementación del modelo predictor, y de realizar el procesamiento de las respuestas del Canvas se necesita un tratamiento de la data. Para esto, y todas las herramientas posteriores, se usa el lenguaje de programación Python mediante el uso de biblioteca pandas, que se especializa en el manejo y análisis de estructuras de datos (Pandas, 2024).

Las principales medidas del tratamiento previo es la limpieza y división de datos. A diferencia de los datos simulados, los datos reales pueden reunir distintos tipos de errores y/o complicaciones que pueden llevar a problemas durante el procesamiento del *corpus*.

En el proceso de limpieza de datos, se eliminan las columnas irrelevantes, como ids territoriales o de agentes operadores, para simplificar la visualización y reducir el uso de memoria. A continuación, se revisan errores comunes como

faltas ortográficas. También es crucial identificar problemas menos comunes que puedan surgir debido a la variabilidad en la escritura de los textos. Para el caso de las faltas ortográficas, se trabaja con la biblioteca SpellChecker, que es un corrector ortográfico para varios idiomas, entre ellos el español (Norvig, 2023).

Además, se debe dividir la data, se consideran 18 archivos independientes separados en dos grupos. En el primero grupo están los nueve archivos que fueron evaluados, y que contienen las respuestas para cada una de las secciones del Canvas (un archivo por sección), estos son usados en el entrenamiento y el testeo. En el segundo grupo, están los nueve archivos restantes que por distintas razones no fueron evaluados, pero sirven en la validación del modelo.

3.2 Implementación del modelo

A grandes rasgos, la implementación del modelo consta de tres etapas secuenciales: el pre-entrenamiento, el testeo, y la validación. Para cada una de dichas etapas, se toma como base tanto la lógica computacional, como el código en Python usado en el entrenamiento con data simulada en el memoria de Conejeros (2023), en conjunto a la información entregada en la investigación de Cabezas (2023).

Por lo anterior, se propone un predictor de texto con base en BETO, con sus resultados optimizados en Optuna, y con validaciones por Matriz de Confusión y Validación Cruzada. Este predictor toma la data de las secciones del Modelo Canvas que Sercotec pide a los postulantes de sus programas, en particular de Capital Semilla Emprende.

3.2.1 Entrenamiento

Es la primera etapa, se ingresa un paquete de los datos etiquetados (texto y nota) de cada archivo, se procesa el texto de entrada, y se espera que el

modelo BETO extraiga la estructura lingüística de dichos textos y los aprenda. Cabe recordar que BETO ya tiene una preparación previa de análisis con textos en español.

Para entrenar el modelo se usa la biblioteca PyTorch, un marco de aprendizaje (Akiba et al., 2019). Está diseñada para proyectos de aprendizajes automáticos, y resulta una herramienta muy adecuada para aplicaciones de DL como ésta.

Es en este punto, dónde el modelo convierte texto en variables cuantitativas al vectorizar los datos ingresados. Mustapic (2024) explica que la modelización de sistemas PLN trabaja mediante la generación de los valores TF-IDF derivados de la tokenización de palabras, dónde:

- *Term Frecuency* o TF (Frecuencia de término): es la frecuencia relativa de un término (*t*) dentro de un documento (*d*). Se calcula dividiendo el número de veces que aparece *t* en *d*, por el número total de términos del documento.
- *Inverse Document Frequency* o IDF (Frecuencia Inversa del documento): mide la cantidad de información que proporciona un término. Se calcula tomando el logaritmo del cociente entre el número total de documentos (*N*) por el número de documentos que contienen el término.
- TF-IDF: es el producto de ambas frecuencias, y muestra la relevancia del término, siendo una puntuación más alta la que denota mayor relevancia y una puntuación más baja la que denota menor relevancia.

Los valores de TF-IDF nos dicen el peso que tiene cada palabra dentro de la respuesta dada (documento). Así, es a partir de esta variable que el modelo genera las redes necesarias para el procesamiento computacional.

3.2.2 Testeo

Es la segunda etapa, en que BETO recibe un segundo paquete de datos etiquetados. El modelo revisa como responder a la estructura lingüística ya estudiada, y se va ajustando poco a poco para coincidir de mejor manera con la nota según vaya procesando los datos.

El proceso de cambio de parámetros internos del modelo es asistido por el optimizador Optuna, con el fin de que se garantice el mejor resultado posible. Producto de este cambio de parámetros a este proceso también se le denomina ajuste fino o *fine-tuning*. El objetivo de Optuna, es la minimización de la función de perdida (Loss), un menor loss indica una mejor predicción de los datos.

El set de hiperparámetros para el entrenamiento se fija de acuerdo con los valores que resultaron en la investigación de Conejeros (2023), y sus intervalos de ajuste se muestran en la **¡Error! No se encuentra el origen de la referencia.**.

Tabla 2: Rango de hiperparámetros definidos para Optuna.

Hiperparámetro	Rango de valores
Learning rate	]0,000001: 0,0001[
Epochs	]1: 5[
Batch size	[8; 16; 32; 64]

Fuente: Extraído de Conejeros (2023).

Dónde:

- *Learning rate* (o tasa de aprendizaje): controla cuánto se ajustan los pesos del modelo con respecto al error calculado durante el entrenamiento. Si es bajo permite ajustes pequeños y precisos, pero puede hacer el entrenamiento más lento. Mientras que si crece puede acelerar el entrenamiento, pero corre el riesgo de que el modelo no converja o se aleje del óptimo.

- *Epoch* (o época): se refiere a cuántas veces el modelo recorre el conjunto completo de datos de entrenamiento durante el proceso. En general, si se entrena durante más épocas puede mejorar el rendimiento del modelo, pero se vuelve un entrenamiento más lento y podría llevar al modelo a sobre ajustarse.
- *Batch size* (o tamaño de lote): se refiere al número de ejemplos de entrenamiento que se utilizan en una iteración para calcular el error y actualizar los pesos del modelo. Un lote más grande puede acelerar el entrenamiento al aprovechar mejor los recursos computacionales, pero también puede requerir más memoria. Un tamaño de lote más pequeño puede proporcionar una convergencia más estable y se utiliza cuando hay limitaciones de memoria o para evitar mínimos locales.

Estos hiperparámetros son críticos para ajustar el rendimiento y la eficiencia del modelo durante el entrenamiento en DL con PLN. El pre-entrenamiento y testeo son procesos que se deben repetir en los nueve primeros archivos. La división de paquetes de datos con y sin etiqueta se hace de forma aleatoria y automática a través de la biblioteca sklearn. Con datos etiquetados, se hace referencia a los que se utilizan en aprendizaje supervisado, o sea, que incluye datos de entrada y resultados correctos, permitiéndole al modelo aprender con el tiempo. (IBM, 2023). En este caso, los datos de entrada están dada por el *corpus* de la respuesta a una sección del Canvas, y los datos de resultado corresponden a la calificación con la que se evaluó cada respuesta.

De manera general, las notas de las evaluaciones son una variable cuantificable. En contraposición a esto, la evaluación del Canvas que hace Sercotec recoge notas puntuales en un espectro acotado de opciones, por ejemplo, para el caso de la sección Clientes, la nota solo puede ser una entre las cuatro opciones del conjunto {1, 3, 5, 7}. Así, cada nota posible se toma como una única clase. Por lo tanto, se puede decir que, al calificar las respuestas, se está abordando un problema de clasificación.

3.2.3 Validación

Finalmente, se llega a la tercera etapa que corresponde a la validación del modelo. En este paso, se usan los archivos de datos no evaluados como datos fuera de muestra. Estos paquetes se ingresan en la matriz de confusión. Con esta herramienta se evalúa la efectividad del modelo mediante el contraste de resultados.

Matriz de confusión: es una tabla que nos permite ver qué tan confundido está un modelo al momento de la clasificación, mostrándonos tanto los aciertos como desaciertos cometidos para cada una de las categorías (Sotaquirá, 2022).

Para usar esta matriz, se debe tomar el set de prueba y se clasifica según el modelo entrenado. Seguido a ese, se realiza el conteo de los aciertos y desaciertos por cada categoría. Luego, se organiza este conteo como en la Tabla 3. Las columnas representan las categorías a las que realmente pertenece cada dato, y las filas representan las categorías predichas por el modelo.

Tabla 3: Matriz de confusión para dos clases.

		Valores Actuales	
		Positivo (1)	Negativo (0)
Predicción	Positivo (1)	Verdaderos Positivos (VP o TP)	Falsos Positivos (FP)
	Negativo (0)	Falsos Negativos (FN)	Verdaderos Negativos (VN o TN)

Fuente: Elaboración propia.

La matriz representa las cuatro posibles combinaciones de los datos:

- **Verdaderos positivos (VP)** o True Positive (TP): es cuando el modelo predice que los datos pertenecen a la clase positiva, y los datos reales efectivamente se encuentran en la clase positiva.
- **Verdadero Negativo (VN)** o True Negative (TN): es cuando el modelo predice que los datos pertenecen a la clase negativa, y los datos reales efectivamente se encuentran en la clase negativa.

- **Falso Negativo (FN)** o False Negative: es cuando el modelo predice que los datos pertenecen a la clase negativa, pero en la realidad los datos pertenecen a la clase positiva.
- **Falso Positivo (FP)** o False Positive: es cuando el modelo predice que los datos pertenecen a la clase positiva, pero en la realidad los datos pertenecen a la clase negativa.

Esta idea se extiende a la cantidad de n clases que se tienen en el modelo, si se vuelve al ejemplo de los Clientes del Canvas, se tendría 4 clases.

Con la matriz de confusión construida se ve en detalle el desempeño del modelo para las *n* categorías. Se toma el modelo con mejor resultado en el testeo, y contrasta el número de FN y FP con resultados de matrices con diversas configuraciones de parámetros. Si la cantidad de Falsos de dicho modelo es menor que el de las otras configuraciones, se evidencia que el modelo si es válido.

Métricas:

La calidad de los resultados del entrenamiento está dada por métricas dependientes de los resultados de la Matriz de confusión, y son las siguientes:

- Exactitud (*Accuracy*): mide la proporción de predicciones correctas sobre el total de predicciones realizadas, según la Ecuación (1).

$$\text{Exactitud} = Accu = \frac{PC}{TP}, con\ Accu \in [0,1] \quad (1)$$

Dónde,

PC: es el total de predicciones correctas, según la Ecuación (2).

$$PC = VP + VN \quad (2)$$

TP: es el total de predicciones, según la Ecuación (3).

$$TP = VP + FP + VN + FN \quad (3)$$

- Precisión: mide la proporción de verdaderos positivos sobre el total de positivos predichos. Buen indicador en caso de alto número de falsos positivos, según la Ecuación (4).

$$\text{Precisión} = Pres = \frac{VP}{VP + FP}, con\ Pres \in [0,1] \quad (4)$$

- Exhaustividad o Sensibilidad (*Recall*): mide la proporción de verdaderos positivos sobre el total de positivos reales. Buen indicador en caso de alto número de falsos negativos, según la Ecuación (5).

$$\text{Exhaustividad} = Rec = \frac{VP}{VP + FN}, con\ Rec \in [0,1] \quad (5)$$

- Medida F1 (F1 *Score*): es una métrica alternativa, junta la Precisión y la Exhaustividad en un solo valor. Buen indicador si se busca un balance entre dichas medidas, según la Ecuación (6).

$$F1 = 2 \times \frac{Pres \times Rec}{Pres + Rec} = \frac{2VP}{2VP + FP + FN}, con\ F1 \in [0,1] \quad (6)$$

Para todas estas métricas se tiene que mientras mayor y más cercana a 1 sea, es mejor. Además, se agregan dos métricas independientes a las anteriores: La Perdida de Entrenamiento y la Pérdida de Validación. La función de pérdida es necesaria para estas dos medidas:

- Pérdida (*Loss*): es la función de pérdida utilizada en problemas de clasificación, mide la discrepancia entre las distribuciones de probabilidad predichas por el modelo y las distribuciones de probabilidad verdaderas para una única instancia, según la Ecuación (7).

$$Loss(\gamma, \varphi) = -\sum_{i=1}^{C} \gamma_i \times \log(\varphi_i), Loss \in [0, \infty[, i \in \{1, \dots, C\} \quad (7)$$

Dónde,

C: es el número de clases.

γ: es la es el vector de clases verdaderas.
φ: es el vector de probabilidades predichas para cada clase.
γ_i: es la etiqueta verdadera para la clase *i*.
φ_i: es la probabilidad predicha para la clase *i*.

- Perdida de Entrenamiento (*Training Loss*): mide el error del modelo en el conjunto de datos de entrenamiento. Se calcula en cada iteración del entrenamiento y refleja cuánto se desvía la predicción del modelo de los valores reales, según la Ecuación (8).

$$\text{Training Loss} = TL = \frac{1}{N}\sum_{j=1}^{N} Loss\left(\gamma^{(j)}, \varphi^{(j)}\right), VT \in [0, \infty[, j \in \{1, \dots, N\} \qquad (8)$$

Dónde,

N : es el número total de ejemplos en un conjunto de entrenamiento.
$\gamma^{(j)}$: es la etiqueta verdadera para el *j*-ésimo ejemplo de *N*.
$\varphi^{(j)}$: es la predicción del modelo para el *j*-ésimo ejemplo de *N*.

- Pérdida de Validación (*Validation Loss*): es similar al TL, pero ya no se calcula sobre los datos de entrenamiento, sino que sobre los datos de validación, según la Ecuación (9).

$$\text{Validation Loss} = VL = \frac{1}{M}\sum_{j=1}^{M} Loss(\gamma^{(j)}, \varphi^{(j)}), VL \in [0, \infty[, j \in \{1, \dots, M\} \qquad (9)$$

Dónde,

M : es el número total de ejemplos en el conjunto de validación.
$\gamma^{(j)}$: es la etiqueta verdadera para el *j*-ésimo ejemplo de *M*.
$\varphi^{(j)}$: es la predicción del modelo para el *j*-ésimo ejemplo de *M*.

Para Loss, TL y VL, un valor 0, indicaría predicción perfecta, por lo que durante el entrenamiento y validación se buscan valores pequeños cercanos a 0.

4. Datos

Algo fundamental en la generación de modelos de ML o DL, es la necesidad de trabajar con una buena base de datos con la que realizar el entrenamiento. Esto contempla uno o varios archivos con un gran volumen de datos, mientras más datos mejor. Además, que dichos datos estén bien tratados, o sea, con el número mínimo de errores y todo bien etiquetado. Que es una situación bastante idealizada, y que por lo general escapa de la realidad.

Para resolver problemas del mundo real mediante proyectos de ML o DL, es crucial utilizar datos reales en lugar de simulados, lo que representa un gran desafío. La integridad y calidad de los datos son fundamentales, ya que los datos generados por personas son especialmente propensos a errores. Esto resalta la necesidad de una gestión efectiva de la calidad de los datos y de la adaptabilidad necesaria para abordar los problemas específicos de cada proyecto.

Los datos de alta calidad son un activo valioso que facilita el proceso de modelado y análisis en distintos proyectos. Sin embargo, incluso los datos considerados malos, ya sea por problemas de lecturas, errores o datos faltantes, pueden proporcionar información valiosa una vez que se someten a un proceso de limpieza y preprocesamiento adecuado.

4.1 Data de Sercotec

Los datos que se usan para el entrenamiento del modelo son datos reales aportados por el equipo de Sercotec. La información está contenida en el archivo Data Canvas Semilla Innominado UdeC, está en formato texto y tienen un peso computacional de 562 MBs (Sercotec, 2023c).

El archivo contiene 36 columnas con información de antiguas postulaciones al programa CSE, además de cada respuesta del postulante al Canvas. De toda la

información, las columnas más relevantes para el modelo son las que contienen el texto de respuesta y la nota. Se pueden ver en el Anexo B, Tabla 26.

El análisis de los datos se realiza mediante Python. El primer paso en su tratamiento fue conservar únicamente las columnas que contribuyen al propósito de la investigación el id de la pregunta, el texto de la respuesta y la nota.

El archivo contiene un total de 600.764 filas, cada una correspondiente a una pregunta respondida del Canvas. Cada nuevo postulante aparece tras cada nueve respuestas. Se identificó un salto de 21 filas no leídas, lo que sugiere que originalmente había 600.785 líneas en el archivo.

4.2 Tratamiento/Depuración

4.2.1 Problemas

En el proceso de análisis de los datos utilizados en esta investigación, se han identificado ciertos desafíos relacionados con la calidad de los datos. Es fundamental abordar estos problemas de calidad de datos para garantizar la confiabilidad y validez de los resultados obtenidos en este estudio. Se tiene que las notas 1, 4, 6 y 7 de la rúbrica de postulaciones de Sercotec (Sercotec, 2022c), no son iguales a las que tienen los datos reales 1, 3, 5 y 7, si bien es algo que no debe afectar el modelo, ya que siguen siendo cuatro clases y solo supone un cambio de nombre en las clases intermedias (3 y 5), se trabaja solo con las notas reales que presentan los datos.

Además de las líneas que no se leen algunas líneas (hay saltos), se observan otros problemas. Por una parte, hay errores en la identificación de las preguntas. Para una misma sección del Canvas hay distintos ids, por lo que a priori, no se puede automatizar la separación de secciones, que es algo fundamental para trabajar el modelo.

Por otro parte, en cierta parte de la data, empieza a verse una décima sección en algunos postulantes. Esta sección son respuestas a un ítem de sustentabilidad añadido en las últimas convocatorias. Así, estos datos no sirven para el entrenamiento ya que no pertenece al Canvas clásico, y hace que el periodo de nueve columnas por postulante se rompa.

Finalmente, se identifican los diferentes problemas propios de textos escritos por personas. Hay faltas ortográficas, errores en los saltos de línea que hace que se junten párrafos y palabras. Además, hay uso tildes y de caracteres especiales como el @ y el # que perjudicarán en la limpieza y tokenización del *corpus*. El abordaje de estos problemas proporciona un gran desafío en el tratamiento previo al entrenamiento del modelo.

4.2.2 Soluciones

Primeramente, respecto al problema de identificación de las preguntas, se plantea trabajar con los primeros índices. Las preguntas con ids entre 15 y 23 son los que concentrar la mayor cantidad de respuestas en la data. Por tanto, son los ids más significativos para realizar el entrenamiento.

Luego se considera que, al estar en pocas convocatorias, y solo en algunos postulantes, hay muy pocos datos que comprenden al ítem de sustentabilidad. Por tanto, no son considerados para el entrenamiento. Estas decisiones implican que disminuya el volumen total de datos previo al entrenamiento.

Finalmente, las faltas ortográficas se corrigen con la versión en español de la biblioteca SpellChecker, se pasa todo el texto a minúsculas y se quitan las tildes.

4.3 Análisis Exploratorio de Datos

4.3.1 Introducción del AED

Sobre los 18 paquetes que se crearon a partir de la data original, y luego se ser depurados. Se implementa un pequeño análisis exploratorio de datos (AED), o EDA en inglés, con el propósito de identificar tendencias, distribuciones, correlaciones entre variables y conocer bien el set de datos con los que se trabaja. Durante el AED, se busca completar algunos de los siguientes objetivos: limpieza de datos, obtención de estadísticas descriptivas, visualización de los datos, investigación de variables y funciones, determinar relaciones y dependencias entre variables, segmentación de datos, generación de hipótesis, y evaluar la calidad y fiabilidad de los datos (Deming et al., 2018). Se entiende que a mayor número de objetivos cumplidos es un mejor AED.

Si el AED resultó positivo para conocer los datos reales de los postulantes al PCSE, y una vez hecha la limpieza de datos y se hayan obtenido las estadísticas descriptivas, se deberían poder contestar las siguientes preguntas:

- ¿Cuántos registros hay?
- ¿Son muy pocos? o ¿Son muchos y falta capacidad computacional para procesarlo?
- ¿Están todas las filas completas, o tenemos variables con valores nulos?
- ¿Qué datos son discretos, y cuáles continuos?
- Si es un problema de tipo supervisado: ¿cuál es la columna de salida? ¿binaria, multiclase?
- ¿Cuáles parecen ser variables importantes?, ¿cuáles podemos descartar?
- ¿Siguen alguna distribución?
- ¿Hay correlación entre variables (características)?
- ¿Existen datos repetidas, mal tipeadas, o duales (mayúsculas/minúsculas, singular/plural)?
- ¿Estamos ante un problema dependiente del tiempo?
- ¿Cuáles son los puntos atípicos que contaminan o desvían las distribuciones?, ¿se pueden eliminar o es importante conservarlos?, ¿son errores de carga o son reales? ¿Tenemos posible sesgo de datos?

4.3.2 Estadísticas descriptivas

Para contestar las preguntas anteriores, se realiza un análisis descriptivo para comprender las características centrales de los datos evaluados, incluyendo la distribución de notas por sección. Luego, se realiza una comparativa con la media, desviación estándar, mediana y moda para cada una de las áreas del Canvas. Estos datos se muestran respectivamente en la Figura 1 y Tabla 4, donde se observa que la sección de Clientes tiene la mayor media (mayor a 5,72), indicando que son respuestas fáciles de contestar y por eso se obtienen mejores notas, contrastando con Elemento diferenciador (o Propuesta de Valor) que tiene la menor media (4,38) entre las secciones. En general, la mediana y la moda sugieren que los datos tienden hacia los valores 5 y 7 en varias secciones, lo que refleja una cierta sobreestimación de las notas 7, que es la nota preponderante en cinco de las nueve secciones. Cabe recordar que para todas las secciones las notas mínima y máxima son 1 y 7 respectivamente, y al ser iguales siempre no se incluyen en la tabla.

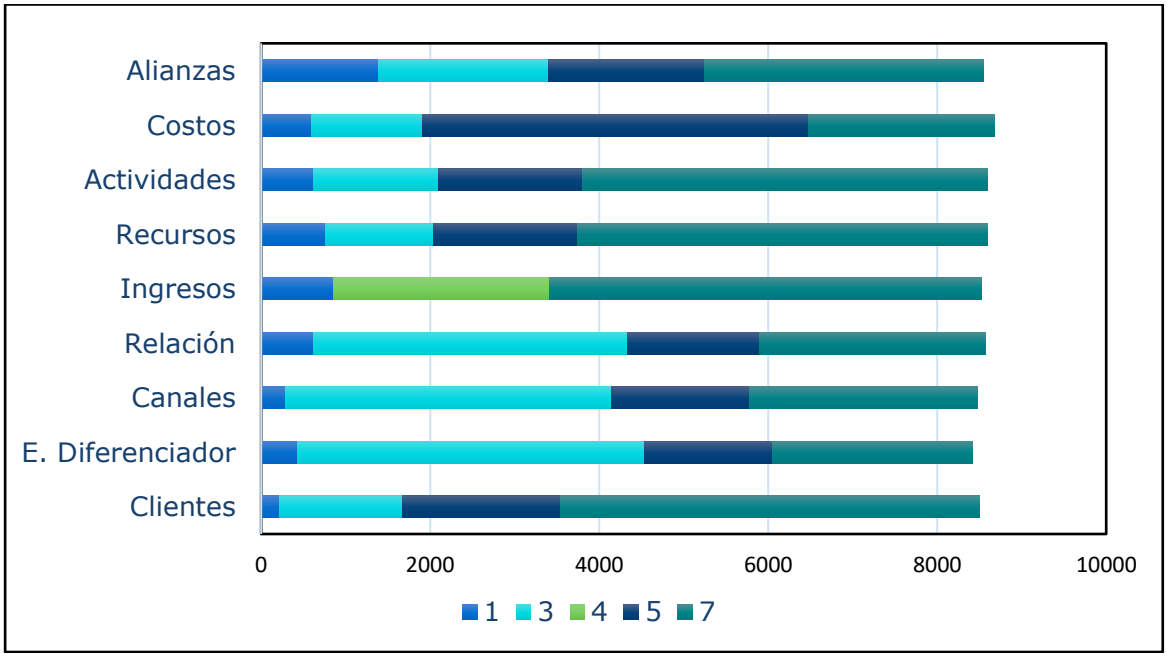


Figura 1: Distribución de notas por sección del Canvas.
Fuente: Elaboración propia.

Tabla 4: Estadísticas descriptivas para la variable nota de los archivos.

Sección de Canvas	Cantidad de datos	Media	Desviación Estándar	Mediana	Moda
Clientes	8503	5,7218	1,7031	7	7
Propuesta de valor	8413	4,3819	1,8736	3	3
Canales	8475	4,5934	1,8621	5	3
Relación con clientes	8572	4,4704	1,9624	3	3
Ingresos[1]	8527	5,4970	2,0148	7	7
Recursos clave	8597	5,4764	2,0118	7	7
Actividades clave	8591	5,4800	1,9618	7	7
Costos	8676	4,9287	1,6477	5	5
Alianzas Clave	8551	4,6545	2,2285	5	7
TOTAL	76905	5,0239			

Fuente: Elaboración propia.

Se observa que para todas las secciones hay pocas notas 1 y 5 en comparación a las notas 3 y 7. Las peores calificaciones se encuentran en la Propuesta de Valor, Canales y Relación con el cliente, donde predomina la nota 3. En contraste a eso, en el resto de las secciones impera el 7, que es el máximo. Se espera que estas tendencias también se vean reflejadas al momento de realizar las predicciones del modelo.

Se registra que un tercio de las secciones: Canales, Costos y Alianzas, tienen cercanía entre su media y mediana (diferencia menor a 0,5), lo que indica una distribución simétrica de los datos. A contraste del resto de secciones que tienen una diferencia mayor de entre 1,28 y 1,52, evidenciando una asimetría en la distribución. Dicha asimetría se ratifica al observar la moda, pues en estas mismas secciones coincide la moda y la mediana, con lo que se tienen demasiadas evaluaciones con la misma nota, y esto puede perjudicar el aprendizaje del modelo. Además, la cantidad de datos a través de las diferentes secciones varía entre 8.413 y 8.676, que es bastante consistente, lo

[1] A diferencia de todas las demás secciones donde las notas son: 1, 3, 5 o 7; en Ingresos las notas que hay son: 1, 4 y 7.

que sugiere que el entrenamiento cubre de manera uniforme las distintas áreas del Canvas.

Para los archivos sin nota, se tiene la información incluida en la Tabla 5. A diferencia de la Tabla 4, en la tabla sin notas no se puede proporcionar un análisis detallado mediante estadísticas. Ésta solo muestra la cantidad de datos por sección, ofreciendo una visión cuantitativa básica sin información sobre la dispersión de los datos, lo que limita la interpretación de la variabilidad y tendencia de los datos en cada sección del Canvas.

Tabla 5: Cantidad de datos de los archivos sin nota.

Sección Canvas	Cantidad de datos
Clientes	33520
Propuesta de Valor	33545
Canales	33471
Relación	33436
Ingresos	33447
Recursos	33394
Actividades	33419
Costos	33426
Alianzas	33494
TOTAL	**301152**

Fuente: Elaboración propia.

La cantidad de datos varían entre 33.394 y 33.545, al igual que en los datos con nota, es bastante consistente y se mantiene uniforme para todas las secciones.

Para la información relativa al texto, se tiene que dar un valor numérico a las respuestas, que es una variable cualitativa. Para ello es bueno una evaluación de la cantidad de palabras por categoría. Esto se muestra en las Figuras 2, 3 y 4.

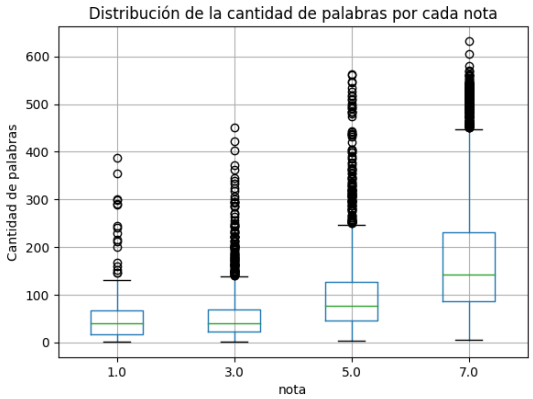


Figura 2: Distribución de palabras por notas, sección Clientes.[2]

Fuente: Elaboración propia.

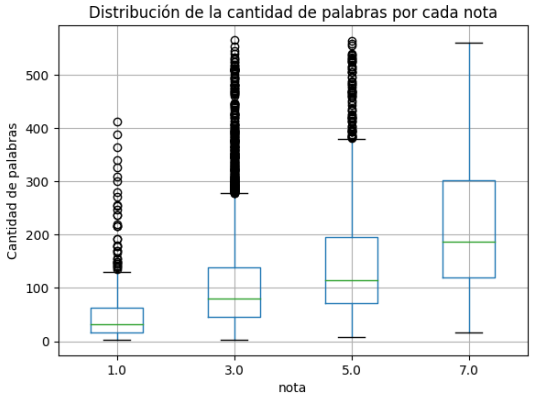


Figura 3: Distribución de palabras por notas, sección Propuesta de Valor.

Fuente: Elaboración propia.

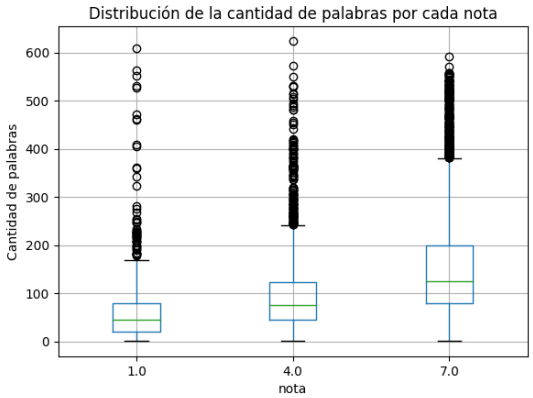


Figura 4: Distribución de palabras por notas, sección Ingresos.

Fuente: Elaboración propia.

[2] Solo se mostrarán tres secciones como ejemplo. Para ver el resto con la distribución de data para entrenar (10%), con y sin stopwords, revisar desde la Tabla 17 a la Tabla 35, Anexo C.

En las Figuras 2, 3 y 4, se observa un claro desequilibrio, donde la nota 7 está mucho más representada que el resto, la representación baja a medida que disminuye la nota. Por lo tanto, se anticipa que las principales discrepancias entre la nota que merece la respuesta y lo predice del modelo, serán por una posible sobrerrepresentación de la clase 7 y una representación casi insignificante de la clase 1.

Se escoge la revisión y análisis de estas tres secciones por sobre el resto, por los siguientes motivos:

- Clientes: es la sección principal del modelo, y es donde se prueba el modelo inicial con data simulada, por lo que se toma como la base del estudio.

- Propuesta de Valor (PdV): es muy importante dentro del Canvas, muestra lo que te diferencia del resto de proyectos. A diferencia de los clientes donde las respuestas siempre son la descripción de personas. La PdV es muy subjetiva, ya que puede describir situaciones, características humanas o de un producto tangible, sensaciones, o hablar de un estándar intangible. Esto hace que el tipo de respuesta varíe mucho y esto afecte el entrenamiento.

- Ingresos: es una variable con respuestas más concisas que PdV, ya que solo describe como monetizaría el proyecto. Pero esta sección es la única donde tiene tres tipos de respuestas: 1, 4 y 7, en vez de 1, 3, 5 y 7 como en las ocho secciones restantes, y eso marca una diferencia en el entrenamiento del modelo.

Palabras más relevantes

A continuación, se presentan las palabras más relevantes según los valores TF-IDF de cada sección del Canvas. Las Tabla 6, 7 y 8 muestran la importancia relativa de diferentes términos clave para cada categoría o clase.

Tabla 6: Valores TF-IDF por categoría, sección Clientes.

Nota	1	3	5	7
clientes	413	1508	2524	5185
personas	1701	5427	8882	17648
productos	1807	5812	9494	18809
negocio	1537	-	-	-
cliente	411	-	-	-
empresas	839	2773	4616	9142
mas	1421	4544	7468	14948
servicio	2073	-	-	-
tipo	2199	-	-	-
segmento	2041	6550	10647	21244
naturales	-	4897	8011	-
particulares	-	5274	-	-
adultos	-	345	-	-
general	-	3558	-	-

Fuente: Elaboración propia.

Tabla 7: Valores TF-IDF por categoría, sección Propuesta de Valor.

Nota	1	3	5	7
calidad	555	3069	2187	3206
productos	3054	15322	11238	16111
producto	3051	15316	11233	16106
servicio	3486	17498	12779	18459
competencia	772	4222	3018	-
clientes	709	3847	2757	4033
mas	2381	12114	8879	12720
cliente	707	3844	2755	4029
trabajo	3704	-	-	-
existe	1647	-	-	-
ademas	-	710	489	758
atencion	-	2128	-	-
valor	-	-	14072	20294
segmento	-	-	-	18263

Fuente: Elaboración propia.

Tabla 8: Valores TF-IDF por categoría, sección Ingresos.

Nota	1	4	7
clientes	1158	2369	4614
calidad	938	1892	3728
pagar	4101	8813	15592
productos	4607	9831	17395
producto	4602	9825	17388
pago	4109	8828	15615
servicio	5254	11234	-
efectivo	2127	4472	8108
estan	2440	-	-
venta	5812	-	-
transferencia	-	12066	21380
pagan	-	8811	15588
segmento	-	-	19676
clientes	1158	2369	4614

Fuente: Elaboración propia.

Se observa duplicidad en los términos cliente y clientes, que son considerados como palabras distintas a pesar de que son lo mismo para

fines del texto, ambas presentan esta relevancia dual en las tres secciones vistas. Ocurre de manera similar con producto y productos, si dichas palabras se toman como una su valor de TF-IDF se sumaría, siendo aún más relevantes para el modelo.

4.3.3 Resultados del AED

Con toda la información anterior, más una visualización previa de las respuestas. Entonces, se pueden responder las preguntas planteadas y dar una buena apreciación respecto a todos los datos involucrados en la investigación.

- ¿Cuántos registros hay?
  - Hay un total de 378.057 registros, donde un 20,34% corresponden a datos evaluados (con nota), y el restante 79,66% son datos sin evaluar (sin nota).
- ¿Son muy pocos? ¿o son muchos y falta capacidad computacional para procesarlo?
  - Para el propósito de este estudio son bastantes datos, incluso excesivos considerando los recursos computaciones que provee Google Colab (12 GBs de RAM) en su versión libre, o un computador de gama media (entre 8 y 16 GBs de RAM, y procesador AMD Ryzen o Intel i entre las series 3 y 7).
  - Es por esto, que el entrenamiento se hace con solo el 10% de los datos con nota (aproximadamente 850 filas/pregunta). Mientras que la predicción solo se realiza con un 1% de los datos sin nota (aproximadamente 330 filas, ver en la Tabla 36 del Anexo D).
- ¿Están todas las filas completas, o tenemos variables con valores nulos?
  - Quitando los datos mal ingresados o saltados por el lector de Python, se tienen que los datos sin nota tienen este campo nulo.

Pero como se explica antes, se toman como datos fuera de muestra.

- ¿Qué datos son discretos, y cuáles continuos?
  - Los datos que interesan en la investigación son tres: id de pregunta, que es un *float* o número real, este nos ayuda a separar los datos por sección, y por sus propiedades de id solo sirve para identificar y no aporta un valor cuantitativo. El texto de respuesta, que es un *string* o hilera de caracteres, que es lo que el modelo debe aprender, y a priori, es una variable no cuantificable. Finalmente, la nota, que es otro *float*, y la única variable donde sí importa su valor, siendo un dato discreto.
- Si es un problema de tipo supervisado: ¿cuál es la columna de salida? ¿binaria, multiclase?
  - Si es de tipo supervisado, y su columna es la nota, siendo de tipo multiclase, donde cada posible nota representa una clase.
- ¿Cuáles parecen ser variables importantes? ¿cuáles podemos descartar?
  - Las variables importantes son las tres dichas anteriormente: id de pregunta, texto de respuesta y nota. Las variables que no se consideraron importantes se quitaron al inicio para evitar sobrecarga de información.
- ¿Siguen alguna distribución?
  - La variable nota se distribuye de manera multinomial, que equivale a una distribución binomial con *k* casos, semejante a lanzar un dado de *k* caras. En este caso con k=4, por cada posible nota. Salvo el caso de la sección ingresos, donde *k* es 3.
- ¿Hay correlación entre variables (características)?
  - Debe existir una gran correlación entre el texto de respuesta y la nota, y esta debe ser positiva, aunque era correlación es difícil de

cuantificar, y se debe procesar el texto para darle un valor numérico a su contenido.

- Por otra parte, puede existir correlación con variables eliminadas como el id del agente operador, o el id del postulante, quien fue el funcionario encargado de evaluar la postulación, y de responder a las preguntas respectivamente. Donde pueden tener correlación positiva o negativa con respecto a la nota, o entre ellos. Aunque esto sería muestra de sesgos, y debería corregirse.

- ¿Existen datos repetidas, mal tipeadas, o duales (mayúsculas/minúsculas, singular/plural)?
  - Si se observan palabras repetidas, y en este caso está bien que ocurran, ya que palabras como: negocio, cliente o valor son casi obligatorias en el contexto de un Canvas. Además, las notas tienen la condición de ser datos repetidos entre los cuatro valores posibles.
  - Se observan datos mal tipeados, pero no repetidos. Dentro de los mal tipeados se evidencian notas 0, o con un espacio, o con un número muy distinto a lo que debería ser, por ejemplo 376 como el valor de una nota, entre muchos otros casos.
  - Además, hay datos duales marcados por el uso de mayúsculas y minúsculas, como: Nuestro y nuestro, o el uso de plural y singular, como: cliente y clientes, como si fueran algo distinto. El primer caso se corrige fácilmente luego de la lectura del archivo original, pasando toda la variable texto a minúsculas para evitar conflictos. Aún permanece la dualidad de palabra en singular versus plural, aunque se espera que sea algo que no influya en el modelo luego de tokenizar.
- ¿Estamos ante un problema dependiente del tiempo?
  - No, en estas variables no influye el tiempo, ni la estacionalidad.

- ¿Cuáles son los puntos atípicos que contaminan o desvían las distribuciones? ¿se pueden eliminar o es importante conservarlos? ¿son errores de carga o son reales?
  - Para la variable nota, se eliminaron los valores atípicos, pues no puede haber nota distinta a las clases existentes. Para el texto de respuesta hay muchos puntos atípicos, como los que se observan en las Figuras 2, 3 y 4, estos no se pueden eliminar, pero si pueden disminuir según la muestra que ingrese al entrenamiento.
- ¿Tenemos posible sesgo de datos?
  - Si se consideran las tres variables importantes para el entrenamiento, no debería haber sesgos. Pero esto puede cambiar si se añade el estudio de otras variables, como, por ejemplo, la variable del agente operador mencionada anteriormente.

5. Resultados

En este capítulo se presenta un análisis de los entrenamientos y resultados del modelo aplicado a la sección Clientes del Canvas. Se exploran diferentes configuraciones de hiperparámetros para evaluar su impacto en las métricas de desempeño del modelo. Además de los hiperparámetros, se hace variar la cantidad de datos con que se entrena el modelo. Así, se fijan los valores para el entrenamiento de las secciones del Canvas restantes. Este análisis proporciona una base de ajuste el modelo con el fin de mejorar su rendimiento en tareas de clasificación de texto.

5.1 Entrenamientos

Se hizo el entrenamiento para las tres secciones del Canvas que se revisaran con más énfasis: Clientes, Propuesta de Valor e Ingresos.

Para cada entrenamiento (Ent.), se realiza una extensa iteración con hiperparámetros fijos, variando el *learning rate*, *epoch* y/o *batch size*. La configuración inicial corresponde a una elección arbitraria entre los valores dados en la Tabla 2.

Tabla 9: Métricas Entrenamiento 1.
Learning rate = 0,0002, Epoch = 4, Batch size = 4, 200 datos.

Epoch	T. Loss	V. Loss	Accuracy	F1	Precision	Recall
1	No log	1,505683	0,520000	0,171053	0,130000	0,250000
2	No log	1,140034	0,520000	0,171053	0,130000	0,250000
3	No log	1,154277	0,520000	0,171053	0,130000	0,250000
4	No log	1,136925	0,520000	0,171053	0,130000	0,250000

Fuente: Elaboración propia.

Tabla 10: Métricas Entrenamiento 2.
Learning rate = 0,0002, Epoch = 4, Batch size = 8, 200 datos.

Epoch	T. Loss	V. Loss	Accuracy	F1	Precision	Recall
1	No log	1,147758	0,560000	0,179487	0,140000	0,250000
2	No log	1,189957	0,560000	0,179487	0,140000	0,250000
3	No log	1,147712	0,560000	0,179487	0,140000	0,250000
4	No log	1,070087	0,560000	0,179487	0,140000	0,250000

Fuente: Elaboración propia.

Tabla 11: Métricas Entrenamiento 3.
Learning rate = 0,0004, Epoch = 8, Bach size = 4, 350 datos.

Epoch	T. Loss	V. Loss	Accuracy	F1	Precision	Recall
1	No log	1,220157	0,603175	0,188119	0,150794	0,250000
2	No log	1,303011	0,603175	0,188119	0,150794	0,250000
3	No log	1,429537	0,111111	0,050000	0,027778	0,250000
4	No log	1,179050	0,603175	0,188119	0,150794	0,250000
5	No log	1,093856	0,603175	0,188119	0,150794	0,250000
6	No log	1,190111	0,603175	0,188119	0,150794	0,250000
7	No log	1,100698	0,603175	0,188119	0,150794	0,250000
8	No log	1,129412	0,603175	0,188119	0,150794	0,250000

Fuente: Elaboración propia.

Entre los Ent. 1, 2 y 3 se observa que a mayor *batch size* mejoran las métricas iniciales. Pero, a lo largo de las épocas estas se mantienen constantes lo que es un indicio de no progreso, esto indica que los paquetes de prueba son muy chicos. Para los próximos entrenamientos se aumentan los valores de *batch size* y de datos, se mantiene un *epoch* de 8, y se varía el *learning rate*.

Tabla 12: Métricas Entrenamiento 4.
Learning rate = 0,00001, Epoch = 8, Batch size = 32, 850 datos.

Epoch	T. Loss	V. Loss	Accuracy	F1	Precision	Recall
1	No log	0,924943	0,596000	0,239932	0,379167	0,278462
2	No log	0,850629	0,656000	0,423520	0,438233	0,420762
3	No log	0,839265	0,668000	0,443890	0,452844	0,443771
4	No log	0,884071	0,652000	0,387607	0,424477	0,393598
5	No log	0,892825	0,656000	0,399102	0,421421	0,399100
6	No log	0,912279	0,656000	0,413979	0,429982	0,412250
7	No log	0,918826	0,644000	0,422687	0,425004	0,424005
8	No log	0,937036	0,632000	0,416211	0,416293	0,418833

Fuente: Elaboración propia.

Tabla 13: Métricas Entrenamiento 5.
Learning rate = 0,00002, Epoch = 8, Bach size = 32, 850 datos.

Epoch	T. Loss	V. Loss	Accuracy	F1	Precision	Recall
1	No log	0,976123	0,610329	0,327280	0,409706	0,340833
2	No log	0,895465	0,638498	0,407383	0,448135	0,400972
3	No log	0,905469	0,629108	0,397523	0,449484	0,387083
4	No log	0,912408	0,638498	0,414937	0,461128	0,405833
5	No log	0,931472	0,643192	0,437831	0,436917	0,440000
6	No log	0,995054	0,643192	0,438597	0,456023	0,431250
7	No log	1,045365	0,638498	0,440074	0,451649	0,435000
8	No log	1,031146	0,647887	0,452889	0,456664	0,453750

Fuente: Elaboración propia.

Tabla 14: Métricas Entrenamiento 6.
Learning rate = 0,0001, Epoch = 8, Batch size = 8, 850 datos.

Epoch	T. Loss	V. Loss	Accuracy	F1	Precision	Recall
1	No log	0,976316	0,652582	0,227962	0,248792	0,266046
2	No log	0,899017	0,666667	0,317312	0,277809	0,385352
3	No log	0,908894	0,643192	0,303628	0,277797	0,348303
4	No log	0,979545	0,661972	0,332063	0,352559	0,333858
5	No log	1,087647	0,615023	0,378124	0,378219	0,391951
6	No log	1,098725	0,652582	0,386945	0,375739	0,401541
7	0.79950	1,230750	0,652582	0,377172	0,387993	0,389861
8	0.79950	1,355646	0,652582	0,363007	0,386900	0,380394

Fuente: Elaboración propia.

De los Ent. 4, 5 y 6, se observa como aumentaron los valores iniciales con la adición de datos. Además, se nota una mejora significativa en la Exactitud, Precisión, Exhaustividad y F1, ya que estás si varían entre épocas, lo que antes no pasaba, pero al alcanzar las últimas épocas estas mejoras se estabilizan sin lograr un avance significativo. Esto sugiere que el modelo podría estar alcanzando un punto de sobreajuste, lo que limita su capacidad para mejorar con los hiperparámetros utilizados (*learning rate* bajo y *batch size* pequeño).

Sin duda, el mejor entrenamiento (Ent. 6) se consigue con la combinación de más datos, un mayor *batch size* y mayor *learning rate*, además de las épocas suficientes para evidenciar el cambio. Es por esto, que el resto de las secciones se entrenan con la misma configuración de hiperparámetros del Ent. 6, y estos quedan fijos como se muestra en la Tabla 15.

Tabla 15: Valor de hiperparámetros definidos para Optuna.

Hiperparámetro	Valor fijo
Learning rate	1e-4
Epoch	8
Batch size	32

Fuente: Elaboración Propia

Aparte de los entrenamientos anterior, se realiza un entrenamiento adicional (Ent. 7), con learning rate = 0.0004, epoch = 4, batch size = 48 y 1700 datos (20% de la data evaluada), para contrastar como cambia el modelo de ser entrenado así. Para este entrenamiento se usó Colab Pro, ya que la versión libre no es capaz de ejecutar el código con dicha cantidad de data real por limitación de memoria RAM, y un procesador más lento.

Tabla 16: Métricas Entrenamiento 7.
Learning rate = 0,0004, Epoch = 4, Batch size = 48, 1700 datos.

Epoch	T. Loss	V. Loss	Accuracy	F1	Precision	Recall
1	No log	1,054475	0,616471	0,231924	0,253856	0,274648
2	No log	0,872927	0,647059	0,308790	0,370265	0,332446
3	No log	0,844013	0,663529	0,386393	0,405369	0,388042
4	No log	0,863073	0,668235	0,400403	0,425257	0,399265

Fuente: Elaboración propia.

En comparación a la Tabla 14, la Tabla 16 refleja un ajuste estratégico de los hiperparámetros, incluyendo un incremento en el *learning rate* y un aumento en el *batch size*. Estos ajustes optimizaron el proceso de entrenamiento, permitiendo al modelo aprender de manera más rápida y manejar la información de forma más eficiente. Como resultado, el modelo mostró un rendimiento más estable y una mayor capacidad de generalización, alcanzando niveles satisfactorios en menos épocas. Estos cambios apuntan a que una mejora en el *hardware* computacional sumado a una configuración más robusta mejora significativamente la eficacia del modelo, incluso con un menor número de iteraciones.

5.2 Resultados de Entrenamientos.

Las métricas de los entrenamientos para las secciones del Canvas, con los hiperparámetros fijos, y con el 10% de la data evaluada según la Tabla 15 y 36 respectivamente, se muestran en las Tablas 17, 18 y 19.

Tabla 17: Métricas para entrenamiento para sección Clientes.

Epoch	T. Loss	V. Loss	Accuracy	F1	Precision	Recall
1	No log	0,976316	0,652582	0,227962	0,248792	0,266046
2	No log	0,899017	0,666667	0,317312	0,277809	0,385352
3	No log	0,908894	0,643192	0,303628	0,277797	0,348303
4	No log	0,979545	0,661972	0,332063	0,352559	0,333858
5	No log	1,087647	0,615023	0,378124	0,378219	0,391951
6	No log	1,098725	0,652582	0,386945	0,375739	0,401541
7	0,79950	1,230750	0,652582	0,377172	0,387993	0,389861
8	0,79950	1,355646	0,652582	0,363007	0,386900	0,380394

Fuente: Elaboración Propia

Tabla 18: Métricas para entrenamiento para sección Propuesta de Valor.

Epoch	T. Loss	V. Loss	Accuracy	F1	Precision	Recall
1	No log	1,006648	0,639810	0,402244	0,428908	0,417890
2	No log	0,956710	0,672986	0,465494	0,470909	0,468708
3	No log	1,231846	0,436019	0,326041	0,344175	0,378332
4	No log	1,212092	0,530806	0,402276	0,437057	0,389124
5	No log	1,520102	0,535545	0,430078	0,463623	0,421386
6	No log	1,857754	0,488152	0,383539	0,407215	0,394990
7	No log	1,785733	0,582938	0,418981	0,424171	0,419071
8	No log	1,920651	0,559242	0,436652	0,490747	0,423368

Fuente: Elaboración Propia

Tabla 19: Métricas para entrenamiento para sección Ingresos.

Epoch	T. Loss	V. Loss	Accuracy	F1	Precision	Recall
1	No log	0,697001	0,714953	0,556604	0,713799	0,523140
2	No log	0,717162	0,686916	0,473796	0,736248	0,465269
3	No log	0,694386	0,714953	0,486605	0,758314	0,476577
4	No log	0,779077	0,738318	0,590086	0,689231	0,562405
5	No log	0,971313	0,714953	0,567651	0,631577	0,543522
6	No log	1,161677	0,663551	0,571040	0,641913	0,550697
7	No log	1,319958	0,672897	0,569910	0,640196	0,545408
8	No log	1,326819	0,682243	0,577694	0,647549	0,553707

Fuente: Elaboración Propia

De las Tablas 16, 17 y 18 se observa lo previsto en la visualización de datos, las mejores métricas están dadas para la sección Ingresos. El hecho de presentar solo tres categorías ayuda a tener mejor exactitud (Accu) y precisión (Pres). Dado que es más fácil clasificar las respuestas medias (3 y 5 para el resto de las secciones), en una sola categoría (4).

Por otro lado, la que presenta las peores métricas fue la sección de Propuesta de Valor, que parte con buena exactitud, pero baja precisión, exhaustividad (Rec) y F1. Pero a medida que pasaron las épocas se estabilizaron, aunque en dicha estabilización la exactitud de PdV baja bastante. A diferencia de la sección Clientes, donde se pudo aumentar el resto de las métricas, manteniendo estable la exactitud, cercana a 0,65.

Desde la perspectiva de buscar el mejor conjunto de métricas. Se puede decir que en general el modelo presenta métricas medias, ya que después de la época 8, la mayoría de las medidas están cerca de 0,5; Accu = [0,56, 0,69], Pres = [0,39, 0,49], Rec = [0,38, 0,42] y F1 = [0,36, 0,44]. Algo alejado de lo que se quiere, para la realización de una predicción confiable, se esperan métricas más cercanas a 1, al menos mayores a 0,75.

Al comparar el Ent. 6 con el 7, se observan mejoras en las métricas, pero la mejora no es tan significativa de cómo se esperaría de un entrenamiento con el doble de datos. Con esto se infiere que la efectividad del modelo tiene más limitaciones por la propia variabilidad de los textos, que por la falta de datos durante el entrenamiento.

Por otro lado, desde la perspectiva de la menor pérdida, esta medida se dio en la segunda época en la sección Clientes, dónde coincide con la máxima exactitud, mismo caso con PdV, y durante la tercera época para Ingresos. Por tanto, sería innecesario fijar *epoch* mayor a 3, puesto que sería tiempo de entrenamiento desperdiciado. Pero al mirar las métricas de precisión, exhaustividad y F1, se marca que son valores muy bajos, cercanos a 0,3, indicando un alto número de falsos positivos y falsos negativos, aunque para este modelo serían categorías mal predichas para cada sección.

5.3 Resultados de la Validación

Se tienen las validaciones dadas por la matriz de confusión para cada sección del Canvas y se distribuyen como muestra en las Figuras 5, 6 y 7. Donde se observan los resultados del modelo después del testeo en matrices de C x C, con *C* igual al número de clases o notas, C=4 en Clientes y PdV, y C=3 en Ingresos. En la diagonal principal se marcan

los resultados que el modelo predice la misma nota que había tenido en la evaluación de Sercotec.

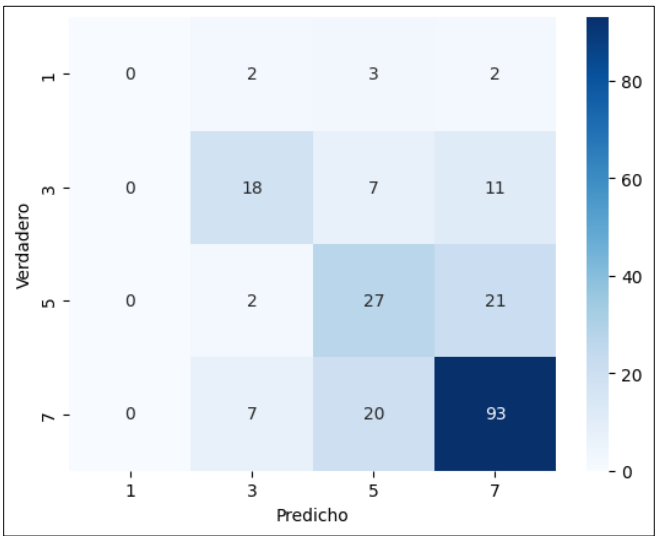


Figura 5: Matriz de confusión, sección Clientes.
Fuente: Elaboración propia.

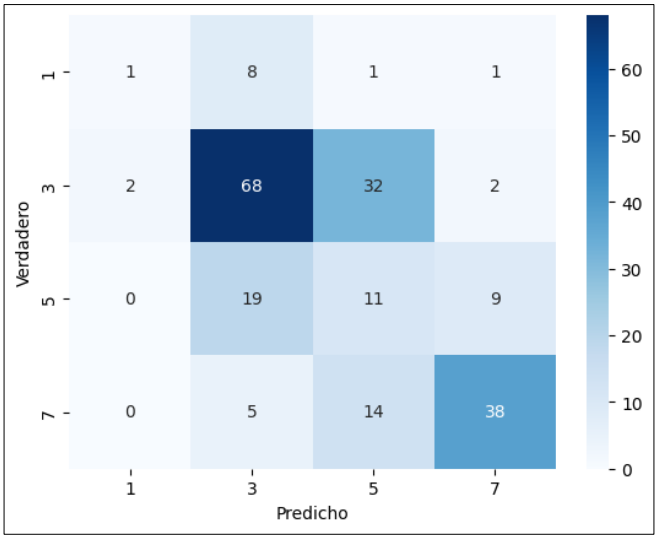


Figura 6: Matriz de confusión, sección Propuesta de Valor.
Fuente: Elaboración propia.

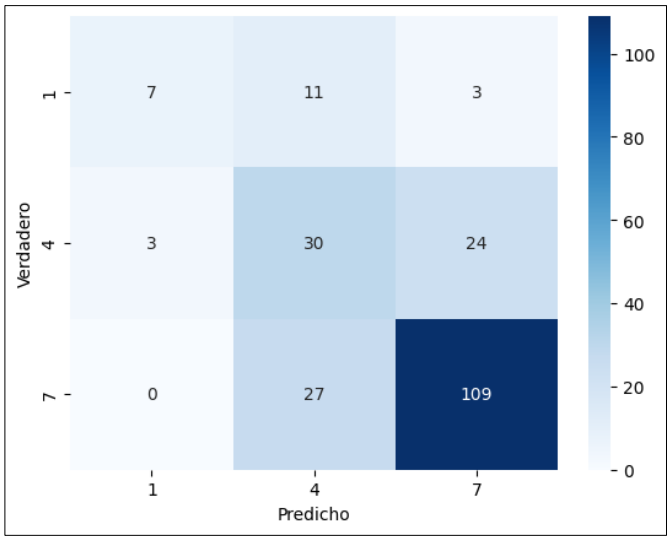


Figura 7: Matriz de confusión, sección Ingresos.
Fuente: Elaboración propia.

Las matrices de confusión confirman lo previsto en los datos del entrenamiento, hay bastantes falsos positivos y falsos negativos. Para el caso de Clientes, hay varios falsos 7, predichos como 5, y viceversa; Además, es preocupante el 0 de 0 en la nota (o clase) 1, que evidencia que habría un mal comportamiento al usar el modelo ante esta clase, pues los textos 7 de clase 1, los predijo como 3, 5, y 7. Se tendrían que revisar dichas calificaciones, para contrastar si la nota en estos casos fue por no respetar la rúbrica de Sercotec, o por un posible error humano a la hora de evaluar.

Lo mismo ocurre con la respuesta a la PdV, donde se presentan muchos falsos positivos y falsos negativos, prediciendo erróneamente notas de 3 como si fueran 5, y viceversa. Es notable que la mayor cantidad de verdaderos positivos se da con la nota 3, a pesar de que la distribución está inclinada hacia la nota 7 y luego 5. Esto indica que la sección tiene un comportamiento único debido a la diversidad de respuestas que pueden proporcionar los postulantes.

Por último, al evaluar las respuestas a Ingreso, queda claro que se comporta similar a Clientes, en cuanto al que la mayor exactitud está en

la nota 7, tal como se esperaba. Además, se da cuenta que mientras menos cantidad de categorías hay, mayor es el número de aciertos en cada una de estas.

5.4 Resultados de Predicciones

Al validar el modelo, se ingresaron los datos no evaluados para realizar una predicción con ellos, y evaluar de manera tangible cómo funciona el modelo y que respuesta da antes los textos sin nota. De los conjuntos de 1% para Clientes, PdV e Ingresos se separa una muestra de los primeros cinco datos, y se muestran en las Tablas 20, 21 y 22. Así, queda explicito que el modelo funciona, dando predicciones verosímiles para cada respuesta. Además, se evidencia la correlación entre la nota y factores como el largo del texto, o las palabras con mayor TF-IDF.

Tabla 20: Predicciones del modelo en data no evaluada, sección Clientes

Texto de respuesta	Nota
1- hombres y mujeres, mayores de 18 anos, de todo chile, que aman la naturaleza y la artesania, que tengan negocio de venta de plantas, en local, ferias y exposiciones, o delivery de regalos, que quieran dar un valor agregado a sus productos. estos clientes compran x mayor. 2- hombres o mujeres (ninos/jovenes/adulto/adulto mayor) son personas que van a la expo (expo artesanos algarrobo) residentes o turistas que buscan un regalo o souvenir o algo para decorar con vida sus espacios. 3- personas que dirigen talleres de arte, colegios o jardines infantiles, profesionales de educacion que ven en el hecho de pintar el desarrollo de habilidad motriz y quieren comprar x mayor para sus estudiantes. para diferentes celebraciones o talleres de arteterapia o manualidades. 4-clientes que nos buscan atraves de redes sociales en nuestro grupo de facebook pasion suculenta chile, o nuestra pagina en concreto deco arte. son personas mayores de 18 anos hombres y mujeres que aman la naturaleza y la decoracion que tienen plantas pero quieren maceteros decorativos de calidad y no desechables para decorar sus espacios.	7
se le esta entregando valor a los clientes que se apasionan por tener un trabajo basicamente artesanal con el sello de identidad de la region, potenciando la economia verde o ecointeligencia, para realzar la riqueza turistica de cada lugar donde se elaboran estos productos. el desarrollo de este negocio va dirigido al los grupos etarios de 15-60 anos comprendiendo que la mayor cantidad de personas en la era digital en chile posee estas edades.	5
personas naturales	3
comercio minorista, y al detalle	3
a todo tipo de cliente que este familiarizado con el comercio electronico... cliente de personalidad independiente, que busca y analiza por si solo las caracteristicas de un producto sin interactuar cara a cara con vendedores para comprar. segmento masculino y femenino de clase media, adulto joven trabajador/a, duena/o de casa, ninos, millenials, etc. en general es la versatilidad, variedad e innovacion del stock de productos, en mi opinion, es lo que determina abarcar mas segmentos de clientes...	7

Fuente: Elaboración propia.

Tabla 21: Predicciones del modelo en data no evaluada, sección PdV.

Texto de respuesta	Nota
porque soy responsable, rapida en la entrega ademas que cada producto es unico y personalizado.	5
porque tendre stock de productos asequibles para todos/as. en diferentes tallas, modelos y variedad. y hare entregas a domicilio, ademas de ofrecer precios y promociones al alcance del bolsillo.	3
porque sera un emprendimiento cercano a la gente ademas de tener productos con su valor al alcance de las personas	5
la adquisicion de la maquinaria gastronomica permitira elaborar productos de alta calidad y sabor, innovando en el sector donde hay gran cantidad de emprendimientos familiares y ofrecer a los clientes que transitan por la zona una nueva oferta gastronomica de pollos asados, papas y/o otros acompanamientos de linea gourmet que no existe en el sector. el producto a ofrecer a los clientes no es solo la elaboracion tipica y tradicional del pollo asado con papas fritas si no que se caracterizara en presentar una carta nueva y atractiva que incluira diversas recetas combinando lo tradicional con otras culturas (pollo estilo americano, pollo mexicano, pollo peruano, etc.) y asi ofrecer una alternativa gastronomica diferente, de calidad y buen servicio a nuestros futuros clientes. la produccion y venta desde mi domicilio permitira un mejor servicio al cliente para solicitar pedido a traves de; llamada, whatsapp y pagina web, retiro en domicilio, despachos de aplicaciones delivery (rappi, uber eats, pedidos ya, glovo) y formas de pago como efectivo, transferencia, debito o crecido a traves de transbank para ofrecer una mayor comodidad y atencion para el cliente.	5
nuestra empresa entrega a nuestros clientes ,servicio de armado y despacho totalmente gratis a diferencia de otro mercado , nuestros productos son realizados a mano lo que le dara una mejor terminacion del productos ya que se realiza pensando en la seguridad que puede entregar nuestro negocio	3

Fuente: Elaboración propia

Tabla 22: Predicciones del modelo en data no evaluada, sección Ingresos.

Texto de respuesta	Nota
porque es un proyecto innovador y tiene como proposito restaurar las especies nativas, para recuperar el ecosistema, dando valor a las propiedades medicinales y las caracteristicas de adaptacion de nuestras especies en la region. por representar el bosque siempre verde siendo una de las especie que consume menos agua, se adapta favorablemente al clima y suelo, que no requiere de poda ni fertilizantes, contribuyendo con la polinizacion, al ser una especie con flor. todos los segmentos de clientes pagaran segun sus necesidades o preferencias, recibiendo boleta o factura, por cualquier medio de pago; -pago efectivo -pago mediante transferencia -pago electronico (tarjeta webpay)	7
cada cluente paga por algo que deceee y yo quiero dar en eso producto como ellos lis quiere casi todos pagan por transferencia	7
porque sus ganancias son mas rentables por una entraga rapida, eficiente y real pagan en efectivo; transferencia u otros.	4
los clientes actuales estan dispuestos a pagar por entretencion. estan dispuestos a adquirir diversas formas de entretencion que permita tener un evento familiar agradable y entretenido para sus hijos. les ofreceremos una forma distinta de celebrar sus eventos, con juegos tipicos y ludicos, que permitiran a ninos entretenerse sanamente y a sus padres volver a sentir la magia de ser nino por un dia. actualmente , nuestros potenciales clientes solo tienen disponibles juegos inflables para entretener a los ninos en sus fiestas, y no existe en la zona una oferta distinta. generalmente, los servicios disponibles son cancelados en efectivo. en primera instancia el patio tendra disponible pagos en efectivo y por transferencia, pero en el corto plazo instauraremos pagos con tarjeta.	7
estamos abiertos a que la cancelacion de los productos sea a traves de pago en efectivo y tambien con debito o a traves de transferencia ya que tambien ofreceriamos el servicio de caja vecina para el pago de cuentas y giro en dinero	7

Fuente: Elaboración propia.

Desde las Tablas 20, 21 y 22 se observa que en general el modelo predijo bien, ya que las predicciones erróneas no varían por más de una categoría. Al levantar la información en una matriz de confusión se presenta la Tabla 23.

Tabla 23: Matriz de confusión para las predicciones del modelo[3].

Notas		Verdadero				
		1	**3**	**4**	**5**	**7**
Predicción	**1**	0	0	0	0	0
	3	1	3	0	0	0
	4	0	0	1	0	0
	5	0	2	0	1	1
	7	0	0	1	1	4

Fuente: Elaboración propia.

Si, en lugar de predecir entre las tres clases de la sección Ingresos, se consideran las cuatro clases de las otras secciones (nota 1, 3, 5 y 7), la clase 4 se reasigna como clase 5 (según la rúbrica de Sercotec), obteniéndose una tabla homologada a cuatro clases, como se muestra en la Tabla 24.

Tabla 24: Matriz homologada para las predicciones del modelo.

Notas		Verdadero			
		1	**3**	**5**	**7**
Predicción	**1**	0	0	0	0
	3	1	3	0	0
	5	0	2	2	1
	7	0	0	2	4

Fuente: Elaboración propia.

Con una exactitud de 0,6, que es un valor similar a las obtenidas en los entrenamientos. Aunque es un dato poco significativo puesto que son solo cinco respuestas por sección, y las categorías están combinadas.

[3] Las notas verdaderas vs predichas fueron puestas de manera manual contrastando las respuestas con la rúbrica de Sercotec (ver Anexo A, Tabla 25).

6. Conclusiones

Esta memoria tenía como objetivo general el entrenar un modelo basado en procesamiento de lenguaje natural para clasificar las respuestas al Canvas de los postulantes al programa Capital Semilla Emprende de Sercotec, donde se utilizaron datos reales provenientes directamente desde la base de datos de postulaciones a Sercotec. Se concluye que el objetivo central fue satisfecho, ya que, en efecto se entrenó un modelo de procesamiento de lenguaje natural bajo la metodología del uso de modelos Transformers, usando de base BETO como modelo pre-entrenado en el procesamiento de lenguaje español.

En lo que respecta netamente al modelo entrenado se puede concluir lo siguiente:

- Eficacia del Entrenamiento: el modelo presenta métricas de desempeño medias o moderadas. Después de la época 8, la exactitud varía entre 0,56 y 0,69, la precisión entre 0,39 y 0,49, la exhaustividad entre 0,38 y 0,42, y la medida F1 entre 0,36 y 0,44, lo cual resulta insuficiente para una predicción completamente confiable. Esta eficacia podría mejorar con un entrenamiento más exhaustivo, utilizando más datos y ajustando mejor los parámetros, aunque su potencial de mejora está limitado, especialmente en relación con los recursos computacionales disponibles durante el entrenamiento.
- Secciones Canvas: como se previsualizó la sección Clientes resultó ser una sección media, es bastante estable en sus métricas y obtuvo un comportamiento intermedio entre Propuesta de Valor e Ingresos. Mientras la sección Propuesta de Valor presenta mucha variabilidad en cuanto a las respuestas, que vuelve más difícil la predicción. La

sección Ingresos presenta la ventaja de tener una categoría menos, por lo que fue más fácil la clasificación.

- Épocas de Entrenamiento: no es necesario fijar un número de épocas alto, ya que la pérdida mínima se alcanzó en la segunda o tercera época. Muchas épocas adicionales resultan en tiempo de entrenamiento desperdiciado.
- Uso de Datos Reales: trabajar con datos reales introduce limitaciones y problemas de distintos tipos, y cada problema afecta de manera directa y negativa al entrenamiento, y por consiguiente a las calificaciones generadas por el modelo.
- Optuna: La iteración y búsqueda de resultados con configuraciones de hiperparámetros utilizando el optimizador Optuna ayudó a ajustar mejor el modelo, aunque no se lograron métricas óptimas.
- Uso del modelo: implementar este modelo (luego de la respectiva integración) como ayuda a las labores actuales de Sercotec, es una estrategia efectiva para la evaluación de postulante. Dónde el modelo predictor haría un primer filtrado de formularios, para que los funcionarios evaluaran definitivamente una carga mucho menor a la actual, tomando solo las notas 5 o superiores, por ejemplo. La combinación de predicciones automáticas con revisión humana asegura que ningún formulario quede sin una evaluación, y que solo las mejores postulaciones ganen los fondos.
- Alta aplicabilidad: este tipo de modelos puede extenderse y aplicarse fácilmente al resto de programas de Sercotec; incluso a otras organizaciones con problemas/situaciones similares. Sin embargo, la calidad de los datos de entrada sigue siendo un factor crítico para mejorar el rendimiento del modelo.
- La principal limitante del modelo pasa principalmente por los recursos computacionales que tiene el equipo donde se ejecuta el

entrenamiento, puesto que con el *hardware* disponible costó llegar a un modelo más eficiente en término de métricas.

- Futuras investigaciones pueden extender los análisis de esta memoria de título de varias maneras, como, por ejemplo: mediante la ejecución de un entrenamiento más exhaustivo con mejores equipos. Mediante la integración del modelo donde las respuestas se procesan juntas, o donde se extienda el modelo a los demás programas de Sercotec.

Respecto al uso de herramientas tecnológicas, como inteligencias artificiales u otros modelos informáticos en general, se concluye lo siguiente:

- Accesibilidad: herramientas como Chat GPT y Dall-E han hecho que la tecnología de IA sea accesible a un amplio número de usuarios.
- IA en optimización de procesos: las inteligencias artificiales (IAs) han demostrado ser herramientas valiosas para la optimización de procesos en diversos sectores, facilitando la automatización de tareas repetitivas y mejorando la eficiencia operativa. En la realización de este estudio se usó la IA Gemini presente en Google Colab, lo que evidencia que el uso y masificación de esta tecnología crea un *Loop* positivo para el desarrollo de más herramientas tecnológicas.
- Herramientas tecnológicas en Sercotec: la integración de IAs, y otras soluciones informáticas en las operaciones de Sercotec puede mejorar significativamente la gestión de proyectos y la asignación de recursos. Algo muy necesario para las organizaciones de la actualidad.
- Alcances y limitaciones: es difícil saber los alcances que pueden tener herramientas digitales como las IAs, puesto que día a día se generan nuevas aplicaciones, y es impredecible hasta dónde llegará su avance

incluso en el corto plazo. Desde el lado de las limitantes se observan 2 focos. Primeramente, se tiene el foco computacional, ya que para la implementación de este tipo de herramientas es fundamental contar con buen *hardware*, y eso es costoso en término de costo de equipos y de tiempo. Además, se tiene la limitación legal, en poco tiempo debería aprobarse la Ley de Inteligencia Artificial en Chile, donde se espera se establezca un marco legal para el desarrollo y uso de IA.

Referencias

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A Next-generation Hyperparameter Optimization Framework* (arXiv:1907.10902). arXiv.

Alarcón, J. M. (2020). Aclarando conceptos: Inteligencia Artificial, Machine Learning, Deep Learning, Big Data y Ciencia de Datos. *Campus MVP*. https://www.campusmvp.es/recursos/post/aclarando-conceptos-inteligencia-artificial-machine-learning-deep-learning-big-data-y-ciencia-de-datos.aspx

Alonso, R. (2024). IA, Machine Learning y Deep Learning, ¿cuál es la diferencia? *Hardzone*. https://hardzone.es/tutoriales/rendimiento/diferencias-ia-deep-machine-learning/

Álvarez, C. (2021). *Aplicación de procesamiento de lenguaje natural sobre una encuesta de satisfacción.* [Memoria de Título para pregrado]. Pontificia Universidad Católica de Chile.

Amat, J. (2020). Machine learning con Python y Scikit-learn. *Ciencia de Datos*. https://cienciadedatos.net/documentos/py06_machine_learning_python_scikitlearn

Arenas, M., Arriagada, G., Mendoza, M., & Prieto, C. (2023). *Una breve mirada al estado actual de la Inteligencia Artificial*.

Baena, P. (2023). ¿Puede la inteligencia artificial optimizar los procesos de tu empresa? *OBS Business School*. https://www.obsbusiness.school/blog/puede-la-inteligencia-artificial-optimizar-los-procesos-de-tu-empresa

Brownlee, J. (2019). 7 aplicaciones del aprendizaje profundo para el procesamiento del lenguaje natural. *Machine Learning Mastery*.

https://machinelearningmastery.com/applications-of-deep-learning-for-natural-language-processing/

Burns, E. (2021). Aprendizaje profundo (deep learning). *Computer Weekly*. https://www.computerweekly.com/es/definicion/Aprendizaje-profundo-deep-learning

Cabezas, M. (2023). *Análisis de patrones sistemáticos en el proceso de evaluación de postulantes a programas SERCOTEC.* [Memoria de Título para pregrado]. Universidad de Concepción.

Cámara de Diputados. (2023). *Propuesta de ley: Ley de Inteligencia Artificial*. https://www.camara.cl/legislacion/ProyectosDeLey/tramitacion.aspx?prmID=16416&prmBOLETIN=15869-19

Cámara de Senadores. (2024). *Boletín 15869-19*. https://www.senado.cl/appsenado/templates/tramitacion/index.php?boletin_ini=15869-19

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2023). *Spanish Pre-trained BERT Model and Evaluation Data* (arXiv:2308.02976). arXiv.

Caro, B. (2021). *Estudio de aplicaciones de la Inteligencia Artificial en el desarrollo de proyectos de ingeniería civil* [Memoria de Título para pregrado]. Universidad de Chile.

Caro, E. (2017). *La Cuarta Revolución Industrial*. Universidad de Sevilla.

Clavijo, C. (2024). *Modelo Canvas: Qué es, para qué sirve, cómo se usa y ejemplos*. https://blog.hubspot.es/sales/modelo-canvas

Conejeros, F. (2023). *Construcción de un predictor de evaluación del Modelo de Negocios para el Programa Capital Semilla Emprende de Sercotec con el modelo BETO* [Memoria de Título para pregrado]. Universidad de Concepción.

Czakon, J. (2023). Optuna vs Hyperopt: ¿Qué biblioteca de optimización de hiperparámetros debería elegir? *Neptune*. https://neptune.ai/blog/optuna-vs-hyperopt

Dallas, J. (2024). La Evolución y el Paisaje Actual de la IA y el Aprendizaje Automático. *iCorps*. https://blog.icorps.com/evolution-of-ai-and-machine-learning

Deming, C., Dekkati, S., & Desamsetti, H. (2018). Exploratory Data Analysis and Visualization for Business Analytics. *Asian Journal of Applied Science and Engineering*, *7*(1).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv.

Donoso, S. (2021). *Entrenamiento y evaluación de modelos pequeños de lenguaje natural* [Memoria de Título para pregrado]. Universidad de Chile.

El Empleo. (2023). *Monotonía laboral, un enemigo silencioso*. El Empleo. https://www.elempleo.com/co/noticias/consejos-profesionales/monotonia-laboral-un-enemigo-silencioso-2551

El Mostrador. (2023). *Ministerio de Ciencia inicia discusión sobre los alcances socioculturales y éticos de la Inteligencia Artificial en Chile*. https://www.elmostrador.cl/destacado/2023/04/19/ministerio-de-ciencia-inicia-discusion-sobre-los-alcances-socioculturales-y-eticos-de-la-inteligencia-artificial-en-chile

Es, S., & Bajaj, A. (2023). Hyperparameter Tuning en Python: Una Guía Completa. *Neptune*. https://neptune.ai/blog/hyperparameter-tuning-in-python-complete-guide

Gayubas, A. (2017). Revolución Industrial. *Enciclopedia Humanidades*. https://humanidades.com/revolucion-industrial/

Giraldo, A., & Orozco, A. (2023). Evolución del procesamiento natural del lenguaje. *TecnoLógicas*, *26*(56).

Gobierno de España. (2022). *Así es MarIA, la primera inteligencia artificial de la lengua española*. https://datos.gob.es/es/blog/asi-es-maria-la-primera-inteligencia-artificial-de-la-lengua-espanola

Gorini, M. (2024). ¿Cuál es la diferencia entre el machine learning y el deep learning? *Bismart*. https://blog.bismart.com/diferencia-machine-learning-deep-learning

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Gonzalez-Agirre, A., Armentano-Oller, C., Rodriguez-Penagos, C., & Villegas, M. (2022). MarIA: Spanish Language Models. *Procesamiento Del Lenguaje Natural*.

Holdsworth, J. (2024). *What is NLP (natural language processing)?* https://www.ibm.com/topics/natural-language-processing

IBM. (2023). *¿Qué es el etiquetado de datos?* https://www.ibm.com/es-es/topics/data-labeling?mhsrc=ibmsearch_a&mhq=etiquetado

IIC. (2023). Modelo de lenguaje español: RigoBERTa. *Instituto de ingeniería del conocimiento*. https://www.iic.uam.es/inteligencia-artificial/procesamiento-del-lenguaje-natural/modelo-lenguaje-espanol-rigoberta/

ISDI. (2023). 10 desventajas de la inteligencia artificial a tener en cuenta. *ISDI Digitalent Group*. https://www.isdi.education/es/blog/desventajas-de-la-inteligencia-artificial

López, J. J., & Gonzales, F. (2021). *Análisis de sentimiento de comentarios en español en Google Play Store usando BERT* [Universidad Nacional de San Agustín de Arequipa]. https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-33052021000300557&lng=en&nrm=iso&tlng=en

Maguregui, C. (2023). Inteligencia artificial: De la ciencia ficción a la realidad. *Portal Educ.Ar*. https://www.educ.ar/recursos/159014/inteligencia-artificial-de-la-ciencia-ficcion-a-la-realidad

Mamani, Z. (2022). Proceso de machine learning para determinar la demanda social de puestos de empleo de profesionales de TI. *Industrial Data*, *25*(2), 275-300.

Martineau, K. (2023). What is generative AI? *IBM*. https://research.ibm.com/blog/what-is-generative-AI

Matich, D. (2001). *Redes Neuronales: Conceptos Básicos y Aplicaciones.* Universidad Tecnológica Nacional.

Maurya, A. (2012). *Running Lean, Second Edition* (2.ª ed.). O'Reilly Media, Inc.

McCarthy, J. (1955). *A proposal for the dartmouth summer research project on artificial intelligence*.

McCulloch, & Pitts. (1943). *A logical calculus of the ideas immanent in nervous activity*.

Mejía-Giraldo, J. F. (2019). Propósitos organizacionales como alternativa para los problemas que proponen los modelos canvas y lean canvas. *Innovar*, *29*.

Mendoza, S. (2024). Historia de la Inteligencia Artificial: Evolución e hitos importantes. *Hiram Noriega*. https://hiramnoriega.com/61641/inteligencia-artificial-historia-evolucion/

Ministerio de Ciencia. (2021). *Chile presenta la primera Política Nacional de Inteligencia Artificial*. https://minciencia.gob.cl/noticias/chile-presenta-la-primera-politica-nacional-de-inteligencia-artificial/

Miró, M. (2023). Qué es un mapa de empatía, cómo crearlo y ejemplos. *Michel Miró*. https://michelmiro.com/que-es-un-mapa-de-empatia-como-crearlo-y-ejemplos/

Mittal, A. (2023). IA generativa: La idea detrás de CHATGPT, Dall-E, Midjourney y más. *Unite*. https://www.unite.ai/es/generative-ai-the-idea-behind-chatgpt-dall-e-midjourney-and-more/

Moreno, A. (2018). Procesamiento del lenguaje natural ¿qué es? *Instituto de Ingeniería del Conocimiento*. https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/

Moya, M. A. (2024). Claves de la nueva Estrategia de Inteligencia Artificial. *El Radar*. https://www.elradar.es/nueva-estrategia-inteligencia-artificial/

Mustapic, B. (2024). Introducción al TF-IDF: Qué es y cómo utilizarlo. *Introducción al TF-IDF: Qué es y cómo utilizarlo*. https://es.semrush.com/blog/tf-idf-es/

Norvig, P. (2023). Pyspellchecker. *Python Package Index*. https://pypi.org/project/pyspellchecker/

Optuna. (2018). *Optuna: Un marco de optimización de hiperparámetros*. Optuna. https://optuna.readthedocs.io/en/stable/

Osterwalder, A., & Pigneur, Y. (2009). *Generación de Modelo de Negocios*. Autoeditado

Pandas. (2024). *Pandas documentation*. https://pandas.pydata.org/docs/index.html#pandas-documentation

Porcelli, A. (2020). *La Inteligencia Artificial y la Robótica: Sus dilemas sociales, éticos y jurídicos*. Universidad de Guadalajara, México; Universidad Nacional de Luján, Argentina.

Sanhueza, N. (2023). Inteligencia artificial: Experiencias comparadas y perspectivas para Chile. *Estado Diario*. https://estadodiario.com/columnas/inteligencia-artificial-experiencias-comparadas-y-perspectivas-para-chile/

Sarmiento-Ramos, J. L. (2020). Aplicaciones de las redes neuronales y el deep learning a la ingeniería biomédica. *Revista UIS Ingenierías*, *19*(4).

Schwab, K. (2016a). *La cuarta revolución industrial*. Debate.

Schwab, K. (2016b). La Cuarta Revolución Industrial: Qué significa, cómo responder. *World Economic Forum*. https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/

Sercotec. (2022c). *Anexos Capital Semilla Emprende zonas rezagadas FNDR Región del Maule 2022*. https://www.sercotec.cl/wp-content/uploads/2022/07/anexos-Bases-Emprende-ZR-Maule-2022.docx

Sercotec. (2022b). *Bases de convocatoria Capital Abeja Emprende «Provincia de Concepción»*.

Sercotec. (2023b). *Crece*. Sercotec. https://www.sercotec.cl/crece/

Sercotec. (2023c). *Data Canvas Semilla Innominado UdeC* [Documento interno].

Sercotec. (2022a). Ministerio de Economía lanza fondo de $4.200 millones para emprendedoras y emprendedores. *Noticias Sercotec*. https://www.sercotec.cl/ministerio-de-economia-lanza-fondo-de-4-200-millones-para-emprendedoras-y-emprendedores/

Sercotec. (2023a). *¿Quiénes Somos?* Sercotec. https://www.sercotec.cl/quienes-somos/

Sotaquirá, M. (2018). Inteligencia Artificial vs. Machine Learning vs. Deep Learning. *Codificando Bits*. https://www.codificandobits.com/blog/ia-vs-ml-vs-dl/

Sotaquirá, M. (2022). Https://www.codificandobits.com/blog/matriz-de-confusion/. *Codificando Bits*. https://www.codificandobits.com/blog/matriz-de-confusion/

UNIR. (2021). Deep learning: En qué consiste, ejemplos y aplicaciones. *UNIR*. https://www.unir.net/ingenieria/revista/deep-learning/

Vaca, A. V., García, G., Montoro, H. M., Aldama, N., Samy, D., Betancur, D., Moreno, A., Guerrero, M., & Barbero, Á. (2022). *RigoBERTa: A State-of-the-Art Language Model For Spanish* (arXiv:2205.10233). arXiv.

Van Zandt, P. (2023). Lean Canvas vs. Business Model Canvas: Aprende la Diferencia. *Ideascale*. https://ideascale.com/blog/lean-canvas-vs-business-model-canvas/

Venegas, R. (2021). Aplicaciones de inteligencia artificial para la clasificación automatizada de propósitos comunicativos en informes de ingeniería. *Revista signos*, *54*(107).

Vera-Cruz, C. (2023). Crecen los casos de uso de la IA en las empresas chilenas. *Computer Weekly*. https://www.computerweekly.com/es/cronica/Crecen-los-casos-de-uso-de-la-IA-en-las-empresas-chilenas

Zelcer, M. (2022). *Machine learning y lógicas semióticas: El caso de la publicidad digital*. Universidad Nacional de las Artes.

Zumarán, R., & Cortés, P. (2021). Emprendedores participaron en taller de Modelo Canvas para postular a fondos concursables. *CFT PUCV*. https://cftpucv.cl/noticias/emprendedores-participaron-en-taller-de-modelo-canvas-para-postular-a-fondos-concursables/

Anexos

Anexo A: Criterio de Evaluación Técnica

Tabla 25: Criterio de Evaluación Técnica, Formulario de Idea de Negocio

N	Criterio Modelo CANVAS	Pregunta Formulario	Criterio de evaluación	Rúbrica	Nota	Ponderación Criterio
1	Clientes	¿Quiénes son los principales clientes? ¿A qué tipo de clientes apunta nuestro negocio? *<u>Tipo</u>: armar grupos de clientes de acuerdo a sus características. Tipos de clientes, con una identificación clara, a los cuales quiere llegar nuestro negocio.*	Descripción del o los tipos de clientes al cual está dirigido su producto/servicio. *<u>Describir</u>: implica nombrar y explicar detalladamente el/los elemento/s solicitados.*	El/la postulante describe las características de al menos 2 tipos de clientes a los cuales enfocará su producto/servicio.	7	12%
				El/la postulante describe las características de al menos 1 tipo de cliente al cual enfocará su producto/servicio.	6	
				El/la postulante solo menciona al cliente o los clientes al cual enfocará su producto/servicio, sin describir las características de los mismos.	4	
				El/la postulante no menciona ni describe tipos de clientes a los cuales enfocará su producto/servicio.	1	
2	Elemento diferenciador	¿Por qué los clientes deberían preferirme por sobre los demás? ¿Por qué los clientes deberían preferir mi producto/servicio por sobre los demás? *<u>Elemento diferenciador</u>: Elemento que ayuda a elegir un producto o servicio por sobre otro de similares características.*	Describe por cada tipo de clientes, cuál es el elemento diferenciador por el cual deberían elegir el producto/servicio. *<u>Describir</u>: implica nombrar y explicar detalladamente el/los elemento/s solicitados.*	El/la postulante describe su elemento diferenciador para todos los tipos de clientes identificados.	7	14%
				El/la postulante describe su elemento diferenciador solo para algunos de los tipos de clientes identificados.	5	
				El/la postulante solo menciona su elemento diferenciador y/o lo describe sin mencionar a qué tipo de cliente pertenece.	3	
				El/la postulante no menciona ni describe el elemento diferenciador de su idea de negocio.	1	
3	Medios de distribución/atención	¿A través de qué medios realizo las ventas a mis clientes? ¿Cuáles son los medios,	Describe los medios necesarios para llegar a los clientes y	El/la postulante describe medios de distribución para todos los tipos de cliente identificados, justificando el por qué lo utilizará.	7	7%

		para dar a conocer mi producto/servicio, que prefieren mi/s tipo/s de clientes? ¿Cuáles son los medios con los que obtendría mayor venta en mi modelo de negocio?	dar conocer el producto/servicio, posibilitando la compra. Además comentar por qué esos medios son los más adecuados (financiera y operacionalmente) respecto a cada tipo de clientes.	El/la postulante describe medios de distribución solo para algunos de los tipos de cliente identificados, justificando el por qué lo utilizará.	5	
				El/la postulante solo menciona o describe medios de distribución, sin mencionar a qué tipo de cliente pertenecen y/o por qué se utilizarán.	3	
				El/la postulante no menciona ni describe medios de distribución, ni tampoco hace referencia a qué tipo de cliente pertenecen.	1	
4	Relación con los clientes	¿Qué relación tiene o espera tener con cada tipo de cliente descrito? ¿Alguno de los medios por los cuales busca relacionarse con el cliente, tiene algún costo asociado?	De acuerdo a los tipos de clientes indicados, establecer cuál o cuáles serán los tipos de relación por cada uno de ellos. La relación con los clientes apunta a fidelizar su compra.	El/la postulante describe y justifica la relación para todos los tipos de cliente identificados.	7	7%
				El/la postulante describe y justifica la relación solo para algunos de los tipos de cliente identificados.	5	
				El/la postulante solo menciona o describe la relación con el cliente, sin mencionar a qué tipo pertenece y/o cuál es su justificación.	3	
				El/la postulante no menciona ni describe la relación con el cliente en ningún tipo (de cliente) descrito.	1	
5	Ingresos	¿Por cuál tipo de producto/servicio estarían dispuestos a pagar más nuestros clientes? ¿Por cuál tipo de producto/servicio pagan actualmente los clientes? ¿Qué tipo de medio de pago prefieren utilizar mis clientes?	Describe qué ingresos recibirá el negocio y a través de qué medios.	El/la postulante describe cada uno de los ingresos de su negocio y a través de qué medios de pago los percibirá.	7	7%
				El/la postulante describe los ingresos de su negocio, sin mencionar través de qué medios de pago los percibirá.	5	
				El/la postulante no describe los ingresos de su negocio y/o solo menciona medios de pago.	3	
				El/la postulante no describe qué ingresos ni tampoco a través de qué medios los percibirá.	1	
6	Elementos clave	¿Qué elementos se debe adquirir para generar mi	Descripción de los elementos	El/la postulante describe al menos 2 elementos clave, necesarios para que su	7	7%

		producto/servicio y entregue a los diferentes tipos de clientes?	clave necesarios para que el producto/servicio se genere y se entregue a los clientes.	producto/servicio llegue a sus clientes.		
				El/la postulante describe al menos 1 elemento clave, necesario para que su producto/servicio llegue a sus clientes.	5	
				El/la postulante no describe elementos claves, necesarios para que su producto/servicio llegue a sus clientes.	1	
7	Acciones / actividades clave	¿Qué acciones se deben realizar para que mi producto/servicio se entregue a los diferentes tipos de clientes?	Descripción de las acciones clave necesarias para que el producto/servicio se entregue a los clientes.	El/la postulante describe al menos 2 acciones clave, necesarias para que su producto/servicio llegue a sus clientes.	7	7%
				El/la postulante describe al menos 1 acción clave, necesarias para que su producto/servicio llegue a sus clientes.	5	
				El/la postulante no describe acciones clave, necesarias para que su producto/servicio llegue a sus clientes.	1	
8	Costos	¿Cuáles son los costos (fijos y variables) para el funcionamiento de los elementos y acciones clave definidos?	Definir cuáles son los costos fijos y variables asociados a los elementos y acciones claves de su negocio.	El/la postulante describe la estructura de costos de su idea de negocio, identificando costos fijos y costos variables de cada elemento y acción clave identificados previamente.	7	7%
				El/la postulante describe la estructura de costos de su idea de negocio, identificando costos fijos y costos variables, sin asociarlos necesariamente a cada elemento o acción clave identificada previamente.	5	
				El/la postulante describe la estructura de costos sin separar entre costos fijos y variables y/o no los asocia a elementos ni acciones claves.	3	
				El/la postulante no es capaz de describir la estructura de costos de su idea de negocio.	1	
9	Alianzas clave	¿Cuáles son las alianzas realizadas o a realizar para mejorar la	Definir cuáles son las actuales o futuras alianzas clave	El/la postulante describe a lo menos 2 alianzas clave que pueden mejorar la satisfacción de sus actuales y/o potenciales clientes.	7	4%

		satisfacción de mis clientes?	(redes de trabajo) que mi negocio debe tener para satisfacer de mejor forma a mis clientes.	El/la postulante describe a lo menos 1 alianza clave que pueda mejorar la satisfacción de sus actuales y/o potenciales clientes.	5	
				El/la postulante no describe alianzas clave destinadas a mejorar la satisfacción de los clientes.	1	
10	Sustentabilidad	¿Qué acciones puedo implementar en mi negocio, desde el punto de vista de la eficiencia energética, energías renovables y economía circular? de manera de hacer mi producto o servicio más sustentable. ¿Tenía ya incorporada alguna de estas acciones en el proceso de mi producto o servicio?	Establecer las acciones de eficiencia energéticas, energías renovables y de economía circular involucradas en el proceso productivo de mi producto/servicio.	La idea de negocio presentada es del tipo sustentable o incorpora en la cadena de desarrollo del producto o servicio, acciones de eficiencia energética y/o de energías renovables y de economía circular.	7	8%
				La idea de negocio presentada, indistinta su naturaleza, incorpora en la cadena de desarrollo del producto o servicio, al menos 1 (una) acción de eficiencia energética o de energías renovables o de economía circular.	4	
				La idea de negocio presentada no incorpora en la cadena de desarrollo del producto o servicio, alguna acción de eficiencia energética y/o de energías renovables o de economía circular, ni tampoco integra en su quehacer actividades que aporten a la sustentabilidad.	1	
11	Coherencia Global de la Idea de Negocio	En este ítem se evaluará la coherencia general de la Idea de Negocio en el formulario de postulación (Canvas), sobre la base de la información incorporada en los correspondientes criterios.	Coherencia en el formulario (Canvas), respecto a los clientes y elemento diferenciador determinados y cómo se refleja los mismos en los demás criterios del modelo.	Se puede observar un **alto** nivel de coherencia en la idea de negocio formulada, desde el punto de vista del/los tipo/s de clientes descrito/s y elemento diferenciador, lo cual se refleja también en los demás criterios de evaluación.	7	20%
				Se puede observar un **alto** nivel de coherencia en la idea de negocio formulada, desde el punto de vista del/los tipo/s de clientes descrito/s y elemento diferenciador, no obstante, éste no se ve claramente reflejado en alguno de los demás criterios de evaluación.	5	

				Se puede observar un nivel de coherencia en la idea de negocio formulada, desde el punto de vista del/los tipo/s de clientes descrito/s y elemento diferenciador, no obstante, éste no se ve reflejado en los demás criterios de evaluación.	4	
				Se puede observar un **bajo** nivel de coherencia en la idea de negocio formulada, desde el punto de vista del/los tipo/s de clientes descrito/s y elemento diferenciador, lo que se ve reflejado también en los demás criterios de evaluación.	3	
				No existe coherencia en la idea de negocio formulada, desde el punto de vista del/los tipo/s de clientes descritos y elemento diferenciador, ni tampoco entre éstos y los demás criterios de evaluación.	1	

Fuente: Sercotec (2022c)

Anexo B: Variables y descripción de la base de datos

Tabla 26: Variables y descripción de la base de datos.

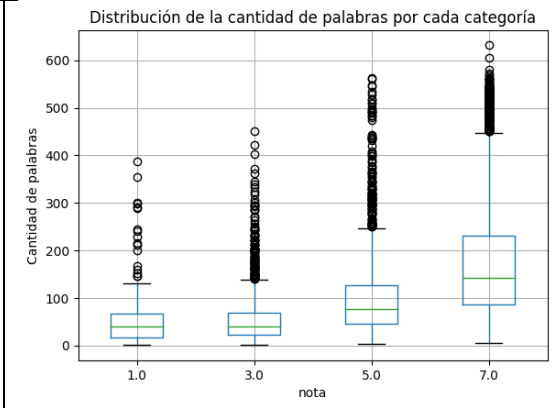
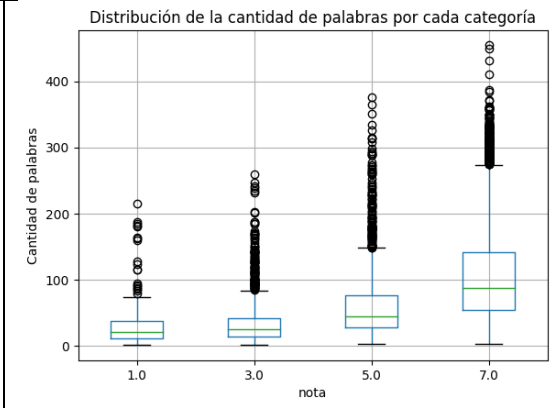
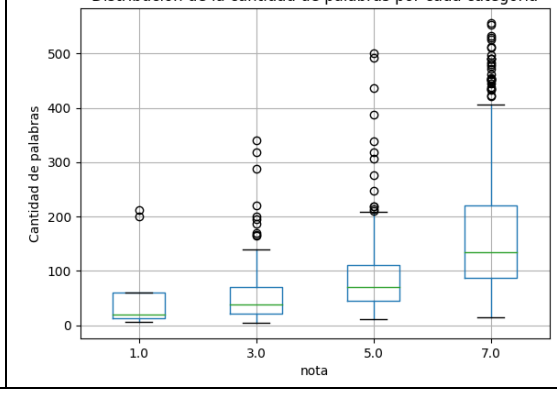
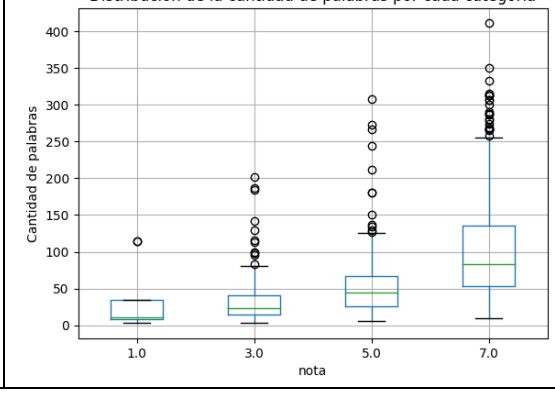
Columna/Variable	Descripción
año_instrumento	Año de la convocatoria
instrumento.id	ID del instrumento
instrumento.desc	Nombre del instrumento
linea_acción.id	ID de la línea de acción
linea_acción.desc	Nombre de la línea de acción
proyecto.id	ID del Proyecto
tipo_presupuesto	Tipo de presupuesto (Presupuesto Sercotec o Extrapresupuestario)
agente.operador.id	ID del agente operador
region.agente.operador.id	ID de la región del Agente Operador
region.agente.operador.desc	Nombre de la región del Agente Operador
postulante_beneficiario.id_beneficiario	ID del postulante/beneficiario
tipo_postulante_beneficiario.desc	Tipo de postulante/beneficiario (Empresa jurídica, Organización gremial, Persona Natural, Rut Extendido)
estado_actual_beneficiario.id	ID del estado del postulante/beneficiario
estado_actual_beneficiario.desc	Descripción del estado del postulante/beneficiario
provincia_postulante_beneficiario.id	ID de la provincia del postulante/beneficiario
provincia_postulante_beneficiario.desc	Descripción de la provincia del postulante/beneficiario
comuna_postulante_beneficiario.id	ID de la comuna del postulante/beneficiario
comuna_postulante_beneficiario.desc	Descripción de la comuna del postulante/beneficiario
clasificacion_ruralidad_comuna	Categoría de clasificación según ruralidad de la comuna, según información del CENSO 2017
sexo_contacto.sexo	Sexo del contacto
nivel_educacional.nivel_educacional	Nivel educacional del contacto
edad_contacto	Edad del contacto
nacionalidad_contacto.desc	Nacionalidad del contacto
actividad_económica_postulación.desc	Actividad económica de la postulación
admisibilidad	Estado de admisibilidad en 2 categorías (Admisible y No Admisible)
puntaje	Puntaje obtenido en la prueba de admisibilidad

puntajecorte	Puntaje de corte contra el que se evalúa la nota del postulante/beneficiario
cod_interno	Clasificador de las preguntas del Canvas
pregunta_id	ID de la pregunta del Canvas
texto.pregunta	Texto de la pregunta
respuesta_id	ID de la respuesta a la pregunta del Canvas
texto.respuesta	Texto de la respuesta
nota	Nota sin ponderar
ponderacion	Ponderación de la pregunta del Canvas
nota_ponderada	Nota ponderada
nota_total_canvas	Nota final al sumar la nota ponderada asociada a cada pregunta del Canvas

Fuente: Entregado por Equipo de Sercotec.

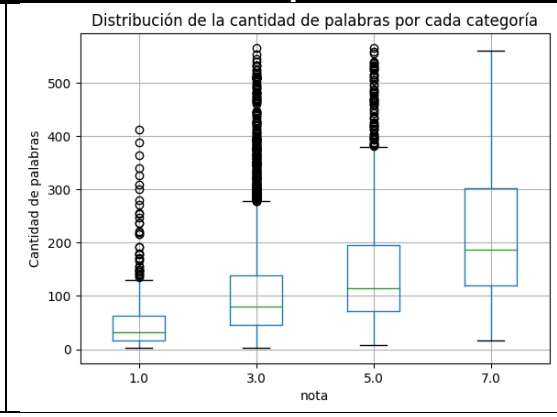
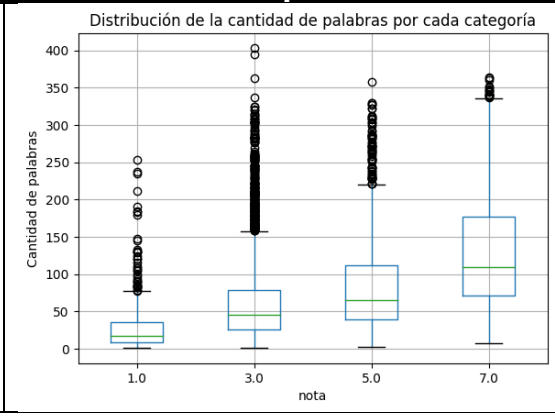
Anexo C: Tablas de distribución de palabras de notas por sección

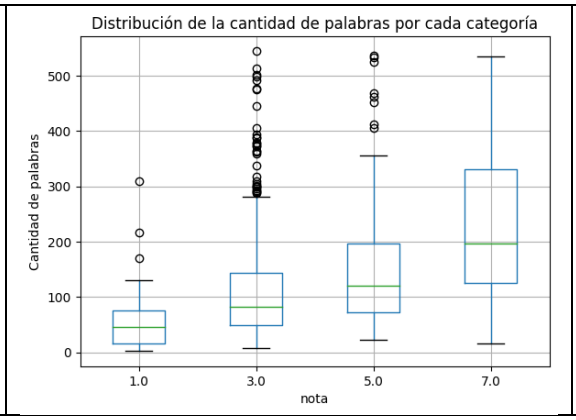
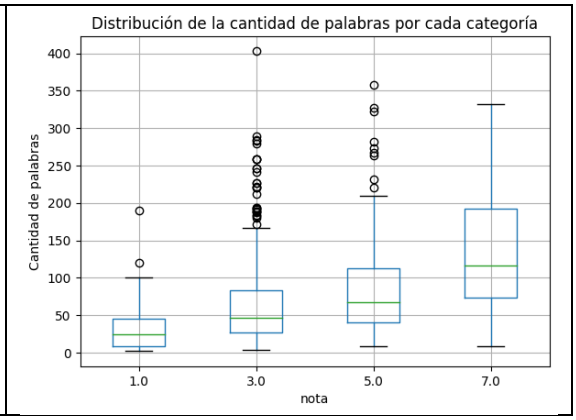
Tabla 27: Distribución de palabras por notas, sección Clientes.

	Con stopwords	Sin stopwords
100% de datos		
10% de datos		

Fuente: Elaboración propia.

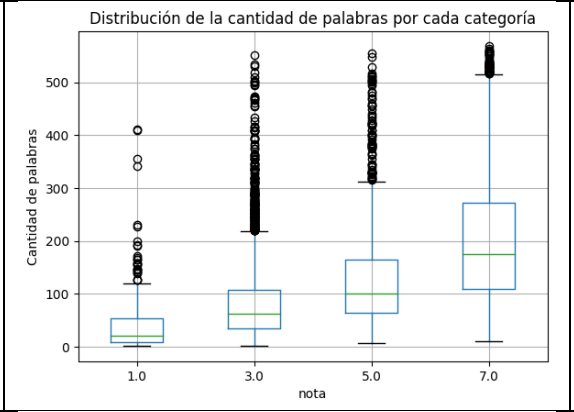
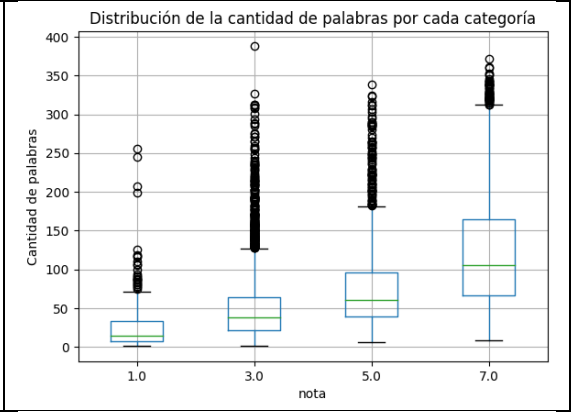
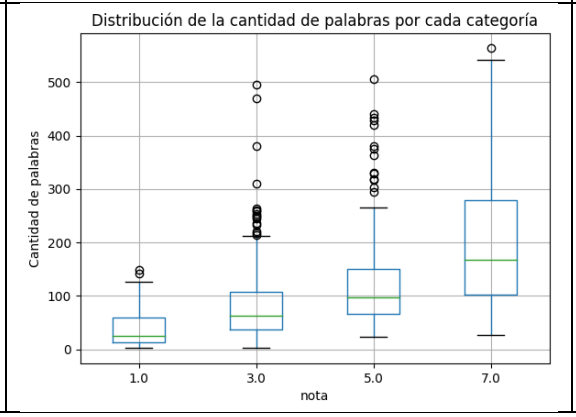
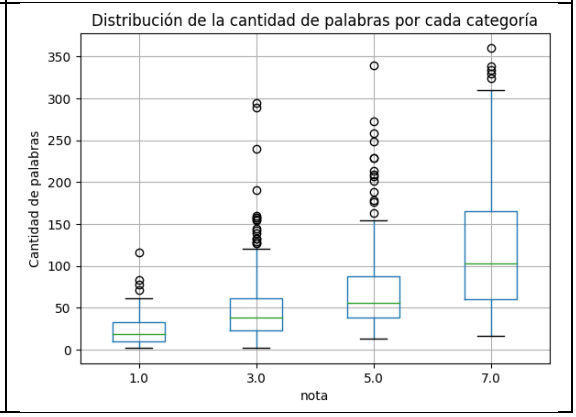
Tabla 28: Distribución de palabras por notas, sección Propuesta de Valor.

	Con stopwords	Sin stopwords
100% de datos		

10% de datos		

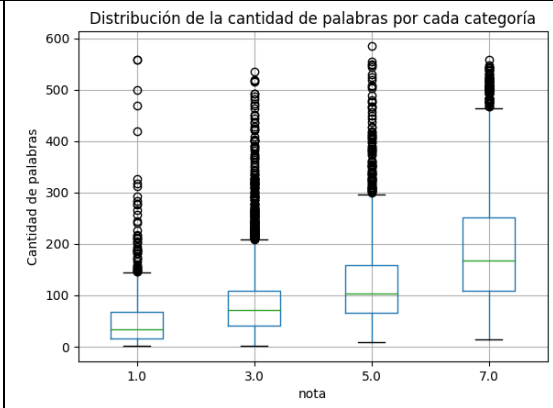
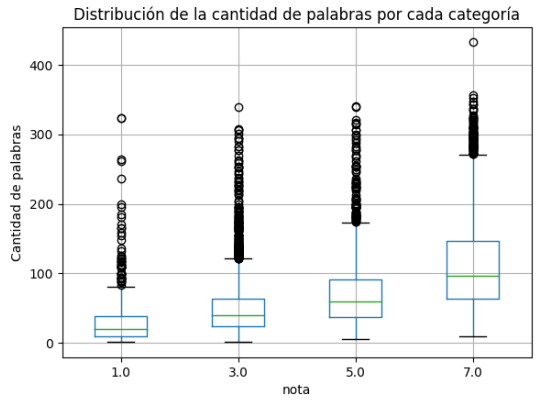
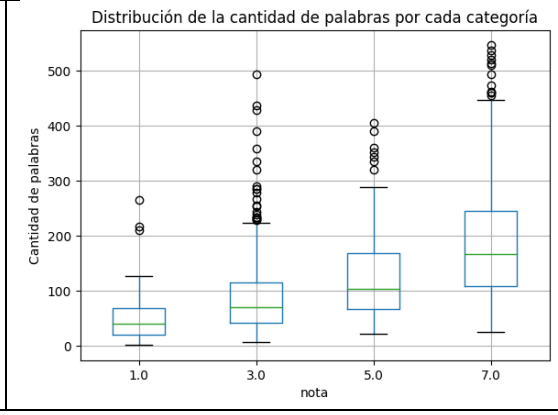
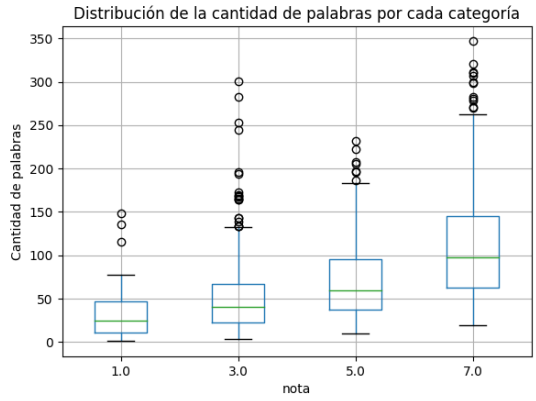
Fuente: Elaboración propia.

Tabla 29: Distribución de palabras por notas, sección Canales.

	Con stopwords	**Sin stopwords**
100% de datos		
10% de datos		

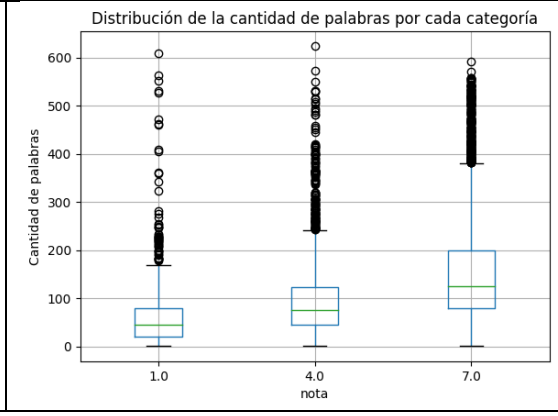
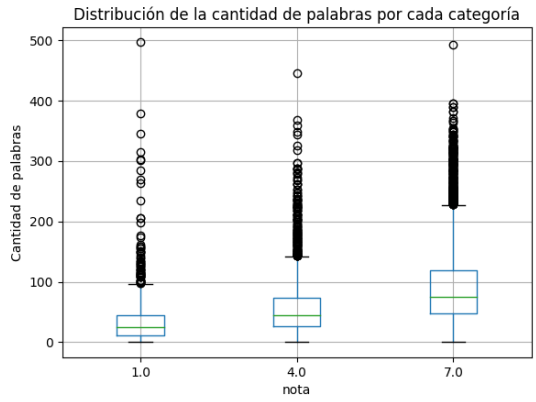
Fuente: Elaboración propia.

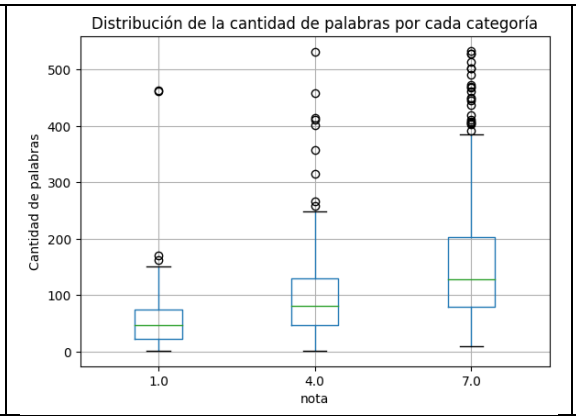
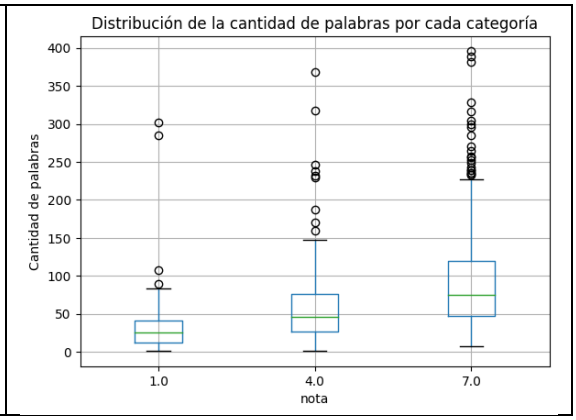
Tabla 30: Distribución de palabras por notas, sección Relación.

	Con stopwords	Sin stopwords
100% de datos		
10% de datos		

Fuente: Elaboración propia.

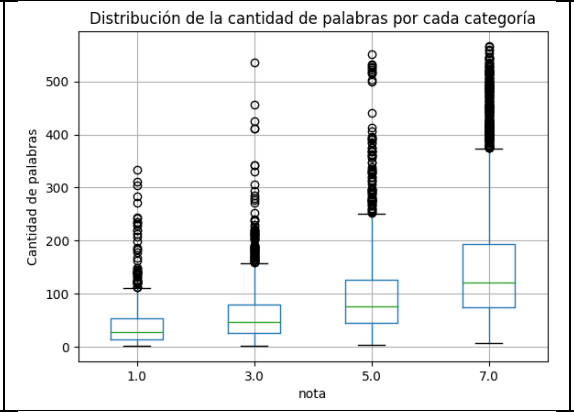
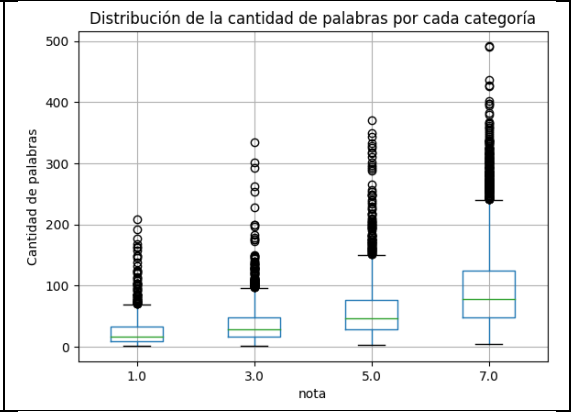
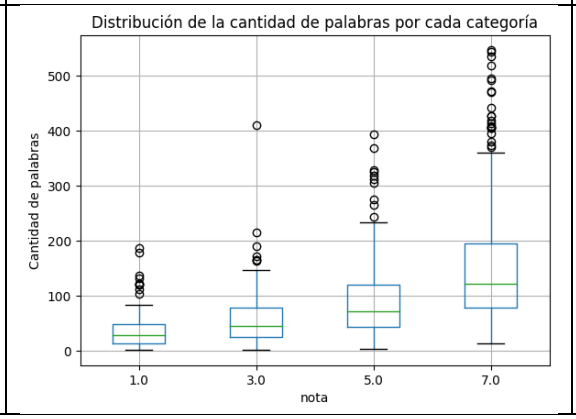
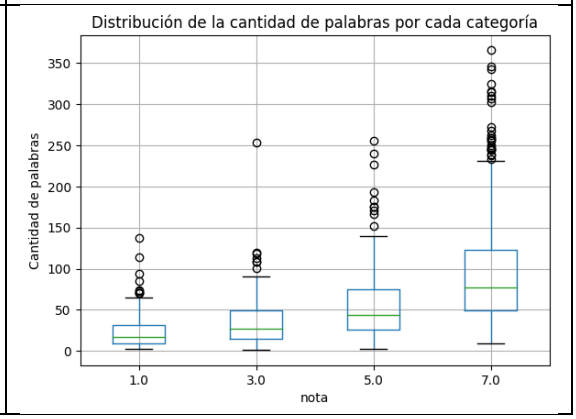
Tabla 31: Distribución de palabras por notas, sección Ingresos.

	Con stopwords	Sin stopwords
100% de datos		

10% de datos		

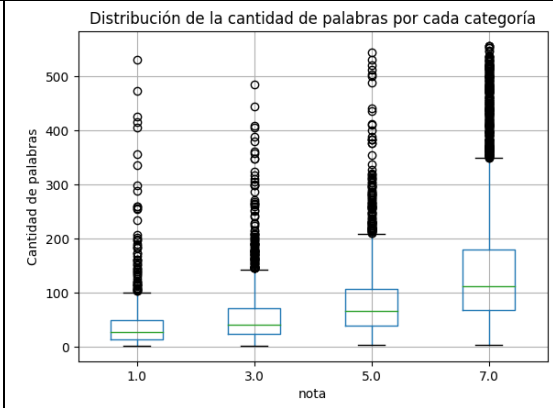
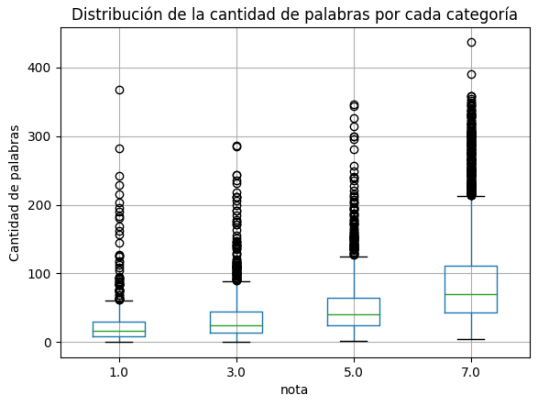
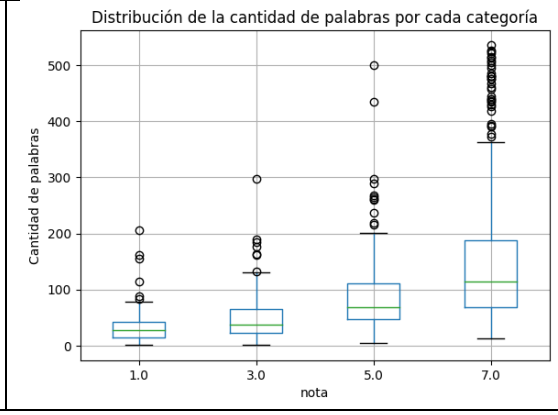
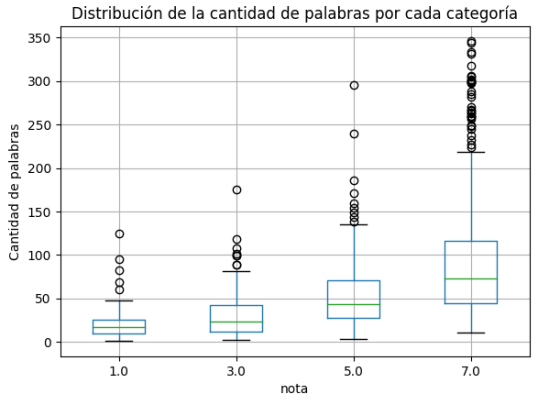
Fuente: Elaboración propia.

Tabla 32: Distribución de palabras por notas, sección Recursos.

	Con stopwords	**Sin stopwords**
100% de datos		
10% de datos		

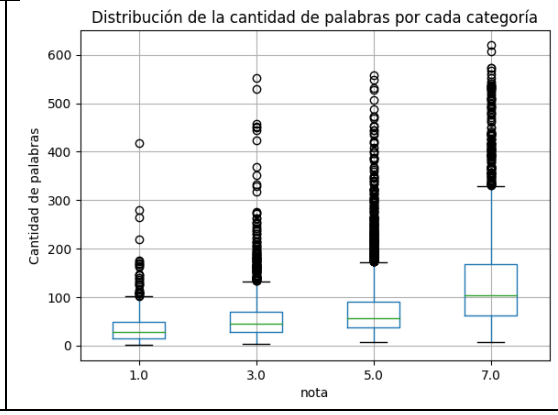
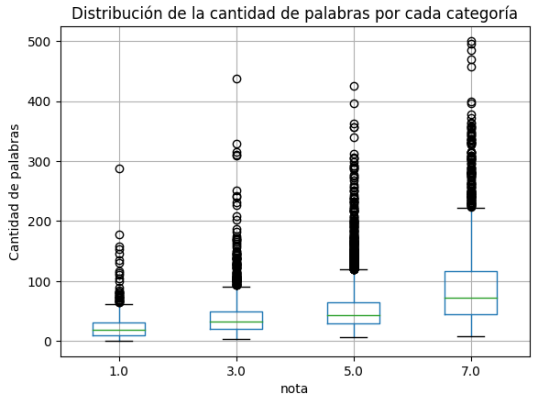
Fuente: Elaboración propia.

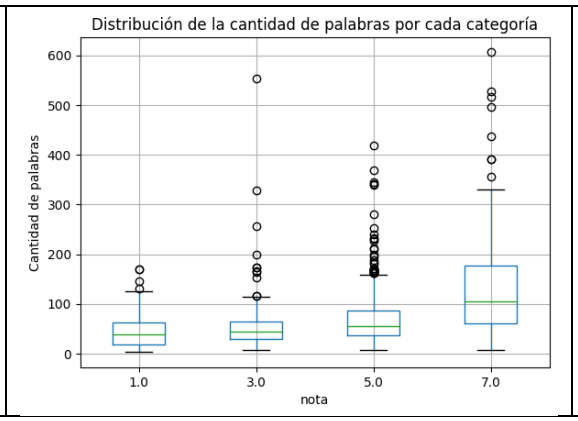
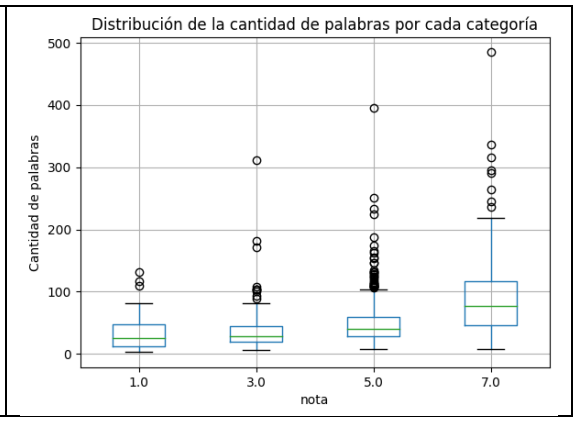
Tabla 33: Distribución de palabras por notas, sección Actividades.

	Con stopwords	Sin stopwords
100% de datos		
10% de datos		

Fuente: Elaboración propia.

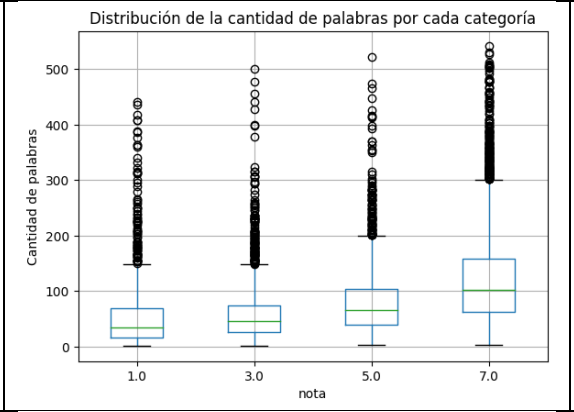
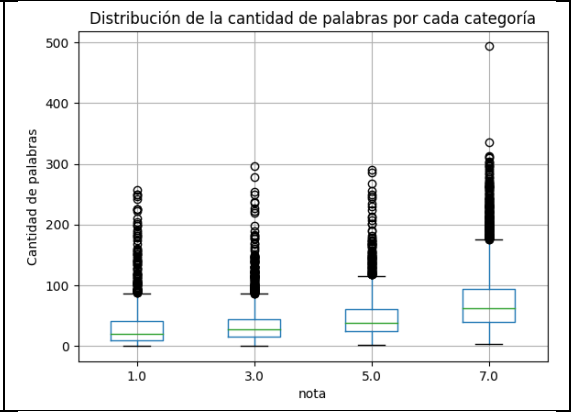
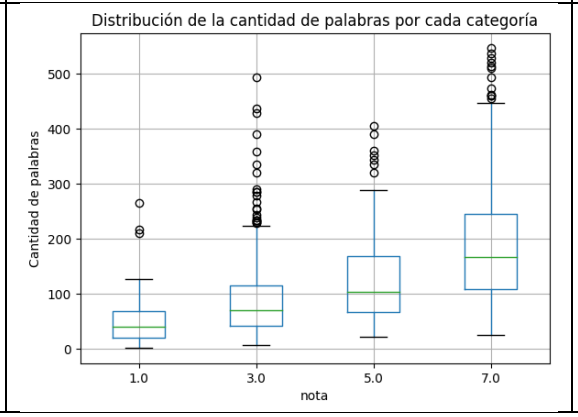
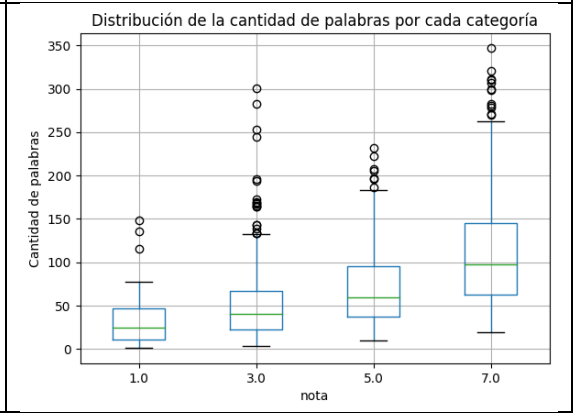
Tabla 34: Distribución de palabras por notas, sección Costos.

	Con stopwords	Sin stopwords
100% de datos		

10% de datos	 

Fuente: Elaboración propia.

Tabla 35: Distribución de palabras por notas, sección Alianzas.

	Con stopwords	Sin stopwords
100% de datos		
10% de datos		

Fuente: Elaboración propia.

Anexo D: Cantidad de datos de evaluación

Tabla 36: Cantidad de datos de evaluación.

Sección del Canvas	Con nota / Sin nota	Cantidad de datos
Clientes	Archivo con nota (10%)	850
	Archivo sin nota (1%)	335
Propuesta de Valor	Archivo con nota (10%)	841
	Archivo sin nota (1%)	335
Canales	Archivo con nota (10%)	848
	Archivo sin nota (1%)	335
Relación con el cliente	Archivo con nota (10%)	857
	Archivo sin nota (1%)	334
Ingresos	Archivo con nota (10%)	853
	Archivo sin nota (1%)	334
Recursos clave	Archivo con nota (10%)	860
	Archivo sin nota (1%)	334
Actividades clave	Archivo con nota (10%)	859
	Archivo sin nota (1%)	334
Costos	Archivo con nota (10%)	868
	Archivo sin nota (1%)	334
Alianzas clave	Archivo con nota (10%)	855
	Archivo sin nota (1%)	335

Fuente: Elaboración propia.