Pontificia Universidad Católica del Ecuador

Facultad de Ingeniería

TEMA:

Aprendizaje Supervisado basado en texto para clasificar patentes de invención dentro de las subclases del sistema estandarizado de Clasificación Internacional de Patentes (CIP)

AUTOR:

Daniel Alejandro Atiencia Garzón

TUTOR:

Jhonny Vladimir Pincay Nieves

TRABAJO PREVIA A LA OBTENCIÓN DEL TÍTULO DE MAGISTER EN SISTEMAS DE INFORMACIÓN MENCIÓN DATA SCIENCE

Quito, Enero – 2024

DEDICATORIA

A Dios, mi esposa y familia cuyo apoyo incondicional ha sido un pilar fundamental en cada paso de mi camino. A mis padres, por su amor y sacrificio constante, que han sido la fuente de mi inspiración y fortaleza. A mi esposa por su apoyo incondicional en todo este proceso.

AGRADECIMIENTO

Al Dr. Jhonny Pincay, por compartir su conocimiento y apoyo durante el desarrollo de este proyecto.

Al Bioq. Fabián Darquea, Director Técnico de Patentes del Servicio Nacional de Derechos Intelectuales por el tiempo y conocimiento brindado.

RESUMEN

El presente proyecto se centra en la aplicación de técnicas de minería de datos para el desarrollo de un modelo de aprendizaje supervisado para la clasificación de patentes dentro del sistema de Clasificación Internacional de Patentes (CIP). Se aborda la clasificación de patentes únicamente en el idioma español debido a que durante la investigación de trabajos similares la mayor parte de esfuerzos están centrados en el idioma inglés y chino. Para lo cual se usó la información del Título y Resumen obtenida de la base de datos PATENTSCOPE que proporciona acceso a solicitudes internacionales de patentes que ya han sido publicadas. La fortaleza que ofrecen los modelos de aprendizaje profundo como las Redes Neuronales para encontrar patrones dentro de un conjunto de datos es importante por lo que se utilizó una Red Neuronal Convolucional Separable que es una variante de la Red Neuronal Convolucional que se enfoca en reducir la cantidad de parámetros y operaciones computacionales requeridas en las capas de convolución reduciendo la sobrecarga computacional. Las redes neuronales pueden deducir el significado de una palabra a partir del orden de estas, debido a esto se realizó una tokenización secuencial a fin de aprovechar las ventajas de su uso. Cuando se trabaja con texto es importante comprender que las palabras del conjunto de datos no son exclusivas del conjunto con el cual se está trabajando, pudiendo aprovechar las relaciones ya establecidas en otros conjuntos de datos. Para esto se usó un modelo preentrenado de Word2vec para trasferir ese aprendizaje previo al modelo y darle una ventaja durante el proceso de entrenamiento. Se espera que, al implementar el proyecto, las clasificaciones realizadas por el modelo puedan orientar de manera adecuada a un analista o investigador, además de ser una herramienta útil para detectar posibles oportunidades de innovación en las diferentes áreas tecnológicas.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1 Capítulo I. Introducción

1.1 Planteamiento del problema

La propiedad intelectual es uno de los pilares fundamentales para la protección de las creaciones intelectuales, en donde se otorga a los creadores derechos exclusivos sobre sus obras o invenciones. En Ecuador la Institución encargada de la protección, gestión y ejercicio del derecho intelectual es el Servicio Nacional de Derechos Intelectuales - SENADI.

La propiedad Intelectual en el Ecuador, se podría decir, que se subdivide en tres grandes campos de acuerdo con la Estructura Orgánica del Servicio Nacional de Derechos Intelectuales, donde por volúmenes de trámites ingresados tenemos en primer lugar a la Propiedad Industrial, encargada del registro de signos distintivos (Marcas de Productos, Marcas de Servicios, Lemas comerciales, entre otros.) y procedimientos de solicitudes de patentes de invención, patentes de modelo de utilidad, diseños industriales y esquemas de trazados de circuitos integrados. En segundo lugar, se encuentra el Derecho de Autor y derechos conexos que gestiona, promueve la protección y difusión de obras literarias, musicales, artísticas, entre otras. Finalmente, en tercer lugar, se encuentran las Obtenciones Vegetales y Conocimientos Tradicionales que garantizan la protección y defensa de las nuevas variedades vegetales para el fomento de la innovación. Como se puede observar la Propiedad Intelectual no sólo incentiva la generación de nuevas ideas y tecnologías, sino que también es un pilar importante para el desarrollo económico y cultural del Ecuador.

En el mundo existe un organismo cuyo objetivo es desarrollar un sistema de propiedad intelectual internacional, la Organización Mundial de la Propiedad Intelectual - OMPI o WIPO por sus siglas en inglés, pretende que este sistema sea equilibrado, accesible, recompense la creatividad, estimule la innovación y contribuya al desarrollo económico. Ecuador como parte de este organismo internacional, utiliza el sistema de Clasificación Internacional de Patentes – CIP, que es un esquema implementado para organizar y homogeneizar las patentes de invención de acuerdo con sus diferentes áreas tecnológicas. Este sistema es una herramienta importante para el acceso a información que se encuentra en el estado de la técnica al proporcionar un lenguaje común, haciendo que inventores, investigadores y empresas puedan identificar tendencias y detectar posibles oportunidades de innovación en las diferentes áreas tecnológicas.

Durante el año 2022, según el Servicio Nacional de Derechos Intelectuales, en el Ecuador se recibieron 532 solicitudes de Patentes de invención y Modelos de Utilidad, de las cuales únicamente 47 corresponden a solicitudes ecuatorianas, siendo un número bajo en comparación con el número de solicitudes internacionales. La ventaja de proteger las innovaciones en el Ecuador es que la concesión de una patente de invención otorga a su solicitante la explotación exclusiva de su innovación durante un período de tiempo determinado, para una patente de invención son 20 años y para el caso de un Modelo de Utilidad es de 10 años (Comisión de la Comunidad Andina, 2000).

El proceso de clasificación de una patente es arduo y complejo donde también puede ser necesario contar con una cantidad significativa de tiempo y poseer el suficiente conocimiento técnico, esto debido a que la Clasificación Internacional de Patentes tiene varios niveles jerárquicos con alrededor de 70.000 subdivisiones (OMPI, 2023). Por lo que, el entrenamiento de un modelo de Aprendizaje Supervisado basado en texto para su clasificación podría simplificar este proceso y ser una herramienta importante para quienes desean realizar búsquedas de las innovaciones que se encuentren en el estado de la técnica además de ser una herramienta importante para los examinadores del Servicio Nacional de Derechos Intelectuales en la gestión de trámites.

El uso de algoritmos para la calificación de texto no es nuevo, en la actualidad estos algoritmos están integrados a una variedad de sistemas de software que procesan textos a grandes escalas. Un ejemplo de esto es el software que se usa para determinar si el correo electrónico que entra a nuestras bandejas debe ser enviado a la bandeja de recibidos o debe ser filtrado a la carpeta de spam (Google Developers, 2023).

Con lo mencionado anteriormente el presente trabajo se ha planteado analizar los textos de la documentación de las patentes y extraer características claves que permitan ser utilizadas para categorizar los documentos dentro de las subclases de la Clasificación internacional de patentes en el idioma español resolviendo a las siguientes interrogantes.

RQ1.¿Es posible clasificar las patentes de invención dentro de las subclases del sistema estandarizado de Clasificación Internacional de Patentes – CIP mediante el título y resumen?

Importancia. Como se mencionó anteriormente, es importante conocer si el proceso de clasificación se lo puede simplificar mediante únicamente el uso del título y resumen de la misma, ya que la información que se encuentra en el documento técnico donde se describe detalladamente la invención, suele contener dibujos, resumen, reivindicaciones y la memoria descriptiva de la invención.

RQ2.¿Se puede utilizar el Procesamiento de Lenguaje Natural para clasificar las Patentes de Invención?

Importancia. En la actualidad el Procesamiento de Lenguaje Natural es utilizado para diferentes fines por lo que es importante probar si se lo puede usar para la clasificación de documentación técnica.

RQ3.¿Es suficiente con el texto del título y resumen de la patente para poder realizar una clasificación óptima?

Importancia. La documentación técnica con la que se debe presentar una patente incluye dibujos y una memoria descriptiva bastante extensa, donde se describe detalladamente la invención presentada por lo que, sí al clasificar una patente de manera óptima usando únicamente el título y resumen simplificará el proceso.

RQ4.¿Qué modelo de machine learning permite tener mejores resultados en la clasificación de patentes?

Importancia. En la actualidad se cuentan con muchos modelos de machine learning por lo que sería importante comparar los resultados obtenidos con diferentes modelos, la información se obtendrá durante la investigación de trabajos relacionados, donde se encontrarán resultados, modelos usados y comparar con el resultado obtenido en el presente proyecto.

En base a lo descrito, este proyecto busca la clasificación de texto mediante el uso del Procesamiento del Lenguaje Natural para asignar los documentos a una o varias categorías ya definidas en la Clasificación Internacional de Patentes basándose en las tecnologías o procedimientos descritos en la documentación técnica presentada, para tratar de simplificar un proceso arduo y ayudando a promover el acceso a innovaciones dentro del estado de la técnica sin necesidad de un conocimiento técnico extenso.

1.1 Objetivos

1.1.1 Objetivo general

Aplicar técnicas y métodos de la ciencia de datos que permitan entrenar un modelo de aprendizaje supervisado basado en texto para clasificar patentes de invención dentro de las subclases del sistema estandarizado de Clasificación Internacional de Patentes (CIP).

1.1.2 Objetivos específicos

- Revisar técnicas y metodologías para el procesamiento de lenguaje natural para la clasificación de documentos.
- Recopilar un conjunto de datos de patentes de invención que contenga título y resumen.
- Utilizar técnicas de procesamiento de lenguaje natural para extraer características clave de las patentes de invención
- Entrenar un modelo de aprendizaje supervisado que permita clasificar las patentes de invención.
- Evaluar el rendimiento del modelo de aprendizaje supervisado utilizado para la clasificación de patentes de invención.

2 Capítulo II. Marco Teórico y revisión de trabajos relacionados

2.1 Clasificación Internacional de Patentes (CIP)

La Clasificación Internacional de Patentes (CIP) es una herramienta que permite organizar de manera eficiente las diversas innovaciones protegidas a lo largo del tiempo mediante las patentes. Este sistema de clasificación tiene una estructura que categoriza tecnologías específicas, donde incluye información jerárquica como sección, clase, subclase, grupo principal y subgrupo.

- **Sección.** Representada por una letra, de la A hasta la H, de la Clasificación Internacional de Patentes.

  Sección A. Necesidades corrientes de la vida.
  Sección B. Técnicas industriales diversas; transportes.
  Sección C. Química; metalúrgica.
  Sección D. Textiles; papel.
  Sección E. Construcciones fijas.
  Sección F. Mecánica; iluminación; calefacción; armamento; voladura.
  Sección G. Física.
  Sección H. Electricidad.

- **Clase.** Representada por la letra de la sección seguida de un número de dos dígitos.

- **Subclase.** Representada por la clase seguida de una letra.

- **Grupo.** Según la Clasificación Internacional de Patentes, un grupo puede dividirse en Grupo Principal y Subgrupos. El Grupo Principal es el cuarto nivel jerárquico y los subgrupos están relacionados o dependen del Grupo Principal.

Debido a que el proceso de clasificación es complicado y para una mejor comprensión de la clasificación de las patentes y sus diferentes niveles se muestra un ejemplo de la página 18 del documento 'Información de patentes: Uso y manejo de bases de datos de patentes' de Manuel Castro Calderón (2023)

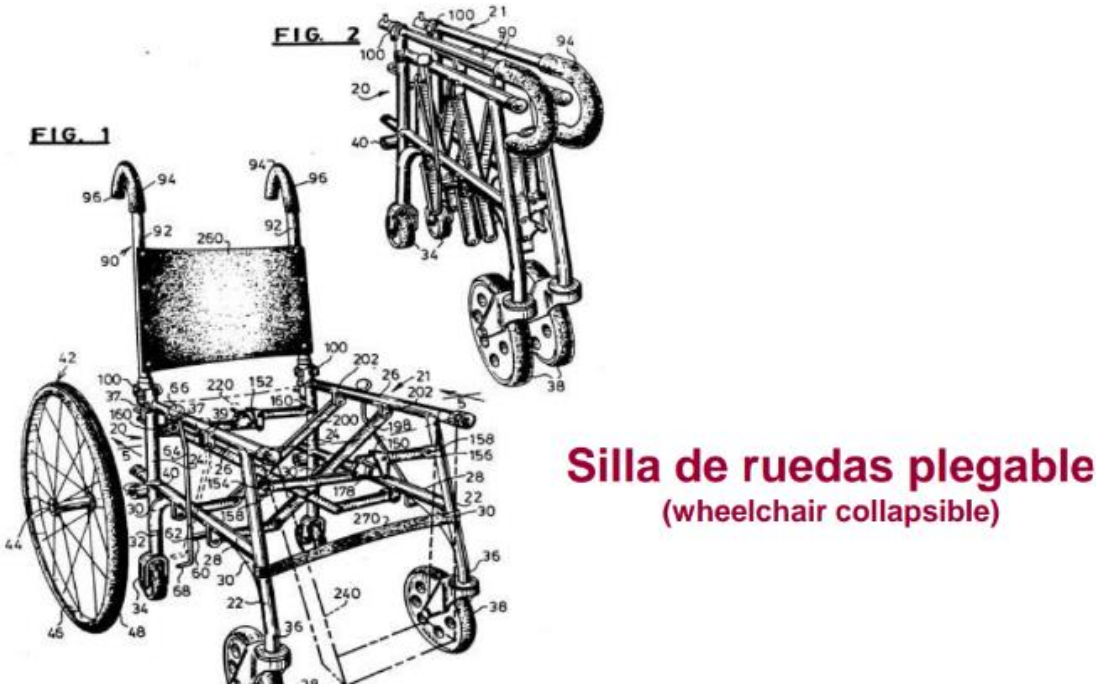
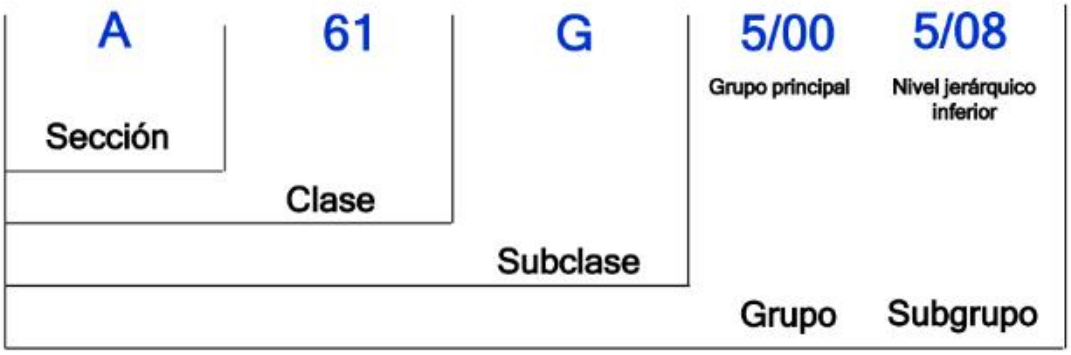


Figura 1. Dibujo Patente Silla de ruedas Plegable
Fuente. Castro Calderón, (2023, P-18).

Simbolo completo de clasificación:



A Necesidades corrientes de la vida

A61 Ciencias médicas o veterinarias; higiene

A61G Medios de transporte o accesorios para enfermos

A61G 5/00 Vehículo con varias ruedas especialmente adaptados para inválidos

A61G 5/08 . plegables

Figura 2. Dibujo Patente Silla de ruedas Plegable
Fuente. Castro Calderón, (2023, P-19).

2.2 Aprendizaje supervisado

El aprendizaje supervisado es considerado como una de las técnicas para minería de datos, cuyo objetivo es descubrir patrones mediante el uso del análisis de datos. El aprendizaje supervisado pretende descubrir las relaciones existentes entre las variables de entrada y las variables de salida como se muestra en la Figura 3.

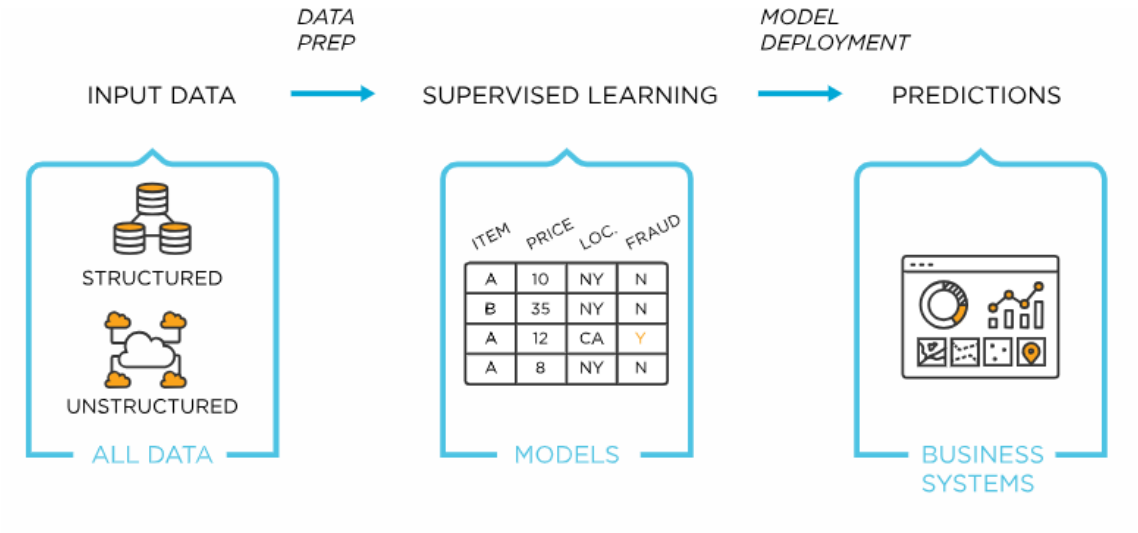


Figura 3. Aprendizaje supervisado
Fuente.TIBCO Software Inc, (2023).

Para el aprendizaje supervisado es necesario considerar los siguientes puntos:

- La disponibilidad de un conjunto de datos que contiene ejemplos de entrada o características y sus etiquetas correspondientes o resultados conocidos. Estas etiquetas son importantes para que el modelo pueda reconocer patrones.

- Elección de un algoritmo de aprendizaje automático apropiado, dentro de los cuales se encuentran Support Vector Machine (SVM), Redes Neuronales, Regresión Lineal o Arboles de Decisión.

- El entrenamiento del modelo, que se da cuando alimentamos al modelo de aprendizaje automático con el conjunto de datos que contiene ejemplos de entrada y salida. Durante el entrenamiento es importante revisar los parámetros para minimizar la diferencia entre las predicciones realizadas y las respuestas reales en los datos de entrenamiento.

- La evaluación del modelo se la debe realizar una vez terminado el entrenamiento, esto se lo realiza utilizando datos de prueba que no se consideraron durante el entrenamiento. Evaluar el modelo nos permite medir qué tan bien está generalizando los datos de entrenamiento a nuevos datos con la finalidad de asegurar que el modelo pueda realizar predicciones precisas. Para esto se utilizan métricas de evaluación

específicas como el F1-score, Error Cuadrático Medio (MSE), coeficiente de determinación $\left(R^2\right)$ o el Accuracy.

- Finalmente, el modelo entrenado y evaluado puede ser utilizado para hacer predicciones de nuevos datos, haciendo predicciones en base a lo aprendido del conjunto de entrenamiento.

2.2.1 Algoritmos de aprendizaje supervisado

IBM (2023) proporciona información relevante sobre algoritmos de aprendizaje supervisado en su página web. Donde para la aplicación del aprendizaje supervisado, se emplean algunos métodos como.

- **Redes Neuronales.** Son principalmente utilizadas en algoritmos de aprendizaje profundo y funcionan imitando la estructura neuronal del cerebro humano mediante el uso de capas y nodos.

- **Naive Bayes.** Este método se basa en el principio de independencia condicional del Teorema de Bayes, utilizado para clasificación, este principio implica que la presencia de una característica no influye en la presencia de otra en cuanto a la probabilidad de un resultado. Este enfoque se utiliza comúnmente en la clasificación de texto, sistemas de recomendación y detección de spam.

- **Regresión Lineal.** Este algoritmo es usado para buscar relaciones entre una variable dependiente y una o más variables independientes. Cuando solo involucra una variable independiente, se conoce como regresión lineal simple, en caso de tener múltiples variables independientes, se denomina regresión lineal múltiple. El objetivo es trazar una línea de mejor ajuste utilizando el método de mínimos cuadrados.

- **Regresión Logística.** Es utilizada cuando la variable dependiente es categórica, es decir, tiene resultados binarios como “verdadero” o “falso”, “si” o “no”. Este modelo al igual que la regresión lineal, busca comprender las relaciones entre los datos, sin embargo, la clasificación logística intenta resolver problemas de clasificación binaria como la detección de spam.

- **K vecinos más cercanos (KNN).** Este es un método no paramétrico que clasifica los datos según su proximidad y semejanza. Supone que los puntos de datos similares son cercanos entre sí. Calcula la distancia entre puntos de datos y luego asigna una categoría basada en el promedio más común. KNN es conocido por su facilidad de uso, pero el tiempo de procesamiento aumenta a medida que crece el conjunto de datos de prueba, lo que lo hace menos adecuado para tareas de clasificación con grandes conjuntos de datos. Se utiliza generalmente en motores de recomendación y reconocimiento de imágenes.

2.3 Minería de texto

La minería de texto es el proceso mediante el cual se convierte información no estructurada en un formato organizado con el fin de identificar patrones y obtener nueva información. Este proceso utiliza técnicas de análisis como las Máquinas de Vectores de Soporte (SVM), Naïve Bayes además de otros algoritmos de aprendizaje profundo.

El texto es uno de los tipos de datos más comunes que se pueden encontrar en bases de datos, por lo que pueden ser categorizados de la siguiente manera:

- **Datos estructurados.** Están en formato tabular y se caracterizan por su facilidad de almacenamiento y procesamiento, siendo beneficioso para su análisis en algoritmos de aprendizaje automático.

- **Datos no estructurados.** No siguen un formato estándar y pueden incluir textos provenientes de redes sociales, reseñas de productos, así como formatos multimedia de audio y video.

- **Datos semiestructurados.** Representan una mezcla entre los datos estructurados y no estructurados. Aunque poseen algún grado de organización, generalmente no cumplen con los requisitos de una base de datos relacional.

Figura 4. Aprendizaje supervisado
Fuente. Forum Huawei, (2022).

Dado que cerca del 80% de los datos mundiales no son estructurados. (Dialani, P.,2020). El uso de herramientas para el análisis de texto y técnicas de procesamiento de lenguaje natural (PLN) permite convertir documentos no estructurados en formatos organizados, facilitando el análisis de los mismos y la generación de información de calidad.

2.3.1 Técnicas de minería de texto

De acuerdo con la página web de IBM (2023), el análisis de texto comprende una serie de actividades que permiten extraer información valiosa de datos no estructurados, que se encuentran en forma de texto. Por lo que al aplicar las diferentes técnicas de minería de texto es importante realizar un preprocesamiento que implica limpieza y transformación de los datos en un formato utilizable para su posterior análisis. Entre las técnicas de minería de texto tenemos:

a. Recuperación de información

La recuperación de información se enfoca en encontrar información o documentos en función a consultas o frases específicas. Estos sistemas de recuperación de información utilizan algoritmos para monitorear las actividades de los usuarios e identificar datos relevantes. Para la recuperación de información, se aplican los siguientes métodos.

- Tokenización. Este método consiste en descomponer un texto extenso en frases y palabras individuales, conocidas como *"tokens".* Estos tokens son importantes para tareas como comparación de documentos

- Lematización. Este proceso implica la eliminación de prefijos y sufijos de las palabras para obtener su forma base o lema, facilitando el análisis y comparación de textos.

b. Procesamiento de lenguaje natural

El procesamiento de lenguaje natural integra la lingüística computacional con modelos estadísticos de aprendizaje automático (machine learning) y de aprendizaje profundo (deep learning). En conjunto, estas tecnologías permiten procesar el lenguaje natural en forma de datos de texto o voz.

c. Extracción de información

Se centra en mostrar información clave al buscar varios documentos. Esta técnica incluye el análisis detallado para obtener información estructurada de textos libres, permitiendo el almacenamiento de atributos clave y relaciones en una base de datos.

2.4 Metodologías de análisis de datos

2.4.1 SEMMA

Desarrollado por una reconocida empresa de software, es un proceso utilizado para seleccionar, explorar y modelar grandes conjuntos de datos. Este proceso facilita la identificación de patrones en los datos empresariales que previamente eran desconocidos. SEMMA se compone de cinco fases fundamentales: una de muestreo, otra de exploración, una de modificación, otra de modelado y finalmente una de valoración.

- **Sample (Muestra).** La etapa de muestreo en la metodología SEMMA es fundamental para el proceso de análisis de datos. En esta fase, se selecciona una muestra representativa del conjunto de datos completo, lo que permite trabajar de manera más eficiente con grandes conjuntos de datos sin comprometer la calidad del análisis.

- **Explore (Explorar).** La exploración de datos es una etapa crucial que sigue al muestreo. Durante esta fase, los analistas examinan los datos de manera visual y estadística para identificar patrones, tendencias y posibles relaciones entre las variables. Esto proporciona una base sólida para la siguiente etapa de modificación y modelado.

- **Modify (Modificar).** La etapa de modificación se centra en abordar problemas de calidad de datos que puedan surgir, como datos faltantes o valores atípicos. Los analistas pueden aplicar técnicas de preprocesamiento, como la imputación de datos faltantes o la transformación de variables, para preparar los datos de manera adecuada para el modelado.

- **Model (Modelar).** La etapa de modelado es central en el proceso SEMMA y consiste en la construcción de modelos predictivos o descriptivos. Aquí es cuando se aplican técnicas de aprendizaje automático o estadísticas para desarrollar modelos que puedan realizar predicciones o describir patrones en los datos, según los objetivos del análisis.

- **Assess (Evaluar).** Aquí se evalúan los modelos creados para determinar su precisión y eficacia. Esto implica la utilización de métricas de rendimiento y validación cruzada para medir qué tan bien se ajusta el modelo a los datos y si cumple con los objetivos establecidos en el análisis.

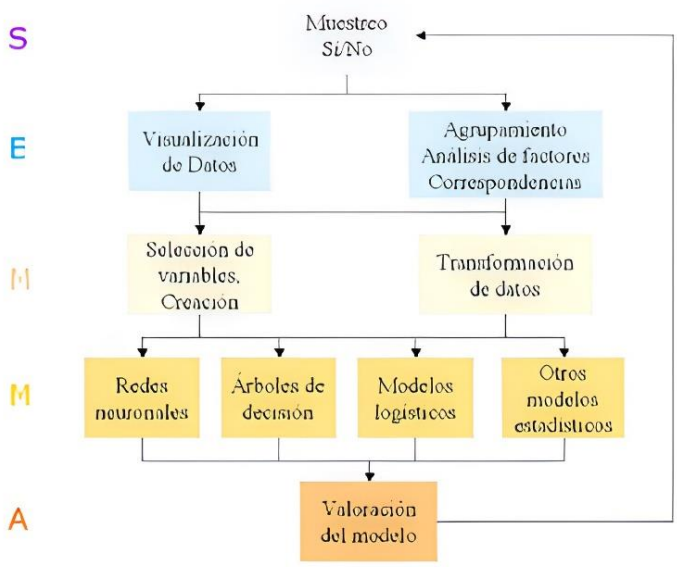


Figura 5. Metodología SEMMA
Fuente. Rodríguez Montequín, Teresa; Álvarez Cabal, J. Valeriano; Mesa Fernández, José Manuel; González Valdés, Adolfo. (s.f.).

2.4.2 CRISP-DM

CRISP-DM es una metodología utilizada para proyectos de minería de datos que va desde la comprensión del negocio hasta el despliegue de modelos. Siendo una metodología flexible que cuenta con varias fases:

- **Comprensión del negocio (Business Understanding).** En esta fase se identifican los problemas y oportunidades que la minería de datos puede abordar, definiendo el alcance del proyecto.

- **Comprensión de los datos (Data Understanding).** Aquí, se recopilan y exploran los datos relevantes para el proyecto. Esto incluye la descripción de su estructura y calidad además de la identificación de posibles problemas.

- **Preparación de datos (Data Preparation).** En esta fase, se limpian, adecuan y preparan los datos para su análisis. Esto implica seleccionar características relevantes, tratar valores faltantes y realizar otras tareas de limpieza y procesamiento antes de pasar a la etapa de modelado.

- **Modelado (Modeling).** En esta etapa, se seleccionan y aplican técnicas para el modelado de datos, como algoritmos de aprendizaje automático, para construir modelos predictivos o descriptivos. La aplicación de los modelos debe ser evaluados y en caso de ser necesarios se realizan ajustes a los modelos.

- **Evaluación (Evaluation).** Se evalúan los modelos desarrollados en la fase anterior para determinar su calidad y su capacidad para cumplir con

los objetivos planteados. Se utilizan métricas y técnicas de evaluación para medir el rendimiento del modelo.

- **Despliegue (Deployment).** En esta etapa se presentan los resultados obtenidos durante la minería de datos en donde se pueden usar herramientas de visualización para que sea más comprensibles los resultados que se obtuvieron en la minería de datos.

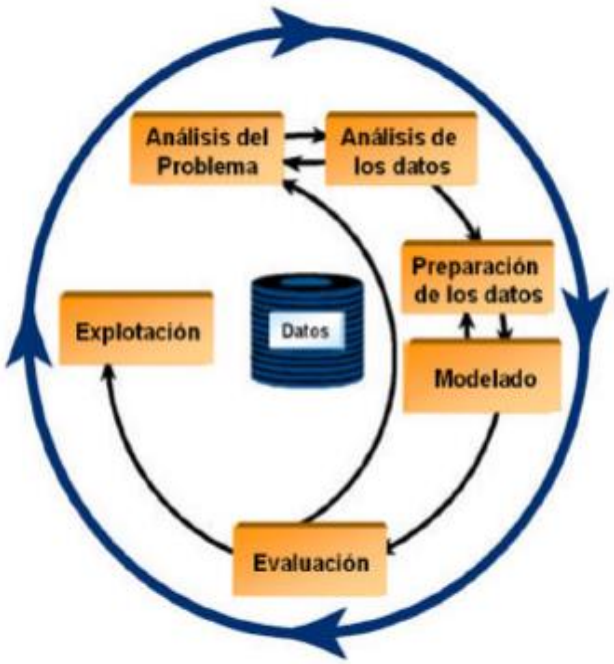


Figura 6. Metodología CRISP-DM
Fuente. Rodríguez Montequín, Teresa; Álvarez Cabal, J. Valeriano; Mesa Fernández, José Manuel; González Valdés, Adolfo. (s.f.).

Como se mostró anteriormente tanto la metodología SEMMA como la metodología CRISP, estructuran los proyectos de Data Mining en fases que se encuentran interrelacionadas, convirtiendo en un proceso iterativo.

La metodología SEMMA se centra más en las características técnicas del desarrollo del proceso, mientras que la metodología CRISP-DM, mantiene una perspectiva más amplia respecto a los objetivos del proyecto.

Las dos metodologías se las resume en la Tabla 1.

	CRISP	SEMMA
Enfoque	Cuenta con un enfoque amplio que aborda todo el proceso de minería de datos, desde la comprensión del negocio hasta el despliegue del modelo.	Tiene un enfoque específico en el desarrollo de modelos de minería de datos, se enfoca particularmente en la manipulación y transformación de los datos antes de aplicar técnicas de modelado.
Fases	Considera 6 fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue.	Considera 5 fases: muestra (sampling), exploración (exploration), modificación (modification), modelado (modeling) y evaluación (assessment).
Interativo	Cuenta con una metodología iterativa que permite retroalimentación y ajustes en cada etapa del proceso. Permite volver a etapas anteriores en función de los resultados y la retroalimentación.	Puede realizar ajustes y modificaciones en las etapas anteriores según los resultados y la retroalimentación obtenida en las fases posteriores.
Origen	Desarrollado por CRISP-DM, una iniciativa conjunta de varias organizaciones y expertos en minería de datos.	Desarrollado por SAS Institute, una empresa de software especializada en análisis de datos y minería de datos.

Tabla 1. Comparación Metodología CRISP VS SEMMA

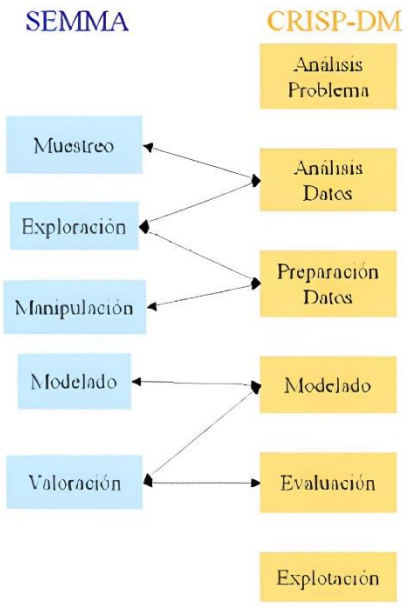


Figura 7. Comparación SEMMA y CRISP-DM
Fuente. Rodríguez Montequín, Teresa; Álvarez Cabal, J. Valeriano; Mesa Fernández, José Manuel; González Valdés, Adolfo. (s.f.).

2.5 Revisión de trabajos similares

El problema de la clasificación de Patentes no es nuevo, la necesidad de automatizar el proceso de clasificación de documentos se lo lleva realizando desde hace algunos años atrás. Entre la documentación revisada la Tabla 2 presenta un resumen con la metodología utilizada, el nivel de clasificación y el mejor resultado del método de clasificación utilizado.

Documento	Método de clasificación	Nivel de clasificación	Idioma	Medida de Evaluación
Aiolli, F., Cardin, R., Sebastiani, F., Sperduti, A.: Preferential text classification: Learning algorithms and evaluation measures. Information Retrieval 12(5), 559–580 (2009)	GPLM (Generalized preference learning model)	Subclase	Inglés	Mejor resultado 0,5298
Chen, Y.L., Chang, Y.C.: A three-phase method for patent classification. Information Processing and Management 48(6), 1017–1030 (2012)	SVM y KNN	Subgrupo	Inglés	Mejor resultado 0,36
Haykin, S.: Neural Networks: A Comprehensive Foundation. Prentice Hall (1994)	SVM	Grupo	Inglés	Mejor resultado 0,30
Teodoro, D., Gobeill, J., Pasche, E., Ruch, P., Vishnyakova, D., Lovis, C.: Automatic IPC encoding and novelty tracking for effective patent mining. In: Proceedings of the 8th NTCIR Workshop Meeting, pp. 309–317. National Institute of Informatics Japan (2010)	KNN	Subclase	Inglés	Mejor resultado 0,68
Verberne, S., D'hondt, E.: Patent classification experiments with the Linguistic Classification System LCS in CLEF-IP 2011. In: Proceedings of CLEF 2011 (Notebook Papers/Labs/Workshop) (2011)	Winnow	Subclase	Inglés	Mejor resultado 0,74

Tabla 2. Revisión de Trabajos Similares

Como se muestra, mientras más específica la clasificación intenta ser, el resultado es más bajo. De la documentación revisada el mejor resultado encontrado es 0,74 obtenido cuando la clasificación se realiza a nivel de subclase. Adicionalmente se evidencia que no existen investigaciones enfocadas al idioma español por lo que este documento aborda una brecha significativa en la investigación de este campo. Muchas de las investigaciones y aplicaciones existentes de Procesamiento de Lenguaje Natural en la Clasificación de Patentes se han centrado en idiomas como el inglés y el chino por lo que al enfocar este documento en el idioma español se amplía el alcance de la protección de propiedad intelectual y fomenta un ecosistema de innovación equitativo y diverso.

3 Capítulo III. Marco metodológico

3.1 Materiales

Los materiales que se usarán en el desarrollo del proyecto se basan en las necesidades que se identifican durante el desarrollo de este.

- **PATENTSCOPE.** Base de datos que proporciona acceso a solicitudes internacionales de patentes.

- **ANACONDA.** Software de distribución libre y abierta de los lenguajes Python y R, utilizada en ciencia de datos, y aprendizaje automático.

- **Tensorflow.** Biblioteca de código abierto para aprendizaje automático a través de un rango de tareas, y desarrollado por Google para satisfacer sus necesidades de sistemas capaces de construir y entrenar redes neuronales para detectar y descifrar patrones y correlaciones.

- **CUDA Deep Neural Network (cuDNN).** Es una biblioteca acelerada por GPU para redes neuronales profundas. cuDNN proporciona implementaciones altamente optimizadas para rutinas estándar como convolución hacia adelante y hacia atrás, atención, agrupación y normalización.

- **Word2vec.** Es una técnica para el procesamiento de lenguaje que utiliza un modelo de red neuronal para aprender asociaciones de palabras a partir de un gran corpus de texto.

- **Python.** Lenguaje de programación que se utilizará para aplicar algoritmos de aprendizaje profundo.

3.2 Métodos

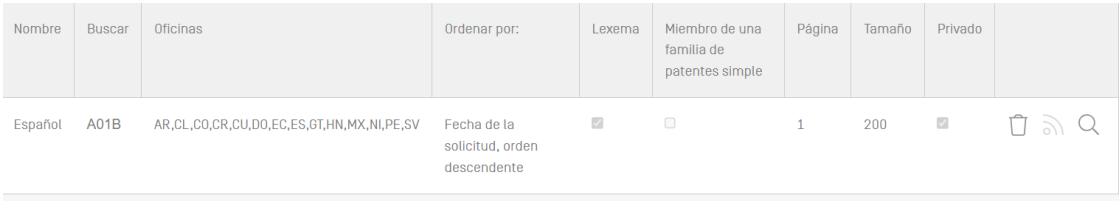
3.2.1 Metodología

La metodología usada para la clasificación de patentes a nivel de subclase, se apoya en la metodología CRISP-DM, la cual proporciona una estructura clara para la planeación y gestión de la minería de datos.

De acuerdo con las etapas que establece la metodología se inicia con la comprensión del negocio y de los datos, para esto se analizará el proceso de registro de patentes en Ecuador y se realizará un acercamiento con los responsables de la clasificación de las patentes para poder comprender la estructura de los documentos y entender el proceso de clasificación, esto con el objetivo de asegurar la validez de los datos que se considerarán.

Durante la segunda etapa se continuará con la preparación de los datos, para lo cual se extraerá la información de Título y Resumen de las Patentes dentro de

la base de datos PATENTSCOPE, donde se usará un filtro de países donde el idioma hablado sea el español.



Nombre	Buscar	Oficinas	Ordenar por:	Lexema	Miembro de una familia de patentes simple	Página	Tamaño	Privado	
Español	A01B	AR,CL,CO,CR,CU,DO,EC,ES,GT,HN,MX,NI,PE,SV	Fecha de la solicitud, orden descendente	☑	☐	1	200	☑	

Figura 8. Captura de pantalla filtros base de datos PATENTSCOPE
Fuente: World Intellectual Property Organization, (2023).

Este proceso se realizará por cada subclase para crear la fuente de datos con la cual se trabajará en los experimentos posteriores. Posteriormente se unirá todos los archivos descargados en un solo archivo, en donde se realizará una exploración de los textos con la finalidad de revisar que el idioma de los textos descargados se encuentre en idioma español, esto debido a que el proyecto se enfoca en la clasificación de documentos que se encuentren en este idioma y demás actividades que permitirán preparar los textos a fin de que los mismos tengan la estructura necesaria para poder aplicar el algoritmo seleccionado.

Durante la tercera etapa se continuará con el modelamiento, donde se pretende aplicar la guía de *"Text Classification"* de Google.

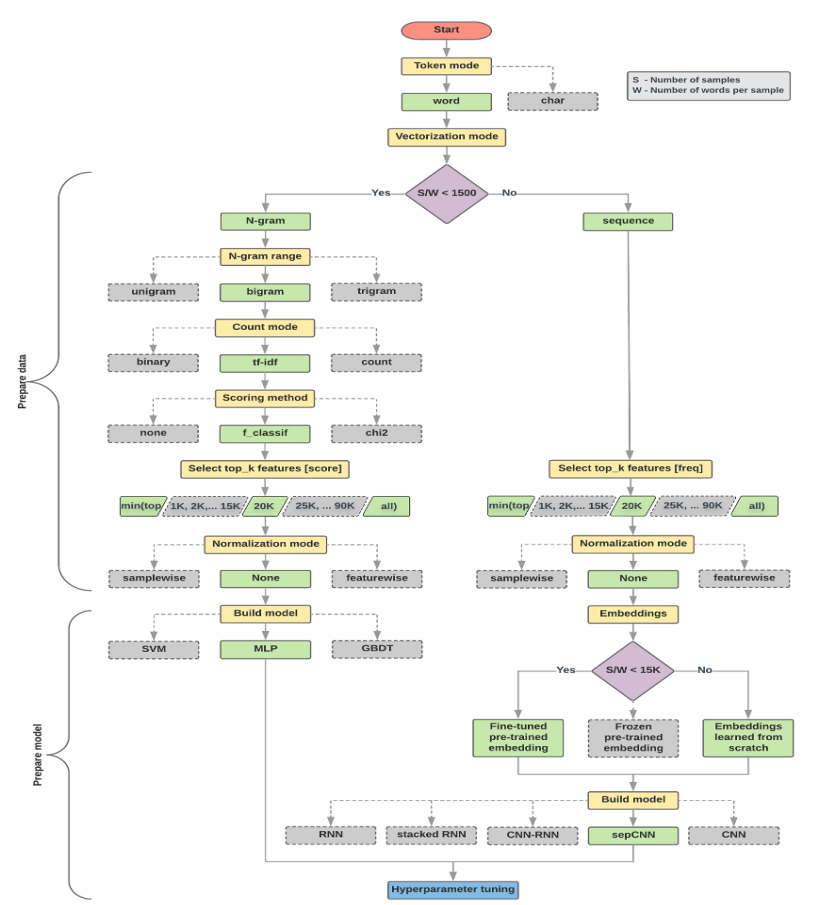


Figura 9. Elección de Modelo
Fuente: Google Developers, (2023).

Donde el diagrama de flujo propone varios modelos según la relación entre el número de muestras y el número de palabras por muestra. Se usará este flujo como punto de partida para la construcción de los experimentos posteriores.

3.2.2 Métodos y técnicas

Para conocer la situación de la clasificación de patentes se usará la técnica de observación y entrevista a través de la cual se espera entender los datos, la documentación y la información que se maneja.

Los datos, como se mencionó anteriormente, se obtendrán directamente de PATENTSCOPE, en formato Excel, para finalmente unir todos los archivos descargados en una sola base de datos en formato *"Comma Separated Values"* (.csv). Para la construcción de la base de datos, solo se usó un filtro de países cuyo idioma oficial es español; no se usó ningún filtro de fecha por lo que la información obtenida corresponde al año 1.919 hasta el mes de agosto del año 2023.

La transformación del texto es una etapa importante al trabajar con textos, y en especial cuando se trabaja con textos de patentes debido al lenguaje técnico que se suele usar para describir las mismas, por lo que se realizará una limpieza y normalización de los textos en Python. Donde se realizan actividades como.

- Revisar el idioma de los textos descargados, esto debido a que el proyecto se enfoca en la clasificación de documentos que se encuentren en el idioma español.
- Eliminar valores duplicados y celdas vacías.
- Normalizar los textos, convirtiendo todo a minúsculas, eliminando puntuación, caracteres especiales, numeraciones irrelevantes.
- Analizar la necesidad de eliminar los *"stop words"*, ya que se puede perder contextos en las oraciones.
- Tokenizar, es el proceso de dividir, para el caso del proyecto, el texto del título y resumen en unidades más pequeñas (para este caso palabras) llamadas *"tokens"*. Estos *"tokens"* determinarán el vocabulario del conjunto de datos.
- Vectorizar, una vez se convierte el texto en *"tokens"* se deben convertir estas palabras en vectores numéricos, para poder usar el modelo elegido.

Una Red Neuronal Convolucional puede aprender de las relaciones de los tokens como modelos de secuencia, por lo que es probables que las palabras del conjunto de datos no sean exclusivas del conjunto. Permitiendo que la Red Neuronal aprenda de la relación entre palabras de otros conjuntos de datos. Esto se puede realizar transfiriendo una incorporación aprendida de otro conjunto de datos a la capa de incorporación del modelo que se quiere entrenar. Google Developers. (2023). Estas incorporaciones se conocen como "***pre-trained embeddings***" y existen varias disponibles como GLove y Word2Vec que han sido entrenadas con varias páginas web y documentos científicos.

Con la información antes mencionada y debido al volumen de datos con los que se cuenta, se han realizado tres experimentos.

1. En el primer experimento, la red neuronal se entrenará únicamente con el vocabulario tokenizado de la base de datos construida donde se permite que la red ajuste todas las ponderaciones.
2. En el segundo experimento se realizará la incrustación de "***pre-trained embeddings***" con los pesos de la capa de incorporación inmovilizados, permitiendo que el resto de la red aprenda.
3. En el tercer experimento se realizará la incrustación de "***pre-trained embeddings***" y se permite que la capa de incorporación también aprenda y realice ajustes a todas las ponderaciones de la red.

Para evaluar los resultados se utilizará la métrica *"**accuracy**"* donde se evaluarán tres resultados: Como se mencionó anteriormente una patente puede estar en varias subclases, es decir las etiquetas no son exclusivas, por lo que el primer resultado evaluará si la etiqueta principal coincide con la probabilidad más alta (top), el segundo resultado evaluará si la etiqueta principal se encuentra dentro de las tres probabilidades más altas (top 3) y finalmente el tercer resultado evaluará si la etiqueta principal se encuentra dentro de las cinco probabilidades más altas (top 5).

Finalmente, en este proyecto además de utilizar el *"**accuracy**"* para evaluar los resultados, también se recurrirá a la evaluación cualitativa en la que se buscará la opinión de gente relacionada al proceso de clasificación.

4 Capítulo IV. Resultados

4.1 Aplicación de técnicas de Minería de Datos

Esta actividad se la ejecutará en base a la metodología CRISP-DM donde se realizará la comprensión de los datos, preparación de los datos, modelado y evaluación.

4.1.1 Recopilación de los datos iniciales

Los datos fueron recopilados a través de la base de datos de PATENTSCOPE, creando un archivo .xlsx por cada subclase. Para la descarga de la información se utilizaron los siguientes filtros:

- **Oficinas.** Argentina, Chile, Colombia, Costa Rica, Cuba, República Dominicana, Ecuador, El Salvador, Guatemala, Honduras, México, Nicaragua, Perú, España.
- **Idiomas.** Español.
- **Fecha.** No se realiza ningún filtro de fecha.

Una vez descargados los archivos .xlsx generados por cada subclase, se unieron en un solo archivo .csv. El detalle de las columnas del archivo generado se presenta en la Figura 9.

```
# Mostrar los nombres de las columnas
nombres_columnas = df.columns.tolist()
rint("Nombres de las columnas:")
for columna in nombres_columnas:
print(columna)
```

```
Nombres de las columnas:
Application Id
Application Date
Publication Date
Country
Title
Abstract
I P C
Subclase
Clase
```

Figura 10. Estructura base de datos

4.1.2 Descripción de los datos

De las diferentes subclases que componen el archivo con el cual se va a trabajar en el proyecto, se describe cada uno de los campos.

- **Application Id.** Contiene el identificador de la solicitud de patente, cada solicitud tiene un número asignado que lo distingue de las demás.
- **Application Date.** Fecha en que se presentó la solicitud de patente, indicando la fecha en la que oficialmente inició el proceso de registro.

- **Publication Date.** Fecha en que se publicó la solicitud de patente, indica la fecha en que la información sobre la patente se hace pública.
- **Country.** País en el que ha sido solicitado la patente. El código sirve para identificar el país.
  - Argentina, AR
  - Chile, CL
  - Colombia, CO
  - Costa Rica, CR
  - Cuba, CU
  - República Dominicana, DO
  - Ecuador, EC
  - El Salvador, SV
  - Guatemala, GT
  - Honduras, HN
  - México, MX
  - Nicaragua, NI
  - Perú, PE
  - España, ES
- **Title.** Título de la invención, proporciona una breve idea sobre lo que trata la patente.
- **Abstract.** Breve resumen de la invención, donde se ofrece una visión general de lo que cubre la patente.
- **I P C.** Clasificación Internacional de Patentes asignada por la oficina de propiedad intelectual.

- **Subclase.** Etiqueta principal dentro de la Clasificación Internacional de Patentes asignada por la oficina de propiedad intelectual.
- **Clase.** Se refiere a la sección general bajo la cual se clasifica la patente dentro de la Clasificación Internacional de Patentes.
  - Sección A. Necesidades corrientes de la vida.
  - Sección Técnicas industriales diversas; transportes.
  - Sección Química; metalúrgica.
  - Sección D. Textiles; papel.
  - Sección E. Construcciones fijas.
  - Sección F. Mecánica; iluminación; calefacción; armamento; voladura.
  - Sección G. Física.
  - Sección H. Electricidad.

4.2 Exploración de los datos

4.2.1 Recopilación de los datos iniciales

Una vez consolidados todos los reportes obtenidos de cada subclase se revisó que la información obtenida, parte desde 1.919 hasta agosto 2.023. Donde se obtuvieron 642 archivos en formato .xls, cada uno con información específica de cada subclase.

En base a la información obtenida se construyó un solo archivo integrado que se usará para el análisis de datos, mismo que permitirá una mayor comprensión de

las relaciones entre las diferentes características de los textos y la etiqueta asignada (subclase).

```
df = pd.read_csv(ruta_archivo)
df.describe
```

```
<bound method NDFrame.describe of         Application Id Application Date Publication Date Country  \
0            ES4902887       13.12.1963       16.03.1963      ES
1            ES4894830       18.11.1963       16.01.1964      ES
2            ES4894835       16.11.1963       16.02.1964      ES
3            ES4894838       16.11.1963       16.01.1964      ES
4            ES4902137       27.12.1962       01.03.1963      ES
...                ...              ...              ...     ...
1089430    MX399830113       21.10.2021       15.12.2021      MX
1089431    MX397713257       13.07.2021       16.08.2021      MX
1089432    MX399961274       08.10.2021       04.01.2022      MX
1089433    MX402867644       02.12.2021       23.06.2022      MX
1089434    CR247677810       13.03.2019       03.05.2019      CR

                                                    Title  \
0        PERFECCIONAMIENTOS EN MÁQUINAS AGRÍCOLAS DE HE...
1        PERFECCIONAMIENTOS EN MÁQUINAS AGRÍCOLAS  DE G...
2                  PERFECCIONAMIENTOS EN MOTOCULTORES
3              PERFECCIONAMIENTOS EN MÁQUINAS AGRÍCOLAS
4        PERFECCIONAMIENTOS INTRODUCIDOS EN LAS GRADAS ...
...                                                    ...
1089430  FABRICACION DE VIA A TRAVES DE SILICIO EN DISP...
1089431           FABRICACION DE UN DISPOSITIVO CUANTICO.
1089432  CUBITS TRANSMON CON ESTRUCTURAS DE CONDENSADOR...
1089433  FABRICACION DE ESTRUCTURAS DE CHIP INVERTIDO T...
1089434  DISPOSITIVO ELECTRÓNICO PARA EL AHORRO DE COMB...

                                              Abstract       I P C  \
0        Perfeccionamientos en máquinas agrícolas de he...        A01B
1        Perfeccionamientos en máquinas agrícolas de gr...        A01B
2        Perfeccionamientos en motocultores, del tipo a...        A01B
3        Perfeccionamientos en máquinas agrícolas, cara...        A01B
4        Perfeccionamientos introducidos en las gradas ...        A01B
...                                                    ...         ...
1089430  En una primera capa superconductora (316) depo...  H10N 69/00
1089431  En una fase de enmascaramiento, se forma un pr...  H10N 69/00
1089432  Un cúbit incluye un sustrato y una primera est...  H10N 69/00
1089433  Se forma un dispositivo informático cuántico (...  H10N 69/00
1089434  Un dispositivo electónico para el ahorro de co...  H99Z 99/00

        Subclase Clase
0           A01B     A
1           A01B     A
2           A01B     A
3           A01B     A
4           A01B     A
...          ...   ...
1089430     H10N     H
1089431     H10N     H
1089432     H10N     H
1089433     H10N     H
1089434     H99Z     H

[1089435 rows x 9 columns]>
```

Figura 11. Resumen base de datos

Como se observa en la Figura 11. El texto recolectado contiene información necesaria para cumplir el objetivo del proyecto, inicialmente se cuenta con un total de 1’089.435 registros. Donde la distribución de patentes por año se presenta en la siguiente figura.

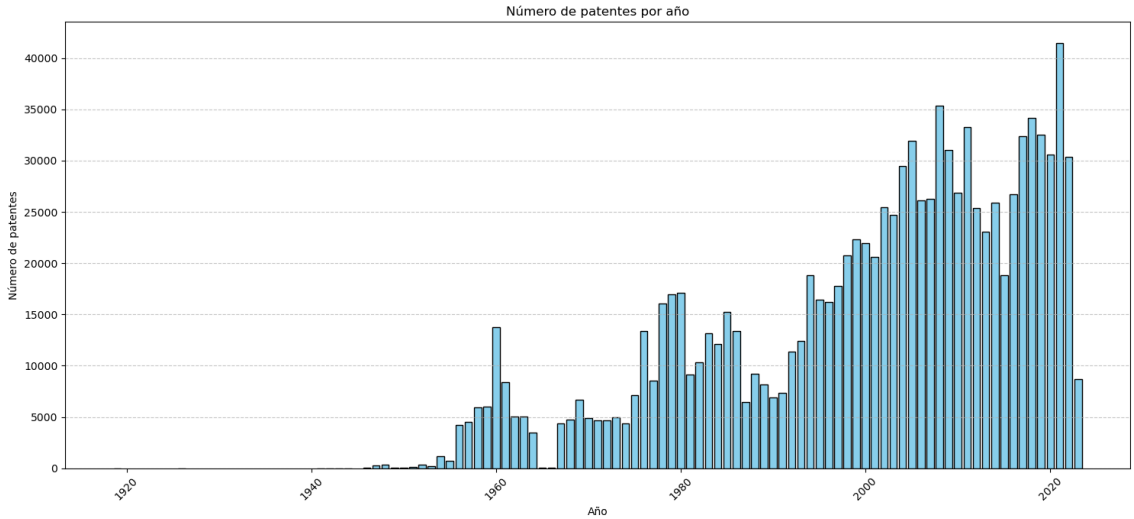


Figura 12. Número de patentes por año

Como se muestra en la Figura 12. La mayor parte de las patentes que componen la base de datos, se encuentran entre los periodos 1.980 y 2.023. Donde, de los países seleccionados, se observa que la mayor parte de las solicitudes publicadas corresponden a España en primer lugar, en segundo lugar, México y en tercer lugar Argentina.

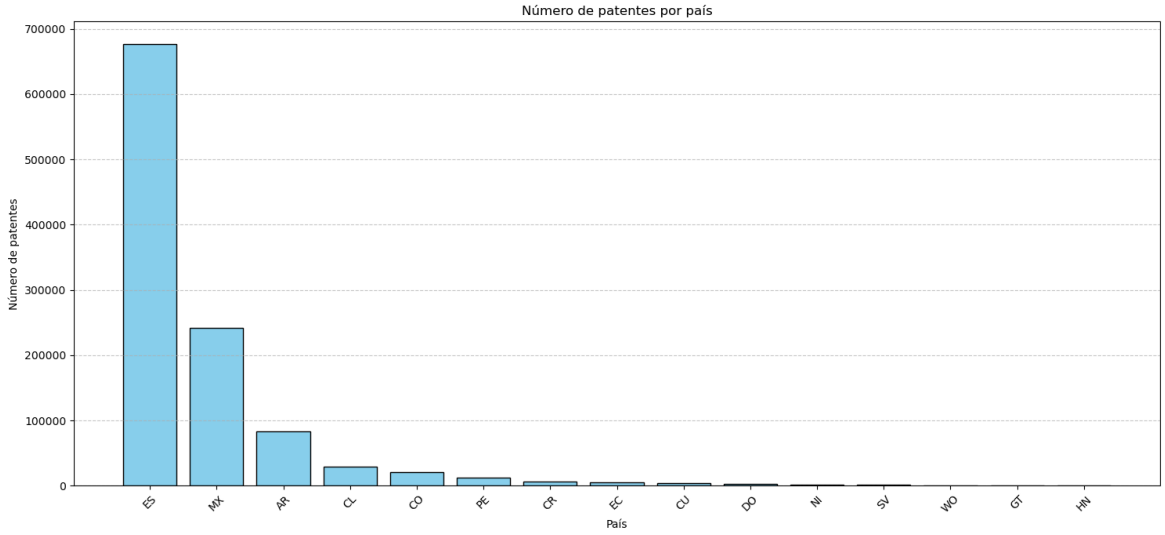


Figura 13. Número de patentes por país

Al trabajar con información etiquetada es importante conocer la distribución de las clases, ya que la recomendación general es que no exista un desequilibrio entre los temas o como en el caso del proyecto de subclases. Es decir, se debe contar con una cantidad comparable de muestras en cada subclase.

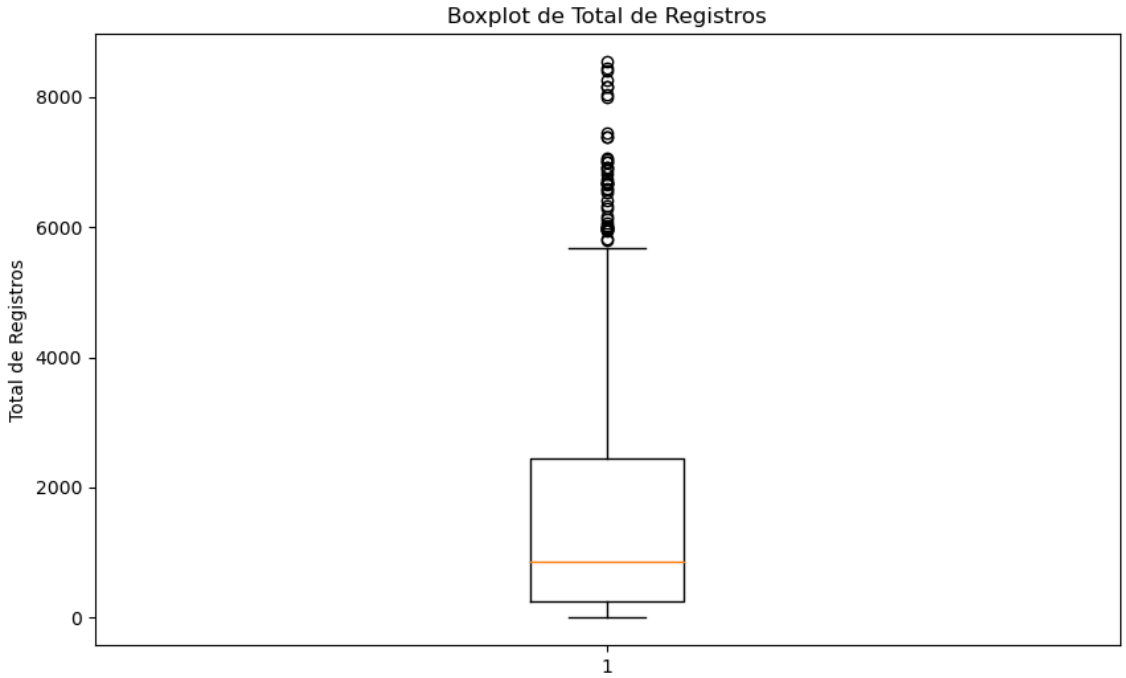


Figura 14. Distribución de las subclases

La figura anterior muestra la distribución de las subclases, es decir el número de ejemplos que se tendrá por cada etiqueta (subclase), como se puede observar existe un desequilibrio de clases que deberá ser analizado.

	Total de Registros
count	642.000000
mean	1696.939252
std	1957.336177
min	1.000000
25%	250.250000
50%	859.500000
75%	2451.750000
max	8537.000000

Figura 15. Descripción de la distribución de las subclases

De acuerdo con la metodología detallada en el punto 3.2.1 donde se menciona el flujo de trabajo que se utilizará, es importante obtener algunas métricas clave que serán utilizadas para determinar el tipo de modelo que se utilizará durante el proceso de entrenamiento. Para lo cual se necesita conocer:

- **Ratio de número de ejemplos sobre el número de palabras por ejemplo (S/W) de Title.** Para el cálculo del ratio de número de ejemplos sobre el número de palabras por ejemplo (S/W) de la columna *"Title"* se obtuvo el siguiente resultado.

```
df['word_count'] = df['Title'].apply(lambda x: len(str(x).split()))

# Calculando S y W
S = len(df)
W = df['word_count'].mean()

# Calculando S/W
SW_ratio = S/W

print(f"S/W ratio es: {SW_ratio:.2f}")
```

```
S/W ratio es: 107918.24
```

Figura 16. S/W de Title

- **Ratio de número de ejemplos sobre el número de palabras por ejemplo (S/W) de Abstract.** Para el cálculo del ratio de número de ejemplos sobre el número de palabras por ejemplo (S/W) de la columna *"Abstract"* se obtuvo el siguiente resultado.

```python
df['word_count'] = df['Abstract'].apply(lambda x: len(str(x).split()))

# Calculando S y W
S = len(df)
W = df['word_count'].mean()

# Calculando S/W
SW_ratio = S/W

print(f"S/W ratio es: {SW_ratio:.2f}")
```

```
S/W ratio es: 7271.33
```

Figura 17. S/W de Abstract

- **Distribución de número de palabras en el texto del conjunto de datos.** Para revisar la distribución del número de palabras en el conjunto de datos, se lo realizó mediante un gráfico de barras, mismo que se presenta en la Figura 18.

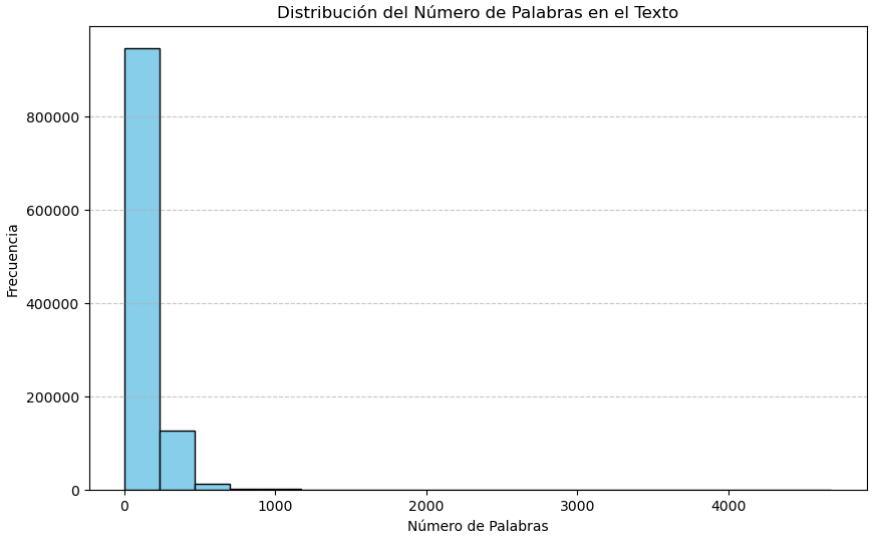


Figura 18. Distribución del número de palabras

4.2.2 Verificación de la calidad de los datos

La base de datos recolectada posee gran cantidad de información por lo que es importante realizar algunas verificaciones de esta; iniciando con la revisión de valores faltantes o nulos, esto con la finalidad que la base de datos pueda ser utilizada de manera adecuada y cumpla con el objetivo del proyecto. De le revisión realizada se obtuvieron los siguientes resultados.

```
# contando registros nulos en cada columna
null_count = df.isnull().sum()
print(null_count)
```

```
Application Id         0
Application Date     393
Publication Date       0
Country                0
Title                  5
Abstract            1477
I P C                  0
Subclase               0
Clase                  0
dtype: int64
```

Figura 19. Valores faltantes o nulos

De acuerdo con los resultados obtenidos se puede observar la existencia de valores faltantes en las variables *"Application Date"* y *"Abstract".* Al ser la variable *"Abstract"* importante para el desarrollo del proyecto, la misma debe ser tratada. Mientras que los valores faltantes o nulos encontrados en la *variable "Application Date"* no tienen ninguna afectación al mismo.

Finalmente se realizaron algunas revisiones aleatorias al conjunto de datos en donde se encontró que, a pesar de haber realizado el filtro de países donde el idioma principal es el español, existen patentes en diferentes idiomas por lo que al plantearse en el presente proyecto analizar los textos de la documentación de las patentes y extraer características claves que permitan ser utilizadas para categorizar los documentos dentro de las subclases de la Clasificación internacional de patentes en el idioma español, se deberá realizar un tratamiento a la selección del idioma dentro del conjunto de datos.

4.3 Preparación de datos

Para una adecuada preparación de los datos es necesario realizar actividades como selección de estos, limpieza del conjunto de datos, construcción de nuevos datos en caso de ser necesario. Estas actividades no se realizaron de manera secuencial, sino que se las fueron realizando conforme se identificaba la necesidad de transformación o limpieza.

4.3.1 Selección de datos

Para la ejecución de esta actividad se consideran los objetivos del proyecto definidos en el punto 1.1.1 y 1.1.2. Seleccionando del conjunto de datos las variables con las que se trabajará, siendo "*Application Id", "Title", "Abstract", "IPC", "Subclse", "Clase"*.

4.3.2 Limpieza de datos

Las actividades requeridas para la limpieza de los datos fueron realizadas en Python mediante Jupyter Notebook; todas las actividades que se detallan posteriormente se las realizó con la finalidad de exportar un conjunto de datos en formato .csv que integre toda la información necesaria para el posterior análisis. Durante el proceso de limpieza de datos se identificaron y ejecutaron varias actividades clave para garantizar la calidad del conjunto de datos como:

- **Valores faltantes/nulos.** Tal y como se observó en la Figura 14. existen 1.477 valores faltantes o nulos que son eliminados de la variable "Abstract" dentro del conjunto de datos.
- **Filtrado de idioma.** Conforme a los señalado anteriormente, de las revisiones aleatorias realizada al conjunto de datos se encontró que existen idiomas diferentes al español en el conjunto de datos por lo que para solucionar este problema se utilizó el paquete de Python llamado "langdetect", que permite detectar 55 idiomas distintos. Para lo cual se desarrolló el siguiente código.

```
df = pd.read_csv(archivo_csv)

def detectar_idioma(row):
    try:
        # Detectar el idioma del texto usando langdetect
        idioma = detect(row['Abstract'])

        # Agregar el idioma detectado a la columna 'Idioma Detectado'
        row['Idioma Detectado'] = idioma

        return row
    except Exception as e:
        print(f"Error al detectar idioma: {e}")
        row['Idioma Detectado'] = 'desconocido'
        return row

# Aplicar la función a cada fila del DataFrame
df = df.apply(detectar_idioma, axis=1)
```

Figura 20. Detección de idioma

- **Conversión a minúsculas.** Una vez filtrados los textos que contenían únicamente información en español del conjunto de datos, se transformaron los textos de la columna *"Title"* y *"Abstract"* a minúsculas para lo cual se usó el siguiente código.

```
# Convertir las columnas 'Title' y 'Abstract' a minúsculas
df = pd.read_csv(ruta_archivo)
df['Title'] = df['Title'].str.lower()
df['Abstract'] = df['Abstract'].str.lower()
```

Figura 21. Convertir a minúsculas

Se verifica que los textos se encuentren en minúsculas teniendo como resultado el mostrado en la Figura 22.

```
print(df.head())
```

```
  Application Id Application Date Publication Date Country  \
0      ES4902887       13.12.1963       16.03.1963      ES
1      ES4894830       18.11.1963       16.01.1964      ES
2      ES4894835       16.11.1963       16.02.1964      ES
3      ES4894838       16.11.1963       16.01.1964      ES
4      ES4902137       27.12.1962       01.03.1963      ES

                                          Title  \
0  perfeccionamientos en máquinas agrícolas de he...
1  perfeccionamientos en máquinas agrícolas  de g...
2                perfeccionamientos en motocultores
3          perfeccionamientos en máquinas agrícolas
4  perfeccionamientos introducidos en las gradas ...

                                        Abstract I P C Subclase Clase
0  perfeccionamientos en máquinas agrícolas de he...  A01B     A01B     A
1  perfeccionamientos en máquinas agrícolas de gr...  A01B     A01B     A
2  perfeccionamientos en motocultores, del tipo a...  A01B     A01B     A
3  perfeccionamientos en máquinas agrícolas, cara...  A01B     A01B     A
4  perfeccionamientos introducidos en las gradas ...  A01B     A01B     A
```

Figura 22. Revisión conversión a minúsculas

- **Textos en contenedores.** De las revisiones aleatorias realizadas se pudo observar que existe texto sin relevancia dentro del texto recolectado que se encuentra dentro de símbolos como, (), [], {}, <>. Para lo cual se usó el siguiente código.

```
# Eliminar contenido dentro de paréntesis
df['Title'] = df['Title'].str.replace(r'\(.*?\)', '', regex=True).str.strip()
df['Abstract'] = df['Abstract'].str.replace(r'\(.*?\)', '', regex=True).str.strip()
```

Figura 23. Eliminar paréntesis y su contenido

```
# Eliminar contenido dentro de corchetes
df['Title'] = df['Title'].str.replace(r'\[.*?\]', '', regex=True).str.strip()
df['Abstract'] = df['Abstract'].str.replace(r'\[.*?\]', '', regex=True).str.strip()
```

Figura 24. Eliminar corchetes y su contenido

```
# Eliminar contenido dentro de llaves
df['Title'] = df['Title'].str.replace(r'\{.*?\}', '', regex=True).str.strip()
df['Abstract'] = df['Abstract'].str.replace(r'\{.*?\}', '', regex=True).str.strip()
```

Figura 25. Eliminar llaves y su contenido

```
# Eliminar contenido dentro de <>
df['Title'] = df['Title'].str.replace(r'\<.*?\>', '', regex=True).str.strip()
df['Abstract'] = df['Abstract'].str.replace(r'\<.*?\>', '', regex=True).str.strip()
```

Figura 26. Eliminar <> y su contenido

Las líneas de código anteriormente mostradas permitieron eliminar cualquier texto dentro de símbolos como, (), [], {}, <> (incluidos los simbolos) en las columnas 'Title' y 'Abstract' del conjunto datos. De la misma manera se detectó que existen listas de números y letras además de textos que no agregan valor a la información que será analizada.

```
# Lista de palabras/patrones a eliminar
palabras_a_eliminar = ['a\)', 'b\)', 'c\)', 'd\)', 'e\)', 'f\)', 'g\)', 'h\)', 'i\)', 'j\)', 'k\)', 'l\)', 'm\)', 'n\)', 'o\)',
                       'p\)','q\)', 'r\)', 's\)', 't\)', 'u\)', 'v\)', 'w\)', 'x\)', 'y\)', 'z\)',
                       '1\)', '2\)', '3\)', '4\)', '5\)', '6\)', '7\)', '8\)', '9\)', '10\)',
                       '1\.', '2\.', '3\.', '4\.', '5\.', '6\.', '7\.', '8\.', '9\.', '10\.',
                       '1\,', '2\,', '3\,', '4\,', '5\,', '6\,', '7\,', '8\,', '9\,', '10\,',
                       'i\.', 'ii\.', 'iii\.', 'iv\.', 'v\.', 'vi\.', 'vii\.', 'viii\.', 'ix\.', 'x\.']

# Crear una expresión regular que busque estas palabras rodeadas de posibles espacios en blanco
patron = r'\s*(' + '|'.join(palabras_a_eliminar) + r')\s*'

# Eliminar las palabras/patrones de la columna 'Title'
df['Title'] = df['Title'].str.replace(patron, ' ', regex=True).str.strip()

# Eliminar las palabras/patrones de la columna 'Abstract'
df['Abstract'] = df['Abstract'].str.replace(patron, ' ', regex=True).str.strip()
```

Figura 27. Eliminar listas de números y letras

```
# Lista de palabras/patrones a eliminar
palabras_a_eliminar = ['1', '2', '3', '4', '5', '6', '7', '8', '9', '10', 'anexos', 'ver anexos', 'fig.', 'resumen',
                       'la presente invención se refiere a', 'la presente invención', 'la invención se refiere a',
                       'la invención', 'invención']

# Crear una expresión regular que busque estas palabras rodeadas de posibles espacios en blanco
patron = r'\s*(' + '|'.join(palabras_a_eliminar) + r')\s*'

# Eliminar las palabras/patrones de la columna 'Title'
df['Title'] = df['Title'].str.replace(patron, ' ', regex=True).str.strip()

# Eliminar las palabras/patrones de la columna 'Abstract'
df['Abstract'] = df['Abstract'].str.replace(patron, ' ', regex=True).str.strip()
```

Figura 28. Eliminar palabras

Finalmente se eliminan caracteres especiales de los textos analizados.

```
#Eliminar caracteres especiales
caracteres_especiales = r'[!@#$%^&*()\-=+\[\]{}|:;\"<>,.?/]'

df['Title'] = df['Title'].str.replace(caracteres_especiales, ' ', regex=True).str.strip()
df['Abstract'] = df['Abstract'].str.replace(caracteres_especiales, ' ', regex=True).str.strip()
```

Figura 29. Eliminar caracteres especiales

- **Outliers.** Una vez se eliminaron los valores vacíos y/o nulos además de filtrar los textos que se encuentran únicamente en el idioma español. Se realiza nuevamente la revisión de outliers. Teniendo el siguiente resultado.

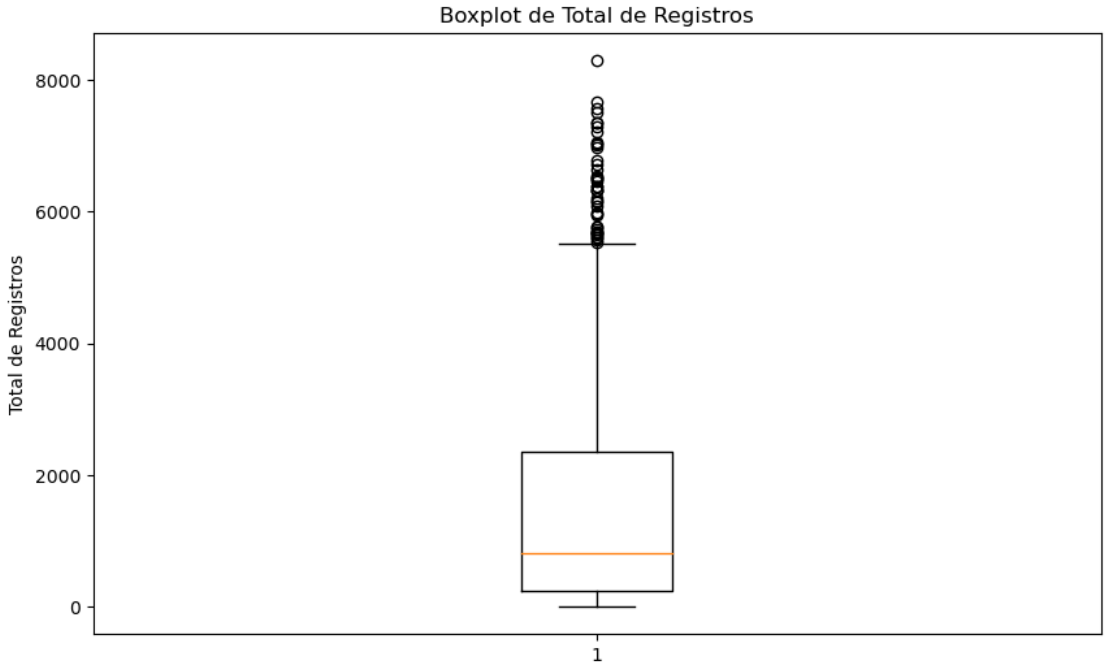


Figura 30. Distribución de las subclases

	Total de Registros
count	641.000000
mean	1607.000000
std	1843.795046
min	1.000000
25%	240.000000
50%	818.000000
75%	2355.000000
max	8291.000000

Figura 31. Descripción de la distribución de las subclases

De acuerdo a los resultados obtenidos en la exploración del conjunto de datos inicial, presentados en la Figura 14 y 15. Se realiza una comparación del conjunto de datos inicial y el procesado, los mismos que se presentan en la Tabla 3.

	Conjunto de datos inicial	Conjunto de datos procesado
count	642	641
mean	1.696,94	1.607,00
std	1.957,34	1.843,80
min	1	1
25%	250,25	240,00
50%	859,50	818,00
75%	2.451,75	2355,00
max	8.537,00	8291,00
Total de registros	1’089.435,00	1’030.087,00

Tabla 3. Comparación conjunto de datos inicial y procesado

Con la finalidad de mitigar el desbalance existente entre las etiquetas (sublcases) del conjunto de datos, se trabajará con la información donde se encuentra concentrado el mayor número de registros, que corresponde a los valores entre el 25% y 75% del conjunto de datos procesado. Por lo que se tomaron las siguientes medidas.

a. Se eliminaron las subclases que tengan menos de 240 registros.
b. Las subclases que cuenten con más de 2.355 registros se realizará una selección al azar de 2.355 registros para que el máximo de registros por subclase sea el antes señalado.

Para ejecutar las medidas señaladas anteriormente se utilizó el siguiente código.

```
# Primero, filtramos las subclases con menos de 240 conteos
conteos_subclase = df['Subclase'].value_counts()
subclases_filtradas = conteos_subclase[conteos_subclase >= 240].index
df_filtrado = df[df['Subclase'].isin(subclases_filtradas)]

# Luego, seleccionamos al azar 2355 registros de las subclases con más de 2355 conteos
def seleccionar_aleatoriamente(subdf):
    if len(subdf) > 2355:
        return subdf.sample(n=2355, random_state=42)
    return subdf
```

Figura 32. Filtrado de conjunto de datos

Una vez realizadas las acciones anteriores se obtuvo el siguiente resultado.

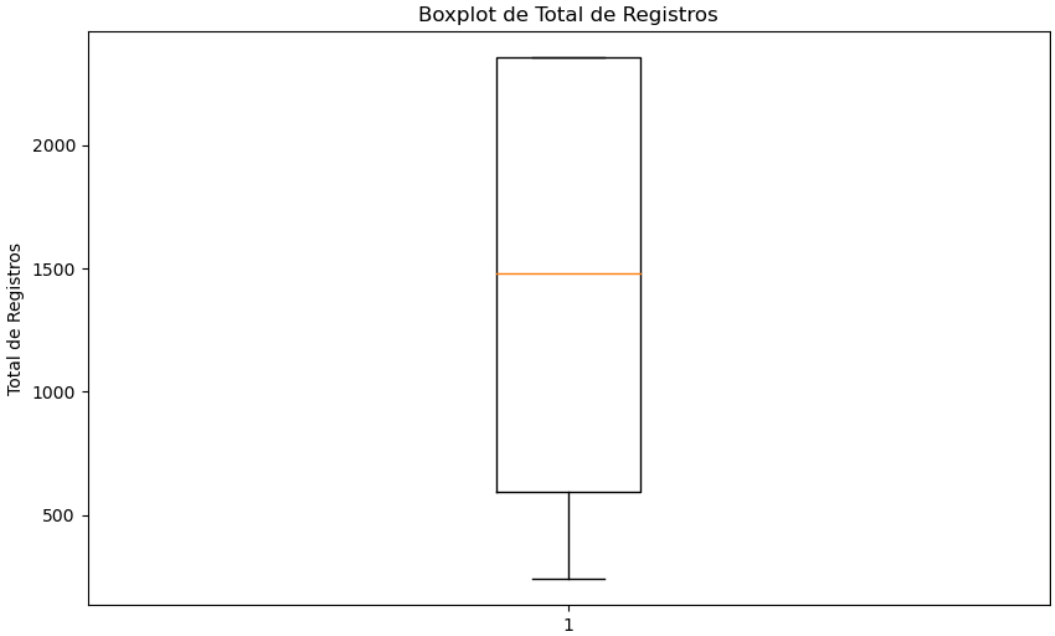


Figura 33. Distribución de las subclases conjunto de datos filtrado

	Total de Registros
count	481.000000
mean	1444.313929
std	808.012334
min	240.000000
25%	596.000000
50%	1479.000000
75%	2355.000000
max	2355.000000

Figura 34. Descripción de la distribución de las subclases del conjunto de datos filtrado

Como se puede observar en la distribución de las subclases, durante la limpieza y filtrado se han perdido 161 etiquetas. El resumen de los dos conjuntos de datos se presenta en la Tabla 4.

	Conjunto de datos procesado	Conjunto de datos filtrado
count	641	481
mean	1.607,00	1.444,31
std	1.843,80	808,01
min	1	240
25%	240,00	596
50%	818,00	1.479,00
75%	2.355,00	2.355,00
max	8.291,00	2.355,00
Total de registros	1’030.087,00	694.715,00

Tabla 4. Comparación conjunto de datos procesado y filtrado

4.3.3 Construcción de datos.

Con la finalidad de poder entrenar el modelo con la mayor cantidad de texto disponible dentro del conjunto de datos, se creó una nueva variable dentro del conjunto de datos llamada “Texto”, misma que corresponde a la unión de las características “Title” y “Abstract” , para lo cual se usó la siguiente línea de código.

```
#Crear una nueva columna que combine el título y abstract
df['Texto'] = df['Title'] + ' ' + df['Abstract']
```

Figura 35. Construcción de columna Texto

Se revisa que esté creada la nueva columna.

```
df = pd.read_csv(ruta_archivo)
print(df.head())
```

```
  Application Id Application Date Publication Date Country  \
0      ES5262911       27.11.1970       01.02.1971      ES
1    ES194415764       08.04.2009       28.02.2017      ES
2      ES5183363       06.04.1983       16.10.1983      ES
3     ES31283499       07.04.1978       16.11.1957      ES
4     CO10399651       29.12.1993       07.06.1995      CO

                                          Title  \
0  un mango para barras de palas  horquillas y pa...
1  disco para uso agrario  concretamente usado pa...
2       maquina extractora de piedra  perfeccionada
3  apero oscilante para labores agricolas entre c...
4   dispositivo oleo neumatico de control de cargas

                                        Abstract                  I P C  \
0  un mango para barras de palas  horquillas y pa...                A01B 1/22
1  disco metálico previsto específicamente para u...   A01B 15/16; A01B 23/06
2  máquina extractora de piedras perfeccionada  q...              A01B 43/00
3  apero oscilante para labores agrícolas entre c...             A01B /; A01B
4  dispositivo oleo neumatico de control de carga...  A01B 63/114; A01B 15/20

  Subclase Clase                                          Texto
0     A01B     A  un mango para barras de palas  horquillas y pa...
1     A01B     A  disco para uso agrario  concretamente usado pa...
2     A01B     A  maquina extractora de piedra  perfeccionada má...
3     A01B     A  apero oscilante para labores agricolas entre c...
4     A01B     A  dispositivo oleo neumatico de control de carga...
```

Figura 36. Revisión de columna Texto

4.3.4 Formateo de datos.

Para poder entrenar el modelo es necesario transformar los textos a un formato que el modelo pueda entender. Por lo que se han realizado las siguientes acciones.

- **Mezcla de datos.** Es necesario revisar que el conjunto de datos que se usará para el entrenamiento no tenga un orden específico, ya que al momento de dividir el conjunto de datos en entrenamiento y validación no representará la distribución general de los datos. Como se puede observar en la Figura 36. Existe un orden determinado del conjunto de datos que está dado por la "Subclase". Para solucionar este problema se utilizó la siguiente línea de código.

```
#Mezclar datos, en la base se encuentran oredenados.
df = df.sample(frac=1, random_state=52).reset_index(drop=True)
```

Figura 37. Mezcla de datos

- **Creación de Vocabulario.** Con la finalidad de cumplir con el objetivo del proyecto se utilizará todo el vocabulario generado en la nueva variable creada llamada "Texto" es posible optimizar el proceso de aprendizaje del modelo descartando palabras que son poco frecuentes o irrelevantes. En este sentido se exploró el vocabulario generado teniendo los resultados presentados a continuación.

```
df = pd.read_csv(ruta_archivo)
print(df.head())
```

```
  Word  Frequency
0   de   11826700
1   la    3872121
2   un    3194395
3    y    2946998
4   en    2876655
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 411446 entries, 0 to 411445
Data columns (total 2 columns):
 #   Column     Non-Null Count   Dtype
---  ------     --------------   -----
 0   Word       411444 non-null  object
 1   Frequency  411446 non-null  int64
dtypes: int64(1), object(1)
memory usage: 6.3+ MB
```

Figura 38. Creación de vocabulario

Como se observan existen 411.446 palabras en la nueva variable "Texto"; no se ha considerado el eliminar las llamadas "stop words" debido a que se puede perder información importante que pueden caracterizar los

textos de ciertas subclases en particular aquellas que utilicen elementos de la tabla periódica para describir las patentes, ya que realizando una exploración de la frecuencia de palabras se observó que varios elementos de la tabla periódica y sus combinaciones pueden ser confundidos como “stop words” como por ejemplo:

- La. Lantano
- O. Oxígeno
- Al. Aluminio
- Y. Itrio
- Es. Einstenio
- Si. Silicio

- **Asignación de Token.** Modelos de aprendizaje profundo como las Redes Neuronales Convolucionales pueden deducir el significado de una palabra a partir del orden de estas, por lo que realizar una tokenización secuencial es importante para el entrenamiento del modelo.

- **Vectorización y normalización.** Una vez realizada la tokenización secuencial se debe convertir las secuencias en vectores numéricos y normalizarlos ya que los modelos de secuencia a menudo tienen una capa de incorporación como su primera capa. Esta capa aprende a convertir secuencias de índices de palabras en vectores de incorporación de palabras durante el proceso de entrenamiento, de modo que cada índice de palabras se asigne a un vector denso de valores reales que representan la ubicación de esa palabra en el espacio semántico. Google Developers. (2023).

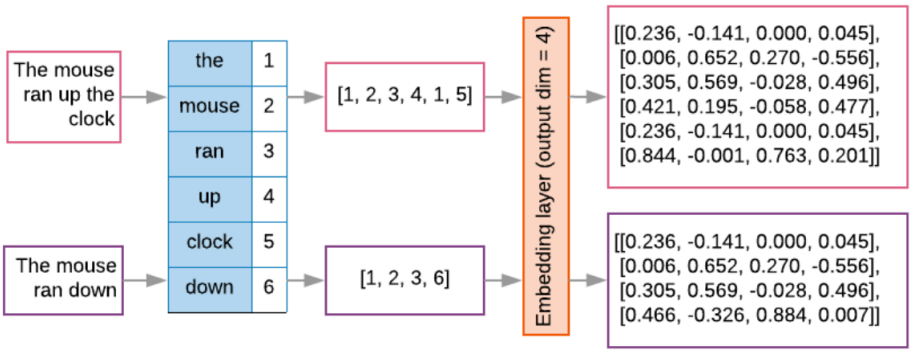


Figura 39. Capa de incorporación
Fuente: Google Developers, (2023).

Como se observó en la Figura 38. se cuenta con un vocabulario total de 411.446 palabras donde se limitó el largo de los textos de cada uno de los registros a 400 palabras, conforme a la distribución obtenida en la Figura 18. Mientras que el número de dimensiones por palabra será de 300; existen diferentes opiniones en cuanto al número de dimensiones recomendadas por palabra que van desde las 100 hasta las 400. Debido a la capacidad computacional con la que se cuenta se mantendrá el número de dimensiones en 300.

- **Codificación de etiquetas.** Como se indicó anteriormente es necesario cambiar los textos a números para poder entrenar y evaluar un modelo por lo que las etiquetas también deben ser transformadas a números. Para esto se utilizó la siguiente línea de código.

```
# Codificar etiquetas
etiqueta = LabelEncoder()
df['etiqueta'] = etiqueta.fit_transform(df['Subclase'])
map_etiqueta = dict(zip(etiqueta.classes_, etiqueta.transform(etiqueta.classes_)))
```

Figura 40. Codificación de etiquetas

Obteniendo el siguiente resultado.

```
# Mostrar mapeo de etiquetas
print(map_etiqueta)
```

```
{'A01B': 0, 'A01C': 1, 'A01D': 2, 'A01F': 3, 'A01G': 4, 'A01H': 5, 'A01J': 6, 'A01K': 7, 'A01M': 8, 'A01N': 9, 'A01P': 10, 'A21
B': 11, 'A21C': 12, 'A21D': 13, 'A22B': 14, 'A22C': 15, 'A23B': 16, 'A23C': 17, 'A23D': 18, 'A23F': 19, 'A23G': 20, 'A23J': 21,
'A23K': 22, 'A23L': 23, 'A23N': 24, 'A23P': 25, 'A24B': 26, 'A24C': 27, 'A24D': 28, 'A24F': 29, 'A41B': 30, 'A41C': 31, 'A41D':
32, 'A41F': 33, 'A41G': 34, 'A41H': 35, 'A42B': 36, 'A43B': 37, 'A43C': 38, 'A43D': 39, 'A44B': 40, 'A44C': 41, 'A45B': 42, 'A4
5C': 43, 'A45D': 44, 'A45F': 45, 'A46B': 46, 'A47B': 47, 'A47C': 48, 'A47D': 49, 'A47F': 50, 'A47G': 51, 'A47H': 52, 'A47J': 5
3, 'A47K': 54, 'A47L': 55, 'A61B': 56, 'A61C': 57, 'A61D': 58, 'A61F': 59, 'A61G': 60, 'A61H': 61, 'A61J': 62, 'A61K': 63, 'A61
L': 64, 'A61M': 65, 'A61N': 66, 'A61P': 67, 'A61Q': 68, 'A62B': 69, 'A62C': 70, 'A62D': 71, 'A63B': 72, 'A63C': 73, 'A63D': 74,
'A63F': 75, 'A63G': 76, 'A63H': 77, 'B01D': 78, 'B01F': 79, 'B01J': 80, 'B01L': 81, 'B02C': 82, 'B03C': 83, 'B03D': 84, 'B04B':
85, 'B04C': 86, 'B05B': 87, 'B05C': 88, 'B05D': 89, 'B06B': 90, 'B07B': 91, 'B07C': 92, 'B08B': 93, 'B09B': 94, 'B21B': 95, 'B2
1C': 96, 'B21D': 97, 'B21F': 98, 'B21H': 99, 'B21J': 100, 'B21K': 101, 'B22C': 102, 'B22D': 103, 'B22F': 104, 'B23B': 105, 'B23
C': 106, 'B23D': 107, 'B23K': 108, 'B23P': 109, 'B23Q': 110, 'B24B': 111, 'B24C': 112, 'B24D': 113, 'B25B': 114, 'B25C': 115,
'B25D': 116, 'B25G': 117, 'B25H': 118, 'B25J': 119, 'B26B': 120, 'B26D': 121, 'B26F': 122, 'B27B': 123, 'B27C': 124, 'B27F': 12
5, 'B27G': 126, 'B27K': 127, 'B27L': 128, 'B27M': 129, 'B27N': 130, 'B28B': 131, 'B28C': 132, 'B28D': 133, 'B29B': 134, 'B29C':
135, 'B29D': 136, 'B29K': 137, 'B30B': 138, 'B31B': 139, 'B31D': 140, 'B31F': 141, 'B32B': 142, 'B41F': 143, 'B41J': 144, 'B41
M': 145, 'B41N': 146, 'B42C': 147, 'B42D': 148, 'B42F': 149, 'B43K': 150, 'B43L': 151, 'B43M': 152, 'B44B': 153, 'B44C': 154,
'B44D': 155, 'B44F': 156, 'B60B': 157, 'B60C': 158, 'B60D': 159, 'B60G': 160, 'B60H': 161, 'B60J': 162, 'B60K': 163, 'B60L': 16
4, 'B60P': 165, 'B60Q': 166, 'B60R': 167, 'B60S': 168, 'B60T': 169, 'B60W': 170, 'B61B': 171, 'B61C': 172, 'B61D': 173, 'B61F':
174, 'B61G': 175, 'B61L': 176, 'B62B': 177, 'B62D': 178, 'B62H': 179, 'B62J': 180, 'B62K': 181, 'B62M': 182, 'B63B': 183, 'B63
C': 184, 'B63H': 185, 'B64C': 186, 'B64D': 187, 'B64F': 188, 'B64G': 189, 'B65B': 190, 'B65C': 191, 'B65D': 192, 'B65F': 193,
'B65G': 194, 'B65H': 195, 'B66B': 196, 'B66C': 197, 'B66D': 198, 'B66F': 199, 'B67B': 200, 'B67C': 201, 'B67D': 202, 'B68G': 20
3, 'C01B': 204, 'C01C': 205, 'C01D': 206, 'C01F': 207, 'C01G': 208, 'C02F': 209, 'C03B': 210, 'C03C': 211, 'C04B': 212, 'C05C':
213, 'C05F': 214, 'C05G': 215, 'C06B': 216, 'C07B': 217, 'C07C': 218, 'C07D': 219, 'C07F': 220, 'C07H': 221, 'C07J': 222, 'C07
K': 223, 'C08B': 224, 'C08C': 225, 'C08F': 226, 'C08G': 227, 'C08J': 228, 'C08K': 229, 'C08L': 230, 'C09B': 231, 'C09C': 232,
'C09D': 233, 'C09J': 234, 'C09K': 235, 'C10B': 236, 'C10G': 237, 'C10J': 238, 'C10L': 239, 'C10M': 240, 'C11B': 241, 'C11C': 24
2, 'C11D': 243, 'C12C': 244, 'C12G': 245, 'C12M': 246, 'C12N': 247, 'C12P': 248, 'C12Q': 249, 'C12R': 250, 'C14B': 251, 'C14C':
252, 'C21B': 253, 'C21C': 254, 'C21D': 255, 'C22B': 256, 'C22C': 257, 'C22F': 258, 'C23C': 259, 'C23F': 260, 'C23G': 261, 'C25
B': 262, 'C25C': 263, 'C25D': 264, 'C30B': 265, 'D01D': 266, 'D01F': 267, 'D01H': 268, 'D02G': 269, 'D03C': 270, 'D03D': 271,
'D04B': 272, 'D04H': 273, 'D05B': 274, 'D06B': 275, 'D06C': 276, 'D06F': 277, 'D06L': 278, 'D06M': 279, 'D06N': 280, 'D06P': 28
1, 'D07B': 282, 'D21B': 283, 'D21C': 284, 'D21D': 285, 'D21F': 286, 'D21H': 287, 'E01B': 288, 'E01C': 289, 'E01D': 290, 'E01F':
291, 'E01H': 292, 'E02B': 293, 'E02D': 294, 'E02F': 295, 'E03B': 296, 'E03C': 297, 'E03D': 298, 'E03F': 299, 'E04B': 300, 'E04
C': 301, 'E04D': 302, 'E04F': 303, 'E04G': 304, 'E04H': 305, 'E05B': 306, 'E05C': 307, 'E05D': 308, 'E05F': 309, 'E05G': 310,
'E06B': 311, 'E06C': 312, 'E21B': 313, 'E21C': 314, 'E21D': 315, 'F01B': 316, 'F01C': 317, 'F01D': 318, 'F01K': 319, 'F01L': 32
0, 'F01M': 321, 'F01N': 322, 'F01P': 323, 'F02B': 324, 'F02C': 325, 'F02D': 326, 'F02F': 327, 'F02K': 328, 'F02M': 329, 'F02N':
330, 'F02P': 331, 'F03B': 332, 'F03D': 333, 'F03G': 334, 'F04B': 335, 'F04C': 336, 'F04D': 337, 'F04F': 338, 'F15B': 339, 'F16
B': 340, 'F16C': 341, 'F16D': 342, 'F16F': 343, 'F16G': 344, 'F16H': 345, 'F16J': 346, 'F16K': 347, 'F16L': 348, 'F16M': 349,
'F16N': 350, 'F16S': 351, 'F17C': 352, 'F21L': 353, 'F21S': 354, 'F21V': 355, 'F22B': 356, 'F23C': 357, 'F23D': 358, 'F23G': 35
9, 'F23J': 360, 'F23L': 361, 'F23N': 362, 'F23Q': 363, 'F24B': 364, 'F24C': 365, 'F24D': 366, 'F24F': 367, 'F24H': 368, 'F24S':
369, 'F25B': 370, 'F25C': 371, 'F25D': 372, 'F25J': 373, 'F26B': 374, 'F27B': 375, 'F27D': 376, 'F28D': 377, 'F28F': 378, 'F41
A': 379, 'F41B': 380, 'F41C': 381, 'F41G': 382, 'F41H': 383, 'F41J': 384, 'F42B': 385, 'F42C': 386, 'F42D': 387, 'G01B': 388,
'G01C': 389, 'G01D': 390, 'G01F': 391, 'G01G': 392, 'G01J': 393, 'G01K': 394, 'G01L': 395, 'G01M': 396, 'G01N': 397, 'G01P': 39
8, 'G01R': 399, 'G01S': 400, 'G01T': 401, 'G01V': 402, 'G02B': 403, 'G02C': 404, 'G02F': 405, 'G03B': 406, 'G03C': 407, 'G03F':
408, 'G03G': 409, 'G04B': 410, 'G05B': 411, 'G05D': 412, 'G05F': 413, 'G05G': 414, 'G06F': 415, 'G06K': 416, 'G06N': 417, 'G06
Q': 418, 'G06T': 419, 'G07B': 420, 'G07C': 421, 'G07D': 422, 'G07F': 423, 'G08B': 424, 'G08C': 425, 'G08G': 426, 'G09B': 427,
'G09D': 428, 'G09F': 429, 'G09G': 430, 'G10D': 431, 'G10K': 432, 'G10L': 433, 'G11B': 434, 'G11C': 435, 'G16H': 436, 'G21C': 43
7, 'G21F': 438, 'H01B': 439, 'H01C': 440, 'H01F': 441, 'H01G': 442, 'H01H': 443, 'H01J': 444, 'H01K': 445, 'H01L': 446, 'H01M':
447, 'H01P': 448, 'H01Q': 449, 'H01R': 450, 'H01S': 451, 'H01T': 452, 'H02B': 453, 'H02G': 454, 'H02H': 455, 'H02J': 456, 'H02
K': 457, 'H02M': 458, 'H02N': 459, 'H02P': 460, 'H02S': 461, 'H03F': 462, 'H03G': 463, 'H03H': 464, 'H03J': 465, 'H03K': 466,
'H03M': 467, 'H04B': 468, 'H04H': 469, 'H04J': 470, 'H04L': 471, 'H04M': 472, 'H04N': 473, 'H04Q': 474, 'H04R': 475, 'H04S': 47
6, 'H04W': 477, 'H05B': 478, 'H05H': 479, 'H05K': 480}
```

Figura 41. Etiquetas codificadas

4.4 Modelado.

4.4.1 Técnica de modelado.

Como se mencionó anteriormente para le selección de la técnica de modelado se utilizará la propuesta que se encuentra en la guía de *"Text Classification"* de Google que se muestra en la Figura 9. Por lo que se calcula el ratio S/W de la nueva variable "Texto" obteniendo como resultado.

```
# Calculando S y W
S = len(df)
W = df['total_word_count'].mean()

# Calculando S/W
SW_ratio = S/W

print(f"S/W ratio 'Texto' es: {SW_ratio:.2f}")
```

```
S/W ratio 'Texto' es: 4488.09
```

Figura 42. Ratio S/W de "Texto"

Debido a que el resultado es 4.488 se hará uso del de la Red Neuronal Convolucional Separable que es una variante de la Red Neuronal Convolucional que se enfoca en reducir la cantidad de parámetros y operaciones computacionales requeridas en las capas de convolución por lo que se reduce la sobrecarga computacional, siendo útil para el uso en dispositivos con recursos limitados.

4.4.2 Diseño de Prueba.

Para realizar el proceso de entrenamiento, de los 694.715 registros con los que cuenta la base de datos se utilizará el 80% para el entrenamiento del modelo mientras que el 20% restante se utilizará para evaluar el mismo.

4.4.3 Construcción del Modelo y Evaluación.

Para la construcción del modelo se utilizaron los siguientes hiperparámetros.

Parámetro de aprendizaje	Valor
Función de pérdida	sparse_categorical_crossentropy
Optimizador	Adam
Tasa de aprendizaje	0,001
Ciclos de entrenamiento	1,000
Tamaño del lote	128
Interrupción anticipada	val_loss', patience=3
Tasa de retirados	0.2
Dimensiones de incorporación	Incorporaciones de Word2vec con 300 dimensiones

Tabla 5. Parámetros de aprendizaje

Con los parámetros antes señalados se realizaron tres experimentos, de los cuales se presentará el resumen de la arquitectura de la Red Neuronal utilizada y los resultados obtenidos. Como se mencionó anteriormente, la etiqueta (Subclase) no es exclusiva, por lo que una patente puede tener varias etiquetas. En este contexto se presentará los resultados donde la etiqueta principal coincide con la probabilidad más alta (top), el segundo resultado evaluará si la etiqueta principal se encuentra dentro de las tres probabilidades más altas (top 3) y finalmente el tercer resultado evaluará si la etiqueta principal se encuentra dentro de las cinco probabilidades más altas (top 5).

1. En el primer experimento, la red neuronal se entrenará únicamente con el vocabulario tokenizado de la base de datos construida donde se permite que la red ajuste todas las ponderaciones.

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 400, 300)          123433800

 dropout (Dropout)           (None, 400, 300)          0

 separable_conv1d (Separable  (None, 400, 256)         78556
 Conv1D)

 separable_conv1d_1 (Separab  (None, 400, 256)         67072
 leConv1D)

 max_pooling1d (MaxPooling1D  (None, 133, 256)         0
 )

 separable_conv1d_2 (Separab  (None, 133, 512)         132864
 leConv1D)

 separable_conv1d_3 (Separab  (None, 133, 512)         265216
 leConv1D)

 global_average_pooling1d (G  (None, 512)              0
 lobalAveragePooling1D)

 dropout_1 (Dropout)         (None, 512)               0

 dense (Dense)               (None, 481)               246753

=================================================================
Total params: 124,224,261
Trainable params: 124,224,261
Non-trainable params: 0
```

Figura 43. Resumen experimento

Del primer experimento se obtuvieron los siguientes resultados.

- **Top**

```
loss, accuracy = model.evaluate(x_test, y_test)
print(f"Test Accuracy: {accuracy * 100:.2f}%")
```

```
4342/4342 [==============================] - 16s 4ms/step - loss: 2.5169 - accuracy: 0.5153
Test Accuracy: 51.53%
```

Figura 44. Resultado top experimento 1

- **Top3**

```
top_3_preds = np.argsort(predictions, axis=1)[:, -3:]
```

```
# Verifica si las etiquetas verdaderas están en los top 3 predicciones
correct_preds = np.any(top_3_preds == y_test[:, np.newaxis], axis=1)
```

```
# Calcula la precisión de top 3
top_3_accuracy = np.mean(correct_preds)
print(f"Top 3 Accuracy: {top_3_accuracy * 100:.2f}%")
```

```
Top 3 Accuracy: 70.89%
```

Figura 45. Resultado top 3 experimento 1

- **Top5**

```
top_5_preds = np.argsort(predictions, axis=1)[:, -5:]
```

```
# Verifica si las etiquetas verdaderas están en los top 5 predicciones
correct_preds = np.any(top_5_preds == y_test[:, np.newaxis], axis=1)
```

```
# Calcula la precisión de top 5
top_5_accuracy = np.mean(correct_preds)
print(f"Top 5 Accuracy: {top_5_accuracy * 100:.2f}%")
```

```
Top 5 Accuracy: 77.53%
```

Figura 46. Resultado top 5 experimento 1

2. En el segundo experimento se realizará la incrustación de "***pre-trained embeddings***" con los pesos de la capa de incorporación inmovilizados, permitiendo que el resto de la red aprenda.

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 400, 300)          123433800

 dropout (Dropout)           (None, 400, 300)          0

 separable_conv1d (Separable  (None, 400, 256)         78556
 Conv1D)

 separable_conv1d_1 (Separab  (None, 400, 256)         67072
 leConv1D)

 max_pooling1d (MaxPooling1D  (None, 133, 256)         0
 )

 separable_conv1d_2 (Separab  (None, 133, 512)         132864
 leConv1D)

 separable_conv1d_3 (Separab  (None, 133, 512)         265216
 leConv1D)

 global_average_pooling1d (G  (None, 512)              0
 lobalAveragePooling1D)

 dropout_1 (Dropout)         (None, 512)               0

 dense (Dense)               (None, 481)               246753

=================================================================
Total params: 124,224,261
Trainable params: 790,461
Non-trainable params: 123,433,800
_________________________________________________________________
```

Figura 47. Resumen experimento 2

Del primer experimento se obtuvieron los siguientes resultados.

- **Top**

```
loss, accuracy = model.evaluate(x_test, y_test)
print(f"Test Accuracy: {accuracy * 100:.2f}%")
```

```
4342/4342 [==============================] - 16s 4ms/step - loss: 1.9392 - accuracy: 0.5447
Test Accuracy: 54.47%
```

Figura 48. Resultado top experimento 2

- **Top3**

```
top_3_preds = np.argsort(predictions, axis=1)[:, -3:]
```

```
# Verifica si las etiquetas verdaderas están en los top 3 predicciones
correct_preds = np.any(top_3_preds == y_test[:, np.newaxis], axis=1)
```

```
# Calcula la precisión de top 3
top_3_accuracy = np.mean(correct_preds)
print(f"Top 3 Accuracy: {top_3_accuracy * 100:.2f}%")
```

```
Top 3 Accuracy: 75.74%
```

Figura 49. Resultado top 3 experimento 2

- **Top5**

```
top_5_preds = np.argsort(predictions, axis=1)[:, -5:]
```

```
# Verifica si las etiquetas verdaderas están en los top 5 predicciones
correct_preds = np.any(top_5_preds == y_test[:, np.newaxis], axis=1)
```

```
# Calcula la precisión de top 5
top_5_accuracy = np.mean(correct_preds)
print(f"Top 5 Accuracy: {top_5_accuracy * 100:.2f}%")
```

```
Top 5 Accuracy: 82.42%
```

Figura 50. Resultado top 5 experimento 2

3. En el tercer experimento se realizará la incrustación de "***pre-trained embeddings***" y se permite que la capa de incorporación también aprenda y realice ajustes a todas las ponderaciones de la red.

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 400, 300)          123433800

 dropout (Dropout)           (None, 400, 300)          0

 separable_conv1d (Separable  (None, 400, 256)         78556
 Conv1D)

 separable_conv1d_1 (Separab  (None, 400, 256)         67072
 leConv1D)

 max_pooling1d (MaxPooling1D  (None, 133, 256)         0
 )

 separable_conv1d_2 (Separab  (None, 133, 512)         132864
 leConv1D)

 separable_conv1d_3 (Separab  (None, 133, 512)         265216
 leConv1D)

 global_average_pooling1d (G  (None, 512)              0
 lobalAveragePooling1D)

 dropout_1 (Dropout)         (None, 512)               0

 dense (Dense)               (None, 481)               246753

=================================================================
Total params: 124,224,261
Trainable params: 124,224,261
Non-trainable params: 0
```

Figura 51. Resumen experimento 3

Del primer experimento se obtuvieron los siguientes resultados.

- **Top**

```
loss, accuracy = model.evaluate(x_test, y_test)
print(f"Test Accuracy: {accuracy * 100:.2f}%")
```

```
4342/4342 [==============================] - 18s 4ms/step - loss: 1.8683 - accuracy: 0.5718
Test Accuracy: 57.18%
```

Figura 52. Resultado top experimento 3

- **Top3**

```
top_3_preds = np.argsort(predictions, axis=1)[:, -3:]
```

```
# Verifica si las etiquetas verdaderas están en los top 3 predicciones
correct_preds = np.any(top_3_preds == y_test[:, np.newaxis], axis=1)
```

```
# Calcula la precisión de top 3
top_3_accuracy = np.mean(correct_preds)
print(f"Top 3 Accuracy: {top_3_accuracy * 100:.2f}%")
```

```
Top 3 Accuracy: 77.70%
```

Figura 53. Resultado top 3 experimento 3

- **Top5**

```
top_5_preds = np.argsort(predictions, axis=1)[:, -5:]
```

```
# Verifica si las etiquetas verdaderas están en los top 5 predicciones
correct_preds = np.any(top_5_preds == y_test[:, np.newaxis], axis=1)
```

```
# Calcula la precisión de top 5
top_5_accuracy = np.mean(correct_preds)
print(f"Top 5 Accuracy: {top_5_accuracy * 100:.2f}%")
```

```
Top 5 Accuracy: 84.06%
```

Figura 54. Resultado top 5 experimento 3

4.5 Evaluación

Los resultados obtenidos del modelo de clasificación de patentes, indican una capacidad prometedora, reflejada bajo la métrica “accuracy” con un porcentaje de 84% en el experimento tres en el top 5. Se comparan todos los resultados obtenidos en cada uno de los experimentos en la siguiente tabla.

	Experimento 1	Experimento 2	Experimento 3
Top	51.53%	54,47%	57,18%
loss	2,51	1,93	1,86
Top3	70,89%	75,74%	77,70%
Top5	77,53%	82,42%	84,06%

Tabla 6. Resumen de resultados

Como se muestra en la Tabla 6 el mejor resultado obtenido corresponde al experimento 3.

Como parte importante de la evaluación de los resultados obtenidos, además de utilizar el *"accuracy"* para evaluar los resultados, también se recurrió a la evaluación cualitativa en la que se buscó la opinión técnica sobre los resultados arrojados por el modelo. Para lo cual, se extrajo del conjunto de datos utilizado para la evaluación del modelo, 10 patentes de manera aleatoria para que un experto pueda evaluar los resultados obtenidos por el modelo. Teniendo como resultado que se *"…evidencia que la clasificación propuesta por el Modelo difiere en muchos casos en una o más clasificaciones adicionales a las asignadas por la Oficina de propiedad intelectual, estas clasificaciones mayoritariamente se encuentran en el sector tecnológico de la invención a la que pertenece, por lo que se puede considerar que el Modelo genera clasificaciones muy relacionadas a la invención, a nivel de clases y subclase que pueden orientar al analista o investigador a definir una clasificación más específica a través de la revisión exhaustiva del grupo y subgrupo que más sea aplicable"*. El detalle de la respuesta a la solicitud de una opinión técnica se encuentra en el Anexo 1.

4.6 Despliegue

Con la finalidad de cumplir con la metodología CRISP-DM, se realiza el despliegue del experimento 3 que obtuvo los valores más altos en la métrica "accuray" y mediante el cual se realizó la evaluación cualitativa con un experto. El desarrollo del despliegue se encuentra detallado en el Anexo2.

5 Capítulo V. Conclusiones y Recomendaciones

5.1 Conclusiones

En relación con las preguntas planteadas en el primer capítulo del documento, y de acuerdo con los resultados obtenidos en los experimentos realizados podemos responder que:

RQ1.¿Es posible clasificar las patentes de invención dentro de las subclases del sistema estandarizado de Clasificación Internacional de Patentes – CIP mediante el título y resumen?

Como se ha demostrado en los experimentos realizados, es posible clasificar patentes de invención dentro de las de las subclases del sistema estandarizado de Clasificación Internacional de Patentes – CIP mediante el título y resumen, conforme a los resultados de los experimentos realizados tanto en la evaluación cuantitativa como cualitativa, el modelo está generalizando de manera aceptable nueva información. Contribuyendo a que el proceso de clasificación se lo pueda simplificar.

RQ2.¿Se puede utilizar el Procesamiento de Lenguaje Natural para clasificar las Patentes de Invención?

El Procesamiento de Lenguaje Natural, es posible usarlo en la clasificación de Patentes, como se explica en el desarrollo del proyecto, se realizó una tokenización secuencial con la finalidad de que el modelo pueda deducir el significado de las palabras a partir del orden en el que se encuentran, permitiendo que pueda generalizar nuevos casos de mejor manera. Demostrando que esta rama de la Inteligencia Atificial puede ser utilizada para la clasificación de documentación técnica.

RQ3.¿Es suficiente con el texto del título y resumen de la patente para poder realizar una clasificación óptima?

De acuerdo a los resultados obtenidos, es suficiente con el texto del título y resumen de la patente para poder realizar una clasificación óptima, de acuerdo a la evaluación cualitativa realizada por personal técnico, es posible que el modelo este considerando información adicional a la considerada por los analistas y realizando clasificaciones adicionales válidas, por lo que se podría estar clasificando una patente de manera eficiente usando únicamente el título y resumen.

RQ4.¿Qué modelo de machine learning permite tener mejores resultados en la clasificación de patentes?

Actualmente existe varios modelos de machine learnig que se pueden usar para este fin, de acuerdo con los resultados obtenidos en la revisión de trabajos similares, las Redes Neuronales Convolucionales parecerían, por el momento, ser una de las mejores opciones.

5.2 Recomendaciones

- Es necesario continuar con la evaluación cualitativa del modelo debido a que el número de solicitudes revisadas por el experto no es representativo a la muestra.

- El desbalance existente en el conjunto de datos trabajado puede estar afectando el resultado del modelo, por lo que es recomendable continuar con el trabajo en el desequilibrio de las etiquetas.

- Se recomienda que en proyectos futuros se pueda realizar la tokenización secuencial por letra en lugar de la tokenización secuencial por palabras, esto podría ayudar a comparar resultados sobre la eficiencia y la precisión de los modelos, ya que suelen existir errores tipográficos al momento de ingresar la información de las patentes por lo que una tokenización secuencial por letras puede limitar este problema.

- Se recomienda poner en producción el modelo debido a que, con los resultados obtenidos, puede ser una herramienta útil para orientar de manera adecuada a un analista o investigador, además de ayudar a detectar posibles oportunidades de innovación en las diferentes áreas tecnológicas.

6 Bibliografía

Aiolli, F., Cardin, R., Sebastiani, F., Sperduti, A.: Preferential text classification: Learning algorithms and evaluation measures. Information Retrieval 12(5), 559–580 (2009)

Aiolli, F., Cardin, R., Sebastiani, F., Sperduti, A.: Preferential text classification: Learning algorithms and evaluation measures. Information Retrieval 12(5), 559–580 (2009)

Castro Calderón, M. (2023). Información de patentes: Uso y manejo de bases de datos de patentes [PDF]. Obtenido de https://www.indecopi.gob.pe/documents/20791/204308/06.-24_04_14+-+Uso+y+manejo+de+base+de+datos.pdf/3eae0e54-d850-4aec-bb17-2a3140a6c872

Chen, Y.L., Chang, Y.C.: A three-phase method for patent classification. Information Processing and Management 48(6), 1017–1030 (2012)

Comisión de la Comunidad Andina. (2000). DECISIÓN 486 - Régimen Común sobre Propiedad Industrial. Lima, Perú.

Dialani, P. (2020, 29 de octubre). The Future of Data Revolution will be Unstructured Data. Analytics Insight. Obtenido de https://www.analyticsinsight.net/the-future-of-data-revolution-will-be-unstructured-data/

Forum Huawei. (2022). Tipos de datos en Big Data. Publicado el 12 de marzo de 2022 a las 20:02:38. Obtenido de https://forum.huawei.com/enterprise/es/Tipos-de-datos-en-Big-Data/thread/667228429046661120-667212895836057600

Google Developers. (2023). Text Classification. Obtenido de https://developers.google.com/machine-learning/guides/text-classification

Haykin, S.: Neural Networks: A Comprehensive Foundation. Prentice Hall (1994)

IBM. (2023). Algoritmos de aprendizaje supervisado. Obtenido de https://www.ibm.com/es-es/topics/supervised-learning

OMPI (Organización Mundial de la Propiedad Intelectual). (2023). Clasificación Internacional de Patentes. Obtenido de https://www.wipo.int/classifications/ipc/es/preface.html

Rodríguez Montequín, Mª Teresa; Álvarez Cabal, J. Valeriano; Mesa Fernández, José Manuel; González Valdés, Adolfo. (s.f.). Metodologías para la realización de proyectos de Data Mining. Obtenido de

https://www.aeipro.com/files/congresos/2003pamplona/ciip03_0257_0265.2134.pdf

SENADI (Servicio Nacional de Derechos Intelectuales) (2023). SENADI en Cifras. Obtenido de https://www.derechosintelectuales.gob.ec/senadi-en-cifras/

Teodoro, D., Gobeill, J., Pasche, E., Ruch, P., Vishnyakova, D., Lovis, C.: Automatic IPC encoding and novelty tracking for effective patent mining. In: Proceedings of the 8th NTCIR Workshop Meeting, pp. 309–317. National Institute of Informatics Japan (2010

TIBCO Software Inc. (2023). What is Supervised Learning? Obtenido de https://www.spotfire.com/glossary/what-is-supervised-learning

Opinión Técnica sobre Clasificación de Patentes de Invención

ANEXO 1

Quito, martes 26 de diciembre de 2023

Servicio Nacional de Derechos Intelectuales

Quito - Ecuador

Bioq. Walter Fabián Darquea Chugcho
Director Técnico de Patentes

Asunto: Solicitud de Opinión Técnica sobre Clasificación de Patentes de Invención

Estimado Director Técnico del Servicio Nacional de Derechos Intelectuales,

Me dirijo a usted en calidad de Maestrante de la Pontificia Universidad Católica del Ecuador, carrera Sistemas de Información con mención en Data Science, con el propósito de solicitar una opinión técnica sobre la clasificación de patentes de invención utilizando un modelo de redes neuronales desarrollado como proyecto de titulación de la maestría antes mencionada.

Contexto:

Para el desarrollo del proyecto de titulación se ha aplicado un modelo de redes neuronales con el objetivo de clasificar patentes de invención de manera eficiente y precisa a nivel de subclase. El modelo ha demostrado un nivel de precisión, medido por la tasa de exactitud (accuracy), de aproximadamente el 75% en el top 3 de clasificaciones propuestas. Dicho modelo ha sido entrenado utilizando un conjunto de datos robusto y diverso que incluye una amplia gama de resúmenes y títulos patentes de invención tomados de PATENTSCOPE.

Motivo de la Solicitud:

Dado que la clasificación de patentes de invención es parte de las tareas que realiza la Oficina de Propiedad Intelectual, estoy interesado en obtener una opinión técnica de su parte, en relación con los resultados arrojados por el modelo propuesto. Lo que se desea conocer es que si el modelo realiza clasificaciones diferentes a las presentadas por la Oficina de Propiedad Intelectual, estas recomendaciones podrían considerarse válidas.

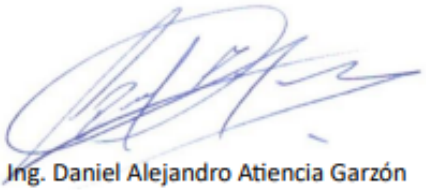
En este sentido, he adjuntado 10 ejemplos de patentes de invención de nuestra base de datos de prueba para su revisión detallada. Estos ejemplos se han seleccionado aleatoriamente para evaluar la validez de las clasificaciones propuestas por nuestro modelo en comparación con las clasificaciones existentes en la base de datos de la Oficina de Propiedad Intelectual.

Agradezco enormemente, si pudieran colaborar con esta evaluación técnica utilizando los ejemplos proporcionados. Estoy dispuesto a proporcionar acceso a los datos utilizados para el entrenamiento del modelo, así como a cualquier otra información que puedan requerir para llevar a cabo esta revisión de manera efectiva.

Agradezco de antemano su tiempo y consideración.

Quedo a su disposición para cualquier consulta adicional.

Atentamente,



Ing. Daniel Alejandro Atiencia Garzón

Ci:1720930039

Anexo: Ejemplos

EJEMPLOS

1. Métodos y productos para controlar la mosca de la seda y mosca jorobada en maíz. se relaciona con métodos para producir plantas de maíz materiales de planta y semillas que proporcionan un medio para controlar y defender contra la mosca de seda de maíz y mosca jorobada y plantas de maíz materiales de planta y semilla mejorada producida por estos métodos moscas de seda de maíz y moscas jorobadas son insectos altamente perjudiciales y destructivos que dañan plantas de maíz materiales de planta y semilla durante sus etapas de crecimiento en muchas regiones geográficas diferentes a lo largo del mundo los métodos inventivos utilizan marcadores moleculares para incorporar genes que sirven de medios para controlar y defender contra la mosca de seda de maíz y mosca jorobada en líneas congeniales parentales de maíz masculinas y femeninas usadas para producir híbridos de maíz o en al menos una de las líneas parentales de maíz de un híbrido.

Application ID: CL312991573	
Clasificación Oficina de PI	**Clasificación Modelo**
A01H	A01H Probabilidad: 0.86
C12Q	C12N, Probabilidad: 0.10
	PORCENTAJE TOTAL: 96%

2. Vehículo eléctrico como unidad punta de suministro de energía. automóvil con al menos un motor eléctrico un acumulador de energía para proporcionar energía motriz para el motor eléctrico con un conector enchufable conectado al acumulador de energía para la conexión a una red de suministro y con un control que comprende un reloj para el control del flujo de corriente a través de la red de suministro y del acumulador de energía en el que el control permite una flujo de corriente desde el acumulador de energía a la red de suministro y el control comprende un dispositivo para el registro de la cantidad de carga en el acumulador de energía e interrumpe el flujo de corriente desde el acumulador de energía a la red de suministro al alcanzar un valor umbral prefijado de la energía de carga residual restante de manera que se garantiza una cantidad de energía suficiente para un trayecto determinado y porque con el control están previstos medios de entrada acoplados mediante los cuales un usuario del vehículo puede ajustar el tiempo o el periodo temporal en el que se puede llevar a cabo al menos parcialmente una descarga del acumulador y con ello una alimentación de la energía a la red de suministro.

Application ID: ES5708536	
Clasificación Oficina de PI	**Clasificación Modelo**
B60L	B60L Probabilidad: 0.68
H02J	B60K Probabilidad: 0.16

	H02J Probabilidad: 0.09
	PORCENTAJE TOTAL: 93%

3. Perfeccionamientos en hornos para el tratamiento de polvo deamolado y similares. perfeccionamientos en hornos para el tratamiento de polvo de amolado y similares para eliminar contaminantes del mismo y producir material metálico refundible caracterizados porque dichos hornos se forman mediante una cámara de combustión alargada con un eje ligeramente inclinado con respecto a la horizontal medios para hacer girar la cámara alrededor de dicho eje medios para alimentar material al extremo superior de la cámara y para recoger el material al extremo superior de la cámara y para recoger el material descargado del extremo inferior de la cámara y por lo menos un quemador para caldear en general a lo largo de la longitud de la cámara.

Application ID: ES31560538	
Clasificación Oficina de PI	**Clasificación Modelo**
C22B	C22B Probabilidad: 0.20
F27B	F27D Probabilidad: 0.12
	F27B Probabilidad: 0.11
	PORCENTAJE TOTAL: 43%

4. Procedimiento de una sola etapa para separar componentes de la biomasa. un procedimiento para la obtención de celulosa de la biomasa en una forma susceptible de sufrir hidrólisisenzimática comprendiendo dicho procedimiento poner en contacto la biomasa con una mezcla de un disolvente orgánico inmiscible en agua un ácido y un catalizador de sal metálica disuelto en la solución de agua ácida a una temperatura en el intervalo y una presión en el intervalo de a bar y filtrando la mezcla de reacción a presión para separar la lignina disuelta la hemicelulosa en la fase de disolvente y acuosa respectivamente y dejando atrás la celulosa pura en el que el catalizador de sal metálica se selecciona de sulfato de cobre sulfato ferroso sulfato de amonioferroso sulfato de níquel sulfato de sodio y cloruro férrico y en el que el catalizador está presente en el intervalo del al de agua ácida.

Application ID: ES105851259	
Clasificación Oficina de PI	**Clasificación Modelo**
D21C	C08B Probabilidad: 0.42
C12P	D21C Probabilidad: 0.41
C13K	D01F Probabilidad: 0.06
C12M	PORCENTAJE TOTAL: 89%

5. Estructura de anclaje y método para la construcción de la misma. en una estructura de un cuerpo de anclaje se inserta una vaina deformada hecha de acero y que tiene un cuerpo tubular y cuya pared periférica presenta una con ración irregular en un agujero excavado se inserta un material de tensión por lo menos en la vaina deformada varios tubos de acero se cubren sobre por lo menos una parte extrema de una pared periférica exterior de la vaina deformada o se insertan en por lo menos una parte extrema de una pared periférica interior de la vaina deformada para reforzar esta en el agujero excavado se introduce un endurecedor y se deja un espacio dentro de la vaina deformada para formar una sección de fijación también se presenta un método para la construcción del cuerpo de anclaje.

Application ID: ES5468252	
Clasificación Oficina de PI	**Clasificación Modelo**
E02D	E21D Probabilidad: 0.74
A47K	E02D Probabilidad: 0.04
	PORCENTAJE TOTAL: 78%

6. Mejoras en calefactores para nidos de animales domésticos. mejoras en calefactores para nidos de animales domésticos consisten en partir de una caja portátil el fondo de la cual posee dos aberturas principales y que comunican con aberturas y de la nidera y una tercera abertura intermedia menor que comunica con el exterior entre cuyas aberturas y hay una resistencia con un termostato de seguridad conectado a un avisador sonoro y otro luminoso entre las aberturas y está situado un ventilador que establece una corriente de aire desde las aberturas y hasta la abertura es posible trasladar fácilmente el grupo calefactor de un nido a otro.

Application ID: ES31819642	
Clasificación Oficina de PI	**Clasificación Modelo**
F24C	A01K Probabilidad: 0.67
A01K	F24C Probabilidad: 0.06
	PORCENTAJE TOTAL: 73%

\7. Módulos de circuito integrado y tarjetas inteligentes que incorporan los mismos. un módulo de circuito integrado para una tarjeta inteligente tanto con interfaces de contacto como sin contacto comprendiendo el módulo de ci un sustrato no conductor que tiene una pluralidad de agujeros de enlace simple y un par de agujeros de enlace múltiple extendiéndose ambos desde un primer a un segundo lado del sustrato en el que cada agujero de enlace múltiple es más grande que cada agujero de enlace simple y está con rado para recibir múltiples elementos conductores en el que cada agujero de enlace simple está con rado para recibir un único elemento conductor una pluralidad de almohadillas de contacto conductoras que incluyen un primer par de las mismas dispuestas en el primer lado del sustrato un primer chip de ci dispuesto en el segundo lado del sustrato una pluralidad de primeros elementos conductores que atraviesan el enlace simple y el par de agujeros de enlace múltiple y que conectan

eléctricamente al menos algunas de las almohadillas de contacto al primer chip de ci en el que los primeros elementos conductores incluyen un primer par de primeros elementos conductores que atraviesan respectivamente el par de agujeros de enlace múltiple y que conectan eléctricamente el primer par de almohadillas de contacto al primer chip de ci y caracterizado por un encapsulante depositado en el primer chip de ci los primeros elementos conductores y el primer par de almohadillas de contacto en el que un borde del encapsulante depositado en el primer par de almohadillas de contacto distribuye cada uno del par de agujeros de enlace múltiple en un primer canal de enlace encapsulado y un segundo canal de enlace no encapsulado contiguo que terminan respectivamente en una primera y una segunda área de enlace contigua en cada una del primer par de almohadillas de contacto de tal manera que el primer par de primeros elementos conductores atraviesan respectivamente el primer canal de enlace del par de agujeros de enlace múltiple y además de tal manera que el encapsulante sella la primera área de enlace y expone la segunda área de enlace a través del segundo canal de enlace para proporcionar una superficie para establecer una conexión eléctrica con el primer chip de ci en el que el encapsulante separa las áreas de enlace primera y segunda entre sí sin necesitar la presencia del sustrato entre las mismas

Application ID: ES321304225	
Clasificación Oficina de PI	**Clasificación Modelo**
G06K	H01L Probabilidad: 0.52
H05K	H05K Probabilidad: 0.22
	G06K Probabilidad: 0.16
	PORCENTAJE TOTAL: 90%

8. Sistema eléctrico que utiliza ca de alta frecuencia y que tiene cargas conectadas inductivamente y fuentes de alimentación y luminarias correspondientes. un sistema de distribución de potencia para distribuir potencia ca de alta frecuencia desde una fuente de potencia limitada actual el sistema comprende un par trenzado de conductores alargados con rados para conectarse a la fuente de alimentación los extremos de los conductores más alejados estando la fuente de alimentación conectada entre sí donde los conductores entre las vueltas adyacentes del par trenzado son separables entre sí para definir una abertura entre ellos un elemento de toma de potencia que está con rado para ser insertado al menos parcialmente a través de la abertura para que la energía eléctrica pueda acoplarse inductivamente desde los conductores al elemento de toma de corriente donde otro conductor se embobine al menos parcialmente alrededor del elemento de toma de corriente caracterizado porque el sistema comprende además un circuito de rectificación síncrono conectado al conductor adicional para convertir la potencia de corriente alterna de alta frecuencia en el conductor adicional a un voltaje regulado por cc para potencia una carga y un rectificador en el circuito de rectificación síncrono para cortocircuitar de manera controlable el circuito de rectificación para detener la transmisión de potencia a la carga.

Application ID: ES217481931	
Clasificación Oficina de PI	**Clasificación Modelo**
H05B	H02J Probabilidad: 0.44
H02J	H02M Probabilidad: 0.32
H01F	H01F Probabilidad: 0.04
H02M	PORCENTAJE TOTAL: 80%

9. Procedimiento para la descarga de un acumulador de energía eléctrica. procedimiento para la descarga de un acumulador de energía eléctrica que está conectado con un circuito electrónico mediante un primer conductor eléctrico y un segundo conductor eléctrico en donde está proporcionado un tiristor para la descarga del acumulador de energía eléctrica en donde en el procedimiento debido a un error que ocurre en el circuito electrónico una corriente de descarga del acumulador de energía eléctrica comienza a fluir desde el acumulador de energía eléctrica a través del primer conductor eléctrico al circuito electrónico y a través del segundo conductor eléctrico de regreso al acumulador de energía eléctrica debido a la corriente de descarga alrededor del primer conductor eléctrico y del segundo conductor eléctrico se genera un campo magnético que se modifica con el tiempo el cual atraviesa el material semiconductor del tiristor por el campo magnético que cambia con el tiempo una corriente es inducida en el material semiconductor del tiristor caracterizado porque el tiristor se activa mediante esta corriente inducida.

Application ID: ES320284468	
Clasificación Oficina de PI	**Clasificación Modelo**
H02M	H02J Probabilidad: 0.69
H03K	H02M Probabilidad: 0.12
	H01M Probabilidad: 0.12
	PORCENTAJE TOTAL: 93%

10. método y sistema de procesamiento de un macro. un método para realizar comunicaciones de datos dentro de un sistema a base de macros caracterizado porque comprende las etapas de recibir información a base de macros que contienen variables intercaladas a partir de un nodo del cliente recuperar en respuesta a la información a base de macros recibida y las variables intercaladas contenidas en el mismo datos que representen al menos una fila de macros particular y que modifiquen los datos recuperados como una función de las variables contenidas en los datos a base de macros recibidos transmitir los datos recuperados modificados al nodo del cliente por lo que proporcionan al nodo del cliente con información reflectiva del estado previo del sistema a base de macros.

Application ID: MX33434	
Clasificación Oficina de PI	**Clasificación Modelo**
G06F	H04L Probabilidad: 0.31
	G06F Probabilidad: 0.30
	H04J Probabilidad: 0.07
	PORCENTAJE TOTAL: 68%

Quito a 09 de enero de 2024

Señor Ingeniero
Daniel Atiencia

De mis consideraciones

Asunto: Solicitud de Opinión Técnica sobre Clasificación de Patentes de Invención

En atención a su requerimiento mediante oficio sin número del 26 de diciembre de 2023, mediante el cual solicita una opinión técnica sobre la clasificación de patentes de invención utilizando un modelo de redes neuronales desarrollado como proyecto de titulación de la maestría por usted cursada, informo a usted que dicha opinión se encuentra como anexo a la presente comunicación .

Con sentimientos de consideración y estima

WALTER FABIAN DARQUEA CHUGCHO

Firmado digitalmente por WALTER FABIAN DARQUEA CHUGCHO
DN: cn=WALTER FABIAN DARQUEA CHUGCHO c=EC o=SECURITY DATA S.A. 2 ou=ENTIDAD DE CERTIFICACION DE INFORMACION
Motivo:Soy el autor de este documento
Ubicación:
Fecha:2024-01-09 14:05:05:00

Fabián Darquea Ch
Director Técnico de Patentes-SENADI.

OPINION TÉCNICA SOBRE LA CLASIFICACION APLICADO POR MODELO DE REDES NEURONALES

El presente documento abarca la opinión técnica sobre la clasificación de patentes de invención utilizando un modelo de redes neuronales desarrollado como proyecto de titulación de la maestría cursada por el Ing, Daniel Atiencia, y en la cual se evalúa la validez de las clasificaciones propuestas por el modelo en comparación con las clasificaciones existentes en la base de datos de la Oficina de Propiedad Intelectual a la que pertenece el documento tomado como ejemplo.

Se debe enfatizar que la clasificación de patentes asignada a un trámite es una actividad que tiene un carácter subjetivo pudiendo identificarse una o más clasificaciones para un mismo invento, dependiendo en muchos casos del criterio del analista o de la Oficina.

A continuación se expone los comentarios sobre las clasificaciones de cada uno de los 10 ejemplos propuestos.

EJEMPLOS

1. Métodos y productos para controlar la mosca de la seda y mosca jorobada en maíz. se relaciona con métodos para producir plantas de maíz materiales de planta y semillas que proporcionan un medio para controlar y defender contra la mosca de seda de maíz y mosca jorobada y plantas de maíz materiales de planta y semilla mejorada producida por estos métodos moscas de seda de maíz y moscas jorobadas son insectos altamente perjudiciales y destructivos que dañan plantas de maíz materiales de planta y semilla durante sus etapas de crecimiento en muchas regiones geográficas diferentes a lo largo del mundo los métodos inventivos utilizan marcadores moleculares para incorporar genes que sirven de medios para controlar y defender contra la mosca de seda de maíz y mosca jorobada en líneas congeniales parentales de maíz masculinas y femeninas usadas para producir híbridos de maíz o en al menos una de las líneas parentales de maíz de un híbrido.

Application ID: CL312991573	
Clasificación Oficina de PI	**Clasificación Modelo**
A01H	A01H Probabilidad: 0.86
C12Q	C12N, Probabilidad: 0.10
	PORCENTAJE TOTAL: 96%

Comentarios.-

La clasificación oficial A01H y C12Q corresponde al sector tecnológico de la invención. La clasificación adicional C12N propuesta por el Modelo se refiere a microorgansimos, enzimas, ingeniería genética. Esta clasificación se aproxima de manera muy general, en virtud de que engloba tecnología que tiene actividad biocida y repelente de plagas, no obstante la clasificación se considerará válida cuando el grupo y subgrupo se relacionen con la invención.

2. Vehículo eléctrico como unidad punta de suministro de energía. automóvil con al menos un motor eléctrico un acumulador de energía para proporcionar energía motriz para el motor eléctrico con un conector enchufable conectado al acumulador de energía para la conexión a una red de

suministro y con un control que comprende un reloj para el control del flujo de corriente a través de la red de suministro y del acumulador de energía en el que el control permite una flujo de corriente desde el acumulador de energía a la red de suministro y el control comprende un dispositivo para el registro de la cantidad de carga en el acumulador de energía e interrumpe el flujo de corriente desde el acumulador de energía a la red de suministro al alcanzar un valor umbral prefijado de la energía de carga residual restante de manera que se garantiza una cantidad de energía suficiente para un trayecto determinado y porque con el control están previstos medios de entrada acoplados mediante los cuales un usuario del vehículo puede ajustar el tiempo o el periodo temporal en el que se puede llevar a cabo al menos parcialmente una descarga del acumulador y con ello una alimentación de la energía a la red de suministro.

Application ID: ES5708536	
Clasificación Oficina de PI	**Clasificación Modelo**
B60L	B60L Probabilidad: 0.68
H02J	B60K Probabilidad: 0.16
	H02J Probabilidad: 0.09
	PORCENTAJE TOTAL: 93%

Comentario

B60L y H02J son coincidentes en la clasificación oficial y la realizada por el Modelo. La clasificación adicional B60K propuesta por el Modelo no está muy relacionada con el objeto de la invención, ya que esta se refiere a un sistema de montaje de varios componentes de un vehículo sin especificar el tipo del mismo, en tanto que la invención se refiere a vehículos eléctricos y elementos que lo integran, es decir, no define con precisión el sector tecnológico por lo que podría considerarse irrelevante.

3. Perfeccionamientos en hornos para el tratamiento de polvo deamolado y similares. perfeccionamientos en hornos para el tratamiento de polvo de amolado y similares para eliminar contaminantes del mismo y producir material metálico refundible caracterizados porque dichos hornos se forman mediante una cámara de combustión alargada con un eje ligeramente inclinado con respecto a la horizontal medios para hacer girar la cámara alrededor de dicho eje medios para alimentar material al extremo superior de la cámara y para recoger el material al extremo superior de la cámara y para recoger el material descargado del extremo inferior de la cámara y por lo menos un quemador para caldear en general a lo largo de la longitud de la cámara.

Application ID: ES31560538	
Clasificación Oficina de PI	**Clasificación Modelo**
C22B	C22B Probabilidad: 0.20
F27B	F27D Probabilidad: 0.12
	F27B Probabilidad: 0.11
	PORCENTAJE TOTAL: 43%

Comentario

La clasificación que se diferencia de la clasificación oficial es la F27D, esta clasificación se encuentra dentro del campo tecnológico de la invención por lo que su proximidad es alta considerándola como una clasificación válida.

4. Procedimiento de una sola etapa para separar componentes de la biomasa. un procedimiento para la obtención de celulosa de la biomasa en una forma susceptible de sufrir hidrólisis enzimática comprendiendo dicho procedimiento poner en contacto la biomasa con una mezcla de un disolvente orgánico inmiscible en agua un ácido y un catalizador de sal metálica disuelto en la solución de agua ácida a una temperatura en el intervalo y una presión en el intervalo de a bar y filtrando la mezcla de reacción a presión para separar la lignina disuelta la hemicelulosa en la fase de disolvente y acuosa respectivamente y dejando atrás la celulosa pura en el que el catalizador de sal metálica se selecciona de sulfato de cobre sulfato ferroso sulfato de amonio ferroso sulfato de níquel sulfato de sodio y cloruro férrico y en el que el catalizador está presente en el intervalo del al de agua ácida.

Application ID: ES105851259	
Clasificación Oficina de PI	**Clasificación Modelo**
D21C	C08B Probabilidad: 0.42
C12P	D21C Probabilidad: 0.41
C13K	D01F Probabilidad: 0.06
C12M	PORCENTAJE TOTAL: 89%

Comentario

La clasificación que se diferencia de la clasificación oficial es la C08B y D01F. C08B se relaciona con productos de celulosa mientras que D01F se relaciona con la fabricación de fibras artificiales, no es aplicable esta clasificación para la invención descrita

La clasificación más idónea es la D21C relacionada con la producción de celulosa.

5. Estructura de anclaje y método para la construcción de la misma. en una estructura de un cuerpo de anclaje se inserta una vaina deformada hecha de acero y que tiene un cuerpo tubular y cuya pared periférica presenta una con ración irregular en un agujero excavado se inserta un material de tensión por lo menos en la vaina deformada varios tubos de acero se cubren sobre por lo menos una parte extrema de una pared periférica exterior de la vaina deformada o se insertan en por lo menos una parte extrema de una pared periférica interior de la vaina deformada para reforzar esta en el agujero excavado se introduce un endurecedor y se deja un espacio dentro de la vaina deformada para formar una sección de fijación también se presenta un método para la construcción del cuerpo de anclaje.

Application ID: ES5468252	
Clasificación Oficina de PI	**Clasificación Modelo**
E02D	E21D Probabilidad: 0.74
A47K	E02D Probabilidad: 0.04
	PORCENTAJE TOTAL: 78%

Comentario

La clasificación más idónea es E02D que se refiere a estructuras subterráneas o subacuáticas, la cual ha sido propuesta por la clasificación oficial y el Modelo. La clasificación E21D no es aplicable para este ejemplo ya que esta agrupa a invenciones relacionadas a espacios físicos subterráneos y no a objetos como el sistema de anclaje de presente caso.

6. Mejoras en calefactores para nidos de animales domésticos. mejoras en calefactores para nidos de animales domésticos consisten en partir de una caja portátil el fondo de la cual posee dos aberturas principales y que comunican con aberturas y de la nidera y una tercera abertura intermedia menor que comunica con el exterior entre cuyas aberturas y hay una resistencia con un termostato de seguridad conectado a un avisador sonoro y otro luminoso entre las aberturas y está situado un ventilador que establece una corriente de aire desde las aberturas y hasta la abertura es posible trasladar fácilmente el grupo calefactor de un nido a otro.

Application ID: ES31819642	
Clasificación Oficina de PI	**Clasificación Modelo**
F24C	A01K Probabilidad: 0.67
A01K	F24C Probabilidad: 0.06
	PORCENTAJE TOTAL: 73%

Comentario

La clasificación F24C y A01K es similar en ambos casos y se consideran las más adecuadas para definir el objeto de la invención descrito en el presente ejemplo, esto es sistemas de calefacción para crías de animales.

7. Módulos de circuito integrado y tarjetas inteligentes que incorporan los mismos. un módulo de circuito integrado para una tarjeta inteligente tanto con interfaces de contacto como sin contacto comprendiendo el módulo de ci un sustrato no conductor que tiene una pluralidad de agujeros de enlace simple y un par de agujeros de enlace múltiple extendiéndose ambos desde un primer a un segundo lado del sustrato en el que cada agujero de enlace múltiple es más grande que cada agujero de enlace simple y está con rado para recibir múltiples elementos conductores en el que cada agujero de enlace simple está con rado para recibir un único elemento conductor una pluralidad de almohadillas de contacto conductoras que incluyen un primer par de las mismas dispuestas en el primer lado del sustrato un primer chip de ci dispuesto en el segundo lado del sustrato una pluralidad de primeros elementos conductores que atraviesan el enlace simple y el par de agujeros de enlace múltiple y que conectan eléctricamente al menos algunas de las almohadillas de contacto al primer chip de ci en el que los primeros elementos conductores incluyen un primer par de primeros elementos conductores que atraviesan respectivamente el par de agujeros de enlace múltiple y que conectan eléctricamente el primer par de almohadillas de contacto al primer chip de ci y caracterizado por un encapsulante depositado en el primer chip de ci los primeros elementos conductores y el primer par de almohadillas de contacto en el que un borde del encapsulante depositado en el primer par de almohadillas de contacto distribuye cada uno del par de agujeros de enlace múltiple en un primer canal de enlace encapsulado y un segundo canal de enlace no encapsulado contiguo que terminan respectivamente en una primera y una

segunda área de enlace contigua en cada una del primer par de almohadillas de contacto de tal manera que el primer par de primeros elementos conductores atraviesan respectivamente el primer canal de enlace del par de agujeros de enlace múltiple y además de tal manera que el encapsulante sella la primera área de enlace y expone la segunda área de enlace a través del segundo canal de enlace para proporcionar una superficie para establecer una conexión eléctrica con el primer chip de ci en el que el encapsulante separa las áreas de enlace primera y segunda entre sí sin necesitar la presencia del sustrato entre las mismas

Application ID: ES321304225	
Clasificación Oficina de PI	**Clasificación Modelo**
G06K	H01L Probabilidad: 0.52
H05K	H05K Probabilidad: 0.22
	G06K Probabilidad: 0.16
	PORCENTAJE TOTAL: 90%

Comentario

La clasificación H01L, la cual difiere de la clasificación oficial, se encuentra en el ámbito tecnológico del objeto de la invención descrito en el presente ejemplo. Las invenciones relacionadas con electricidad y circuitos tienen varios campos de clasificación, por lo que la clasificación más precisa va a estar determinada por el grupo y subgrupo de clasificación; no obstante H01L define el sector tecnológico macro de la invención y si se consideraría una clasificación válida.

8. Sistema eléctrico que utiliza ca de alta frecuencia y que tiene cargas conectadas inductivamente y fuentes de alimentación y luminarias correspondientes. un sistema de distribución de potencia para distribuir potencia ca de alta frecuencia desde una fuente de potencia limitada actual el sistema comprende un par trenzado de conductores alargados con rados para conectarse a la fuente de alimentación los extremos de los conductores más alejados estando la fuente de alimentación conectada entre sí donde los conductores entre las vueltas adyacentes del par trenzado son separables entre sí para definir una abertura entre ellos un elemento de toma de potencia que está con rado para ser insertado al menos parcialmente a través de la abertura para que la energía eléctrica pueda acoplarse inductivamente desde los conductores al elemento de toma de corriente donde otro conductor se embobine al menos parcialmente alrededor del elemento de toma de corriente caracterizado porque el sistema comprende además un circuito de rectificación síncrono conectado al conductor adicional para convertir la potencia de corriente alterna de alta frecuencia en el conductor adicional a un voltaje regulado por cc para potencia una carga y un rectificador en el circuito de rectificación síncrono para cortocircuitar de manera controlable el circuito de rectificación para detener la transmisión de potencia a la carga.

Application ID: ES217481931	
Clasificación Oficina de PI	**Clasificación Modelo**
H05B	H02J Probabilidad: 0.44
H02J	H02M Probabilidad: 0.32
H01F	H01F Probabilidad: 0.04
H02M	PORCENTAJE TOTAL: 80%

Comentario:

La clasificación del Modelo es coincidente con la clasificación en clase y subclase de la asignada por la oficina de propiedad intelectual en la cual se presentó la solicitud de patente. La clasificación se relaciona con dispositivos y mecanismos de conversión de energía, mismo que corresponde al sector tecnológico en el cual se enmarca la patente descrita en el presente ejemplo, por lo que se concluye que la clasificación asignada por el Modelo es válida.

9. Procedimiento para la descarga de un acumulador de energía eléctrica. procedimiento para la descarga de un acumulador de energía eléctrica que está conectado con un circuito electrónico mediante un primer conductor eléctrico y un segundo conductor eléctrico en donde está proporcionado un tiristor para la descarga del acumulador de energía eléctrica en donde en el procedimiento debido a un error que ocurre en el circuito electrónico una corriente de descarga del acumulador de energía eléctrica comienza a fluir desde el acumulador de energía eléctrica a través del primer conductor eléctrico al circuito electrónico y a través del segundo conductor eléctrico de regreso al acumulador de energía eléctrica debido a la corriente de descarga alrededor del primer conductor eléctrico y del segundo conductor eléctrico se genera un campo magnético que se modifica con el tiempo el cual atraviesa el material semiconductor del tiristor por el campo magnético que cambia con el tiempo una corriente es inducida en el material semiconductor del tiristor caracterizado porque el tiristor se activa mediante esta corriente inducida.

Application ID: ES320284468	
Clasificación Oficina de PI	**Clasificación Modelo**
H02M	H02J Probabilidad: 0.69
H03K	H02M Probabilidad: 0.12
	H01M Probabilidad: 0.12
	PORCENTAJE TOTAL: 93%

Comentario

La clasificación H02M es la más cercana al objeto de la patente y se refiere a aparatos para conversión de energía, esta clasificación es semejante tanto en la clasificación oficial como aquella asignada por el Modelo. La clasificación H02J que difieren en el Modelo, se considera válida ya que se ubica en el sector tecnológico de la invención, esto es sistemas de almacenamiento de energía y generación, conversión o distribución de energía eléctrica; por su parte, la clasificación H01M no se relaciona directamente porque se refiere a la conversión de energía química en energía eléctrica, no siendo esto el objeto de la invención.

10. método y sistema de procesamiento de un macro. un método para realizar comunicaciones de datos dentro de un sistema a base de macros caracterizado porque comprende las etapas de recibir información a base de macros que contienen variables intercaladas a partir de un nodo del cliente recuperar en respuesta a la información a base de macros recibida y las variables intercaladas contenidas en el mismo datos que representen al menos una fila de macros particular y que modifiquen los datos recuperados como una función de las variables contenidas en los datos a base de macros recibidos transmitir los datos recuperados modificados al nodo del cliente por lo

que proporcionan al nodo del cliente con información reflectiva del estado previo del sistema a base de macros.

Application ID: MX33434	
Clasificación Oficina de PI	**Clasificación Modelo**
G06F	H04L Probabilidad: 0.31
	G06F Probabilidad: 0.30
	H04J Probabilidad: 0.07
	PORCENTAJE TOTAL: 68%

Comentario

La clasificación internacional idónea es G06F presentada tanto por la oficina de propiedad intelectual así como por el Modelo y se refiere a la trasmisión digital de datos eléctricos. Las clasificaciones adicionales H04L y H04J propuesta por el Modelo se enmarca en el sector tecnológico de la invención y se relacionan con transmisión de información digital pudiendo considerar como válida, no obstante la clasificación específica estará determinada por los grupos y subgrupos que forman parte del sistema de clasificación de patentes.

CONCLUSION.-

La clasificación internacional CIP es un sistema jerárquico, conformados por secciones, grupos, subgrupos, clases y subclases, que permite a las patentes ubicarlas en sectores tecnológicos específicos permitiendo que la recuperación de información en la búsqueda de este tió de información se realice de forma sistematizada. Esta clasificación es definida esencialmente por los inventores, sus solicitantes y también por la oficina de propiedad intelectual correspondiente.

Del análisis de los ejemplos presentados en este documento, se evidencia que la clasificación propuesta por el Modelo difiere en muchos casos en una o más clasificaciones adicionales a las asignadas por la Oficina de propiedad intelectual, estas clasificaciones mayoritariamente se encuentran en el sector tecnológico de la invención a la que pertenece, por lo que se puede considerar que el Modelo genera clasificaciones muy relacionadas a la invención, a nivel de clases y subclase que pueden orientar al analista o investigador a definir una clasificación más específica a través de la revisión exhaustiva del grupo y subgrupo que más sea aplicable.

-Fin del documento-

Despliegue del Modelo

ANEXO 2

In [1]:
```
import numpy as np
import tensorflow as tf
from keras.models import load_model
from keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
import pickle
```

In [2]:
```
# Carga el modelo guardado
modelo = load_model(r'C:\Users\modelo_3_incrustaciones.h5')
```

In [3]:
```
#cargar tokenizer del modelo
with open(r'C:\Users\danie\tokenizer.pickle', 'rb') as handle:
    tokenizer = pickle.load(handle)
```

In [4]:
```
#cargar etiquetas
ruta_archivo = r"C:\Users\mapa_etiquetas.pkl"
with open(ruta_archivo, 'rb') as archivo:
    map_etiqueta = pickle.load(archivo)
```

In [5]:
```
# Tokenizador y configuración utilizados durante el entrenamiento
MAX_SEQUENCE_LENGTH = 400
```

In [6]:
```
# Ejemplo de texto de patentes para clasificar
nuevas_patentes = ["automóvil con al menos un motor eléctrico  un acumulador de energía para proporcionar energía motriz para el motor eléctrico  con un conector enchufable conectado al acumulador de energía para la conexión a una red de suministr
```

In [7]:
```
secuencias = tokenizer.texts_to_sequences(nuevas_patentes)
datos = pad_sequences(secuencias, maxlen=MAX_SEQUENCE_LENGTH)
```

In [8]:
```
predicciones = modelo.predict(datos)
```

```
1/1 [==============================] - 3s 3s/step
```

In [9]:
```
for i, texto in enumerate(nuevas_patentes):
    print(f"Texto de la patente: {texto}")
    top_5_indices = np.argsort(predicciones[i])[::-1][:5]  # Obtiene los índices de las 5 clasificaciones más altas
    for j in top_5_indices:
        nombre_categoria = next(clase for clase, etiqueta in map_etiqueta.items() if etiqueta == j)
        print(f"Clasificación {j+1}: Categoría {nombre_categoria}, Probabilidad: {predicciones[i][j]}")
    print("\n")
```

```
Texto de la patente: automóvil con al menos un motor eléctrico  un acumulador de energía para proporcionar energía motriz para el motor eléctrico  con un conector enchufable conectado al acumulador de energía para la conexión a una red de suminist
ro y con un control que comprende un reloj para el control del flujo de corriente a través de la red de suministro y del acumulador de energía  en el que el control permite una flujo de corriente desde el acumulador de energía a la red de suminist
ro y el control comprende un dispositivo para el registro de la cantidad de carga en el acumulador de energía e interrumpe el flujo de corriente desde el acumulador de energía a la red de suministro al alcanzar un valor umbral prefijado de la ener
gía de carga residual restante  de manera que se garantiza una cantidad de energía suficiente para un trayecto determinado y porque con el control están previstos medios de entrada acoplados mediante los cuales un usuario del vehículo puede ajusta
r el tiempo o el periodo temporal en el que se puede llevar a cabo  al menos parcialmente  una descarga del acumulador y  con ello  una alimentación de la energía a la red de suministro.
Clasificación 165: Categoría B60L, Probabilidad: 0.6831147074699402
Clasificación 164: Categoría B60K, Probabilidad: 0.16171158850193024
Clasificación 457: Categoría H02J, Probabilidad: 0.09835444390773773
Clasificación 331: Categoría F02N, Probabilidad: 0.011404245160520077
Clasificación 461: Categoría H02P, Probabilidad: 0.007573160342872143
```