Universidad del Desarrollo
Facultad de Ingeniería

Deep Learning para caracterizar la innovación en Chile

Uso de redes neuronales para clasificación de proyectos de innovación de CORFO

POR: FABIÁN ALEXIS ORTEGA VEGA

Capstone project presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA:

Dra. Daniela Opitz y Sr. Tomás Fontecilla

Noviembre 2022

SANTIAGO

Dedico este trabajo a mi madre y a mis abuelos, quienes me han apoyado durante toda mi vida y que sin su cariño nada sería posible.

AGRADECIMIENTO

A Javiera Alvear, por su invaluable apoyo y tiempo dedicado a escucharme, a Rodrigo González por sus consejos y discusiones. A mi jefa Brenda Rain, por confiar, creer en mí y compartir sus conocimientos para que pudiese llevar a cabo este proyecto.
Finalmente, a mi profesor guía, Tomás Fontecilla, agradecer sus consejos y sobre todo, el tiempo dedicado.

TABLA DE CONTENIDO

1. Resumen

Varios autores mencionan la innovación tecnológica como uno de los principales motores del crecimiento económico de largo plazo, por este motivo es importante contar con información acerca del estado de la innovación en Chile. En este punto, CORFO tiene un rol relevante al ser parte fundamental del sistema de ciencia y tecnología e innovación y ser la agencia que concentra los fondos públicos que financian iniciativas de este tipo. El problema recae en que la información de los proyectos que postulan a los concursos de financiamiento público vienen en forma no estructurada a través de un formulario de postulación y en formato de texto libre, lo que hace costoso poder obtener información o realizar análisis debido al gran volumen de datos. En este trabajo se plantea hacer uso de técnicas de procesamiento de lenguaje natural (NLP) y *deep learning* para abordar de manera automática esta tarea y tener información para caracterizar los proyectos de innovación que postulan a CORFO. Para esto, se busca clasificar proyectos por su mercado de llegada de la innovación, por su tipo (producto, proceso o servicio) y por último, identificar si los proyectos contienen componentes de sostenibilidad. Para esto se construyó dos tipos de modelos, el primero utiliza Word2Vec más una red neuronal convolucional (CNN) que cumple un rol de línea base. Y un segundo modelo, que utiliza técnicas del estado del arte en NLP y está basado en una red neuronal *Transformers* del tipo representación de codificador bidireccional de *Transformers* (BERT) pre entrenada en español (BETO). Este modelo supera en desempeño a Word2Vec + CNN, alcanzando un *accuracy* del 70% para la clasificación del mercado de llegada, 73% para el tipo de innovación, y un *accuracy* de 80% para clasificar proyectos por componente sostenible. Finalmente, se utilizan estos modelos para caracterizar los proyectos postulados entre 2019 y primer semestre de 2022.

1. Introducción

La relación entre la innovación y el crecimiento económico es un tópico ampliamente estudiado. Desde los tiempos de Schumpeter y su teoría de la destrucción creativa (Schumpeter, 1942) es que se ha planteado que la innovación industrial junto con el avance tecnológico son uno de los principales motores del crecimiento económico de largo plazo. En este contexto, Chile posee una de las tasas de innovación más bajas a nivel OCDE (OCDE, 2018), la cual viene a la baja desde el año 2013, y en donde sólo el 21% de las empresas chilenas realiza algún tipo de actividad innovativa (ENI, 2020).

A nivel institucional, en Chile existe la Corporación de Fomento de la Producción (CORFO) la cual tiene por objetivo promover la competitividad y diversificación productiva del país, a través del fomento del emprendimiento y la innovación.

Es en esta agencia en donde se concentra la mayoría de la demanda por financiamiento público para proyectos de innovación, la cual la convierte en la principal fuente de información del sistema nacional de innovación y desde donde se pueden obtener indicios del estado de la innovación en Chile.

Esta información es obtenida a través de formularios de postulación a fondos de financiamiento, y son una fuente importante para conocer el estado en que están las empresas que realizan proyectos de innovación. A pesar de que estos formularios están digitalizados y se realizan a través de una plataforma que permite su salida a una base de datos estructurada, la mayoría de la información relevante se encuentra en formato de texto libre, lo que hace difícil y muy costoso en términos de recursos humanos poder analizar y obtener conocimiento de las innovaciones postuladas.

Este estudio busca caracterizar la demanda por innovación utilizando los datos de texto abierto de los formularios de postulación a los distintos programas de innovación de CORFO, haciendo uso de técnicas de procesamiento de lenguaje natural (NLP) y de *deep learning* para generar modelos de clasificación que permitan identificar el mercado de llegada de las innovaciones, el tipo de innovación, y si es que la innovación tiene componentes de sostenibilidad.

El desarrollo de este trabajo es un avance para conocer y saber más sobre el estado del ecosistema de innovación en Chile, poder obtener información relevante de los distintos sectores económicos que están demandando innovación, identificar de manera temprana avances en la novedad de las innovaciones, identificar cambios en tendencias tecnológicas, tener acercamientos al nivel de complejización de la matriz productiva por sector económico e incluso generar herramientas para poder conectar demanda de innovación con oferta de innovación.

1.1 CORFO dentro del Sistema de CTCI

La Corporación de fomento de la producción tiene un rol fundamental dentro del sistema nacional de ciencia, tecnología, conocimiento e innovación (CTCI) . La política nacional de CTCI incluye dentro de su componente de innovación, la innovación de base científico-tecnológica, la cual está definida como "el desarrollo experimental y actividades de carácter científico-tecnológicas que pueden llevar a la generación de productos, procesos o servicios nuevos o sustancialmente mejorados, en etapas previas a su comercialización, la cual está basada en conocimiento científico y que puede requerir tiempos de investigación y desarrollo más largos previos a su llegada al mercado. Este componente está articulado por el Ministerio de de Ciencia, tecnología, conocimiento e Innovación" [1], en donde CORFO tiene el mandato de ejecutar programas que son financiados a través del fondo de innovación para la competitividad de ciencia y tecnología (FICYT), en donde se encuentran instrumentos como Crea y Valida I+D+i, Crea y Valida I+D+i Colaborativo, o Alta tecnología.

También, dentro del componente de innovación, está el segundo relacionado a la innovación empresarial, la cual está definida en la política nacional de CTCI como "Aquella que nace con el propósito de crear valor a través de la transformación de ideas y/o conocimientos en nuevos recursos o bienes mejorados, servicios y/o procesos que difieran significativamente de los previamente existentes en la empresa y que hayan sido introducidos en el mercado, y que tiene como objetivo aumentar la competitividad de las empresas, a través de la generación de valor agregado".

[1] Política Nacional de Ciencia, Tecnología, Conocimiento e Innovación, 2020.

2. Trabajo Relacionado

El uso de metodologías de *deep learning* y NLP para tareas de clasificación de texto es un tema ampliamente abordado en la literatura (Minaee et al., 2021). Investigadores han hecho uso de estos métodos para realizar tareas como clasificación de artículos de prensa (Lavanya et al., 2021), análisis de sentimiento (Alshari et al., 2017), detección y filtro de spam (Hu et al., 2017), entre otros. En trabajos similares (Tagarev et al., 2019) hacen uso técnicas de *deep learning* como Word2vec y GloVe para clasificar empresas en función de la descripción pública de cada compañía. Posterior a esto, (Slavov et al., 2019) utiliza modelos pre-entrenados como BERT, XLNet y ULMfit para abordar el mismo problema de clasificación de empresas, en donde encuentran que el modelo pre-entrenado BERT es el que obtiene los mejores resultados en esta tarea. Cabe mencionar que los modelos BERT están pre entrenados con corpus compuestos principalmente por texto en inglés, y que a pesar de haber algunos modelos BERT multilenguaje, hasta la fecha sólo existe un modelo pre entrenado en español, BETO (Cañete et al., 2020).

3. Hipótesis y Objetivos

3.1 Hipótesis

Es factible clasificar los proyectos postulados a la gerencia de innovación de CORFO utilizando *deep learning*, haciendo uso de los datos de texto libre contenidos en los formularios de postulación.

3.2 Objetivo general

El objetivo de este trabajo es entrenar un modelo de clasificación de texto que, a partir de la información contenida en los formularios de postulación a los distintos programas de innovación de CORFO, permita clasificar los proyectos según mercado al que está dirigida la innovación, el tipo de innovación, y si la innovación posee características sostenibles, con el fin de caracterizar el estado de la innovación en Chile.

3.3 Objetivos específicos

Para el desarrollo de este trabajo se definieron los siguientes objetivos específicos:

1. Entrenar un modelo de clasificación de texto usando técnicas de NLP y de aprendizaje profundo que permita identificar mercados de destino de proyectos de innovación, el tipo de innovación (Producto, proceso o servicio) y si el proyecto postulado posee componentes de triple impacto (económico, medioambiental y/o social).
2. Que los modelos realizados tengan un accuracy superior al 60%.
3. Caracterizar los proyectos postulados entre 2019 y 2022 haciendo uso de los modelos de clasificación.

4. Datos y Metodología

4.1. Datos

El set de datos utilizados corresponde a la base de proyectos adjudicados de la gerencia de Innovación de Corfo, compuesta por los proyectos ganadores de las distintas convocatorias entre 2010 y 2022, y las cuales cuentan con un etiquetado de las variables de interés realizada por juicio experto.

El set de datos posee 8373 observaciones en donde cada observación corresponde a un proyecto, y se tienen campos como el título del proyecto, el objetivo general y específico del proyecto, un resumen que da una descripción general de lo que busca generar el proyecto, además de los campos de interés como el mercado objetivo, que indica el mercado de llegada de la innovación, tipo de innovación, que indica si el proyecto corresponde a un producto, proceso o servicio, el campo sostenible, que indica si el proyecto cuenta con componentes de triple impacto (económico, medioambiental y/o social). Además de un set de datos con cerca de 12.000 proyectos postulados y no adjudicados de los cuales no se cuenta con los campos a clasificar, y que es a los cuales se le busca caracterizar.

La descripción de los campos a clasificar es la siguiente:

- **Mercado objetivo**: Corresponde a los mercados en los cuales se puede comercializar o dar utilidad a la innovación que se está realizando. Está construida en base al CIIU4 y simplificada para un mejor entendimiento.
  Cuenta con 24 categorías, las cuales se pueden revisar en el anexo 1.

- **Tipo de innovación**: De acuerdo con el Manual de Oslo (OECD, 2018), la innovación de producto "corresponde a la introducción de un bien o servicio nuevo o mejorado, en cuanto a sus características, o al uso al que se destina".
  De esta forma, la innovación se puede clasificar en **innovación en productos**, en donde la innovación corresponde a un bien tangible, **innovación en servicios**, en donde el producto es un intangible, y para la **innovación de procesos**, el Manual de Oslo (OECD, 2018) la define como la introducción de un nuevo o significativamente mejorado, proceso de producción o de distribución.
  Este campo posee 3 categorías, Innovación de producto, de proceso y de servicio.

- **Clasificación de sostenibilidad:** corresponde a aquellos proyectos que generan beneficios sociales y/o medioambientales, adicionales al económico.
  Este campo es una clasificación binaria en función de si el proyecto cumple o no con características de sostenibilidad.

4.2. Metodología

La metodología propuesta es utilizar representaciones de palabras en vectores numéricos (*word embedding*) usando el resumen de los proyectos para entrenar modelos de clasificación para cada una de los campos a clasificar.

Como modelo de línea base se utilizó Word2Vec más una red neuronal convolucional (CNN) , y luego se generó un modelo de clasificación usando una red neuronal del tipo Representación de Codificador Bidireccional de *Transformers* (BERT) en su versión en español (BETO).

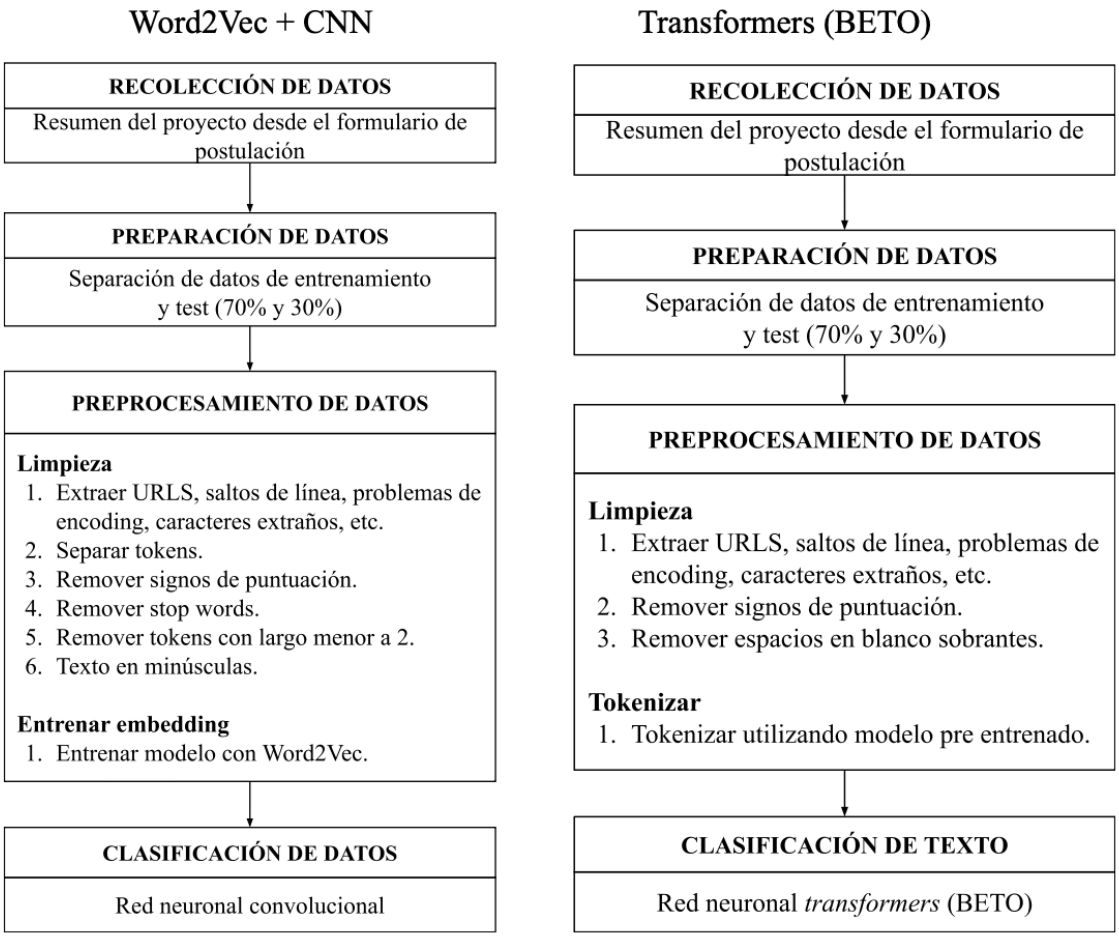


Figura 1: Metodología. Fuente: Elaboración propia.

4.2.1 Word2Vec + CNN

El modelo Word2Vec es un tipo de red neuronal que hace uso de arquitecturas CBOW (Bolsa de palabras continuas) y Skip-gram para calcular las representaciones vectoriales de las palabras (Mikolov, 2013). Los vectores de palabras se sitúan de manera que las palabras que tienen contextos similares en el corpus se sitúan muy cerca unas de otras en el espacio vectorial. La idea subyacente es que palabras que se usan en contextos similares, tienen significados similares.

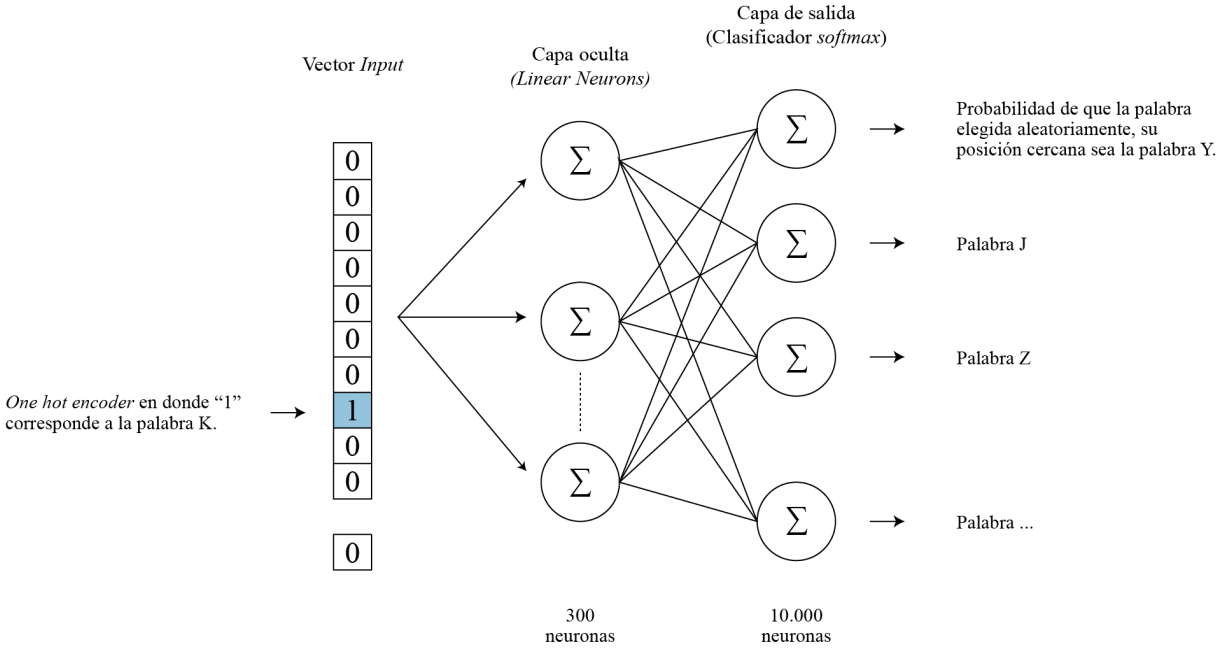


Figura 2: Arquitectura Word2Vec. Fuente: Elaboración propia en base a (Mikolov, 2013)

Para construir el modelo, primero se generó la representación vectorial del texto usando Word2Vec como input para una red neuronal convolucional. El vocabulario utilizado consiste en 79.637 palabras. La red neuronal construida está compuesta por 5 capas, se utilizó una tasa de aprendizaje de 0.001, un tamaño del *batch* de [16, 32, 64] y un optimizador Adam, siguiendo la arquitectura planteada por (Mikolov, 2013) y (Hughes et al., 2017).

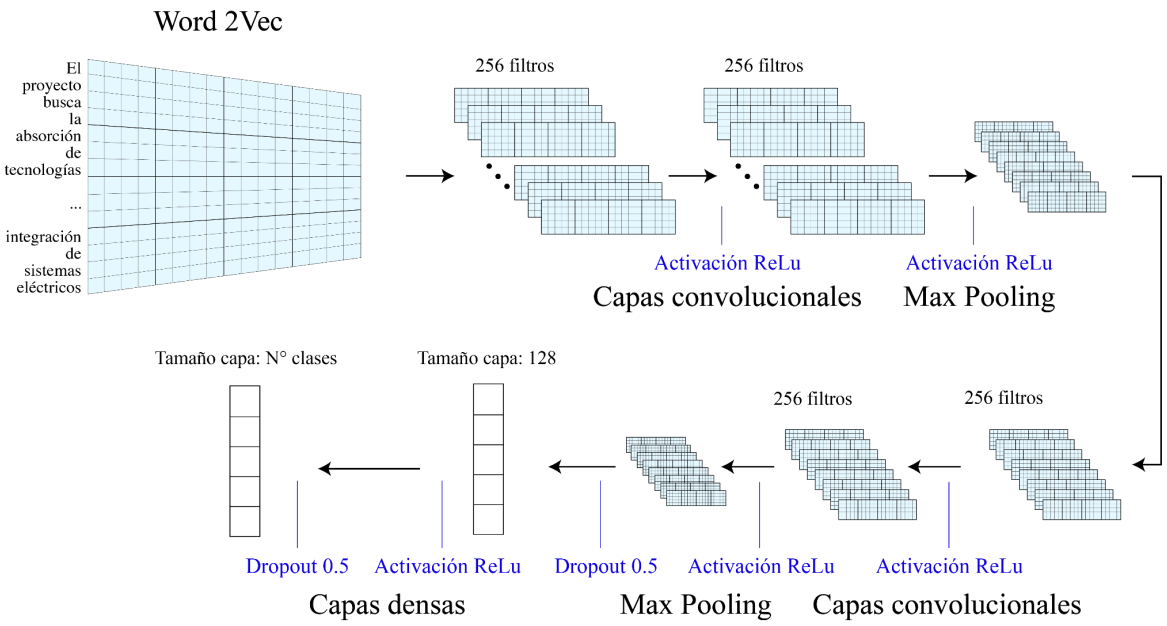


Figura 3: Arquitectura modelo Word2Vec + CNN. Fuente: Elaboración propia en base a (Hughes et al., 2017)

La arquitectura de la red es la misma para las 3 variables de interés, y solo cambia la última capa para adaptar la dimensión en función de las categorías a clasificar según cada campo (una dimensión de 24 para mercado objetivo, 3 para tipo de innovación, 2 para sostenible).

4.2.2 Transformers (BETO)

Un modelo tipo *Transformer* (Vaswani et al., 2017) es una red neuronal que aprende contexto mediante el seguimiento de relaciones en datos secuenciales. Busca resolver problemas del tipo secuencia a secuencia (*seq 2 seq*) pero tiene una variedad de aplicaciones dentro del NLP.

Dentro de los modelos *Transformers*, están los modelos BERT, que son una representación del *encoder* bidireccional de *Transformers* pre entrenados mediante una combinación de modelado lingüístico enmascarado (*Masked language modeling*) y de predicción de la siguiente frase. El vocabulario está construido en base a un corpus que comprende el Toronto Book Corpus y Wikipedia en inglés. Para texto en español existe BETO (Cañete et al., 2020), que es un modelo con arquitectura tipo BERT entrenado en un corpus completamente en español usando fuentes como Wikipedia y *OPUS project*.

Para la implementación del modelo BETO, se siguió lo planteado en (Cañete et al., 2020) y por (Vaswani et al., 2017), con una estructura BERT estándar con 12 capas de autoatención, 16 cabezas de atención (*Multi-Head Attention*) y con una dimensión del *embedding* oculto de 768. Se utilizó una tasa de aprendizaje de [3e-4, 1e-4, 5e-5, 3e-5], un tamaño del *batch* de [8, 16, 32, 64], [1, 2, 3, 4, 5] épocas, un optimizador *AdamW* (Loshchilov & Hutter, 2017) y un largo de tokens de [128, 224, 324, 448] para el *embedding* de entrada.

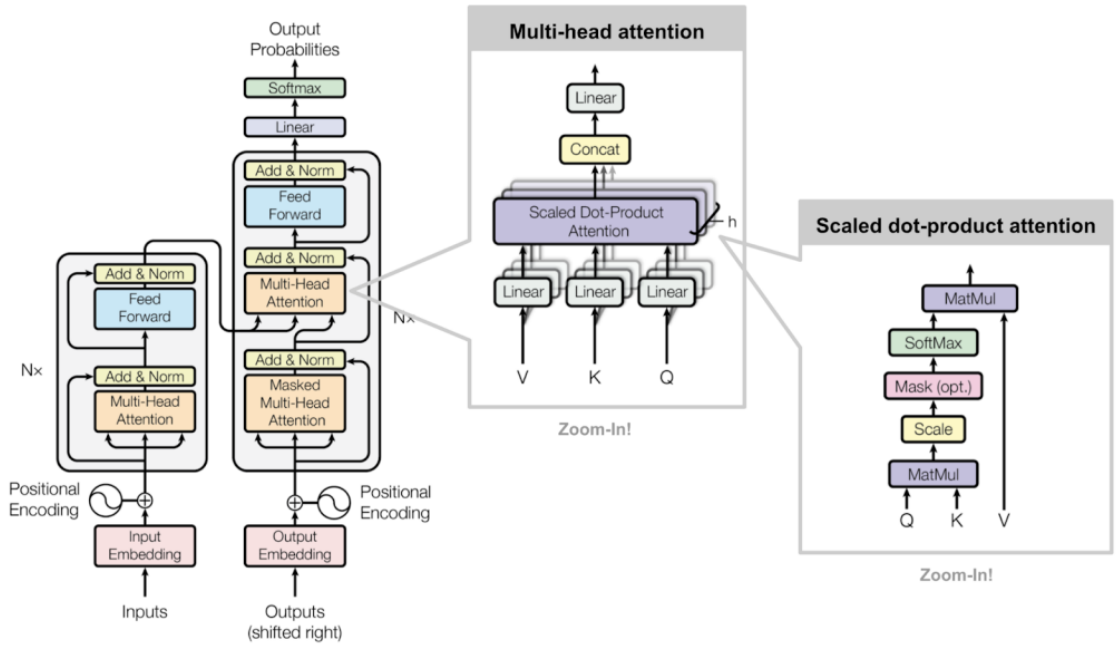


Fig. 4: Arquitectura modelo Transformers. Fuente: Attention is all you need, 2017.

Al igual que en el modelo de línea base, la arquitectura es similar para los 3 modelos. Los cambios principales se dan en el modelo de clasificación de mercado objetivo, esto debido a que la variable de interés tiene clases desbalanceadas por lo que se le añade sesgo a la función de pérdida para intentar obtener una mejor generalización, además de los cambios en la dimensionalidad de la salida de las probabilidades según variable a clasificar.

Análisis exploratorio de datos

Del análisis exploratorio de datos se seleccionaron las variables a utilizar como entrada del modelo. Se realizó un análisis de la distribución de frecuencias de palabras para los campos objetivo, objetivos específicos y resumen, encontrando que resumen es el campo que mayor información reúne en términos de vocabulario, seguido por objetivo, y por último, el campo

objetivos específicos. Sumado a lo anterior, se realizó un análisis de la frecuencia de tokens junto con nubes de palabras para identificar los tokens con mayor frecuencia. Y por último se analizó la distribución de las variables a clasificar, encontrando que el campo "mercado objetivo" está desbalanceado, en donde categorías como "Finanzas", "Sector Público" y "Asociaciones y organizaciones no empresariales ni gubernamentales" poseen menos de 100 observaciones. Ver anexo 2 para más detalle.

5. Resultados

Para comparar el rendimiento de los modelos, estos se entrenaron en la misma división de datos de entrenamiento y prueba para obtener resultados comparables.

5.1 Word2Vec + CNN

Mercado Objetivo

Para el modelo de línea base de mercado objetivo (24 clases), el desempeño suele ser inversamente proporcional a la cantidad de observaciones por clase, con algunas excepciones, alcanzando un *accuracy* de 0.54 y un *f1-score* ponderado de 0.48.

El modelo es bueno clasificando los mercados Agrícola, Alimentos, Educación y servicios conexos, Minería, Pesca y Acuicultura; Salud y Farmacéutica, pero tiene un mal desempeño llegando a tener un *accuracy* de 0 para clases como Asociaciones y Orgs. no gubernamentales, Energía, Finanzas, Forestal, Recursos Hídricos, TIC's, Turismo y Vitivinícola. El detalle de las métricas por clase se puede ver en el anexo 3.

Accuracy: 0,54

	Precision	***Recall***	***F1-score***	**Número observaciones**
macro avg	0.26	0.29	0.26	2477
weighted avg	0.46	0.54	0.48	2477

Tabla 1: Métricas del modelo w2v + CNN para mercado objetivo

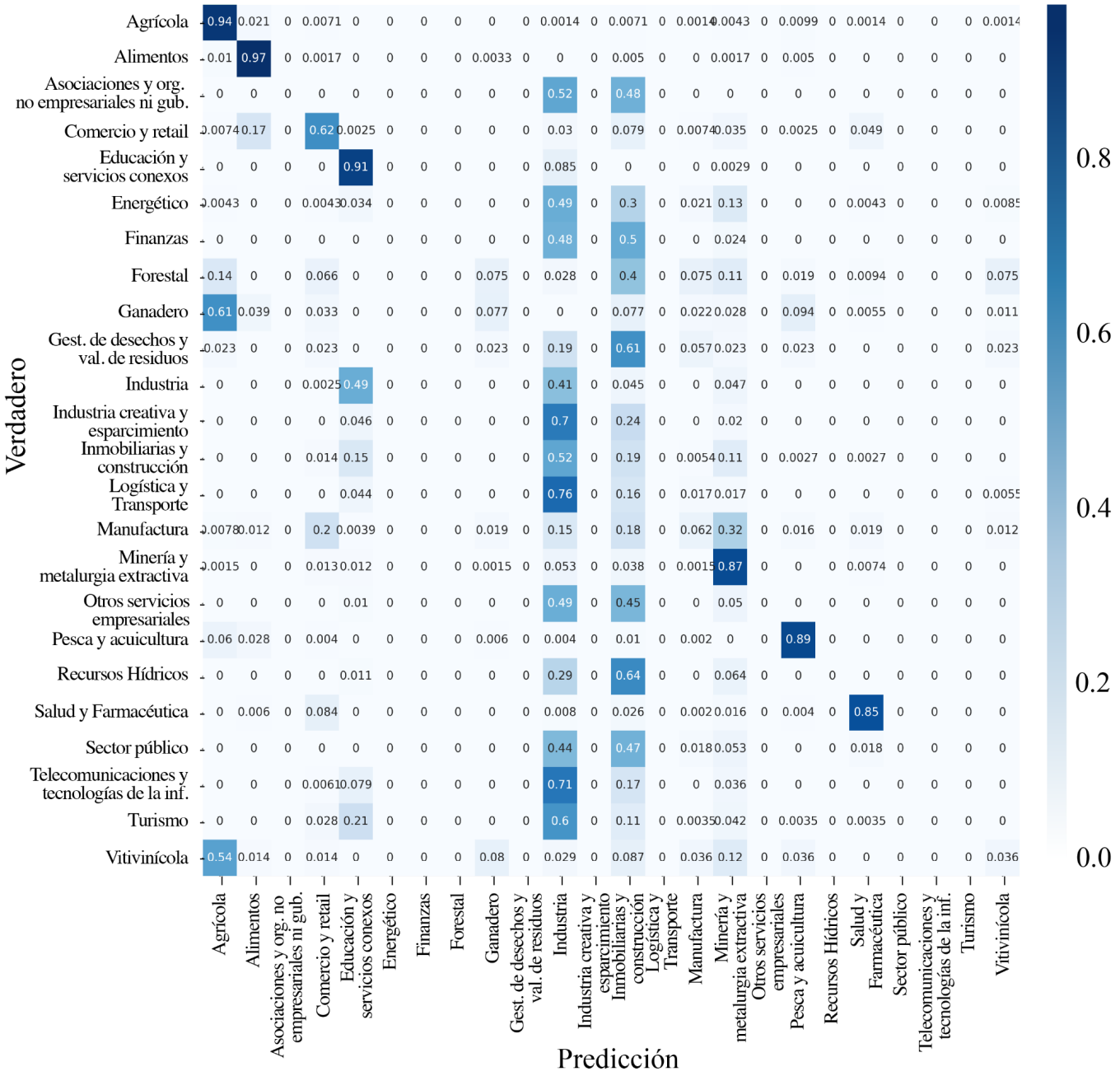


Fig. 5: Matriz de confusión modelo Word2vec para clasificación de mercado objetivo.

Tipo de innovación

El modelo Word2Vec para el tipo de innovación (3 clases) tuvo un *accuracy* de 0.62 y un *f1-score* ponderado de 0.59. El modelo de línea base tiene un buen desempeño para clasificar la clase Producto (clase 1) y la clase Servicio (2), pero no logra clasificar de manera correcta la clase Proceso, clasificándola de manera errónea como Producto.

Clase	Precision	Recall	F1-score	Support
0	0.46	0.17	0.25	620
1	0.63	0.77	0.69	1080
2	0.64	0.74	0.69	992
Accuracy			0.62	2692
Macro avg	0.58	0.56	0.54	2692
Weighted avg	0.59	0.62	0.59	2692

Tabla 2: Métricas por clase - Modelo Tipo de innovación.

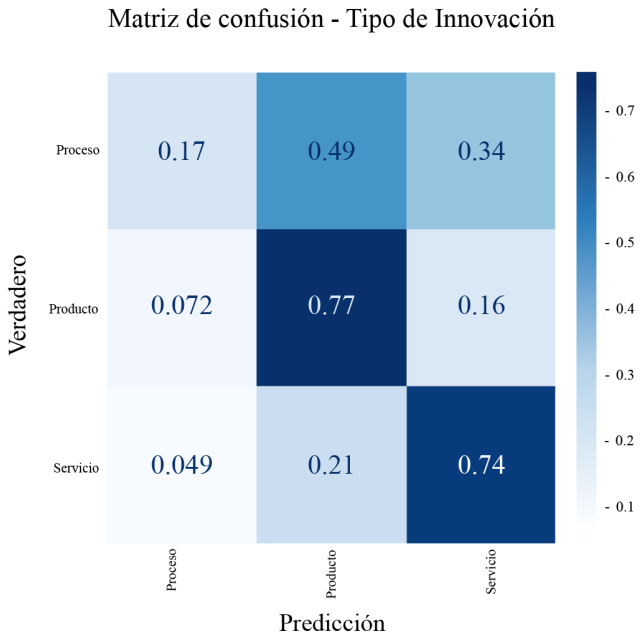


Fig. 6: Matriz de confusión para modelo w2v + CNN Tipo de innovación.

Clasificación de sostenibilidad

Para el modelo de clasificación de sostenibilidad, el modelo de línea base alcanza un accuracy de 0,593 y un f1-score 0,593. En donde la clase “Sí”, es clasificada de manera correcta el 64% de las veces, y la clase “No”, es clasificada de manera correcta el 56% de las veces.

Clase	Precision	Recall	F1-score	Support
0 (No)	0.68	0.56	0.61	1454
1 (Sí)	0.51	0.64	0.57	1057
accuracy			0.59	2511
macro avg	0.60	0.60	0.59	2511
weighted avg	0.61	0.59	0.59	2511

Tabla 3: Métricas por clase, modelo Word2Vec para clasificación de sostenibilidad

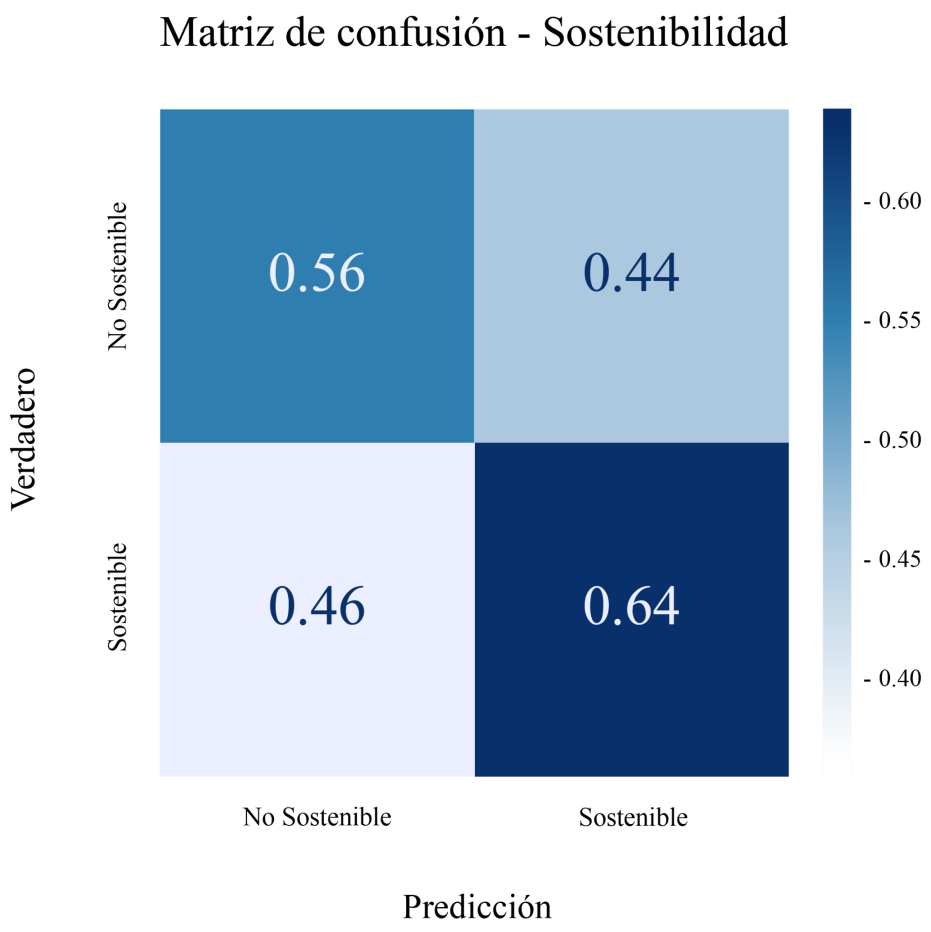


Fig. 7: Matriz de confusión modelo Word2vec para clasificación de sostenibilidad.

5.2 Transformers (BETO)

Mercado objetivo

Para el modelo BETO se realizó una serie de experimentos siguiendo lo indicado en la sección de metodología. Los resultados de estos se pueden ver en el anexo 4. En donde se obtuvo que el modelo con mejores resultados en el proceso de entrenamiento fue el de una tasa de aprendizaje de 0,00005, con 5 épocas, con un tamaño del *batch* de 16 y con 448 tokens de largo máximo por entrada.

Las métricas de este modelo son un 30% superior al modelo de línea base, alcanzando un *accuracy* de 0,704 y un *F1-score* ponderado de 0,702.

	precision	***recall***	***f1-score***	***support***
accuracy			0.70	2477
macro avg	0.62	0.65	0.62	2477
weighted avg	0.72	0.70	0.70	2477

Tabla 4: Métricas modelo BERT para mercado objetivo

Al analizar el desempeño por clase (ver anexo 5), vemos que el modelo BETO tiene un desempeño superior en 20 de 24 clases respecto del modelo de línea base, a excepción de las clases Alimentos (Clase 1), Comercio y Retail (Clase 3), y Salud y farmacéutica (Clase 19) en donde tiene un rendimiento inferior del 24%, 43% y 9% respectivamente.

De las 20 clases en donde el modelo BETO es superior al modelo de línea base, sólo en 6 clases tiene un desempeño inferior al 0,5 medido en relación al *f1-score*. Estás clases son Asociaciones y organizaciones. no empresariales ni gubernamentales (clase 2), Comercio y Retail (clase 3), Manufactura (clase 14), Otros servicios empresariales (clase 16), Sector público (Clase 20) y Tecnologías de la información y comunicación (clase 21). A excepción de la clase 3, 14 y 21, las clases 2, 16 y 20 poseen menos de 100 observaciones.

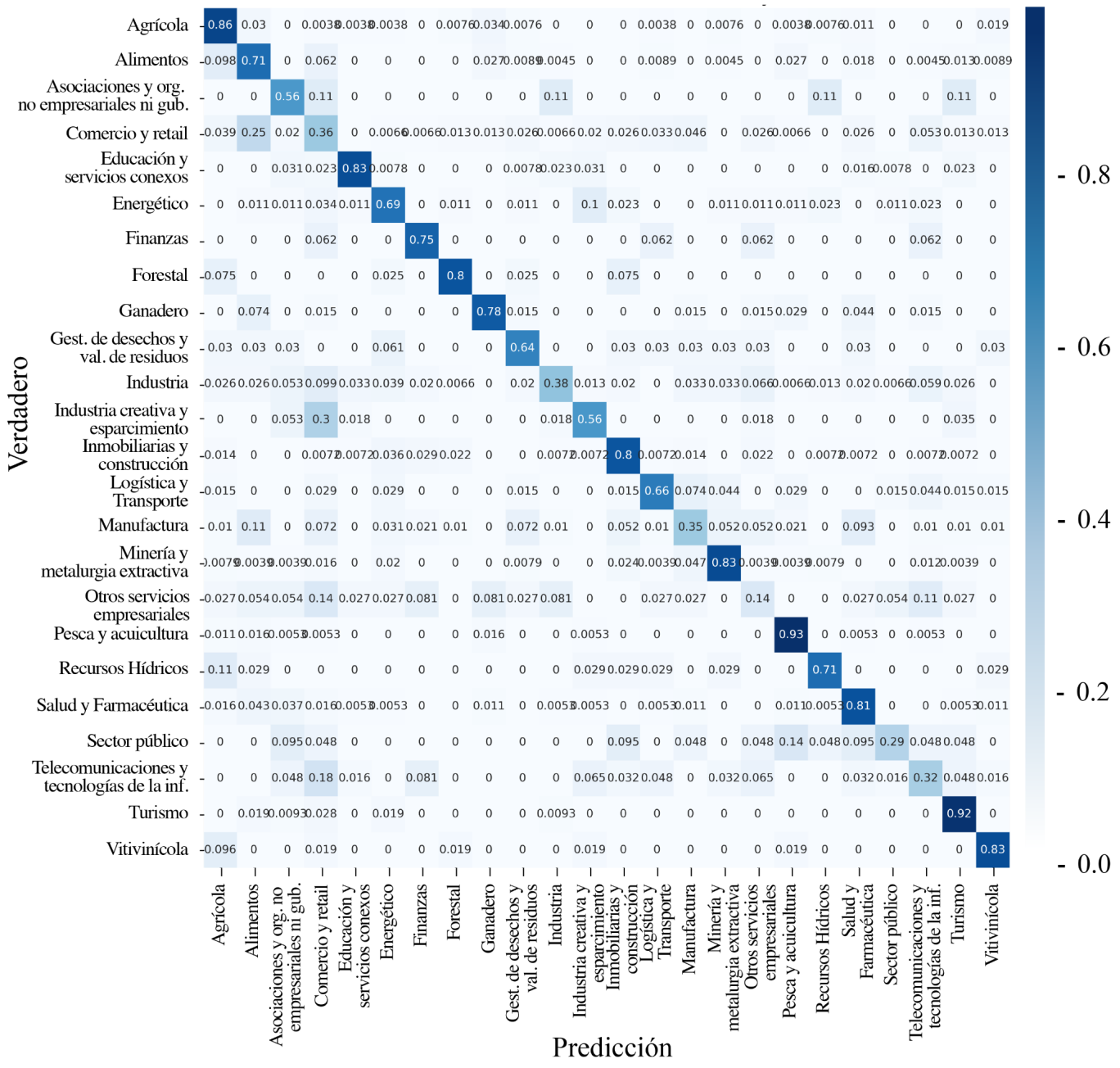


Fig. 8: Matriz de confusión modelo BETO para mercado objetivo.

También se observa que el modelo suele clasificar de manera incorrecta la clase Comercio y Retail (Clase 3), confundiéndola en un 25% de las veces con el mercado de Alimentos (Clase 1). Un problema similar sucede con el mercado Telecomunicaciones y tecnologías de la información, en donde el 18% de las veces la clasifica como Comercio y Retail. Al analizar más en detalle el vocabulario de estas clases, se identifica que los tokens en común entre Comercio y Retail respecto a Alimentos alcanzan el 62%, lo que explicaría en gran medida

porque el modelo no logra diferenciar de forma correcta entre ambas clases. En el caso de Telecomunicaciones y tecnologías de la información, el vocabulario común respecto Comercio y Retail es de un 78% lo que también podría explicar porque el modelo no es capaz de distinguir. Esto queda más claro al visualizar los *embeddings* de la última capa oculta del modelo. Para esto se construyó un clasificador BERT binario entre las clases "Comercio y Retail" y "Alimentos", un segundo clasificador entre "Telecomunicaciones y tecnologías de la información" y "Comercio y Retail", y por último, un clasificador de referencia, entre "Minería" y "Salud y farmacéutica" (en donde el modelo original tiene buen desempeño). A estos modelos se les aplicó una reducción de dimensionalidad de 768 a 2 dimensiones usando Análisis de Componentes Principales (PCA) (Bro & Smilde, 2014).

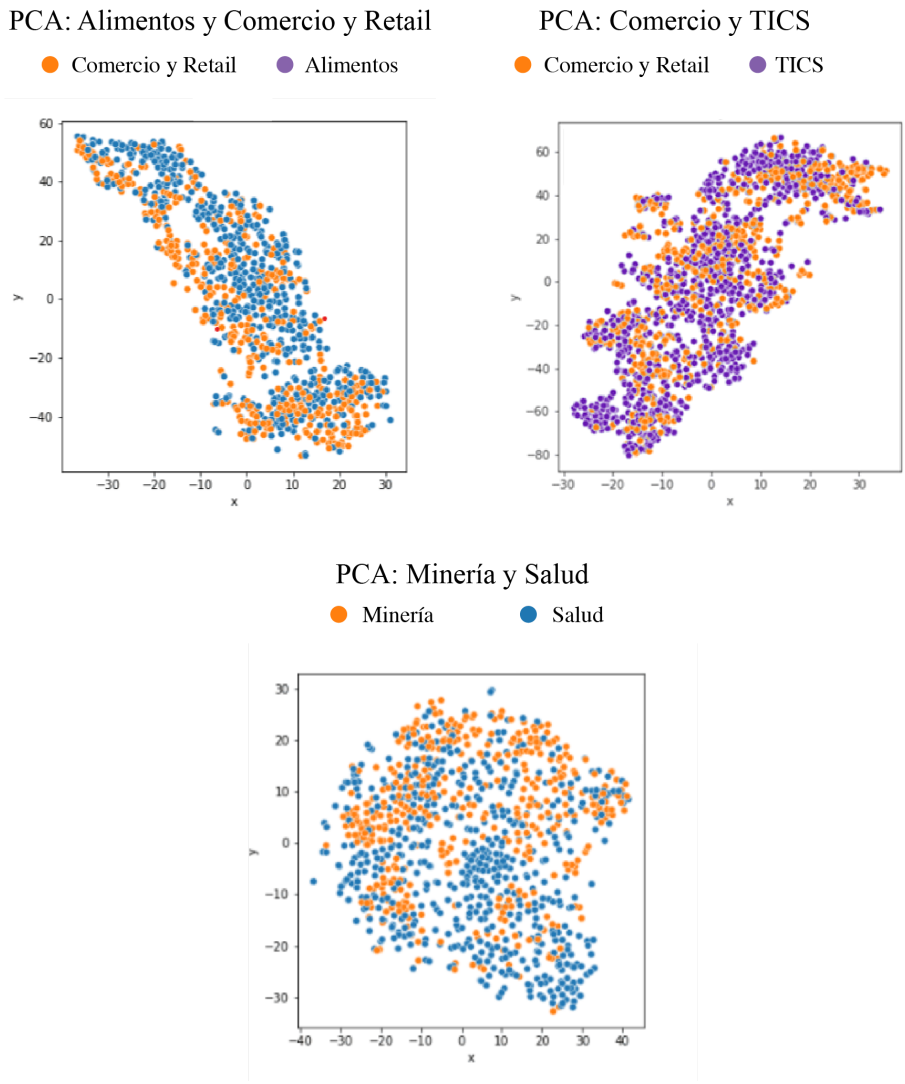


Figura 9: PCA de embeddings modelo BERT

Similitud en vocabulario

Nube de Palabras: Mercado Alimentos

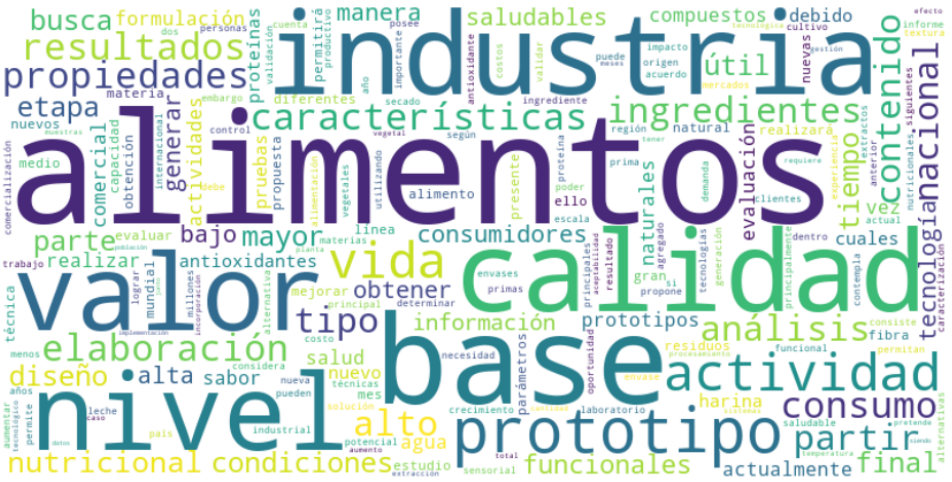


Figura 10: Nube de Palabras de texto para categoría Mercado Alimentos

Nube de Palabras: Mercado Comercio y Retail

Figura 11: Nube de Palabras de texto para categoría Mercado Comercio y Retail

Tipo de innovación

Para el modelo BETO de tipo de innovación se usó una arquitectura similar a la del modelo de mercado objetivo, pero se modificaron algunos parámetros como el número máximo de tokens. El desempeño de este modelo fue de 16% superior al modelo de línea base, alcanzando un *accuracy* de 0,73 y un *f1-score* ponderado de 0,71.

clase	***precision***	***recall***	***f1-score***	***support***
0 (Proceso)	0.61	0.6	0.60	620
1 (Producto)	0.78	0.76	0.77	1080
2 (Servicio)	0.75	0.78	0.77	992
accuracy			0.73	2692
macro avg	0.71	0.71	0.71	2692
weighted avg	0.73	0.73	0.73	2692

Tabla 5: Métricas modelo BERT para tipo de innovación

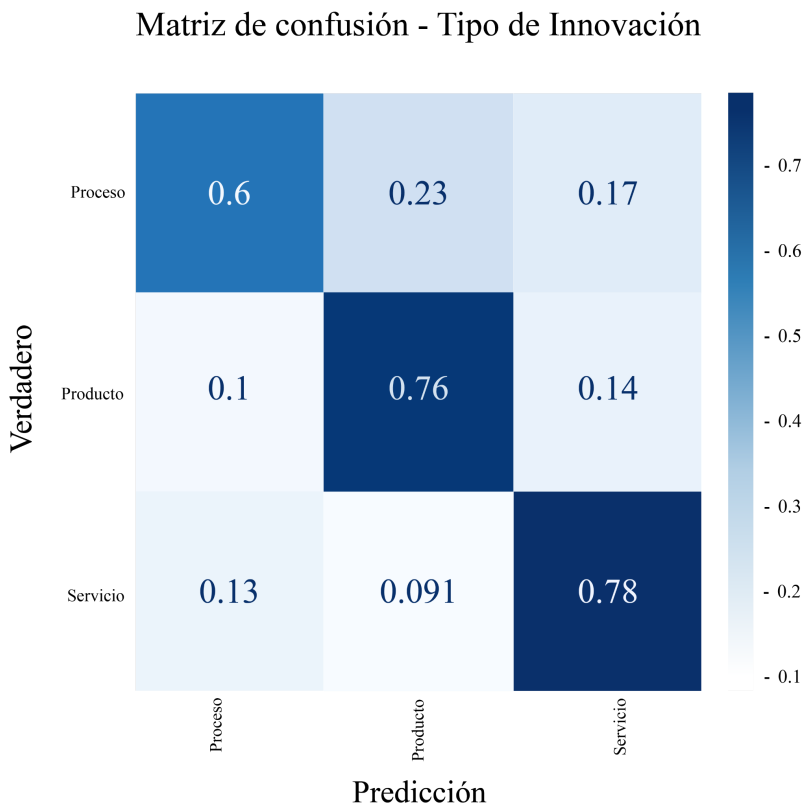


Figura 12: Matriz de confusión modelo BETO para tipo de innovación.

Clasificación de sostenibilidad

El modelo BETO para clasificación de sostenibilidad obtuvo un desempeño de un 35% superior al modelo de línea base, alcanzando un *accuracy* de un 0,804 y un *F1-score* ponderado de 0,8.

clase	***precision***	***recall***	***f1-score***	***support***
0 (No)	0.85	0.83	0.84	1454
1 (Sí)	0.74	0.75	0.84	1057
accuracy			0.80	2511
macro avg	0.79	0.79	0.79	2511
weighted avg	0.80	0.80	0.80	2511

Tabla 6: Métricas modelo BERT para clasificación de sostenibilidad.

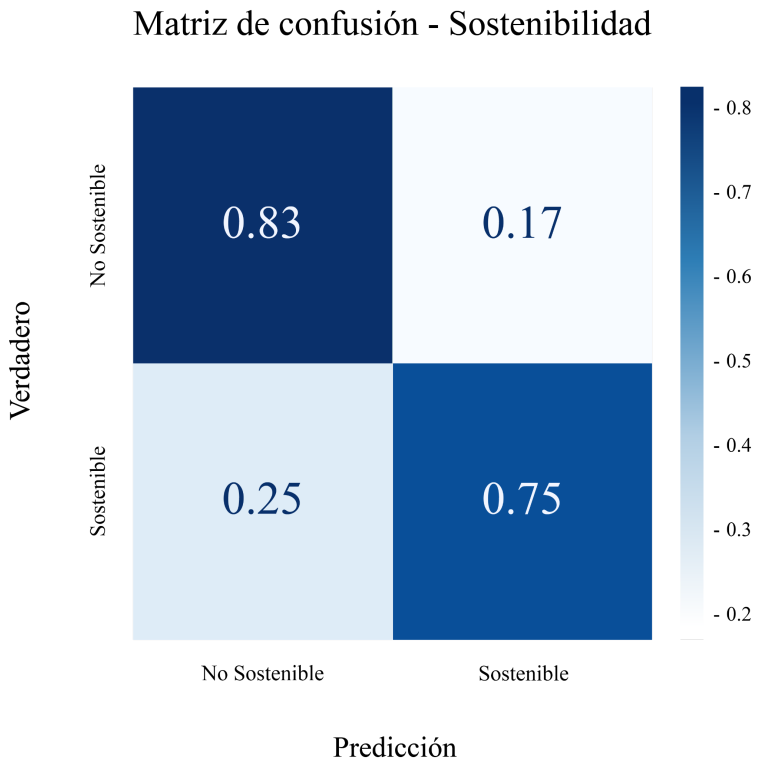


Figura 13: Matriz de confusión modelo BETO para clasificación de sostenibilidad.

5.3 Caracterización de la Innovación

En esta sección se busca realizar una caracterización de cerca de 12.000 proyectos postulados a programas de innovación de CORFO entre enero de 2019 y abril de 2022, haciendo uso de los modelos de clasificación. A pesar de que los modelos poseen un margen de error se busca obtener una caracterización general y de forma agregada que dé indicios del estado del ecosistema de innovación.

Mercado Objetivo

Con las clasificaciones realizadas con el modelo BETO de mercado objetivo, se seleccionaron tres de las principales sectores económicos para caracterizar la relación entre oferta y demanda de las innovaciones y analizar el flujo entre el sector al que pertenece la empresa con el mercado de llegada de la innovación.

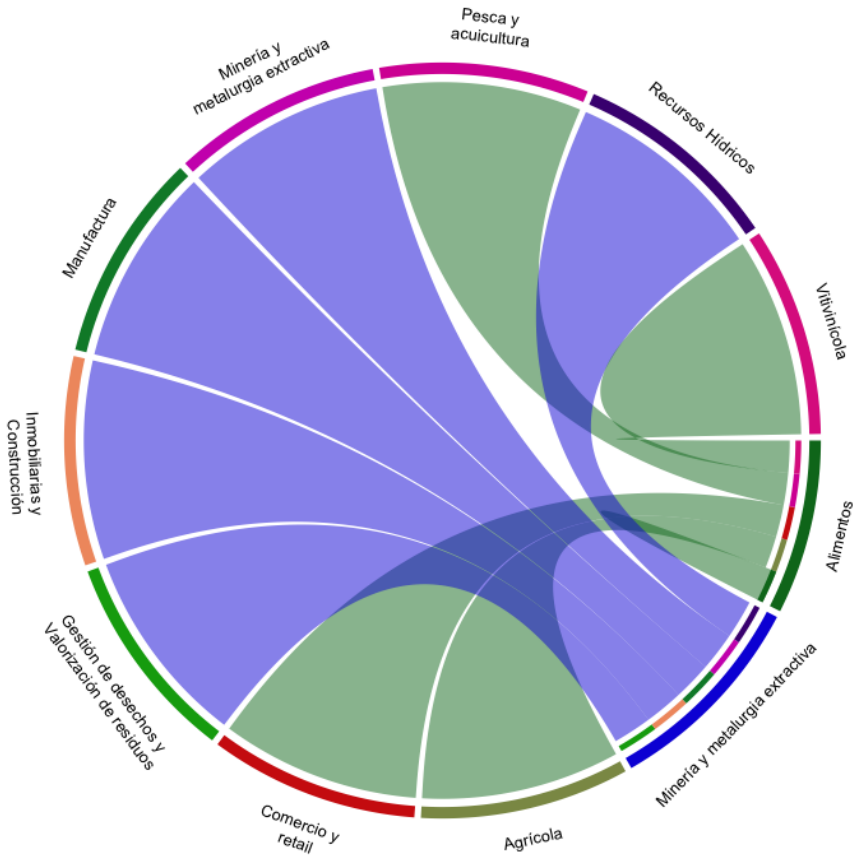


Figura 14: Mercados de llegada del sector Alimentos y Minería y metalurgia extractiva.

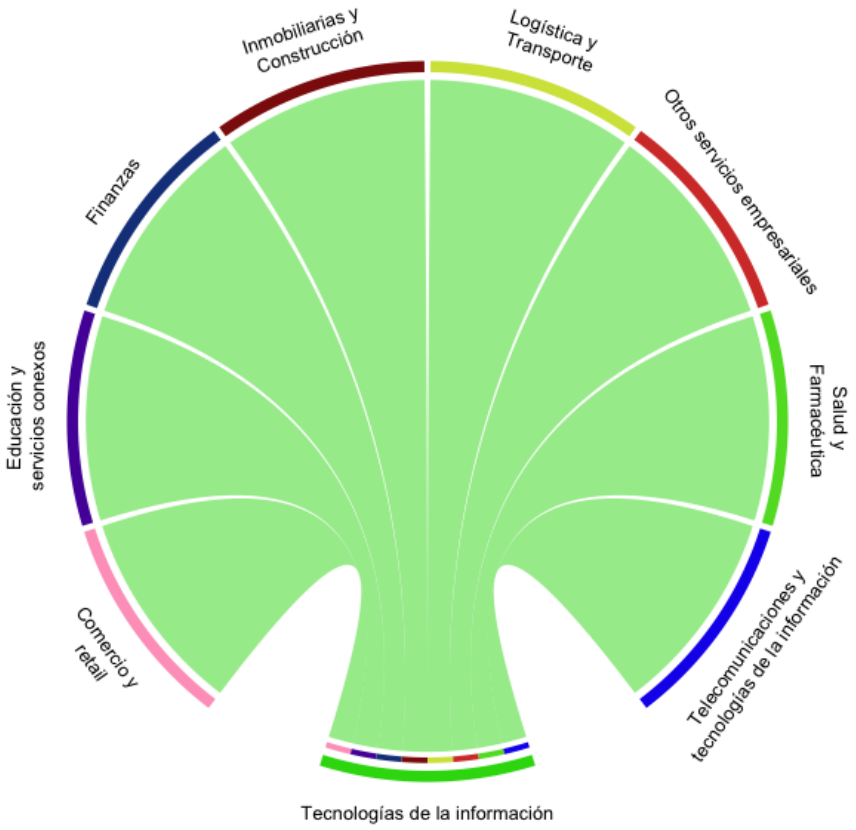


Figura 15: Mercado de llegada de innovaciones del sector Tecnologías de la información.

De este análisis se encontró que las empresas pertenecientes a la minería están demando principalmente proyectos de innovación que están dirigidos hacia la minería, pero también innovaciones relacionadas a recursos hídricos, manufactura y Gestión de desechos. A su vez, el sector Alimentos está innovando en el mercado de Alimentos, y también en el mercado Agrícola, Comercio y Retail, vitivinícola y el mercado objetivo de Pesca y Acuicultura.

También, las empresas pertenecientes al sector de Tecnologías de la información están innovando en su símil de mercado objetivo, Telecomunicaciones y tecnologías de la información, pero también en finanzas, logística y transporte, Salud y farmacéutica, Educación, e Inmobiliarias y construcción, con un fuerte componente de proyectos relacionados al uso de aprendizaje de máquinas.

Tipo de Innovación

Conocer qué tipo de innovación están demandando las industrias permite tener una aproximación al nivel de desarrollo de su matriz productiva. Por ejemplo, innovar en productos permite aumentar la complejidad económica de un sector, e innovar en procesos permite a una industria tener una mayor sofisticación (Fritsch & Meschede, 2001). Por ejemplo, la minería es un sector intensivo en innovación en procesos, a diferencia del sector de alimentos (menos sofisticado) que es más intensivo en productos. Ver figura 16.

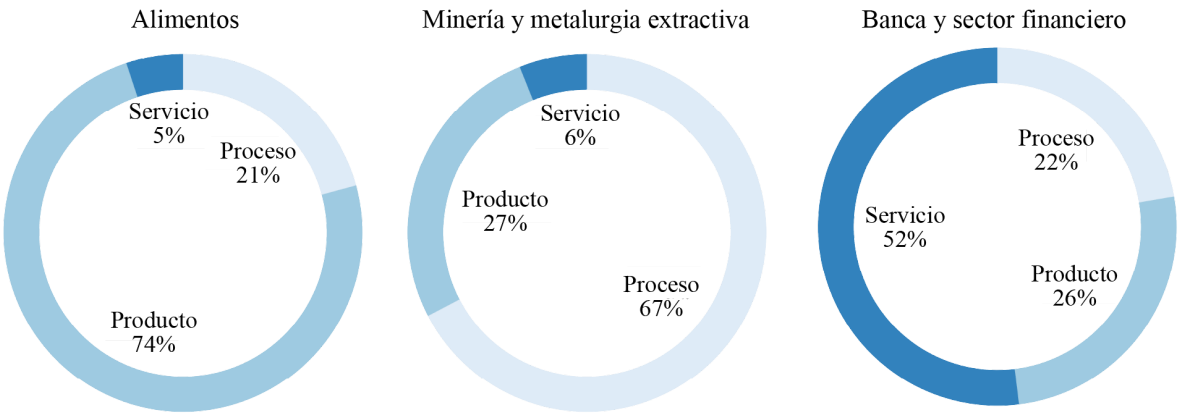


Fig. 16: Tipo de innovación por sector Alimentos, Minería y metalurgia extractiva, y Banca y sector financiero.

Al caracterizar el tipo de innovación por sector económico al que pertenecen las empresas postulantes se pueden encontrar hallazgos interesantes como, por ejemplo, el sector alimentos es intensivo en innovaciones en producto, esto puede dar indicios de la madurez, la competitividad y también de características subyacentes al sector. En cambio, el sector Minería y metalurgia extractiva es intensivo en innovación en procesos, lo que da indicios de un sector más maduro y más sofisticado, se podría inferir que las innovaciones en este sector buscan hacer más eficientes los procesos y que es un sector que compite en costos más que en productos. Como último ejemplo, se analizó Banca y sector financiero, el cual es intensivo en

innovaciones en servicios, lo cual podría parecer trivial ya que es la oferta propia del sector, pero también, más de un 25% de las innovaciones de este sector son en productos, lo que indicaría que las empresas podrían estar intentando ampliar su oferta.

Sostenibilidad

Con la clasificación realizada se puede caracterizar los sectores económicos que más hacen innovación con componente de sostenibilidad. Dentro de estos destacan sectores que por su naturaleza están más ligados a la eficiencia o reutilización de recursos, como Gestión de desechos, el sector Hídrico o el Energético, pero también aparecen sectores como Manufactura de no metálicos, en donde el 58% de las postulaciones son de manufactura sostenible. En este grupo destacan proyectos enfocados en la producción con materiales reutilizados o procesos de bajo impacto ambiental. También destaca el sector Minería y metalurgia extractiva, con un claro foco en procesos sostenibles y el manejo de relaves para disminuir su impacto ambiental.

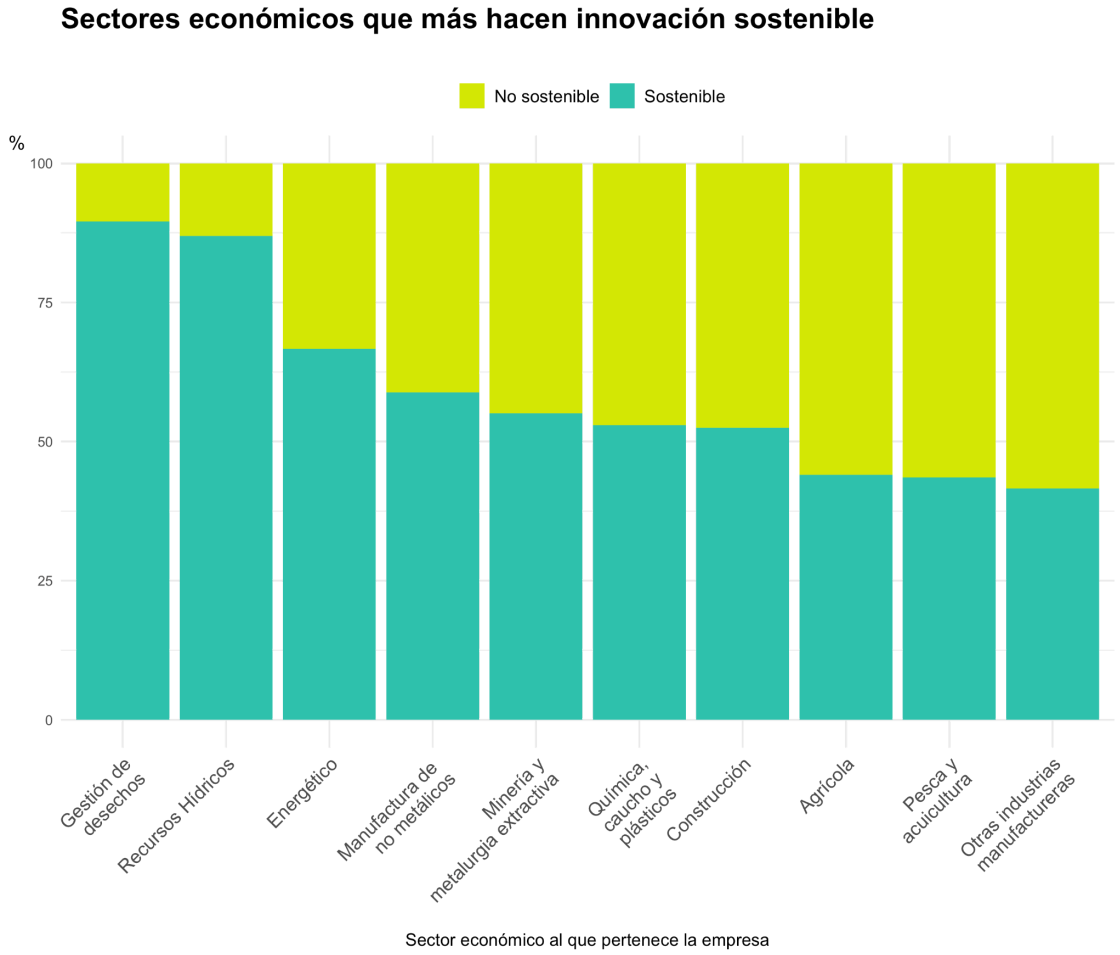


Figura 17:Sectores económicos por componente de sostenibilidad.

6. Conclusiones

Los avances en NLP en los últimos años han evidenciado una clara mejora en diversas tareas, entre ellas, la clasificación de texto. Esto se ve reflejado en el desempeño del modelo *Transformers* (2017) en comparación al modelo Word2Vec (2013).

El análisis de los resultados muestra que el modelo *Transformers* (BETO) tiene un desempeño superior en un 30% respecto al modelo Word2Vec para la clasificación de mercado objetivo en base al texto abierto del resumen del proyecto alcanzando un *accuracy* de 0,70. En términos de clasificación, el modelo *Transformers* supera al modelo de línea base en 20 de las 24 categorías de mercado objetivo. Esto se cumple para la clasificación de tipo de innovación, y de sostenibilidad, en donde el modelo *Transformers* tiene un desempeño superior de un 16% y 35% respectivamente en términos de *accuracy*. El análisis también muestra que tanto Word2Vec como *Transformers* tienen problemas en clasificar de manera correcta clases con pocas observaciones, en especial los modelo de clasificación de mercado objetivo, lo que indica que podría existir un margen de mejora en las métricas obtenidas para esas clases más pequeñas si es que se contase con un set de datos más balanceado y con masa crítica suficiente para las clases con menos observaciones, lo que escapa del alcance de este estudio. A este limitante se suma el dominio del corpus que se utilizó para pre-entrenar BETO, a pesar de ser un corpus grande, este está entrenado principalmente sobre Wikipedia, por lo que el dominio de ciertos tópicos no es profundo. Esto deriva en que hay *tokens* que el modelo no logra identificar, en especial, cuando son palabras muy técnicas o específicas de algún área de conocimiento.

Hay algunas direcciones claras para continuar el estudio y mejorar el desempeño de los modelos, una de ellas es trabajar para mejorar el balance de los datos, como por ejemplo, incluir técnicas de *oversampling* para clases con un bajo número de observaciones, o generación de datos sintéticos con técnicas como SMOTE (Chawla et al. 2002), o traducción inversa (Xie et al., 2019) aplicados sobre el *embedding* de los modelos. Otra dirección podría ser utilizar técnicas de “Destilado de conocimiento” (Knowledge Distillation) (Hinton et al., 2015) para aumentar el área de dominio de conocimiento del modelo a temas más específicos, similar a lo hecho con modelos enfocados en ciencia como SciBERT (Beltagy et al., 2019) o en leyes como Legal-Bert (Chalkidis et al., 2020).

7. Bibliografía

Alshari, E., Azman, A., & Doraisamy, S. (2017). Improvement of sentiment analysis based on clustering of Word2Vec features. *28th international workshop on database and expert systems applications (DEXA)*, (pp. 123-126).

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. arXiv.

Bro, R., & Smilde, A. (2014). Principal component analysis. Analytical methods, 6(9), 2812-2831.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J., Kang, H., & Pérez, J. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. PML4DC at ICLR.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. arXiv preprint.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Fritsch, M., & Meschede, M. (2001). Product innovation, process innovation, and size. Review of Industrial organization, 19(3), 335-350.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv, 2(7).

Hu, W., Du, J., & Xing, Y. (2017, Feb). Spam filtering by semantics-based text classification. Eighth International Conference on Advanced Computational Intelligence (ICACI), (pp. 89-94). IEEE.

Hughes, M., Kotoulas, S., & Suzumura, T. (2017). Medical Text Classification using Convolutional Neural Networks. In Informatics for Health: Connected Citizen-Led Wellness and Population Health. IOS Press, (pp. 246-250).

Lavanya, P. M. (2021). Deep learning techniques on text classification using Natural language processing (NLP) in social healthcare network: A comprehensive survey. 3rd International Conference on Signal Processing and Communication (ICPSC), IEEE(3), (pp. 603-609).

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv.

Mikolov, T. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning--based text classification: a comprehensive review. CM Computing Surveys (CSUR), 54(3), 1-40.

Ministerio de Ciencia, Tecnología, Conocimiento e Innovación. Resultados Encuesta Nacional de Innovación (ENI) 2017-2018. (2020).
OECD. (2018). Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation. OECD Publishing.
Schumpeter, J. (1942). Creative destruction. Capitalism, socialism and democracy, 825, 82-85.
Slavov, S., Tagarev, A., Tulechki, N., & Boytcheva, S. (2019). Company Industry Classification with Neural and Attention-Based Learning Models. Big Data, Knowledge and Control Systems Engineering (BdKCSE). IEEE, 1-7.
Tagarev, A., Tulechki, N., & Boytcheva, S. (2019, September). Comparison of Machine Learning Approaches for Industry Classification Based on Textual Descriptions of Companies. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), (pp. 1169-1175).
Vaswani, A., Shazeer, N., Parmar, N., & Uszkoreit, J. (2017). Attention is all you need. Advances in neural information processing systems, (30).

8. Anexos

8.1 Anexo 1: Definición mercado objetivo

Mercado objetivo	**Descripción**
Agrícola (excepto cultivo de uvas)	Proyectos que busquen atender (vender productos u optimizar procesos) a el cultivo y cuidado frutícola (excepto uvas), de hortalizas, de oleaginosas, leguminosas, flores ornamentales, insumos agrícolas, o plantas de uso medicinal, aromático o industriales; incluidos servicios de apoyo a la actividad agrícola tales como mejoramiento genético de las especies descritas u otros servicios, sensores, entre otros.
Alimentos (excepto producción de vino y derivados)	Proyectos que busquen atender (vender productos u optimizar procesos) a empresas dedicadas a cualquier proceso de elaboración y/o conservación de alimentos de origen animal o vegetal, incluyendo la elaboración de bebidas y productos de tabaco. Incluye alimentos para animales sin fines productivos (domésticos) y servicios o productos de apoyo a la producción de alimentos.
Asociaciones y organizaciones no empresariales ni gubernamentales o comunidades específicas	Proyectos que busquen atender (vender productos, crear u optimizar procesos) en organizaciones o comunidades específicas, como, por ejemplo, hogares de ancianos, comunas o localidades específicas que no estén contenidas en ninguna de las clasificaciones anteriores.
Comercio y retail	Proyectos que busquen atender (vender productos u optimizar procesos) a empresas que se dedican a la venta de productos sin transformación. Incluye comercio minorista, mayorista y reparación de productos.
Construcción e Inmobiliarias	Proyectos que busquen atender (vender productos u optimizar procesos) a empresas dedicadas a la preparación del terreno para construcción, a la construcción de edificios y sus instalaciones especializadas o a la construcción de infraestructura pública como carreteras, puertos u otras obras de ingeniería civil. Se incluyen servicios de apoyo a la construcción; Proyectos que busquen atender (vender productos u optimizar procesos) a la actividad inmobiliaria, específicamente en empresas dedicadas a la venta y compra de bienes raíces, alquiler de éstos o prestación de servicios

	inmobiliarios, como tasación.
Educación y servicios conexos	Proyectos que busquen atender (vender productos u optimizar procesos) en el rubro de la educación. Incluye organizaciones que impactan desde la educación preescolar a superior, la enseñanza deportiva, cultural o recreativa, y servicios de apoyo a la educación
Energético	Proyectos que busquen vender u optimizar procesos en empresas generadoras, transmisoras y/o distribuidoras de energía eléctrica proveniente de distintas fuentes (Hidrocarburos, Hidroeléctrica, Solar, Eólica, Biomasa, entre otras) o empresas que refinen, importen, transporten y/o distribuyan combustibles y otros derivados del petróleo, biocombustibles líquidos, biogás, gas natural, kerosene, gasolinas, diésel o similares.
Finanzas	Proyectos que busquen atender (vender productos u optimizar procesos) a instituciones bancarias (Bancos y sociedades financieras); a instituciones no bancarias como aseguradoras, reaseguradoras, bolsas de comercio, A.F.P, fondos mutuos, entre otras. Se incluyen servicios de apoyo a la industria de las finanzas.
Forestal	Proyectos que busquen atender (vender productos u optimizar procesos) a empresas que se dediquen a la silvicultura y/o la producción de productos madereros primarios (Tableros, chapas, postes, astillas y similares), pastas de madera, papel, cartón y artículos o envases de papel y cartón. Se incluyen servicios de apoyo a la actividad forestal, tales como mejoramiento genético de especies nativas o no nativas, entre otros.
Ganadero	Proyectos que busquen atender (vender productos u optimizar procesos) a empresas dedicadas a la cría y reproducción de animales no acuáticos y que generan productos destinados principalmente a la industria alimentaria, de vestuario o calzado; incluidos servicios de apoyo a la ganadería tales mejoramiento de alimentación, mejoramiento genético, farmacéutica para recursos ganaderos, entre otros.
Gestión de desechos y Valorización de residuos	Proyectos que busquen vender u optimizar procesos en empresas que se dediquen a recoger, tratar o eliminar desechos peligrosos y/o no peligrosos; a la recuperación de materiales; a la revalorización de residuos para la formación de nuevos materiales o realizar actividades de descontaminación
Industria creativa y esparcimiento	Proyectos que busquen vender productos u optimizar procesos en organizaciones que se dediquen a la industria creativa, específicamente a actividades artísticas y culturales tales como música, literatura y libros, audiovisuales y producción de espectáculos que incluya artes escénicas, artes visuales, fotografía, artesanías y aspectos artísticos de la arquitectura y diseño. En esparcimiento, proyectos que busquen atender (vender productos u optimizar procesos) a organizaciones que atiendan otros intereses culturales, recreativos y de entretenimiento al público general, como museos, espectáculos en vivo, juegos de azar y actividades deportivas y recreativas. Se excluyen actividades con fines turísticos.
Logística y Transporte	Proyectos que busquen atender (vender productos u optimizar procesos) a empresas dedicadas al almacenaje y distribución de materias primas, insumos y existencias de otras empresas como también al transporte de pasajeros.
Manufactura	Proyectos que busquen atender (vender productos u optimizar procesos) a empresas que se dediquen a la fabricación de maquinarias y equipos de uso industrial específico o genérico en distintas industrias; Proyectos que busquen atender (vender productos u optimizar procesos) a empresas que se dediquen a la fundición de metales ferrosos y no ferrosos o fabricación de productos de metal como partes, recipientes y estructuras de metal para diversos rubros (excepto maquinaria equipo) productos que generen, distribuyan o permitan consumir energía eléctrica como tableros, motores; cableado y electrodomésticos; equipos electrónicos; equipos de transporte como partes de buques, locomotoras, vehículos y aeronaves; fabricación de equipo de medición, prueba, navegación y control para diversos fines industriales y no industriales.
Minería y metalurgia extractiva	Proyectos que busquen atender (vender productos u optimizar procesos) a empresas dedicadas a la minería del cobre, otros metales y su refinación (incluida refinación de relaves); a la extracción de combustibles fósiles como carbón y petróleo; a la explotación de canteras de rocas o minerales no metálicos. Incluidos servicios de apoyo a la minería.

Multisectorial	Proyectos cuyo resultado puede ser comercializado en el corto plazo en más de una industria de las descritas en las demás categorías.
Otros servicios empresariales	Proyectos que busquen atender (vender productos u optimizar procesos) a empresas dedicadas a servicios de apoyo empresarial basados en el conocimiento tales como actividades de consultoría de gestión empresarial; de contabilidad; jurídicas; de arquitectura y diseño; investigación y desarrollo, análisis técnicos; publicidad; marketing; gráfica y estudios de mercado o servicios de apoyo administrativos como seguridad, arrendamiento, selección de personal, limpieza, entre otros.
Pesca y acuicultura	Proyectos que busquen atender (vender productos u optimizar procesos) a empresas dedicadas a la captura, caza, sego, recolección o producción de recursos hidrobiológicos tales como peces, crustáceos, moluscos, algas y otros organismos marinos, siempre con fines comerciales. Se incluyen servicios o actividades de apoyo a la pesca y cultivo, tales como mejoramiento de alimentación, mejoramiento genético, farmacéutica para recursos hidrobiológicos, entre otros.
Recursos Hídricos	Proyectos que busquen vender u optimizar procesos en consumidores o empresas que captan, tratan y suministran recursos hídricos a otras industrias y a domicilio, como también la captación, tratamiento y eliminación de aguas residuales
Salud y farmacéutica (en humanos)	Proyectos que busquen atender (vender productos u optimizar procesos) a instituciones públicas o privadas prestadoras de servicios de salud médicos u odontológicos, o bien, a empresas que se dediquen a la fabricación y/o preparación de productos medicinales; cosméticos; farmacéuticos y botánicos.
Sector público	Proyectos que busquen atender (vender productos u optimizar procesos) a instituciones pertenecientes al estado y que estén orientadas a mejorar la administración pública, la defensa nacional, el orden público, la promulgación de leyes, entre otros. Se excluyen las instituciones ministeriales relacionadas a mercados objetivo de los demás mercados descritos.
Telecomunicaciones y tecnologías de la información	Proyectos que busquen atender (vender productos u optimizar procesos) a empresas que se dediquen al rubro de las telecomunicaciones; actividades de tecnologías de la información; procesamiento de datos y otros servicios de información tales como programación informática, servidores, consultorías y actividades relacionadas a ellas.
Turismo	Proyectos que busquen atender (vender productos u optimizar procesos) a empresas dedicadas a la Hotelería, restaurantes, agencias de viaje, operadores turísticos y actividades relacionadas.
Vitivinícola	Proyectos que busquen atender (vender productos u optimizar procesos) a la industria vitivinícola, específicamente en empresas que se dediquen desde el cultivo de uvas y/o hasta la manufactura y envase de productos vitivinícolas. Se incluye el sub rubro pisquero.

Tabla 7: Descripción de clases de mercado objetivo.

8.2 Anexo 2: Análisis Exploratorio de Datos

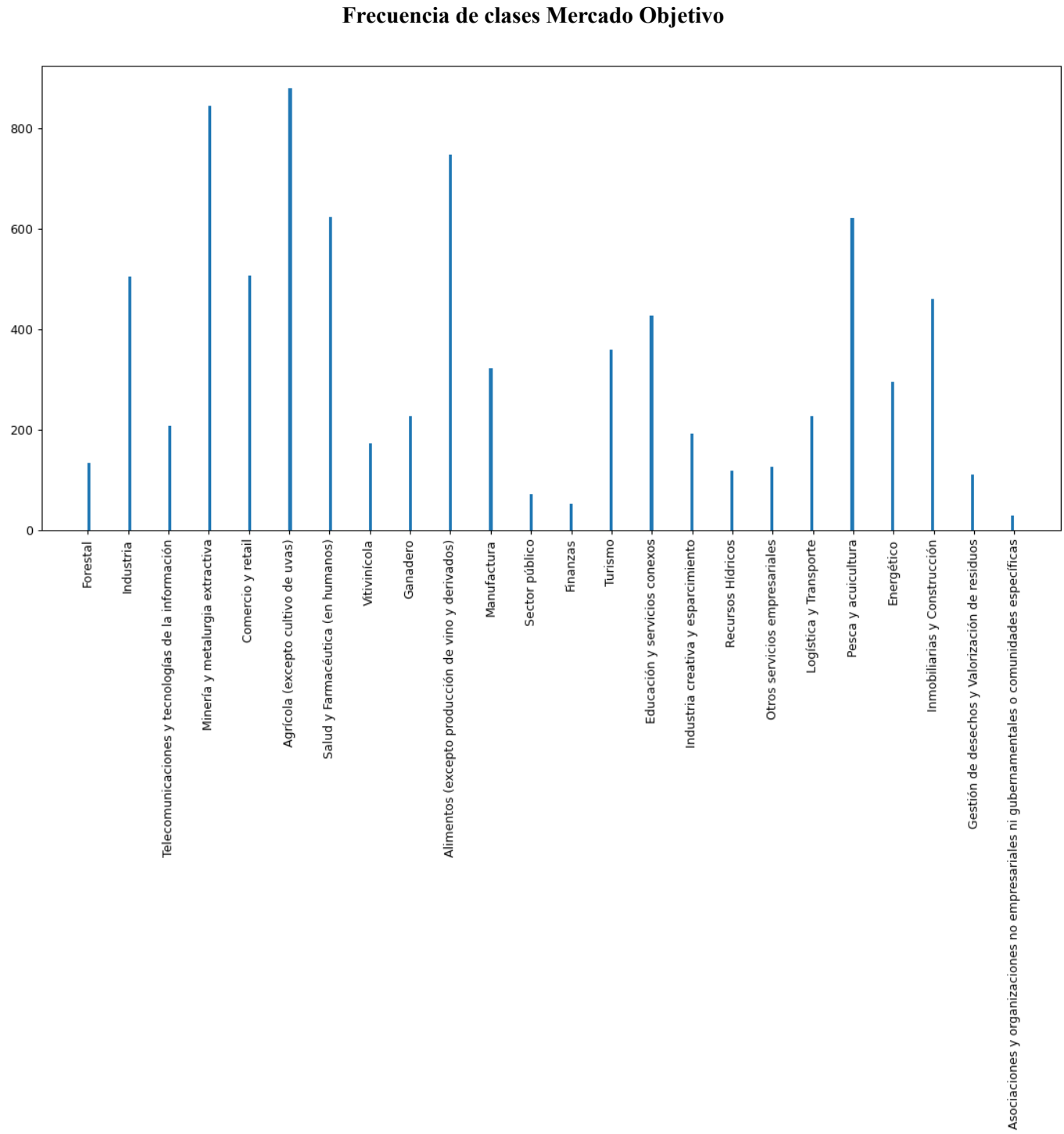
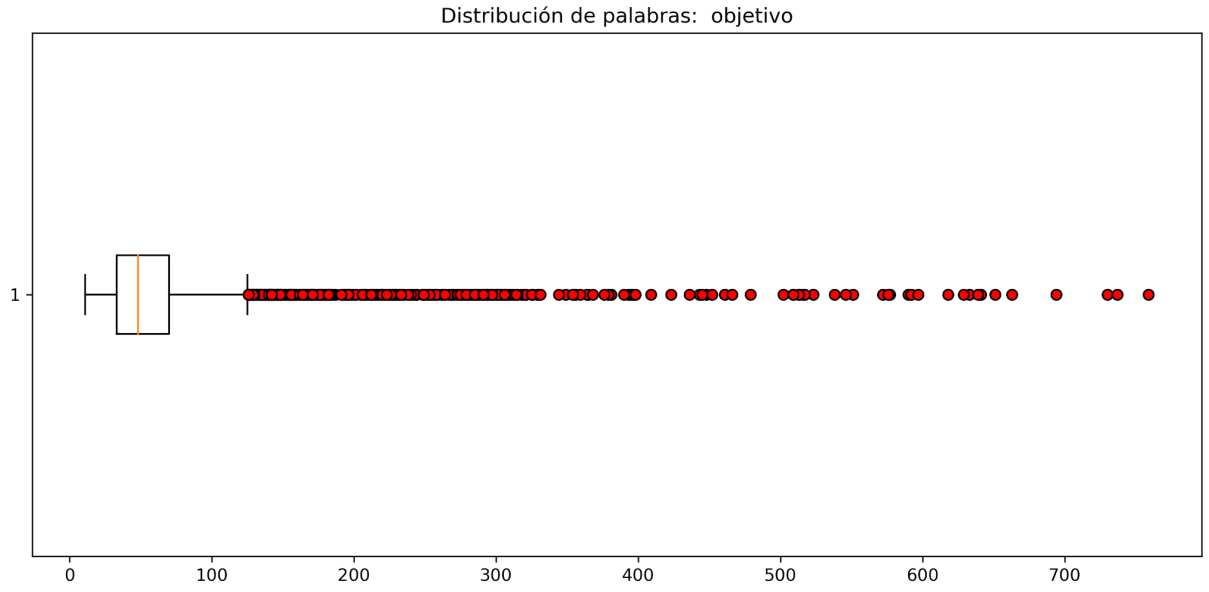
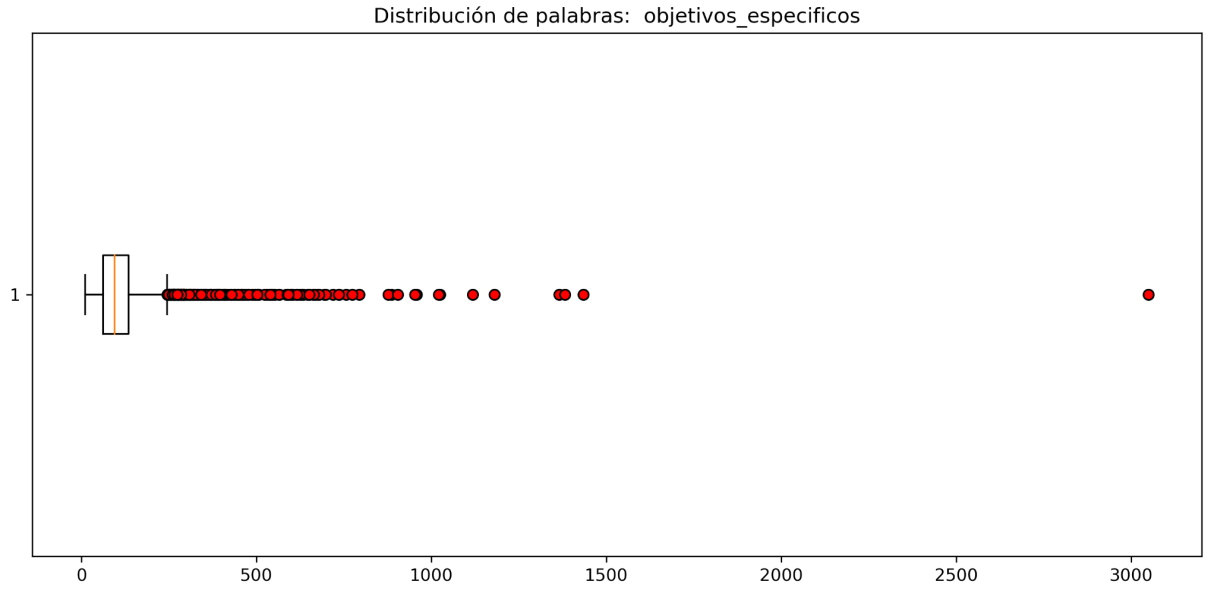


Figura 18: Frecuencia de clases Mercado Objetivo

Distribución de palabras por variable de interés





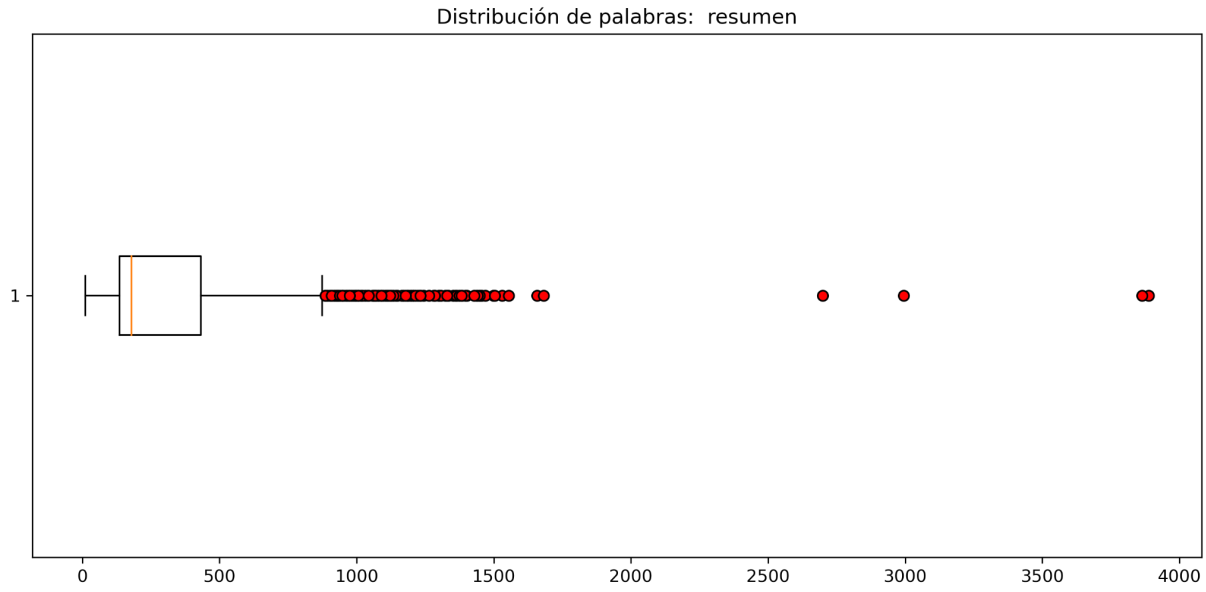


Figura 19:Gráficos de caja para distribución de palabras por objetivo, objetivos específicos y resumen del proyecto.

Distribución de proyectos por cantidad de tokens: campo Resumen del proyecto

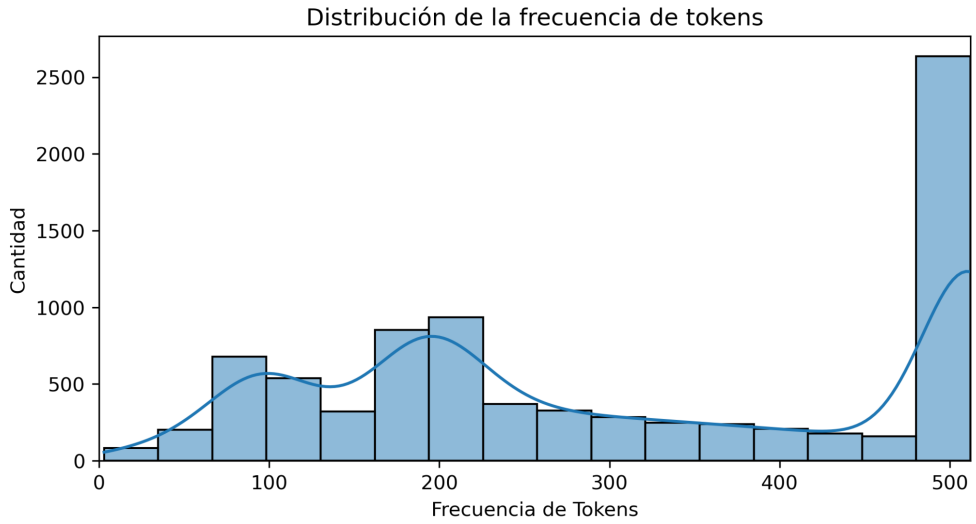


Figura 20: Cantidad de proyectos según largo de tokens.

8.3 Anexo 3: Word2Vec

clase	*precision*	*recall*	*f1-score*	*support*
0	0.73	0.94	0.82	264
1	0.84	0.97	0.90	224
2	0.00	0.00	0.00	9
3	0.64	0.62	0.63	152
4	0.46	0.91	0.61	128
5	0.00	0.00	0.00	88
6	0.00	0.00	0.00	16
7	0.00	0.00	0.00	40
8	0.30	0.08	0.12	68
9	0.00	0.00	0.00	33
10	0.13	0.41	0.20	152
11	0.00	0.00	0.00	57
12	0.10	0.19	0.13	138
13	0.00	0.00	0.00	68
14	0.28	0.06	0.10	97
15	0.68	0.87	0.76	254
16	0.00	0.00	0.00	37
17	0.91	0.89	0.90	187
18	0.00	0.00	0.00	35
19	0.92	0.85	0.89	187
20	0.00	0.00	0.00	21
21	0.00	0.00	0.00	62
22	0.00	0.00	0.00	108
23	0.21	0.04	0.06	52
accuracy			0.54	2477
macro avg	0.26	0.29	0.26	2477
weighted avg	0.46	0.54	0.48	2477

Tabla 8: Experimentos de entrenamiento Modelo Word2Vec + CNN para clasificación de mercado objetivo.

8.4 Anexo 4: Experimentos modelo BETO - Mercado Objetivo

learning rate	*epochs*	*batch_size*	*max_len*	*accuracy*	*f1_score*
0,00005	3	16	256	0,542149	0,495108
0,00005	3	16	320	0,603889	0,500897
0,00005	3	16	320	0,551128	0,525848
0,00005	3	16	320	0,601796	0,550728
0,00005	3	16	448	0,698022	0,696301
0,00005	5	32	128	0,546581	0,528641
0,00005	5	32	128	0,536015	0,534121
0,00005	5	32	192	0,576665	0,541401
0,00005	5	32	208	0,568677	0,528959
0,00005	5	16	216	0,666936	0,667073
0,00005	5	32	224	0,588888	0,567957
0,00005	5	16	320	0,595633	0,543338
0,00005	5	16	320	0,598013	0,554014
0,00005	5	16	448	0,696811	0,695178
0,00005	5	16	448	0,680993	0,680171
0,00005	5	16	448	0,682277	0,679039
0,00005	5	16	448	0,67824	0,674226
0,00005	5	16	448	0,70004	0,695524
0,00005	5	16	448	0,693985	0,691224
0,00005	**5**	**16**	**448**	**0,704078***	**0,702408***
3,5E-06	10	16	208	0,573829	0,567447
3,5E-06	15	32	224	0,652193	0,54188
3,5E-06	20	16	208	0,659574	0,668255

Tabla 9: Experimentos entrenamiento Modelo BETO

8.5 Anexo 5: Métricas por clase - Modelo BETO Mercado Objetivo

Clase	***precision***	***recall***	***f1-score***	***support***
0	0.80	0.86	0.83	264
1	0.65	0.71	0.68	224

2	0.12	0.56	0.20	9
3	0.36	0.36	0.36	152
4	0.90	0.83	0.86	128
5	0.66	0.69	0.68	88
6	0.40	0.75	0.52	16
7	0.74	0.80	0.77	40
8	0.68	0.78	0.73	68
9	0.45	0.64	0.52	33
10	0.81	0.38	0.52	152
11	0.54	0.56	0.55	57
12	0.79	0.80	0.79	138
13	0.70	0.66	0.68	68
14	0.48	0.35	0.40	97
15	0.91	0.83	0.87	254
16	0.13	0.14	0.13	37
17	0.88	0.93	0.91	187
18	0.68	0.71	0.69	35
19	0.81	0.81	0.81	187
20	0.46	0.29	0.35	21
21	0.36	0.32	0.34	62
22	0.80	0.92	0.85	108
23	0.73	0.83	0.77	52
accuracy			0.70	2477
macro avg	0.62	0.65	0.62	2477
weighted avg	0.72	0.70	0.70	2477

Tabla 10: Métricas por clase Modelo BETO para clasificación de mercado objetivo