Método Interpretable de Clasificación Multietiqueta en Documentos usando Algoritmos Genéticos

Kelvin Paul Pucho Zevallos

Orientador: Prof. Yuber Elmer Velazco Paredes

UNSA - Universidad Nacional de San Agustín de Arequipa
Septiembre de 2023

Índice general

1. Introducción **1**

1.1. Planteamiento del Problema 2

1.1.1. Descripción de la Realidad Problemática 2

1.1.2. Problema Principal 2

1.2. Objetivos 3

1.2.1. Objetivo Principal 3

1.2.2. Objetivos Específicos 3

1.3. Justificación e Importancia de la Investigación. 3

1.3.1. Justificación 3

1.3.2. Importancia 4

1.3.3. Alcance 5

2. Trabajos Relacionados **6**

2.0.1. Enfoques basados en metadatos 6

2.0.2. Enfoques basados en contenido 7

2.0.3. Criterios de Evaluación 9

2.0.4. Análisis crítico de la revisión de la literatura 9

3. Marco Teórico **12**

3.1. Introducción 12

3.2. Conjunto de Datos 12

3.3. Aprendizaje Automático 13

3.3.1. Aprendizaje Automático en Clasificación multietiqueta de Documentos 14

3.3.2. Inteligencia Computacional 15

3.3.3. Inteligencia Artificial Explicable mediante word clouds 16

4. Metodología propuesta **17**

4.0.1. Recopilación y preprocesamiento de datos 17

4.0.2. Técnicas de limpieza y preparación de los documentos 17

4.0.3. Algoritmo de clasificación multietiqueta optimizado con Algoritmo Genético 18

4.0.3.1. Descripción detallada del algoritmo propuesto 18

4.0.4. Integración de word clouds en el proceso de clasificación 19

4.0.5. Integración de Algoritmos Evolutivos y Word Clouds 20

4.0.6. Selección 20

4.0.7. Cruzamiento 21

4.0.8. Mutación 21

4.0.9. Visualización 21

4.0.10. Integración de word clouds en el proceso de clasificación 22

4.0.10.1. Generación de las word clouds a partir de los documentos 22

Bibliografía **24**

Resumen

Título del Resumen

Capítulo 1

Introducción

El aprendizaje automático, como campo de estudio, se dedica al desarrollo de métodos que permiten extraer patrones significativos a partir de conjuntos de datos. El objetivo principal es utilizar estos patrones para tomar decisiones inteligentes y fundamentadas. En el ámbito de la clasificación, el aprendizaje automático se ha utilizado ampliamente para abordar este problema en diversas áreas. Sin embargo, es importante destacar que el enfoque de clasificación basado en el aprendizaje automático tiene sus limitaciones. Tradicionalmente, se ha utilizado para la predicción de clases predefinidas, lo que implica que solo puede clasificar nuevos datos en categorías previamente conocidas. Esto deja fuera muchos otros problemas de clasificación que pueden surgir en diferentes dominios y situaciones. En contraste, la inteligencia computacional se centra en resolver problemas que requieren un nivel de inteligencia comparable al de los seres humanos o animales. Este enfoque busca abordar desafíos complejos que no se pueden resolver fácilmente mediante métodos tradicionales. Entre los temas de interés en el campo de la inteligencia computacional se encuentran las redes neuronales, los sistemas difusos y los algoritmos genéticos. Estas áreas de estudio ofrecen herramientas y técnicas para resolver problemas que van más allá de la clasificación tradicional y que requieren un razonamiento más sofisticado.

En este proyecto en particular, se ubica en el cruce entre el aprendizaje automático y la explicable inteligencia computacional. Al emplear sistemas difusos y algoritmos genéticos, se busca abordar los desafíos de clasificación de documentos en el campo de Ciencias de la Computación. Esto es especialmente relevante dado que los investigadores generan una gran cantidad de información científica, y se estima que dicha información se duplica aproximadamente cada cinco años. La necesidad de manejar y organizar eficientemente esta creciente cantidad de datos es evidente, además de conocer el porque dicho documento pertenece a una o varias categorías y es aquí donde las herramientas provenientes del aprendizaje automático y la inteligencia computacional pueden desempeñar un papel crucial. Al utilizar sistemas difusos y algoritmos genéticos, este proyecto busca mejorar la capacidad de clasificación de documentos, y empleando una manera de explicabilidad permitiendo a los investigadores gestionar y analizar de manera más efectiva la información científica y saber el porque un documento pueda pertenecer a una o varias categorías. Esto puede tener un impacto significativo en el avance de la investigación en Ciencias de

la Computación, al facilitar la identificación de tendencias, la extracción de conocimientos relevantes.

1.1. Planteamiento del Problema

1.1.1. Descripción de la Realidad Problemática

El proceso de clasificación manual de documentos de diversas áreas se vuelve laborioso y propenso a errores, especialmente al lidiar con grandes volúmenes de documentos. La amplia variedad de disciplinas y temas dentro de un campo específico, como la Ciencia de la Computación, complica aún más la identificación de la temática principal de cada artículo. Además, clasificar documentos con múltiples etiquetas implica asignar varias categorías a un solo documento, lo que añade complejidad al proceso.

Los enfoques tradicionales, basados en características estáticas, pueden clasificar adecuadamente en función de términos de frecuencia, pero dependen de umbrales o características específicas para determinar si un documento pertenece a una o varias categorías. Sin embargo, estos algoritmos de aprendizaje automático no ofrecen una explicación clara de cómo llegaron a sus resultados o de por qué el documento pertenece a esas categorías específicas.

Por tanto, surge la necesidad de desarrollar una técnica avanzada que optimice estos modelos y permita explicar el proceso de clasificación. Es aquí donde entran en juego los algoritmos genéticos, que abordan este problema. Al emplear algoritmos genéticos, logramos mejorar la precisión de la clasificación de documentos con múltiples etiquetas y, al mismo tiempo, proporcionamos explicaciones comprensibles sobre por qué un documento ha sido clasificado de cierta manera. Este enfoque de inteligencia artificial explicable no solo mejora la precisión en la clasificación, sino que también permite a los usuarios determinar si un documento justifica su clasificación, lo que es esencial para la confianza y toma de decisiones informadas. La utilización de algoritmos genéticos se ha mostrado como una valiosa solución para abordar estos desafíos y brindar una comprensión clara y fundamentada del proceso de clasificación de documentos con múltiples etiquetas.

1.1.2. Problema Principal

Se busca resolver es el proceso laborioso y propenso a errores que representa la clasificación manual de documentos de diversas áreas, especialmente cuando se trata de grandes volúmenes de documentos. La amplia variedad de disciplinas y temas dentro de un campo específico, como la Ciencia de la Computación, dificulta aún más la tarea de identificar la temática principal de cada artículo. Además, el desafío se complica al tener que clasificar documentos con múltiples etiquetas, lo que implica asignar varias categorías a un solo documento. Por otro lado se sabe que enfoques tradicionales, basados en características estáticas, pueden clasificar adecuadamente en función de términos de

frecuencia, pero dependen de umbrales o características específicas para determinar la pertenencia del documento a las categorías. Sin embargo, estos algoritmos de aprendizaje automático no ofrecen una explicación clara de cómo llegaron a sus resultados o por qué el documento pertenece a esas categorías específicas.

1.2. Objetivos

1.2.1. Objetivo Principal

Proponer un nuevo método para la clasificación mutietiqueta de documentos usando algoritmos genéticos y considerando características de interpretabilidad.

1.2.2. Objetivos Específicos

- Analizar el contenido de los documentos para realizar el preprocesamiento de datos.
- Utilizar algoritmos genéticos para la modificación de pesos de las palabras para cada etiqueta
- Definir un método basado en pesos y cantidad de palabras para la generación de una nube de palabras
- Realizar el análisis experimental del método propuesto y validar los resultados con trabajos relacionados.

1.3. Justificación e Importancia de la Investigación.

1.3.1. Justificación

El uso de la inteligencia artificial explicable ha sido ampliamente explorado y aplicado en diversos campos, y uno de ellos es la clasificación de documentos en Ciencias de la Computación. La inteligencia artificial explicable se ha convertido en una herramienta invaluable para extraer patrones y conocimientos útiles a partir de grandes volúmenes de datos, permitiendo tomar decisiones inteligentes basadas en esos patrones identificados. Sin embargo, aunque la inteligencia artificial explicable ha demostrado su eficacia en muchos casos, la clasificación de documentos es un desafío complejo que presenta diversas problemáticas. Una de las limitaciones del enfoque de clasificación basado en inteligencia artificial explicable es su tendencia a depender de la interpretación de los resultados para explicar las decisiones tomadas por el modelo. Esto significa que el modelo de inteligencia artificial explicable puede clasificar documentos en categorías basadas en explicaciones claras y comprensibles que se generan junto con las predicciones. Si bien esto es útil en

muchas aplicaciones, hay situaciones en las que se necesita abordar problemas de clasificación más complejos, donde las explicaciones pueden ser ambiguas o insuficientes para tomar decisiones fundamentadas.

Aquí es donde entra en juego la inteligencia computacional. A diferencia del enfoque de aprendizaje automático, la inteligencia computacional se centra en resolver problemas que requieren un grado de inteligencia similar al que poseen los seres humanos y otros animales. La inteligencia computacional se basa en diferentes técnicas y algoritmos inspirados en la naturaleza y en el comportamiento de los sistemas biológicos.

Dentro del campo de la inteligencia computacional, se han investigado y desarrollado varias técnicas que resultan relevantes para la clasificación de documentos. Las redes neuronales, por ejemplo, han demostrado ser muy efectivas para el procesamiento y la clasificación de información en textos. Estas redes, inspiradas en el funcionamiento del cerebro humano, son capaces de aprender y reconocer patrones complejos en los datos, lo que las convierte en una herramienta valiosa para la clasificación de documentos en Ciencias de la Computación. Otra técnica relevante en el ámbito de la inteligencia computacional es la utilización de sistemas difusos. Estos sistemas permiten modelar y representar conocimientos vagos o imprecisos, lo que resulta especialmente útil cuando las categorías de clasificación no son claramente definidas. Los sistemas difusos se basan en la lógica difusa, que permite trabajar con valores intermedios y grados de pertenencia, lo que brinda mayor flexibilidad a la hora de realizar la clasificación de documentos ya que es posible que un articulo científico pertenezca a mas de una categoría.

Además, la computación evolutiva es otra área de interés en el campo de la inteligencia computacional para la clasificación de documentos. Esta técnica se basa en algoritmos genéticos y procesos de selección natural para buscar soluciones óptimas en un espacio de búsqueda amplio y complejo. La computación evolutiva puede utilizarse para optimizar la selección de características relevantes en los documentos, mejorar la precisión de los modelos de clasificación y adaptarse a cambios en los datos a lo largo del tiempo.

1.3.2. Importancia

La clasificación de documentos en el campo de Ciencias de la Computación se ha vuelto cada vez más crucial debido al aumento exponencial en el volumen de información científica generada por los investigadores. Con la duplicación de la información científica cada cinco años, resulta indispensable contar con métodos eficientes para organizar y clasificar estos documentos de manera efectiva.

En este contexto, el uso del aprendizaje automático y la explicable inteligencia computacional se ha destacado como una solución prometedora. Los sistemas difusos, por ejemplo, ofrecen una forma de abordar la naturaleza vaga e imprecisa de muchos documentos científicos. Al permitir la representación y el manejo de conocimientos inciertos, los sistemas difusos pueden ayudar a clasificar documentos en categorías adecuadas, incluso cuando los límites entre estas categorías no son claros. Esto contribuye a mejorar la precisión y la eficacia de los procesos de clasificación, evitando la ambigüedad y per-

mitiendo una organización más coherente de la información científica. Además se debe avanzar hacia soluciones de inteligencia artificial que sean más precisas, interpretables y útiles en contextos prácticos.

1.3.3. Alcance

El alcance de este proyecto de investigación se centra en la aplicación del aprendizaje automático y la inteligencia computacional para abordar los desafíos de clasificación de documentos en el campo de Ciencias de la Computación. El objetivo principal es superar los límites de la clasificación tradicional utilizando herramientas como sistemas difusos y algoritmos evolutivos. Estas técnicas permiten representar la información genética de las soluciones y utilizar la frecuencia de términos extraídos de los documentos como cromosomas para la clasificación. Además, se busca aprovechar la visualización de datos mediante word clouds basadas en la frecuencia de términos para identificar los términos clave y las tendencias en los artículos científicos. El límite de investigación se encuentra en la eficacia y eficiencia de los algoritmos evolutivos y la configuración de parámetros óptimos para obtener resultados precisos y relevantes en la clasificación de documentos de Ciencias de la Computación.

Capítulo 2

Trabajos Relacionados

En esta sección, se centra en el análisis crítico de todos los enfoques actuales, dado que cada estudio de investigación depende del estudio previo que ya se ha realizado en este campo. En la comunidad investigadora sobre la clasificación de documentos ha propuesto una serie de nuevas ideas para clasificarlos debido al aumento diario de documentos publicados. Los enfoques más avanzados propuestos en la literatura pueden dividirse en dos grandes categorías.

1. Enfoques basados en el contenido.
2. Enfoques basados en metadatos.

Los enfoques basados en los metadatos se centran en los metadatos del articulo y contienen el titulo, el abstract, los keywords, los datos del autor, etc., y estan disponibles públicamente mientras que los enfoque basados en contenido del articulo de investigación que contiene la introducción, la metodología y la conclusión, y el cual la mayoría de los artículos no están disponibles. Este es el motivo por el cual los investigadores muestran una mayor preferencia por los metadatos en lugar del contenido en sí.

2.0.1. Enfoques basados en metadatos

Los métodos actuales que se basan en metadatos utilizan la información descriptiva de los artículos de investigación para categorizar los documentos de estudio. Estos metadatos, como el título, autor, palabras clave, términos generales y categorías, están ampliamente disponibles y de fácil acceso, a diferencia del contenido completo de los datos, que no se encuentra disponible de forma gratuita. Esta circunstancia motiva a la comunidad de investigadores a centrarse en los metadatos de libre acceso de los documentos de investigación en lugar de en su contenido.

El siguiente paper [Yohan et al., 2014] propone una técnica de procesamiento de lenguaje natural para encontrar entidades de nombre y clasificarlas en sus respectivas

categorías, este método evalúa utilizando diferentes de datos como Newspaper y Telugu wiki, concluyendo con un rango de precisión que oscila entre 0.79 y 0.94.

Otro de los artículos [Godbole and Sarawagi, 2004] propone un enfoque para mejorar la clasificación de documentos utilizando información estructural y de citas. Este modelo combina los datos estructurales, como el título y el resumen, con las citas provenientes de otros artículos de investigación. Mediante esta combinación, lograron obtener resultados significativos en la clasificación de documentos.

Una de las propuestas basado en la lógica difusa [Sajid et al., 2016] para la categorización de artículos de investigación en el ámbito de la informática. Eligieron los conjuntos de datos JUCS debido a su cobertura en todas las áreas del campo de la informática. Después de realizar una evaluación exhaustiva del método, los resultados demostraron que este lograba una precisión del 0,91 y una medida F del 0,96 para la clasificación de una sola etiqueta.

Para la clasificación de documentos, el articulo siguiente [Wang and Desai, 2007] propone otro método de extracción de metadatos. Este enfoque propone un sistema ”post hoc”para clasificar los documentos. Realiza dos fase, la extracción de metadatos basados en plantillas, y la clasificación basada en los metadatos. Ellos evaluaron con un conjunto de datos diversificado del Centro de Información Técnica de Defensa (DTIC), que contiene un millón de datos de artículos científicos, tesis doctorales, trabajos de investigación de conferencias, revistas, diapositivas y documentos jurídicos, etc. Los resultados revelaron que este enfoque predice la categoría del documento 0,83 veces correctamente.

La clasificación multietiqueta de documentos científicos [Ali and Asghar, 2018], propusieron basada en características de metadatos. Este enfoque utiliza dos características de metadatos (título y palabras clave). Este método utiliza diferentes medidas de similitud para encontrar la relevancia entre los documentos y las etiquetas. Este método utiliza un clasificador basado en PSO para clasificar los documentos. Este enfoque se evalúa en dos conjuntos de datos diferentes de artículos de investigación (JUCS y ACM). Los resultados del estudio revelan que el método alcanza una precisión de hasta el 0,78 %.

2.0.2. Enfoques basados en contenido

Los enfoques basados en el contenido se basan en el contenido de los artículos de investigación. En un estudio realizado en 2015 por el articulo [Le and Ho, 2015] se examinaron todos los enfoques existentes para la selección de características en la clasificación de textos. En este estudio, se analizaron todos los métodos de selección y reducción de características, debido a que el contenido de un articulo es extenso y contiene palabras sin significado y estés se clasifican en dos categorías principales: wrapper y filter. Segun los investigadores se encontró que el rendimiento del método filter es considerablemente superior al del método wrapper, ya que filter no depende del algoritmo de clasificación utilizado. Por lo tanto en la literatura, la mayoría de los investigadores utilizan la técnica filter para clasificar documentos.

En 2016, el articulo [Zhou, 2016] propusieron un enfoque basado en el contenido utilizando algoritmos de Bayes ingenuo y regresión logística. Utilizaron dos conjuntos de datos diversificados de de ciencias de la computación, como: (1) CiteSeerX, (2) arXiv. Los resultados obtenidos son La puntuación F1 de los conjuntos de datos arXiv y CiteSeerX es de 0,95 y 0,75 respectivamente.

El articulo,[Zong et al., 2015] propusieron la similitud semántica en diferentes características para la clasificación de texto. El experimento se realiza en dos conjuntos de datos diferentes como (1) Routers-10 (2) 20-Newsgroups. Llevando a cabo una serie de experimentos en los conjuntos de datos Routers-10 y 20-Newsgroups y aplicó el algoritmo Support Vector Machine (SVM) logrando una puntuación F de 0,76 para 20-Newsgroups y 0,91 para los conjuntos de datos Router.

Otro de los articulos [Chekima et al., 2012] propuso un agente categorizador de documentos basados en la jerarquía del ACM mediante un Naive Bayes Classifer. Tras realizar un experimento con 1000 artículos de informática, obtuvieron un 91 % de precisión.

Similar articulo [Zong et al., 2015] donde documentos de texto basado en el clasificador SVM el articulo [Cai and Hofmann, 2004] usa este método pero con la diferencia del conjunto de datos WIPO-alpha Collection.

Otro método de clasificación jerárquica de texto multietiqueta [Baker and Korhonen, 2017], en el que se utiliza un modelo de red neuronal para la clasificación. Los resultados se han evaluado utilizando datos biomédicos. Los resultados concluyen que la clasificación a nivel de documento funciona mejor que la clasificación a nivel de frase.

Este enfoque basado en el contenido es el propuesto por [Rodrigues and Santos, 2009] contiene dos pasos principales, (1) Crear un conjunto de datos de un documento en forma de jerarquía multi-etiqueta, estos documentos fueron extraídos de la biblioteca digital ACM. (2) Desarrollar un modelo de clasificación de texto multietiqueta combinando varios algoritmos de clasificación. Los datos que extraen para representar las caracteristicas de un documento son el titulo, el resumen, las palabras clave. Los algoritmos de clasificación que usan son la relevancia binaria, conjunto de etiquetas de potencia, optimización mínima secuencial y Naive Bayes Multinomial, etc. Después de realizar un experimento exhaustivo, los resultados revelaron que la relevancia binaria combinada con Naive Bayes Multinomial tiene un rendimiento extraordinario y alcanza una medida f de 0,88 en comparación con otros clasificadores utilizados individualmente y combinados.

Un método para la categorización eficiente de texto multietiqueta de los artículos de investigación dichos en este articulo [Jindal, 2018]. Este texto se enfoca en utilizar el análisis léxico y semántico para categorizar documentos de texto con múltiples etiquetas. Se utiliza la taxonomía IEEE para identificar los tokens en los artículos de investigación, y luego se analizan las relaciones entre ellos utilizando la base de datos léxica WordNet. A continuación, se realiza la clasificación basada en la taxonomía IEEE. El método fue evaluado en 150 artículos de informática, y los resultados mostraron una precisión de hasta el 0,75.

2.0.3. Criterios de Evaluación

Los criterios de evaluación surgidos tras el estudio de la literatura relacionada con la clasificación de documentos indican la existencia de diversas variables a considerar. Se identifican diferentes tipos de datos utilizados, provenientes tanto de metadatos como del contenido de los documentos.

- **Tipo de datos:** Se utilizan diferentes fuentes de datos para la clasificación, las fuentes de datos pueden ser de dos tipos: metadatos y contenido de documentos.
- **Tipo de clasificación:** Se utilizan dos tipos de clasificación: clasificación binaria y clasificación multi clase.
- **Conjuntos de datos:** Se utilizan diferentes conjuntos de datos en el enfoque y también se presenta la cantidad de datos utilizada.
- **Algoritmo de clasificación:** Se utiliza diferentes tipos de algoritmos o metodologías en el enfoque para la clasificación de documentos.
- **Resultados:** Después de realizar experimentos, se presentan los resultados obtenidos.
- **Medidas de evaluación:** Existen diferentes medidas para evaluar el rendimiento de un clasificador, como la precision, la recall, la f-score y la accuracy.

Después de una exhaustiva revisión de la literatura, este estudio llegó a la conclusión de que todos los enfoques mencionados anteriormente utilizan tanto el contenido como los metadatos de los documentos. Por lo tanto, las fuentes de datos se dividen en dos categorías principales.

1. **Enfoques basados en el contenido:** Los enfoques basados en el contenido utilizan el contenido general o el texto del documento y clasifican los documentos en sus respectivas clases únicas o múltiples a las que pertenecen.
2. **Enfoques basados en los metadatos:** El enfoque basado en los metadatos utiliza los metadatos de los documentos de investigación y clasifica los documentos en sus respectivas clases únicas o múltiples a las que pertenecen.

2.0.4. Análisis crítico de la revisión de la literatura

En el enfoque basado en el contenido, se utiliza el contenido general del documento, mientras que en el enfoque basado en metadatos, se utilizan los metadatos del artículo. Los metadatos del artículo están disponibles libremente, mientras que el contenido del artículo de investigación necesita suscripción para acceder a los datos.

Hay diferentes conjuntos de datos utilizados por los investigadores en la literatura como:

- arXiv
- JUCS
- ACM
- Newspaper
- 20-Newsgroups
- Teluguwiki
- CiteSeerX
- Reuters-21578

Existen diferentes algoritmos utilizados para la clasificación de artículos de investigación como:

- Expectation Maximization (EM)
- A Naive Bayes Classifier
- SVM
- Fuzzy Based Rule Merger
- PSO based classifier
- Logistic regression

En el área de clasificación de trabajos de investigación, la mayoría de los enfoques realizan clasificación de etiqueta única y existen muy pocos enfoques que han realizado clasificación de etiqueta múltiple. Con el aumento del número de funciones, su complejidad aumenta, por lo que la optimización de funciones es una tarea importante para la clasificación de documentos. En la literatura, casi todos los investigadores han realizado combinaciones manuales de funciones en lugar de optimizarlas.

Este estudio tiene como objetivo clasificar los artículos de investigación utilizando metadatos como características individuales. Para comparar nuestra técnica hemos utilizado Ali y Asghar [Khan et al., 2019]. Han realizado una clasificación de etiquetas múltiples utilizando la jerarquía ACM. Utilizaron conjuntos de datos JUCS. Realizaron una clasificación multietiqueta. Su enfoque obtiene una puntuación de hasta el 91 % en el conjunto de datos JUCS. Este estudio se enfoca en usar Metadatos individualmente para la clasificación del documento de investigación y consta de dos pasos principales: extracción de características y clasificación difusa. En el paso de extracción de características, los autores utilizan la frecuencia de términos (TF) para extraer características del título y las palabras clave de cada documento. En el paso de clasificación difusa, los autores utilizan un clasificador difuso para asignar cada documento a una o más categorías. En

el paso el proceso de TF es una matriz que contiene las frecuencias de términos de cada documento por cada regla, donde cada uno de ellas se puede evaluar y preprocesar para tomar como una poblacion de cada documento y aplicar una optimización en función de lograr un buen resultado de la precision y recall mediante programación evolutiva.

Capítulo 3

Marco Teórico

3.1. Introducción

El análisis crítico de los enfoques ya presentados en la revisión de la literatura revela que la comunidad de clasificación de artículos de investigación ha propuesto diversas técnicas para categorizar estos artículos en categorías únicas y múltiples. Las principales observaciones extraídas de la revisión de la literatura que han motivado y respaldan el marco propuesto son las siguientes: 1) No se ha llevado a cabo ningún estudio que utilice meta heurísticas para optimizar modelos de clasificación de multietiquetas y 2) Mostrar una visualización de palabras que determinen a que categoría pertenece un documento. Estas observaciones nos han llevado a proponer una técnica para abordar los problemas mencionados anteriormente. El marco propuesto clasifica los artículos de investigación en un sistema de clasificación JUCS predefinido mediante una evaluación exhaustiva de los metadatos. Además, se utiliza Algoritmos Genéticos, para la optimización de funciones mediante meta heurísticas. En este capítulo se aborda el marco teórico necesario para entender el método de clasificación de documentos de Ciencia de la Computación multietiqueta propuesto en la presente tesis.

3.2. Conjunto de Datos

La selección de conjuntos de datos es un paso crucial para realizar una evaluación exhaustiva del sistema propuesto. Con el fin de evaluar la técnica propuesta de manera precisa, se ha realizado una cuidadosa selección de un conjunto de datos diversificado que abarca artículos de investigación en el campo de las Ciencias de la Computación. Este conjunto de datos incluye artículos de investigación del Journal of Universal Computer Science (JUCS) [Wang and Desai, 2007]. La elección del conjunto de datos JUCS se debe a que contiene documentos que abarcan múltiples áreas del campo de la informática, lo cual es fundamental para lograr una evaluación integral. En la siguiente sección se ofrece una descripción detallada de este conjunto de datos seleccionado.

Cuadro 3.1: Descripción de las características y registros

Características	Registros
Número total de documentos de investigación	1863
Porcentaje de artículos de investigación de etiqueta única	51 %
Porcentaje de artículos de investigación multietiqueta	49 %
Número total de clases en el nivel raíz	13
Metadatos de un documento de investigación	Title Keyword

En la Tabla 3.1 se presenta la cantidad de documentos utilizados en el estudio, los cuales serán particionados para el entrenamiento y las pruebas. Específicamente, el 70 % de los documentos se destinará al entrenamiento del modelo, mientras que el 30 % restante se reservará para las pruebas.

Además, se utilizarán dos metadatos importantes en este análisis: los títulos y las palabras clave (Keywords) de los documentos. Estos metadatos desempeñarán un papel fundamental en la investigación.

3.3. Aprendizaje Automático

El aprendizaje automático es un campo de estudio que implica el uso de algoritmos y modelos estadísticos para permitir a los sistemas informáticos mejorar su rendimiento en una tarea específica a través de la experiencia [Costa, 2021] . Es un subconjunto de la inteligencia artificial que se centra en el desarrollo de sistemas que puedan aprender de los datos y tomar decisiones basadas en ellos. Los algoritmos de aprendizaje automático pueden clasificarse a grandes rasgos en tres categorías: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo [Costa, 2021] . En este estudio nos enfocamos en el aprendizaje supervisado que consiste en entrenar un modelo sobre datos etiquetados, en los que se conocen las variables de entrada y salida, para predecir la variable de salida para nuevos datos de entrada [Costa, 2021]. El proceso de aprendizaje implica la presentación de un conjunto de datos de entrenamiento al modelo, que luego se ajusta a los patrones y características presentes en los datos para realizar predicciones precisas cuando se enfrenta a nuevos datos, es decir, al conjunto de prueba.

Además, en el aprendizaje automático, existen diferentes tipos de algoritmos según el tipo de tarea que se desee abordar, como algoritmos de clasificación para problemas de asignación a categorías predefinidas y algoritmos de regresión para problemas de estimación de valores numéricos.

El aprendizaje automático ha demostrado ser útil en una amplia variedad de aplicaciones, como la clasificación de correos electrónicos como spam o no spam, el reconocimien-

to de voz y de imágenes, la clasificación de documentos, la recomendación de productos y servicios en línea, y mucho más. Su capacidad para extraer patrones significativos de grandes cantidades de datos ha revolucionado numerosas industrias y continúa siendo una de las áreas más activas de investigación en el campo de la inteligencia artificial.

3.3.1. Aprendizaje Automático en Clasificación multietiqueta de Documentos

En el estado del arte se encuentran algunos modelos o algoritmos de aprendizaje automático para la clasificación multietiqueta que utilizan la frecuencia de términos como los siguientes:

- **Métodos de transformación de problemas:** Estos métodos transforman el problema de clasificación multietiqueta en uno o más problemas de clasificación binaria. Un enfoque consiste en utilizar la frecuencia de términos como una representación de características para cada documento y aplicar algoritmos de clasificación binaria como la regresión logística o las máquinas de vectores de soporte (SVM) para predecir la presencia o ausencia de cada etiqueta. [[Deutschman, 2021], [Bogatinovski et al., 2022]]
- **Métodos de adaptación de algoritmos:** Los métodos de adaptación de algoritmos generalizan los algoritmos de etiqueta única para hacer frente directamente a los datos de etiquetas múltiples. En este contexto, la frecuencia de términos puede utilizarse como representación de características, y algoritmos como MLkNN (Multi-Label k-Nearest Neighbors) o RAkEL (Random k-Labelsets) pueden aplicarse para realizar una clasificación multietiqueta [Bogatinovski et al., 2022]. Es importante señalar que estos métodos no son exclusivos de la frecuencia de términos y pueden aplicarse también con otras representaciones de características. La elección de la representación de rasgos depende del problema específico y de las características del conjunto de datos. En resumen, los métodos de transformación de problemas y los métodos de adaptación de algoritmos se utilizan habitualmente en tareas de clasificación multietiqueta. Estos métodos pueden utilizar la frecuencia de términos como representación de características para clasificar documentos con etiquetas múltiples.[Bogatinovski et al., 2022]

La clasificación multietiqueta de documentos es una técnica de aprendizaje automático que permite asignar múltiples etiquetas a un documento, en lugar de limitarse a una sola etiqueta. A medida que crece la cantidad de información disponible en el ámbito de la ciencia de la computación, resulta cada vez más desafiante clasificar y organizar eficientemente los documentos relacionados con múltiples temáticas o aspectos. La clasificación multietiqueta se presenta como una solución efectiva para esta problemática, ya que permite asignar de manera precisa y exhaustiva las etiquetas que describen las diferentes dimensiones de un documento.

Para lograr la clasificación multietiqueta de documentos, se utilizan algoritmos de aprendizaje automático, como clasificadores basados en árboles de decisión, máquinas de

soporte vectorial (SVM), redes neuronales o métodos de aprendizaje profundo, entre otros. Estos algoritmos aprenden a partir de un conjunto de datos de entrenamiento que contiene documentos previamente etiquetados [Shin et al., 2006]. El objetivo es que el clasificador aprenda patrones y relaciones entre las características de los documentos y las etiquetas correspondientes, de manera que pueda generalizar y asignar adecuadamente las etiquetas a nuevos documentos no vistos previamente.

La clasificación multietiqueta de documentos tiene aplicaciones significativas en diferentes áreas de la ciencia de la computación, como la organización de grandes repositorios de documentos académicos, la clasificación de noticias o artículos científicos por temas relacionados, la indexación y búsqueda eficiente de documentos en motores de búsqueda y la recomendación personalizada de contenido basado en múltiples aspectos de interés para los usuarios.

La importancia de la clasificación multietiqueta de documentos radica en su capacidad para manejar la complejidad y la diversidad temática que existe en los documentos relacionados con la ciencia de la computación, como el ejemplo del Computing Classification System (CCS). La clasificación multietiqueta permite una representación más precisa y detallada de los documentos que abordan múltiples aspectos o temáticas interrelacionadas. Esto es fundamental en un campo tan amplio y diverso como la ciencia de la computación, donde los documentos pueden cubrir una amplia gama de temas, como inteligencia artificial, seguridad informática, redes neuronales, bases de datos, entre otros.

3.3.2. Inteligencia Computacional

La Inteligencia Computacional es un subcampo de la Inteligencia Artificial (IA) que se centra en el desarrollo de algoritmos inspirados en sistemas biológicos para resolver problemas complejos del mundo real [Wik, 2023a]. Abarca una serie de tecnologías, como el aprendizaje automático, la lógica difusa y la computación evolutiva [[Wik, 2023a], [Xu et al., 2021], [Eiben and Smith, 2015]] . La Inteligencia Computacional se utiliza en diversos campos, como las finanzas, la sanidad y la ingeniería, para desarrollar sistemas inteligentes capaces de aprender de los datos y tomar decisiones basadas en ellos [Xu et al., 2021]. La computación evolutiva es un subconjunto de la Inteligencia Computacional que utiliza procesos evolutivos inspirados en la teoría de la evolución biológica para resolver problemas complejos de optimización [Eiben and Smith, 2015]. Los algoritmos de computación evolutiva son solucionadores de problemas de ensayo y error basados en poblaciones con un carácter metaheurístico o de optimización estocástica. Estos algoritmos implican técnicas que implementan mecanismos inspirados en la evolución biológica, como la reproducción, la mutación, la recombinación, la selección natural y la supervivencia del más apto. Las soluciones candidatas al problema de optimización desempeñan el papel de individuos de una población, y la función de coste determina el entorno en el que ”viven”las soluciones. Existen varios algoritmos utilizados en la computación evolutiva, como los algoritmos genéticos, la programación evolutiva, la programación genética y los modelos de inteligencia de enjambre, como la optimización de colonias de hormigas o la optimización de enjambres de partículas [[Xu et al., 2021], [Eiben and Smith, 2015]] . Los

algoritmos genéticos son uno de los más populares en la computación evolutiva. Se inspiran en el proceso de selección natural y utilizan una población de soluciones candidatas para evolucionar hacia una solución óptima. El algoritmo funciona creando una población inicial de soluciones, evaluando su idoneidad, seleccionando las mejores y aplicando operadores genéticos como el cruce y la mutación para crear nuevas soluciones.

3.3.3. Inteligencia Artificial Explicable mediante word clouds

La Inteligencia Artificial Explicable (XAI) es un subcampo de la Inteligencia Artificial (IA) que se centra en el desarrollo de algoritmos que puedan proporcionar resultados transparentes e interpretables a los usuarios. El objetivo de la XAI es hacer que la IA sea más transparente y comprensible para los seres humanos, especialmente en los casos en que las decisiones tomadas por los sistemas de IA pueden tener consecuencias importantes [[Wik, 2023b], [IBM, 2021]]. Los algoritmos XAI siguen los principios de transparencia, interpretabilidad y explicabilidad. La idea es crear un conjunto de técnicas que produzcan modelos más explicables manteniendo altos niveles de rendimiento [IBM, 2021]. Las nubes de palabras son una técnica de visualización que permite representar las palabras más frecuentes de un texto. Son una herramienta útil para resumir grandes cantidades de datos textuales e identificar las palabras clave y los conceptos más importantes [Overgaag, 2023] . Las nubes de palabras pueden utilizarse en XAI para visualizar las características o atributos más importantes que contribuyen a la clasificación de un documento o conjunto de datos. Analizando la nube de palabras, los usuarios pueden comprender mejor cómo toma decisiones el sistema de IA e identificar posibles sesgos o errores en el sistema [Overgaag, 2023]. Por lo tanto la XAI es importante porque ayuda a los usuarios de sistemas basados en IA a actuar con mayor eficacia al mejorar su comprensión de cómo razonan esos sistemas. La XAI puede mejorar la experiencia del usuario de un producto o servicio ayudándole a confiar en que la IA toma buenas decisiones [Wik, 2023b] . Además, los sistemas XAI se han centrado principalmente en hacer que los sistemas de IA sean comprensibles para los profesionales de la IA y no para los usuarios finales, y sus resultados sobre la percepción de estos sistemas por parte de los usuarios han sido desiguales.

Esta visualización de datos con nubes de palabras basada en la frecuencia de términos de artículos de ciencia de la computación es una técnica valiosa para identificar los términos más relevantes y las tendencias en el contenido de estos artículos. La frecuencia de términos se refiere al número de veces que aparece un término específico en los artículos, lo que permite medir su importancia y relevancia [Castella and Sutton, 2014]. Las word clouds, o nubes de palabras, son representaciones visuales de los términos más frecuentes, donde los términos comunes se presentan en mayor tamaño y los menos frecuentes en menor tamaño. Estas visualizaciones permiten una rápida identificación de los términos clave.

Capítulo 4

Metodología propuesta

La arquitectura propuesta en este artículo representa un enfoque innovador y efectivo para la clasificación multietiqueta de documentos de ciencia de la computación. La combinación de algoritmos evolutivos para la optimización del modelo y la visualización explicativa mediante el uso de word clouds permite obtener resultados de clasificación precisos y comprensibles. Este trabajo contribuye al avance de la clasificación de documentos y ofrece una herramienta valiosa para investigadores, profesionales y usuarios interesados en explorar grandes conjuntos de datos de ciencia de la computación de manera efectiva y eficiente.

4.0.1. Recopilación y preprocesamiento de datos

En esta sección, se selecciono el conjunto de datos de la Revista de Informática Universal (JUCS) [Shin et al., 2006], reconocida por su amplio espectro de artículos científicos en el campo de la informática. El conjunto de datos del (JUCS) presenta información relevante de artículos y para este proyecto se extrajo los titulos, palabras claves y el abstract de los articulos debido a que el enfoque es basado en metadatos por la gran cantidad de datos disponibles públicamente.

4.0.2. Técnicas de limpieza y preparación de los documentos

En esta etapa se aplicaron diversas técnicas de preprocesamiento de datos con el fin de garantizar la calidad y consistencia de los documentos utilizados en el modelo de clasificación multietiqueta.

- Eliminación de stopwords: Se realizaron filtrados para eliminar palabras comunes y poco informativas, como artículos y preposiciones, que no contribuyen significativamente a la clasificación multietiqueta.

- Eliminación de puntuación: Se eliminaron signos de puntuación, como comas, puntos y guiones, para evitar la interferencia de estos elementos en el proceso de clasificación.
- Normalización del texto: Se aplicaron técnicas de normalización, como la conversión a minúsculas y la eliminación de caracteres especiales, para asegurar una representación homogénea y coherente de los documentos.
- Tokenización: Se dividieron los textos en unidades más pequeñas, como palabras o n-gramas, para facilitar el análisis y el posterior procesamiento del modelo de clasificación multietiqueta.
- Stemming: Se utilizaron técnicas de Stemming para reducir las palabras a sus formas base, lo que permitió agrupar variantes de una misma palabra y simplificar el proceso de clasificación.

4.0.3. Algoritmo de clasificación multietiqueta optimizado con Algoritmo Genético

La optimización basada en algoritmos genéticos con algoritmos de clasificación multietiqueta es través de establecer y alterar términos de frecuencia de documentos que pertenecen a categorías como individuos en una población. Cada individuo representa una tabla de caracteristicas de las cuales esta compuesta por las categorias y los metadatos por las palabras mas frecuentes de los documentos en las categorias asociadas.

El proceso de optimización comienza con una población inicial aleatoria y alterada de los pesos de cada frecuencia. Luego, se evalúan y clasifican de acuerdo con una métrica de rendimiento (como la precisión y el recall) en un conjunto de validación.

A continuación, los algoritmos genéticos intervienen en el proceso evolutivo. Los clasificadores con mejor rendimiento tienen más probabilidades de sobrevivir y transmitir sus características a la siguiente generación. Además, se aplican operadores genéticos como la selección, el cruzamiento y la mutación para diversificar y mejorar la población.

Este ciclo de evaluación, selección y reproducción continúa durante varias generaciones hasta que se alcanza una convergencia deseada o se cumple un criterio de terminación predefinido.

4.0.3.1. Descripción detallada del algoritmo propuesto

En esta etapa se establece el Fuzzy Based Rules Merger (FBRM) Algorithm o el Algoritmo de fusión de reglas difusas (FBRM), el cual su función es extraer los títulos y las palabras claves de la dataset preprocesada y hallar el termino de frecuencia para cada categoría según sus metadatos o la unión de ellos, esto nos permite ver la frecuencia

de términos los documentos por categoría. La representación se basa en asignar valores numéricos a las categorías y palabras más frecuentes, generando así un cromosoma compuesto por aproximadamente 60000 genes. Se emplea una estrategia de mutación que consiste en modificar el peso de una palabra hacia atrás y hacia adelante en un 10 %. Este proceso de mutación se aplica para generar una población inicial.Se establecerán un total de 10 poblaciones mediante un enfoque de prueba y error. Cada población se evaluará utilizando un conjunto de datos de prueba para determinar si se produce una mejora en la macro accuracy promedio del clasificador multietiqueta. La mutación de los pesos de las palabras en un 10 % permite explorar diferentes combinaciones y ajustar gradualmente los valores para encontrar la configuración óptima. El objetivo principal de este enfoque es mejorar la accuracy macro promedio del clasificador multietiqueta al ajustar los pesos de las palabras en función de su frecuencia y relevancia. Al asignar valores específicos a las categorías y palabras más frecuentes, se espera que el algoritmo genere una población diversa y, eventualmente, converja hacia una configuración que mejore el rendimiento del clasificador en términos de precisión.

Para validar la eficacia de este enfoque propuesto, se realizarán experimentos comparativos utilizando diferentes conjuntos de datos de prueba y se analizará la mejora en la accuracy macro promedio obtenida en cada caso. Además, se explorarán otras métricas de evaluación y se considerarán posibles ajustes adicionales en el proceso de mutación para optimizar aún más los resultados.

4.0.4. Integración de word clouds en el proceso de clasificación

Además del algoritmo propuesto, se incorpora la utilización de nubes de palabras como una herramienta de visualización para representar de manera más concreta las categorías a las que pertenecen los documentos de ciencia de la computación. Las nubes de palabras, también conocidas como "word clouds", son representaciones gráficas que muestran las palabras más frecuentes en un texto, donde el tamaño de cada palabra se basa en su frecuencia o relevancia. En el contexto de este algoritmo, las nubes de palabras se generan a partir de los términos y pesos asignados a cada categoría. Cada categoría está asociada con un conjunto específico de palabras clave y, mediante el proceso de mutación y generación de poblaciones, se ajustan los pesos de estas palabras. La generación de las nubes de palabras permite visualizar qué palabras tienen una mayor influencia en la asignación de una categoría en particular. Esta visualización más concreta aporta una comprensión intuitiva de cómo se relacionan las palabras clave con las categorías y cómo influyen en la clasificación de los documentos de ciencia de la computación. Al observar las nubes de palabras, los investigadores y científicos de datos pueden identificar patrones, tendencias y enfoques temáticos relevantes en el campo.

Además de la visualización de las palabras clave, las nubes de palabras también pueden ayudar a identificar palabras o conceptos prominentes en cada categoría. Esto puede ser útil para comprender la naturaleza y el contenido de los documentos de ciencia de la computación en un nivel más profundo. Por ejemplo, si una categoría tiene palabras clave relacionadas con .ªprendizaje automático", redes neuronalesz .ªlgoritmos de clasificación",

es evidente que la categoría está relacionada con el campo de la inteligencia artificial y el procesamiento de datos.

4.0.5. Integración de Algoritmos Evolutivos y Word Clouds

En esta etapa se incorporan dos tecnologías utilizadas para la clasificación y visualización. En la parte evolutiva, los datos de entrada se representan mediante poblaciones de frecuencia de términos, como se muestra en la Figura 1.

	Names	Metadata	system	knowledg	comput	model	inform	manag	design	languag	...	semiautomata	luaproc	beyond	lzw	coupl	bigbatch	scorm	bilinear	scholarli	ciphertext
0	A	tf_keywords	1	16	0	0	4	8	1	0	...	0	0	0	0	0	0	0	0	0	0
1	A	tf_title	6	24	1	2	10	16	0	1	...	0	0	0	0	0	0	0	0	0	0
2	A	tf_keywords_title	7	40	1	2	14	24	1	1	...	0	0	0	0	0	0	0	0	0	0
3	B	tf_keywords	1	0	1	1	0	0	1	0	...	0	0	0	0	0	0	0	0	0	0
4	B	tf_title	0	0	5	1	0	0	4	0	...	0	0	0	0	0	0	0	0	0	0
5	B	tf_keywords_title	1	0	6	2	0	0	5	0	...	0	0	0	0	0	0	0	0	0	0
6	C	tf_keywords	13	2	1	6	3	2	4	1	...	0	0	0	0	0	0	0	0	0	0
7	C	tf_title	11	5	6	7	3	4	4	3	...	0	0	0	0	0	0	0	0	0	0
8	C	tf_keywords_title	24	7	7	13	6	6	8	4	...	0	0	0	0	0	0	0	0	0	0
9	D	tf_keywords	17	3	6	19	3	2	7	13	...	0	0	1	0	0	1	0	0	0	0
10	D	tf_title	24	4	18	25	10	7	11	29	...	0	1	0	0	1	0	0	0	0	0
11	D	tf_keywords_title	41	7	24	44	13	9	18	42	...	0	1	1	0	1	1	0	0	0	0
12	E	tf_keywords	0	0	1	2	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
13	E	tf_title	2	0	0	3	1	0	0	0	...	0	0	0	0	0	0	0	1	0	0
14	E	tf_keywords_title	2	0	1	5	1	0	0	0	...	0	0	0	0	0	0	0	1	0	1
15	F	tf_keywords	17	0	19	8	1	0	3	8	...	0	0	0	0	0	0	0	0	0	0
16	F	tf_title	29	0	49	19	3	1	4	19	...	1	0	0	0	0	0	0	0	0	0
17	F	tf_keywords_title	46	0	68	27	4	1	7	27	...	1	0	0	0	0	0	0	0	0	0
18	G	tf_keywords	7	0	7	0	0	0	2	1	...	0	0	0	0	0	0	0	0	0	0
19	G	tf_title	8	0	19	2	1	0	2	3	...	0	0	0	0	0	0	0	0	0	0
20	G	tf_keywords_title	15	0	26	2	1	0	4	4	...	0	0	0	0	0	0	0	0	0	0
21	H	tf_keywords	34	36	6	16	18	22	14	4	...	0	0	0	0	0	0	0	0	0	0
22	H	tf_title	46	63	30	27	40	39	23	12	...	0	0	0	0	0	0	0	0	1	0
23	H	tf_keywords_title	80	99	36	43	58	61	37	16	...	0	0	0	0	0	0	0	0	1	0
24	I	tf_keywords	14	8	2	2	7	4	7	3	...	0	0	0	0	0	0	0	0	0	0
25	I	tf_title	22	19	15	12	12	10	6	9	...	0	0	0	0	0	0	0	0	0	0
26	I	tf_keywords_title	36	27	17	14	19	14	13	12	...	0	0	0	0	0	0	0	0	0	0
27	J	tf_keywords	10	6	1	4	6	2	1	0	...	0	0	0	0	0	0	0	0	0	0
28	J	tf_title	8	10	6	9	11	3	5	2	...	0	0	0	0	0	0	0	0	0	0
29	J	tf_keywords_title	18	16	7	13	17	5	6	2	...	0	0	0	0	0	0	0	0	0	0
30	K	tf_keywords	10	8	3	9	8	7	5	3	...	0	0	0	0	0	0	0	0	0	0

Figura 4.1: Matriz de Cromosomas de Términos de Frecuencia por Categoría

El cromosoma representara la matriz completa en forma de un arreglo 1-dimensional de la siguiente forma:

$$C = p1, p2, p3, ..., pn \tag{4.1}$$

Donde:
C: representa el cromosoma
p: pesos de cada palabra

4.0.6. Selección

Para llevar a cabo la selección por torneos de cada cromosoma, se utiliza un clasificador fuzzy como primer paso. Este clasificador procesa cada documento de entrenamiento

para determinar las etiquetas que posiblemente corresponden a las categorías a las que pertenece el documento. Luego, se calcula el promedio de accuracy del macro average, que es una métrica utilizada para evaluar un clasificador multietiqueta. Este valor obtenido se utiliza como función fitness para determinar la selección de los padres. En este proceso, se selecciona aquel cromosoma que tenga el mayor valor posible para asegurar una mejor calidad en la selección de los padres.

4.0.7. Cruzamiento

En la parte del cruzamiento de un algoritmo genético, se puede utilizar una estrategia de cruzamiento por PLX para valores numéricos. Esta técnica se basa en la selección de puntos de corte aleatorios en los cromosomas de los padres, y a partir de ellos, se generan dos hijos. En el caso de valores numéricos, se realiza una operación de interpolación lineal para obtener los valores de los hijos entre los puntos de corte seleccionados. De esta manera, se garantiza que los valores numéricos de los hijos se encuentren dentro del rango establecido por los padres, lo que permite mantener la diversidad y explorar nuevas soluciones en el espacio de búsqueda. Esta estrategia es ampliamente utilizada en problemas de optimización numérica y ha demostrado ser efectiva en la mejora del rendimiento del algoritmo genético.

4.0.8. Mutación

En la parte de mutación de un algoritmo genético, se busca introducir variabilidad en los cromosomas para explorar nuevas soluciones en el espacio de búsqueda. Para lograr esto, se establece un rango de mutación para cada gen del cromosoma. Por ejemplo, si un valor es 10, que representa el peso de una palabra en una categoría, este se mutará entre su 10 % hacia atrás y hacia adelante, es decir, entre 9 y 11. Esto significa que el valor original del gen puede aumentar o disminuir en un 10 % de su valor actual. Este rango de mutación se establece para evitar cambios demasiado grandes en los valores de los genes, lo que podría llevar a soluciones no viables o a una pérdida de información importante. Al limitar la mutación a un rango específico, se asegura que los cambios sean graduales y controlados, lo que permite explorar nuevas soluciones sin comprometer la calidad de las soluciones ya encontradas.

4.0.9. Visualización

En un algoritmo de clasificación de documentos, se busca proporcionar una explicabilidad clara y concisa del por qué un documento pertenece a ciertas categorías. Para lograr esto, se utiliza una nube de palabras que lista los pesos más frecuentes de las palabras en el documento. Estos pesos se representan en una escala de colores, lo que permite una fácil interpretación visual de la información. Los pesos más altos se representan en

colores más oscuros, mientras que los pesos más bajos se representan en colores más claros. De esta manera, se puede identificar rápidamente las palabras más relevantes para la categorización del documento. Además, la nube de palabras proporciona una vista general del contenido del documento, lo que facilita la comprensión y la interpretación de los resultados del algoritmo. En resumen, la parte visual del algoritmo de clasificación de documentos es esencial para proporcionar una explicabilidad clara y concisa de los resultados obtenidos, lo que permite una toma de decisiones informada y precisa.

4.0.10. Integración de word clouds en el proceso de clasificación

4.0.10.1. Generación de las word clouds a partir de los documentos

El análisis de frecuencia de términos implica calcular y examinar la frecuencia de ocurrencia de los términos en los artículos de ciencia de la computación y que estas se tomaran como parte de la población para datos de entrada para el entrenamiento del algoritmo genéticos, y mediante los ensayos de prueba y error se estarán revelando aquellos términos más relevantes y comunes en el corpus de documentos que. Esto proporciona una comprensión más profunda de los temas y áreas de enfoque principales en esta disciplina. Antes de realizar el análisis, es importante realizar un preprocesamiento de datos, que incluye técnicas como la eliminación de stopwords, la normalización de palabras, stemming, y la eliminación de caracteres especiales. Estas técnicas aseguran que los datos estén limpios y preparados para el análisis de frecuencia de términos, el aprendizaje mediante algoritmos evolutivos y la generación de las word clouds.

Bibliografía

[IBM, 2021] (2021). What is the explainable artificial intelligence (xai)?

[Wik, 2023a] (2023a). Evolutionary computational.

[Wik, 2023b] (2023b). Explainable artificial intelligence.

[Ali and Asghar, 2018] Ali, T. and Asghar, S. (2018). Multi-label scientific document classification. *Journal of Internet Technology*, 19(6):1707–1716.

[Baker and Korhonen, 2017] Baker, S. and Korhonen, A. (2017). Initializing neural networks for hierarchical multi-label text classification. In *BioNLP 2017*, pages 307–315.

[Bogatinovski et al., 2022] Bogatinovski, J., Todorovski, L., Džeroski, S., and Kocev, D. (2022). Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications*, 203:117215.

[Cai and Hofmann, 2004] Cai, L. and Hofmann, T. (2004). Hierarchical document categorization with support vector machines. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 78–87.

[Castella and Sutton, 2014] Castella, Q. and Sutton, C. (2014). Word storms: Multiples of word clouds for visual comparison of documents. In *Proceedings of the 23rd international conference on World wide web*, pages 665–676.

[Chekima et al., 2012] Chekima, K., On, C. K., Alfred, R., Soon, G. K., and Anthony, P. (2012). Document categorizer agent based on acm hierarchy. In *2012 IEEE International Conference on Control System, Computing and Engineering*, pages 386–391. IEEE.

[Costa, 2021] Costa, C. D. (2021). Machine learning books you must read in 2020.

[Deutschman, 2021] Deutschman, Z. (2021). Multi-label text classification.

[Eiben and Smith, 2015] Eiben, A. E. and Smith, J. E. (2015). *Introduction to evolutionary computing*. Springer-Verlag Berlin Heidelberg.

[Godbole and Sarawagi, 2004] Godbole, S. and Sarawagi, S. (2004). Discriminative methods for multi-labeled classification. In *Advances in Knowledge Discovery and Data Mining: 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004. Proceedings*, volume 8, pages 22–30. Springer Berlin Heidelberg.

[Jindal, 2018] Jindal, R. (2018). A novel method for efficient multi-label text categorization of research articles. In *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, pages 333–336. IEEE.

[Khan et al., 2019] Khan, A. M., Shahid, A., Afzal, M. T., Nazar, F., Alotaibi, F. S., and Alyoubi, K. H. (2019). Swics: Section-wise in-text citation score. *IEEE Access*, 7:137090–137102.

[Le and Ho, 2015] Le, N. H. N. and Ho, B. Q. (2015). A comprehensive filter feature selection for improving document classification. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 169–177.

[Overgaag, 2023] Overgaag, A. (2023). What is explainable ai (xai)?

[Rodrigues and Santos, 2009] Rodrigues, F. and Santos, A. P. (2009). Multi-label hierarchical text classification using the acm taxonomy. In *Text Mining and Applications (TeMA'09) Track of EPIA09*.

[Sajid et al., 2016] Sajid, N., Afzal, M., and Qadir, M. (2016). Multi-label classification of computer science documents using fuzzy logic. *Journal of the National Science Foundation of Sri Lanka*, 44(2).

[Shin et al., 2006] Shin, K., Abraham, A., and Han, S. (2006). Enhanced centroid-based classification technique by filtering outliers. In *Text, Speech and Dialogue: 9th International Conference, TSD 2006, Brno, Czech Republic, September 11-15, 2006. Proceedings*, volume 9, pages 159–163. Springer Berlin Heidelberg.

[Wang and Desai, 2007] Wang, T. and Desai, B. C. (2007). Document classification with acm subject hierarchy. In *2007 Canadian Conference on Electrical and Computer Engineering*, pages 792–795. IEEE.

[Xu et al., 2021] Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., Wu, Y., Dong, F., Qiu, C.-W., Qiu, J., Hua, K., Su, W., Wu, J., Xu, H., Han, Y., Fu, C., Yin, Z., Liu, M., Roepman, R., Dietmann, S., Virta, M., Kengara, F., Zhang, Z., Zhang, L., Zhao, T., Dai, J., Yang, J., Lan, L., Luo, M., Liu, Z., An, T., Zhang, B., He, X., Cong, S., Liu, X., Zhang, W., Lewis, J. P., Tiedje, J. M., Wang, Q., An, Z., Wang, F., Zhang, L., Huang, T., Lu, C., Cai, Z., Wang, F., and Zhang, J. (2021). Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, 2(4):100179.

[Yohan et al., 2014] Yohan, P. M., Sasidhar, B., Basha, S. A. H., and Govardhan, A. (2014). Automatic named entity identification and classification using heuristic based approach for telugu. *International Journal of Computer Science Issues (IJCSI)*, 11(1):173.

[Zhou, 2016] Zhou, T. (2016). *Automated identification of computer science research papers*. PhD thesis, University of Windsor (Canada).

[Zong et al., 2015] Zong, W., Wu, F., Chu, L. K., and Sculli, D. (2015). A discriminative and semantic feature selection method for text categorization. *International Journal of Production Economics*, 165:215–222.