



Doctorado en Ciencias Informáticas

“Metodología integral de recomendación de
revisores en revistas científicas sobre múltiples
factores”

Directora de Tesis: Dra. Patricia San Martín
(CONICET-UNR)

Codirector de Tesis: Dr. Carlos Neil (FTI-UAI)

Doctorando: Mg. Guillermo Decoppet

Índice

Introducción	12
Problemática	12
Objetivos de la investigación	15
General.....	15
Específicos	15
Metodología	15
Estructura de la tesis doctoral	16
1 Mapeo Sistemático	18
1.1 Preguntas de investigación	18
1.2 Métodos de revisión	18
1.3 Elección de fuentes de datos	18
1.4 Criterios de inclusión y exclusión	20
1.5 Síntesis de datos extraídos.....	22
1.5.1 ¿Qué características generales definen agrupamientos iniciales del conjunto de artículos científicos seleccionados?.....	22
1.5.2 ¿Qué factores están incluidos en las propuestas de asignación automática de revisores?	24
1.5.3 ¿Qué modelos y técnicas de PLN son utilizadas en las propuestas?	28
1.5.4 ¿Cuáles son las metodologías utilizadas para las diferentes propuestas?	30
1.5.5 ¿Cuáles son las diferentes fuentes de datos utilizadas para cada uno de los factores incluidos en las diferentes soluciones?.....	44
1.6 Breves consideraciones	47
2 Fundamentos teóricos-metodológicos	50
2.1 Consideraciones previas	50
2.1.1 Síntesis del mapeo de metodologías	51
2.2 Perspectiva socio-técnica cultural	56
2.3 Argumentos para la discusión	58
2.4 Resultados de la encuesta.....	61

2.5	Breves reflexiones analíticas	64
2.6	Hacia una metodología integral	67
2.6.1	Globalizador de factores: Producción científica	67
2.6.2	Globalizador de factores: Autoridad académica del Investigador	75
2.6.3	Globalizador de factores: Revisiones anteriores	75
2.6.4	Definición de flexibilidad de la fórmula integral	76
3	Construcción metodológica integral	78
3.1	Adquisición, preprocesamiento y guardado de datos	78
3.1.1	Identificación y elección de fuentes de datos	78
3.1.2	Extracción de datos de las fuentes de información	79
3.1.3	Preprocesamiento y guardado de datos	80
3.2	Producción científica	82
3.2.1	Cálculo de Experiencia	83
3.2.2	Aplicación de Interés del investigador (Antigüedad)	91
3.2.3	Aplicación de Aporte de los autores (Orden)	92
3.2.4	Control de Conflicto de intereses	93
3.2.5	Cálculo de Puntaje de producción científica	94
3.2.6	Corte por no alcanzar el Puntaje mínimo de producción científica	95
3.3	Autoridad Académica	96
3.3.1	Aplicación de índices bibliométricos	96
3.3.2	Aplicación de titulación	99
3.3.3	Cálculo de Puntaje de autoridad académica	100
3.3.4	Gráfica de los procesos de autoridad académica	101
3.4	Revisiones anteriores	101
3.4.1	Registración de revisiones anteriores	102
3.4.2	Aplicación de Controles temporales y tiempos de no consideración	103
3.4.3	Aplicación de Penalizaciones (temporales - totales)	107
3.4.4	Parámetros generales de revisiones	110
3.4.5	Cálculo de Puntaje de revisiones anteriores	111

3.4.6	Gráfica de los procesos de revisiones realizadas	112
3.5	Fórmula de la metodología con parámetros integrales y con múltiples perfiles de configuración	113
3.5.1	Múltiples perfiles de configuración de la metodología	114
4	Prototipo y pruebas de laboratorio	115
4.1	Definiciones de tecnología.....	116
4.2	Definiciones asociadas a la metodología	116
4.2.1	Adquisición, preprocesamiento y guardado de datos	116
4.2.2	Producción Científica	136
4.3	Laboratorio de comparativas de partes representativas de artículos científicos	143
4.4	Análisis de técnicas de PLN y evaluación de expertos	150
4.5	Desambiguación y extracción de información de perfiles de Google Scholar	158
4.5.1	Autoridad Académica	165
4.5.2	Revisiones anteriores	166
4.5.3	Integración de la metodología con parámetros integrales y múltiples perfiles de configuración.	167
4.6	Validación experimental de la metodología propuesta en la revista IRICE ..	167
4.6.1	Resultados.....	170
	Conclusiones	173
	Aportes de la metodología propuesta:	174
	Líneas futuras de investigación:.....	177
	Bibliografía	179
	Anexo A	191
	Anexo B	200
	Anexo C	204
	Anexo D.....	206

Índice de tablas

Tabla 1 Preguntas de investigación y motivación	19
Tabla 2 Cadenas de búsqueda por cada repositorio y buscadores de artículos científicos	20
Tabla 3 Criterios de inclusión y de exclusión	21
Tabla 4 Agrupamiento de factores.....	28
Tabla 5 Agrupamiento de técnicas y modelos de PLN	31
Tabla 6 Recorte de factores incluidos en artículos científicos, desde 3 hacia arriba	32
Tabla 7 Recorte de factores incluidos en artículos científicos (hasta 2).....	38
Tabla 8 Relación entre técnicas y modelos de PLN, y modelo general	43
Tabla 9 Agrupamiento de las diferentes fuentes de información	45
Tabla 10 Partes de textos consideradas para evaluar experiencia	53
Tabla 11 Tipo de conflictos de intereses declarados	53
Tabla 12 Fuentes declaradas para "Autoridad académica"	54
Tabla 13 Estadísticas generales de datos guardados en el proceso de extracción	121
Tabla 14 Agrupación de registros de autores CONICET por nivel de titulación	122
Tabla 15 Estadísticas por acceso a la descarga	122
Tabla 16 Estadísticas por idioma en recursos, resúmenes y palabras claves	123
Tabla 17 Estadísticas del proceso de control y limpieza de palabras claves.....	125
Tabla 18 Diferencias en títulos entre idiomas según diferentes bibliotecas utilizadas	125
Tabla 19 Diferencias en resúmenes entre idiomas según diferentes bibliotecas utilizadas	126
Tabla 20 Diferencias en palabras claves entre idiomas según diferentes bibliotecas utilizadas	126
Tabla 21 Estadística comparativas iniciales y finales luego del proceso de depuración y limpieza	128
Tabla 22 Información importante para el cálculo de corte cuando no hay introducción	133
Tabla 23 Técnicas y metodologías de representación de PLN por cada parte de texto.....	141
Tabla 24 Preprocesamientos, cálculos y tipo de representación según diferentes técnicas y modelos de PLN	142
Tabla 25 Comparativa entre partes representativas del artículo científico con centro en texto completo, denominado en la tabla como archivo	144
Tabla 26 Comparativa entre partes representativas del artículo científico centrado en resumen	146

Tabla 27 Comparativa entre partes representativas del artículo científico con centro en texto completo	148
Tabla 28 Comparativas entre partes representativas del artículo científico con centro en archivo. Marcados en amarillo los que obtuvieron más puntaje por cada experto	152
Tabla 29 Comparativas entre artículos no repetidos entre textos completos y asociación a mejor metodología.....	154
Tabla 30 Comparativas entre artículos no repetidos entre textos completos de artículos y resúmenes	155
Tabla 31 Comparativas entre artículos no repetidos entre textos completos de artículos y resúmenes	156
Tabla 32 Comparativas entre mejor metodología elegida por cada experto y las diferentes técnicas de PLN sobre resúmenes, y entre todas las técnicas del informe de priorización y y las diferentes técnicas de PLN sobre resúmenes.	157
Tabla 33 Comparativas entre mejor metodología elegida por cada experto en su informe de evaluación y las diferentes técnicas de PLN sobre mejor representación, y entre todas las técnicas del informe de priorización y y las diferentes técnicas de PLN sobre mejor representación.	158
Tabla 34 Detalle discriminado por devoluciones de perfiles de investigador por cada nombre incluido en registros.....	159
Tabla 35 Estadísticas de perfiles confirmados sobre autores en el RI.....	160
Tabla 36 Estadísticas de perfiles por autor	160
Tabla 37 Estadísticas de autores por perfil	161
Tabla 38 Estadísticas de perfiles confirmados por indicadores	161
Tabla 39 Estadística de cantidad de perfiles candidatos y Google Scholar confirmados	162
Tabla 40 Estadística de cantidad de perfiles candidatos y coincidencia exacta entre nombres RI y Google Scholar confirmados	162
Tabla 41 Autores discriminados por cantidad de autores en cada publicación y verificación de orden basado en nombres	164
Tabla 42 Cantidad total de artículos en diferentes idiomas, luego del proceso de depuración inicial	191
Tabla 43 Estadísticas de reconocimiento de palabras que definen áreas de artículos científicos por búsqueda de títulos de cada área y asociado a los diferentes idiomas reconocidos en los recursos	191
Tabla 44 Resumen para análisis de los resultados obtenidos en los diferentes idiomas	198

Tabla 45 <i>Información del proceso de detección de ubicación de diferentes áreas promedio con definición de umbral para corte y muestra de artículos fuera de umbral</i>	200
Tabla 46 <i>Perfiles confirmados luego de primera etapa de búsqueda de perfiles</i>	204
Tabla 47 <i>Devolución de perfiles previa y posterior a la modificación de títulos</i>	204
Tabla 48 <i>Perfiles confirmados por cada etapa de procesamiento</i>	205

Índice de Figuras

Figura 1. <i>Distribución de trabajos encontrados por año</i>	22
Figura 2 <i>Distribución de artículos por tipo de solución propuesta</i>	24
Figura 3 <i>Factores incluidos en la extracción</i>	27
Figura 4 <i>Diagrama de preprocesamiento y guardado de datos</i>	82
Figura 5 <i>Diagrama del proceso de cálculo de experiencia</i>	96
Figura 6 <i>Diagrama del proceso de cálculo de autoridad</i>	101
Figura 7 <i>Diagrama del proceso de cálculo y cortes de revisiones</i>	112
Figura 8 <i>Recorte de un informe enviado a los expertos</i>	151

Agradecimientos

Quiero expresar mi más sincero agradecimiento a todas las personas e instituciones que hicieron posible la realización de esta tesis, especialmente a quienes, con su apoyo y comprensión, me acompañaron en los momentos más difíciles del proceso.

A mi directora de tesis, por su orientación, apoyo constante y valiosas sugerencias a lo largo de todo este proceso. Su experiencia y dedicación fueron fundamentales para el desarrollo de este trabajo.

También deseo agradecer a IRICE (CONICET-UNR) por brindarme los recursos y el espacio necesarios para llevar a cabo esta investigación. De manera especial, al equipo editorial de la Revista IRICE, cuya colaboración permitió aplicar y validar la metodología de este estudio. Además, a través de sus encuestas, aportó datos esenciales para la toma de decisiones metodológicas. También al investigador externo reconocido por su desempeño internacional en gestión y evaluación de sistemas de CyT que colaboró en esta etapa. Y a quienes evaluaron la encuesta para poder mejorar el instrumento.

A la Universidad, por permitirme realizar el doctorado y continuar con mi formación académica y especialmente a mi Codirector. Asimismo, agradezco a todas las personas de la institución que, con su apoyo y orientación, fueron clave en el desarrollo de este trabajo.

A los expertos que colaboraron evaluando un aspecto clave de esta investigación, aportando sus conocimientos y valiosas apreciaciones. Todos ellos, profesionales de primer nivel en cada una de sus áreas de conocimiento, hicieron una contribución invaluable a este trabajo.

A mi hijo, por estar a mi lado, brindándome su compañía, apoyo, comprensión y afecto en cada momento. A mis padres, por ser mi guía, mi ejemplo y mi mayor fuente de inspiración en esfuerzo y perseverancia. A mis amigos, por el valor de su amistad, su generosidad y por los momentos compartidos, que siempre llevaré conmigo.

Finalmente, a todas las personas que, desde distintos ámbitos personales, laborales y educativos, contribuyeron a mi crecimiento personal y aportaron valores y enseñanzas que siempre llevaré conmigo.

Resumen

Un problema recurrente que afrontan los editores de revistas científicas es la elección adecuada de revisores pares para evaluar los trabajos enviados por la comunidad académica en vistas a su posible publicación. En la región latinoamericana una cantidad significativa de editores científicos realizan esta tarea de manera artesanal sin contar con herramientas automatizadas de apoyo. A los fines de subsanar esta vacancia, la presente tesis doctoral propone una metodología innovadora utilizando técnicas validadas de Procesamiento del Lenguaje Natural. Se toma como caso de aplicación el repositorio institucional CONICET Digital (Argentina) y se tiene principalmente en cuenta aspectos tales como la concordancia entre la temática del artículo y de los evaluadores seleccionados, la priorización de la experiencia del revisor en la temática en el orden de recomendación, la actualidad del conocimiento, la detección del conflicto de interés, la desambiguación de nombres, la autoridad y la realimentación con información histórica de revisiones pasadas. En esta dirección, se desarrolló un primer prototipo que realiza el procesamiento de la información necesaria con dicha metodología, a los fines de ofrecer el informe de recomendación. En síntesis, el aporte original de esta investigación se centra en una propuesta metodológica para el desarrollo de una solución dinámica con las características expuestas, que se inscribe como una herramienta asistiva cuyos destinatarios son los editores expertos de revistas científicas que, como sistema con técnicas de IA posibilita una búsqueda ágil y adecuada de revisores para aplicar al corpus de trabajos científicos a evaluar.

Palabras Claves

Modelos de recuperación, Proceso de búsqueda, Proceso de selección, Inteligencia Artificial, Procesamiento del Lenguaje Natural, Revisión por pares.

Abstract

A recurring problem faced by editors of scientific journals is the appropriate selection of peer reviewers to evaluate papers submitted by the academic community for possible publication. In Latin America, a significant number of scientific editors perform this task in an artisanal manner without automated support tools. To address this gap, this doctoral thesis proposes an innovative methodology using validated Natural Language Processing techniques. The institutional repository CONICET Digital (Argentina) is used as a case study and primarily considers aspects such as the concordance between the article's topic and the selected reviewers' subject matter, the prioritization of reviewer expertise in the order of recommendation, the currency of knowledge, the detection of conflicts of interest, name disambiguation, authority, and feedback with historical information from past reviews. To this end, a first prototype was developed that processes the necessary information with this methodology to offer the recommendation report. In summary, the original contribution of this research focuses on a methodological proposal for the development of a dynamic solution with the characteristics described above, which is designed as an assistive tool aimed at expert editors of scientific journals. As a system with AI techniques, it enables a quick and appropriate search for reviewers to apply to the corpus of scientific works to be evaluated.

Keywords

Retrieval models, Search process, Selection process, Artificial Intelligence, Natural Language Processing, Peer reviewing.

Introducción

Problemática

Los trabajos enviados a revistas científicas deben ser evaluados para determinar si reúnen las condiciones necesarias para ser publicados, esto en general, es conocido como revisión por pares ciegos. La problemática de elegir los mejores revisores pares es de gran complejidad (Cuevas & Mestanza, 2002). La revisión por pares fue creada a principios del siglo XVIII por la Asociación Médica Británica y desde entonces se ha convertido en un factor clave para garantizar la calidad de las publicaciones científicas. Más allá de algunos cuestionamientos, es el método que se utiliza hasta la actualidad (Z.-Y. Tan et al., 2019). Asignar revisores a artículos científicos no es tarea sencilla para los comités editores de revistas científicas y los responsables de eventos académicos ya que es necesaria una correspondencia entre revisores y artículos. Esto demanda un tipo de precisión basado en el conocimiento y la experiencia del revisor sobre los temas que aborda el trabajo (Liu et al., 2014). La calidad de un evento académico y de una revista científica en gran medida está dada por la calidad de las revisiones, por lo cual dicha asignación debe ser tal que maximice la calidad de las evaluaciones que forman parte de la revisión (K. Li et al., 2017). A esta problemática también se le suma el problema del tiempo en que debe ser realizada la revisión. Dada esta situación, brindar herramientas para mejorar el proceso en tiempo y forma es de suma utilidad (Charlin & Zemel, 2013).

El aumento escalar de artículos de investigación y manuscritos publicables ha provocado una importante demanda sobre la cantidad de revisores expertos que solicitan los distintos eventos y publicaciones académicas. El principal problema es que una desacertada asignación de revisor puede producir una merma en la calidad o podría ocasionar el rechazo de un excelente trabajo científico. Estas situaciones han generado la necesidad de pensar en qué medida los desarrollos tecnológicos actuales podrían aportar algún tipo de solución a esta problemática (Protasiewicz et al., 2016).

Frente al crecimiento a nivel mundial de la cantidad de científicos y la producción científica, además de la creciente especialización del conocimiento, encontrar los mejores revisores para trabajos presentados en congresos y revistas científicas se ha vuelto muy complejo y laborioso. El crecimiento de la Inteligencia Artificial (IA) ha colaborado significativamente en la creación de propuestas automáticas de recomendación de revisores.

Diferentes artículos científicos abordan la problemática en cuestión. Desde principios del siglo XXI hasta la actualidad se puede observar que la necesidad de revisión por pares de artículos tanto de las revistas científicas como de los congresos están vinculadas. En la revisión

de la literatura se han relevado diferentes metodologías, modelos, estrategias, técnicas, algoritmos, sistemas, indicadores y enfoques en búsqueda de la mejor solución.

La importancia de la revisión por pares está dada en que es el proceso de validación central de los artículos científicos (Martínez, 2012) permitiendo verificar los contenidos y la metodología en los trabajos a publicar, el cumplimiento de estándares de calidad, la validez y la claridad en la redacción (Abad, 2017). Entre las primeras propuestas se puede mencionar una basada en documentos autodistribuidos (Riggs & Wilensky, 2001) y un método de calificación iterativa fundamentado en el filtrado colaborativo (Rigaux, 2004). Los enfoques de recomendación colaborativa sustentan sus predicciones en la detección de comportamientos similares entre usuarios. Existen dos enfoques principales, uno está basado en usuarios, en donde se calcula la similitud entre las preferencias de los usuarios y el otro, en los elementos donde la similitud se calcula entre estos. En esta dirección, se buscan elementos que tengan calificaciones similares de diferentes usuarios.

En 2005 se presentó otra propuesta alternativa utilizando la tecnología en todo el proceso de distribución y revisión (Rodríguez et al., 2005), proponiendo un modelo de revisión por pares que separa la distribución de la revisión y propone la utilización de los repositorios de OAI (Open Archive Initiative) para distribuir y por otro lado, un mecanismo de revisión que sea incluido en el protocolo OAI-PMH (Open Archive Initiative-Protocol for Metadata Harvesting) que propone la construcción de un modelo de comunicación en el ámbito académico basado en el repositorio y se sugiere crear una red de coautoría utilizando datos bibliográficos del DBLP (Proyecto de Biblioteca y Bibliografía Digital, del inglés Digital Bibliography & Library Project).

Otra solución (Rodríguez & Bollen, 2008) utilizó una estructura de red de coautoría con algoritmo de enjambre de partículas pero en una forma discreta relacionada a los algoritmos de andador aleatorio del análisis de la cadena de Markov y aumentando la cantidad de partículas. Se implementó el concepto de partículas de energía negativas para evitar el conflicto de intereses. Para la red de coautoría también se utilizaron los datos del DBLP y se validó usando datos de una conferencia conjunta del 2005 sobre Bibliotecas digitales.

Balog et al. (2006) plantean la búsqueda de expertos en diferentes organizaciones proponiendo dos maneras, ambas usando técnicas de lenguaje probabilístico útiles para la recuperación de información. Los autores asumen que los documentos se extraen de un depósito heterogéneo de documentos y la recomendación se realiza sobre la base de filtrado de contenido, proponiendo dos modelos. Uno de estos consiste en modelar el conocimiento de un experto sobre la base de sus documentos asociados. El otro, clasifica los documentos en tipos de conocimientos sobre determinados temas y procede a la búsqueda de expertos según sus documentos asociados. El resultado de la investigación fue que el segundo modelo dio mejores resultados en casi todas las pruebas.

Mimno & McCallum (2007) probaron varios métodos para definir la correspondencia de un revisor a un documento a evaluar. La propuesta utilizó modelos de lenguaje sin mezcla con suavizado Dirichlet, de AT (Autor-Tema), y APT (Autor-Persona-Tema) con el objetivo de encontrar expertos en lugar de documentos relevantes. En AT los documentos en el corpus están obligados a tener un solo autor y en APT los documentos por cada autor son fragmentados en uno o más grupos y cada grupo tiene una distribución de temas por separado. En este último modelo han obtenido los mejores resultados. La investigación implementó una base de datos denominada Rexa, que fue elaborada sobre trabajos de investigación descargados de internet, que extrae los autores, títulos y citas de los archivos en PDF.

Otra propuesta (Yang et al., 2009), menciona el concepto de inteligencia colaborativa utilizando Wikipedia, proponiendo una base de conocimiento incremental. Divide el problema en partes. Se inicia con el modelado de dominio clasificando la propuesta, representándola como una distribución de probabilidad multinomial en base a las palabras clave de la propuesta. Luego, continúa con el emparejamiento de expertos utilizando WCN (Wikipedia Concept Network), donde calcula la relación correlacionada entre expertos y propuesta midiendo la relación semántica entre la propuesta de expertos y las publicaciones y realiza la clasificación del experto con las puntuaciones de sus publicaciones. Finalmente, se arma de manera incremental la EKD (del inglés, Expert Knowledge Database) tomando a Wikipedia para modelar los caracteres de los dominios y clasificar la propuesta en dominios relacionados.

Todos estos artículos históricos dan cuenta de la posibilidad de tener diferentes miradas en la búsqueda de una solución al problema de recomendación de revisores (RAP, del inglés “reviewer assignment problem”). En vistas de la evolución de la IA y más específicamente del Procesamiento del Lenguaje Natural (PLN), en tanto sus métodos y técnicas, es probable el impacto en las actuales metodologías.

Un aspecto importante en la revisión por pares es encontrar los revisores adecuados en un lapso razonable. Por otra parte, la gran cantidad de artículos académicos publicados a nivel global tiene un crecimiento permanente, satisface entonces, la definición de big data, por lo cual el problema de recomendar revisores es de alcance mundial. Además, existe una gran dificultad en encontrar revisores con experiencia, teniendo en cuenta el conflicto de intereses (Salinas et al., 2020). Cabe destacar que es una tarea de alta complejidad detectar cuales son los candidatos a revisores que tengan una coincidencia en los intereses de investigación con el trabajo científico a revisar, entonces, para evitar ciertas subjetividades propias de estos procesos paulatinamente se van refinando diversas metodologías y sistemas validados científicamente.

En base a lo expuesto, se pretende que la metodología propuesta colabore en una adecuada selección y priorización de revisores de artículos científicos teniendo en cuenta

múltiples factores y la retroalimentación del equipo editorial. Al utilizar técnicas avanzadas del PLN se espera aportar una metodología innovadora y adecuada al ámbito de recomendación de revisores que pueda ser aplicada a un prototipo. La misma debe fundamentarse en los avances de PLN, la inclusión de múltiples factores intervinientes y retroalimentación en el proceso de recomendación para lograr la mejor recomendación que sirva de base para la toma de decisiones en la elección del revisor adecuado. La mayor limitación observable es la que posee cualquier metodología en este tipo de desarrollo que, al basarse en los avances, y siendo esta una rama del conocimiento con una dinámica de desarrollo constante es posible que nuevos conocimientos a futuro puedan mejorar la propuesta metodológica. En prospectiva, esto solicita indefectiblemente un permanente reestudio del estado del conocimiento para sopesar en que medida se debe realizar una readaptación de la metodología teniendo en cuenta avances que puedan mejorar el rendimiento y la efectividad de la recomendación.

Objetivos de la investigación

General

Proponer una metodología integral que utilice IA para el PLN a los fines de contribuir y apoyar la selección de revisores para artículos científicos incorporando una significativa multiplicidad de factores que contemplan los comités editoriales en dicha tarea.

Específicos

- Relevar analíticamente las variables intervinientes en la selección de revisores de artículos científicos definiendo cuáles de estas son tratadas en los enfoques metodológicos de las herramientas automatizadas de recomendación en la temática.
- Efectuar un reestudio del PLN con la finalidad de seleccionar los enfoques metodológicos más adecuados a la problemática de recomendación de revisores científicos.
- Elaborar una metodología configurando variables múltiples aplicadas al caso de revisores para artículos científicos en el campo de las Ciencias de la Educación a nivel de primer prototipo.
- Evaluar el primer prototipo de la metodología de recomendación propuesta con miembros de comité editoriales de revistas científicas del campo de las Ciencias de la Educación.

Metodología

La presente tesis tiene un enfoque propositivo. El desarrollo teórico y metodológico de la investigación se basa en un mapeo sistemático bibliométrico y la elección de un caso contextualizado. En el trayecto de la tesis doctoral se han implementado instrumentos y técnicas

cualitativas adhiriendo a un marco socio-técnico (Thomas et al., 2019). La elaboración del prototipo se basó en un enfoque iterativo, en el que se desarrollaron y probaron componentes de manera progresiva, corrigiendo y ajustando continuamente cada parte antes de integrar el sistema completo. Además, se realizaron pruebas integrales que permitieron identificar nuevas opciones y mejoras, lo que favoreció la evolución constante del prototipo. (Boehm, 1986)

El caso de referencia que tomó la presente investigación, es la Revista IRICE¹ que publica el Instituto Rosario en Ciencias de la Educación (IRICE) dependiente del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) y la Universidad Nacional de Rosario (UNR). Esta Revista cuenta con más de 25 años de trayectoria y en la actualidad, la búsqueda de revisores la realiza el consejo editorial sin ningún tipo de herramientas de apoyo para la toma de decisiones². Esta situación es también compartida por la Red de Revistas Académicas en Educación de la Argentina³. Por lo cual se estima que el caso es modélico y en consultas a la mencionada Red se ha detectado un alto interés en los posibles resultados de esta investigación en la comunidad de editores científicos de la disciplina.

El trabajo de campo con editores se desarrolló en dos etapas. En la primera, para la construcción de los fundamentos teóricos y la metodología propuesta, se llevaron a cabo entrevistas y encuestas semiestructuradas a investigadores con experiencia en comités editoriales. La segunda etapa consistió en la evaluación y ajustes del prototipo recomendador, en la cual se realizaron consultas y evaluaciones contextualizadas a dichos expertos. Mayores detalles sobre los instrumentos y técnicas implementadas se desarrollan en los capítulos que conforman este escrito. Cabe mencionar que todos estos insumos fueron fundamentales en el desarrollo socio-técnico cultural aquí propuesto.

Estructura de la tesis doctoral

A partir de lo expuesto, la presente tesis doctoral se inscribe en el ámbito de recuperación de información, específicamente como búsqueda de expertos, valiéndose de las técnicas validadas del PLN dentro del ámbito de la IA. En este sentido, el escrito está dividido en seis capítulos.

Luego de esta introducción, se aborda un mapeo sistemático del estado del conocimiento. El segundo capítulo desarrolla los fundamentos teóricos-metodológicos que sustentan los aspectos socio-técnicos culturales de la metodología innovadora propuesta. La

¹ Sitio oficial Revista IRICE: <https://ojs.rosario-conicet.gov.ar/index.php/revistairice/>

² Cabe mencionar que esta tesis doctoral tiene como antecedente una tesis de maestría de mi autoría defendida y acreditada en el año 2020 en el marco del Máster en Inteligencia Artificial de la Universidad Internacional de la Rioja (UNIR)(España) titulada "Recomendador de revisores de artículos de la revista IRICE".

³ RED de RED: <https://saieinvestigacionedu.wixsite.com/saieargentina/reddered>

misma se desarrolla en el tercer capítulo. En el cuarto, que integra la segunda etapa del trabajo de campo se presenta la aplicación de la metodología en un prototipo y se detallan los testeos y evaluaciones de laboratorio. Finalmente, se abordan las conclusiones, posibles líneas de trabajo a futuro, la bibliografía referenciada y los anexos documentales.

A modo de breve síntesis, lo propuesto integra un estado del conocimiento actualizado de PLN y la inclusión de múltiples factores: experiencia, conflicto de intereses, autoridad, intereses del investigador y retroalimentación de revisiones anteriores. Además, permite trabajar con datos de fuentes heterogéneas. En esta dirección, resulta imprescindible tener en cuenta en una recomendación aspectos relevantes tales como conflicto de intereses, desambiguación de nombres, autoridad. A la fecha, se ha observado en las diferentes metodologías, baja inclusión de retroalimentación del historial de revisiones. Además, cabe mencionar que, la mayoría de los métodos actuales utilizan la representación de características que no incluyen representación semántica. En esta dirección, se presenta un modelo que permite elegir la mejor representación del texto de cada producción científica disponible para la configuración de la representación semántica del artículo.

1 Mapeo Sistemático

Este capítulo da cuenta del primer objetivo específico de la presente tesis doctoral donde se realizó una revisión sistemática de la producción científica asociada a la problemática de recomendación de revisores con metodologías automatizadas publicados entre 2018 y 2023 (Decoppet & San Martín, 2024). En esta dirección, se aplicó un procedimiento sistemático en sitios de divulgación científica seleccionado finalmente veinticinco artículos de revistas científicas y conferencias. Este estudio analiza el dominio de las soluciones de recomendación de revisores, sus métodos, factores y los conjuntos de datos utilizados. El aporte se centra en un detallado estado del arte, específicamente los tipos diversos de soluciones propuestas y las similitudes entre los diferentes métodos, lo cual brinda información comparativa de alto nivel fundamental para el desarrollo de posibles soluciones.

1.1 Preguntas de investigación

El objeto de este trabajo es recopilar y analizar críticamente las producciones científicas asociadas a la problemática de recomendación automática de revisores en base al dominio de las soluciones de recomendación de revisores, sus métodos, los conjuntos de datos usados como fuente de información. En base a ello se plantean las siguientes preguntas (Tabla 1):

1.2 Métodos de revisión

El presente mapeo sistemático de la literatura adscribe a la metodología propuesta por (Kitchenham et al., 2009). Se estableció un protocolo bajo esos lineamientos teniendo en cuenta tres etapas elementales: 1) elección de las fuentes de artículos académicos primarios, 2) determinación de las cadenas de búsqueda a utilizar para buscar artículos académicos y 3) aplicación de criterios de inclusión y exclusión de artículos científicos que sirvan de base a este trabajo.

1.3 Elección de fuentes de datos

Se seleccionaron los principales repositorios y buscadores de artículos científicos a nivel mundial, determinando que los más destacados son: 1) ACM Digital Library, 2) Google Scholar, 3) IEEE Xplore y 4) ScienceDirect

Sobre la metodología presentada en (Kitchenham et al., 2009), se definieron los términos sobre los que se hacen las búsquedas. En primera instancia se buscaron los mapeos

sistemáticos encontrando solo el artículo de Aksoy, Yanik, y Amasyali, (Aksoy et al., 2023) donde seleccionaron 103 producciones científicas según criterios definidos. Luego se buscaron diversas palabras que puedan estar presentes en la mayoría de los artículos académicos que tratan sobre esta temática. En atención a toda esta información se seleccionaron las siguientes palabras: *Paper; Reviewer; System, Recommender; Assignment, Problem*

Tabla 1

Preguntas de investigación y motivación

Preguntas de investigación	Motivación
P 1 - ¿Qué características generales definen agrupamientos iniciales del conjunto de artículos científicos seleccionados?	Extraer y comenzar a clasificar información general sobre los artículos seleccionados para luego profundizar en las siguientes preguntas.
P 2 - ¿Qué factores están incluidos en las propuestas de asignación automática de revisores?	Determinar qué aspectos de la problemática son tenidos en cuenta al momento de plantear una solución.
P 3 - ¿Qué modelos y técnicas de PLN son utilizadas en las propuestas?	Precisar cuál es el grado de incorporación de las innovaciones del PLN en las propuestas
P 4 - ¿Cuáles son las metodologías utilizadas en las diferentes propuestas?	Establecer qué métodos, técnicas y procedimientos alternativos se utilizaron en cada propuesta.
P 5 - ¿Cuáles son las diferentes fuentes de datos utilizadas en cada uno de esos factores incluidos en las diferentes soluciones?	Identificar las fuentes de información utilizadas que puedan servir de insumo para otros trabajos que pretendan dar una solución a esta problemática

También se incluyeron palabras alternativas para la búsqueda tales como: *Peer; Algorithm; Methodology*

Posteriormente teniendo en cuenta los términos mencionados se realizó una búsqueda genérica en todos los repositorios seleccionados. Dada la diversidad de los resultados obtenidos con la aplicación de los distintos filtrados, se decidió efectuar diferentes cadenas de búsqueda adaptadas a cada repositorio, a los fines de obtener el mejor resultado posible (Tabla 2).

Tabla 2*Cadenas de búsqueda por cada repositorio y buscadores de artículos científicos*

Repositorios y buscadores de artículos científicos	Cadenas de búsqueda
Google Scholar	"paper reviewer" and "assignment problem" and (system OR algorithm OR methodology)
ACM Digital Library	("paper" OR "peer") AND "reviewer" AND "assignment problem" AND (system OR algorithm OR methodology)
IEEE Xplore	paper AND reviewer AND (assignment OR recommender) AND (problem OR system)
ScienceDirect	"paper reviewer" AND (assignment OR recommender) AND (problem OR system)

1.4 Criterios de inclusión y exclusión

A partir de los artículos obtenidos en el proceso de búsqueda en cada repositorio, fue necesario realizar una selección de los más cercanos a los objetivos planteados. Para ello, se definieron diferentes criterios de inclusión y exclusión con el fin de poder realizar esta tarea de manera sistemática que se muestran en la siguiente tabla 3. Luego, se fundamentan brevemente los criterios de inclusión consignados.

Breve fundamentación de los criterios de inclusión:

- ✓ Artículos en idioma inglés: un significativo porcentaje de la información en esta problemática del conocimiento circula en dicha lengua. Artículos en Acceso Abierto (AA) en otros lenguajes son muy pocos y no aportan resultados novedosos en relación a su aporte original al conocimiento.
- ✓ Artículos publicados en revistas científicas o en Congresos con revisión por pares: esta restricción es necesaria para asegurar la calidad académica de los textos. Sin revisión este requisito el contenido puede resultar de dudosa calidad científica.
- ✓ Artículos publicados en el período 2018-2023: es un área de conocimiento nueva y muy ligada a los avances a las técnicas de PLN y de IA, por lo cual el estado del arte siempre está ligado a estos avances y conocimientos de mayor

antigüedad son muy valiosos para tomarlos como referencia, pero pueden estar desactualizados. El corte en 2023 es la fecha de realización del estudio.

- ✓ Artículos científicos de texto completo que responden directa o indirectamente a las preguntas de investigación: la problemática abordada tiene una multiplicidad muy diversa de enfoques en los trabajos de investigación que tratan la temática de la asignación de revisores, por lo cual no todos se ajustan a las preguntas de investigación planteadas en el presente mapeo. En esta dirección se seleccionaron los que responden al criterio expuesto.

Tabla 3

Criterios de inclusión y de exclusión

Criterios de inclusión	Criterios de exclusión
• Artículos en idioma inglés. • Artículos publicados en revistas científicas o en Congresos con revisión por pares. • Artículos publicados en el período 2018 - 2023. • Artículos científicos de texto completo que responden directa o indirectamente a las preguntas de esta investigación.	• Estudios en idiomas diferentes del inglés, duplicados y aquellos a los que no se tiene acceso. • Artículos publicados en Acceso Abierto sin revisión por pares. • Artículos anteriores al año 2018. • No se consideran tesis de postgrado, libros, propuestas de trabajo a futuro u otro material que no aplique a la categoría a artículos científicos. • Artículos que si bien pueden ser cercanos a la temática no se relacionan con las preguntas de investigación • Asociados solamente al proceso de la restricción de asignación por carga de trabajo y/o cobertura. • Artículos que no tienen un desarrollo suficientemente adecuado referido a su metodología de extracción y asignación de revisores.

Cabe señalar que el criterio centrado en artículos con revisión de pares da por supuesto que trabajos de mayor extensión tales como tesis de postgrado y otros materiales acreditan por lo general su comunicación de resultados más significativos en artículos científicos evaluados, requerimiento solicitado por lo general, antes de la evaluación final de estas investigaciones de postgrado. En libros se encuentran en general compilaciones de trabajos que generalmente han sido publicados en formato artículo. Por otra parte, como el objeto de esta investigación se centra en la temática de recomendación de revisores para revistas científicas, en las mismas se pondera dicho formato.

Al utilizar las cadenas de búsqueda siguiendo el procedimiento del protocolo de revisión, búsqueda y criterios de selección, se accedió a 524 trabajos, a los que se les aplicaron los criterios de inclusión y exclusión expuestos en un procedimiento iterativo de dos etapas: 1ra. y 2da. del filtro. En la primera etapa se examinaron los títulos y los resúmenes, quedando 74 artículos seleccionados. Luego, al aplicar el 2do. filtro que consistía en la lectura completa del artículo se seleccionaron 25.

1.5 Síntesis de datos extraídos

La presente síntesis se estructura a partir de las preguntas de investigación mencionadas.

1.5.1 ¿Qué características generales definen agrupamientos iniciales del conjunto de artículos científicos seleccionados?

De los 25 artículos del período 2018-2023, en la figura 1 se puede observar la cantidad de trabajos registrados por año.



Figura 1.

Distribución de trabajos encontrados por año

También se analizó dónde fueron publicados, contando 11 en revistas científicas y 14 en eventos académicos.

En el siguiente análisis se ingresa específicamente a la problemática de recomendación de revisores con métodos automáticos. Este problema se puede dividir inicialmente entre la etapa de extracción con representación y la de asignación. En la etapa de extracción con

representación se toman las características que van a representar por un lado los documentos a ser revisados y por otro, los posibles revisores del artículo. La etapa de la asignación se refiere al proceso de selección de los revisores elegidos. Si bien los 25 artículos tratan la problemática, no todos tratan la asignación de igual manera. Se observan varias formas de abordar la extracción con representación que se explicarán más adelante, pero en cuanto a la asignación, en primera instancia es posible clasificarla en dos tipos diferentes: 1) sin restricción y 2) con restricción.

Cuando se refiere al tipo 1, sin restricción, el proceso de selección se va haciendo solo en base a los resultados de la extracción y puede pensarse en un proceso de única vez referido a cómo buscar los revisores para un artículo científico en particular, caso que se puede dar en una revista con referato. Luego se corre el mismo proceso para otro artículo y así sucesivamente.

Cuando se refiere a asignación con restricción (tipo 2), estas vienen dadas en relación a que ya se cuenta con un grupo de revisores y los mismos se tienen que asignar a una determinada cantidad de artículos científicos. Este proceso, a diferencia del caso anterior, es un proceso múltiple de varias asignaciones a la vez y es común en los eventos científicos. En general se toma registro de la carga de trabajo de los revisores y la cobertura como restricciones. La carga de trabajo de los revisores se refiere a cuantos artículos van a ser asignados en simultáneo para que realice la revisión. La cobertura expresa, en general, la cantidad de revisores que deben ser asignados a cada artículo. También se puede corresponder a la necesidad de cubrir con revisores adecuados todos los temas que trata un manuscrito. Si bien esto se presenta en términos generales, las metodologías en cada tipo pueden tener sus propias variaciones. Todos los artículos que tratan la extracción, también tratan la asignación que puede ser sin restricción (Dong et al., 2020) (T. Pradhan et al., 2021) (Jin et al., 2019) (Yong et al., 2021) (Kreutz & Schenkel, 2022) (Kalmukov, 2020b) (Medakene et al., 2019) (Mittal et al., 2020) (Patil & Mahalle, 2021a) (Latypova, 2023) (Kusumawardani & Khairunnisa, 2018) (Ishag et al., 2019) (Alkazemi, 2018) (Patil & Mahalle, 2021b) (Kalmukov, 2020a) (Al-Zboon et al., 2019) (Duan et al., 2019) y con restricción (Nguyen et al., 2018) (Abduljaleel & Al, 2021) (Kreutz et al., 2022) (D. K. Pradhan et al., 2020) (C. Yang et al., 2020) (Patil et al., 2023) (Xu et al., 2022) (Mirzaei et al., 2019).

Cabe mencionar que los artículos (Stelmakh et al., 2021) (Yeşilçimen & Yıldırım, 2019) (Lian et al., 2018) (M. Tan et al., 2021) (Kobren et al., 2019) (de Givry et al., 2020) (Guo et al., 2018) del período estudiado (2018-2023) solo tratan la asignación con restricción por lo cual no fueron incluidos. Los mismos solo se encargan de esta parte del problema y parten de una priorización que ya viene dada, tomándose como insumo necesario.

En la Figura 2 se puede observar cómo se agrupan los artículos según cómo tratan el problema en base a la clasificación definida anteriormente.

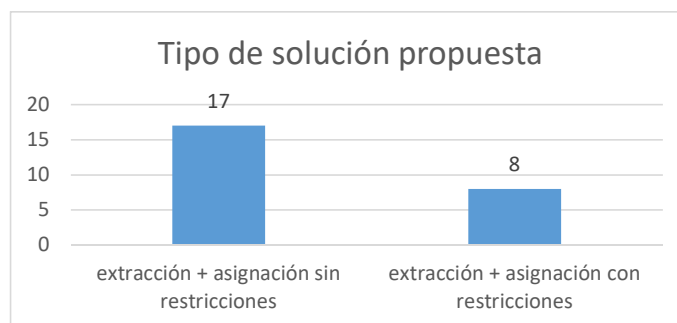


Figura 2

Distribución de artículos por tipo de solución propuesta

El análisis presentado en este capítulo se va a centrar en los procesos de extracción y asignación sin restricción, quedando exceptuado el análisis de asignación con restricción porque no responde al objeto de estudio necesario para metodología. La asignación con restricción refiere a las conferencias en donde hay una cierta cantidad específica de revisores a los que deben repartirse equitativamente todos los trabajos científicos presentados, en cambio en revistas científicas el proceso es individual por cada manuscrito presentado.

1.5.2 ¿Qué factores están incluidos en las propuestas de asignación automática de revisores?

Todas las propuestas metodológicas que abordan la extracción de información al basarse en datos clave y relevantes que influyen directamente en el proceso de selección de revisores, ofrecen una información de diferente tipo porque toman solo una parte del problema total. Esto permite establecer una clasificación, por lo cual esa información categorizada será denominada como factores. El tratamiento de cada factor es diferente y depende del tipo. Algunos de estos pueden poner restricciones a la asignación y otros que representan variables cualitativas se los somete a un proceso de codificación o transformación numérica según la interpretación de cada metodología. De esta manera, son incluidos en una fórmula que tiene impacto directo en diferentes proporciones en la selección de revisores. Esos factores intervinientes según el relevamiento realizado en el corpus seleccionado en el presente análisis, son ocho: experiencia, conflicto de intereses, autoridad, diversidad, interés del investigador, *seniority* y revisiones.

En las distintas propuestas se observa diversidad en las formas de nombrarlos y existen agrupamientos de algunos factores como partícipes del cálculo de otros. A pesar de esto, es posible efectuar una definición de cada uno de ellos en términos generales que permite reconocerlos en los diferentes artículos científicos, aunque sean tratados con otra

denominación de base. Así, se consideran aceptables todas las representaciones adecuadas no ambiguas con otros factores.

A continuación, se conceptualizan los ocho factores mencionados:

Experiencia: refiere principalmente a los textos de las producciones científicas de los candidatos a revisores, recabados de diferentes fuentes y de diferentes maneras por cada metodología. También se utiliza como temas de investigación declarados expresamente por los investigadores como en (Alkazemi, 2018). En la representación de los textos de las producciones científicas algunos artículos aclaran expresamente qué partes de los mismos son significativos para representar la experiencia. Por ejemplo, tomando todo el texto como en (Yong et al., 2021); palabras claves en (Ishag et al., 2019); los resúmenes en (Kalmukov, 2020a) y (Kalmukov, 2020b); títulos y resúmenes en (Kreutz et al., 2022) (Xu et al., 2022) (Dong et al., 2020) (Kreutz & Schenkel, 2022) (T. Pradhan et al., 2021) (Duan et al., 2019) y en las pruebas de (Mirzaei et al., 2019); títulos, resúmenes y palabras claves en (C. Yang et al., 2020) y (Latypova, 2023); títulos, resúmenes e introducción en (Patil & Mahalle, 2021a) y (Patil & Mahalle, 2021b); títulos, resúmenes, palabras claves, introducciones y conclusiones como en (Patil et al., 2023). Finalmente, solo los títulos como en (Al-Zboon et al., 2019) o dividiendo todos los artículos científicos en cinco partes: título, palabras claves, resúmenes, referencias y el resto del texto como en (Abduljaleel & Al, 2021)

Conflicto de intereses: representa las situaciones donde pueden existir intereses de distinta índole en el revisor que podrían provocar la pérdida de objetividad evaluativa. Tal puede ser el caso de ser autor del artículo a revisar, coautor en otro artículo científico de los autores del artículo a revisar, entre otros. En algunos casos se agregan controles de casos que verifican si tienen la misma afiliación como en (Patil & Mahalle, 2021a) que agrega al nombrar la afiliación, refiriéndose al mismo instituto, universidad o industria además de supervisor doctoral del autor del manuscrito. En (Latypova, 2023) también se controla la afiliación. En (Kreutz & Schenkel, 2022) se agrega que no sean coautores de otros revisores. Además se puede verificar que no sean coautores de coautores (D. K. Pradhan et al., 2020). En (Xu et al., 2022) agrega el concepto conflictos de competición que es el caso en que el revisor y el artículo a revisar comparten el mismo área de investigación en un evento científico y podría provocar la pérdida de objetividad. En (Al-Zboon et al., 2019) agrega relaciones familiares.

Autoridad: refiere al prestigio académico del investigador. En base a la mayoría de los artículos analizados se puede entender como el impacto académico de la producción del investigador reflejado en índices de las métricas como número de citas u otros factores de impacto bibliométrico, lo cual se puede tomar como un indicador objetivo propio de la comunidad académica. En (Patil et al., 2023) (T. Pradhan et al., 2021) y (Ishag et al., 2019)

utilizan número de citas e índice h, en (Kreutz & Schenkel, 2022) usan el promedio de índice h de los revisores calculado sobre los *papers* relevantes en relación al manuscrito y el promedio de las citas en estos artículos científicos. En (Patil & Mahalle, 2021b) se utiliza índice h. En (Latypova, 2023) se utilizan para autoridad, autoridad global sin ninguna relación con el tema del manuscrito y autoridad local sobre el tema del manuscrito a revisar, ambos se expresan por el número de referencias a los artículos del revisor. En (Nguyen et al., 2018) se usa el puntaje de calidad, en el que se integran el número de doctorados supervisados, libros y capítulos de libros escritos, artículos publicados en revistas y conferencias además del índice h. En (Yong et al., 2021) se utiliza como titulación o nivel de estudio del candidato a revisor.

Diversidad: este factor es aplicado en los artículos de diferentes maneras, pero refiere a la búsqueda de una evaluación justa, integral y equitativa tratando de incorporar revisores basados en lo que conforma la diversidad para cada uno de los autores de los trabajos analizados. En (Kreutz & Schenkel, 2022) define la diversidad como la medida que asegura que la experiencia de los revisores sea distribuida a áreas lo más diferentes posible. En (Kreutz et al., 2022) engloba *background*, ubicación y *seniority* (que será definido más adelante) . En *background* buscan diversidad en que al menos un revisor debe trabajar en el mundo académico y uno en la industria (pudiendo ser el mismo), en ubicación geográfica se recomienda como deseable incluir revisores de ubicaciones geográficas diferentes. En (Ishag et al., 2019) la diversidad la toman en ubicación geográfica de sus afiliaciones, referidas a países. En (T. Pradhan et al., 2021) se utiliza caminata aleatoria definida en inglés como *Random walk with restart* (RWR) con reinicio para elegir revisores, lo que da como posibilidad la obtención de revisores diversos.

Interés del investigador: denominado en algunos artículos como *freshness* o *recency*, está asociado a la elección de artículos científicos dentro de un determinado rango de tiempo reciente u otorgar más peso sobre otros artículos a los que sean más cercanos en un rango de tiempo, o por escalas de pesos según rangos de tiempos cercanos. El sentido de elegir esa denominación se basa en que a través del tiempo puede ir cambiando el tema de estudio del investigador, lo que lleva a pensar que sus producciones más cercanas refieren a su interés más cercano de investigación. Algunos artículos expresan el interés de investigación como factor de experiencia, pudiendo observarse en (Jin et al., 2019) y (Yong et al., 2021) pero que en realidad tratan sobre los temas de interés de investigación. En (Nguyen et al., 2018) se refieren al vector de interés de tema para representar el área de experiencia del revisor pero utilizan el factor de impacto del investigador como puntaje de actualidad refiriéndose a los artículos publicados en los últimos años del revisor y definiéndolo como promedio ponderado de los artículos publicados por dicho académico en un rango que abarca los últimos años. En (T. Pradhan et al.,

2021) toman los artículos científicos más cercanos en el tiempo con mayor peso por sobre los otros, para armar la red de revisores, lo mismo que en (Kreutz & Schenkel, 2022). En (Patil et al., 2023) consideran que el candidato a revisor debe estar activo en áreas de investigación del manuscrito y se hace un cálculo considerando la diferencia del año actual con el año de publicación en la fórmula. En (Patil & Mahalle, 2021) toman en cuenta la actualidad de las publicaciones de los revisores.

Seniority: refiere a la categoría académica alcanzada por el investigador. En (Kreutz et al., 2022) se usa este término dentro de diversidad como ya se explicó anteriormente y se exige que cada conjunto de revisores contenga al menos un investigador categoría *senior*. En (Kreutz & Schenkel, 2022) expresan como deseable que la revisión de cada manuscrito la realicen al menos un investigador *senior* y un *junior*.

Revisiones: este factor es poco tenido en cuenta y toma como referencia el historial de revisiones anteriores de los candidatos como en (Latypova, 2023).

En síntesis, la utilización de los factores en el corpus de artículos seleccionados se puede observar en la Figura 3. El más utilizado es Experiencia y en orden decreciente Conflicto de intereses, Autoridad, Diversidad, Interés del investigador, *Seniority* y por último como ya se adelantó, Revisiones.

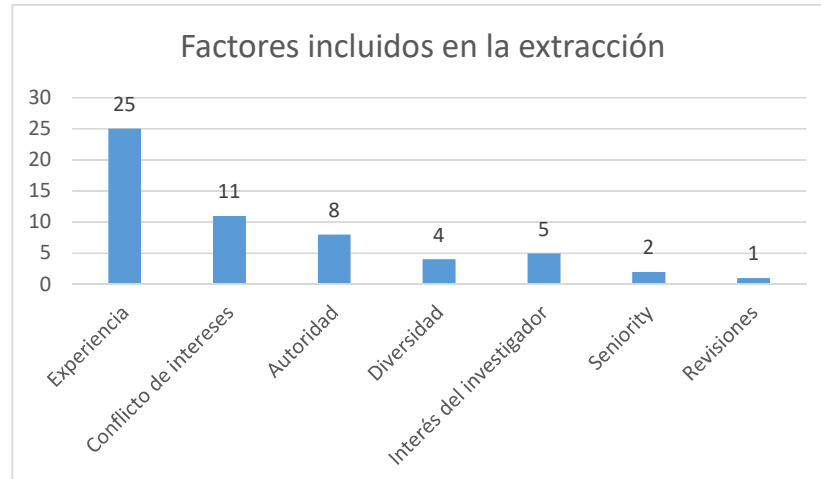


Figura 3

Factores incluidos en la extracción

En la siguiente Tabla 4 se puede observar cómo se agrupan los factores: once solo toman Experiencia y cabe señalar que todos los artículos que tratan un solo factor toman el mencionado. En el caso de los artículos que tratan dos factores, cinco toman Experiencia y Conflicto de intereses y uno toma Experiencia y Autoridad. Desde los artículos que toman tres

factores hasta el máximo de seis, existe una diversidad en los factores a incluir, lo único que se puede inferir es que el factor común a tomar siempre es Experiencia, variando todos los otros.

Tabla 4

Agrupamiento de factores

Cantidad de factores	Artículos científicos	Tipo de factores
1	11	solo experiencia
2	5	experiencia y conflicto de intereses
2	1	Experiencia y autoridad
3	2	diferentes factores, el único común es experiencia
4	4	
5	1	
6	1	

1.5.3 ¿Qué modelos y técnicas de PLN son utilizadas en las propuestas?

Siendo la recomendación el resultado de un proceso con la mayor automatización posible, una vez que se definen los factores de los que se extrae información se desea saber si se utilizan técnicas de PLN en el procesamiento y cuáles son las utilizadas en las diferentes metodologías.

La técnica más utilizada en los artículos es Latent Dirichlet Allocation (LDA) (Blei, s. f.). La misma consiste en un modelo probabilístico generativo para colecciones de datos discretos que se emplea para descubrir temas latentes en un conjunto de documentos considerando que estos tratan sobre múltiples temas. En este sentido, buscan descubrir las palabras que reflejan la existencia de diferentes temas en cada documento del conjunto analizado. El modelo fue utilizado para procesar el factor de experiencia en (Nguyen et al., 2018) (D. K. Pradhan et al., 2020) (C. Yang et al., 2020) (Patil et al., 2023) (Medakene et al., 2019) (Patil & Mahalle, 2021a) (Patil & Mahalle, 2021b) (Al-Zboon et al., 2019). Luego le sigue la técnica Term Frequency-Inverse Document Frequency (TF-IDF) que es una técnica de modelos de espacio vectorial que por un lado miden la frecuencia en que una palabra aparece en un documento y por otro, mide la frecuencia inversa que aparece en todos los otros documentos. De esta manera evalúa la importancia de una palabra en un documento en relación con todos los otros y fue utilizada en (Abduljaleel & Al, 2021) (Kreutz et al., 2022) (Latypova, 2023). En (Kalmukov, 2020b) se utiliza TF-IDF o BM-25 (Robertson et al., 1994) que es otra técnica de modelo de espacio vectorial utilizada para la recuperación de información y en relación a TF-IDF agrega mayor sensibilidad a la frecuencia de términos y a la longitud del documento. TF-IDF y BM-25 son modelos de espacio

vectorial, y más específicamente modelos de vectores escasos que representan los documentos como bolsas de palabras ignorando el orden y la sintaxis de las palabras y que no tienen información semántica incluida en las representaciones. En (Kalmukov, 2020a) usan BM-25. En (Duan et al., 2019) usan TF-IDF y un algoritmo de red neuronal como Bidirectional encoder representations from Transformers (BERT) (Devlin et al., 2019) o Convolutional Neural Network (CNN) (Krizhevsky et al., 2012) o Bidirectional Long Short Term Memory (BiLSTM) (Sherstinsky, 2020). BERT es un modelo de lenguaje basado en Transformers desarrollado por Google con un enfoque bidireccional por la forma que fue entrenado que a diferencia de los anteriores contiene representación semántica contextualizada en la conformación de los vectores densos que representan las palabras. Las CNN son un tipo de redes neuronales que por su arquitectura fueron diseñadas principalmente para manejo de imágenes, pero pueden ser utilizadas para otro tipo de problemas vinculados al PLN. Las BiLSTM son un tipo particular de redes neuronales recurrentes (RNN) que agregan memoria interna y que fueron diseñadas para trabajar con datos secuenciales y también fueron utilizadas en el PLN. El tipo particular de RNN utilizada es la Long Short Term Memory (LSTM) que son capaces de recordar información relevante y, además fue diseñada con bidireccionalidad para procesar los datos en ambos sentidos a la vez. Cabe aclarar que CNN como otras técnicas se pueden usar para problemas de PLN, aunque no es lo común, siendo lo más utilizado RNN. Sin embargo, se agrega en este análisis por ser una técnica más aplicada al mismo nivel que otras más específicas de PLN.

En T. Pradhan et al. (2021) usan LDA y una variación aplicada a la recomendación de revisores de TF-IDF que denominan Term Frequency-Inverse Reviewer Frequency (TF-IRF). En (Kusumawardani & Khairunnisa, 2018) usan Author-Topic Modeling (ATM) que extiende el modelo LDA para incorporar información sobre la autoría de los documentos y así poder capturar la relación entre temas y autores (Rosen-Zvi et al., 2004). En (Jin et al., 2019) usan Author-Subject-Topic Modeling (AST) (Mou et al., 2015), este modelo agrega una capa de asunto adicional al ATM. Esa capa supervisada adicional permite compartir temas entre autores y agrupar documentos bajo varias distribuciones de temas, lo que propone una mejoría con respecto a solo agrupar documentos sobre la distribución de temas de un solo autor. En (Yong et al., 2021) usan Word2Vec (Mikolov et al., 2013) para extraer las palabras claves semánticas principales. Word2Vec es un modelo desarrollado por Google para generar *word embeddings* que son representaciones vectoriales de palabras que contienen información semántica. Esto se realiza sobre la base del entrenamiento con grandes cantidades de texto no etiquetado para aprender esas representaciones utilizando ventanas de texto deslizables. En (Kreutz & Schenkel, 2022) utilizan en los artículos científicos seleccionados: TF-IDF, Doc2Vec (Le & Mikolov, 2014) o BERT y para los cálculos usan TF-IDF y LDA. Doc2vec es un desprendimiento de Word2Vec que al igual que este, a través de un algoritmo no supervisado captura representaciones vectoriales

densas de documentos completos en lugar de palabras individuales. En (Dong et al., 2020) plantean una red neuronal del tipo unidad recurrente cerrada (GRU) (Cho et al., 2014) bidireccional de dos niveles con un mecanismo de atención. Las GRU son una mejora sobre las redes neuronales recurrentes (RNN) tradicionales para abordar problemas de desvanecimiento de gradientes y mejorar la eficiencia computacional. En (Mirzaei et al., 2019) usan Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 2013) que es una extensión de Latent Semantic Analysis(LSA) que define un proceso generativo probabilístico por el cual, cada documento está asociado a una distribución de probabilidad sobre los temas y cada tema está asociado a una distribución de probabilidad sobre las palabras, capturando la probabilidad de que un documento trate ciertos temas y de que una palabra aparezca en ciertos temas. En (Xu et al., 2022) aplican la técnica de Prompt tuning (Yao et al., 2022) que es un método simple pero eficaz de comunicarse con modelos largos de lenguaje (LLM) para que realicen tareas específicas, ajustando la redacción de las instrucciones para obtener los resultados deseados, en contraposición al preentrenamiento o “fine tuning”. En (Mittal et al., 2020) se expone el modelo de temas sin aclarar cuál es la técnica. Cabe mencionar que dos de los artículos no tienen técnicas de PLN. En (Ishag et al., 2019) se utilizan las palabras claves de los artículos científicos de los revisores y de los manuscritos sin procesamiento. En (Alkazemi, 2018) se declaran las palabras claves de las propuestas como interés de investigación del revisor y las declaradas en los manuscritos de las disponibles en una taxonomía. En la siguiente Tabla 5 se detalla el agrupamiento de técnicas y modelos utilizados.

1.5.4 ¿Cuáles son las metodologías utilizadas para las diferentes propuestas?

En este apartado se registran ocho artículos cuya metodología para el cálculo de la priorización para la asignación (sin tener en cuenta las restricciones) se basa en la integración de múltiples factores en una fórmula. Si se establece una relación entre los artículos científicos involucrados con los diferentes factores como muestra la Tabla 6, de los 7 factores discriminados, a partir de la cantidad de tres seleccionados se aplican este tipo de soluciones.

Tabla 5

Agrupamiento de técnicas y modelos de PLN

Agrupamiento	Cant.	Técnica y/o modelos de PLN utilizado	Artículos
Modelos de temas	12	LDA, ATM, AST, PLSA	(Nguyen et al., 2018) (D. K. Pradhan et al., 2020) (C. Yang et al., 2020) (Patil et al., 2023) (Mirzaei et al., 2019) (Jin et al., 2019) (Medakene et al., 2019) (Mittal et al., 2020) ⁴ (Patil & Mahalle, 2021a) (Kusumawardani & Khairunnisa, 2018) (Patil & Mahalle, 2021b) (Al-Zboon et al., 2019)
Modelos de vectores escasos	5	TF-IDF, BM-25	(Abduljaleel & Al, 2021) (Kreutz et al., 2022) (Kalmukov, 2020b) (Latypova, 2023) (Kalmukov, 2020a)
Combinación de modelos y técnicas diferentes de PLN	1	LDA y TF-IRF	(T. Pradhan et al., 2021)
	1	(TF-IDF o Doc2Vec o BERT) y (TF-IDF y LDA)	(Kreutz & Schenkel, 2022)
	1	TF-IDF y (BERT o CNN ⁵ o BiLSTM)	(Duan et al., 2019)
Modelos basados en Word embeddings	1	Word2Vec	(Yong et al., 2021)
Modelos con redes neuronales	1	BiGRU	(Dong et al., 2020)

⁴ No se aclara la técnica de modelo de temas utilizado.

⁵ Como ya se aclaró en el texto, se incluye por ser opción entre otras técnicas más específicas de PLN.

Modelos basados en LLM	1	Prompt tuning	(Xu et al., 2022)
No hay modelo	2		(Ishag et al., 2019) (Alkazemi, 2018)

Tabla 6

Recorte de factores incluidos en artículos científicos, desde 3 hacia arriba

Art.	Experiencia	Conflicto de intereses	Autoridad	Diversidad	Interés del investigador	Seniority	Revisiones	Cant. de factores
(Ishag et al., 2019)	Sí	No	Sí	Sí	No	No	No	3
(Patil & Mahalle, 2021b)	Sí	No	Sí	No	Sí	No	No	3
(Nguyen et al., 2018)	Sí	Sí	Sí	No	Sí	No	No	4
(Kreutz et al., 2022)	Sí	Sí	No	Sí	No	Sí	No	4
(Patil et al., 2023)	Sí	Sí	Sí	No	Sí	No	No	4
(Latypova, 2023)	Sí	Sí	Sí	No	No	No	Sí	4
(T. Pradhan et al., 2021)	Sí	Sí	Sí	Sí	Sí	No	No	5
(Kreutz & Schenkel, 2022)	Sí	Sí	Sí	Sí	Sí	Sí	No	6

En Ishag et al. (2019) la metodología propuesta extrae las palabras claves del manuscrito y de los artículos científicos de los revisores. En este caso toma varios conceptos como que el prestigio de un *paper* (Pr) en el que intervienen el factor de diversidad en relación a la afiliación de sus autores y la cantidad de citas del artículo científico; en el impacto de un autor (AI) intervienen el índice *h* del autor como factor de autoridad y el prestigio de una publicación, y en el impacto de palabra clave (KI) intervienen la frecuencia de las palabras claves y el prestigio de

una publicación. Finalmente, en el impacto de una transacción (I) intervienen el prestigio del artículo científico con los impactos de sus autores y palabras claves. Sobre esta base, se formulan dos tipos de patrones de objetivos. Por un lado, los patrones de tema general del investigador (RGP) que incluye solo a los investigadores y por el otro, los patrones de temas específicos del investigador (RSP) que está compuesto de combinaciones de investigadores y palabras claves que representan su área de experiencia.

En Patil & Mahalle (2021b), se puede observar que de los artículos científicos se toman título, resumen e introducción procesados con LDA para extraer el tema más importante de cada documento y así encontrar los expertos más relevantes de cada tema. Luego, se ordenan descendientemente en base a la cantidad de publicaciones en ese dominio. En el cálculo de la metodología se incorpora el factor de interés del investigador, tomando en cuenta este factor asociado a la actualidad de sus publicaciones y la autoridad en base al índice h a los fines de considerar el experto más relevante en el dominio del manuscrito.

En Nguyen et al., (2018) con los artículos presentados se crea el listado de temas con LDA, de la conferencia y todos los posibles temas de investigación de cada revisor son extraídos de diferentes fuentes e ingresados en un diccionario común de términos. El diccionario alinea términos con los temas de la conferencia. Se usa una asignación automática, pero el resultado es examinado por un experto para asegurar la traducción adecuada. Luego, en cada artículo se calcula la relación de temas que tiene con los de un evento académico y como resultado se obtiene la proporción de temas de dicho evento cubiertos por cada artículo. Así, asignar revisores es el proceso de determinación de coincidencias para identificar el conjunto de revisores que satisfacen las necesidades del manuscrito. En este caso, se utiliza una metodología con cuatro indicadores. El primero es la determinación de coincidencia entre los temas cubiertos por el manuscrito y la experiencia del revisor que lo denomina su interés en el tema de investigación y se agregan el factor de autoridad que lo denominan calidad, el factor de interés y la disponibilidad que está asociado a la asignación con restricción. La calidad, como ya se definió, tiene en cuenta el número de doctorados supervisados, libros y capítulos de libros escritos, artículos publicados en congresos y revistas y su índice h. El factor de interés lo define como el factor de impacto promedio ponderado. Esta metodología tiene en cuenta también el conflicto de intereses.

En Kreutz et al. (2022) la finalidad de DiveRS es extender el comité del programa de un evento académico con nuevos revisores no incluidos utilizando un proceso automático y a modo de sugerencia. El proceso genera la asignación teniendo en cuenta dos tareas interconectadas, la primera identifica envíos que no recibirían revisores adecuados utilizando el comité del programa actual y a su vez, genera como segunda tarea, la sugerencia de nuevos revisores para

ampliarlo. Las personas recién incluidas deberían ser capaces de evaluar múltiples manuscritos, y garantizar la diversidad de todo el comité del programa. En esencia, DiveRS se basa en un método de asignación de revisores, que considera a los revisores como un conjunto para cada artículo, con el fin de satisfacer las limitaciones de diversidad. Para la representación de manuscritos y trabajos de revisores usa TF-IDF y para calcular la similitud aplica la técnica de similitud de coseno. Así, establece una similitud mínima entre un manuscrito y sus revisores y para la asignación toma en cuenta la diversidad teniendo considerando el *background*, ubicación y *seniority* cuantificando todos los puntos incluidos e integrándolos en una fórmula. También verifica conflicto de intereses.

En Patil et al. (2023) se toma los títulos, resúmenes, palabras claves y conclusiones de los manuscritos y los artículos de los revisores y los procesa con LDA para encontrar todos los dominios temáticos y para asignar un conjunto de revisores que puedan cubrir todos los temas que trata el manuscrito. Por cada publicación de revisor genera un único documento que colectivamente representa todas las publicaciones del revisor y selecciona los temas más relevantes de cada revisor ordenándolos descendientemente hasta una cierta cantidad a ser considerados en base a un parámetro. Luego, se realiza un cálculo de similitud entre el manuscrito y los revisores para encontrar los más adecuados. De esta manera se obtiene la experiencia de los revisores en cada uno de los temas del manuscrito. La asignación final la realiza un algoritmo a través de una suma ponderada de cada uno de los valores de experiencia, autoridad e interés del investigador. Para la autoridad se aplica una fórmula que integra número de publicaciones, índice h y número de citaciones. En la fórmula de cálculo del interés del investigador utiliza como parámetro un determinado lapso de años a tener en cuenta artículos científicos desde el año actual y asociados a la temática. También se controla el conflicto de intereses. A esta metodología se la conoce en idioma castellano como “Sistema de asignación de revisores a artículos científicos proactivo no supervisado”.

Latypova (2023), selecciona los mejores revisores tomando el título, el resumen y las palabras claves del manuscrito y los revisores, procesándolos con TF-IDF y realizando el cálculo de similitud de coseno, y los ordena de forma descendente hasta un cierto número dado como parámetro. Luego realiza la asignación haciendo una evaluación multicriterio expresada por un indicador realizando una suma ponderada de diferentes parámetros relacionados a sus valores máximos. En este caso, se toma la experiencia relacionada al valor máximo de la misma. La autoridad se evalúa tomando el indicador de Autoridad Global (GA) y el indicador de Autoridad Local (AL). Para calcular GA, toma el número de referencias del revisor sobre el de todos los revisores y en el cálculo de AL, toma el número de referencias a los artículos de los revisores sobre el tema del manuscrito. El factor de revisión lo realiza a través de la definición de calidad

de trabajo en el rol de revisor y lo hace con tres indicadores. RE toma la cantidad de años que es revisor y RA toma todos los días de revisión sobre la cantidad de revisiones. AQ representa el número de manuscritos donde el equipo editorial estuvo de acuerdo con la revisión sobre el total de revisiones solicitadas. Entonces, cada uno de los indicadores tiene un peso y luego de realizar el cálculo, los revisores se ordenan descendentemente y se asignan desde el mayor valor hasta un parámetro de cantidad de revisores a asignar definido como parámetro. En esta metodología también se controla el conflicto de intereses.

En Pradhan et al. (2021), se propone un modelo denominado TCRRec con una red multicapa que integra tres capas esenciales que se describen a continuación:

Capa de red de temas: se identifican publicaciones de los investigadores y utilizan el título, el resumen para extraer temas con LDA. Con la finalidad de darle mayor importancia a los artículos más cercanos, se le asigna mayor peso en la fórmula agrupándolos por cada revisor en relación al año de publicación, con un número máximo de años hacia atrás desde el actual. También para capturar el contenido lexicográfico se toma el título y el resumen de todas las publicaciones de cada revisor y se procesa aplicando TF-IRF (que es una variación de TF-IDF), lo cual cambia documento por revisor en la sigla. Esto último se debe a que todas las producciones de cada revisor se fusionan en un solo documento. El proceso unifica las dos fórmulas anteriores en una sola fórmula, sumando las respectivas similitudes de coseno por separado con normalización ponderada con pesos complementarios. Para la construcción de la red de temas y los pesos entre estos, se realiza un cálculo de los grupos relacionados. La detección de comunidad de la red de temas se hace por los pesos y así se forma la red de temas individuales. La relación entre la red de temas y la red de citas se efectiviza por los temas de interés que son aquellos que tienen mayor peso en el manuscrito. Los artículos científicos que tienen mayores pesos en los temas de interés son los elegidos como candidatos.

Capa de red de citas: se estiman las de coautoría ya que se infiere que artículos similares en citas tratan temas similares. Esto colabora en considerar artículos que no están relacionados por los temas sino por citaciones similares, ya que probablemente traten temas afines, lo cual se entiende como un enlace indirecto. La red de citas se asocia a la red de revisores porque el conjunto de revisores asociados son los autores de los artículos candidatos.

Capa de red de revisores: agrega otros atributos para hacer la recomendación final. En el tratamiento del tema Autoridad, identifican el candidato a revisor empleando características como índice h, citación, además de puntaje de co-citación, y algunas otras características como similitud de co-autor. Para la asignación se realiza un acercamiento multicapa integrando red de temas, red de citaciones y red de revisores, utilizando la técnica de caminata aleatoria con reinicio (RWR) en la red de revisores final, lo que es considerado como factor de Diversidad. En

esta capa se aplica una versión personalizada del algoritmo RWR. El mecanismo de puntaje elimina los artículos menos relevantes y reduce la complejidad computacional del sistema. En la metodología se incluye el control de conflicto de intereses. Las redes de revisores tienen en cuenta otros atributos que asocian revisores. Algunos de ellos tienen similitud con el manuscrito, pero diversidad entre los mismos. El uso de caminata aleatoria se implementa para reducir el cálculo porque evita calcular la similitud del documento y todos los revisores. En esta red cada borde entre dos revisores existe si se cumple una de las siguientes condiciones: 1) si hay una similitud temática mayor a un umbral definido o 2) son coautores en al menos un artículo científico.

En Kreutz & Schenkel (2022), se proponen dos etapas, la primera se divide en tres partes. En la primera parte se hace la búsqueda de expertos, usando TF-IDF, BERT o Word2Vec; en la segunda se hace el cálculo de similitud de coseno entre el manuscrito y los papers de los revisores y se ordenan los papers por puntaje de forma descendente. Aquí se realiza un primer filtro de conflicto de intereses por coautor de autor. En la tercera parte, para obtener la lista final de revisores realiza un cálculo tomando técnicas de votos (se consideraron 13 técnicas), teniendo en cuenta un umbral de consideración y usando técnicas de consideración de puntajes totales.

Cabe mencionar que los autores citados, en la segunda etapa, usan LDA y TF-IDF para representar los *papers*, que los une y efectúa un cálculo por similitud de coseno. Luego, ordena los revisores en forma descendente y toma a los primeros más similares según parámetro. Sobre lo anterior, efectúa el cálculo de experiencia, autoridad, diversidad, interés de investigación y *seniority*. En experiencia considera que los revisores deben tener conocimientos sólidos de los términos y temas del manuscrito respaldados por numerosas publicaciones. Específicamente, la presentación debe ser similar a las publicaciones escritas por los revisores y el número de artículos debe ser alto. Para autoridad, los revisores deben tener reconocimiento en el área a la que pertenece el manuscrito y proponen dos puntuaciones para medirla: el índice h promedio de los revisores calculado sobre artículos relevantes para el manuscrito y el número promedio de citas obtenidas en estos artículos. A estos índices los suma con una normalización ponderada complementaria a 1 (uno). A la Diversidad la toma como una medida para garantizar que la experiencia de los revisores se distribuya en áreas lo más separadas posible. Esto permite que las revisiones cubran múltiples aspectos del manuscrito. En esta dirección, toma el interés de investigación midiendo el ajuste de los revisores y el manuscrito con respecto a su desarrollo temporal priorizando los más cercanos. También, combinan las publicaciones de los revisores con respecto a su antigüedad en un vector de longitud normalizada donde se ponderan con mayor fuerza los artículos recientes con relación a los más alejados en el tiempo. Para *seniority*

la metodología considera como deseables a los conjuntos de revisores que no estén formados únicamente por investigadores experimentados. Debe estar incluido al menos un investigador senior que esté familiarizado con la metodología del artículo y es deseable tener un grupo diverso en términos de antigüedad, y aclara que se debería incluir al menos un investigador junior. Al final, elabora una recomendación teniendo en cuenta todos los valores resultados de los cálculos, multiplicándolos. En las pruebas se realizó lo siguiente: para cada uno de los artículos publicados por posibles revisores, los conjuntos de datos tuvieron que contener representaciones vectoriales tf-idf, Doc2Vec, LDA y BERT de su título y resumen, cuando estén disponibles. Para los manuscritos enviados, estos cuatro tipos de representación de documentos están contenidos para el texto completo, así como solo para las secciones de investigación del artículo (que constan de todas las secciones excluyendo el resumen, la introducción, el trabajo relacionado, la conclusión, las referencias y los agradecimientos). En síntesis, son cuatro las metodologías aplicadas al texto completo y al recorte del texto. Cabe mencionar que se consideraron únicamente documentos en inglés para la construcción de los conjuntos de datos.

En la Tabla 7 se puede observar que los métodos restantes tienen solo el factor Experiencia y otros, agregan factor Conflicto de intereses. No hay un agrupamiento único en estos casos y hay pocas coincidencias metodológicas que permitan agrupamientos por lo cual se tratan como técnicas separadas.

Dong et al. (2020) utilizan una red de dos capas GRU bidireccionales con un mecanismo de atención en dos niveles. Una trabaja sobre los tokens codificando oraciones y la otra lo hace sobre las oraciones codificando el documento. Así, toma los títulos y resúmenes y los procesa. La recomendación es la selección del revisor que tenga más etiquetas de campo de investigación que posee el artículo a revisar usando una asignación simple de revisor basada en múltiples etiquetas (MLBRA)

En Jin et al. (2019) aplican el modelo AST que es una extensión del AT que introduce una capa asunto cuya tarea es supervisar la generación de temas jerárquicos y permitir compartirla entre autores. Se proponen dos versiones: AST1 y AST2. La primera asigna a cada documento una distribución de tema en la capa de asunto (agrupación suave) y la segunda a cada documento asigna una etiqueta de asunto de acuerdo con la distribución de categorías (agrupación rígida). En las pruebas realizadas AST1 obtuvo mejor rendimiento que AST2.

En Kalmukov (2020b) se utilizaron las técnicas de modelo espacio vectorial TF-IDF, con variaciones en la ponderación de términos de los documentos y BM-25 para representar los documentos de los cuales se tomó el resumen. Se utilizó similitud de coseno para efectuar la correspondencia entre manuscrito y revisores y los mejores resultados se consiguieron con BM-25.

Yang et al., (2020) usan LDA para representar los manuscritos y los artículos científicos de los revisores. La parte del texto que utiliza son el título, el resumen y las palabras claves. A partir de esto, asigna los revisores de acuerdo a la similitud temática con el manuscrito.

Tabla 7

Recorte de factores incluidos en artículos científicos (hasta 2).

Art.	Experiencia	Conflicto de intereses	Autoridad	Diversidad	Interés del investigador	Seniority	Revisiones	Cant. de factores
(Dong et al., 2020)	Sí	No	No	No	No	No	No	1
(Jin et al., 2019)	Sí	No	No	No	No	No	No	1
(Kalmukov, 2020b)	Sí	No	No	No	No	No	No	1
(C. Yang et al., 2020)	sí	No	No	No	No	No	No	1
(Duan et al., 2019)	sí	No	No	No	No	No	No	1
(Medakene et al., 2019)	sí	No	No	No	No	No	No	1
(Mittal et al., 2020)	sí	No	No	No	No	No	No	1
(Mirzaei et al., 2019)	Sí	No	No	No	No	No	No	1
(Kusumawardani & Khairunnisa, 2018)	Sí	No	No	No	No	No	No	1
(Alkazemi, 2018)	Sí	No	No	No	No	No	No	1
(Kalmukov, 2020a)	Sí	No	No	No	No	No	No	1
(Abduljaleel & Al, 2021)	Sí	Sí	No	No	No	No	No	2
(D. K. Pradhan et al., 2020)	Sí	Sí	No	No	No	No	No	2
(Patil & Mahalle, 2021a)	Sí	Sí	No	No	No	No	No	2
(Xu et al., 2022)	Sí	Sí	No	No	No	No	No	2
(Al-Zboon et al., 2019)	Sí	Sí	No	No	No	No	No	2

En Duan et al., (2019) se propone un modelo predictivo de clasificación con casos positivos (título y resumen del mismo artículos) y casos negativos generados sobre títulos y resúmenes de diferentes artículos. El entrenamiento se realiza de esa manera para disminuir la

función de pérdida, y luego se utiliza para predecir posibles revisores. Esta metodología propone una forma de asignación de revisor basado en modelos de pares de oraciones (SPM-RA), modelando dichas relaciones de pares mediante la supervisión de información. La información de supervisión hace que el modelo aprenda con precisión las características de campo de los revisores y manuscritos. En primer lugar, construye la formación establecida por la relación de campo entre título y resumen. Así, obtiene un conjunto de datos con una muestra TF-IDF y usa diferentes arquitecturas de redes neuronales (como BERT, CNN o BiLSTM) a los fines de construir el modelo de pares de oraciones para entrenar la relación de campo entre el título y el resumen del artículo (modelo supervisado) y, luego poder predecir la similitud entre revisores y manuscritos entrenando el modelo.

En Medakene et al. (2019) la metodología está formada por dos componentes. Por un lado, se toman los manuscritos y los artículos científicos de los revisores procesándolos con LDA. Para calcular la similitud le resta a la unidad la distancia de Jensen-Shanon entre revisores y manuscritos, que es la raíz cuadrada de la divergencia de esta unidad de distancia. Por el otro lado, realiza el cálculo de la similitud de referencias armando los conjuntos y comparando por cada revisor contra las referencias del artículo científico. Se utiliza la medida de índice de Jaccard para calibrar la similitud y diversidad del conjunto de muestras. Luego, se unifican las dos fórmulas anteriores, sumando los valores de similitud de experiencia y de referencias con normalización ponderada con pesos complementarios. En la metodología los autores incluyen el control de conflicto de intereses.

En Mittal et al. (2020) sobre las palabras más importantes extraídas de los *papers* de los revisores y manuscritos, se construye un gráfico difuso con dichas palabras claves y se aplican medidas de centralidad en el gráfico difuso, creando así conjuntos difusos para las claves seleccionadas y sus pesos. Con Wordnet se calcula la distancia entre los conjuntos difusos representados por los *papers* de los autores y el de los candidatos a revisores. Finalmente, el principio de extensión difuso es aplicado sobre los conjuntos para seleccionar los tres principales expertos para revisar la propuesta. Es decir, de los candidatos a revisores que resultan con la menor distancia al artículo a revisar (haciendo un recorte de un artículo solo), son seleccionados solo tres según esta metodología.

Mirzaei et al. (2019) aplican un método que agrupa las publicaciones de los revisores en áreas de investigación latente. Para utilizar la información latente de este tipo de área, es necesario abordar dos subproblemas: 1) identificar las áreas de investigación latentes en las publicaciones de los revisores y, 2) mejorar la coincidencia entre revisores y artículos. Con el fin de obtener las áreas de investigación latentes, se agrupan todas las publicaciones de los revisores utilizando k-medias con distancia coseno. También se utiliza la indexación semántica

latente para reducir la dimensionalidad de estos documentos en el espacio temporal y así mejorar la calidad de la agrupación. Cada revisor está asociado a una serie de áreas de investigación latentes a las que pertenecen sus publicaciones. Para extraer temas de la unión de todos los perfiles de revisores se utilizó el método de modelado de temas PLSA. El objetivo final es asignar cada artículo a un grupo de revisores con experiencia diversa, quienes conjuntamente, cubran al máximo los aspectos del artículo.

En Kusumawardani & Khairunnisa (2018) se tomaron los textos seleccionando desde el resumen hasta las conclusiones, representados con el modelo autor-tema que es una extensión de LDA. Este modelo agrega una capa con autores sobre la distribución de temas. Para seleccionar posibles revisores los autores utilizan el cálculo de la distancia de Hellinger en función de la similitud de sus vectores de probabilidad de tema. En este trabajo se sostiene que esta métrica es adecuada y tendría una buena *performance* de rendimiento para enfoques probabilísticos.

Alkazemi (2018) organiza la propuesta de artículos para su evaluación y los revisores en pistas categorizadas según lo definido por un árbol de dominios de investigación jerárquicos. De esta manera, los revisores crean sus perfiles de intereses de investigación en el registro y tomando esto como base, se calcula la similitud mediante el coeficiente de similitud de Jaccard entre las palabras clave del revisor y las de la propuesta.

Kalmukov (2020a) propone una metodología que se divide en tres partes. La primera consiste en la búsqueda y extracción de publicaciones anteriores de los revisores en Internet obteniendo sus títulos y resúmenes. La segunda refiere al cálculo de factores de similitud a través de técnicas de PLN. En la práctica solo se probaron modelos de espacio vectorial y BM-25 arrojó los mejores resultados probando sobre los resúmenes. La tercera realiza una evaluación de la precisión de los factores de similitud. La misma está conformada por conversión de indicadores de similitudes a niveles considerando nivel de experiencia (Alto, Medio y bajo) determinados automáticamente. Además, incluye un armado de análisis de correlación de Pearson entre los niveles de determinación automática y los explicitados por los revisores. Los resultados presentan informes y estadísticas detalladas y comparativas, resaltando las calculadas erróneamente.

En Abduljaleel et al. (2021) se usa TF-IDF para representar el texto dividido en el título, las palabras claves, los resúmenes, las referencias y el resto del texto. La medida de similitud propuesta está basada en realizar la similitud de coseno de cada parte y asignarle un peso y concluye calculando la medida de similitud total ponderada. Se realiza también la detección de conflicto de intereses.

En D. K. Pradhan et al. (2020) el puntaje de experiencia del revisor se calcula en varios pasos. Se utiliza LDA para extraer el conjunto de temas de cada revisor. El puntaje inicial es obtenido usando un modelo de espacio de vectorial, comparando este archivo de revisor con manuscrito. Este puntaje es obtenido realizando un producto escalar entre las publicaciones del revisor y el manuscrito, luego de normalizar y suavizar los datos obtenidos. La experiencia declarada del revisor es considerada una verdad fundamental y es combinada con el puntaje inicial y un algoritmo de predicción supervisada se corre para obtener el puntaje final. Se calcula la distancia de coautoría para detectar el conflicto de intereses para determinar si es autor, coautor de un autor, coautor de un coautor o si no existe relación. La asignación final busca maximizar la similitud del tema y minimizar el posible conflicto de intereses.

En Patil & Mahalle (2021a) el proceso consta de tres pasos: 1) muestreo de Gibbs con LDA para representar el título, el resumen, la introducción del manuscrito y los artículos científicos de los revisores, usándose para extraer los temas de los manuscritos y la experiencia de los revisores en base a sus publicaciones con referato, 2) análisis de tema que se ocupa de encontrar la relación entre el título, el resumen, la introducción, las palabras claves y la conclusión como contenido de supervisión para interpretar el área de investigación del artículo. La relación del tema, definido como área de investigación, entre el título y el resumen se utiliza para entrenar el modelo. Este modelo se aplica para calcular la similitud entre un manuscrito y las publicaciones del revisor y 3) medida de relevancia donde se calcula el índice de similitud entre las publicaciones de cada revisor y el artículo utilizando los pesos de temas principales y ordenándolos de forma descendente para obtener los cinco revisores más relevantes. También se controla el conflicto de intereses y en base al ranking se asignan los revisores apropiados al manuscrito.

En Xu et al. (2022) con el uso de *prompt tuning* se obtienen los dominios de investigación por cada artículo introduciendo el título y el resumen. La plantilla se utiliza para procesar el texto original con fichas adicionales, mientras que el verbalizador proyecta las etiquetas originales en palabras del vocabulario para su predicción. Esto se efectúa con los manuscritos y los artículos científicos de los revisores calculando el porcentaje de cobertura que corresponde al porcentaje de dominios de investigación que los revisores comparten entre ellos. El otro puntaje que se utiliza en la metodología corresponde al de tendencia que utiliza "The Toronto Paper Matching System" (TPMS). Este puntaje indica la experiencia del revisor y la disposición a realizar la revisión. También se trata la importancia de un orden en la asignación de revisores y se proponen agrupamientos para ordenarlos. Los autores unifican las dos fórmulas anteriores en una sola fórmula, sumando el puntaje de cobertura y el de tendencia suministrado por TPMS con una normalización ponderada con pesos complementarios.

Los autores Al-Zboon et al. (2019) aplican LDA para extraer los temas de los títulos de los artículos, luego utilizan una base de datos gráfica para almacenar el contenido usando un lenguaje de consultas específico para esta tarea y recuperar contenido. Los datos son guardados como gráficos de nodos con las relaciones entre ellos. Se utilizan cuatro etiquetas: Artículo, Autor, Tema y Lugar; usando el lenguaje de consultas se busca a los 10 autores activos más relevantes en ese tema. En una primera instancia son considerados candidatos a revisores hasta que se haga un control de conflicto de intereses. Se aplica el algoritmo de predicción para encontrar revisores candidatos con dicho conflicto, quitarlos del listado y para definir la lista final de revisores recomendados.

El artículo de Yong et al. (2021) no fue incluido en los grupos previos porque se considera que la metodología puede tener diferentes reglas que tomen diferentes factores aunque en el ejemplo desarrollado en la publicación solo toman dos factores y excluye Conflicto de intereses. La metodología propuesta no toma múltiples factores en el cálculo, sino que los usa como regla. Se aplica una herramienta de extracción de texto y una herramienta de segmentación para tomar palabras clave de registros del texto completo, seleccionando las palabras centrales más importantes que fueron extraídas utilizando el algoritmo Word2Vec. Estas fuentes de reglas se combinan con diferentes requisitos definidos para generar conclusiones mediante el uso de razonamiento basado en reglas. El procedimiento configura el motor de reglas que analiza las conexiones semánticas de palabras claves y hace la inferencia basada en la base de conocimientos y los datos de entrada. A saber, primero se especifica la ontología del manuscrito y de los revisores que consiste en un conjunto de definiciones conceptuales, propiedades y relaciones, luego se aplica el algoritmo y las reglas con la finalidad de tomar un artículo y un revisor como entrada, ejecutar la regla y generar el resultado como verdadero o falso. Los autores en su artículo desarrollaron un ejemplo que se considera relevante sobre la realidad de reglas que puede tener la metodología en un caso real, tomando el nivel de estudios como regla de factor de autoridad, el campo de investigación y las palabras claves extraídas como factor de experiencia para verificar las reglas de similitud. Asimismo, en el trabajo aclaran que podrían incluirse en estas reglas condiciones basadas en otros factores que complejizaría la metodología.

En la Tabla 8 se presenta una relación entre las técnicas y modelos de PLN utilizadas y una abstracción en un modelo general del proceso que abarca la extracción, representación y asignación. Cabe aclarar que cada uno de los modelos tiene una complejidad en su desarrollo y que toda abstracción es una generalización donde es probable la pérdida en parte del complejo entramado de cálculos, fórmulas y algoritmos que se llevan adelante hasta la asignación. La generalización presentada debe entenderse solo a nivel orientativo de los procedimientos seguidos en los pasos generales a los fines de obtener la solución.

Tabla 8

Relación entre técnicas y modelos de PLN, y modelo general

Total	3	1	1	1	8	1
clustering con modelo de temas	0	0	0	0	0	0
modelo de gráfico difuso sobre modelo	0	0	0	0	0	0
modelo supervisado que combina experiencia	0	0	0	0	(D. K. Pradhan et al., 2020)	0
modelo que combina modelo VSM y declaración de revisores	0	0	0	0	0	0
modelo de espacio vectorial	(Latypova, 2023)	0	0	0	0	0
similitud entre palabras claves	0	0	0	0	0	0
razonamiento basado en reglas	0	0	0	0	0	0
modelo supervisado de pares de oraciones	0	(Duan et al., 2019)	0	0	0	0
modelo de etiquetas jerárquicas	0	0	0	0	0	0
modelo de temas con capas	0	0	(Jin et al., 2019)	0	0	0
modelo de temas, base de datos gráfica y	0	0	0	0	(Al-Zboon et al., 2019)	0
Modelo de temas y	0	0	0	0	(Patil & Mahalle, 2021a)	0
modelo de temas	0	0	0	(Kusumawardani & 0	(C. Yang et al., 2020)	0
modelo largo de lenguaje	0	0	0	0	0	0
combinación de modelo de temas	0	0	0	0	(Medakene et al., 2019)	0
cálculo de diferentes factores	(Abdujaleel & Al, 2021) (Kreutz et al., 2022)	0	0	0	(Nguyen et al., 2018) (Patil et al., 2023) (Patil & Mahalle, 2021b)	(T. Pradhan et al., 2021)
	TF-IDF	TF-IDF y red neuronal (BERT u otro CNN, BiLSTM)	AST model	AT model	LDA	LDA y TF-IRF

análisis podría resultar útil para determinar las fuentes de información de futuras propuestas metodológicas.

Tabla 9
Agrupamiento de las diferentes fuentes de información

Orden	Fuentes de datos	para evaluación	Información de artículos y perfiles	Cantidad total
1	Digital Bibliography & Library Project (DBLP)	(Patil et al., 2023), (Patil & Mahalle, 2021b)	(Nguyen et al., 2018), (Kreutz et al., 2022), (Kreutz & Schenkel, 2022), (Kalmukov, 2020b), (Medakene et al., 2019), (Patil & Mahalle, 2021a), (Patil & Mahalle, 2021b), (Kalmukov, 2020a), (Al-Zboon et al., 2019)	10
2	Neural Information Processing Systems (NIPS)	(T. Pradhan et al., 2021), (Medakene et al., 2019), (Patil et al., 2023), (Patil & Mahalle, 2021a),		4
3	Aminer		(Nguyen et al., 2018), (T. Pradhan et al., 2021), (T. Pradhan et al., 2021), (D. K. Pradhan et al., 2020)	4
3	Special Interest Group on Information Retrieval (SIGIR)	(C. Yang et al., 2020), (Xu et al., 2022), (Mirzaei et al., 2019)	(C. Yang et al., 2020), (Xu et al., 2022), (Mirzaei et al., 2019)	3
4	Google Scholar		(Abduljaleel & Al, 2021), (Patil et al., 2023), (Ishag et al., 2019)	3
4	ResearchGate	(Patil et al., 2023)	(Nguyen et al., 2018), (Patil & Mahalle, 2021a)	3
5	Arvix	(Duan et al., 2019), (Medakene et al., 2019)	(Duan et al., 2019)	2
5	CiteSeer	(Patil et al., 2023)	(Patil & Mahalle, 2021a)	2
5	CompSysTech database	(Kalmukov, 2020b), (Kalmukov, 2020a)		2
5	Conference on Artificial Intelligence of AAAI	(Patil et al., 2023), (Patil & Mahalle, 2021a)		2

5	Interspeech	(Patil et al., 2023), (Patil & Mahalle, 2021a)		2
5	Semantic Scholar		(Kalmukov, 2020b), (Kalmukov, 2020a)	2
5	The International Conference on Learning Representations (ICLR)	(Medakene et al., 2019; Stelmakh et al., 2021)		1
6	motor de búsqueda especializado en el ámbito de la medicina y las ciencias de la salud (PubMed)	(Mirzaei et al., 2019)	(Mirzaei et al., 2019)	1
6	Academic journal "MOIT"	(Latypova, 2023)		1
6	ACM Digital Library	(Dong et al., 2020)	(Dong et al., 2020)	1
6	Association for Computational Linguistics (ACL)	(Medakene et al., 2019)		1
6	base sintética (a partir de los datos de desambiguación procesados por conocido editor académico)	(Duan et al., 2019)	(Duan et al., 2019)	1
6	Conference on Computational Natural Language Learning (CONLL)	(Medakene et al., 2019)		1
6	Conference on Database Systems for Business, Technology and Web (BTW)	(Kreutz & Schenkel, 2022)		1
6	Conference on Information and Knowledge Management (CIKM)	(C. Yang et al., 2020)		1
6	Conferencia Internacional sobre la Teoría de la Recuperación de Información (ICTIR)	(Kreutz et al., 2022)		1
6	Congreso Internacional de la Asociación Catalana de Inteligencia Artificial (CCIA)	(Nguyen et al., 2018)		1
6	conjunto de datos de Mushroom (http://www.philippe-fournier-viger.com/spmf/)	(Ishag et al., 2019)		1
6	Database of School of Artificial Intelligence University of Chinese Academy of Sciences	(Yong et al., 2021)		1
6	elibrary.ru (base cienciométrica abierta).		(Latypova, 2023)	1
6	Evento Semantic Evaluation (SemEVAL)	(Mittal et al., 2020)		1

6	Internacional Conference on BUiness & Information Management (ICBIM)	(D. K. Pradhan et al., 2020)		1
6	International Conference on Artificial Intelligence and Information Technology (ICA2IT)	(Medakene et al., 2019)		1
6	International Conference on New Trends in Information and Communications Technology Applications (NTICT)	(Abduljaleel & Al, 2021)		1
6	material cursos: sitios personales y universitarios		(Medakene et al., 2019)	1
6	National Natural Science Foundation of China	(Jin et al., 2019)		1
6	Open Access Journal of Information Systems (OAJIS)	(Kusumawardani & Khairunnisa, 2018)	(Kusumawardani & Khairunnisa, 2018)	1
6	Papers independientes	(Al-Zboon et al., 2019)		1
6	The European Conference on Information Retrieval (ECIR)	(Kreutz & Schenkel, 2022)		1
6	The Molecular Machine Learning Conference (MOL)	(Kreutz & Schenkel, 2022)		1
6	The Semantic Scholar Open Research Corpus (S2ORC)	(Xu et al., 2022)	(Xu et al., 2022)	1
6	the Umm Al-Qura University	(Alkazemi, 2018)	(Alkazemi, 2018)	1
6	The World Wide Web Conference (WWW)	(C. Yang et al., 2020)		1
6	Wanfang Data		(Jin et al., 2019)	1
6	Web de ciencia de Clarivate Analytics	(Ishag et al., 2019)		1

1.6 Breves consideraciones

A partir de un análisis exhaustivo del estado del conocimiento, el presente capítulo ha abordado un mapeo sistemático de las soluciones automatizadas de recomendación de revisores tomando el período 2018-2023. La metodología implementada para dicho relevamiento siguió 5 preguntas clave de investigación que se respondieron a partir de la selección final de 25 artículos. En este recorrido, es posible sintetizar distintos aspectos de la problemática que resultan relevantes tanto en lo comparativo de las metodologías como en lo que aporta de información fundada para la posible propuesta metodológica de esta tesis doctoral.

En principio se reconocieron siete factores, siendo “Experiencia” el factor que está presente en todos los artículos, seguido por “Conflicto de intereses” y “Autoridad”. Un aspecto a considerar, es que solo una propuesta trata el factor “Revisiones”, siendo a juicio del autor de esta tesis doctoral una de las claves necesarias para evaluar la calidad de la trayectoria de revisión del investigador que solo se puede identificar de la información histórica de sus revisiones. Con respecto al factor *Seniority* pareciera que existe cierta superposición sobre el factor “Autoridad” y en los dos casos que se utiliza directa o indirectamente está asociado a “Diversidad”.

Sobre este último factor, cabe señalar una amplitud de interpretaciones de parte de los autores que arroja una significativa variedad (intradiversidad) de componentes. Todo lo cual provoca una notoria dificultad para establecer una conceptualización común al respecto. De lo analizado, se observa que cada aplicación práctica pareciera basarse en el concepto de diversidad en sí, pero en su desarrollo. Cada autor busca una forma de determinarlo según su propia interpretación. En este sentido, se estima que es más bien una perspectiva que se quiere incorporar a la metodología, que un factor interviniente en la línea de lo planteado en este trabajo.

En cuanto a las técnicas y modelos de PLN, 12 veces se utilizaron modelos de temas y 5 veces modelos de espacio vectorial. En un solo caso se efectuó una combinación de modelo de temas y de espacio vectorial. Es decir, el 72 % de los trabajos usan estos modelos. Lo común entre estos modelos es que no tienen incorporada información semántica, tampoco de las palabras por separado, ni del contexto. De las otras técnicas tratadas, se observan redes neuronales recurrentes, Word *embeddings* y Transformers, siendo este último el modelo más avanzado de los tres. Solo un artículo utilizó modelos largos de lenguaje, con la técnica *prompt tuning*.

En cuanto a analizar la metodología usada en cada artículo, se puede observar que hay una relación directa entre el uso de varios factores y la metodología. Cada factor es tratado en primera instancia de manera independiente con sus propios cálculos y se incorpora a una fórmula integral que tiene en cuenta todos los factores. Es llamativo el alto grado de diferenciación en las integraciones propuestas en cada artículo. El factor de conflicto de intereses es tomado como una restricción. En síntesis, no se observan mayores agrupamientos, solo hay dos que corresponden a su vez, a dos artículos cada uno que usan el modelo de temas y el modelo de espacio vectorial sin combinaciones con otros métodos. En esta dirección, no se profundiza en la asignación con restricción. Las restantes técnicas que se contabilizan alrededor del 50 % tienen metodologías muy diversas.

En cuanto a las fuentes de información hay una gran dispersión, encontrándose 42 distintas, siendo la más utilizada DBLP en diez artículos diferentes. Luego se agrupan desde cinco artículos hasta uno, en general se usan varias fuentes por artículo para encontrar tanto artículos de los revisores como perfiles de investigadores. Esto da cuenta a nivel mundial que no hay una base centralizada, ni existe una uniformidad que contenga toda la información. Es así que, según las necesidades que impulsaron la construcción de las diferentes metodologías, se seleccionaron diferentes fuentes de información según su utilidad relativa a los propósitos.

De todo lo expuesto, es posible afirmar que un significativo porcentaje de propuestas proponen métodos diversos para la resolución del problema de recomendación de revisores. Asimismo, se registra muy baja utilización de las técnicas más avanzadas de PLN. También hay un impacto directo entre la cantidad de factores a incluir y la metodología de resolución desarrollada, aunque siempre con matices diferentes. Siendo las metodologías más complejas y avanzadas las que incluyen múltiples factores. En este sentido, se desprende que esta sería la orientación más adecuada para la resolución de esta problemática diversa y compleja porque toma cada una de las variables que influyen en una recomendación final situada.

No obstante, estos avances, es importante siempre tener en cuenta que la recomendación de procesos automáticos como los descritos son información valiosa que se le acerca al equipo editorial para que sean ellos quienes tomen la decisión final al respecto.

2 Fundamentos teóricos-metodológicos

2.1 Consideraciones previas⁶

En acuerdo con Manovich (2020) las sociedades contemporáneas están mediatizadas no solo porque en la producción y circulación discursiva, los medios y tecnologías de comunicación tienen un rol clave, sino también porque las prácticas y las instituciones sociales se han estructurado paulatinamente en torno a artefactos y lenguajes de comunicación. Las dinámicas de publicación en Ciencia y Tecnología (CyT) no han sido ajenas a estos procesos que, potenciados por la globalización han generado un significativo aumento escalar de manuscritos sujetos a evaluación para su publicación. Esta situación ha provocado una importante demanda sobre la cantidad de personas que puedan realizar las evaluaciones dada la multiplicidad de eventos académicos y publicaciones de referencia.

La literatura sobre este tema, pone de manifiesto la directa relación entre la calidad de una revista científica o evento y el grado de excelencia del proceso de revisión que lo sustenta. En esta dirección, elegir las personas más adecuadas para efectuar la revisión de resultados de investigación se presenta desde hace varias décadas como un proceso de gran complejidad no exento de tensiones (Cuevas & Mestanza, 2002). Esta asignación calificada se inscribe como uno de los aspectos clave que garantizan posteriormente la consistencia de los dictámenes que orientan las decisiones de hacer público los manuscritos (K. Li et al., 2017).

Si bien (Ortiz Núñez, 2024) plantea los diversos cuestionamientos que se suscitan en el método “por pares ciegos”, especialmente en los sesgos y limitaciones que todo proceso humano conlleva, este sistema con más de dos siglos y medio de existencia se ha impuesto progresivamente en el ámbito mundial de la CyT y se continúa reconociendo en la práctica actual con algunas modalidades diferenciadas.

La tarea de asignar posibles revisores buscando cierta correspondencia con los textos propuestos es el gran desafío de los consejos editoriales y de quienes son responsables científicos de los eventos académicos. Esta selección debe tener en cuenta diversas precisiones tales como el grado de conocimiento en la temática del manuscrito (Liu et al., 2014), además de otros factores intervinientes pautados en las políticas editoriales. Dada esta situación, entre las propuestas y recomendaciones que buscan mejorar la calidad y fiabilidad de las publicaciones, se menciona brindar una diversidad de herramientas tecnológicas y estrategias organizacionales

⁶ El capítulo refiere con mínimas adecuaciones a la publicación (Autor & Autor, 2025)

para mejorar el proceso en tiempo y forma (Charlin & Zemel, 2013) (Ortiz Núñez, 2024) (Protasiewicz et al., 2016).

En este escenario, los avances de la IA han sido significativos al posibilitar la generación de sistemas digitales capaces de recomendar posibles revisores. Como ya se anticipó, desde principios de siglo XXI hasta el presente, numerosos artículos científicos han abordado esta problemática de búsqueda automática para la evaluación de trabajos científicos (Aksoy et al., 2023). En esta dirección, la presente tesis doctoral pretende abordar algunas de las problemáticas socio-técnicas que deben ser consideradas a los fines de fundamentar el desarrollo de una metodología integral de recomendación automatizada de revisores que sea flexible a las políticas y prácticas de un Consejo Editorial (CE) en referencia a la elección de pares evaluadores.

Para desarrollar este tópico se ha tomado como referencia una revista científica en Ciencias de la Educación publicada en Argentina que, acredita más de 25 años de trayectoria ininterrumpida e índices en bases de datos científicas reconocidas. Con varias características similares a otras de su tipo, su CE no utiliza a la fecha ningún tipo de herramienta automatizada que asista en la toma de decisiones de recomendación.

El marco teórico socio-técnico cultural que se presenta a continuación, es resultado de una elaboración participativa situada que permitió consensuar aspectos que sustentan la propuesta metodológica de desarrollo tecnológico a tal fin.

2.1.1 Síntesis del mapeo de metodologías

El capítulo anterior dio cuenta de un relevamiento estratégico del estado de conocimiento en el tema (Decoppet & San Martín, 2024). En el análisis de la bibliografía de referencia se clasificaron y cuantificaron aspectos generales, factores recurrentes en la asignación automática, modelos y técnicas de PLN, metodologías y fuentes de datos utilizadas por cada uno de los factores incluidos en las distintas soluciones propuestas. En atención a los límites del presente capítulo, se mencionará solo lo más significativo de dicho análisis a los fines de poder avanzar en las argumentaciones.

El estudio abordó una amplia diversidad en lo referente al desarrollo metodológico de cada artículo. Sin embargo, fue posible observar algunas coincidencias tales como la identificación de algunos factores recurrentes en las distintas propuestas. Estos factores son elementos clave que influyen en la evaluación de diversas dimensiones o criterios dentro de una metodología e incluyen: “Experiencia” (considerado en los 25 artículos), seguido de “Conflicto de intereses” y “Autoridad académica”. En menor medida son recurrentes “Diversidad”, “Interés del investigador” y “Seniority”. Un séptimo factor, “Revisiones” sólo es tratado en una

propuesta, cuestión que será abordada más adelante ya que resulta un aspecto ponderado en el presente trabajo. Por otra parte, en relación a “Seniority” pareciera que existe cierta superposición con “Autoridad académica” y en los casos que se utiliza está asociado a “Diversidad” de forma directa o indirecta. Sobre este último factor, se observa una alta variedad de interpretaciones en los escritos (intradiversidad) y se pueden identificar distintas componentes, dada esta situación no sería conveniente establecer una modelización universal al respecto, lo cual sería en sí mismo un contrasentido en términos teóricos.

Con respecto a la utilización de distintas técnicas de PLN, 23 artículos las aplican para seleccionar candidatos a revisores y se evalúan como eficientes en ese aspecto. En cuanto a la metodología desarrollada en cada artículo, se verifica una relación directa entre el uso de varios factores y lo metodológico. Cada factor es tratado de manera independiente con sus propios cálculos y es incorporado a una fórmula integral que los aborda en su totalidad. Es importante destacar que, los cálculos de cada factor y las integraciones propuestas son diferentes en cada artículo. Finalmente, se observa que las metodologías más complejas y avanzadas son las que incluyen múltiples factores. Este aspecto, es considerado en la presente tesis como la orientación más adecuada para la resolución de esta problemática en un contexto situado.

En relación a los factores considerados, en “Experiencia” se pondera el grado de conocimiento de académico de las personas a través de sus trabajos científicos publicados. La Tabla 10 muestra que partes son utilizadas para evaluar la experiencia considerando lo declarado textualmente en los 25 artículos analizados. Cabe señalar que en ninguno de los artículos analizados se explicita las relaciones de las partes y su pertinencia para medir la experiencia, por lo cual adoptando una perspectiva socio-técnico cultural, se ha consignado como un aspecto clave a definir tanto por miembros expertos en selección de evaluadores como por la aplicación de herramientas específicas que puedan ponderar similitudes entre textos.

Varios de estos artículos abordaron el factor 'Conflicto de intereses', analizando distintos aspectos, como se ilustra en la Tabla 11. No obstante, desde una perspectiva socio-técnico-cultural, es fundamental investigar qué otros tipos de conflictos podrían ser considerados relevantes por un CE, dado que este factor se entiende como una restricción para efectuar una elección determinada.

El factor 'Autoridad académica' está relacionado con el prestigio en el ámbito académico y, en la mayoría de los estudios revisados, se aborda a través de diversos índices bibliométricos. A continuación, se presentan las diferentes fuentes citadas por los autores analizados, las cuales fueron abordadas utilizando distintas metodologías. La tabla 12 muestra las diferentes fuentes que se declaran y se tratan con diversas metodologías.

Tabla 10*Partes de textos consideradas para evaluar experiencia*

Partes del texto consideradas	Cantidad de artículos
Todo el texto	1
Palabras claves	1
Resúmenes	2
Títulos y resúmenes	7
Títulos, resúmenes y palabras claves	2
Títulos, resúmenes e introducción	2
Títulos, resúmenes, palabras claves, introducciones y conclusiones	1
Títulos	1
Dividiendo todos los artículos científicos en cinco partes: título, palabras claves, resúmenes, referencias y el resto del texto	1

Tabla 11*Tipo de conflictos de intereses declarados*

Tipo de conflicto	Cantidad de artículos
relación de coautoría	7
misma afiliación	4
relaciones familiares	1
relación de tutoría	2
conflictos de competición	1

Tabla 12*Fuentes declaradas para "Autoridad académica"*

Artículos	Índice h	Nº de citas	Doctorados supervisados, libros y capítulos de libros, artículos en revistas y conferencias	titulación o nivel de estudio
4	X	X		
1	X			
1		X		
1	X		X	
1				X

La tabla precedente resume un agrupamiento de impacto de la producción científica, de la actividad de investigación y grado académico. Se observa la preponderancia del agrupamiento impacto, lo cual se reconoce como un indicador de la valoración de la comunidad científica de la producción de quienes investigan. En este sentido, se considera más adecuado dicho agrupamiento destacando que la metodología que lo utiliza, lo plantea de ese modo. Otro de los artículos tiene en cuenta la titulación o nivel de estudios como restricción, lo cual podría significar otra manera de sopesar la “Autoridad académica”, no obstante, se propone que podría ser usado no solo como restricción sino como puntaje.

El factor “Interés del investigador”, relevado en 5 artículos se denomina en algunos casos como *freshness* o *recency* y refiere a la preferencia por seleccionar artículos científicos dentro de un período de tiempo reciente. Esto puede implicar otorgar más peso a las publicaciones más cercanas en el tiempo, o utilizar escalas de ponderación según la proximidad temporal. La elección de esta denominación supone que los intereses de investigación en lo académico pueden variar con el tiempo, lo que sugiere que las producciones más recientes darían cuenta de las actuales áreas de interés.

Latypova (2023), propone el factor “Revisiones” evaluando a las personas expertas a partir de tres indicadores: 1) el lapso temporal de su experiencia como revisor, 2) la cantidad de días totales dedicados a la revisión en relación con el total de revisiones solicitadas y 3) la proporción de evaluaciones efectuadas avaladas de forma positiva por el equipo editorial en relación con el total de revisiones solicitadas.

En los 25 artículos se revela que no existe una única fórmula que responda a la totalidad de los posibles requerimientos de los comités editoriales de las revistas en CyT en circulación. Si

bien se consideran patrones comunes, el contexto de práctica de cada comité podría requerir una herramienta flexible basada en una metodología que le posibilite tener en cuenta la influencia de cada factor de forma individual y en el cálculo final. Todo lo cual sería más beneficioso y significativo a la hora de tomar la decisión de quienes serán los pares ciegos asignados a los diferentes manuscritos.

En síntesis, los desarrollos metodológicos informáticos estudiados exploran soluciones tecnológicas innovadoras en base a fórmulas cerradas con poca adaptabilidad general a las definiciones y criterios de las diferentes revistas científicas. Esta problemática se puede observar, por ejemplo, en Patil et al. (2023) cuando presenta como único grado de flexibilidad una asignación final que se efectúa a través de una suma ponderada de “Experiencia”, “Autoridad académica” e “Interés del investigador”. Asimismo, en este último proceso (Latypova, 2023) realiza una evaluación multicriterio expresada por un indicador aplicando una suma ponderada de diferentes parámetros relacionados a sus valores máximos. En T. Pradhan et al., (2021a) se extraen los temas de las publicaciones de los investigadores y su contenido lexicográfico, se calculan las similitudes y luego se suman con normalización ponderada de pesos complementarios.

Otras metodologías, como la propuesta por Medakene et al. (2019), han integrado los valores de similitud de experiencia y referencias mediante una normalización ponderada de pesos complementarios. Por su parte, Abduljaleel et al. (2021) sugieren una representación del texto dividida en título, palabras clave, resúmenes, referencias y el resto del contenido, en la que el procedimiento calcula la similitud de cada sección, asigna un peso y luego calcula la medida de similitud total ponderada. En el trabajo de Xu et al. (2021), se identifican los dominios de investigación, calculando por un lado el puntaje de cobertura y, por otro, la experiencia del revisor y su disposición a realizar la revisión, lo que denominan como puntaje de tendencia. Posteriormente, la metodología combina ambas fórmulas, sumando el puntaje de cobertura y el valor de tendencia, con una normalización ponderada de pesos complementarios.

A partir de lo expuesto, es posible afirmar que cada una de las metodologías analizadas abordó la problemática con el objetivo de ofrecer una solución flexible. No obstante, se identificó un obstáculo significativo: las adecuaciones situacionales necesarias para definir los diversos CE siguen siendo muy limitadas. Esto resalta la necesidad de enfoques más adaptables que consideren la diversidad de escenarios en los que se aplican estas metodologías. Asimismo, ninguno de los estudios analizados aborda de manera explícita la contribución de cada autor en artículos científicos de autoría múltiple. Esta omisión podría atribuirse a la dificultad de obtener dicha información, aunque en la actualidad algunas revistas han comenzado a exigir explicaciones más detalladas sobre este aspecto. Esta tendencia sugiere un posible cambio en

las prácticas editoriales, lo que podría fomentar una mayor transparencia en la asignación de autorías.

2.2 Perspectiva socio-técnica cultural

En vistas a lo que puede aportar en la actualidad una tecnología, es importante tener en cuenta que la recomendación de procesos automáticos es solo una información de referencia a la que puede acceder un CE. Desde este posicionamiento, la decisión final en la elección de quienes podrían realizar con calidad científica, compromiso ético y, en tiempo y forma dicha tarea es responsabilidad ineludible de dicho grupo de trabajo.

Ortiz Núñez (2024), pone en foco varias alternativas para la mejora y fortalecimiento del proceso de evaluación por pares doble ciegos aclarando que es necesario considerar entre los distintos factores de orden socio-técnico, las diferencias emergentes según los contextos de práctica. En relación a la temática de este trabajo el autor destaca entre otros, el valor de brindar directrices de evaluación claras y objetivas, promover el trabajo colaborativo, la formación específica y orientaciones en las tareas de evaluación, y el vínculo permanente de retroalimentación entre comunidad científica y sociedad. Esto lo asocia también al logro de un efectivo sostenimiento de las políticas y acciones promovidas por la Ciencia Abierta en cuanto a la rigurosidad y circulación del conocimiento en los distintos campos del saber. En síntesis, la mejora del complejo proceso de evaluación por pares ciegos requiere indefectiblemente de la participación de una amplia comunidad integrada por investigadores, académicos y editores, que como red socio-técnica responsable y colaborativa evite estratégicamente los posibles sesgos que puedan resultar excluyentes de valiosas contribuciones científicas.

Desde una visión filosófica sobre este necesario diálogo que habilita el proceso del conocimiento, (Meñica, 2023) plantea:

La clave del proceso de conocimiento, del progresivo manifestarse de la cosa es precisamente su alteridad. Lo otro, cuyo manifestarse suscita la pregunta, inicia el proceso que permite acceder a una visión más amplia. Cambia nuestro modo de ver, modifica el fundamento de nuestra subjetividad. Ese es el verdadero proceso del pensar que caracteriza a la razón humana. (...) El sistema formal del pensamiento es reproducible mecánica o electrónicamente. Los otros elementos que configuran la racionalidad no son transferibles (p. 591).

Llegado a este punto, el devenir histórico del proceso que culmina en una publicación científica validada por evaluaciones de pares ciegos da cuenta de los distintos estilos de época del fenómeno mediático (Verón, 2015). En la actualidad, es deseable que se genere en el marco de una red socio-técnica cultural colaborativa con finalidades y tareas consensuadas

fundamentada en una ética no excluyente, que ejerce presencia responsable en las interacciones plurales y utiliza tecnologías heterogéneas en el contexto físico-virtual donde se llevan adelante las actividades. Esta visión estratégica se encuadra en las recomendaciones de la UNESCO (2021) sobre la importancia de la Ciencia Abierta ante los complejos e interdependientes desafíos ambientales y socioeconómicos a los que se enfrentan la población y el planeta.

Desde un marco más general, (Andrés et al., 2023) proponen la noción de Dispositivo Intermedial Dinámico (DID) para este tipo de redes considerando entre otros, el enfoque constructivista de la tecnología (Thomas et al., 2019). La perspectiva DID aporta en lo teórico-metodológico interdisciplinar una propuesta analítica multidimensional compuesta por cuatro dimensiones con sus variables e indicadores específicos que posibilita guiar aspectos tanto de la construcción como de la Sostenibilidad de la red socio-técnica cultural para llegar a concretar sus fines. Las dimensiones que configuran este modelo analítico se las denomina “Político-Institucional”, “Social”, “Tecnológica” e “Intermedial”. Las mismas son complementarias entre sí, no se plantean con una relación jerárquica ni adscriben a un orden secuencial prefijado.

La potencialidad de este marco teórico y metodológico implica tener en cuenta aspectos que incumben tanto a las dinámicas de agencia social en la producción y reconocimiento como aquellos procesos de apropiación creativa de tecnologías. En este sentido, se entiende que la participación en un grupo bajo un enfoque metodológico interdisciplinar que aborde un desarrollo o mejora de una tecnología es clave para la producción o adecuación contextualizada de herramientas (Thomas, 2012).

Cabe mencionar que el DID adhiere a los lineamientos de acceso abierto y código abierto y ciencia abierta. Específicamente en todo lo concerniente a la publicación científica validada, las distintas herramientas digitales que se ofrecen tanto de código abierto como de pago presentan para un mismo requerimiento algunas diferencias en sus prestaciones o configuración como así también existe una significativa variedad según las distintas tareas del proceso. En principio, la apropiación creativa de tecnología implica poder realizar un detallado análisis de las prestaciones de cada artefacto indagando si hay posibilidades abiertas que habiliten formas alternativas de uso y de configuración. Esto es insoslayable ya que las políticas editoriales y normas de cada publicación presentan variaciones de distinto tipo.

Es importante tener en cuenta lo dificultoso que puede resultar a un CE lo concerniente a la extracción de características con diferentes técnicas y modelos de PLN aplicados en una metodología de desarrollo tecnológico para la selección automatizada de revisores. Esto incumbe a una investigación pormenorizada de las diferentes alternativas en base a un primer prototipo que permita compararlas. No obstante, resulta ineludible considerar todo el

conocimiento experto de dicho consejo para validar parte de las propuestas metodológicas que se inscriben en el marco teórico-metodológico del DID como abiertas y situadas. Todo lo cual implica una metodología con posibilidades de parametrización que permita además la elección de cuáles serían los factores que podrían influir en una correcta recomendación según los consensos del grupo responsable y la consecuente determinación de los pesos que explicitan los niveles de importancia al cálculo final integrado.

Esta perspectiva socio-técnico cultural se alinea con las directrices mundiales que promueven el desarrollo sostenible, ya que abona a la generación de procesos de innovación dando lugar a la creación de nuevas metodologías de diseño, desarrollo y productos tecnológicos promoviendo la participación y agencia colaborativa de agrupamientos diversos (Organización de las Naciones Unidas, 2015).

2.3 Argumentos para la discusión

A los fines de avanzar en una propuesta metodológica flexible y adaptativa en la temática de recomendación automática de revisores, se considera el antecedente del mapeo analítico ya mencionado y se toma como referencia situada una revista científica cuya incumbencia se encuentra en el campo de las Ciencias Sociales e inicia una nueva gestión editorial a mediados del año 2022 bajo el marco teórico metodológico del DID. La misma, se estructura en lo organizacional con una dirección (que por reglamento la ejerce quien dirige en instituto de investigación) y un CE compuesto por 11 nuevos miembros, ocho pertenecientes a otras instituciones (cinco nacionales y tres de otros países), un consejo científico con un porcentaje de miembros extranjeros mayor al 60%, un consejo de redacción de tres miembros profesionales y dos colaboraciones en la producción editorial, estos últimos pertenecientes al propio centro de investigación.

La publicación en acceso abierto cuenta con un Open Journal System (OJS) alojado en un servidor de dependencia institucional con personal informático dedicado. Si bien la revista acredita una larga trayectoria de circulación en formato impreso, desde hace 10 años la publicación es únicamente digital, registrándose una transición de tres años donde convivieron los dos formatos. Las reuniones de la Dirección con el CE se llevan adelante en modalidad híbrida utilizando una herramienta de videoconferencia, las comunicaciones por e-mail se realizan desde el correo de la revista y se utiliza en su potencialidad el OJS. La dinámica de las distintas actividades se efectiviza en los miembros que conforman y sostienen el DID Revista, lo cual posibilita un intercambio dialógico enriquecedor y resolutorio entre quienes participan.

En consideración de los resultados del mapeo internacional que orientaron algunas definiciones a considerar en lo metodológico y en relación a la elaboración de una propuesta socio-técnica cultural innovadora, la dirección de la revista planteó al CE la necesidad de especificar con mayor detalle los acuerdos sobre la elección de revisores. En esta dirección fue necesaria la creación de un instrumento encuesta que abordara además de algunos temas generales, preguntas particulares y posibles parametrizaciones. Esto cumpliría una doble función: por un lado, posibilitaría avanzar en la finalidad planteada y por otro, habilitaría discutir en profundidad ejerciendo una presencia responsable, algunos aspectos de la actual política editorial, visibilizar debilidades a subsanar, revisar normativas y otros aspectos que pueden atraer a nuevos autores y ampliar las audiencias⁷.

El instrumento utilizado fue una encuesta semiestructurada con opciones múltiples, compuesta por 14 preguntas. Esta permitía realizar observaciones abiertas, aplicar ponderaciones porcentuales en tablas en algunas de las preguntas y brindaba la posibilidad de añadir, modificar o eliminar datos cuando fuera necesario. Al finalizar, se invitaba a los participantes a proporcionar comentarios sobre si algún aspecto relevante no había sido incluido. Esto permitió acceder a un análisis integral de datos tanto de orden cuantitativo como cualitativo.

En lo tecnológico, el instrumento se desarrolló *ad hoc* en el gestor de contenidos Drupal, incorporando los módulos necesarios para el funcionamiento deseado incluyendo el logueo individual con acceso a la generación de un solo formulario “Encuesta” por cuenta de usuario. El mismo podía ser editado cuantas veces fuera necesario antes del envío final y solo podía ser accedido por la cuenta correspondiente. Desde un menú por pasos se presentaban algunas de las preguntas con tablas que posibilitaban ingresar las parametrizaciones por rangos, además de habilitar otras opciones de acuerdo a las respuestas ingresadas. También se incorporó una breve explicación del proyecto e información de ayuda, un archivo adicional con información de cómo completar las parametrizaciones con rangos (tablas) y un redireccionamiento a la página de información una vez realizado el logueo, para acceder a las encuestas.

El instrumento fue evaluado por cinco expertos: uno perteneciente al Comité Editorial (CE) de la Revista IRICE, otro miembro del CE de una revista extranjera de alto impacto en un ámbito similar, y tres miembros del equipo técnico de la revista, con experiencia en otras publicaciones y estudios en el campo de las Ciencias de la Información. Los expertos realizaron sugerencias de mejora, lo que resultó en la ampliación de la asistencia para el llenado de tablas,

⁷ Primera etapa del trabajo de campo de la presente tesis.

ajustes sintácticos en las preguntas para mejorar su comprensión y pequeños cambios en el orden de presentación de las mismas.

Del análisis de cómo se podrían asociar los distintos factores intervinientes en el instrumento encuesta se indagó en primer lugar, la relación entre estos para determinar si alguno cumplía el rol de disparador de la metodología de recomendación. De ese análisis, “Conflicto de intereses” se descarta como inicial ya que se controla una vez que se tienen seleccionados posibles evaluadores. Estas personas seleccionadas deben tener Autoridad académica pero la literatura señala que tiene solo influencia en la modificación de una selección. Lo mismo sucede con “Revisiones” e “Interés del investigador” que pueden igualmente influir en la priorización final, al igual que el factor agregado correspondiente al aporte de autoría. Cabe mencionar que, esa modificación en la priorización podría provocar la restricción de inclusión de algunas de las personas seleccionadas, como sucede también con “Conflicto de intereses”. Del mapeo realizado se desprende que quienes podrían realizar la evaluación deben tener conocimiento en las temáticas del artículo a revisar. En consecuencia, esto permite seleccionar un grupo de artículos que se aproximan al manuscrito sin considerar la autoría de los mismos. Así, “Experiencia” es el factor disparador de la selección de revisores a recomendar. Además, 'Experiencia' se basa en los artículos publicados por los candidatos a revisores. Tanto los 'Intereses del Investigador' como el 'Aporte de los autores' están asociados a esos artículos, lo que establece una relación directa entre ellos.

Finalmente, la encuesta se organizó a partir de la siguiente secuencia: “Experiencia” (con “Intereses del investigador” y “Aporte de autores”), “Conflicto de intereses”, “Autoridad académica”, y “Revisiones anteriores”, siendo indistinto el orden elegido de estos últimos tres factores. “Autoridad académica” se conformó tomando consideraciones referidas a indicadores bibliométricos y las titulaciones también con orden indistinto.

En relación al historial de revisiones de posibles candidatos, puede haber una variabilidad de eventos importantes a considerar según las distintas revistas. Es fundamental establecer acuerdos sobre diversos aspectos del proceso de revisión, tales como la cantidad de evaluaciones que se pueden pedir por año a un revisor, el tiempo de descanso entre solicitudes, la posibilidad de pedir revisiones simultáneas. Además, si es necesario evaluar la necesidad de aplicar penalizaciones por rechazos y revisiones de mala calidad.

Por último, caber reiterar que, parametrizar el factor diversidad es justamente efectuar una reducción semiótica del término por lo cual no se consideró pertinente incorporarlo en este primer instrumento que buscó avanzar hacia requerimientos posibles de un desarrollo tecnológico automatizado.

2.4 Resultados de la encuesta

La encuesta en su última versión fue respondida voluntariamente por 11 investigadores con trayectoria en consejos editoriales, 8 pertenecientes a la revista, 2 con funciones jerárquicas dentro del propio Instituto de Investigación y un investigador externo, reconocido por su desempeño internacional en gestión y evaluación de sistemas de CyT. A continuación, se sintetizará lo más significativo de las respuestas obtenidas. Cabe mencionar que en algunos datos numéricos aportados por los encuestados se va a realizar el análisis del promedio y de la moda, definiendo esta última como el número que se presenta con más frecuencia en un conjunto de datos.

La primera pregunta plantea si es determinante la proximidad temática entre el artículo a revisar y las producciones científicas de quienes podrían ser revisores en vistas a considerar su idoneidad en esa tarea. El 100% acuerda positivamente, sin embargo, una observación advierte que dependería de la necesidad y urgencia de la revisión y otra, que de haber múltiples revisores sería conveniente que al menos uno no tuviera excesiva proximidad temática.

El siguiente paso aborda el problema referido a qué partes de un artículo científico permitiría detectar con mayor eficacia el tema central del mismo. En este caso se observa cierta inconsistencia en las respuestas ya que, por ejemplo, en la opción “Contenido compuesto por: Título + Palabras claves + Resumen” el promedio general resulta de un 95% frente a “Contenido compuesto por Texto completo incluyendo todo, menos bibliografía” que arroja un promedio de 89,91%.

Ante la interrogación si podría afectar la calidad de la evaluación el tiempo transcurrido entre la solicitud de revisión y las últimas publicaciones relacionadas con la temática del artículo a evaluar, se obtuvo una mayor diversidad de respuestas. Siete investigadores opinan que sí, dos eligen otra opción, pero marcan cierta relevancia y dos dicen que no. Entre las observaciones de quienes responden en positivo se expresa que depende de la disciplina, de la evolución del conocimiento en el tema del artículo, que debe acreditar actividad en los últimos 12 meses en la temática, y no más de cinco años si es cercano al tema. En cuanto a la tabla que refleje la pérdida de actualidad del conocimiento según la antigüedad de las publicaciones, si bien se completan con diferentes valores, se puede observar una similitud en la tendencia.

La cuarta pregunta se centra en la posición de autoría relacionada a la variación del aporte de cada persona. El 45,5% eligió otra opción y el 18,18% respondió negativamente realizando observaciones referidas principalmente a diferentes criterios en los grupos de investigación y distintas tradiciones en los diversos campos disciplinares. En este ítem no se registró ninguna tabla que haya sido completada como dato adicional.

La siguiente pregunta de carácter abierto plantea el conflicto de interés. Las razones expuestas en las opiniones fueron variadas predominando los factores de proximidad intelectual y laboral, colaboración científico-académica y conocimiento personal. Sin embargo, cabe mencionar otros aportes recabados de mayor complejidad como la enemistad, cuestiones de posicionamiento paradigmático, resultados del artículo que significando un avance pueden competir entre grupos, cuestiones de interés empresarial, entre otros.

Sobre el impacto de las citas que puede tener en el perfil de quienes pueden evaluar, la mayoría le asignó un valor relativo observando, por ejemplo, que indican una cierta influencia en la comunidad académica pero que esto no asegura la calidad de la labor de revisión. Otra opinión señala que los sistemas de citas no emplean criterios puramente académicos ya que intervienen cuestiones institucionales y económicas que suelen otorgar cierta artificiosidad a los indicadores bibliométricos. Con respecto al nivel de estudio la coincidencia fue total con respecto al requerimiento de formación de postgrado y un alto porcentaje asignó el nivel de doctorado como lo más deseable. En relación a las tablas que reflejan una ventaja relativa basada en los valores del indicador elegido y en el nivel de estudio, pareciera no haber claridad en las respuestas acerca de qué se entiende como ventaja relativa.

En cuanto al tiempo máximo que un CE debería otorgar para la devolución de una evaluación experta de manuscrito, las opiniones oscilan entre un mínimo de 21 y un máximo de 90 días, con un promedio de 49 y una moda igual a 60. Esto es un tema que presenta con alta frecuencia dificultades, ya que más allá de las normas de cada revista, la devolución de los artículos evaluados dentro de los tiempos de las normas pautadas para el proceso editorial muchas veces no se materializa.

La pregunta nueve refiere a la influencia que podría tener la baja aceptación de pedidos de evaluación a una persona experta para poder efectuar una nueva solicitud. En este caso, siete respuestas fueron afirmativas, tres negativas y una respondió otra opción, expresando que depende de las razones por las cuales se excuse. Entre las observaciones positivas se pone en valor el seguimiento de las actuaciones, incluyendo la predisposición. Asimismo, se propone la posibilidad de una segunda y tercera vez de revisión. En las respuestas negativas se hace mención a tener en cuenta los motivos que pueden generar la no aceptación, lo cual se vincula con la respuesta de las razones que fundamentan la excusación. En relación a la elaboración de una tabla de penalización se pueda observar una tendencia similar, pero con amplitud de valores.

Acerca de considerar pertinente solicitar a una misma persona, más de una revisión de manera simultánea dentro del período asignado para la entrega de la misma, el 81,82% respondió negativamente. Quien planteó otra opción manifiesta que sólo en el caso que resulte

difícil hallar revisores en temáticas muy especializadas, por lo cual se estima que esto amerita ser discutido ante una singular situación.

También se indagó sobre el tiempo de descanso deseable entre revisiones. Las respuestas fueron positivas en un 100%, el valor mínimo fue de dos meses y el máximo de doce, con un promedio de 5,89 y una moda igual a 6. Entre las observaciones, se sugiere evitar la recurrencia de pedidos a quienes realizan el trabajo en tiempo y forma ya que el mismo no es pago, otra opinión también adhiere a evitar la sobrecarga, pero lo justifica en función de proteger la calidad de las evaluaciones, no obstante, aclara que no se necesitaría un tiempo prolongado. A la interrogación sobre el número de revisiones de diferentes artículos que podría solicitar un CE a un mismo revisor en un año, se registró un valor mínimo de 1 y un máximo de 4, con una moda y promedio igual a 2.

La pregunta trece indaga acerca de la necesidad de evaluar la consistencia de las revisiones recibidas. En caso de ser positiva la respuesta, se solicita proponer una tabla de calificaciones comparativa estimando un porcentaje a favor o en contra para considerar a dicha persona experta. También se da la opción de definir si hay una calificación neutra y si existen algunas que pueden ser valoradas como no satisfactorias. Acerca del envío de revisiones que no reúnan las condiciones esperadas, se preguntó sobre la cantidad que podría tener una misma persona y si puede considerarse un período de no consideración (penalización) a futuro. Y si fuera así, que lapso de penalización se estima adecuado. Sólo se registró una respuesta negativa referida a la necesidad de evaluar la consistencia de las revisiones, sin expresar opinión.

En las tablas de asignación comparativa de la mencionada pregunta, se relevaron valores entre un 20% y 50% de asignación positiva. Si bien la tabla presentaba cierta complejidad pudo establecerse en general una tendencia, pero por la amplitud de posibilidades sin restricciones que ofrecía dicha tabla, una respuesta evidencia una interpretación no precisa. Sobre la penalización se observa acuerdo, pero las opiniones son diversas en cuanto a la cantidad de revisiones no satisfactorias para aplicar la penalización (oscilan entre 1 y 4) y también en los períodos de tiempo se proponen diferentes parametrizaciones de penalización con una amplitud de valores entre 3 y 36 meses siendo la mayor penalización “tiempo indeterminado”. Al observar estos resultados, es imprescindible que el CE pueda consensuar con detenimiento estos aspectos. Por otra parte, en atención a la ética editorial es ineludible informar a quienes realizan la revisión de los mecanismos de selección que se ponen en obra.

La última pregunta solicita que se asigne a los ítems a) “Producción científica” que engloba “Experiencia”, “Intereses de los investigadores” y “Aporte de los autores”, b) “Autoridad académica” y c) “Revisiones anteriores”, un porcentaje con un valor total máximo de 100. En la totalidad de respuestas hubo dos cuyos datos se completaron de manera errónea, por lo cual

no se consideran. Los datos finales son: “Producción científica” resulta en promedio de 39,44%, “Autoridad académica” promedia 30,56% y “Revisiones anteriores” 30%, siendo las modas respectivas 40, 30 y 30. Si bien los valores asignados son muy cercanos, la tendencia privilegia la “Producción científica”.

Como ya se anticipó, el final de la encuesta invitaba a realizar algún otro aporte o expresar un posicionamiento sobre la problemática. Se registraron cinco contribuciones, dos de estas ponderan favorablemente el tratamiento de la problemática y agradecen haber participado de la encuesta. Otra contribución señala la diversidad que puede haber en usos y costumbres de otras comunidades disciplinares tomando en cuenta el contexto de la presente encuesta. Dos opiniones se expresan en la problemática del desempeño de la persona experta, se señala sintéticamente que hay criterios cualitativos que quizás sean difíciles de cuantificar (o de automatizar. En este sentido, se advierte, que alguien con importante número de publicaciones no necesariamente puede desempeñarse satisfactoriamente en la tarea de evaluación de manuscritos, por lo cual la variable referida al desempeño de una persona en dicha labor debe ser bien ponderada ya que su habilidad en revisiones en las futuras solicitudes debe ser monitoreada bajo la responsabilidad del editor en jefe (o el CE). En otras palabras, se pone en valor la gestión editorial en su conjunto.

Finalmente, se puede inferir que la encuesta movilizó a quienes participaron a realizar un ejercicio reflexivo en la temática con cierto nivel de complejidad, especialmente en el armado de tablas. A nivel general, las tablas correspondientes al valor de representatividad no tuvieron mayores inconvenientes de interpretación, caso contrario se observó en las que demandaron interpretar las ventajas relativas. De acuerdo a la observación de valores de estas últimas se infiere que algunas de las respuestas arrojan una interpretación como valor de representatividad.

2.5 Breves reflexiones analíticas

A partir de lo expuesto a los fines de avanzar en una propuesta metodológica de recomendación de revisores automatizada se plantea imprescindible atender a los aspectos socio-técnicos culturales de la problemática. Esto implica que, a sabiendas de la diversidad de contextos editoriales y sus políticas, las determinaciones que se puedan establecer ante los distintos factores que integran la metodología estarán sujetos a variables criterios de pertinencia y/o valoración dependiendo, entre otros, de la posibilidad de encontrar personas adecuadas a la tarea. Esto constituye un argumento central del presente enfoque metodológico, devenido de la perspectiva DID que pone en foco el dinamismo de lo social y el devenir permanente de los cambios tecnológicos.

En síntesis, la metodología que se propone se puede dividir en tres globalizadores de factores: 1) “Producción científica”, 2) “Autoridad académica”, y 3) “Revisiones anteriores”. Asimismo, cada factor tiene sus variables que representan diferentes partes del problema en su conjunto, como por ejemplo lo referido en la encuesta al tiempo de descanso entre solicitudes, cantidad de envíos, entre otros.

La producción científica se ve influenciada por diversos factores, entre ellos la experiencia en temáticas similares al manuscrito, el interés del investigador reflejado en la antigüedad de sus publicaciones, el aporte de los autores y la posible existencia de un conflicto de intereses. Es importante destacar que la experiencia y el conflicto de intereses no son opcionales, ya que condicionan directamente la rigurosidad y validez del proceso de selección.

Por otro lado, la autoridad académica puede considerarse un criterio adicional y está determinada por indicadores bibliométricos y titulaciones. Ambos factores son opcionales, lo que permite cierta flexibilidad en la evaluación. En caso de no disponer de datos sobre algunos posibles candidatos, se recomienda establecer un valor por defecto para cada factor, asegurando así un proceso de selección más equitativo.

“Revisiones anteriores”, en caso de utilizarse, se considera con tres factores: Cálculos (donde se obtiene el valor de revisiones anteriores), Cortes (en base a determinados valores de control diferentes al de cálculo), y Cálculos y cortes. Este último realiza las 2 opciones anteriores. En este sentido, si se realiza el cálculo sería necesario definir el valor a incluir si la persona no registra revisiones anteriores.

La confirmación del 100% en referencia a la importancia de la similitud temática vinculada a la idoneidad de quien realizaría la evaluación solicita sopesar la eficiencia de diferentes técnicas de PLN. Dado que este factor es obligatorio en la metodología para la detección automática más eficaz del tema central del artículo científico, se hace necesario verificar el análisis automatizado sobre cada texto con una evaluación de una persona experta en la temática, a los fines de definir la técnica de PLN más adecuada.

Sobre la antigüedad de las publicaciones científicas, en general hay una valoración positiva hacia tener presente este factor y puede inducirse un valor por defecto. La posición de la autoría en el caso de estudio no es representativa, sin embargo, puede ser importante para otros equipos editoriales de distintas disciplinas o revistas científicas que puedan tener criterios diferentes.

En cuanto a “Conflicto de intereses”, se verifica en el caso de estudio alto grado de unanimidad de opinión, pero teniendo en cuenta diferentes circunstancias. Las dos agrupaciones con mayor presencia en los aportes de la encuesta, es decir colaboraciones con

otros investigadores e institutos y universidades en donde trabajan podrían ser considerados en procesos automáticos.

Sobre la influencia de las citas, en general parecería no incidir en el contexto situado del estudio, sin embargo, *h-index* como opción fue la más nombrada entre otras posibilidades. En referencia a la titulación de las personas expertas, en general predomina como deseable el nivel de doctorado. La diferencia de interpretación de la ventaja relativa de estos dos puntos anteriores, da cuenta de la necesidad de volver a discutir este tema al interior del CE del caso referenciado.

En relación a algunos datos solicitados en la encuesta, la mayoría acuerda que no se puede pedir más de una revisión y se registra una aceptación del 100 % sobre el tiempo de descanso entre solicitudes, pudiéndose establecer un valor al respecto. Teniendo en cuenta la consulta sobre la cantidad de evaluaciones que se pueden solicitar a una misma persona en un año, fue posible establecer ese valor y hay una coincidencia indirecta con el período de descanso y solicitudes simultáneas seleccionadas por quienes participaron. Cabe mencionar el consenso que se registró en referencia a lo necesario que puede resultar evaluar la consistencia de las revisiones recibidas. También la mayoría del CE de la revista estudiada acuerda con algún tipo de penalización por rechazo de revisiones y por aquellas con baja consistencia, lo cual permite realizar una parametrización. Sin embargo, es necesario que dicho CE llegue a un acuerdo para poder evaluar la dispersión de dichas penalizaciones.

Como último paso, en el marco de la perspectiva socio-técnica cultural es posible atribuir los porcentajes de relevancia de los globalizadores como valores de representatividad. Más allá de los diversos contextos disciplinares y políticas de cada publicación, la encuesta visibiliza como en un CE de una misma institución pueden coexistir distintas interpretaciones o diversidad de criterios que deben ser discutidos y consensuados de forma dinámica según las singularidades que se presenten.

Esto ratifica la flexibilidad que solicita la consideración o no de los factores que conforman la metodología para otorgar una sostenibilidad calificada al proceso de recomendación automatizado. Como se pudo observar en los resultados de la encuesta, las dimensiones puestas en juego de lo Institucional, Social, Técnico e Intermedial están presentes cuando se realiza un análisis multidimensional y complejo de la problemática que da lugar a la innovación tecnológica. Esta innovación debe brindar un servicio adecuado a los requerimientos situados. Todo lo cual implica en la cuantificación en los casos que correspondiere, otorgar la posibilidad de modificación en el contexto de las políticas consensuadas por los diversos CE de revistas en CyT. Siendo esto independiente a la propuesta de valores por defecto aplicable a un caso situado que puede ser también modificado cuando el CE lo desee.

Finalmente, para discutir con mayor profundidad las fortalezas y debilidades de la metodología integral propuesta, es esencial poner en obra un primer prototipo de laboratorio que pueda ser utilizado por los destinatarios. Estos, a partir de su experiencia proporcionarán retroalimentación que contribuirá al avance de investigaciones interdisciplinarias bajo la perspectiva socio-técnica cultural orientadas al desarrollo de estos sistemas automatizados, con mayores posibilidades de ser flexibles y adaptables a la diversidad de contextos editoriales, abriendo así nuevos desafíos a las prácticas de publicación científica.

2.6 Hacia una metodología integral

Llegado a este punto, se especifica la siguiente estructura secuencial para la metodología integral propuesta que será detallada en este apartado.

Producción científica:

- Experiencia
- Interés del investigador
- Aporte de los autores
- Conflicto de intereses⁸

Autoridad académica del investigador

- Índices bibliométricos
- Titulaciones

Revisiones anteriores

- Simultaneidad de solicitudes de revisiones
- Tiempo de espera entre solicitudes de revisiones
- Revisiones por año
- Rechazo de pedidos previos de revisores
- Evaluación de calidad de las revisiones realizadas
- Rechazos por revisiones de mala calidad

2.6.1 Globalizador de factores: Producción científica

Resulta prioritario evaluar la concordancia entre el artículo a ser revisado y sus posibles revisores. Se espera seleccionar a aquellos con mayor conocimiento relevante asociado al artículo. Para lograr esto, es necesario determinar el nivel de conocimiento de los candidatos.

⁸ Cabe señalar que, se ha incluido el factor de conflicto de intereses en la producción científica porque para evaluar dicho conflicto es necesario primero obtener la lista de artículos de los candidatos a revisor y luego de que se crea, priorizar y depurar la lista de los candidatos.

En esta dirección, mayor grado de conocimiento asociado al artículo, mayor posibilidad de ser elegido como revisor. Esto se vincula principalmente al factor experiencia lo cual implica una significativa complejidad en el cálculo ya que se vinculan múltiples factores como interés del investigador que evalúa la antigüedad de la producción científica y aporte de los autores que busca determinar cuál es el grado de contribución de cada autor en los artículos. Además del conflicto de intereses que se incluye como una restricción, que puede realizarse solo cuando se ha depurado un grupo inicial de posibles revisores.

Factor experiencia:

La experiencia es central en la evaluación y selección de los candidatos más idóneos. A continuación, se presentan aspectos relacionados a este factor:

Secciones representativas del artículo

Cabe señalar que a la fecha no existe una base de datos a nivel mundial actualizada que agrupe todos los temas en los que son expertos todos los investigadores, con el armado de taxonomías muy bien definidas para esta tarea. Los proyectos que existen no satisfacen por ahora este requerimiento. Ante esta situación, los equipos editoriales recurren por lo general a los expertos del consejo científico de su revista para propongan investigadores con el mayor prestigio posible que traten temas lo más cercano posible al artículo a revisar. Si bien no se deja de valorar el alto nivel académico que tienen todos los grupos intervinientes de las revistas científicas, esto resulta muchas veces dificultoso. Esto conlleva, por ejemplo, a la necesidad de leer las producciones científicas de una gran cantidad de investigadores para encontrar los mejores revisores, lo cual es un proceso muy largo. Entonces, se plantea el otro interrogante sobre cuáles son las secciones representativas de los artículos científicos para comprender efectivamente de qué trata sin necesidad de su lectura completa.

Sobre este problema, podría inferirse que el título es una información muy resumida y concisa de su contenido, el resumen tiene la finalidad de comunicar sintéticamente el aporte significativo del artículo completo y las palabras claves sirven para poder comunicar diferentes tópicos que aborda. El título, el resumen y las palabras claves tienen un sentido de abarcar todo lo que trata el artículo de diferentes maneras y en forma resumida. Además, sin quitar importancia a la bibliografía podríamos inferir que para esta tarea no sería necesaria. Es necesario tener en cuenta que toda reducción o síntesis puede llevar consigo una pérdida de información valiosa contenida en el artículo, por lo que se considera que toda representación parcial del texto debe ser analizada, comparada y cuantificada en el presente estudio para determinar si alguna puede ser lo suficientemente representativa. Además, debe tener en cuenta el esfuerzo necesario de la tarea como una de las variables a analizar comparada con la

representación total de los artículos debido a la necesidad de leer o procesar todos los artículos científicos disponibles. Todo lo cual es una gran cantidad de información con afectación a la carga de todos los artículos y al tiempo de lectura o procesamiento necesario por cada uno. Es decir, si con todo el texto del artículo se lograra el mayor grado de efectividad en la tarea y con la de cualquiera de las tres representaciones lograrán un valor apenas inferior, no quedaría dudas que por la economía de esfuerzo sería la solución adecuada.

Llegado a este punto existen cuatro opciones de mínima para caracterizar un artículo científico sin tomarlo de forma completa: 1) artículo completo sin bibliografía, resumen, palabras claves, ni título, 2) título, 3) palabras claves y 4) resumen. Y las combinaciones que pueden realizarse entre estas diferentes partes.

En los casos del mapeo sistemático que explicitaron esta problemática se determinó que la mayoría se basó en títulos y resúmenes. Sin embargo, los mismos no justifican la elección de este criterio. Por tal motivo y para ratificar o rectificar esta selección la encuesta abordó la pregunta sobre diferentes partes agrupadas o no del artículo que permiten detectar eficazmente los temas que tratan y su porcentaje de representatividad. En algunas respuestas de los participantes se observó ciertas inconsistencias que globalmente dieron cuenta de la necesidad de continuar indagando este tópico en el prototipo propuesto como segunda etapa del trabajo de campo.

Técnicas de PLN para extracción de características

Una vez visto el tema de la representación, se debe considerar la importancia de poder capturar y medir el conocimiento que trata un manuscrito y el que poseen sus posibles revisores para lograr una adecuada vinculación. Para acceder a mayores posibilidades de elección calificada de un revisor, la cantidad de revisores debe ser amplia, por lo cual habría que contar con una cantidad significativa de artículos científicos, que representa un gran esfuerzo para los CE. Esta tarea amerita la asistencia de los avances en PLN que posibilitaría automatizar el proceso implementando una técnica adecuada.

Bajo esta premisa, en el marco del contexto planteado en este trabajo, se descarta el uso de modelos basados en LLM debido a limitaciones tanto prácticas como metodológicas. En particular, estos modelos presentan severas dificultades en términos de auditabilidad y reproducibilidad del proceso de selección, ya que su funcionamiento interno resulta considerablemente menos interpretable en comparación con los modelos vectoriales tradicionales. Asimismo, su implementación a gran escala conlleva costos computacionales elevados, manifestados tanto en la infraestructura requerida para despliegues locales como en el consumo de tokens en servicios externos. Finalmente, la dependencia estructural de datos

preentrenados introduce una incertidumbre adicional respecto a la trazabilidad de los criterios de similitud empleados, lo que lleva a priorizar enfoques más controlables y transparentes. Esto no quita, sin embargo, que los MLL puedan ser utilizados dentro de la metodología general con otros fines específicos y delimitados en los que demuestren ser muy valiosos.

Se trata entonces de tomar la mejor representación de los textos de los artículos científicos y en base a definiciones metodológicas, encontrar las mejores técnicas de PLN para una extracción de características adecuada que permita cuantificarlos y compararlos en un proceso automático eficaz y eficiente. De esta manera, se podrían determinar los posibles revisores que acrediten publicaciones de artículos similares al artículo a evaluar, con la menor carga de procesamiento y de tiempo.

En el análisis de las técnicas de PLN empleadas en el conjunto de artículos del mapeo sistemático, se identificó que los modelos de temas y los modelos de espacio vectorial fueron los más utilizados. El uso del primer tipo de modelo se basó en analizar los principales tópicos tratados en los artículos, con el objetivo de seleccionar aquellos más cercanos en dichas temáticas generales. Aunque esta aproximación es válida, se propone considerar la totalidad del texto en la representación elegida. Esto implica que la representación debe contener, de manera implícita, la combinación de temas abordados en el manuscrito, sin necesidad de descomponerlos explícitamente en conjuntos de tópicos específicos. En lugar de ello, se busca una representación única para cada documento, con un nivel de abstracción adecuado que permita reflejar la información contenida en su totalidad. Esto facilitará la comparación de dicha representación con las de otros artículos científicos en la base de conocimiento, que también integran implícitamente combinaciones de temas, con el objetivo de identificar similitudes de manera efectiva.

Otras metodologías relevadas realizan una búsqueda de revisores tomando la similitud de citas bibliográficas (Medakene et al., 2019). En Salinas et al. (2020) se realiza una construcción de conjunto de artículos, visualizando y explorando una red de citas bibliográficas tomando una base de datos de referencia que incluye artículos, autores y citas de fuentes seleccionadas. Autores como Rodríguez & Bollen (2008) plantearon una estructura de red de coautoría con algoritmo de enjambre de partículas pero en una forma discreta relacionada a los algoritmos de andador aleatorio del análisis de la cadena de Markov. Una de las hipótesis de estos modelos es que, si un autor fue citado en una de las producciones científicas, alguna parte del artículo se relaciona con ese artículo citado y eso permitiría que pueda ser considerado revisor del artículo. Cabe considerar que esta no es una regla unívoca por lo cual se estima la necesidad de caracterizar ambas producciones y compararlas.

A juicio del autor de esta tesis, la metodología de búsqueda de revisores considerando solo la bibliografía sería un camino acotado de búsqueda de revisores que podría conducir a

dejar fuera expertos que no fueron citados pero que acreditan publicaciones similares al artículo en cuestión. Entonces, la propuesta de este trabajo no se centra en una búsqueda directa de artículos y autores citados, sino que los considera en el contexto de los artículos y autores obtenidos mediante la extracción de datos de repositorios científicos relevantes para la tarea. No se limita únicamente a las citas, ya que la evaluación abarca una cantidad mucho mayor de artículos, todos los cuales compiten para seleccionar a los mejores revisores posibles.

Otro aspecto clave es garantizar que la elección de un candidato a revisor sea coherente con el conocimiento relevante asociado al manuscrito. Para lograrlo, se busca identificar los artículos de su autoría que sean más similares al artículo en revisión. La evaluación de esta coincidencia se realiza analizando, de manera individual, cada una de las publicaciones de los investigadores incluidos en la base de conocimiento. Aquellos que hayan publicado trabajos más similares al manuscrito a revisar tienen mayores probabilidades de ser seleccionados como revisores. Esta decisión se toma bajo la premisa de que el conocimiento es altamente granular y un artículo representa una combinación muy específica de saberes. Por lo tanto, priorizar los temas generales abordados en el artículo o usar una representación única por candidato a revisor no garantizaría capturar la profundidad y especificidad del conocimiento que se presenta en su combinación exacta, hasta en sus más pequeños detalles. Esto descarta la opción de trabajar con temas generales o representaciones únicas por revisor, y requiere la búsqueda de técnicas capaces de capturar esa granularidad, generando una representación precisa de cada artículo científico.

En síntesis, la selección de la técnica de PLN más adecuada, en el marco de los lineamientos establecidos por la metodología, debe ser el resultado de un análisis detallado realizado por expertos capaces de evaluar el rendimiento de cada una de las técnicas posibles. Esta evaluación será explicitada en el próximo capítulo correspondiente a la segunda etapa del trabajo de campo.

Técnicas de similitud

Luego de obtener una representación del documento a partir de una técnica de PLN que permitió la extracción de las características, se debe utilizar una técnica de similitud para comparar todas las representaciones de los artículos de la base de conocimiento con el manuscrito determinando cuáles son los más parecidos y así proseguir con el proceso de selección de revisores.

A lo largo del tiempo se han propuesto distintas medidas tales como la distancia euclidiana tomada de la geometría euclidiana que mide la distancia en línea recta entre dos puntos, la distancia de Manhattan similar a la euclidiana, el coeficiente de Jaccard que se utiliza

para medir cuánto se superponen dos conjuntos en términos de elementos compartidos en relación con el total de elementos únicos en ambos conjuntos, la distancia de Hamming que define el número de posiciones en la que los valores son diferentes, la divergencia de Kullback-Leibler que es una medida no simétrica entre distribuciones de probabilidad de los vectores y similitud de coseno que toma los ángulos entre dos vectores entre varias otras (Lin et al., 2014) y la similitud de coseno que es una medida utilizada comúnmente para determinar la similitud entre dos vectores en un espacio multidimensional. La justificación de la técnica de similitud a utilizar es un aspecto clave de la metodología.

Nombres de los autores

Uno de los problemas centrales en la implementación de una metodología es la elección de las fuentes de datos. Una posibilidad poco probable es contar con datos totalmente depurados y adaptados a la metodología. En una significativa cantidad de casos, las fuentes son heterogéneas y presentan los datos en diferentes formatos y estructuras, por lo cual debe realizarse un preprocesamiento que permita su depuración para poder ser utilizados.

Cuando los datos no están correctamente sistematizados y especialmente, cuando se usan fuentes heterogéneas, uno de los grandes problemas refiere a las diferentes formas que se presentan los nombres de los autores en sus publicaciones científicas. Ya sea por variaciones de uso del propio autor o por diversos requerimientos de formato donde deben ser publicados. Lo deseable es contar con una estructura de datos de autores bien conformada y cada vez que se necesita seleccionar y/o elegir a un autor, se tomen los datos cargados en el registro correspondiente. Si bien no existe a la fecha una base internacional de nombres universalmente utilizada, que posibilite la obtención de un formato único de nombres y apellidos de cada autor de todos los artículos científicos, se relevan algunas bases con distintos grados de sistematización donde se registra la información de las publicaciones científicas y los autores según requerimientos definidos por dichos contenedores de este tipo de información.

Dada esta situación, es muy común que los nombres de los autores figuren en diferentes formatos como, por ejemplo: con todos los nombres y apellidos, con abreviaturas de nombres intermedios, con abreviaturas del primer nombre o todos, con un apellido en caso de poseer varios, con todos los apellidos, con diferente orden de aparición de nombres y apellidos, y otros formatos en base a toda la variedad posible. Cabe también aclarar que utilizar los nombres como identificadores unívocos de una persona resulta inapropiado dado que puede haber dos o más personas con mismos nombres y apellidos, no solo por temas de formato.

A partir de lo expuesto, la desambiguación de nombres es un proceso muy importante dentro de la metodología para realizar una correcta recomendación. En este sentido, el objetivo

buscado es la minimización del error de considerar un autor como dos o más autores diferentes o la consideración de dos o más autores como si fueran uno.

Idiomas

El idioma es uno de los pasos que se trabaja en el preprocesamiento y sus definiciones son muy importantes, particularmente en esta metodología afectan directamente al globalizador de factores de conocimiento científico, motivo por el cual se incluye en esta sección. Es necesario considerar que los artículos científicos se escriben en diferentes idiomas cuyos autores registran diversas nacionalidades. Entonces, para efectuar una comparación de artículos en diferentes idiomas se deben tomar las representaciones de los textos y compararlas para determinar si los contenidos son similares.

En la actualidad no todas las técnicas de PLN pueden trabajar con diferentes idiomas. Una manera posible de solucionar este problema sería aplicar una técnica que posibilite llevar todos los textos a un solo idioma y luego así compararlos.

Umbral de conocimiento

La necesidad de considerar el umbral de conocimiento se basa en poder definir los límites de especificidad de un artículo. Esto se vuelve relevante cuando se registra más de un artículo de un posible revisor cercano al artículo a revisar. Entonces, es necesario definir cuál es la ventaja relativa que tiene ese candidato a revisor con múltiples artículos sobre los otros cercanos al artículo a revisar. Si bien el hecho de tener priorizado más de un artículo da una ventaja sobre quienes tienen uno solo, la misma es relativa y está asociada a los valores de similitud de sus artículos. A partir de ello, se debe sopesar si se considera todo el conocimiento de cada artículo en su representación para definir esa ventaja relativa o hay una parte de conocimiento que es específico sobre el total del escrito. A juicio del autor de esta tesis, lo correcto sería considerar solo esa parte específica. De allí, la importancia de definir el umbral del conocimiento, debido a que permitiría considerar solo dicha parte.

Aquí se presenta la dificultad de cómo saber cuándo un conocimiento es de un tipo específico y cuando pasa a ser general u otro tipo de conocimiento específico diferente. Definir esto en términos prácticos presenta una gran dificultad. Kreutz & Schenkel, (2022) lo definen en términos deterministas. Sin embargo, si bien un valor determinista puede servir para simplificar una decisión, se vuelve difícil fundamentar cómo se llegó a ese valor específico.

Entonces, resulta actualmente imposible definir ese valor real, pero sería posible plantear otra solución alternativa, formulada de manera relativa y asociada al número de

artículos incluidos en los cálculos más allá de la mencionada si se plantea de manera relativa asociándolo a un número de artículos a ser considerados para los cálculos. Por ejemplo, si se determina un número de artículos cercanos a ser considerados, se puede definir como umbral de conocimiento a la representación del artículo más alejada de aquellos que son considerados.

Factor Interés del investigador (antigüedad)

El motivo de su inclusión de este factor es la posible variabilidad en el tiempo de temas de estudio de quienes investigan, por lo cual se asume que sus publicaciones recientes reflejan de manera más precisa sus actuales intereses. Por tal motivo, se le asigna menor peso a las publicaciones más antiguas que pueden diferir de los intereses actuales. En este sentido, cabe tener en cuenta que no todos los campos de conocimiento son afectados de igual manera por su antigüedad. En el caso de que un campo disciplinar solicite un conocimiento actualizado, este factor debe definir cuál es la tendencia en la pérdida de conocimiento teniendo en cuenta el tiempo de desactualización. Si bien no se puede captar en detalle la especificidad de cada caso, esto refleja solo la tendencia. Asimismo, se deben considerar los diferentes criterios editoriales respecto a lo que considera un conocimiento actualizado.

Factor aporte de los autores (orden)

Este factor de la presente metodología no mencionado en la bibliografía relevada, busca poder conocer cuál es el aporte de cada uno de los autores de un artículo científico. El mismo está asociado al orden en que aparecen los autores en la publicación y resulta difícil de determinar porque puede haber diferentes lógicas de ordenamiento en las diversas culturas disciplinares. En el caso que se conozca esa tendencia general de priorizar los aportes en el orden autorial en la disciplina en cuestión, su utilización se estima adecuada.

Factor conflicto de intereses

El factor conflicto de intereses se considera imprescindible en este tipo de metodologías por la posible afectación que puede tener en el resultado de una revisión. Para realizar este proceso de detección de este factor, se requiere la previa desambiguación de nombres para detectar todos los posibles nombres correspondientes a un mismo individuo o diferenciar investigadores con el mismo nombre que influyen negativamente en su detección.

Del relevamiento del mapeo sistemático y del análisis de la encuesta se observa que lo más significativo es la relación de coautoría y la misma afiliación. Luego le siguen relaciones familiares, estudiante-supervisor, amistades, etc. Cabe mencionar que acceder de forma

automática a la información de relaciones de coautoría y de afiliación no resulta problemático, caso contrario se presenta en las otras aplicando este tipo de procedimiento.

2.6.2 Globalizador de factores: Autoridad académica del Investigador

El prestigio del investigador es mencionado como Autoridad académica, por lo cual se entiende que aquellos investigadores con mayor prestigio tienen una ventaja relativa con respecto a los otros. Existen diferentes tipos de información de base para calcular este factor. Dada esta situación, sumado a la diferencia conceptual con los otros factores, se incluye por separado como globalizador. Para su cálculo se utiliza lo siguiente:

- *Índices bibliométricos*: miden la actividad científica y tienen en cuenta el impacto de la producción científica del investigador, efectuando cálculos en base a las citas de sus publicaciones.
- *Títulos*: Las titulaciones académicas refieren a la más alta alcanzada por el candidato a revisor.

Tal como fuera explicado en el capítulo anterior estas opciones colaboran a la hora de definir ventajas comparativas y restricciones. Un punto importante a tener en cuenta es de donde obtener dicha información para incorporar al sistema. Todo lo cual requiere un estudio previo de las posibilidades de obtener estos datos en contenedores de información científica.

2.6.3 Globalizador de factores: Revisiones anteriores

Al igual que el ítem precedente, revisiones anteriores refiere a un solo factor, pero la multiplicidad de datos que lo influyen y su marcada diferencia conceptual marcada con los otros amerita un globalizador de factores por separado.

Revisiones anteriores tiene en cuenta la diversidad de posibilidades de este tipo de problemática y se definen ventajas comparativas a los fines de lograr una correcta distribución de la carga de trabajo entre una significativa cantidad de posibles revisores. Se trata de evitar una exigencia mayor a los investigadores más calificados balanceando así el esfuerzo. A continuación, se detallan los eventos requeridos a registrar en la metodología propuesta:

- *Solicitudes de revisiones y rechazos*: se debe registrar con fecha la aceptación o rechazo de una solicitud de revisión.

- *Cantidad de Rechazos:* al sobrepasarse una cierta cantidad de solicitudes rechazadas, se puede definir una penalización parcial o total. Es posible establecer diferentes rangos de tiempo asociados a diferentes penalizaciones.
- *Cantidad de revisiones simultáneas:* es necesario considerar una cantidad de revisiones simultáneas a solicitar a quien realiza la evaluación tomando el plazo comprendido entre el período de solicitud y su entrega. Los pedidos simultáneos solo podrían tener un carácter excepcional ante una singular circunstancia.
- *Tiempo de descanso entre solicitudes de revisiones:* se puede definir un tiempo en el que no se soliciten revisiones a una misma persona, luego de la entrega de su último informe de revisión.
- *Cantidad de revisiones por período:* se puede considerar una cantidad máxima de solicitudes de revisión a una misma persona limitada a un período determinado.
- *Calificación de las revisiones:* esta tarea relevante para el equipo editorial puede tener diversos indicadores, como por ejemplo la consistencia de la revisión, su grado de completitud, entrega dentro de los tiempos estipulados, entre otros. La metodología permite incluir anotaciones y aclaraciones en observaciones para mayores detalles. Para poder realizar la calificación, se debe hacer una registración en base a una parametrización definida para la evaluación.
- *Malas calificaciones:* En caso de sobrepasar una cierta cantidad de revisiones con calificaciones deficientes se puede definir un tiempo de penalización total o parcial, pudiéndose establecer rangos de tiempo diferenciados asociados al tipo de penalización.

Vale mencionar que las penalizaciones, calculables por período de tiempo o de exclusión total, son simples opciones de la metodología y no se definen en un contexto negativo, sino con el único fin de mejorar el proceso de selección de revisores. Estas penalizaciones pueden aplicarse tanto por rechazos como por calificaciones deficientes.

2.6.4 Definición de flexibilidad de la fórmula integral

La metodología intenta ser lo más flexible dentro de cada uno de los factores, en las consideraciones para el cálculo y también en la disponibilidad de diferentes parámetros que aumenten esa adaptabilidad buscada. No solo en valores, sino que también en operadores a ser utilizados. Además de la posibilidad de definir cuáles son los factores que se van a utilizar y qué opciones internamente se van a implementar.

La variabilidad de la fórmula integral es total, lo que permite una mayor adaptabilidad a la mayor cantidad de equipos editoriales posible.

También incluye la posibilidad de poder definir diferentes perfiles de configuraciones para poder adaptar la búsqueda a las posibilidades reales de conseguir revisores pudiendo, por ejemplo, iniciar con una configuración más estricta y llegar a configuraciones menos rigurosas hasta conseguir los revisores deseados. Este punto, como ya se anticipó, ha surgido del análisis del instrumento encuesta.

3 Construcción metodológica integral

En este capítulo se describe la metodología desarrollando las definiciones metodológicas asociadas a la producción científica, la autoridad académica y las revisiones anteriores, incluyendo los procesos, tablas, variables metodológicas y parámetros numéricos y textuales. También se presenta el armado de la fórmula de integración final de múltiples factores con inclusión de parámetros que se denominaron “integrales”, junto a la utilización de múltiples perfiles.

En principio se aborda una temática que se considera importante ya que trata la adquisición y preprocesamiento de datos. Si bien en el capítulo anterior se han realizado algunas anticipaciones al respecto, se presentan con mayores detalles en el siguiente apartado.

3.1 Adquisición, preprocesamiento y guardado de datos

En esta etapa, se identifican, obtienen, preparan y guardan los datos necesarios para su posterior análisis, asegurando su calidad y adecuación mediante técnicas de limpieza y transformación.

3.1.1 Identificación y elección de fuentes de datos

Se deben definir las fuentes de datos de donde extraer los artículos científicos y los autores de estos, además de los títulos, los resúmenes y las palabras claves entre otros datos relevantes como paso inicial del procesamiento de esta etapa. No todos los textos completos de los artículos científicos se encuentran en acceso abierto, por lo cual en algunos casos solo se trabajará con los datos informados.

Las fuentes de datos seleccionadas deben posibilitar la realización de los procesos específicos para la extracción primaria que permitan luego las transformaciones necesarias para el procesamiento adecuado que permita su utilización.

Para el globalizador de conocimiento científico se necesita contar como mínimo con los siguientes datos:

- Título
- Autores
- Resumen
- Palabras claves
- Fecha de la publicación
- Texto del artículo (en los casos que esté disponible)

Para el globalizador de factores de autoridad académica se necesita contar con los siguientes datos:

- Indicadores bibliométricos
- Titulaciones

Para el globalizador de factores de revisiones anteriores se necesita contar con los datos de revisiones anteriores. El contenedor de estos datos y la carga por parte del equipo editorial está definido en la metodología en el apartado correspondiente, por lo tanto, no es necesario detallarlo en esta etapa.

Con respecto al conflicto de intereses se necesita contar con:

- Información de donde extraer los artículos científicos y los autores de estos (información obtenida previamente)
- Información de afiliación de los investigadores.

También se necesita contar con una base de perfiles de investigadores actualizada para los casos que se necesite desambiguar autores.

3.1.2 Extracción de datos de las fuentes de información

Una vez elegidas las fuentes de datos, se procede a la extracción de información. En este proceso se automatiza la extracción de datos, y pueden definirse tres alternativas válidas:

- *Descarga de contenedor de datos:* si se cuenta con la información en un formato específico de guardado de datos, se descarga y se arman los procesos de extracción en base a la estructura.
- *Conexión por medio de API:* se conecta a la Application Programming Interface y en base a la estructura de los datos de intercambio se procede a armar los diferentes procedimientos.
- *Web scraping:* con esta técnica se extrae información de sitios web y en base a la estructura de las páginas web donde están publicados los datos, se arman los procesos de extracción de estos.

En esta etapa dependiendo de las decisiones de implementación se pueden tomar dos decisiones principales, a saber:

- En línea: directamente pasar los datos al procesamiento sin su guardado en esta etapa.
- Guardado intermedio de la información: pudiendo realizarse un guardado primario [en crudo] para un posterior preprocesamiento.

3.1.3 Preprocesamiento y guardado de datos

En esta etapa se realizan todos los pasos necesarios para depurar los datos de manera que puedan ser utilizados en las subsiguientes. Un preprocesamiento calificado procura mejores resultados en las etapas posteriores. Cabe tener en cuenta que en el caso de las publicaciones científicas por sus características hay pasos de preprocesamiento de textos que resultan innecesarios, las excepciones al respecto son de muy baja incidencia.

El preprocesamiento se debe ajustar a una serie de procesos necesarios según el caso, ya que no se releva un modelo universal. Los definidos en la presente metodología son los siguientes:

Control de datos extraídos: la extracción debe incluir un proceso de control que garantice la calidad de la información que conformará la base de conocimiento. Durante la importación de los datos hacia dicha base, se deben realizar las siguientes verificaciones:

- Verificación de estructura: Comprobar que la estructura de los datos extraídos se ajusta al formato esperado o definido previamente.
- Validación de la extracción: Confirmar que los datos han sido extraídos de manera completa y correcta, sin pérdida ni distorsión de información.
- Análisis de discrepancias: Evaluar si existe alguna diferencia entre los datos extraídos y los datos que se esperaba obtener.
- Control estadístico inicial: Cuantificar los datos extraídos y realizar una clasificación preliminar. Esta información debe formar parte de una estadística inicial que sirva como base para el control y como evidencia del estado inicial del proceso.

Control, depuración y traducción de datos caracterizados y estructurados: se procesa sobre los datos obtenidos de títulos de artículos, resúmenes y palabras claves y se propone la siguiente secuencia:

- Control de corrección de los datos: en base a las estadísticas anteriores y chequeos internos de los datos incorporados como duplicaciones y datos vacíos. En caso de encontrar errores, registrarlos.
- Control de idioma de los datos: en base a la determinación del idioma sobre el que se va a trabajar, se procede al control y detección de idiomas en el que están los diferentes datos, si se cuenta con la información del idioma en que figuran los datos utilizarlo contra la detección para reconocer diferencias. Se debe tener un registro de los datos controlados. El idioma base es un parámetro de la metodología que se denomina "Parámetro de Idioma P_i ".

- Depuración de datos: En base a los controles anteriores y al detectarse diferencias se debe evaluar si se van a corregir los datos erróneos en caso de ser posible y los restantes se eliminarán de la base de conocimiento o si se eliminan todas las diferencias directamente.
- Traducción de los datos: de todos aquellos que no estén en el idioma base.

Determinación, conversión de formatos del texto completo del artículo, control y depuración: se debe identificar cuál es el formato en el que vienen los artículos de investigación completos y migrar a archivo de texto plano en el caso de ser necesario. Generalmente los archivos que contienen los artículos científicos están en PDF, por lo cual lo más común es migrarlos a archivos con formato TXT. En este proceso debe contar con un control de errores en la conversión y registración para determinar los pasos a seguir. A continuación, se lista dicho proceso:

- Determinación y conversión de formatos: si los datos no están en texto plano se debe elegir y realizar un proceso de conversión para llevarlos a dicho formato.
- Registro de errores en el proceso de conversión de formato: es un control indispensable si se realizó la conversión de formato.
- Determinación de datos no válidos: registro de control para la verificación de la validez de los datos resultantes luego del proceso de conversión.
- Depuración de datos: En base a los controles anteriores y al detectarse diferencias se debe evaluar que se aplica a esto. En síntesis, se debe definir:
 - limpieza de la base de conocimiento
 - estudio de los casos para determinar otras maneras de extracción más adecuada para poder contar con dichos datos.

Determinación de las partes representativas y traducción de los textos de los artículos completos: en este proceso se define un orden de tareas de limpieza y conversión buscando la mayor eficiencia posible, por lo cual en esta metodología los siguiente no es representativo⁹:

- Bibliografía: este apartado debe borrarse porque no produce información válida para el vector correspondiente.
- Parte inicial del artículo: una vez identificada la parte previa al cuerpo del artículo se quitan los títulos, autores, resúmenes y palabras claves.

⁹ Este proceso de selección se explicará en detalle en 4.2.1 que incluye la generación de vectores sobre técnicas de PLN

- Formato impreso del artículo: se quita toda la información de formato impreso como pies de página, números de hojas, saltos de líneas por estructura de diseño, además de los guiones de división de palabras en el final de línea y efectuar su reunificación.
- Caracteres especiales u otra información: se deben sacar correos electrónicos, páginas web, también otras informaciones que no puedan tener significación para el vector de contenido.
- Traducción de los textos: definir cuál es el idioma base y traducir todos los otros textos a dicho idioma.

Guardado de la información: se debe definir cómo y dónde se guardan los datos depurados en la “Base de conocimiento” de la metodología que va a servir de insumo para las siguientes etapas. Estas decisiones se asocian más a cuestiones de implementación.

En la gráfica de la figura 4 se pueden observar los procesos de esta etapa de la metodología: los parámetros intervinientes, los insumos y los resultados correspondientes.

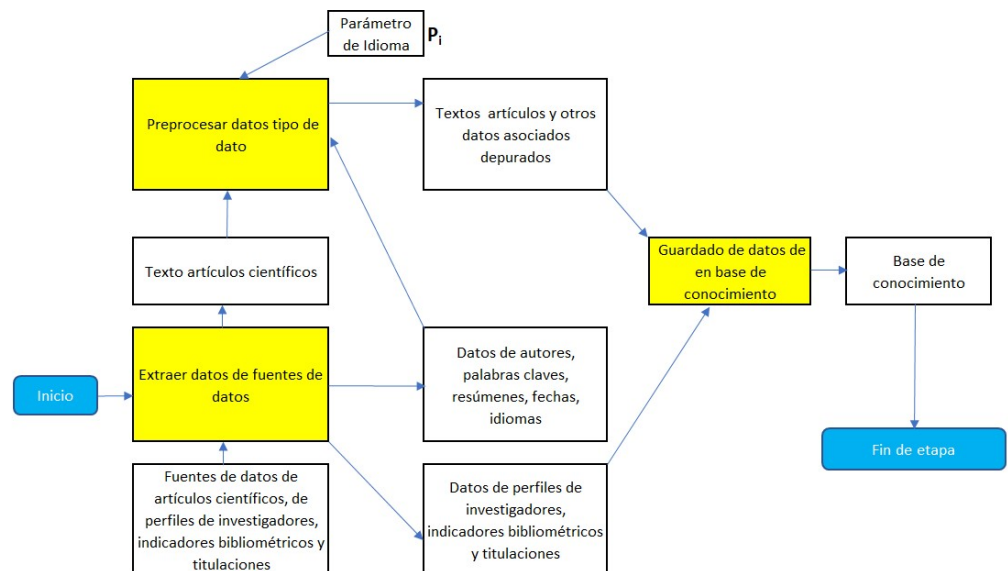


Figura 4

Diagrama de preprocesamiento y guardado de datos

3.2 Producción científica

La producción científica como un globalizador de factores es considerada para la metodología como ya se expuso, en asociación al factor experiencia y es afectada por los

factores de interés del investigador y aporte de cada autor con aplicación de la restricción de control del conflicto de intereses.

3.2.1 Cálculo de Experiencia

El objetivo de esta etapa es extraer los datos necesarios de acuerdo a las partes de texto representativas de los artículos científicos, tanto del manuscrito como de las producciones de los posibles revisores. Se busca generar una representación de características adecuada que contenga el conocimiento de cada artículo científico para luego compararlos con el manuscrito y así obtener un valor numérico representativo de la similitud entre cada artículo científico y el manuscrito. Posteriormente se los agrupa y se procede a desambiguar los nombres de los autores para realizar cálculo de su experiencia sobre el índice de similitud de los documentos. A continuación, se desarrolla en detalle este proceso.

Parte representativa del artículo científico

Resulta primordial definir los datos que van representar el artículo científico. Este dato se guarda en el **Parámetro de parte representativa del artículo científico: PPR_{ac}**

En el capítulo anterior, se explicitó la necesidad de evaluar las opciones dentro del prototipo de prueba de la metodología, quedando la opción por defecto abierta para su modificación por parte de los editores como también otras configuraciones técnicas.

Las partes a evaluar son:

- Títulos
- Palabras claves
- Resúmenes
- Resúmenes y títulos
- Resúmenes, títulos y palabras claves
- Contenido del artículo sin títulos, resúmenes, palabras claves y bibliografía.

Los datos de la parte representativa del artículo científico son extraídos de la base de conocimiento, al igual que los representativos de las publicaciones de los candidatos y de la carga del manuscrito a revisar.

Generación de vectores sobre técnicas de PLN

La metodología debe permitir la elección de varias técnicas de PLN para la generación de vectores representativos y definir una técnica por defecto que se guarda en el **Parámetro de técnica de PLN: $P_{t_{pln}}$**

En esta etapa se le aplica la técnica elegida de PLN contenida en el parámetro de técnica PLN y se obtiene una representación vectorial que pueda captar el conocimiento que trata cada artículo científico, sea del manuscrito o de la base de conocimiento de los candidatos a revisores.

A continuación, se procede a detallar los tipos de modelos que pueden ser útiles para encontrar las representaciones buscadas y las técnicas más comunes:

Se proponen técnicas que se basen en 3 tipos de modelos de PLN:

Modelos de vectores dispersos (*sparse vectors models*) tienen representaciones de las palabras que forman parte de los textos, pero no incluyen ninguna información semántica. Estos modelos tampoco tienen en cuenta el orden de las palabras, por eso se les denomina de bolsa de palabras (*bag of words*).

Los modelos utilizados para el recupero de información son:

- TF-IDF: *Term frequency – Inverse document frequency* (frecuencia de término – frecuencia inversa de documento). TF refleja la frecuencia en que aparece un término en un documento. IDF determina con qué frecuencia aparece en los otros documentos. A mayor frecuencia menor importancia del término, disminuyendo su peso y aumentando el peso de los que aparecen poco. De esta manera determina cuales son los términos más relevantes del documento.
- BM25: es otra técnica dentro de los modelos de espacio vectorial que en relación a TF-IDF agrega mayor sensibilidad a la frecuencia de términos y a la longitud del documento.

Modelos *Word embeddings*: a diferencia del modelo anterior, las palabras son representadas por vectores densos que incluyen información semántica. Los vectores están formados por números reales y se entrenan sobre la base de un corpus de texto sin etiquetar dando por resultado los vectores densos. En este modelo las palabras de significación semántica parecidas se encuentran cercanas en el espacio vectorial. Los modelos de este tipo más destacados son:

- Word2Vec (Mikolov et al., 2013): creado por Google, su entrenamiento está basado en ventanas de palabras alrededor de una palabra central deslizable por todo el corpus. Existen dos tipos de arquitecturas: CBOW (*Continuous Bag of Words*), que

predice la palabra central en base a las palabras cercanas hacia atrás y hacia delante y, *Continuous Skip-Gram Model* que a partir de una palabra central trata de predecir las palabras cercanas hacia atrás y hacia delante.

- GloVe(Pennington et al., 2014), creado por la Universidad de Stanford, su nombre deriva de Global Vectors. A diferencia de Word2Vec que se basa en las palabras del contexto de una palabra dada, usa estadísticas globales de todo el corpus de entrenamiento.
- FastText (Bojanowski et al., 2017), creado por Facebook con parte del equipo que creó Word2Vec, se diferencia de este porque el vector de la palabra es formado por los vectores de las subpalabras que conforman las palabras y las denomina *n-gramas*, permitiendo de este modo formar vectores con sentido de palabras que no hayan estado presentes en el entrenamiento.

Modelos Document Embeddings: Su objetivo es capturar el significado de un documento completo con representación semántica, permitiendo comparaciones y análisis más allá del nivel de palabras individuales. En esta clasificación se consideran únicamente modelos previos a la aparición de los Transformers, los cuales serán tratados por separado debido a sus características distintivas que justifican su análisis independiente.

- Doc2vec(Le & Mikolov, 2014) es una extensión del modelo Word2Vec para generar vectores de documentos. A diferencia de Word2Vec se entrena directamente con los documentos a comparar, por lo que no requiere un modelo preentrenado para generar los vectores. Esto permite su aplicación en distintos idiomas sin necesidad de ajustes previos.

Aunque existen otros modelos para la generación de embeddings de documentos, Doc2Vec es uno de los más conocidos y utilizados debido a su capacidad de entrenarse directamente sobre los datos sin depender del idioma. Por esta razón, se ha seleccionado como el único modelo en este trabajo.

Dentro del proceso de trabajo de los tres tipos de modelos mencionados, es común hacer preprocesamientos en los textos:

- El primer proceso es quitar las denominadas *stop words* o palabras vacías que no tienen ningún significado determinante sobre los textos y se utilizan en la construcción de las frases.
- El segundo proceso consiste en simplificar la cantidad de palabras y así aumentar la frecuencia, tratando de encontrar mayor asociación y frecuencia de palabras que eran diferentes pero que guardan algún tipo de relación. Para esto se puede utilizar alguna de las siguientes técnicas:

- Stemming: es un proceso sencillo y de fácil realización por librerías de PLN por el cual se extrae la raíz de una palabra.
- Lematización: en este proceso se busca el lema de base de las palabras utilizadas, requiere un proceso más complejo que el anterior y solicita tener los lemas de todas las palabras y las diferentes variaciones que puede tener la misma.

Entre las dos técnicas se opta por la segunda porque permite mantener la base de las palabras utilizadas y en las metodologías que generen vectores por palabra con significación semántica posibilita generarlas adecuadamente¹⁰.

Modelos Transformers: desde lo publicado por Vaswani et al. (2017), resulta el estado del arte en una gran cantidad de tareas del PLN y para el desarrollo de la IA en general. A diferencia de Word embeddings que generaban un vector por palabra y que no podían resolver el problema de los homónimos, los Transformers permiten la generación de vectores contextualizados. Dos técnicas representativas de estos modelos son BERT (Devlin et al., 2019) y GPT (Radford & Narasimhan, 2018).

- BERT (*Bidirectional Encoder Representations from Transformers*) es generado por Google. Tiene la capacidad para captar el contexto y la relación de las palabras por su enfoque bidireccional, debido a que puede procesar la información de izquierda a derecha y de derecha a izquierda.
- GPT (*Generative Pre-trained Transformer*) es un modelo de lenguaje desarrollado por OpenAI. A diferencia de BERT, está diseñado para generar texto completo y coherente de manera autónoma de izquierda a derecha. No es bidireccional en el sentido de BERT, pero su estructura de atención le permite considerar el contexto tanto a la izquierda como a la derecha de cada palabra durante el entrenamiento, lo que colabora en la captura de relaciones complejas y contextuales en el texto.

Cálculo de similitud y ordenamiento

¹⁰ Otro tipo de modelo que podría haberse utilizado es “Sentence Embeddings” que generan representaciones de oraciones. Se ha optado por usar tanto Word Embeddings como Document Embeddings porque los Word Embeddings capturan el significado semántico de las palabras y permiten construir representaciones precisas de los documentos, mientras que los Document Embeddings son más adecuados para capturar la semántica global del texto completo. Aunque los Sentence Embeddings pueden ser útiles para comparar oraciones, no son tan comunes para la comparación de documentos completos, ya que se enfocan en fragmentos más pequeños y no capturan satisfactoriamente las relaciones a nivel de documento. Por todo lo cual Word Embeddings y Document Embeddings son técnicas más utilizadas en este contexto, tomando en cuenta que es una etapa anterior a Transformers en donde no se incluye información de contexto en las representaciones, tema que cambiaría mucho la consideración. Por lo cual, los modelos previamente mencionados resultan suficientes para la tarea a realizar en esta etapa de evolución de la tecnología.

Esta parte está asociada a la comparación entre el vector del manuscrito y todos los vectores de los artículos de la base de conocimiento para encontrar el valor de similitud. Para ello se usa una técnica de similitud que está contenida en el **Parámetro de técnica de similitud: (P_{ts})**. Este parámetro va a contener una sola técnica, la “Similitud de coseno” que ha funcionado satisfactoriamente en diferentes casos (Zahrotun, 2016; Thada & Jaglan, 2013; D’hondt et al., 2010; Li, 2013). Si bien va a tener un solo valor, se usa cómo parámetro para no obturar a futuro la posibilidad de ingresar otros valores. Una vez aplicada la técnica de similitud, se deben ordenar todos los artículos de forma descendente por su valor de similitud. Las ventajas que justifican la elección de la mencionada técnica son las siguientes:

- ✓ Independencia de la magnitud: la similitud de coseno mide la orientación de los vectores y no su magnitud, lo que permite comparaciones precisas independientemente del tamaño de los documentos.
- ✓ Rendimiento en espacios de alta dimensionalidad: funciona correctamente en espacios vectoriales de alta dimensionalidad como los vectores de palabras o documentos.
- ✓ Rendimiento con vectores escasos y densos: es adecuada tanto para vectores escasos (con gran cantidad de “0”) como para vectores densos que contienen información semántica.
- ✓ Eficiencia computacional: es una técnica computacionalmente eficiente, lo que permite manejar grandes volúmenes de datos de manera efectiva.

Aplicación de punto de corte

Luego del paso anterior se debe hacer un corte para seleccionar los más cercanos en base a un parámetro que se ha denominado **Parámetro de punto de corte: P_{pc}**

Este valor debe ser introducido al momento de realizar una búsqueda de revisores y va a representar la cantidad de artículos priorizados que se va a seleccionar en la metodología.

Desambiguación de nombres

Como ya fue explicitado, pueden surgir situaciones de ambigüedad en los nombres de los autores. En la propuesta metodológica si solo se trabaja sobre una base de conocimientos correctamente sistematizada, con los datos de autores codificada y artículos correctamente asociados no sería necesario correr este proceso. Sin embargo, si se utilizan fuentes de información con datos de autores no sistematizados, que solo contienen los nombres de los

misimos en los artículos, se plantea esta problemática primordial y, en consecuencia, se propone un método de contrastación.

El proceso debe iniciarse aplicándolo a todos los autores y artículos publicados en la base de conocimiento principal, con el objetivo de establecer asociaciones con los perfiles de los investigadores en una segunda base de conocimiento verificada, basándose en la coincidencia de artículos escritos por los mismos autores en ambas bases. Una vez realizada esta asociación, se lleva a cabo el proceso de desambiguación. Este enfoque proporciona el mayor nivel de confianza en los resultados.

En una segunda instancia, y con un nivel de confianza menor, si no es posible desambiguar el nombre del investigador mediante la verificación de artículos científicos publicados, se plantean búsquedas alternativas. Un ejemplo de estas es la búsqueda de información sobre coautorías en las bases de perfiles seleccionadas para la desambiguación. Como se mencionó anteriormente, el éxito de este proceso dependerá de la información disponible en la práctica y de cómo esté estructurada.

En síntesis, la finalidad buscada es desambiguar la mayor cantidad de nombres de investigadores, dado que la totalidad sería lo óptimo.

Definición de tipo de umbral de conocimiento a utilizar

Según lo ya expuesto, el umbral de conocimiento es la consideración del conocimiento específico que tratan los artículos científicos. Se toma como premisa que los artículos más cercanos entre sí son los que tratan las mismas temáticas y se calcula la ventaja relativa que tiene un candidato sobre los otros cuando tiene más de un artículo cercano al artículo a revisar. La idea principal es asociarlo a un número de artículos a ser considerados para los cálculos, siendo el umbral el valor representativo del artículo más alejado de ese conjunto. La metodología también permite poner ese valor de umbral de forma manual según criterio del equipo editorial otorgando mayor flexibilidad al modelo. El valor obtenido en el cálculo del umbral se denomina “Valor de umbral de conocimiento”.

Los parámetros asociados a umbral de conocimiento son los siguientes:

- **Parámetro de cálculo de umbral de la similitud:** PC_{us} . define cómo se realizará el cálculo del umbral, a saber:

- Automático: se toma de umbral el menor valor de similitud del documento de los artículos seleccionados, para ello verifica si está activado el cálculo de antigüedad de los artículos (interés del investigador) para considerarlo en este procedimiento.
- manual: se ingresa manualmente el valor de umbral para los cálculos

- **Parámetro de cálculo manual de umbral de similitud: PCM_{us}** , solo se utiliza en el caso de ingresar en Tipo la opción manual. El valor a ingresar debe ser mayor o igual a 0 y menor o igual a 1.

Cálculo de Puntaje de experiencia genérica

La experiencia genérica de un candidato a revisor es la resultante del cálculo de su experiencia sobre el índice de similitud de los documentos con aplicación del umbral de conocimiento en todos los documentos priorizados del mismo candidato. Y se refleja en la variable metodológica **Puntaje de experiencia genérica: P_{eg}**

Hay diferentes opciones de cómo configurar el umbral para calcular el puntaje de experiencia genérica asociada directamente a los textos. Todos los cálculos se hacen luego de generar los vectores representativos del documento y compararlos entre sí. El primer dato a tener en cuenta tiene las siguientes opciones:

- Cálculo automático asociado a umbral (PC_{us} = “automática”): puede basarse en dos tipos de cálculos teniendo en cuenta los factores asociados activos. Las opciones varían si toma el valor de experiencia con o sin antigüedad. En los dos casos está asociado a los factores que tome de base y no incluye en el cálculo el factor de orden, por el cual, si este factor estuviera activado y el valor calculado fuera menor al del umbral, no se tomaría ese dato. En el cálculo automático el valor de umbral a tomar siempre es mayor a “0”.
- Cálculo manual asociado umbral (PC_{us} = “manual”): se debe ingresar manualmente el valor de PCM_{us} entre el rango de “1” (uno) mayor proximidad posible al documento (el artículo del candidato a revisor es igual al artículo a revisar) y “0” (cero) que es el menor valor posible. En el caso de colocar “0” se desactiva la lógica del uso de umbral y se sumaría el valor comparativo total de todos los documentos de un mismo revisor.

Para el cálculo también se define el **valor residual** que es el valor que se va a considerar del documento para el cálculo de la Experiencia Genérica. En primer lugar, se calcula este valor en los documentos del mismo autor:

- Valor residual del 1er.documento de un candidato a revisor en orden de proximidad al documento a revisar es:
 - o $Vr_{doc(1)} = V. similitud_{doc(1)}$

- Valor residual individual de los documentos del mismo autor, desde el segundo en adelante es:

$$\bigcirc \quad Vr_{doc(i)} = V. similitud_{doc(i)} - V. umbral \text{ para } i > 1$$

Una vez que se tiene el valor residual se utiliza el **Parámetro de tipo de cálculo de similitud: PTC_s** que define cómo se va a realizar el cálculo entre los documentos de un mismo autor para obtener la Experiencia Genérica. Y los valores posibles son:

- suma_simple
- promedio_total
- promedio_ponderado

En base a este parámetro se presentan las fórmulas de las tres opciones para el cálculo de la Experiencia Genérica:

- Suma simple (si **PTC_s**= “suma_simple”): se suman directamente todos los valores residuales de los documentos.

$$\text{Suma simple} \quad SS = \sum_{i=1}^n Vr_{doc(i)}$$

$$P_{eg} = SS$$

- Promedio total (si **PTC_s**= “promedio_total”): se suman todos los valores acumulativos residuales de cada valor desde el inicio y se divide por la cantidad de valores (promedio de suma anidada).

$$\text{Promedio total} \quad PT = \frac{\sum_{i=1}^n (\sum_{x=1}^i Vr_{doc(x)})}{n}$$

$$P_{eg} = PT$$

- Promedio ponderado (si **PTC_s**= “promedio_ponderado”): se genera una función recursiva que inicia con el primer valor definido como el primer valor residual y luego se aplica la función recursiva, tomando dos veces el valor anterior de la función, sumándole el valor residual correspondiente y dividiendo todo por 2. Esta función produce una leve moderación o suavizado en el resultado de la función anterior. Se logra haciendo el promedio cada vez que toma un nuevo valor y no al final como en la función del promedio total (promedio ponderado recursivo).

$$\text{Promedio ponderado} \quad \begin{cases} PP_{(1)} = Vr_{doc(1)} \\ PP_{(i)} = \frac{(2 \cdot PP_{(i-1)}) + Vr_{doc(i)}}{2} \text{ para } i > 1 \end{cases}$$

$$P_{eg} = PP$$

3.2.2 Aplicación de Interés del investigador (Antigüedad)

En el interés del investigador se considera que los últimos artículos publicados revelan los temas en los que está interesado en la actualidad y se asigna mayor peso sobre los otros. Este es un factor que puede ser considerado en esta metodología y se ve reflejado en el valor del **Parámetro de antigüedad: P_a**

El valor de pérdida de representatividad del conocimiento por desactualización puede ser definido por el equipo editorial utilizando una lógica basada en rango de años de antigüedad y valor de representatividad del artículo según la **tabla de cálculo de antigüedad T_{ca}** . Cada rango se define por los años de antigüedad, establecidos mediante un año de inicio y un año de fin, y se asocia a un valor de representatividad (valor de P_a) con valores comprendidos entre el rango 0 - 1, siendo 1 la mayor representatividad total conservada y 0 no poseer ningún valor de representatividad por estar completamente desactualizado.

Para la inclusión o no del cálculo se usa el siguiente parámetro: **Parámetro de cálculo de antigüedad: PC_a**

Los valores posibles son: S, que habilita el cálculo de antigüedad, y N, que lo deshabilita y asume un valor de representatividad igual a 1.

Procedimiento:

En primer lugar, se determina si corresponde aplicar el cálculo de antigüedad utilizando el parámetro PC_a :

Si $PC_a = "N"$, entonces se asigna directamente $P_a = 1$.

Si $PC_a = "S"$, se habilita el procedimiento de cálculo de P_a para cada uno de los artículos correspondientes a los candidatos a revisor priorizados y se realiza el siguiente paso:

Determinación de P_a :

a. Obtener el valor de Antigüedad

$$\text{Antigüedad} = \text{Fecha actual} - \text{Fecha publicación}$$

b. Extraer P_a en base al valor calculado de Antigüedad dentro del rango ra de la tabla T_{ca}

Formalmente, si T_{ca} es una tabla donde $T_{ca} = \{(ra_{(1)}, P_{a(1)}), (ra_{(2)}, P_{a(2)}), \dots, (ra_{(n)}, P_{a(n)})\}$ donde cada $ra(i)$ representa un rango de años determinado por un año de inicio y un

año de fin según el registro i , y $P_{a(i)}$ es el valor de representatividad asociado, entonces se selecciona el $P_{a(i)}$ correspondiente al rango en el que se encuentra la Antigüedad.

$$P_a = \begin{cases} P_{a(1)} & \text{si Antigüedad} \in ra_{(1)} \\ P_{a(2)} & \text{si Antigüedad} \in ra_{(2)} \\ \vdots & \vdots \\ P_{a(n)} & \text{si Antigüedad} \in ra_{(n)} \end{cases}$$

3.2.3 Aplicación de Aporte de los autores (Orden)

El aporte de los autores asociado al orden de aparición de los nombres de los candidatos a revisores se ve reflejado en esta metodología en el **Parámetro de aporte de los autores: P_{au}** y su valor puede ser considerado por el equipo editorial en una lógica basada en rango de posiciones, y el valor de representatividad del aporte del investigador según la **tabla de cálculo de aporte de los autores T_{cau}** . Cada rango se define por el orden de los autores, establecidos mediante una posición de inicio y una de fin, y se asocia a un valor de representatividad (valor de P_{au}) con valores comprendidos entre el rango 0 – 1, siendo 1 la mayor representatividad total conservada y 0 no poseer ningún valor de representatividad por la posición que ocupa.

Para la inclusión o no del cálculo se usa el siguiente parámetro: Parámetro de cálculo del aporte de los autores: PC_{au}

Los valores posibles son: S, que habilita el cálculo de aporte, y N, que lo deshabilita y asume un valor de representatividad igual a 1.

Procedimiento:

Se determina si corresponde aplicar el cálculo de aporte de los autores utilizando el parámetro PC_{au} :

Si $PC_{au} = "N"$, entonces se asigna directamente $P_{au} = 1$.

Si $PC_{au} = "S"$, se habilita el siguiente paso para cada uno de los artículos correspondientes a los candidatos a revisor priorizados.

Determinación de P_{au} :

a. Obtener el orden en el que aparece el autor:

Orden = Posición del autor en la lista de autores

b. Extraer P_{au} en base al valor de Orden dentro del rango ro de la tabla T_{ca}

Formalmente, si T_{cau} es una tabla donde $T_{cau} = \{(ro_{(1)}, P_{au(1)}), (ro_{(2)}, P_{au(2)}), \dots, (ro_{(n)}, P_{au(n)})\}$ donde cada $ro_{(i)}$ representa un rango de posiciones determinados por posición de inicio y fin según el registro i , y $P_{au(i)}$ es el valor de representatividad asociado, entonces se selecciona $P_{au(i)}$ correspondiente al rango en el que se encuentra la posición.

$$P_{au} = \begin{cases} P_{au(1)} & \text{si } Orden \in ro_{(1)} \\ P_{au(2)} & \text{si } Orden \in ro_{(2)} \\ \vdots & \vdots \\ P_{au(n)} & \text{si } Orden \in ro_{(n)} \end{cases}$$

3.2.4 Control de Conflicto de intereses

Una vez que se procede a la desambiguación de nombres, ya se podría verificar **el conflicto de intereses: Col A** a partir de las consideraciones ya expuestas, se realiza el control de coautoría de cada candidato a revisor con la posibilidad de definirlo en 3 niveles y el control de afiliación. Sobre cada uno de los autores de los artículos priorizados para la búsqueda de revisores, se sigue el siguiente procedimiento:

Procedimiento:

1. Verificación de coautoría

Nivel 1: Verificación de autores directos

- Se busca el candidato a revisor en la lista de autores del manuscrito a revisar:
 - Si es uno de los autores: se marca como conflicto de intereses.
 - Se continúa con la verificación de Nivel 2.

Nivel 2: Verificación de coautoría directa

- Se busca el candidato a revisor en la lista de coautores directos de los autores del manuscrito a revisar.
 - Si aparece como coautor directo: se marca como conflicto de intereses.
 - Se continúa con la verificación de Nivel 3.

Nivel 3: Verificación de coautoría indirecta

- Se busca el candidato a revisor en la lista de coautores de coautores de los autores del manuscrito a revisar.
 - Si aparece como coautor indirecto: se marca como conflicto de intereses.
 - Se continúa con la verificación de afiliación.
- 2. Verificación de afiliación: Para este requerimiento se debe contar con datos de afiliación de los autores del manuscrito a revisar.
 - Verificación de datos de afiliación de candidato a revisor
 - Si no hay datos: no se realiza este control.
 - Si hay datos disponibles: se procede al siguiente paso.
 - Comparar la afiliación del autor del manuscrito con la del candidato a revisor:
 - **Si hay coincidencia de afiliación:** se marca como conflicto de intereses.

Otros Motivos de Conflicto de Intereses: todos los otros posibles motivos que se detallaron en el mapeo sistemático y la encuesta. Se deben identificar estos motivos seleccionados y establecer procesos de carga manual o semiautomático.

Consideraciones Adicionales: Todos los posibles conflictos de intereses identificados deben ser analizados en detalle por el equipo editorial.

3.2.5 Cálculo de Puntaje de producción científica

Esta fórmula se logra con la integración de las variables metodológicas Puntaje de genérica P_{eg} , del Parámetro de antigüedad P_a y del Parámetro de aporte de los autores P_{au} . Se debe tener en cuenta que la integración de estas se hace a nivel de documento, por lo cual se integran directamente al valor residual del documento de los candidatos a revisor. En el caso que el Parámetro de cálculo de antigüedad PC_a no esté habilitado se le asigna el valor 1 y de la misma manera, si el Parámetro de cálculo de aportes de los autores PC_{au} no está habilitada también se le asigna el valor 1. La variable metodológica a utilizar es **Puntaje de producción científica: P_{pc}**

En el caso del cálculo de Suma Simple (si $PTC_s = \text{"suma_simple"}$):

$$\text{Suma simple: } SS = \sum_{i=1}^n Vr_{doc(i)} * P_{a(i)} * P_{au(i)}$$

$$P_{pc} = SS$$

En el caso del cálculo de Promedio Total (si $PTC_s = \text{"promedio_total"}$):

$$\text{Promedio total: } PT = \frac{\sum_{i=1}^n (\sum_{j=1}^n Vr_{doc(x)} * P_{a(i)} * P_{au(i)})}{n}$$

$$P_{pc} = PT$$

En el caso del cálculo de Promedio Ponderado (si $PTC_s = \text{"promedio_total"}$):

$$\text{Promedio ponderado: } \begin{cases} PP_{(1)} = Vr_{doc(1)} * P_{a(1)} * P_{au(1)} \\ PP_{(i)} = \frac{(2 * PP_{(i-1)}) + Vr_{doc(i)} * P_{a(i)} * P_{au(i)}}{2} \end{cases} \text{ para } i > 1$$

$$P_{pc} = PP$$

Por último, en los casos de conflicto de intereses se marcan como posibles conflictos de intereses.

3.2.6 Corte por no alcanzar el Puntaje mínimo de producción científica

Este corte se produce cuando el valor del cálculo resulta menor al del umbral definido, ya sea de manera automática o manual. Es producido porque el Parámetro de antigüedad P_a define en determinado rango con valor "0" y el cálculo termina resultando en dicho valor o porque el cálculo que incluye el Parámetro de aporte de los autores P_{au} al ser considerado arroja un valor menor al umbral de consideración.

Finalmente, la gráfica de la Figura 5 sintetiza los procesos llevados a cabo en esta etapa de la metodología con todos los parámetros intervinientes, los insumos y los resultados en cada paso.

Para ello se utiliza:

Índice bibliométrico

Se debe elegir cuál es el índice que se va a utilizar de base, y se le denomina **Parámetro de tipo de índice bibliométrico: PT_{ib}**

Pudiendo ser, entre otros:

Citas totales: recuento de todas las citas que tiene la totalidad de los artículos científicos del candidato.

Índice h (h-Index): indicador que considera los artículos publicados por un autor y la cantidad de citas por cada artículo, considerando aquellos artículos que tienen al menos la misma cantidad de citas que la cantidad de artículos publicados.

Índice i10 (i10-index): en este caso, el número al lado de la i indica la cantidad de citas que deben tener al menos cada artículo científico a ser considerado, en este caso 10 citas.

Cada una de las opciones anteriores puede ser delimitada para un rango de tiempo, es posible que en algunas disciplinas tenga mayor importancia los índices de los últimos años a la fecha, por ejemplo, poner un valor de 5 años como corte para el cálculo.

El nombre del candidato debe ser coincidente al de la fuente de datos a utilizar o se debe hacer una conversión previa.

Sobre el valor del índice bibliométrico seleccionado del candidato a revisor, el objetivo es aumentar las posibilidades de ser considerado revisor en base al **Parámetro del índice bibliométrico: P_{ib}**

Lo primero a definir en este parámetro es el índice bibliométrico a tomar de base. Luego de elegir P_{ib} , se definen los rangos de valores que corresponde a un valor de representatividad en la **tabla de cálculo bibliométrico T_{cb}** .

Cada rango se define por el valor del índice, establecidos mediante valor de inicio y de fin y se asocia un valor de representatividad (valor de P_{ib}) con valores comprendidos entre el rango (1 a n)

En caso de no poder acceder a la información de índices bibliométricos de algunos candidatos, la propuesta metodológica considera poner un valor de representatividad que indica esta situación.

Para la inclusión o no del cálculo se usa el siguiente parámetro:

Parámetro de cálculo sobre índices bibliométricos: PC_{ib}

Valores posibles:

S – Habilita el cálculo sobre índices Bibliométricos

N – Deshabilita el cálculo sobre índices Bibliométricos

Procedimiento:

Decidir la inclusión del cálculo sobre índices bibliométricos usando PC_{ib} :

Si $PC_{ib} = "N"$, no se habilita el procedimiento de determinación de P_{ib}

Si $PC_{ib} = "S"$, habilita el procedimiento de determinación de P_{ib}

En el caso de estar habilitado, sobre cada uno de los candidatos a revisor priorizados se realiza lo siguiente:

Determinación de P_{ib} :

- Obtener el tipo de indicador bibliométrico: PT_{ib} = Tipo de indicador bibliométrico elegido
- Obtener valor del indicador bibliométrico: $V_{indicador}$ = Valor del indicador bibliométrico del candidato a revisor
- Decidir si corresponde seguir con el procedimiento en base al valor de $V_{indicador}$:

Si $V_{indicador}$ es nulo, se utiliza la variable metodológica **valor por defecto del Parámetro del índice bibliométrico VDP_{ib}** . Por defecto es 1 pero puede modificarse. Entonces, $P_{ib} = VDP_{ib}$

Si $V_{indicador}$ no es nulo, extraer P_{ib} en base al valor del indicador $V_{indicador}$ dentro del rango ri de la tabla T_{cb}

Formalmente, si T_{cb} es una tabla donde $T_{cb} = \{(ri_{(1)}, P_{ib(1)}), (ri_{(2)}, P_{ib(2)}), \dots, (ri_{(n)}, P_{ib(n)})\}$ donde cada $ri_{(i)}$ representa un rango de valores por valor de inicio y fin del indicador según el registro i , y $P_{ib(i)}$ es el valor de representatividad asociado, entonces se selecciona $P_{ib(i)}$ correspondiente al rango en el que se encuentra el valor del indicador.

$$P_{ib} = \begin{cases} P_{ib(1)} & \text{si } v_{indicador} \in ri_{(1)} \\ P_{ib(2)} & \text{si } v_{indicador} \in ri_{(2)} \\ \vdots & \vdots \\ P_{ib(n)} & \text{si } v_{indicador} \in ri_{(n)} \end{cases}$$

3.3.2 Aplicación de titulación

Sobre la titulación académica, se busca aumentar las posibilidades de ser considerado revisor en base al uso del **Parámetro de Título: P_t**

Se utiliza la **tabla de cálculo de título T_α** con registración de títulos y su correspondiente valor de representatividad.

Se definen rangos en base al título de los candidatos, de la siguiente manera:

- Título: descripción de título (según esté consignado en la fuente de datos a utilizar o conversión previa de los mismos)
- Orden: orden jerárquico de importancia de los títulos entre el rango (1 a n) siendo 1 el de menor importancia
- Valor de representatividad (valor de P_t): valores comprendidos entre el rango (0 a n)

Como en el caso anterior, existe la posibilidad de no contar con la información de títulos de algunos candidatos a revisores por lo cual la propuesta metodológica considera poner un valor de representatividad para estos casos que indica que dicha situación.

Para la inclusión o no del cálculo se usa el siguiente parámetro:

Parámetro de cálculo sobre titulaciones: PC_t

Valores posibles:

S – Habilita el cálculo de Titulaciones

N – Deshabilita el cálculo de Titulaciones

Procedimiento:

Decidir la inclusión del cálculo sobre titulaciones usando PC_t :

Si $PC_t = "N"$, no habilita el procedimiento de determinación de P_t

Si $PC_t = "S"$, habilita el procedimiento de determinación de P_t

En el caso de estar habilitado, sobre cada uno de los candidatos a revisor priorizados se realiza lo siguiente:

Determinación de P_t :

- a. Obtener el título del candidato a autor: $V_{\text{título}} = \text{Título del candidato a revisor}$
- b. Decidir si corresponde seguir con el procedimiento en base al valor de $V_{\text{título}}$:

Si $V_{\text{título}}$ es nulo, se utiliza la variable metodológica **valor por defecto del Parámetro de título VDP_t** , que contiene el valor en el caso que sea nulo. Por defecto es 1 pero se puede modificar. Entonces, $P_t = VDP_t$

Si $V_{\text{título}}$ no es nulo, extraer P_t en base al valor de título $V_{\text{título}}$ buscarlo en el campo título t de la tabla T_{ct}

Formalmente, si T_{ct} es una tabla donde $T_{ct} = \{(t_{(1)}, P_{t(1)}), (t_{(2)}, P_{t(2)}), \dots, (t_{(n)}, P_{t(n)})\}$ donde cada $t_{(i)}$ representa un valor de título en base al registro i sobre el campo título y $P_{t(i)}$ es el valor de representatividad asociado, entonces se selecciona $P_{t(i)}$ correspondiente al valor en el que se encuentra el título.

$$P_t = \begin{cases} P_{t(1)} & \text{si } i v_{\text{título}} = t_{(1)} \\ P_{t(2)} & \text{si } i v_{\text{título}} = t_{(2)} \\ \vdots & \vdots \\ P_{t(n)} & \text{si } i v_{\text{título}} = t_{(n)} \end{cases}$$

Aplicación de corte por no alcanzar titulación mínima requerida

Se puede aplicar un corte por titulación. Para hacerlo se define dentro de la tabla de cálculo de título T_{ct} . En cada uno de los títulos que no quieran ser considerados, según el valor del campo de orden, se debe colocar el Valor de representatividad en 0 (cero). De esta manera se activa el corte y se marca en el informe final de la metodología.

3.3.3 Cálculo de Puntaje de autoridad académica

Se ve reflejada en la variable metodológica **Puntaje de autoridad académica: P_{aa}**

Esta variable contiene el resultado final de los cálculos entre los valores obtenidos de los parámetros de índice bibliométrico y de titulación. Se pueden establecer diferentes tipos de cálculos entre estos y para ellos se utiliza el **Parámetro de cálculo entre índices bibliométricos y titulación: PC_{ibt}**

Valores posibles que modifican el cálculo del Puntaje de autoridad total:

Sí Parámetro de cálculo sobre titulaciones: $PC_t = "S"$ y Parámetro de cálculo sobre índices bibliométricos: $PC_{ib} = "S"$

- suma: $PC_{ibt} = P_{ib} + P_t$
- promedio: $PC_{ibt} = (P_{ib} + P_t)/2$
- producto: $PC_{ibt} = P_{ib} * P_t$

Si Parámetro de cálculo sobre titulaciones: $PC_t = "S"$ y Parámetro de cálculo sobre índices bibliométricos $PC_{ib} = "N"$, entonces $PC_{ibt} = P_{ib}$

Si Parámetro de cálculo sobre titulaciones: $PC_t = "N"$ y Parámetro de cálculo sobre índices bibliométricos: $PC_{ib} = "S"$, entonces $PC_{ibt} = P_t$

3.3.4 Gráfica de los procesos de autoridad académica

La Figura 6 muestra todos los procesos de esta etapa de la metodología incluyendo los parámetros intervinientes, los insumos y los resultados en cada paso.

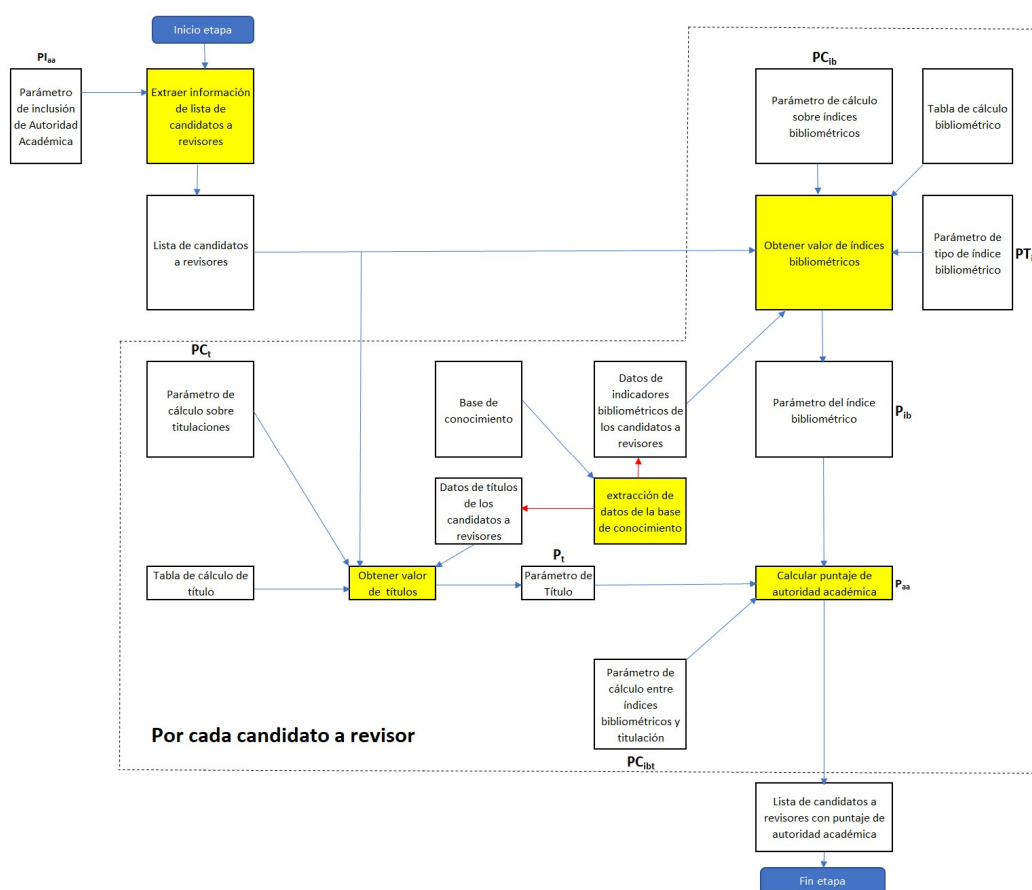


Figura 6

Diagrama del proceso de cálculo de autoridad

3.4 Revisiones anteriores

Este globalizador de la metodología está basado en el historial de revisiones anteriores y lo conforma toda la información de retroalimentación del equipo editorial. Primero se evalúa el **Parámetro de inclusión de revisiones: PI,**

Valores posibles:

S – Habilita los cálculos y/o los cortes de revisiones

N – Deshabilita los cálculos y/o los cortes de revisiones

En base a este parámetro se procede o no con los puntos siguientes.

3.4.1 [Registración de revisiones anteriores](#)

Resulta necesario contar con un repositorio de información histórica de revisiones anteriores cargado por el equipo editorial. Este repositorio, **tabla de revisiones T_r**, debe registrar al menos los eventos requeridos por la metodología para poder tomar decisiones sobre las solicitudes de revisiones actuales y futuras en base a toda la información recolectada.

Registración de la solicitud de revisión

En primera instancia, en la tabla de revisiones se debe registrar la solicitud de la revisión del artículo en cuestión al revisor asignado.

- Artículo sobre el que se solicita revisores
- Nombres del revisor asignado
- Fecha de solicitud

Registración de la revisión no aceptada

En caso que la revisión no sea aceptada se debe registrar en los datos iniciales de la registración y deben ser:

- Aceptación de la revisión (N- no aceptada)

Registración de la revisión aceptada

De igual manera, registrar:

- Aceptación de la revisión (S- aceptada)
- Fecha de aceptación

Registración de la revisión entregada

Sobre los datos de registración aceptada, se debe ingresar:

- Fecha de entrega

Antes de hacer la registración de la “calificación de la revisión” entregada es necesario desarrollar el contenido siguiente:

Una vez analizada la revisión por el equipo editorial para evaluar su pertinencia y otras cuestiones según el criterio de cada revista, la evaluación debe ser registrada en base a un esquema de calificaciones. Este esquema va a determinar una priorización o no de un candidato

a revisor para ser tenido en cuenta para la tarea. Además, puede dar cuenta de posibles penalizaciones por falta de calidad de la revisión enviada.

Se definen rangos en base a las diferentes calificaciones posibles en la **tabla de calificaciones T_c** , de la siguiente manera:

- Valor numérico de la calificación desde: valor de rango menor (0 en adelante).
- Valor numérico de la calificación hasta: valor de rango mayor (0 en adelante).
- Coeficiente: definido entre el rango (0 a n en base a criterio del equipo editorial)
- Mala calidad: se indica la calidad de la revisión (N- al menos calidad aceptable, S- mala calidad)

El coeficiente de la tabla se denomina **Coeficiente de evaluación de revisión CE_r** , y va a ser el valor que va a tomarse para el cálculo específico del indicador de revisión total. En la metodología es importante tomarlo como un valor de ventaja o desventaja relativo.

Registración de la calificación de la revisión entregada

En base a lo expuesto, en el registro correspondiente de la tabla de revisiones se realiza lo siguiente:

Se toman los datos de la registración de la revisión realizada y se agrega:

- Evaluación: valor numérico o descripción de la calificación en base a la tabla de calificaciones confeccionada según criterio del equipo editorial.
- Observaciones: lo que se considere oportuno sobre la evaluación.

Cabe señalar que se debe tener en cuenta a aquellas personas que nunca han sido convocadas por lo cual es importante definir cuál es el valor asignado. Para estos casos se propone 1, como valor neutro y con esa base se recomienda consignar los coeficientes a favor y en contra.

3.4.2 Aplicación de Controles temporales y tiempos de no consideración

Se identifican eventos que disparan tiempos de no consideración como inhabilitaciones temporales, que pueden estar determinados por diferentes motivos.

Pedidos simultáneos

En la solicitud de revisión a un evaluador puede darse el caso de simultaneidades provenientes de la misma revista, por lo cual este parámetro posibilita poner un límite a esa simultaneidad. Así, la selección de revisiones a evaluar se efectúa sobre las que fueron

solicitadas aún sin devolución. Se utiliza para ello el **Parámetro de revisiones sobre pedidos simultáneos: P_{rps}**

Los valores posibles son:

- Un valor igual o mayor a 1 indica la cantidad de revisiones simultáneas que se pueden solicitar a un revisor.
- Un valor igual a 0 determina que no se controla el número de revisiones simultáneas.

Las solicitudes de revisiones simultáneas mayores al valor 1, generalmente responden a situaciones excepcionales, por ejemplo, que no se pueda conseguir un académico cercano en conocimiento al manuscrito a revisar.

Procedimiento:

Determinación de inclusión de cálculo de aplicación de control de pedidos simultáneos usando P_{rps} :

Si $P_{rps} = 0$, no se aplica cálculo de aplicación de control de pedidos simultáneos

Si $P_{rps} > 0$, continua con el procedimiento de cálculo de aplicación de control de pedidos simultáneos

En el caso de estar habilitado, sobre cada uno de los candidatos a revisor priorizados se realiza lo siguiente:

Cálculo de aplicación de control de pedidos simultáneos:

- a. Obtener la cantidad de pedidos abiertos de revisiones del candidato a revisor de la tabla T_r :

$V_{abiertos}$ = Cantidad de pedidos abiertos del candidato a revisor

- b. Cálculo de correspondencia de restricción:

Si $P_{rps} > V_{abiertos}$, el candidato a revisor no tiene restricción de consideración por pedidos simultáneos

Si $P_{rps} \leq V_{abiertos}$, el candidato a revisor no puede ser considerado en base a límite de pedidos simultáneos.

Tiempo de descanso

Luego de entregada una revisión, el equipo editorial puede definir un tiempo en el cual no se vuelve a solicitar revisiones a dicha persona. Este parámetro busca evitar también la

sobrecarga de las solicitudes a un mismo revisor y toma la fecha de la última revisión entregada (en meses) desde cuando se calcula este control.

Parámetro de revisiones sobre tiempo de descanso: P_{rtd}

Los valores posibles son:

- Un valor igual o mayor a 1 indica la cantidad de meses de descanso de solicitud de revisiones a un revisor desde la entrega de la última revisión.
- Un valor igual a 0 determina que no se aplica este parámetro.

Procedimiento:

Determinación de inclusión de cálculo de aplicación de control de tiempo de descanso usando P_{rtd} :

Si $P_{rtd} = 0$, no se aplica el control de tiempo de descanso

Si $P_{rtd} > 0$, continua con el procedimiento de cálculo de aplicación de tiempo de descanso

En el caso de estar habilitado, sobre cada uno de los candidatos a revisor priorizados se realiza lo siguiente:

Cálculo de aplicación de control de tiempo de descanso:

- a. Obtener la fecha de entrega de la última revisión del candidato a revisor de la tabla T_r :

$F_{U_{revisión_entregada}}$ = Fecha de la última revisión aceptada del candidato a revisor

- b. Decidir si corresponde seguir con el procedimiento en base al valor de $F_{U_{revisión_entregada}}$:

Si $F_{U_{revisión_entregada}}$ es nulo, no se aplica el control de tiempo de descanso

Si $F_{U_{revisión_entregada}}$ no es nulo, continua con el cálculo de aplicación de control de tiempo de descanso

- c. Obtener la fecha actual:

F_{actual} = fecha actual

- d. Cálculo de correspondencia de penalización:

Si $F_{actual} > (F_{U_{revisión_entregada}} + P_{rtd} \text{ meses})$, el candidato a revisor no tiene restricción de consideración

Si $F_{\text{actual}} \leq (F_{\text{revisión_entregada}} + P_{\text{td}} \text{ meses})$, el candidato a revisor no puede ser considerado por restricción de consideración en base al tiempo de descanso definido entre pedidos de revisiones hasta que venza el plazo.

Revisiones por período

Este parámetro busca limitar la cantidad de revisiones que se le puede solicitar a un revisor por año, a los fines de evitar una distribución la labor evaluativa más equitativa. Para ello se usa el **Parámetro de revisiones sobre pedidos por período: P_{rpp}**

Los valores posibles son:

- Un valor igual o mayor a 1 indica la cantidad de revisiones por período que se puede pedir a un revisor.
- Un valor igual a 0 determina que no se controla el número de revisiones por período.

Procedimiento:

Determinación de inclusión de cálculo de aplicación de control de revisiones por período usando P_{rpp}

Si $P_{\text{rpp}} = 0$, no se aplica cálculo de aplicación de control de revisiones por período

Si $P_{\text{rpp}} > 0$, continua con el procedimiento de cálculo de aplicación de control de revisiones por período

En el caso de estar habilitado, sobre cada uno de los candidatos a revisor priorizados se realiza lo siguiente:

Cálculo de aplicación de control de revisiones por período:

- a. Obtener la cantidad de pedidos de revisiones del corriente año del candidato a revisor de la tabla T_r :

$V_{\text{año}} = \text{Cantidad de pedidos de revisiones del corriente año del candidato a revisor}$

- b. Cálculo de correspondencia de no consideración:

Si $P_{\text{rpp}} > V_{\text{año}}$, el candidato a revisor no tiene restricción de consideración por revisiones por período.

Si $P_{\text{rpp}} \leq V_{\text{año}}$, el candidato a revisor no puede ser considerado por restricción de consideración en base a cantidad de revisiones por período hasta el próximo.

3.4.3 Aplicación de Penalizaciones (temporales - totales)

Algunos eventos disparan penalizaciones sobre los candidatos a revisor. Estas pueden ser:

- Totales: No vuelven a ser considerados.
- Temporales: Luego de un tiempo determinado de suspensión de la posibilidad de ser considerados revisores, pueden volver a ser candidatos.

Penalización por rechazos de revisiones

La metodología posibilita realizar penalizaciones a los candidatos en base a la cantidad de rechazo de las solicitudes, lo cual afecta la posibilidad de ser considerados.

Esta penalización está contenida en el **Parámetro de revisiones sobre penalización de rechazadas: P_{rpr}**

Se utiliza la **tabla de cálculo para rechazos T_{cr}** en la que se establecen los rangos de penalización de la siguiente manera. Cada rango se define por la cantidad de rechazos, establecidos por la cantidad de rechazos de inicio y de fin, y se asocia a un valor de penalización (valor de P_{rpr}) con valores comprendidos entre el rango de 1 a n y "T". El valor numérico indica la cantidad de meses de penalización, "T" indica que es por tiempo indeterminado.

Para la habilitación de esta penalización se utiliza el siguiente parámetro:

Parámetro de habilitación de revisiones sobre penalización de rechazadas: PH_{rpr}

Teniendo dos opciones:

S – Habilita el control de penalización por revisiones rechazadas.

N – Deshabilita el control de penalización por revisiones rechazadas.

Procedimiento:

Decidir la inclusión del cálculo de penalización usando PH_{rpr} :

Si $PH_{rpr} = "N"$, no se aplica la penalización por rechazos de revisiones.

Si $PH_{rpr} = "S"$, se habilita el procedimiento de penalización por rechazos de revisiones.

En el caso de estar habilitado, sobre cada uno de los candidatos a revisor priorizados se continúa:

Determinación de P_{rpr} :

a. Obtener la cantidad de rechazos del candidato a revisor de la tabla T_r :

$$V_{\text{rechazos}} = \text{Cantidad de rechazos del candidato a revisor}$$

b. Decidir si corresponde seguir con el procedimiento en base al valor de $V_{rechazos}$:

Si $V_{rechazos} = 0$, no se aplica la penalización por rechazos de revisiones

Si $V_{rechazos} > 0$, extraer P_{rpr} en base al valor de cantidad de rechazos $V_{rechazos}$ dentro del rango rr de la tabla T_{cr} .

Formalmente, si $T_{cr} = \{(rr_{(1)}, P_{rpr(1)}), (rr_{(2)}, P_{rpr(2)}), \dots, (rr_{(n)}, P_{rpr(n)})\}$ donde cada $rr_{(i)}$ representa un rango de cantidad de rechazos determinado por una cantidad de inicio y de fin según el registro i , y $P_{rpr(i)}$ es el valor de penalización asociado, entonces se selecciona el $P_{rpr(i)}$ correspondiente al rango en el que se encuentra la cantidad de rechazos.

$$P_{rp} = \begin{cases} P_{rpr(1)} & \text{si } V_{rechazos(1)} \in rr_{(1)} \\ P_{rpr(2)} & \text{si } V_{rechazos(2)} \in rr_{(2)} \\ \vdots & \vdots \\ P_{rpr(n)} & \text{si } V_{rechazos(n)} \in rr_{(n)} \end{cases}$$

Determinación de penalización por tipo de penalización en base al valor de P_{rpr} :

Si P_{rpr} es "T" (Tiempo Indeterminado), el candidato está penalizado por tiempo indeterminado.

Si P_{rpr} es un valor numérico, se realiza la verificación de correspondencia de penalización:

1. Obtener fecha del último rechazo en tabla T_r :

$FU_{rechazo}$ = fecha del último rechazo del candidato a revisor

2. Obtener fecha actual: F_{actual} = fecha actual

3. Cálculo de correspondencia de penalización:

Si $F_{actual} > (FU_{rechazo} + P_{rpr} \text{ meses})$, no se aplica la penalización por rechazos de revisiones

Si $F_{actual} \leq (FU_{rechazo} + P_{rpr} \text{ meses})$, candidato penalizado hasta que termine el período de penalización.

Penalización por revisiones de baja calidad

La metodología permite realizar una penalización en base a la cantidad de revisiones no aceptables que tiene un revisor. Esta penalización está contenida en el **Parámetro de revisiones sobre penalización por mala calidad: P_{rpmc}**

Se utiliza la **tabla de cálculo para malas calificaciones T_{cmc}** en la que se establecen los rangos de penalización de la siguiente manera: Cada rango se define por la cantidad de malas revisiones, establecidos mediante una cantidad de rechazos de inicio y de fin, y se asocia a un valor de penalización (valor de P_{rpmc}) con valores comprendidos entre el rango de 1 a n y "T". El valor numérico indica la cantidad de meses de penalización, "T" indica que es por tiempo indeterminado.

Para la habilitación de esta penalización se utiliza el siguiente parámetro:

Parámetro de habilitación de control de revisiones de mala calidad: PH_{rmc}

Teniendo dos opciones:

S – Habilita el control de penalización por revisiones de mala calidad.

N – Deshabilita el control de penalización por revisiones de mala calidad.

Procedimiento:

Decidir la inclusión del cálculo de penalización usando PH_{rpmc} :

Si $PH_{rmc} = "N"$, no se aplica la penalización por revisiones de mala calidad

Si $PH_{rmc} = "S"$, habilita el procedimiento de penalización por revisiones de mala calidad

En el caso de estar habilitado, sobre cada uno de los candidatos a revisor priorizados se realiza lo siguiente:

Determinación de P_{rpmc} :

- a. Obtener la cantidad de revisiones de mala calidad del candidato a revisor de la tabla T_r asociada a los registros correspondientes a malas calificaciones de la tabla T_c :

$V_{mala_calidad} =$ Cantidad de revisiones de mala calidad del candidato a revisor

- b. Decidir si corresponde seguir con el procedimiento en base al valor de $V_{mala_calidad}$:

Si $V_{mala_calidad} = 0$, no se aplica la penalización por rechazos de revisiones

Si $V_{mala_calidad} > 0$, extraer P_{rpmc} en base al valor de cantidad de revisiones de mala calidad $V_{mala_calidad}$ dentro del rango rm de la tabla T_{cmc} :

Formalmente, si $T_{cmc} = \{(rm_{(1)}, P_{rpmc(1)}), (rm_{(2)}, P_{rpmc(2)}), \dots, (rm_{(n)}, P_{rpmc(n)})\}$ donde cada $rm_{(i)}$ representa un rango de cantidad de revisiones de mala calidad

determinada por una cantidad de malas de inicio y de fin según el registro i, y $P_{rpmc(i)}$ es el valor de penalización asociado, entonces se selecciona el $P_{rpmc(i)}$ correspondiente al rango en el que se encuentra la cantidad de malas revisiones.

$$P_{rpmc} = \begin{cases} P_{rpmc(1)} & \text{si } V_{mala_calidad(1)} \in rm_{(1)} \\ P_{rpmc(2)} & \text{si } V_{mala_calidad(2)} \in rm_{(2)} \\ \vdots & \vdots \\ P_{rpmc(n)} & \text{si } V_{mala_calidad(n)} \in rm_{(n)} \end{cases}$$

Determinación de penalización por tipo de penalización en base al valor de P_{rpmc} :

Si P_{rpmc} es "T" (Tiempo Indeterminado), el candidato está penalizado por tiempo indeterminado.

Si P_{rpmc} es un valor numérico, realizar verificación de correspondencia de penalización:

1. Obtener la fecha de la última revisión de mala calidad de la tabla T_r :

$FU_{mala_calidad}$ = fecha de la última revisión de mala calidad del candidato a revisor

Obtener la fecha actual: F_{actual} = fecha actual

2. Cálculo de correspondencia de penalización:

Si $F_{actual} > (FU_{mala_calidad} + P_{rpmc} \text{ meses})$, no se aplica la penalización por revisiones de mala calidad

Si $F_{actual} \leq (FU_{mala_calidad} + P_{rpmc} \text{ meses})$, el candidato está penalizado hasta que venza el plazo.

3.4.4 Parámetros generales de revisiones

Existen parámetros que son específicos del cálculo general de revisiones:

- Inclusión de parámetro de cálculo y cortes: existe la posibilidad de definir si se va a tener en cuenta el globalizador de factores de revisiones anteriores con las opciones de cálculo y cortes. Se realiza a través del **Parámetro de cálculo de revisiones con cálculo y cortes: PC_{rec}**

Los valores posibles:

- cálculo-cortes: Habilita las revisiones anteriores y las opciones de realizar el cálculo y los cortes.

- cálculo: Habilita las revisiones anteriores y las opciones de realizar solo el cálculo, sin cortes.
- cortes: Habilita las revisiones anteriores y las opciones de realizar solo los cortes, sin el cálculo.
- Tipo de cálculo entre valores del coeficiente de evaluación de revisiones: se pueden establecer diferentes operaciones entre las diferentes evaluaciones de las revisiones realizadas. Se realiza a través del **Parámetro de cálculo entre valores del coeficiente de evaluación de revisiones: PC_{cer}**

Los valores posibles son:

- Suma
- Promedio
- Producto

3.4.5 Cálculo de Puntaje de revisiones anteriores

Se ve reflejado en la variable metodológica **Puntaje de revisiones anteriores: P_{ra}** y contiene el valor derivado de los cálculos de la metodología dentro del Globalizador de factores de revisiones anteriores. Se ve afectado por los controles que pueden surgir de la aplicación de los siguientes parámetros: Parámetro de cálculo de revisiones con cálculo y cortes junto al Parámetro de cálculo entre valores del coeficiente de evaluación de revisiones.

El Coeficiente de evaluación de revisiones es la base fundamental del cálculo de este Indicador, cuyos datos individuales son obtenidos de todos los puntajes de evaluaciones de calidad de todas las revisiones realizadas por el candidato a revisor.

Si el Parámetro de cálculo de revisiones con cálculo y cortes tiene los valores cálculo-cortes o cálculo, según sea el valor del Parámetro de cálculo entre valores del coeficiente de evaluación de revisiones se realizan las siguientes operaciones para obtener el Puntaje de revisiones total:

- Si es Suma: $P_{ra} = \sum_{i=1}^n CE_{r(i)}$
- Si es Promedio: $P_{ra} = \frac{\sum_{i=1}^n CE_{r(i)}}{n}$
- Si es Producto: $P_{ra} = \begin{cases} PCE_{r(1)} = CE_{r(1)} \\ PCE_{r(i)} = PCE_{r(i-1)} * CE_{r(i)} \text{ para } i > 1 \end{cases}$

En la gráfica de la Figura 7 se pueden observar todos los procesos llevados a cabo en esta etapa de la metodología con todos los parámetros intervinientes, los insumos y los resultados en cada paso.

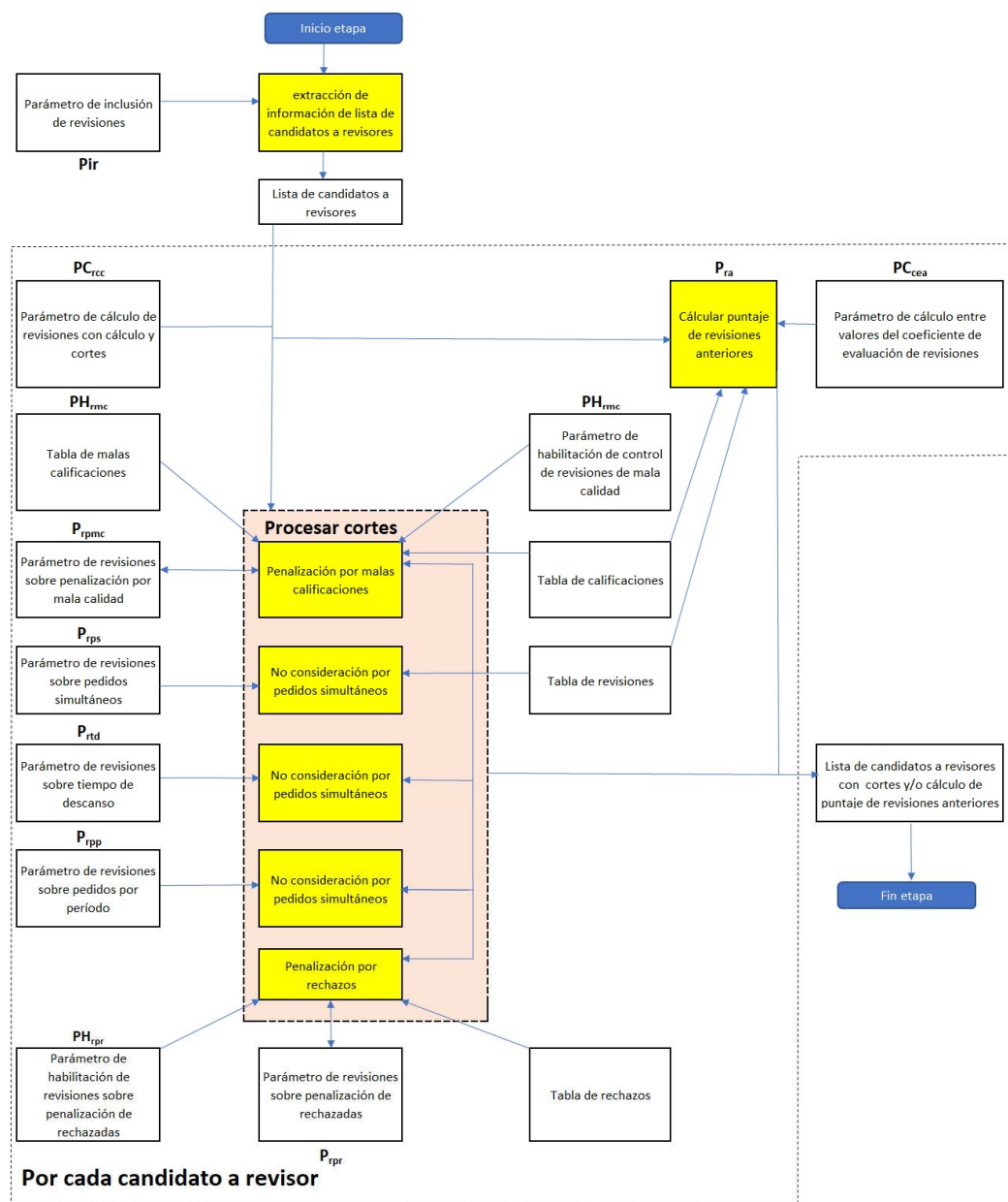


Diagrama del proceso de cálculo y cortes de revisiones

3.5 Fórmula de la metodología con parámetros integrales y con múltiples perfiles de configuración

En esta sección describe la fórmula general final con todos los parámetros integrales que se aplican y la posibilidad de incluir múltiples perfiles de configuración.

Hay 2 tipos parámetros integrales que controlan la relación de los globalizadores de factores, a saber:

Peso relativo de globalizadores de factores

El peso relativo se aplica a cada globalizador y define en término generales, la proporcionalidad de su valor relacionado con los otros globalizadores de factores. Este valor se multiplica al puntaje de cada globalizador de factores. Los parámetros son los siguientes:

- **Parámetro integral de peso de producción científica PIP_{pc}**
- **Parámetro integral de peso de autoridad académica PIP_{aa}**
- **Parámetro integral de peso de revisiones anteriores PIP_{ra}**

El rango de valores que puede asumir cada uno de los 3 parámetros integrales se define entre 0 y 100, con una restricción: la suma del valor de estos aplicado a los 3 globalizadores de factores debe dar 100, que representa el valor límite máximo que puede tomar el globalizador de factores.

$$PIP_{pc} + PIP_{aa} + PIP_{ra} = 100$$

Peso relativo de la unidad de globalizadores de factores

El peso relativo se asigna a cada globalizador y define la proporción del Indicador del globalizador que refleja el valor relativo de la unidad del Indicador asociado. Considerando el peso relativo del globalizador de factores, se determina el número máximo de unidades que puede tener el Indicador asociado. Este valor se multiplica por el puntaje de cada globalizador de factores. Los parámetros son los siguientes:

- **Parámetro integral de peso de unidad de producción científica $PIPU_{pc}$**
- **Parámetro integral de peso de unidad de autoridad académica $PIPU_{aa}$**
- **Parámetro integral de peso de unidad de revisiones anteriores $PIPU_{ra}$**

El rango de valores reales que puede asumir cada uno de los 3 parámetros integrales es de 0 a 1.

Si bien se puede aplicar de forma individual a cada globalizador, la proporcionalidad en el cálculo debe ser tenida en cuenta. Si bien cada globalizador tiene cálculos independientes de los otros y distintas representaciones que responden a diferentes lógicas, es preciso considerar

una proporcionalidad adecuada en la fórmula, por lo cual es recomendable utilizar igual valor de peso de unidad en cada globalizador.

Fórmula con los parámetros integrales asociados

La fórmula general de un posible revisor $R_{(i)}$ de un Manuscrito P sería:

$$R_{(i)}P = PIP_{pc} * PIPU_{pc} * P_{pc} + PIP_{aa} * PIPU_{aa} * P_{aa} + PIP_{ra} * PIPU_{ra} * P_{ra}$$

Teniendo en cuenta las restricciones siguientes:

$$PIP_{pc} * PIPU_{pc} * P_{pc} \leq PIP_{pc}$$

$$PIP_{aa} * PIPU_{aa} * P_{aa} \leq PIP_{aa}$$

$$PIP_{ra} * PIPU_{ra} * P_{ra} \leq PIP_{ra}$$

Cabe mencionar que la fórmula total desplegada tiene en cuenta la activación de los cálculos de Autoridad académica y Revisiones anteriores. Según se desactive alguna de estas o las dos se anulan los respectivos cálculos.

Una vez aplicada la fórmula general se procede a ordenar de forma descendente todos los candidatos a revisores, teniendo como productos dos listados, a saber:

- Listado de posibles revisores ordenados de forma descendente.
- Listado de candidatos que tienen conflicto de interés y/o cortes por análisis de revisiones anteriores, con el/los motivos que hubiere para ser incluidos en el listado

3.5.1 Múltiples perfiles de configuración de la metodología

Finalmente, se genera un nivel superior de agrupamiento reflejado en el índice k que representa diferentes perfiles de configuración con sus propios valores de parámetros, parámetros integrales y variables metodológicas de toda la fórmula completa.

$$R_{(i)}P_k = PIP_{pc,k} * PIPU_{pc,k} * P_{pc,k} + PIP_{aa,k} * PIPU_{aa,k} * P_{aa,k} + PIP_{ra,k} * PIPU_{ra,k} * P_{ra,k}$$

$$R_{(i)}P_k = \begin{cases} PIP_{pc,1} * PIPU_{pc,1} * P_{pc,1} + PIP_{aa,1} * PIPU_{aa,1} * P_{aa,1} + PIP_{ra,1} * PIPU_{ra,1} * P_{ra,1} & \text{si } k = 1 \\ PIP_{pc,2} * PIPU_{pc,2} * P_{pc,2} + PIP_{aa,2} * PIPU_{aa,2} * P_{aa,2} + PIP_{ra,2} * PIPU_{ra,2} * P_{ra,2} & \text{si } k = 2 \\ \vdots & \\ PIP_{pc,n} * PIPU_{pc,n} * P_{pc,n} + PIP_{aa,n} * PIPU_{aa,n} * P_{aa,n} + PIP_{ra,n} * PIPU_{ra,n} * P_{ra,n} & \text{si } k = n \end{cases}$$

4 Prototipo y pruebas de laboratorio

En el marco de la metodología de la presente tesis doctoral, se desarrolló un prototipo de laboratorio cuya finalidad es validar lo original de la metodología propuesta, ya que resulta útil para probar diferentes alternativas. En este sentido, en las pruebas de laboratorio se registran valores por defecto producto del análisis de los datos presentados en los capítulos precedentes para determinar otros valores por defecto de índole más técnica, que sean más adecuados a los procesos intervinientes. Por lo tanto, no se intenta resolver los problemas de un sistema funcional, sino verificar paulatinamente todas las operaciones respetando la metodología, extrayendo información valiosa sobre datos reales para mostrar información estadística, brindar informes intermedios que sirvan para realimentar la propuesta y así poder evaluar los resultados. En este proceso resulta fundamental brindar informes que permitan ser evaluados por expertos con el objeto de optimizar en la medida de lo posible todos los pasos de un informe final de recomendación de revisores.

Es importante precisar, con respecto al corpus de datos utilizados en esta investigación, que el propósito central no consistió en abordar exhaustivamente todas las problemáticas potenciales derivadas de la extracción y análisis de la totalidad de información disponible en un sistema funcional completo. Por el contrario, se ha adoptado un enfoque selectivo que prioriza la recopilación y procesamiento de información suficiente y necesaria para evaluar con rigor la metodología implementada.

Esta delimitación metodológica responde a criterios de pertinencia investigativa y eficiencia analítica. Ex profeso se han excluido del presente análisis otras complejidades específicas inherentes al desarrollo de sistemas funcionales, las cuales, si bien relevantes en otros contextos, no contribuyen directamente a los objetivos centrales establecidos en la introducción de este trabajo.

Las dimensiones no contempladas en el presente estudio pueden ser abordadas en investigaciones posteriores, una vez evaluada la efectividad y eficiencia de la metodología propuesta. Este análisis secuencial permitirá construir sobre bases empíricamente validadas, garantizando que las futuras indagaciones sobre aspectos más complejos del sistema se fundamenten en una comprensión sólida de los elementos metodológicos centrales aquí examinados. Esta aproximación escalonada resulta coherente con los objetivos investigativos de esta tesis doctoral y establece las bases para un programa de investigación progresivo y sistemático.

4.1 Definiciones de tecnología

A nivel tecnológico se tomaron diferentes definiciones:

Lenguaje de programación: se ha utilizado Python en la versión 3.11 con todas las librerías necesarias para hacer las tareas de extracción de información, procesamiento, guardado y confección de informes necesarios. Su elección se basa en que proporciona las librerías necesarias para hacer el *pipeline* de PLN, secuencia estructurada de pasos a través de los cuales se procesan y analizan textos para aplicar modelos de lenguaje y terminar comparando textos. Además, permite la incorporación de técnicas y modelos de IA necesarios para el desarrollo, optimización y automatización de los procesos de análisis textual. Cada paso del *pipeline* realiza una tarea específica, y el resultado de cada etapa se utiliza como entrada para la siguiente, asegurando un flujo eficiente de procesamiento de datos.

La computadora utilizada tiene un sistema operativo GNU Linux de distribución Debian 12 Bookworm sin interfaz gráfica con conexión a Internet. Al sistema se accede por consola desde la pantalla del servidor o desde terminales remotas por SSH.

Una de las definiciones importantes generales del proceso fue utilizar software disponible que pueda ser utilizado localmente en los procesos, y no utilizar APIs (del inglés Application Programming Interface) propietarias y/o que tienen una restricción de uso en el tiempo, además del hecho de que no se cuenta con presupuesto para afrontar estas erogaciones.

4.2 Definiciones asociadas a la metodología

Dentro de cada paso se tomaron decisiones prácticas asociadas a una implementación de la metodología propuesta

4.2.1 Adquisición, preprocesamiento y guardado de datos

Identificación y elección de fuentes de datos

Luego de evaluar la necesidad de datos que tiene la metodología, se resolvió utilizar como base principal de alimentación de artículos científicos, el Repositorio Institucional (RI) del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) denominado “CONICET Digital”. Además de la relevancia internacional reconocida al CONICET como organismo de CyT, este repositorio es de acceso abierto y su misión es reunir, registrar, divulgar, preservar y dar acceso público a la producción científico tecnológica de la institución.

En la actualidad posee más de 200.000 títulos. Los tipos de documentos que almacena son:

- Artículos

- Capítulo de libros
- Libros
- Trabajos en Eventos científicos tecnológicos
- Tesis

El conocimiento lo clasifica según las siguientes disciplinas:

- Ciencias Biológicas
- Ciencias de la Tierra y relacionadas con el Medio Ambiente
- Ciencias Médicas y de la Salud
- Ciencias Naturales y Exactas
- Ciencias Físicas
- Ciencias Sociales
- Ingenierías y Tecnologías
- Historia y Arqueología
- Humanidades
- Sociología

Dado que el trabajo de campo mayormente implicó al Consejo Editorial de la Revista IRICE (CONICET-UNR) el contenido seleccionado se inscribe dentro del área de Ciencias Sociales en el campo de las Ciencias de la Educación con sus subcategorías. Dado el objetivo de esta tesis, el tipo de documento seleccionado fue Artículos.

Sobre el Repositorio CONICET Digital vale señalar algunas consideraciones:

- Las publicaciones de procedencia corresponden a diferentes revistas, congresos, etc. Por lo cual los formatos de los artículos son muy diversos. No obstante, la clasificación de la información es centralizada e incluye los metadatos necesarios para la metodología, a saber:
 - Títulos
 - Autores
 - Resúmenes
 - Palabras Claves
 - Fecha de publicación
 - Textos completos del material (excepto aquellos con embargos de lapso activo)

Como se anticipó, tiene los recursos clasificados por la temática específica a la que se pretende acceder y agrega el idioma que tiene utilidad para este chequeo en el preprocesamiento, lo cual permite tener una ventaja de la clasificación centralizada y la

posibilidad de acceder a diversas fuentes de recursos publicados en diferentes lugares asemejando fuentes heterogéneas de datos.

Otro aspecto positivo, es que está organizando una base centralizada de autores codificada de CONICET lo que garantiza una parte de la desambiguación directamente por sistema y permite:

- Certificar que es el autor de más de un artículo en el caso que tenga más de un artículo publicado.
- Acceder a información de titulaciones y afiliaciones de los investigadores. Si bien no es un registro completo de la totalidad de autores (ya que en asociación varios de ellos no son CONICET), es información adicional que sirve y se presenta valiosa para chequear la metodología propuesta.
- En relación a la Revista IRICE (CONICET-UNR) el repositorio CONICET es adecuado ya que proporciona una base de autores calificados como posibles revisores de una revista de prestigio que se inscribe en ese marco institucional.

Además de la fuente de datos principal de artículos científicos e información relacionada de autores desambiguados en el caso de ser investigadores CONICET y con información de titulaciones y afiliaciones. Dicha información se considera en esta etapa de pruebas de laboratorio como suficiente.

Al no estar todos los investigadores incluidos se necesita otra fuente de información para desambiguar y contemplar también la necesidad de indicadores bibliométricos, por lo cual es necesaria una base de perfiles de investigadores que facilite ambos procesos. Entonces, se buscó una base con:

- Impacto internacional
- Información actualizada y dinámicamente actualizable a diario o en intervalos cortos de tiempo.
- Usada por una importante cantidad de investigadores del ámbito específico definido.

De todas las posibilidades estudiadas se definió el uso de los perfiles de Google Scholar. Este es un motor de búsqueda de Google especializado en la búsqueda de contenido y bibliografía científico-académica. Los perfiles de Google Scholar permiten que los autores muestren sus publicaciones académicas, acceder a quienes están citando sus artículos y cuáles son sus indicadores bibliométricos. A los investigadores les permite añadir variantes de nombres diferentes que se usan para las publicaciones de artículos, afiliación, correo de verificación, áreas de interés y página principal como opcional. Asimismo, permite búsquedas automáticas de

artículos que el investigador debe seleccionar como propias o cargarlas manualmente. También sobre actualizaciones, al encontrar nuevas publicaciones del autor brinda la posibilidad de cargarlos automáticamente o enviar un mail para confirmar. También, posibilita definir que el perfil sea público de modo que pueda aparecer en las búsquedas de Google Scholar.

Extracción de datos de las fuentes de investigación

Una vez elegida la fuente de datos utilizada se siguen los procesos de la metodología comenzando por la extracción de esos datos. De las posibilidades de acceso a la información se establece que el acceso por *Web scraping* es adecuado. Vista la factibilidad, se armaron las rutinas de descarga con chequeos de datos ya obtenidos y descargados para la actualización:

Datos a ser utilizados: Se extraen del RI CONICET los datos asociados al artículo principalmente y derivados disponibles:

- ✓ Título, autores, resumen, palabras claves.
- ✓ Asociaciones a características de la publicación como fecha de publicación e idiomas.
- ✓ Datos de la carga de los perfiles de los autores con identificación CONICET.
- ✓ Archivos de los artículos en caso de estar disponibles.

Se extraen los datos asociados a la desambiguación de autores con todos los indicadores bibliométricos:

- ✓ Nombres de los investigadores en los perfiles.
- ✓ Información de publicaciones.
- ✓ Indicadores bibliométricos.

Cabe aclarar que, en la etapa inicial solo se guarda la información del repositorio RI que contiene la información de base que debe ser depurada en el procesamiento. Los datos asociados a Google Scholar tienen otra opción de descarga en el prototipo y es una tarea posterior al preprocesamiento inicial del otro contenido.

Uno de los desafíos del desarrollo fue poder bajar todos los perfiles necesarios, principalmente para desambiguación, tratando de evitar los bloqueos de Google. Esto provocó que se tengan en cuenta los tiempos de penalización y bloqueo de los procesos que obtenían los perfiles de Google Scholar. Esto se debe considerar en un desarrollo funcional debido a la alta demanda de búsqueda de perfiles actualizados en el proceso de recomendación. En el caso

del prototipo es suficiente para obtener un estado inicial que posibilite probar la metodología. Cabe acotar que existen soluciones de pago que se pueden utilizar para saltar esas barreras.

Guardado de datos

Como primera instancia cabe aclarar que, si bien en la metodología el proceso de guardado de datos aparece posteriormente al preprocesamiento, en el prototipo ese proceso se va a realizar constantemente para probar y mejorar en el desarrollo cada uno de los pasos de la metodología.

Los datos extraídos son guardados como archivos con extensión CSV y los archivos que contienen los artículos presentados en formato PDF, son descargados en un directorio.

En esta primera instancia se descargan los datos disponibles en el RI CONICET

Se utilizan los IDs de los recursos que forman parte de la ruta de la url tanto sea para artículos, como para perfiles de investigadores CONICET cargados. Los datos son guardados en archivos CSV, dentro del directorio csv, de la siguiente manera:

En un archivo se guardan el ID del recurs, título del artículo, datos asociados a la publicación como fecha e idioma entre otra información. En otros se guardan separados los resúmenes y las palabras claves (también se decidió guardar las clasificaciones del artículo en el repositorio). Aparte se guardan los autores tomando un autor por registro, por ID del recurso ~~id_recurso~~ con orden de aparición en el artículo e ID CONICET en caso de existir y en otro archivo, los datos de los autores del CONICET con información adicional cargada sobre el autor por ID CONICET.

Los archivos PDFs de los artículos se guardan en el directorio *files*.

No solo la información de las fuentes de datos se guarda, también se guardan en archivos CSV, los parámetros del sistema y las tablas con rangos que se usan en la metodología.

Preprocesamiento

Como ya fue explicado en el capítulo anterior, en esta etapa se realizan todos los pasos necesarios para convertir los datos crudos en depurados para ser utilizados en las siguientes etapas de la metodología.

En el prototipo y en base a las definiciones de la metodología, este proceso es extenso en el control y en el desarrollo por la importancia de contar con los mejores datos posibles para las siguientes etapas.

A continuación, se detallan cada uno de los pasos realizados:

A) Control de datos extraídos

Se controló la estructura de datos propuesta y se observó que es adecuada. Luego se realizó la estadística de control de datos extraídos, en base al requerimiento de la metodología como muestra la tabla 13:

Tabla 13

Estadísticas generales de datos guardados en el proceso de extracción

Descripción	Cantidad
Cantidad de recursos en recursos.csv:	3880
Registros de autores en autores.csv:	8610
autores agrupados por recursos: (sin diferencias):	3880 (sin diferencias)
autores agrupados por nombres:	3488 (dato a priori de nombres diferentes)
Registros de resúmenes en resúmenes.csv:	3880 (sin diferencias)
Registros de clasificaciones en clasificaciones.csv:	4374
clasificaciones agrupadas por recursos:	3880 (sin diferencias)
Registros de palabras claves en palabras-claves.csv:	16219
palabras claves agrupadas por recurso:	3872 (diferencia de faltante de 8 recursos sin palabras claves)
Registros de autores CONICET en autores-conicet.csv:	823
autores CONICET: agrupados por id_conicet:	823 (no hay registros duplicados)

En la Tabla 14 se puede observar la agrupación de registros de autores CONICET por nivel de titulación:

Tabla 14*Agrupación de registros de autores CONICET por nivel de titulación*

Nivel de titulación	Cantidad
Terciario no universitario	2
Universitario de grado	36
Universitario de posgrado/doctorado	600
Universitario de posgrado/especialización	2
Universitario de posgrado/maestría	8
autores sin datos en grado	175

De esta primera instancia se registraron 8 registros con diferencias en las palabras claves y, se procedió a registrar los recursos con diferencias quitándolos de la base de conocimiento de trabajo. El porcentaje de pérdida no es significativo.

Con respecto a la disponibilidad de los artículos científicos para su descarga, se presenta el registro en la Tabla 15:

Tabla 15*Estadísticas por acceso a la descarga*

Descargar	3785
Solicitar	94 (se debe solicitar porque no son de libre descarga)
Inicial sin datos	1 (no fue publicado el enlace de descarga)

Según la etiqueta de idioma correspondiente a la publicación de la información de cada artículo del RI se publica la información guardada que corresponde. Al realizarse la registración de las palabras claves por cada palabra, se registra en cada línea el idioma correspondiente. En la Tabla 16 se pueden observar las estadísticas por idioma en recursos, resúmenes y palabras claves.

Con respecto al control, depuración y traducción de títulos, resúmenes, palabras claves y archivos de los artículos completos, si bien a nivel teórico se plantean diferenciados los datos y categorizados, en el caso particular del procesamiento del prototipo es conveniente realizar los controles de estos datos y de los archivos que contienen los artículos para luego hacer una limpieza de los registros en forma conjunta y sacar la estadística correspondiente. A posteriori es necesario efectuar la traducción de los títulos, resúmenes y palabras claves.

Tabla 16

Estadísticas por idioma en recursos, resúmenes y palabras claves

RECURSOS Tamaño dataframe: 3880		RESÚMENES Tamaño dataframe: 3880		PALABRAS CLAVES Tamaño dataframe: 11334	
Idioma	Cantidad	idioma	Cantidad	idioma	Cantidad
Catalán	3	Catalán	3	Catalán	11
Español	3444	Español	3444	Español	14330
Frances	9	Frances	9	Frances	37
Inglés	370	Inglés	370	Inglés	1642
Italiano	2	Italiano	2	Italiano	7
Portugués	52	Portugués	52	Portugués	192

Consideraciones previas al proceso

La determinación de base del idioma castellano se fundamenta en relación al caso revista IRICE y teniendo en cuenta la gran disponibilidad de recursos en este idioma en las estadísticas anteriores.

Es clave el proceso de limpieza de datos para la prueba de la metodología ya que los problemas de diferencias deben ser resueltos limpiándolos de la base de conocimiento. Mientras la cantidad de datos sea suficiente para la prueba y las depuraciones no afecten dicha cantidad, se logra el objetivo de contar con datos depurados. La base de esta verificación atiende a las posibles inconsistencias en los datos extraídos como duplicaciones y omisiones, diferencias de idiomas e información no útil para los siguientes procesos. En base a esa premisa se procede a definir las acciones a realizar en el proceso teniendo en cuenta las definiciones de la metodología.

En primera instancia se realiza el paso **B.1: Control de corrección de los datos**, mediante:

- Datos duplicados o nulos: se va a verificar la posibilidad de que existan palabras claves duplicadas por algún error de carga, que se hayan omitido o que en el proceso de migración por cuestiones de formato se hayan cargado erróneamente y sean nulas.
- Control sobre artículos no descargados: paso previo al proceso de conversión de archivos de artículos se va a chequear la existencia de archivos no descargados y que debieron descargarse. Se verifica sobre los datos extraídos, tomando la base de recursos y comparándolo con los PDFs de los artículos descargados. Se puede dar que ante un problema en la descarga por diferentes circunstancias propias del servidor donde se guardan los datos o ajenos al sistema como problemas de conexión a internet o servidor de extracción fuera

de servicio no se pueda descargar algún artículo. En este punto se verifica ese dato. Se registra que no se cuenta con el archivo PDF del artículo para depurar la base.

Las diferencias de Idiomas en títulos, resúmenes, palabras claves se abordan en base al punto **B.2: Control de idioma de los datos** de la metodología:

- La base de conocimientos ya mencionada tiene una etiqueta de idiomas por lo cual se puede constatar con ella el idioma detectado y así poder definir acciones. Se procede a constatar el idioma publicado con el idioma detectado, se corren diferentes procesos para detectarlo. En el caso de detectarse diferencias se procede a limpiar dichos registros de la base de conocimiento para contar con registros depurados para los siguientes procesos. Las mismas se procesan en conjunto al final de todos los pasos para hacer un consolidado de diferencias y luego limpiarlos de la base de conocimiento.

Como fuera definido en la metodología, en la etapa de preprocesamiento se evalúa el paso **C. Proceso de determinación y conversión de textos**:

C.1: Determinación y conversión de formatos: En este caso particular y para poder trabajar con los archivos de los artículos es necesario realizar el proceso de conversión de los archivos PDFs a archivos de texto, determinando que la extensión del formato destino sea TXT.

C.2: Registro de errores en el proceso de conversión se registran los errores del proceso de extracción de datos de los archivos PDFs. No solo se pueden presentar problemas en el momento de la conversión, también los datos extraídos pueden no tener información que pueda ser de utilidad debido a codificaciones específicas.

C.3.: Determinación de datos no válidos.

Depuración de los datos: considerando los puntos **B.3: Depuración de datos** y **C.4: Depuración de datos** ya mencionados en la metodología se toman las registraciones de diferencias detectadas anteriormente para proceder a la limpieza de la base y generar las estadísticas de los cambios.

Finalmente, teniendo en cuenta el punto **B.4: Limpieza y traducción del texto de los artículos** y tomando el lenguaje base se traducen todos los textos que están en otro idioma.

En base a los procedimientos y consideraciones expuestas se obtiene lo siguiente:

- ✓ Datos duplicados o nulos: de la revisión de este punto, se llegó al resultado descrito en la Tabla 17 sobre palabras claves.

Tabla 17

Estadísticas del proceso de control y limpieza de palabras claves

Cantidad Palabras claves	16219
Cantidad Palabras claves duplicadas	9
Cantidad Palabras claves vacías	19
Cantidad Palabras claves duplicadas y vacías	28
Cantidad Palabras claves corregidas	16191

- ✓ Control sobre artículos no descargados: No dio diferencias
- ✓ Diferencias **de Idiomas en títulos, resúmenes, palabras claves**: Se comprobó con 4 librerías diferentes. Se utilizaron las bibliotecas cld2, cld3, langdetect y fastText.

Se verificó la cantidad de diferencias entre los posibles idiomas que figuran en las etiquetas y la cantidad de diferencias en general. Lo deseable es que las diferencias sean mínimas, pero en el caso de haber diferencias, lo recomendable es que las diferencias totales no tomen idiomas fuera de los idiomas elegibles.

1 - En Títulos

Tamaño dataframe palabras a evaluar: 3880

Se pueden observar las diferencias entre idiomas elegibles e idiomas totales en la Tabla 18.

Tabla 18

Diferencias en títulos entre idiomas según diferentes bibliotecas utilizadas

Procesos	cld2	cld3	Langdetect	fastText
Cantidad de diferencias entre idiomas elegibles	58	53	59	44
Cantidad de diferencias de idiomas	906	111	60	44

Se observa que la menor variabilidad y coincidencia entre diferencias totales y entre los idiomas elegibles se presenta en el método de fastText.

2 - En Resúmenes

Tamaño dataframe palabras a evaluar: 3880

Se pueden observar las diferencias entre idiomas elegibles e idiomas totales en la Tabla 19.

Tabla 19*Diferencias en resúmenes entre idiomas según diferentes bibliotecas utilizadas*

Procesos	cld2	cld3	langdetect	fastText
Cantidad de diferencias entre idiomas elegibles	86	86	87	86
Cantidad de diferencias de idiomas	91	88	89	88

En este caso, se observa una equivalencia entre las diferentes librerías, con poca variabilidad y coincidencia entre diferencias totales y entre los idiomas elegibles. Es decir que con las diferentes librerías se obtienen resultados muy parecidos. La interpretación de la variabilidad se presenta porque existe una cantidad de texto adecuada para que puedan trabajar las diferentes librerías, esto no pasa con títulos y palabras claves dado el poco texto. Este dato es importante porque se puede inferir que para el texto completo del artículo todas las librerías podrían ser útiles.

3 - En Palabras Claves

En este caso se toman todas las palabras claves juntas de cada recurso, a diferencias del guardado que se realizó una palabra clave por registro.

Tamaño conjunto de palabras claves a evaluar agrupadas por recurso: 3872 (8 menos que son las palabras claves vacías)

Se pueden observar las diferencias entre idiomas elegibles e idiomas totales en la Tabla 20.

Tabla 20*Diferencias en palabras claves entre idiomas según diferentes bibliotecas utilizadas*

Procesos	cld2	cld3	langdetect	fastText
Cantidad de diferencias entre idiomas elegibles	213	207	222	210
Cantidad de diferencias de idiomas	1782	805	242	211

Hay dos modelos que tuvieron la menor variabilidad en diferencias totales que son langdetect y fastText, este último con menor diferencia. Y en entre las elegibles y relacionadas a las totales, fastText tiene menor variabilidad con las totales, lo cual es un comportamiento más adecuado.

En síntesis, de las cuatro librerías utilizadas la que tuvo un comportamiento más efectivo y menos inestable en relación con los idiomas encontrados fue fastText en relación con títulos, resúmenes y palabras claves. En una inspección visual a priori se pudo detectar que hay etiquetas con error de idiomas, por lo cual no se procede a corregirlo sino a depurar la base con procesos automatizados. Así, se procedió a fundir, detectando las diferencias coincidentes en

los mismos registros de artículos dando como resultado la eliminación de 266 registros con diferencias en idiomas sea en títulos, resúmenes y/o palabras claves.

Es importante tener en cuenta la diversidad del RI ya explicitada. Esto dificulta un proceso general de extracción eficiente de todos los datos por lo cual es posible que no se pueda contar con la información de todos los artículos. Igualmente se aclara que no se necesita para esta etapa un proceso de excelencia, sino uno suficiente para extraer la mayor cantidad de datos posibles para las pruebas.

Dada esta situación, se utiliza un paquete para la conversión, en este caso fue pdfminer3 que en las pruebas realizadas ha demostrado robustez para la tarea y en el caso de detectar problemas en la conversión que impiden que se realice, se registra en un archivo de errores. Resultado: se registraron 51 PDFs que tenían problemas para ser abiertos.

Archivos TXT que no tienen datos: se estableció un proceso de detección de errores en los casos que se generaron los archivos TXT, es decir previo al proceso de conversión, en base a los idiomas elegibles y que no sean detectados como indefinidos, ya sea por no tener datos o por codificación sin texto. Para hacer este desarrollo fue necesario ajustar constantemente los valores de cálculo hasta lograr la configuración adecuada.

Se empleó el modelo lid.176.bin de los recursos disponibles de fastText para la identificación de lenguajes en textos largos, dado que su rendimiento en términos de velocidad y precisión es óptimo para las necesidades de este trabajo. Se toma un umbral definido en el 25 % del documento. Resultado: Se detectaron 113 PDFs que no contienen datos. Cantidad de títulos repetidos: 1

Depuración de los datos: Teniendo en cuenta los puntos **B.3: Depuración de datos** y **C.4: Depuración de datos** de la metodología y tomando las registraciones de diferencias detectadas anteriormente se procede a la limpieza de la base y se generan estadísticas de los cambios. Resultado final de corrección de datos con estadísticas comparativas se muestran en la Tabla 21.

Visto las modificaciones de los datos totales depurados, la base resultante es adecuada para realizar las pruebas de la metodología.

Tabla 21*Estadística comparativas iniciales y finales luego del proceso de depuración y limpieza*

Recursos	Cantidad inicial	Cantidad final
Totales	3880	3505
Totales (solo descargables)	3785	3415

Recursos por tipo	Cantidad inicial	Cantidad final
Descargables directamente	3785	3415
Solicitar	94	90
Sin datos	1	0

Recursos totales por idioma	Cantidad inicial	Cantidad final
Catalán	3	1
Español	3444	3136
Francés	9	3
Inglés	370	336
Italiano	2	1
Portugués	52	28

Recursos descargables por idioma	Cantidad inicial	Cantidad final
Catalán	3	1
Español	3419	3114
Francés	9	3
Inglés	301	269
Italiano	1	0
Portugués	52	28

Registros de autores	Cantidad inicial	Cantidad final
Totales	8610	8337
Totales (solo descargables)	7775	7515

Registros de autores diferentes	Cantidad inicial	Cantidad final
Totales	3488	3375
Totales (solo descargables)	3217	3107

Autores CONICET asociados a recursos	Cantidad inicial	Cantidad final
Totales	1423	1376
Totales (solo descargables)	1307	1259

Autores CONICET con datos académicos	Cantidad inicial	Cantidad final
Totales	880	834
Totales (solo descargables)	785	738

Datos cargados de Estudios Autores CONICET de todos los artículos publicados	Cantidad inicial	Cantidad final
Terciario no universitario	2	2
Universitario de grado	36	32
Universitario de posgrado/doctorado	600	530
Universitario de posgrado/especialización	2	1
Universitario de posgrado/maestría	8	7

Datos cargados de Estudios Autores CONICET de todos los artículos publicados descargables	Cantidad inicial	Cantidad final
Terciario no universitario	2	2
Universitario de grado	35	31
Universitario de posgrado/doctorado	567	496
Universitario de posgrado/especialización	2	1
Universitario de posgrado/maestría	7	6

Traducción de títulos, resúmenes y palabras claves: seguidamente se procedió a la traducción utilizando librerías *open source* en consonancia con la perspectiva de los Dispositivos Intermediales Dinámicos. Luego de efectuar una búsqueda exhaustiva, se eligió unos proyectos publicados en Hugging Face, empresa de Estados Unidos que genera y comparte herramientas de Inteligencia Artificial desarrolladas por otros, abiertos al uso de la comunidad¹¹. De esa manera se seleccionaron los siguientes modelos:

- Modelo facebook/nllb-200-distilled-600M, basado en el proyecto NLLB (No Language Left Behind) (Team et al., 2022), que se centra en la traducción automática multilingüe. Desarrollado por Meta AI y centrado en la concepción de un modelo que pueda traducir entre 200 idiomas. Su objetivo es proporcionar un soporte robusto para la traducción de idiomas poco representados en los sistemas de traducción automática, abarcando una amplia gama de lenguas y dialectos. Arquitectura basada en el modelo *distilled* (destilado) que ha sido optimizado para ser más ligero y rápido, mientras mantiene un buen rendimiento en términos de calidad de traducción. Tiene 600 millones de parámetros, por lo cual es menos pesado que algunas de las versiones más grandes, pero efectivo para tareas de traducción.

¹¹ Así es como se comparten modelos preentrenados para diferentes usos, basados en los últimos avances en la Inteligencia Artificial. <https://huggingface.co/>

- Modelo facebook/m2m100_418M de Hugging Face, basado en el proyecto M2M-100 (Many-to-Many Multilingual Machine Translation) (Fan et al., 2020), que es otro modelo de traducción automática multilingüe. Fue desarrollado por Facebook AI y permite traducciones entre varios idiomas sin necesidad de un idioma intermedio. Fue diseñado para realizar traducciones entre 100 idiomas de manera directa, sin recurrir al inglés como lengua intermedia. Su arquitectura está basada en la arquitectura de Transformers, lo que le permite captar relaciones contextuales complejas en el texto para mejorar la calidad de la traducción. Tiene 418 millones de parámetros optimizando el equilibrio entre rendimiento y eficiencia.

Para la determinación de las partes representativas y traducción de los textos de los artículos completos, se siguieron los lineamientos definidos en la metodología para realizar esta tarea. Así, una vez que se tiene la base depurada se procede a trabajar con los textos de los artículos científicos.

Entre las consideraciones previas al proceso se busca la mejor caracterización de los artículos científicos trabajando con el texto para observar diferentes alternativas. Se considera desde no aplicar ningún procesamiento al texto que lo modifique hasta el procesamiento final, lo cual va generando diferentes versiones del documento según los procesamientos parciales con el fin de afinar los desarrollos parciales en pos de una mejor caracterización.

Texto base sin procesar: Es el inicial del artículo luego de la extracción de los archivos PDF y verificación de que tienen información válida.

Se propone ir transformando el texto según una secuencia de modificación de las partes representativas del texto y traducción en base a la metodología en el punto “D”:

Limpieza de bibliografía: cuando se propone obtener una significación semántica del texto, no se piensa en la bibliografía sino en el contenido textual de desarrollo del artículo. Por lo cual se procede en primera instancia a quitarla de acuerdo con el punto **D.1** de la metodología.

Limpieza de parte inicial del artículo: en la base de conocimiento ya se encuentran separados el título, el resumen y las palabras claves, generalmente el inicio del artículo además de tener esta información tiene otro tipo de información que no corresponde a la significación semántica del artículo, como nombres de los autores, mails, afiliaciones y algún otro tipo de información de formato. Por lo cual se propone quitar esta información inicial de acuerdo con el punto **D.2** de la metodología.

Limpieza de maquetado: las diferentes publicaciones de artículos que forman la base de conocimiento corresponden a diversos espacios de publicación científica y cada uno de estos

tiene distintas definiciones de cómo se publican los artículos. Es decir, pueden tener encabezados y pies de páginas con diferente información como nombres de las revistas científicas, números de páginas, información del artículo y autores, entre otras posibilidades. En este sentido, se debe realizar un proceso que detecte y permita la limpieza de este tipo de información de acuerdo con el punto **D.3** de la metodología

Limpieza caracteres especiales y otra información no válida: En el proceso de conversión de archivos pueden aparecer diferentes codificaciones que no representan sentido y deben eliminarse del texto representativo de acuerdo con el punto **D.4** de la metodología

Limpieza idioma no base en el texto: Pueden existir artículos que tengan parte de su contenido en otro lenguaje diferente al principal, generalmente son secciones como resúmenes y palabras claves duplicadas en otro idioma. En esta etapa del proceso se procede a identificarlos y quitarlos del texto base de acuerdo con el punto **D.4** ya mencionado.

Proceso de traducción de artículos en otros idiomas: Luego de los procesos anteriores sobre los artículos que no están etiquetados en idioma base, se los traduce buscando un proceso que permita hacerlo de manera adecuada teniendo en cuenta los últimos avances del PLN en base al punto **D.5** de la metodología.

Como se observa, todos los pasos del proceso son importantes. En primera instancia, el recorte de la bibliografía y la parte inicial del artículo hasta la introducción o capítulo similar del mismo. El primer interrogante que surge es saber dónde está la bibliografía o referencias dentro de un artículo científico. Por convención está al final del artículo y siempre hay un título que a priori lo indica, si bien no hay una norma universal sobre como titular esa parte e incluso otras dentro del artículo. La única manera de saber que otros títulos tiene el contenido bibliográfico, se efectuó abriendo todos los textos posibles y realizando el registro de control dentro del desarrollo. Así, se fue verificando en la mayoría de los casos. Luego de una revisión exhaustiva se pudo registrar una gran cantidad de textos que representan los títulos posibles de esta parte de los artículos científicos.

La posición de este título de la sección de bibliografía permite saber dónde comienza y con ese dato es posible cortar el texto desde el título si se tiene registrada la posición del mismo. Una efectividad cercana al 90 % sería un logro concreto relevante para este proyecto y el objetivo se logró con los procesos descriptos.

Una situación real es que puede haber textos en diferentes idiomas, por lo cual también se debe plantear dividir la búsqueda teniendo en cuenta este aspecto. El objetivo es poder llegar al mayor número de artículos en los diferentes idiomas donde se encuentran esos textos como títulos.

Este mismo problema planteado para bibliografía, se repite en el corte de la parte inicial del texto. El siguiente interrogante que se plantea asociado al anterior es cómo saber dónde se debe hacer el corte de la parte inicial.

Luego de la información inicial de un artículo científico, por ejemplo, título, autores e información relacionada, le sigue por lo general el resumen y las palabras claves. En un alto porcentaje de casos, se presenta el título “Introducción”, situación similar a la de bibliografía. Entonces, se procedió a realizar una revisión pormenorizada de los diferentes títulos posibles que puede llegar a tener. Asimismo, se registraron estos en los diferentes idiomas, logrando una cercanía al 90 % de los títulos posibles.

A diferencia de la bibliografía, en la parte inicial se accede a información adicional que puede colaborar al corte al igual que tener la información de ubicación del título “Resumen” y/o “Palabras claves”. Entonces, en caso de no encontrar el título de la parte inicial introductoria, con los mencionados títulos se puede realizar un corte bastante adecuado.

Otro tema que se presenta en la mayoría de los artículos es que aparecen resúmenes y palabras claves en otros idiomas y pueden aparecer otros títulos en diferentes idiomas, por lo cual se procede a registrarlo también.

Una vez definido todo lo anterior, se procede a registrar la mayor cantidad de posibilidades de cada título de resúmenes, palabras claves, introducción y bibliografía en cada idioma, y los cruces en otros idiomas.

Luego de tener esta información con contadores y registro de ubicación por cada posible palabra en diferentes idiomas se realizará un análisis. En el Anexo A, se muestran las tablas correspondientes a los procesos realizados

Resultados del análisis de los textos

En La detección de título de bibliografía en artículos en idioma castellano se consiguió un 95 %, lo cual es un número bastante adecuado sobre las expectativas buscadas. La muestra fue cercana al total de los artículos realizando una búsqueda minuciosa sobre la variedad de títulos posibles. Con respecto a la parte introductoria se consiguió un 82 % en título de introducción, un 87 % en título de resúmenes y 86% en palabras claves, lo cual permite un corte cercano al doble objetivo de poder realizar un corte de la parte inicial y poder así llegar a un 85 % aproximadamente.

En el idioma inglés las estadísticas resultaron similares. En francés, al registrarse solo tres textos, se logra una efectividad del 100 %. En portugués sobre 20 textos, se logró un 75 % en bibliografía, un 60 % en introducción y un 85% para resúmenes y palabras claves. Cabe mencionar que, no hay estadísticas de catalán e italiano porque no quedaron textos depurados.

Una vez realizado este análisis, se debe tener en cuenta la ubicación de los títulos de las secciones mencionadas. Para ello es necesario considerar el largo de cada artículo y el lugar proporcional en cada texto de los títulos. Así, es posible calcular el valor promedio de los títulos en todos los casos registros. Con esta información se comienza a trabajar un valor porcentual de umbral para cada título posible. En la presente investigación, luego de un número significativo de pruebas se logró definir valores de umbral para cada sección: resúmenes, palabras claves, introducción, bibliografía. En el Anexo B se puede observar la información del proceso de detección de ubicación de diferentes áreas promedios y artículos fuera del umbral.

Resultados del valor promedio de ubicación de títulos

En el caso del título Resumen el valor promedio de ubicación fue 2,8 % y el umbral se definió en el 12 %. La cantidad de textos que quedaron fuera del umbral fueron 22.

En Palabras claves el promedio fue 5,2 % y el umbral definido es del 25 %. La cantidad de textos fuera del umbral fueron 51.

En Introducción el promedio fue 10,34 % y el umbral se fijó en 25 %. La cantidad de textos que quedaron fuera del umbral fueron 10.

En los casos que no hay introducción se realizó la estadística de cuáles serían los mejores cortes, asignando el valor más alto de los posibles, según muestra la siguiente Tabla 22:

Tabla 22

Información importante para el cálculo de corte cuando no hay introducción

Tipo	Idioma	Cantidad
Palabras claves	Español	63
Palabras claves	Inglés	298
Palabras claves	Francés	8
Palabras claves	Portugués	54
Palabras claves	Catalán	0
Palabras claves	Italiano	0
Resumen	Español	24
Resumen	Inglés	32
Resumen	Francés	1
Resumen	Portugués	15
Resumen	Catalán	0
Resumen	Italiano	0
	Total	495

En bibliografía el promedio es 86,3 % y el umbral se fijó en 60 %. La cantidad de textos fuera del umbral fue 25. Estos cálculos de umbrales se lograron ingresando a varios de los artículos científicos que quedaron fuera de los umbrales a los fines de obtener el número más adecuado o cercano al corte erróneo.

Luego, se procedió a la depuración del texto, en diferentes etapas:

A) Texto sin procesar: insumo necesario para las etapas de depuración.

B) Limpieza de bibliografía: una vez obtenidos los datos anteriores, se aplican en un proceso que permite realizar el corte de bibliografía por medio de los títulos detectados, desde el título en adelante (los que están sobre el umbral).

C) Limpieza de parte inicial del artículo: Al igual que en el paso anterior se procede a realizar el corte, desde el inicio del artículo hasta el título de la introducción, tomando solo los que están antes del umbral. En caso de no detectar título de introducción, se utiliza la información anterior registrada de títulos de resúmenes y palabras claves en diferentes idiomas, utilizando el de mayor valor para realizar el corte.

D) Limpieza idioma no base en el texto: se genera un proceso que basado en la metodología de fastText de detección de idiomas, utilizado en pasos anteriores, tomando cada línea del documento. Se detecta el idioma, si no es el base, limpia la línea. Esto se efectúa desde el inicio al fin de cada documento con todos los documentos.

E) Limpieza caracteres especiales y otra información no válida: Divide esta tarea en 3 partes. La primera limpia caracteres especiales que no corresponden a información textual o símbolos asociados a textos. Luego corre varias comprobaciones, por ejemplo:

- ✓ Sacar el guión del medio que divide las palabras por cuestiones de formato impreso al final de las líneas y unir las.
- ✓ Limpiar la cadena (cid: "número") que aparece en varios documentos producto de la transformación del contenido de los PDFs en texto plano.
- ✓ Borrar líneas que tengan solamente un carácter
- ✓ Borrar caracteres asociados a números de páginas
- ✓ Borrar fechas y horas
- ✓ Borrar URLs y mails
- ✓ Borrar datos de ISSN y DOI
- ✓ limpiar espacios y tabulaciones

Varios de estos procesos salen de la verificación visual de una significativa cantidad de documentos de la base de conocimiento. Como resultado de esta verificación visual, se detectó que, durante el proceso de conversión a texto plano de algunos artículos en español, se registraban palabras que no se separan correctamente debido a la falta de una librería que

realice este procedimiento. En primera instancia, se evalúa el impacto de este problema en la muestra. Tras confirmar que no afecta significativamente a la base de conocimiento, se procede a eliminar dichos caracteres. Se toma la decisión de eliminar las palabras largas, estableciendo un límite máximo de 34 caracteres, ligeramente superior a la longitud de la palabra más larga en español.

Finalmente, se eliminan secuencias de texto específicas que forman parte de la información estructural del artículo. También se identifican y eliminan las líneas repetidas, eliminando variabilidad, como los números de página. A continuación, se eliminan la mayoría de las líneas de formato y maquetado propias de los artículos científicos, dependiendo de su lugar de publicación. Además, se suprimen los saltos de línea, dejando el texto de forma continua

F) Limpieza de maquetado: Luego del paso anterior se procede a quitar la estructura del texto del artículo según definiciones de forma de presentación de este en donde sea publicado. Entonces, se sacan los saltos de línea de maquetado para que queden los párrafos separados y se detectan los títulos. También incluye el siguiente proceso para el paso siguiente

Separa oraciones: Este proceso es útil en una de las diversas representaciones del texto, al utilizar la arquitectura de PLN Transformers, que trabaja con secuencias de texto de hasta 512 tokens en los modelos seleccionados. Los tokens están estrechamente relacionados con la cantidad de palabras, incluyendo subpalabras y otros caracteres, y generalmente su cantidad es mayor, pudiendo superar en un 20 a 30 % la cantidad de palabras. Sin embargo, lo importante es que existe un límite máximo para generar las características del texto. Este aspecto se menciona en esta etapa porque la traducción se realiza mediante Transformers entrenados específicamente para esta tarea. Anteriormente, se trabajó con las traducciones de títulos, resúmenes y palabras clave. De estas representaciones, los resúmenes son los más largos en términos de cantidad de palabras, pero no superan el límite máximo necesario para realizar las traducciones.

Es importante señalar que el problema se presenta al trabajar con todo el texto del artículo, por lo que se anticipa desde el principio. Aunque se utilizará para las traducciones, se aprovecha esta etapa para procesar todos los textos, ya que con Transformers también se busca generar una caracterización completa del documento.

Se trabajará en la detección de puntos. Lamentablemente, el punto no solo se utiliza al final de las oraciones, lo que puede generar errores. Para reducir estos problemas, se implementará un control de longitud mínima y máxima de las oraciones durante el proceso. No hay inconveniente en que, con el límite máximo de tokens que pueden ser procesados por Transformers, varias oraciones se agrupen en una sola. El límite mínimo se utiliza para evitar

cortar oraciones por debajo de ese rango, mientras que el máximo se emplea para realizar cortes si no se ha registrado ninguno previamente.

G) Proceso de traducción de artículos en otros idiomas: se utilizan dos modelos de Transformers entrenados para esta tarea. Uno es el facebook/nllb-200-distilled-600M, ya explicado anteriormente, que fue entrenado con textos de hasta 512 tokens. El otro es facebook/mbart-large-50-many-to-many-mmt tiene un máximo de 1024 tokens, que es parte de la familia de modelos mBART (Multilingual BART) (Tang et al., 2020) y se utiliza para la traducción automática multilingüe. Este fue diseñado para realizar traducciones automáticas entre múltiples pares de idiomas (50 idiomas) sin necesidad de un idioma intermedio, lo que mejora la eficiencia y la calidad de las traducciones. Su arquitectura es BART (Bidirectional and Auto-Regressive Transformers), que combina técnicas de codificación y decodificación, permitiendo así un manejo más eficaz de las dependencias contextuales en el texto. Al ser un modelo "large", contiene un gran número de parámetros, lo que le permite captar mejor las complejidades del lenguaje y producir traducciones más precisas.

En la práctica se trabajó con las dos rutinas porque en determinadas líneas de la traducción se observó que se generaban repeticiones extrañas de secuencias de palabras en diferentes líneas y de manera aleatoria. Eso se presentaba en los dos procesos, pero no en las mismas líneas por lo cual cuando se detectaba ese error con uno de los procesos se corría el otro para generar el vector correcto. Así, se pudo reducir la mayoría de los errores. La rutina verificaba el tamaño de cada línea de base y la traducía. Si esta última estaba por debajo o arriba de cierto porcentaje del umbral se realizaba la traducción en el otro modelo y se comparaba nuevamente la cantidad de palabras. En última instancia, la más cercana al tamaño de texto original es la que se guardaba.

4.2.2 Producción Científica

Una vez obtenidos todos los datos depurados, se continúa trabajando con el globalizador "Producción científica" abordando los factores que lo conforman: experiencia, intereses de investigación, aporte de los autores y la integración del control del conflicto de intereses. Cada uno con sus procesos internos a realizar.

Experiencia

Selección de representaciones de contenido textual: luego del preprocesamiento de todas las partes representativas de los artículos científicos se pasa a la extracción de características. Se elige cuál de las partes obtenidas del preprocesamiento se va a usar.

- Título

- Resumen
- Palabras claves
- Artículos científicos en sus diferentes representaciones parciales:
 - Archivos sin procesar - sin traducir
 - Archivos sin biografía - sin traducir
 - Archivos sin biografía y sin introducción - sin traducir
 - Archivos llevados a su idioma, quitando partes de otros idiomas - sin traducir
 - Archivos limpios - sin traducir
 - Archivos limpios y maquetados - sin traducir
 - Archivos limpios, maquetados y traducidos

Cabe mencionar que estas opciones se deben al modo de prueba de la metodología en un primer prototipo para uso de pruebas de laboratorio.

Generación de vectores con técnicas de PLN: se busca generar representaciones del artículo en base a las diferentes partes definidas que pretenden ser más representativas (título, resumen, palabras claves, artículo completo). Las mismas serán generadas con técnicas de PLN para extraer características suficientes para contener el conocimiento que trata cada artículo. Las técnicas utilizadas se han agrupado por tipo.

- Bag of words: Se utiliza TF-IDF, ya que se evaluó el uso del modelo BM-25, pero se descartó su implementación, dado que los parámetros que incluye para la tarea de similitud de texto no son relevantes en este caso. La frecuencia de las palabras está asociada al tamaño del documento, y esta variabilidad se compensa mediante la similitud por coseno, que compara proporciones sin depender de la escala. Por esta misma razón, la longitud del documento no influye significativamente. Además, el modelo BM-25 es más eficaz en tareas de relevancia de consultas dentro de un conjunto de documentos, lo que no se ajusta a los objetivos de este caso.
- Word Embeddings: Se utiliza Word2Vec y fastText. Se desestimó el uso de GloVe porque no se encontraron vectores entrenados en español.
- Document Embeddings: Doc2Vec, aunque se basa en una lógica similar a Word2Vec, se utiliza específicamente para representar documentos en lugar de palabras individuales sin necesidad de modelos preentrenados lo que permitió que sea el modelo más utilizado dentro de este tipo de modelos.
- Transformers: Para generar la representación se debe elegir el tipo de modelo a utilizar. Los modelos de tipo Transformer requieren una cuidadosa

elección para generar representaciones vectoriales (embeddings) efectivas. Analizando las opciones disponibles, es posible afirmar que:

GPT es un modelo autoregresivo optimizado para la generación de texto. Aunque no fue diseñado primariamente para crear embeddings documentales, puede adaptarse a esta tarea mediante: 1) Extracción de representaciones de sus capas internas y 2) Utilización de su capacidad generativa para derivar embeddings.

BERT y sus variantes están específicamente diseñados para tareas de comprensión contextual, lo que los posiciona como opciones superiores para capturar: 1) Significado semántico profundo; 2) Relaciones contextuales a nivel de documento y oración. La principal ventaja de BERT radica en su mecanismo de atención bidireccional, que permite al modelo procesar simultáneamente el contexto anterior y posterior de cada palabra. Esta característica fundamental enriquece significativamente la calidad de los embeddings tanto para oraciones como para documentos completos. En base a estos factores técnicos, la opción más adecuada para la generación de embeddings de alta calidad son los modelos derivados de la arquitectura BERT.

A continuación, se describen de manera general las técnicas de PLN utilizadas para la generación de vectores según las metodologías, ya que es fundamental seleccionar una de ellas. Una característica clave a considerar en las diferentes metodologías es la variación en el proceso de cálculo, especialmente en cuanto a la necesidad de contar con todos los artículos para ser representados en conjunto o si es necesario hacerlo de manera individual para cada tipo de dato a representar.

Los modelos TF-IDF y Doc2vec necesitan hacer el cálculo incluyendo todos los artículos que forman la base de conocimiento. Por lo cual, si se va a realizar una comparativa para un nuevo artículo a ser revisado, se tiene que incluir este y todos los artículos de todos los candidatos a revisores para efectuar el cálculo. A diferencia de los otros métodos que no necesitan hacer este muestreo total para generar el vector de cada documento.

Dentro del cálculo de las diferentes metodologías existen ciertas particularidades. Excepto en las metodologías asociadas con Transformers, que necesitan el contexto de las palabras para ir generando el vector de documento, se puede optar por activar la remoción de stop words o palabras vacías. Asimismo, entre la posibilidad de usar stemming o lematización, se considera más adecuada el uso de esta última al convertir cada palabra a su lema base. Con esto se evita que queden palabras inexistentes como sucede al aplicar stemming.

Por lo tanto, hay 4 opciones:

- ✓ texto sin conversión (sin *stop words*, sin lematización).
- ✓ texto con limpieza de *stop words* (sin lematización).
- ✓ texto con lematización (sin *stop words*).
- ✓ texto con limpieza de *stop words* y lematización.

Es importante aclarar que estas metodologías llamadas *bag of words* no tienen información semántica y se basan en estructurar el texto en tokens o palabras representadas por un vector *one hot encodings*, con un “1” en determinada posición diferente a las demás y todos los demás “0”. Entonces, la última opción de disminuir el tamaño de la muestra con *stop words* y lematización, es altamente recomendada porque no afecta a la significación semántica y acelera los tiempos de procesamiento.

Las opciones que ofrece el prototipo son:

TF-IDF

Word2Vec: Dentro de esta metodología se utilizan los vectores entrenados sobre grandes corpus de textos. Y se tienen en cuenta dos posibilidades en cuanto a las operaciones con vectores:

- Promedio
- Suma

Doc2vec

fastText: Se usan los vectores entrenados del mismo proyecto y se utiliza un método de la librería para hacer el cálculo. Se tienen en cuenta dos posibilidades en cuanto a las operaciones con vectores:

- Promedio
- Suma

Transformers: Se utilizaron diferentes soluciones basadas en el proyecto Sentence Transformers (también conocido como SBERT que es la abreviatura de Sentece-BERT) (Reimers & Gurevych, 2019) que puede ser utilizado para encontrar el grado de similitud entre diferentes textos. Es una modificación del transformer BERT que mantiene la misma precisión que este, pero reduce sustancialmente los tiempos de procesamiento. Se generan vectores a nivel de oración y en base a ellos, con su contribución relativa al documento, se combinan para generar el vector completo del texto.

Para la generación de vectores se utilizaron dos proyectos. El motivo de esta decisión fue evaluar cada uno para detectar el mejor entrenado para la tarea. Cabe esta aclaración porque en “Hugging Face” se pueden utilizar un significativo número de modelos disponibles para la comunidad. Entre los disponibles se requiere que puedan trabajar una cierta longitud de

tokens. En esta dirección, se definió la tarea en un mínimo de 512 tokens para poder incluir oraciones largas (aproximadamente 400 palabras).

Otras alternativas pueden ser Longformer y Big Bird, ya que pueden trabajar con secuencias largas de tokens, pudiendo con el último trabajar hasta con textos completos. Sin embargo, al existir modelos de Sentence Vectors que pueden trabajar con 512 tokens, y teniendo en consideración la longitud de los textos, no fue necesario utilizar ese tipo de soluciones. Los modelos elegidos son:

1) El proyecto **mrm8488/multilingual-e5-large-ft-sts-spanish-matryoshka-768-16-5e**, de Sentence Transformers diseñado para tareas de similaridad semántica y representación de texto multilingüe. Este modelo está enfocado en el español y se basa en la arquitectura E5 (Wang et al., 2024) que se ha diseñado para la generación de *embeddings* de texto. E5 se centra en proporcionar representaciones de alta calidad que capturan las relaciones semánticas entre textos. Tiene un enfoque multilingüe, aunque está optimizado para el español. Además, está ajustado finamente (*fine-tuned*) utilizando el conjunto de datos STS (Semantic Textual Similarity) en español, lo que mejora su desempeño en tareas de similitud semántica. El mismo, genera embeddings de 1024 dimensiones, lo que permite una representación rica y detallada del texto. Está entrenado para entender y generar texto, así como para calcular la similitud entre diferentes fragmentos de texto, lo que lo hace ideal para tareas de clasificación, recuperación de información y análisis semántico. Trabaja sobre textos con una longitud máxima de 512 tokens. En adelante será denominado como “Transformer 1”.

2) El modelo **hiiamsid/sentence_similarity_spanish_es** (hiiamsid, s. f.) está diseñado específicamente para evaluar la similitud semántica entre oraciones en español. Se basa en arquitecturas de transformers como BERT, que están bien adaptadas para tareas de comprensión del lenguaje. El modelo ha sido entrenado utilizando un conjunto de datos que incluye ejemplos de pares de oraciones junto con etiquetas que indican su similitud. Esto le permite aprender a identificar relaciones semánticas entre textos. El vector resultante tiene 768 dimensiones. Al igual que el modelo anterior, trabaja sobre textos con una longitud máxima de 512 tokens. “Transformer 2” será denominado en adelante.

Por cada uno de los dos modelos elegidos, se tienen en cuenta dos posibilidades en cuanto a las operaciones con vectores:

- Promedio
- Suma

A continuación, se abordará la integración de diferentes técnicas de PLN y partes representativas del artículo científico para generar vectores.

Si se efectúa un recorte de las diferentes partes de representación del artículo y las diferentes técnicas tratadas en general. Se obtiene una variabilidad de 36 opciones, como muestra la Tabla 23.

Tabla 23

Técnicas y metodologías de representación de PLN por cada parte de texto

	TF-IDF	Word2Vec	Doc2vec	fastText	Transformer 1	Transformer 2
Título						
Resumen						
Palabras claves						
Título y resumen						
Título, resumen y palabras claves						
Texto artículo procesado						

Cabe aclarar que estas opciones son un recorte de la variabilidad de posibilidades de trabajo dentro de cada metodología. Si se toman las diferentes posibilidades descritas en cada técnica de PLN aplicada, se pueden generar las siguientes representaciones vectoriales de los mismos, según la Tabla 24.

En total son 24 diferentes representaciones por cada parte representativa. Sin embargo, considerando la recomendación de utilización de limpieza de *stop words* y lematización como más adecuada excepto en Transformers, la cantidad de representaciones disminuye resulta 10. En lo que respecta a las 3 últimas filas que son suma de vectores, promedio de vectores y vectores, se realizan a modo de prueba los cálculos de sumas y promedio. A nivel teórico (corroborado más adelante en la práctica en la comparación de diferentes tipos de representaciones en las diferentes técnicas) los valores deberían ser iguales. El promedio y la suma se realizaron en Word2Vec, fastText y los dos tipos de Transformers. Por lo cual la cantidad de representaciones disminuye a 6

Tabla 24

Preprocesamientos, cálculos y tipo de representación según diferentes técnicas y modelos de PLN

Representación de textos	Modelos					
	TF-IDF	Word2Vec	Doc2Vec	fastText	Transformer 1	Transformer 2
sin procesar	X	X	X	X	X	X
Con limpieza de stop words	X	X	X	X		
con lematización	X	X	X	X		
con limpieza de stop words y lematización	X	X	X	X		
suma vectores		X		X	X	X
promedio vectores		X		X	X	X
Vectores	X		X			

De lo explicitado se deduce que por cada parte representativa se accede un máximo de 28 opciones, lo que daría 168 opciones. Al tomar 6 opciones en la simplificación, resultan 36 opciones.

Cabe aclarar que se realiza un recorte en el contenido textual del artículo científico debido al amplio abanico de representaciones textuales (7 guardadas en diferentes etapas), lo cual responde al proceso parcializado de depuración, que es principalmente parte de la evaluación de los resultados de cada etapa. Por lo tanto, se considera que, en primera instancia, la representación más adecuada para realizar la comparación es la última, que corresponde a los archivos limpios, maquetados y traducidos. No obstante, en el prototipo se incluye la opción de realizar la comparación en cada etapa del procesamiento de archivos.

Seguidamente, se realiza una comparación sobre el vector resultante del manuscrito y las producciones de los revisores, el cual se construye a partir de la parte textual más representativa de los artículos científicos. Tras aplicar la técnica de procesamiento de lenguaje natural (PLN) seleccionada, se empleará la medida de similitud de coseno, que evalúa la similitud entre los diferentes vectores resultantes midiendo el coseno del ángulo entre dos vectores en un espacio multidimensional. Esta medida varía entre 1 y -1, donde un valor de 1 indica una coincidencia exacta, mientras que -1 refleja una discrepancia total. Por lo tanto, los artículos de los candidatos cuyo valor de similitud esté más cercano a 1 serán los considerados más similares al manuscrito en cuanto al conocimiento que abordan. Posteriormente, los valores de similitud

se ordenarán de manera descendente y se aplica el punto de corte elegido para seleccionar los candidatos a revisores en orden descendente hasta ese valor.

Luego de este paso se pueden realizar los laboratorios de comparativas de partes representativas de artículos científico y el análisis de técnicas de PLN con evaluación de expertos.

4.3 Laboratorio de comparativas de partes representativas de artículos científicos

Antes del análisis siguiente, se llevaron a cabo diversas pruebas de laboratorio, que incluyeron diferentes alternativas de cálculo de vectores utilizando distintos modelos. En el contexto de esta tesis doctoral, se desea resaltar específicamente la prueba de suma y promedio de vectores en las técnicas de *word embeddings*, la cual arrojó los mismos valores en la similitud por coseno. Este resultado, que desde el punto de vista teórico se esperaba, fue confirmado mediante pruebas prácticas como parte del control del procedimiento de cálculo. Tras realizar estas pruebas, se decidió adoptar el promedio como método a considerar, dado que ofrece los mismos resultados que la suma y garantiza que el valor resultante se mantenga dentro de un rango más controlado.

Del análisis de las similitudes de las partes representativas del artículo científico entre título, resúmenes, palabras claves y texto completo entre introducción y hasta bibliografía sin incluir, tomando como centro está última, la más cercana en similitud de texto es el resumen con un promedio de casos del 95,71 % entre las 6 pruebas con diferentes modelos de PLN, siendo el valor más bajo 89,76 % en TF-IDF y 99,7 % en Word2Vec y fastText. Como puede observarse en la tabla 25.

Para identificar si esa relación cercana se da en sentido inverso, se realizó la comparación con centro en el resumen. El resultado arrojó que tanto el resumen y el cuerpo son más cercanos en ambos sentidos. Para las pruebas en las metodologías que lo permitían se utilizaron las técnicas de stop words y lematización. Como puede observarse en la tabla 26.

Tabla 25

Comparativa entre partes representativas del artículo científico con centro en texto completo, denominado en la tabla como archivo

TF-IDF (con stop words y con lematización)

Partes de textos						
dato 1	dato 2	dato 3	dato 4	Totales	Totales agrupados	Porcentaje agrupados
texto completo	palabras claves	resumen	título	159	196	5,85%
texto completo	palabras claves	título	resumen	37		
texto completo	resumen	palabras claves	título	1321	3005	89,76%
texto completo	resumen	título	palabras claves	1684		
texto completo	título	palabras claves	resumen	27	147	4,39%
texto completo	título	resumen	palabras claves	120		
						100,00%

Word2Vec (promedio y suma) (con stop words, con lematización)

Partes de textos						
dato 1	dato 2	dato 3	dato 4	Totales	Totales agrupados	Porcentaje agrupados
texto completo	palabras claves	resumen	título	1	2	0,06%
texto completo	palabras claves	título	resumen	1		
texto completo	resumen	palabras claves	título	693	3338	99,70%
texto completo	resumen	título	palabras claves	2645		
texto completo	título	palabras claves	resumen	5	8	0,24%
texto completo	título	resumen	palabras claves	3		
						100,00%

Doc2Vec (con stop words, con lematización)

Partes de textos						
dato 1	dato 2	dato 3	dato 4	Totales	Totales agrupados	Porcentaje agrupados
texto completo	palabras claves	resumen	título	18	44	1,31%
texto completo	palabras claves	título	resumen	26		
texto completo	resumen	palabras claves	título	887	3216	96,06%
texto completo	resumen	título	palabras claves	2329		
texto completo	título	palabras claves	resumen	17	88	2,63%
texto completo	título	resumen	palabras claves	71		
						100,00%

fastText (promedio y suma) (con stop words y con lematización)

Partes de textos						
dato 1	dato 2	dato 3	dato 4	Totales	Totales agrupados	Porcentaje agrupados
texto completo	palabras claves	resumen	título	1	2	0,06%
texto completo	palabras claves	título	resumen	1		
texto completo	resumen	palabras claves	título	655	3338	99,70%
texto completo	resumen	título	palabras claves	2683		
texto completo	título	palabras claves	resumen	4	8	0,24%
texto completo	título	resumen	palabras claves	4		
						100,00%

Transformers 1 (promedio y suma)

Partes de textos						
dato 1	dato 2	dato 3	dato 4	Totales	Totales agrupados	Porcentaje agrupados
texto completo	palabras claves	resumen	título	76	92	2,75%
texto completo	palabras claves	título	resumen	16		
texto completo	resumen	palabras claves	título	1289	3113	92,98%
texto completo	resumen	título	palabras claves	1824		
texto completo	título	palabras_claves	resumen	17	143	4,27%
texto completo	título	resumen	palabras claves	126		
						100,00%

Transformers 2 (promedio y suma)

Partes de textos						
dato 1	dato 2	dato 3	dato 4	Totales	Totales agrupados	Porcentaje agrupados
texto completo	palabras claves	resumen	título	7	10	0,30%
texto completo	palabras claves	título	resumen	3		
texto completo	resumen	palabras claves	título	847	3320	99,16%
texto completo	resumen	título	palabras claves	2473		
texto completo	título	palabras claves	resumen	2	18	0,54%
texto completo	título	resumen	palabras claves	16		
						100,00%

Tabla 26

Comparativa entre partes representativas del artículo científico centrado en resumen

TF-IDF (con stop words, con lematización)

Partes de textos						
dato1	dato2	dato3	dato4	Totales	Totales agrupados	Porcentaje agrupados
resumen	texto completo	palabras claves	título	1090	2679	80,02%
resumen	texto completo	título	palabras claves	1589		
resumen	palabras claves	texto completo	título	208	275	8,21%
resumen	palabras claves	título	texto completo	67		
resumen	título	texto completo	palabras claves	305	394	11,77%
resumen	título	palabras claves	texto completo	89		
						100,00%

Word2Vec (promedio y suma) (con stop words, con lematización)

Partes de textos						
dato1	dato2	dato3	dato4	Totales	Totales agrupados	Porcentaje agrupados
resumen	texto completo	palabras claves	título	738	3341	99,79%
resumen	texto completo	título	palabras claves	2603		
resumen	palabras claves	texto completo	título	1	1	0,03%
resumen	título	texto completo	palabras claves	5	6	0,18%
resumen	título	palabras claves	texto completo	1		
						100,00%

Doc2Vec (con stop words, con lematización)

Partes de textos						
dato1	dato2	dato3	dato4	Totales	Totales agrupados	Porcentaje agrupados
resumen	texto completo	palabras claves	título	697	3152	94,15%
resumen	texto completo	título	palabras claves	2455		
resumen	palabras claves	texto completo	título	9	26	0,78%
resumen	palabras claves	título	texto completo	17		
resumen	título	texto completo	palabras claves	120	170	5,08%
resumen	título	palabras claves	texto completo	50		
						100,00%

fastText (promedio y suma) (con stop words, con lematización)

Partes de textos						
dato1	dato2	dato3	dato4	Totales	Totales agrupados	Porcentaje agrupados
resumen	texto completo	palabras claves	título	697	3329	99,43%
resumen	texto completo	título	palabras claves	2632		
resumen	palabras claves	texto completo	título	3	4	0,12%
resumen	palabras claves	titulo	texto completo	1		
resumen	título	texto completo	palabras claves	13	15	0,45%
resumen	título	palabras claves	texto completo	2		
						100,00%

Transformers 1 (promedio y suma)

Partes de textos						
dato1	dato2	dato3	dato4	Totales	Totales agrupados	Porcentaje agrupados
resumen	texto completo	palabras claves	título	779	2957	88,32%
resumen	texto completo	título	palabras claves	2178		
resumen	palabras claves	texto completo	título	21	35	1,05%
resumen	palabras claves	título	texto completo	14		
resumen	título	texto completo	palabras claves	330	356	10,63%
resumen	título	palabras claves	texto completo	26		
						100,00%

Transformers 2 (promedio y suma)

Partes de textos						
dato1	dato2	dato3	dato4	Totales	Totales agrupados	Porcentaje agrupados
resumen	texto completo	palabras claves	título	622	3307	98,78%
resumen	texto completo	título	palabras claves	2685		
resumen	palabras claves	texto completo	título	3	4	0,12%
resumen	palabras claves	título	texto completo	1		
resumen	título	texto completo	palabras claves	20	37	1,11%
resumen	título	palabras claves	texto completo	17		
						100,00%

Del mapeo sistemático se desprende que la mayoría de los artículos que explicitaron que parte representativa tomaron de base, usaron conjuntamente título y resúmenes. Se comparó en base al contenido textual del artículo agregando las opciones de título, palabras claves y resúmenes en conjunto, además de resúmenes solo. Como resultado se obtuvo que el comportamiento es diferente en las diferentes metodologías. En TF-IDF y Doc2Vec arrojó mejores resultados el conjunto de título con resumen y palabras claves, siendo los porcentajes 76,61 % y 41,40 % respectivamente, en este último resumen obtuvo 37,34 %. Y en todos los otros casos resumen dio los mejores resultados con porcentajes entre 61,53 % y 83,75 % con amplia diferencia con respecto a los otros modelos. Como puede observarse en la tabla 27.

Tabla 27

Comparativa entre partes representativas del artículo científico con centro en texto completo

TF-IDF (con stop words, con lematización)

Partes de textos						
dato 1	dato 2	dato 3	dato 4	Totales	Totales agrupados	Porcentaje agrupados
texto completo	resumen	resumen + título	resumen + título + palabras claves	130	173	5,17%
texto completo	resumen	resumen + título + palabras claves	resumen + título	43		
texto completo	resumen + título	resumen	resumen + título + palabras claves	152	610	18,22%
texto completo	resumen + título	resumen + título + palabras claves	resumen	458		
texto completo	resumen + título + palabras claves	resumen	resumen + título	172	2565	76,61%
texto completo	resumen + título + palabras claves	resumen + título	resumen	2393		
						100,00%

Word2Vec (promedio y suma) (con stop words, con lematización)

Partes de textos						
dato 1	dato 2	dato 3	dato 4	Totales	Totales agrupados	Porcentaje agrupados
texto completo	resumen	resumen + título	resumen + título + palabras claves	1833	2060	61,53%
texto completo	resumen	resumen + título + palabras claves	resumen + título	227		
texto completo	resumen + título	resumen	resumen + título + palabras claves	468	817	24,40%
texto completo	resumen + título	resumen + título + palabras claves	resumen	349		
texto completo	resumen + título + palabras claves	resumen	resumen + título	91	471	14,07%
texto completo	resumen + título + palabras claves	resumen + título	resumen	380		
						100,00%

Doc2Vec (con stop words, con lematización)

Partes de textos						
dato 1	dato 2	dato 3	dato 4	Totales	Totales agrupados	Porcentaje agrupados
texto completo	resumen + título + palabras claves	resumen + título	resumen	974	1386	41,40%
texto completo	resumen + título + palabras claves	resumen	resumen + título	412		
texto completo	resumen + título	resumen + título + palabras claves	resumen	458	712	21,27%
texto completo	resumen + título	resumen	resumen + título + palabras claves	254		
texto completo	resumen	resumen + título + palabras claves	resumen + título	510	1250	37,34%
texto completo	resumen	resumen + título	resumen + título + palabras claves	740		
						100,00%

fastText (promedio y suma) (con stop words, con lematización)

Partes de textos						
dato 1	dato 2	dato 3	dato 4	Totales	Totales agrupados	Porcentaje agrupados
texto completo	resumen	resumen + título	resumen + título + palabras claves	1892	2060	61,53%
texto completo	resumen	resumen + título + palabras claves	resumen + título	168		
texto completo	resumen + título	resumen	resumen + título + palabras claves	421	774	23,12%
texto completo	resumen + título	resumen + título + palabras claves	resumen	353		
texto completo	resumen + título + palabras claves	resumen	resumen + título	92	514	15,35%
texto completo	resumen + título + palabras claves	resumen + título	resumen	422		
						100,00%

Transformers 1 (promedio y suma)

Partes de textos						
dato 1	dato 2	dato 3	dato 4	Totales	Totales agrupados	Porcentaje agrupados
texto completo	resumen	resumen + título	resumen + título + palabras claves	2263	2804	83,75%
texto completo	resumen	resumen + título + palabras claves	resumen + título	541		
texto completo	resumen + título	resumen	resumen + título + palabras claves	123	276	8,24%
texto completo	resumen + título	resumen + título + palabras claves	resumen	153		
texto completo	resumen + título + palabras claves	resumen	resumen + título	94	268	8,00%
texto completo	resumen + título + palabras claves	resumen + título	resumen	174		
						100,00%

Transformers 2 (promedio y suma)

Partes de textos						
dato 1	dato 2	dato 3	dato 4	Totales	Totales agrupados	Porcentaje agrupados
texto completo	resumen	resumen + título	resumen + título + palabras claves	1417	2118	63,26%
texto completo	resumen	resumen + título + palabras claves	resumen + título	701		
texto completo	resumen + título	resumen	resumen + título + palabras claves	233	428	12,78%
texto completo	resumen + título	resumen + título + palabras claves	resumen	195		
texto completo	resumen + título + palabras claves	resumen	resumen + título	265	802	23,95%
texto completo	resumen + título + palabras claves	resumen + título	resumen	537		
						100,00%

En vistas a los resultados reportados, se considera que la mejor representación es el texto completo desde la introducción hasta bibliografía y debe ser utilizada en caso de estar disponible. Dentro de las otras representaciones, la más cercana es resumen. Se reconoce una variabilidad significativa de título, resumen y palabras claves en conjunto en TF-IDF, no siendo tan significativa y bastante similar al resumen en Doc2Vec.

4.4 Análisis de técnicas de PLN y evaluación de expertos

A partir de los resultados alcanzados en el Anexo B, se asume que el texto completo es de mayor utilidad que el resumen, se evaluó cuál es la mejor técnica de PLN a incluir en la metodología final con la participación de expertos.

Dado que las métricas no captan las características cualitativas complejas necesarias, desde un enfoque socio-técnico cultural se sostiene que los expertos son los únicos que cuentan

con el conocimiento profundo y contextualizado para determinar con precisión que técnica de PLN está realizando mejor la tarea o se adecúa a los resultados esperados. Cabe considerar que como puede suceder en cualquier otro caso, para evitar en lo posible el riesgo del sesgo de un experto en particular, la participación de un grupo es imprescindible para lograr un resultado mas confiable. En esta dirección se convocaron miembros del IRICE, investigadores y becarios posdoctorales pertenecientes a las distintas líneas de investigación del instituto.

A los mismos se les solicitó que seleccionen tres artículos científicos que no sean de su autoría y que sean cercanos a sus temáticas de estudio dentro la base de conocimientos del RI con la restricción de que estén clasificados en el área de Educación y dentro del período de extracción de la misma. De dicha selección se evaluó que estén escritos en castellano para evitar procesos intermedios de traducción de texto y que en el preprocesamiento tengan lo más correctamente posible un buen corte entre la introducción y la bibliografía sin incluir.

Si se cumplía con estos requisitos se elegían aleatoriamente cualquiera de los elegidos. Seguidamente se muestra en la figura 8 uno de los informes enviado a los expertos

	Id	Método Columna 1	Id	Método Columna 2	Id	Método Columna 3	Id	Método Columna 4	Id	Método Columna 5	Id	Método Columna 6
	222607	Experiencia educativa en virtualidad: percepciones de futuros formadores sobre una agencia en construcción	222607	Experiencia educativa en virtualidad: percepciones de futuros formadores sobre una agencia en construcción	222607	Experiencia educativa en virtualidad: percepciones de futuros formadores sobre una agencia en construcción	222607	Experiencia educativa en virtualidad: percepciones de futuros formadores sobre una agencia en construcción	222607	Experiencia educativa en virtualidad: percepciones de futuros formadores sobre una agencia en construcción	222607	Experiencia educativa en virtualidad: percepciones de futuros formadores sobre una agencia en construcción
1	184604	El compromiso académico estudiantil presente en una educación expandida por el uso de las TIC	211281	Compromiso docente en contextos virtuales. Un estudio en el nivel medio de educación	184604	El compromiso académico estudiantil presente en una educación expandida por el uso de las TIC	184604	El compromiso académico estudiantil presente en una educación expandida por el uso de las TIC	226596	Los recursos digitales como herramientas para repensar la agencia docente en futuros formadores	208093	Mitos acerca de la Educación Inclusiva y la Pandemia: Un Abordaje en la Formación Continua: ¿Barreras y Oportunidades para la Enseñanza?
2	226596	Los recursos digitales como herramientas para repensar la agencia docente en futuros formadores	226596	Los recursos digitales como herramientas para repensar la agencia docente en futuros formadores	226596	Los recursos digitales como herramientas para repensar la agencia docente en futuros formadores	164942	Residencia: Los nuevos trayectos de las prácticas profesionales en el mustado contexto del distanciamiento social	184604	El compromiso académico estudiantil presente en una educación expandida por el uso de las TIC	209651	Compromiso académico en contextos virtuales e híbridos. Adaptabilidad frente al cambio en el ámbito universitario
3	226603	Experiencias educativas que construyen: Compromiso académico en la formación docente universitaria	23450	Educación, Tecnologías y Medios de Comunicación: nuevos modos de interacción a los docentes en contextos escolares	226603	Experiencias educativas que construyen: Compromiso académico en la formación docente universitaria	194366	Tareas grupales en ambientes virtuales. Dificultades percibidas y aprendizajes logrados por estudiantes universitarios	138951	Clase invertida, formación docente y agencia transformadora: un estudio preliminar en pandemia con estudiantes argentinos	226596	Los recursos digitales como herramientas para repensar la agencia docente en futuros formadores

Figura 8

Recorte de un informe enviado a los expertos

Cada experto convocado recibió un reporte basado en el informe del artículo seleccionado, en el cual se eligieron los diez artículos más relevantes de la base de conocimiento en cada técnica de PLN aplicada a los textos completos. Es importante destacar que las metodologías se presentaron de manera aleatoria en cada reporte, sin especificar qué técnica se utilizaba, con el fin de evitar posibles sesgos derivados del conocimiento previo de los participantes sobre la temática. La tarea consistía en evaluar, en una escala del 1 al 6, cada una

de las columnas correspondientes a las técnicas, siendo 1 la más precisa y 6 la menos precisa en la detección de similitudes.

La devolución de los expertos fue muy significativa para esta tesis. Se procedió a puntuar de forma inversa cada uno de los métodos (1 como 6, 2 como 5 y así sucesivamente) y elaborar una tabla con suma, promedio y porcentaje. Tomando en cuenta los informes de los 8 expertos, el valor más alto fue obtenido por Transformer 1 con 4,8 en promedio seguido de Doc2Vec y TF-IDF, ambos con 4. Estos 3 sobre el valor promedio de 3,5. Las otras técnicas en orden descendente fueron Transformer 2, fastText y Word2Vec. Transformers 1 se mantuvo entre las 3 primeras opciones salvo una excepción que fue elegida como la última en priorización. El detalle se puede observar en la tabla 28.

Tabla 28

Comparativas entre partes representativas del artículo científico con centro en archivo.

Marcados en amarillo los que obtuvieron más puntaje por cada experto

	TF-IDF	Word2Vec	Doc2Vec	fastText	Transf. 1	Transf. 2
Experto 1	3	4	5	6	1	2
Experto 2	2	5	4	1	6	3
Experto 3	6	2	3	1	4	5
Experto 4	6	1	4	3	5	2
Experto 5	6	4	2	1	5	3
Experto 6	1	2	4	3	6	5
Experto 7	3	2	6	1	5	4
Experto 8	4	1	6	3	5	2
Experto 9	5	2	3	1	6	4
Experto 10	4	1	3	6	5	2
Total	40	24	40	26	48	32
Promedio	4	2,4	4	2,6	4,8	3,2

Los participantes de la tarea coincidieron en que existe una relación significativa entre el artículo a revisar y las priorizaciones, en un alto porcentaje. Algunos mencionaron que les tocó revisar artículos que se alejaban del tema específico de investigación, al observar la relación entre los artículos priorizados y el artículo seleccionado. Con la ayuda de los expertos, se pudo identificar que, cuando un artículo abarca múltiples temas, y no existen en la base de conocimiento artículos que combinen exactamente esos mismos temas, se priorizan aquellos con temas más generales que aparecen en el artículo elegido. Otro aspecto relacionado con esto es que varios de los miembros de IRICE pertenecen al área de Psicología del desarrollo, y muchas de sus publicaciones y temas de interés se encuentran más clasificados bajo Psicología que bajo Educación en el RI.

La cantidad de expertos consultados mitigó la posibilidad de que exista algún sesgo en la priorización de algún experto en particular, permitiendo el promedio acomodar valores con cierta distorsión.

Aunque se observa una ventaja clara de Transformer 1 sobre las demás técnicas, también se percibe una distribución relativamente homogénea de puntuaciones, con un rendimiento inferior en las técnicas directas basadas en *word embeddings*. Resultó significativo el buen desempeño de TF-IDF, a pesar de ser una técnica que no incorpora información semántica. El modelo Transformer 2 tuvo un rendimiento por debajo de lo esperado, lo que llevó a incorporar dos variantes de Transformers para evaluar su desempeño en comparación con los demás modelos, y así seleccionar el más adecuado. Transformer 1 y TF-IDF fueron elegidas las mejores técnicas 3 veces, Doc2Vec y fastText 2 veces. Llama la atención este último, debido a que el rendimiento global no fue tan bueno como el de las otras 3 técnicas. Con base en el análisis de las distintas técnicas, se recomienda utilizar la técnica de PLN Transformer 1 para trabajar con el texto del artículo, especialmente en el área de educación.

La evidencia del comportamiento distinto de los dos modelos Transformer confirma el acierto de incluir al menos dos modelos y evaluar su rendimiento en la tarea específica.

Transformer 1 presenta un rendimiento superior en tareas de similitud semántica textual en español, lo que probablemente se debe a una optimización más efectiva durante su proceso de *fine-tuning*. Este mejor desempeño podría atribuirse a: 1) La incorporación de un corpus más amplio y diverso de textos en español durante su entrenamiento; 2) Una exposición a mayor variedad de construcciones lingüísticas y contextos culturales propios del idioma y 3) Un ajuste más preciso de los parámetros específicos para captar los matices semánticos del español. Esta robusta base de entrenamiento permite que el modelo comprenda con mayor precisión las relaciones semánticas entre textos en español, resultando en evaluaciones de similitud más acordes con el juicio humano y un mejor rendimiento general en tareas de procesamiento de lenguaje natural específicas para este idioma. Por todo lo expuesto, la propuesta de solución se considera válida, como fue confirmado por los expertos.

Cabe destacar que cada técnica tiene su propio enfoque, lo que implica limitaciones inherentes. Por esta razón, el rol de los expertos es fundamental para seleccionar la técnica más adecuada para cada caso específico. No obstante, el hecho de que muchas técnicas coincidan al identificar artículos similares sugiere que se están detectando similitudes relevantes, independientemente del enfoque utilizado. En consecuencia, se puede inferir que el grado de coincidencia entre las técnicas indica la cercanía de la solución: más coincidencias reflejan mayor cercanía, mientras que menos coincidencias indican una priorización de menor calidad.

Para corroborar esto, en la Tabla 29 se pueden observar los artículos no repetidos de cada técnica de PLN en las hojas de evaluación de cada experto, destacando la técnica que

obtuvo los mejores resultados y la cantidad de artículos repetidos en la priorización seleccionada por cada participante. En la mayoría de los casos, la solución se aproxima al promedio de la cantidad de artículos no repetidos, ubicándose también en los valores extremos, tanto en la menor como en la mayor cantidad. El valor promedio se encuentra cercano al centro de los valores registrados, lo que sugiere que puede haber un indicio claro de la mejor metodología del conjunto a partir de estos datos. De manera similar a cómo se analizó la base de artículos no repetidos, también es posible realizar un análisis sobre los artículos repetidos, obteniendo un resultado similar, pero en sentido opuesto.

Tabla 29

Comparativas entre artículos no repetidos entre textos completos y asociación a mejor metodología

Experto	Metod. 1ra.	artículos no repetidos en textos completos						Suma	Promedio	ubicación del primero en el conjunto
		TF-IDF	Doc2Vec	Word2Vec	fastText	Transformer 1	Transformer 2			
Experto 1	fastText	1	1	3	0	3	0	8	1,33	-
Experto 2	Transf. 1	5	4	1	4	3	3	20	3,33	M -
Experto 3	TF-IDF	3	6	3	1	1	3	17	2,83	M +
Experto 4	TF-IDF	4	3	0	0	3	1	11	1,83	+
Experto 5	TF-IDF	3	4	1	2	4	1	15	2,50	M +
Experto 6	Transf. 1	6	3	1	1	1	1	13	2,17	-
Experto 7	Doc2Vec	3	3	3	5	3	2	19	3,17	M -
Experto 8	Doc2Vec	3	3	0	0	1	2	9	1,50	+
Experto 9	Transf. 1	3	2	1	1	2	1	10	1,67	M +
Experto 10	fastText	2	0	1	1	1	0	5	0,83	M +
SUMA		33	29	14	15	22	14	127	21,17	
PROMEDIO		3,30	2,90	1,40	1,50	2,20	1,40	12,70	2,12	
PROMEDIO 1ras. Metodologías									2,3	

“-” Menor valor del conjunto

“+” Mayor valor del conjunto

“M” Valor promedio con “+” o “-” según esté más cerca del mayor o menor valor del conjunto

En las pruebas de laboratorio y en lo documentado en los informes de expertos se detectó que tomando los textos completos de los artículos y comparando todas las técnicas incluidas existe gran convergencia entre las diferentes técnicas de PLN en la priorización de textos, evidenciándose gran cantidad de repeticiones de artículos entre las diferentes técnicas,

y una disminución cuando se incluían los resúmenes y otras combinaciones que incluyeran a estos. Al analizar las priorizaciones entregadas a los expertos en comparación con las mismas priorizaciones, pero basadas en los resúmenes, se observa lo siguiente:

En base a la tabla 30, hay un crecimiento del 94,49 % en la cantidad de artículos no repetidos que aumenta del 23,52 % en textos de artículos al 45,74 % en resúmenes. Existe una excepción en las priorizaciones generadas para el experto 7 que genera mayor convergencia con el resumen resaltado en color amarillo en la tabla.

Hasta este punto se compararon los textos completos con los resúmenes. Sin embargo, basándose en las pruebas de similitud entre las diferentes representaciones del artículo (utilizando los textos completos como referencia), se determinó que tanto en TF-IDF como en Doc2Vec la mejor representación correspondía a la combinación de títulos + resúmenes + palabras clave.

Tabla 30

Comparativas entre artículos no repetidos entre textos completos de artículos y resúmenes

Experto	artículos no repetidos en textos completos	artículos no repetidos en resúmenes
Experto 1	8	17
Experto 2	20	31
Experto 3	17	21
Experto 4	11	34
Experto 5	15	28
Experto 6	13	32
Experto 7	19	10
Experto 8	9	25
Experto 9	10	37
Experto 10	5	12
Total artículos no repetidos	127	247
representa	23,52%	45,74%
expansión no repetidos		94,49%
Total artículos repetidos	473	353
representa	87,59%	65,37%
retracción		25,37%

Por consiguiente, en estas dos técnicas se reemplazaron los resúmenes por la representación “títulos + resúmenes + palabras clave”, con el objetivo de analizar si esta modificación producía mejores resultados.

Como muestra la tabla 31, no hubo una mejora en el resultado, siendo dispar el comportamiento en cada conjunto de datos. A veces, generando más artículos no repetidos y, a veces menos. El mejor comportamiento en similitud se presentó entre el texto del artículo y una representación del texto en cada técnica, no entre las diferentes mejores representaciones de cada técnica. En esta dirección, se desprende que estar comparando resúmenes en algunas técnicas con títulos, resúmenes y palabras claves de otras técnicas puede estar afectando de manera negativa esa similitud.

Visto lo anterior se realizan pruebas de comparación tal como muestra la tabla 32.

Tabla 31

Comparativas entre artículos no repetidos entre textos completos de artículos y resúmenes

Experto	artículos no repetidos en textos completos	artículos no repetidos en resumen y T+R+PC en TF-IDF y Doc2Vec
Experto 1	8	15
Experto 2	20	31
Experto 3	17	22
Experto 4	11	31
Experto 5	15	30
Experto 6	13	33
Experto 7	19	8
Experto 8	9	28
Experto 9	10	38
Experto 10	5	12
Total artículos no repetidos	127	248
representa	23,52%	45,93%
expansión no repetidos		95,28%
Total artículos repetidos	418	292
representa	77,41%	54,07%
retracción		30,14%

Esta comparación muestra mayor convergencia en relación a la mejor metodología y sobre el promedio, TF-IDF en primer lugar y Transformer 1 en segundo, pero es un dato estadístico que no refleja directamente ser la mejor solución del conjunto.

Si se toman las mejores representaciones de cada técnica, la tabla 33, muestra una mejora en representaciones en cuanto a coincidencias en TF-IDF y en Doc2Vec cuando se utilizan

la representación de título, resumen y palabras claves en vez de resumen. Queda así en las dos comparaciones sobre el promedio Transformer 1 y TF-IDF.

Aunque se trata de una prueba derivada, señala una tendencia clara. A mayor coincidencia, mayor convergencia entre las técnicas, lo que lleva a una mejor identificación de la similitud textual y, en consecuencia, a una mayor similitud temática. Por lo tanto, se puede deducir que el conjunto óptimo a utilizar corresponde a los textos de los artículos científicos con los recortes previamente mencionados. En caso contrario, si la calidad de la priorización disminuye, las representaciones obtenidas a partir de la comparación de diferentes representaciones textuales entre las técnicas de PLN incluidas en el estudio presentan un rendimiento inferior.

Tabla 32

Comparativas entre mejor metodología elegida por cada experto y las diferentes técnicas de PLN sobre resúmenes, y entre todas las técnicas del informe de priorización y y las diferentes técnicas de PLN sobre resúmenes.

Experto	Metod. 1ra.	cant. coincidencias en Metod. 1ra.	relación resumen con 1er. Metodología texto completo							relación resumen con las 6 metodologías de texto completo						
			coincidencia en resumen							coincidencia en resumen						
			TF-IDF	Doc2Vec	Word2Vec	fastText	Transformer 1	Transformer 2	cant. coincidencias	TF-IDF	Doc2Vec	Word2Vec	fastText	Transformer 1	Transformer 2	cant. coincidencias
Experto 1	fastText	3	2	2	2	2	1	0	9	7	3	5	5	6	4	30
Experto 2	Transf. 1	3	2	0	1	1	1	0	5	3	0	3	3	2	1	12
Experto 3	TF-IDF	4	3	3	1	1	3	4	15	3	3	1	1	5	4	17
Experto 4	TF-IDF	3	2	0	2	2	1	0	7	3	1	2	3	1	3	13
Experto 5	TF-IDF	7	4	0	3	2	4	0	13	5	1	4	4	6	3	17
Experto 6	Transf. 1	4	1	1	1	3	2	1	9	3	2	3	4	4	2	18
Experto 7	Doc2Vec	8	5	5	4	4	5	5	28	5	6	5	5	5	6	32
Experto 8	Doc2Vec	4	2	1	0	1	3	3	10	4	3	0	1	4	3	15
Experto 9	Transf. 1	3	1	0	0	0	1	2	4	1	1	0	1	6	3	12
Experto 10	fastText	6	5	3	1	2	4	3	18	9	5	3	5	7	7	36
			27	15	15	18	25	18	118	43	25	26	32	46	36	202

Finalmente, si los vectores de los textos completos convergen y muestran mayores similitudes entre diferentes técnicas en comparación con los vectores generados a partir de los resúmenes, esto sugiere que los textos completos capturan un significado más preciso y completo. El mayor contexto proporcionado por la extensión del texto completo, en contraste

con el resumen, que necesariamente omite información al sintetizar, contribuye a una mayor precisión en la similitud detectada.

Tabla 33

Comparativas entre mejor metodología elegida por cada experto en su informe de evaluación y las diferentes técnicas de PLN sobre mejor representación, y entre todas las técnicas del informe de priorización y las diferentes técnicas de PLN sobre mejor representación.

Experto	Metod. 1ra.	cant. coincidencias en Metod. 1ra.	relación resumen con 1er. Metodología texto completo							relación resumen con las 6 metodologías de texto completo						
			coincidencias en							coincidencias en						
			T+R+PC			Resumen				T+R+PC			Resumen			
			TF-IDF	Doc2Vec	Word2Vec	fastText	Transformer 1	Transformer 2	cant. coincidencias	TF-IDF	Doc2Vec	Word2Vec	fastText	Transformer 1	Transformer 2	cant. coincidencias
Experto 1	fastText	3	3	2	2	2	1	0	10	8	4	5	5	6	4	32
Experto 2	Transf. 1	1	0	0	1	1	1	0	3	1	0	3	3	2	1	10
Experto 3	TF-IDF	4	2	3	1	1	3	4	14	1	2	1	1	5	4	14
Experto 4	TF-IDF	4	4	1	2	2	1	0	10	6	1	2	3	1	3	16
Experto 5	TF-IDF	7	4	0	3	2	4	0	13	5	1	4	4	6	3	17
Experto 6	Transf. 1	4	1	1	1	3	2	1	9	4	3	3	4	4	2	20
Experto 7	Doc2Vec	7	5	6	4	4	5	5	29	5	6	5	5	5	6	32
Experto 8	Doc2Vec	4	3	1	0	1	3	3	11	5	3	0	1	4	3	16
Experto 9	Transf. 1	4	1	1	0	0	1	2	5	1	2	0	1	6	3	13
Experto 10	fastText	7	5	4	1	2	4	3	19	10	4	3	5	7	7	36
			28	19	15	18	25	18	123	46	26	26	32	46	36	206
			+1	+4						+3	+1					

4.5 Desambiguación y extracción de información de perfiles de Google Scholar

Como ya fuera explicitado, la base elegida para desambiguar fue la de perfiles de Google Scholar. En este sentido, se hizo una búsqueda de perfiles posibles en la mencionada base para ver su potencialidad y el resultado se puede observar en la Tabla 34.

El resultado fue que se encontraron posibles resultados positivos en 1238 búsquedas de perfiles candidatos. En esta tarea se debe garantizar que los perfiles encontrados correspondan

con los investigadores ya registrados. Por esto, hay que efectuar un proceso de desambiguación en dos etapas.

En primera instancia se comparan las producciones científicas de los investigadores en la plataforma con las producciones cargadas en el perfil de Google Scholar para encontrar al menos una coincidencia, siendo este proceso el de mayor confianza. Luego, se procede a desambiguar por el apartado de coautores del perfil de Google Scholar verificado en el paso anterior, con los coautores de artículos en los datos extraídos, en este caso, del RI. Este proceso de desambiguación es de menor confianza que el anterior.

Tabla 34

Detalle discriminado por devoluciones de perfiles de investigador por cada nombre incluido en registros¹²

Cantidad de devoluciones en búsqueda de nombres de investigadores	cantidad de perfiles encontrados
0	2241
1	844
2	94
3	60
4	26
5	19
6	17
7	12
8	11
9	9
10 o +	146
cantidad de posibles coincidencias de autores encontrados en Google Scholar	1238

Otra opción posible consiste en trabajar con todos los autores del RI que no tengan un perfil confirmado en Google Scholar. Para ello, se puede realizar una búsqueda de los títulos de sus artículos en Google Scholar, con el fin de encontrar un perfil que corresponda al autor, pero bajo un nombre distinto. Una vez identificada una posible coincidencia, se procede a verificar la concordancia del apellido o de uno de los apellidos. De este modo, no se agregan perfiles nuevos, sino que se desambiguan los nombres, asociándolos a perfiles ya existentes en Google Scholar.

El procedimiento realizado se presenta en el Anexo C.

¹² Sobre los autores en los registros usando la url de búsqueda de perfiles por ejemplo para el investigador (hipotético) identificado con nombre1 nombre2 apellido1 apellido2 la url de búsqueda sería la siguiente: https://scholar.google.com/citations?view_op=search_authors&hl=en&mauthors=nombre1+nombre2+apellido1+apellido2&btnG=

Luego de aplicadas las tres instancias, se obtuvieron 781 perfiles confirmados. Esto resulta de importancia solo si se compara con el total de autores de la extracción de datos del RI y con el total que mostró al menos un perfil para buscar coincidencias (Tabla 35):

Cabe mencionar que el valor 22,39 % correspondiente a los perfiles confirmados sobre el total de autores RI permite no solo desambiguar, sino agregar indicadores bibliométricos relevantes para la aplicación en el prototipo, ya que se accede a suficientes datos para realizar distintas pruebas. Asimismo, resulta significativo el valor de 63,09 % que muestra la tabla anterior. Otras estadísticas que brindan información relevante se agrupan en las siguientes tablas.

Tabla 35

Estadísticas de perfiles confirmados sobre autores en el RI

Total de perfiles confirmados	Cantidad de autores distintos incluidos en RI	Cantidad autores que obtuvieron al menos un resultado en búsqueda de perfil	Porcentaje perfiles confirmados sobre total autores RI	Porcentaje perfiles confirmados sobre autores que obtuvieron al menos un resultado
781	3488	1238	22,39 %	63,09%

La Tabla 36 muestra que hay 10 autores del RI que están relacionados a 2 perfiles de Google. Lo que quiere decir que estos autores cargaron 2 veces su perfil, por el motivo que sea.

Tabla 36

Estadísticas de perfiles por autor

Cantidad de perfiles por autor	Cantidad
1	761
2	10

En estos casos se toma la decisión de usar el perfil cuyos indicadores bibliométricos den valores más altos porque se considera que es el que está más actualizado.

La estadística de cantidad de autores por perfil se registra en la Tabla 37. Este dato relevante demuestra que hay algunos perfiles que tienen asociados más de un autor, lo cual significa que encontró diferentes denominaciones del mismo autor. En el caso particular, registró 50 autores que tienen 2 nombres diferentes en el RI, y 2 autores que tienen 3 nombres diferentes en el RI.

Tabla 37*Estadísticas de autores por perfil*

Cantidad autores por perfil	Cantidad
1	675
2	50
3	2

En síntesis, se encontraron autores que tienen más de un perfil, definiéndose que hacer en estos casos y autores que tienen diferentes denominaciones en el RI. Las dos posibilidades de desambiguación que se pueden dar, para un lado y para el otro. Esto prueba la validez del método para desambiguar autores además de acceder a indicadores bibliométricos útiles a la metodología. En esta dirección, el agrupamiento de datos de acuerdo a si están o no confirmados y tienen indicadores se puede observar en la Tabla 38. Este dato es importante para saber cuántos de los 781 perfiles confirmados tienen indicadores bibliométricos. Solo 30 de ellos no tienen indicadores, representando el 3,84 % de la muestra.

Tabla 38*Estadísticas de perfiles confirmados por indicadores*

Tipo de perfiles	Cantidad
Perfiles confirmados que no poseen indicadores	30
Perfiles confirmados con indicadores	751

A continuación, se hace un análisis de los nombres de los autores en RI asociados a la búsqueda exacta de los mismos en los perfiles de Google Scholar. Cabe aclarar que en el RI los nombres de los autores aparecen por lo general, primero el apellido y luego el nombre y en Google Scholar aparecen de forma inversa. Esto requiere la tarea previa de definir la conversión al formato de Google Scholar teniendo en cuenta también las excepciones, lo cual fue resuelto en el prototipo para la búsqueda de perfiles.

En resumen, de la búsqueda de coincidencias se obtiene el resultado que se muestra en la Tabla 39.

Del análisis de las 643 coincidencias con los perfiles confirmados de Google Scholar se obtiene el resultado que se observa en la Tabla 40.

Tabla 39*Estadística de cantidad de perfiles candidatos y Google Scholar confirmados*

cantidad de perfiles que devuelve la búsqueda general de nombres	Google Scholar confirmados
1	468
2	51
3	25
4	7
5	7
6	6
7	4
8	6
9	6
10	63
Total	643

Tabla 40

Estadística de cantidad de perfiles candidatos y coincidencia exacta entre nombres RI y Google Scholar confirmados

cantidad de perfiles que devuelve la búsqueda	coincidencia exacta entre nombre RI y Google Scholar confirmado
1	409
2	32
3	18
4	3
5	3
6	3
7	1
8	2
9	1
10	20
Total	492

De las 643 coincidencias, se confirmaron 492 nombres de autores, lo que representa un 59 % más de rendimiento en comparación con los 781 nombres obtenidos mediante el método de desambiguación propuesto. Esto demuestra que la metodología propuesta es superior a la simple búsqueda de nombres exactos. Además, evidencia la ineficiencia del uso de nombres exactos como campo clave en la búsqueda de perfiles para desambiguación o en la obtención de datos asociados, como indicadores bibliométricos.

En relación al esquema de guardado, se definen dos momentos importantes principales. La Búsqueda que consiste en la extracción propiamente dicha de los datos y la Confirmación que se define como el proceso interno del prototipo que verifica la correspondencia de los datos de los perfiles a los datos de autores RI.

La búsqueda tiene 2 etapas que son:

- ✓ Búsqueda de posibles perfiles: todos los datos extraídos son cargados en archivos CSV, en el mismo momento que se realiza el proceso de extracción. Se toma la cantidad de perfiles que devuelve cada búsqueda y los datos de acceso a cada perfil encontrado, junto a valores de orden en la aparición.
- ✓ Búsqueda de datos de los perfiles: una vez que se tienen todos los datos anteriores se puede correr el proceso de extracción de información de perfiles donde se obtiene principalmente, artículos publicados, indicadores bibliométricos, cantidad de citas anuales, coautores obtenidos del bloque que tiene esta información y datos estadísticos de la extracción.

Un dato que se tiene en cuenta en esta parte del proceso, es la limitación del ámbito de la búsqueda para tratar de evitar el rechazo de Google en el proceso y los tiempos entre consultas¹³.

La confirmación tiene los tres procesos de validación que ya fueron explicados. Todos los datos confirmados con información de fuente de validación son guardados en un archivo para su uso posterior.

En cuanto a la definición del umbral de conocimiento se guardan las configuraciones en los archivos csv correspondientes a los parámetros del sistema y se utilizan con normalidad. También se realizan los cálculos de experiencia genérica, teniendo en cuenta los valores residuales y los cálculos con la configuración de umbral.

En cuanto a la aplicación del interés del investigador, no se registró ningún inconveniente trabajando sobre los parámetros asociados, la tabla de antigüedad y el cálculo de antigüedad del artículo en base a la fecha cargada en la base de conocimientos.

En referencia al ordenamiento de los nombres de los autores, se había indicado que tiene sentido su utilización si este coincide con el orden de importancia del aporte en el artículo científico. Al no existir una ordenación determinada e indefectible en este sentido, se pueden analizar los datos para determinar la factibilidad y validez de la utilización de la lógica que se propone.

¹³ Esto se aclara ya que dada las restricciones presupuestarias no se trabaja con aplicaciones de pago.

Un análisis posible es determinar si hay una lógica de ordenación asociada al orden alfabético ascendente en la aparición de los autores. La Tabla 41 muestra información de los artículos incluidos en la base de conocimiento, siguiendo lo mencionado.

Tabla 41

Autores discriminados por cantidad de autores en cada publicación y verificación de orden basado en nombres

Número Autores	Registros ordenados	Registros totales	Sobre probabilidad	Probabilidad de 1 sobre:
2	569	1114	12	2
3	124	558	31	6
4	26	285	14	24
5	5	90	4	120
6	0	41	0	720
7	3	33	3	5040
8	1	17	1	40320
9	1	11	1	362880
10	0	8	0	3628800
11	0	3	0	39916800
12	0	1	0	479001600
13	1	3	1	6227020800
14	1	3	1	87178291200
15	0	1	0	1307674368000
Totales	731	2168	68	

Debido a la ordenación alfabética ascendente, no se puede concluir que sea la lógica más utilizada, considerando los valores de probabilidad promedio generados de forma aleatoria. Aunque se observa un ligero aumento cuando hay muchos autores, la cantidad es mínima en relación con el tamaño de la muestra. Por lo tanto, no se detectó una priorización significativa de este tipo de ordenación. Sin embargo, sigue siendo válido sugerir, en términos relativos, la posibilidad de incluir un factor de ordenación basado en la importancia del aporte de los autores al artículo científico-

Respecto a la utilización de este factor, se tomaron los valores de la tabla de parámetros y de la tabla de aportes de los autores en comparación con la ordenación de los autores en la base de conocimiento, observando que funciona de manera adecuada.

Sobre el control de Conflicto de intereses, primero deben desambiguarse los nombres. El hecho de que exista un ID de los autores del CONICET, ayuda al proceso inicial porque no es necesario desambiguar a los mismos, lo cual lo simplifica en parte. Para los otros casos se realiza la búsqueda de los posibles nombres en base al proceso propuesto de desambiguación, con el

fin de encontrar todos los nombres que pueden tener en el RI. El primer conflicto de intereses se presenta cuando los autores del artículo no pueden ser sus propios revisores.

Una vez obtenidos todos los posibles nombres que tiene el autor se busca a quienes han realizado artículos de investigación en coautoría. En este caso hay una consideración más compleja a realizar sobre el real conflicto de intereses, de allí que en el reporte de recomendación se marcan como coautores con posible conflicto de intereses.

Sobre los coautores también se procede a hacer la desambiguación de nombres ya que podrían tener diferentes nombres en la base del RI. Así, es posible marcar correctamente quienes son coautores al confeccionar una lista con todas las ocurrencias. También se procede a hacer una búsqueda de coautores de coautores para marcarlos correctamente. Luego se verifican las afiliaciones incluidas en el RI, para marcar aquellos que tienen misma afiliación dentro de los priorizados.

Con referencia al cálculo de puntaje de la producción científica, se tuvieron en cuenta todos los parámetros intervinientes, las tablas y los valores de puntaje de experiencia genérica, del parámetro de antigüedad y aporte de los autores, efectuándose los cálculos sin inconvenientes.

4.5.1 Autoridad Académica

En este globalizador se desarrollaron todas las funcionalidades sin ningún inconveniente, tanto sea de habilitación del cálculo como de valores internos.

La elección del índice bibliográfico a utilizar se efectuó considerando los indicadores presentes en los perfiles de investigadores de Google Scholar, descargados en los archivos CSV correspondientes. Las opciones fueron:

- ✓ Citas
- ✓ Citas desde (año). 5 años de antigüedad
- ✓ h-index
- ✓ h-index desde (año). 5 años de antigüedad
- ✓ i10-index
- ✓ i10-index desde (año). 5 años de antigüedad
- ✓ Citas manuales: se registra el año desde donde se empiezan a contar las citas anuales.

El procedimiento de aplicación de índices bibliométricos fue explicitado en el capítulo correspondiente a la metodología propuesta. Si el procedimiento está habilitado, se elige el indicador a utilizar y si se cuenta con datos en dicho indicador se busca en la tabla de cálculo

bibliométrico tomando el valor de representatividad que se va a aplicar en la fórmula de cálculo de autoridad académica.

Para aplicar la titulación, se tomaron los valores de título del CSV con información de los autores CONICET. Si el procedimiento está habilitado, se busca el valor en la tabla de cálculo de título y se toma el valor de representatividad que es aplicado luego en la fórmula de cálculo de autoridad académica.

Para el cálculo de puntaje de autoridad académica se realizó el procedimiento en base a los valores de cálculo de indicadores bibliométricos y titulación, junto al parámetro de cálculo entre ambos.

4.5.2 Revisiones anteriores

Se desarrollaron todas las funcionalidades sin ningún inconveniente. Tanto sea de habilitación del cálculo y/o cortes y de valores internos.

Para la registración se realizaron cargas de prueba dentro del CSV correspondiente a las revisiones anteriores y se probaron los procedimientos asociados.

En la aplicación de controles temporales y tiempos de no consideración se cargaron los parámetros por pedidos simultáneos, tiempo de descanso y revisiones por período en el CSV de parámetros de revisiones anteriores. Se probaron todos los procedimientos asociados a los cortes hasta constatar su funcionamiento adecuado.

En relación con la implementación de penalizaciones (temporales y totales), en los archivos CSV correspondientes se cargaron las configuraciones necesarias para habilitar las penalizaciones por rechazos y malas calificaciones. Además, se integraron las tablas de cálculo correspondientes a los rechazos y a las malas calificaciones, utilizando las tablas de revisiones y calificaciones correspondientes. Se llevaron a cabo pruebas exhaustivas de todos los procedimientos asociados a las penalizaciones, hasta garantizar que funcionaran correctamente.

De igual manera en el cálculo de puntaje de revisiones anteriores, se realizaron los mismos con cargas de prueba, tomando del CSV de parámetros de revisiones anteriores el parámetro de cálculo entre valores del coeficiente de evaluación de revisiones. Se probó el procedimiento y se constató su preciso funcionamiento.

4.5.3 Integración de la metodología con parámetros integrales y múltiples perfiles de configuración.

Finalmente, se realizaron los procedimientos asociados a la parte final donde convergen todos los parámetros, tablas y variables metodológicas de los cálculos locales de cada globalizador de factores.

En la fórmula integral de la Metodología, se utilizaron las configuraciones definidas en el archivo CSV de parámetros generales, incorporando los parámetros integrales de peso y peso relativo de la unidad, junto con los puntajes de cada globalizador. Se realizaron pruebas exhaustivas de los procedimientos hasta lograr un cálculo preciso, lo que permitió generar los listados de candidatos a revisores ordenados, así como el listado de candidatos sujetos a restricciones, de acuerdo con los diferentes motivos establecidos por la metodología.

Considerando el requerimiento socio-técnico-cultural de los múltiples perfiles de configuración de la metodología, se crearon tres perfiles de configuración distintos para las pruebas, los cuales fueron sometidos a pruebas exhaustivas hasta asegurar su buen funcionamiento.

Con respecto a la información necesaria para la parametrización, se tomó la información consensuada por el CE de la Revista IRICE. Los mismos resultaron de los intercambios realizados luego de la encuesta ya mencionada. La configuración se muestra en el punto siguiente.

4.6 Validación experimental de la metodología propuesta en la revista IRICE

En el laboratorio de pruebas se implementó el modelo de recomendación de revisores propuesto, aplicándolo a manuscritos de la revista IRICE. Para ello, se utilizó una base de conocimientos disponible, utilizando el texto del artículo y la técnica de Transformer 1, datos obtenidos en los pasos anteriores y se configuraron las parametrizaciones específicas entregadas por el equipo editorial de la revista, contemplando sus criterios y requerimientos particulares.

Los criterios acordados para esta primera prueba fueron:

Globalizador de conocimiento científico: Sí

- Experiencia: Sí
- Intereses del investigador: Sí
 - Tabla de antigüedad

desde año de antigüedad	hasta año de antigüedad	% de representatividad
0	5	100
5	10	50
11	+	0

Aporte de los autores: Sí No

Globalizador de Autoridad académica: Sí

- Indicador bibliográfico: No
- Titulación: Sí
 - Tabla de títulos

Título	% de ventaja comparativa
Doctorado	30 %

Globalizador de revisiones anteriores: Sí

- Revisiones anteriores: cortes - cálculo
- Revisiones simultáneas: 1
- Tiempo de descanso: 6 meses
- Revisiones por período: 2
- Penalización por rechazo de solicitudes de revisiones: Sí
 - Tabla de penalización por rechazos de revisiones

Cantidad de respuestas negativas desde	Cantidad de respuestas negativas hasta	Penalización
3	4	6 meses
5	7	12 meses
8	+	Por tiempo indeterminado

- Tabla de calificaciones de revisiones:

Valor calificación desde	Valor calificación hasta	Porcentaje de variación	Sentido de la variación	Mala calificación
7	10	20	A favor	N
6	6		Neutral	N
5	5	20	En contra	N
0	4	40	En contra	S

- Penalización por malas calificaciones: Sí
 - Tabla de penalización por malas calificaciones

Cantidad de malas calificaciones desde	Cantidad de malas calificaciones hasta	Penalización
2	3	2 meses
4	6	6 meses
7	+	Por tiempo indeterminado

Peso de los Globalizadores:

- Globalizador de conocimiento científico: 40 %
- Globalizador de Autoridad académica: 30 %
- Globalizador de revisiones anteriores: 30 %

Ante la complejidad de esta primera experiencia del CE, también en diálogo interdisciplinar se sugirieron algunas cuestiones que fueron acordadas, a saber:

Elección de umbral: automático con experiencia + antigüedad debido a que el interés del investigador se consideró un factor adecuado a tener en cuenta.

Forma de sumar artículos: por promedio ponderado, así de esta manera la integración de varios artículos tiene un impacto, pero con cierto ajuste producto del cálculo

En la fórmula de cálculo de revisores que tiene como opciones (suma, promedio y multiplicación), se opta por el promedio que promueve cierto ajuste (suavizado) en el cálculo en relación a las otras opciones.

En la fórmula general: definición de peso unitario de cada globalizador en base al 50% del valor (0,5), permitiendo dos unidades por cada uno de ellos.

Los manuscritos seleccionados fueron 19 que debían ser evaluados para su publicación durante el 2025, los mismos fueron enviados a 7 expertos del CE en las temáticas involucradas conjuntamente con los informes proporcionados por el recomendador (Anexo D esquema del informe enviado).

Para validar el proceso se realizó una entrevista en profundidad a cada uno de ellos. La misma indagó en base a cuatro preguntas abiertas los aspectos que se consideraron más significativos para evaluar la metodología propuesta:

1. ¿Encuentra una similitud temática entre los artículos de los candidatos a revisores y el manuscrito?
2. ¿Considera adecuado el listado de recomendación elaborado?
3. ¿Propone alguna modificación en la parametrización?
4. ¿Sugiere algún aspecto que podría mejorarse en esta metodología?

4.6.1 Resultados

Similitud temática

Los resultados revelan una tendencia mayoritariamente positiva respecto a la similitud temática entre los artículos de los candidatos a revisores y el manuscrito evaluado. La totalidad de los participantes reconocieron, de manera explícita o implícita, la existencia de una relación temática entre los documentos analizados.

Tipos de evaluaciones identificadas: en relación a la coincidencia explícita algunos participantes expresaron en forma directa la coincidencia temática entre los documentos. Sobre la evaluación jerárquica se consideró el grupo de candidatos adecuado en términos generales, pero identificaron diferentes niveles de pertinencia dentro del conjunto, estableciendo un orden de relevancia entre los potenciales revisores.

Es importante mencionar que hay factores que influyen en la evaluación como la parametrización del sistema resultante de la decisión del CE, dado que genera modificaciones en el orden de similitud temática, reflejando la idiosincrasia particular de cada CE. En esta dirección también en algunos casos se identificaron similitudes en temas generales más que en la especificidad del artículo evaluado. Este hallazgo confirma las observaciones previas realizadas durante la selección de la técnica de Procesamiento de Lenguaje Natural (PLN) más adecuada. Los informes de expertos revelaron un patrón específico en el funcionamiento del sistema: cuando un artículo abarca múltiples temas y no se encuentran en la base de conocimiento artículos que combinen exactamente esas mismas temáticas, el algoritmo prioriza aquellos documentos que tratan los temas más generales presentes en el artículo objetivo. Esta característica del sistema si bien podría representar una limitación se presenta en esta metodología como una estrategia de adaptación ante la complejidad temática de ciertos manuscritos.

Evaluación del Listado de Revisores

Las respuestas obtenidas en la segunda pregunta evidencian una adecuación general satisfactoria del sistema. La unanimidad de los participantes en considerar el listado como apropiado y funcional indica que la metodología de selección de revisores está correctamente orientada en términos generales. Entre los aspectos positivos se destaca:

- ✓ Funcionalidad del sistema: Todos los participantes validaron la operatividad del listado.
- ✓ Pertinencia de la selección: El contenido propuesto fue valorado positivamente por la totalidad de evaluadores.
- ✓ Direccionamiento correcto: Los resultados sugieren que el enfoque metodológico es adecuado.

Las mejoras propuestas se orientan a que, si bien la funcionalidad core del sistema es apropiada, existen oportunidades de mejora en la presentación y organización de la información que podrían optimizar la experiencia del usuario sin comprometer la calidad de la selección algorítmica.

Evaluación de la Parametrización

La tercera consulta evidenció una aceptación general de la parametrización implementada. La mayoría de los participantes recibieron y comprendieron favorablemente los criterios establecidos, lo que sugiere que los parámetros utilizados resultan apropiados y transparentes para la comunidad evaluadora.

Validación de los Criterios Fundamentales: es importante destacar que ninguna de las observaciones realizadas cuestionó los fundamentos de la parametrización, confirmando la solidez conceptual del marco metodológico desarrollado.

Sugerencias de Mejora Identificadas: algunos participantes propusieron incluir una nota aclaratoria sobre el criterio de ordenamiento. Cabe señalar que toda la información sobre parametrización había sido proporcionada al inicio de la entrevista, previa a las consultas específicas.

También se sugirió evitar repeticiones en el listado para mejorar tanto la precisión como la economía de la propuesta presentada. En este caso se identificó un problema técnico: durante la evaluación se detectó una inconsistencia en la presentación de resultados. Los candidatos con puntaje cero aparecían duplicados tanto en la lista de priorizados (al final) como en la lista de excluidos. Esta duplicación se originaba porque los autores que únicamente contaban con publicaciones de más de 10 años de antigüedad eran automáticamente excluidos por la parametrización, pero el sistema los mostraba en ambos listados. La corrección de esta anomalía se considerará en futuras versiones del sistema.

En síntesis, la parametrización cuenta con una base sólida y aceptación generalizada. Las observaciones recibidas se enfocan exclusivamente en aspectos de presentación y optimización técnica, sin comprometer la validez conceptual del marco metodológico implementado.

Análisis de la Recepción Metodológica

Los entrevistados no identificaron problemas metodológicos fundamentales, lo cual constituye una validación parcial de la estructura y el enfoque empleados en la investigación.

Entre las sugerencias de mejora identificadas se menciona optimizar la presentación visual y estructural de los resultados finales. Esta recomendación será incorporada en las futuras iteraciones del prototipo.

Un entrevistado propuso revisar la ponderación de los insumos textuales analizados, sugiriendo privilegiar las palabras clave por sobre el resumen para optimizar la identificación de similitudes temáticas. En este sentido, es necesario reforzar la información sobre los resultados de las pruebas realizadas dentro de la presente tesis, con el fin de facilitar la comprensión de las decisiones metodológicas adoptadas¹⁴.

Recomendaciones para Futuros Desarrollos: los hallazgos indican la necesidad de generar más instancias de intercambio con el CE para presentar en mayor detalle las decisiones metodológicas y sus fundamentos.

Finalmente, de las opiniones relevadas, la evaluación reveló una aceptación metodológica sólida. Las sugerencias recibidas son puntuales y aplicables, sin constituir un cuestionamiento fundamental del método propuesto. Por el contrario, representan una invitación constructiva a refinar aspectos operativos y técnicos específicos, lo cual fortalece la validez y aplicabilidad de la propuesta metodológica. Además, en las entrevistas se verificó que la metodología facilitó una mejora en el proceso de recomendación respecto al procedimiento habitual, optimizando la selección de revisores en función de criterios objetivos y configurables. Asimismo, se evidenció que el prototipo constituye una herramienta de apoyo útil para la toma de decisiones editoriales, al permitir flexibilizar y adaptar los criterios de asignación de revisores a cada caso. Aunque no se dispone de un modelo previo implementable para realizar una comparación práctica directa, los resultados obtenidos a la fecha respaldan, en términos funcionales y operativos, la validez y aplicabilidad de la metodología propuesta en contextos editoriales reales.

¹⁴ Es importante aclarar que la metodología empleada se seleccionó como la más eficiente basándose en los informes de expertos que evaluaron diferentes priorizaciones de técnicas de PLN sobre textos completos. Esta decisión también se fundamentó en pruebas exhaustivas de la metodología.

Conclusiones

En la presente tesis doctoral se ha explorado el desarrollo y validación de una nueva metodología adaptativa para la recomendación de revisores, fundamentada en múltiples factores. En lugar de proponer un modelo cerrado, se ha avanzado en un diseño que, bajo un enfoque sociotécnico-cultural, propone un sistema con capacidad de adaptarse a los criterios específicos establecidos por el equipo editorial de cada revista científica. El informe de recomendación funciona como una herramienta de apoyo que facilita la toma de decisiones en la asignación de revisores. El mapeo sistemático, el marco teórico y los resultados de la encuesta contribuyeron significativamente a la creación de un modelo flexible y adaptable, que considera los factores clave que influyen en los procesos editoriales de diversas publicaciones científicas.

El análisis de las diferentes partes representativas de los artículos científicos proporcionó información sólida que sirvió de base para tomar decisiones sobre su selección y orientar el proceso de búsqueda de similitud textual, evaluando tanto sus ventajas como desventajas. La incorporación de diversas técnicas de Procesamiento de Lenguaje Natural (PLN) aportó flexibilidad al modelo, permitiendo integrar distintas etapas evolutivas del PLN, desde enfoques tradicionales hasta los más avanzados, como los Transformers, que son fundamentales en los últimos avances de la IA.

La evaluación realizada por expertos permitió validar el rendimiento de estas técnicas en simulaciones de casos reales, lo que facilitó la identificación de la técnica más adecuada para la revista científica que se tomó de referencia para realizar el trabajo de campo. Esta integración de análisis de similitud textual y selección de técnicas de PLN optimiza la recomendación de revisores, proporcionando una ventaja significativa en la búsqueda y asignación de revisores, lo que redundará en una mejora sustancial en la calidad del proceso de selección.

Entre los aspectos más relevantes del proceso de preprocesamiento se incluyen las técnicas de desambiguación de autores, los procesos de traducción, la limpieza de maquetado y el corte de los artículos científicos para separar las partes textuales representativas. Estas tareas, junto con otras orientadas a obtener el material más depurado posible, lograron una efectividad superior al 90% en muestras aleatorias, asegurando que el contenido procesado fuera altamente preciso.

Por otro lado, la adaptabilidad del modelo a los diferentes criterios de los equipos editoriales no puede evaluarse mediante una métrica estándar o comparativa, ya que cada grupo puede tener enfoques distintos, incluso frente a un determinado caso a evaluar se pueden tomar decisiones puntuales dentro de un mismo grupo. En lugar de buscar un "mejor resultado" entre configuraciones, esta flexibilidad refleja la diversidad de interpretaciones y criterios

dentro de los equipos editoriales, posicionando dicha diversidad como una parte fundamental de la solución.

Aportes de la metodología propuesta:

La metodología desarrollada presenta una serie de conceptos y cálculos que complementan y superan las limitaciones de las metodologías estudiadas. Como ya se mencionó, integra criterios flexibles y adaptativos orientados a las necesidades específicas de cada equipo editorial, conformando así una solución adaptable a la diversidad de criterios presentes en las diferentes revistas científicas. Este enfoque es un avance respecto al uso tradicional del factor de diversidad en algunas metodologías estudiadas. A continuación, se detallan los principales aportes organizados en tres grandes áreas globalizadoras:

Globalizador de factores de Conocimiento Científico:

En experiencia:

- Se introduce el concepto de conocimiento específico relativo, basado en la aplicación práctica del umbral dinámico, utilizando los valores de los artículos seleccionados.
- La metodología permite acumular conocimiento relativo de acuerdo con la cantidad de artículos priorizados, en lugar de utilizar una cantidad fija predefinida en la fórmula.
- Permite seleccionar las partes más representativas de un artículo. En contraste con los métodos previos, que definen directamente qué parte del artículo es representativa, aquí el equipo editorial tiene varias opciones para seleccionar el contenido relevante:
 - ✓ Texto completo (excluyendo título, resumen, palabras clave y bibliografía).
 - ✓ Título.
 - ✓ Resumen.
 - ✓ Palabras clave.
 - ✓ Título y resumen.
 - ✓ Título, resumen y palabras clave.
- Se incorpora la capacidad de definir las técnicas de PLN a utilizar, permitiendo que cada equipo editorial elija la más adecuada para su contexto.

En interés del investigador:

- Flexibiliza la incorporación o exclusión de este factor.

- Permite la configuración de rangos de antigüedad para definir el valor representativo más adecuado, según lo decida el equipo editorial, en lugar de aplicar un valor de antigüedad preestablecido.

En aporte de los autores:

- Esta metodología es la única entre las estudiadas que incorpora este factor, y lo hace mediante el orden de publicación de los autores.
- Ofrece flexibilidad en la incorporación de este factor, permitiendo configurar rangos de orden para definir el valor representativo más adecuado en base al criterio editorial, en lugar de aplicar un valor predefinido.
- En general:
 - Implementa el Control de Umbral de Consideración, restringiendo la inclusión de investigadores con artículos priorizados si sus valores caen por debajo del umbral determinado.

Globalizador de factores de autoridad académica:

- Flexibiliza la incorporación o no de este globalizador.

En Indicadores bibliométricos:

- Flexibiliza la incorporación o no de este factor.
- Permite definir cuál es el indicador que se va a utilizar en base a los disponibles.
- Permite la configuración de rangos de consideración de valores de indicadores para definir por rango el valor representativo más adecuado en base a consideración del equipo editorial y no por aplicación directa de un valor predeterminado en la fórmula.

Titulaciones:

- Flexibiliza la incorporación o no de este factor.
- Permite la configuración de valor de representatividad en base a la información disponible según criterio del equipo editorial, además de poder definirlo como restricción si no posee determinado nivel de titulación. A diferencia del artículo estudiado que lo incorpora, no solo lo toma de regla de inclusión o no, sino que influye en el cálculo.

Operación de combinación de indicadores:

- Permite definir cuál es la operación que asocia los valores de indicadores bibliométricos y titulaciones en el caso de estar los dos cálculos habilitados.
- Permite definir los valores por defecto en los casos que no se obtienen datos de Indicadores Bibliométricos y Titulaciones.

Globalizador de revisiones anteriores:

- Permite al equipo editorial decidir si se incorpora o no.
- En caso de utilizarse permite definir si se va a utilizar para:
 - o Cálculo
 - o Corte
 - o Cálculo y corte
- Posibilita evaluar la consistencia de las revisiones con una parametrización con un buen grado de flexibilidad.
- Permite definir si se va a controlar la cantidad de revisiones simultáneas máximas a solicitar a un evaluador y en caso afirmativo ingresar ese valor.
- Permite definir si se va a controlar el tiempo de espera o descanso entre pedidos de revisión y en caso afirmativo ingresar ese valor
- Permite definir si se va a controlar la cantidad de revisiones máximas a solicitar a un evaluador por año calendario y en caso afirmativo ingresar ese valor.
- Permite definir si se va a controlar la cantidad de rechazos a solicitudes de revisiones y en caso afirmativo, definir por parametrizaciones para penalizar por la cantidad de revisiones por rangos asociado al valor representativo de la penalización en meses o de forma permanente.
- Permite definir si se va a controlar la calidad de las revisiones y en caso afirmativo, definir por parametrizaciones para penalizar por la cantidad de malas revisiones por rangos asociado al valor representativo de la penalización en meses o de forma permanente. Además de definir valor por defecto en caso de no tener historial de revisión.
- Permite realizar diferentes cálculos entre todas las revisiones que haya realizado el candidato a revisor y estén registradas

Otras características del modelo:

- La metodología permite definir diferentes pesos para los factores globalizadores, cuya suma debe ser igual a 100. Cada peso también representa el límite máximo de cada globalizador.
- Incorpora el concepto de peso unitario para cada factor, permitiendo definir cuántas unidades se pueden incorporar al cálculo total.
- Ofrece la posibilidad de configurar múltiples perfiles de parametrización para adaptarse a necesidades más complejas de los equipos editoriales, especialmente cuando se enfrentan a dificultades para encontrar revisores y necesitan flexibilizar los requisitos establecidos.

Es importante destacar que, al ser una metodología adaptable a la idiosincrasia de cada CE, incorpora inevitablemente cierto grado de subjetividad en la forma de abordar la problemática. Esta característica impide que los resultados puedan ser medidos objetivamente mediante una métrica única, ya que la flexibilidad del enfoque da lugar a resultados distintos según las particularidades de cada equipo de trabajo.

Los aspectos desarrollados representan características diferenciales de la metodología presentada en relación con las metodologías revisadas en el trayecto investigativo de esta tesis doctoral. Dichas particularidades configuran un marco conceptual orientado a introducir mejoras en los procesos de recomendación de revisores. Estas mejoras fueron validadas experimentalmente a través de un prototipo, que permitió evaluar si la aplicación de la metodología propuesta aporta ventajas significativas respecto a los procedimientos tradicionales de selección de revisores. Asimismo, en línea con los principios establecidos por la Guía de la UNESCO para el uso de IA generativa en educación e investigación (Miao & Holmes, 2023), se adoptaron criterios de transparencia, equidad, responsabilidad y respeto por los derechos de las personas en cada etapa del proceso. Esto incluye consideraciones sobre la obtención ética de datos públicos, la correcta gestión y resguardo de la información, y el uso responsable de tecnologías de IA para apoyar procesos de priorización y selección de revisores en ámbitos académicos y científicos. De este modo, se garantiza la legitimidad, integridad y confiabilidad de la metodología, promoviendo prácticas éticas y responsables en la aplicación de la IA en la gestión de recursos humanos para la investigación.

Líneas futuras de investigación:

Como prospectiva de trabajo, se identifican los siguientes tópicos relevantes que los resultados de esta tesis habilitan para futuras investigaciones y desarrollos:

- Desarrollo de métodos de corte de artículos científicos: Un área de investigación futura podría centrarse en la creación o mejora de métodos automatizados para el corte de artículos científicos, de modo que se pueda obtener un software accesible y eficiente para este propósito. Este método permitiría procesar los artículos de manera más precisa y reducir la intervención manual en el proceso de selección de partes representativas de los textos.
- Análisis más detallado de la obtención de perfiles de investigadores: Sería necesario realizar un análisis más exhaustivo de las diferentes opciones disponibles para la obtención de perfiles de investigadores. Esto incluye explorar el uso de diversas bases de datos, herramientas y métodos que puedan proporcionar información más precisa y variada sobre los investigadores, lo cual puede mejorar la recomendación de revisores.

- Optimización mediante APIs: En el futuro, se debería buscar la optimización del modelo actual mediante el uso de APIs en lugar de depender de métodos de Web Scraping. Esto permitiría una mayor eficiencia y precisión en la obtención de datos, facilitando la integración con fuentes externas y reduciendo el tiempo de procesamiento de la información.

- Mejora de la similitud textual: Una futura línea de investigación podría centrarse en el desarrollo de nuevas metodologías adaptadas para evaluar la similitud temática de los artículos. Mejorar esta capacidad permitiría que el sistema sea aún más preciso a la hora de identificar revisores con conocimientos relevantes para el contenido del artículo.

- Integración de retroalimentación y aprendizaje automático: Un paso importante para mejorar el modelo sería integrar retroalimentación continua mediante técnicas de aprendizaje automático. Este enfoque podría ayudar a ajustar y optimizar diferentes partes del modelo en función de los resultados obtenidos, permitiendo su evolución constante. Además, la incorporación de Modelos de Lenguaje Largos podría mejorar la calidad y profundidad de las recomendaciones de revisores.

- Desarrollo de un modelo funcional accesible para revistas científicas: Finalmente, una meta a largo plazo sería desarrollar un modelo funcional que pueda ser utilizado por diferentes revistas científicas. Esto incluiría la recopilación de retroalimentación de los usuarios para poder identificar áreas de mejora, corregir fallas y evolucionar el sistema, adaptándolo a las necesidades de las comunidades científicas y editoriales.

Finalmente, se espera que esta metodología contribuya significativamente a la recomendación de revisores en diversas revistas científicas, trascendiendo el ámbito regional. Se destaca la adaptabilidad del modelo propuesto frente a la rigidez de los sistemas preestablecidos, ofreciendo una herramienta valiosa y flexible que se ajusta a las políticas editoriales particulares de cada publicación. Aunque las técnicas evolucionan con considerable rapidez, el marco teórico-metodológico fundamentado en la perspectiva sociotécnico-cultural aplicado en esta investigación demuestra, a partir de los resultados obtenidos, que este enfoque representa una mejora sustancial en el proceso de recomendación de revisores. Este modelo posee un alto potencial de aplicación en equipos editoriales que buscan mayor precisión y personalización en la asignación de evaluadores, optimizando así la calidad del proceso editorial científico.

Bibliografía

- Abad, M. F. (2017). El proceso de revisión por pares. En E. Abadal (Ed.), *Revistas científicas: Situación actual y retos de futuro* (pp. 53-71). Universitat de Barcelona.
<http://eprints.rclis.org/32131/>
- Abduljaleel, A. Q., & Al, E. (2021). Reviewer Assignment Using Weighted Matching and Hungarian Algorithm. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(7), Article 7. <https://doi.org/10.17762/turcomat.v12i7.2629>
- Aksoy, M., Yanik, S., & Amasyali, M. F. (2023). Reviewer Assignment Problem: A Systematic Review of the Literature. *Journal of Artificial Intelligence Research*, 76, 761-827.
<https://doi.org/10.1613/jair.1.14318>
- Alkazemi, B. Y. (2018). PRATO: An Automated Taxonomy-Based Reviewer-Proposal Assignment System. *Interdisciplinary Journal of Information, Knowledge, and Management*, 13, 383-396.
- Al-Zboon, S. A., Tawalbeh, S. K., Ai-Jarrah, H., Al-Asa'd, M., Hammad, M., & Al-Smadi, M. (2019). Resolving Conflict of Interests in Recommending Reviewers for Academic Publications Using Link Prediction Techniques. *2019 2nd International Conference on new Trends in Computing Sciences (ICTCS)*, 1-6.
<https://doi.org/10.1109/ICTCS.2019.8923033>
- Andrés, G. D., Martin, P. S. S., & Rodríguez, G. L. (2023). Análisis multidimensional de la sostenibilidad-DID en el contexto físico-virtual. *Cuadernos.info*, 54, Article 54.
<https://doi.org/10.7764/cdi.54.52515>
- authorCorporate:UNESCO. (2021). *Recomendación de la UNESCO sobre la Ciencia Abierta*.
https://unesdoc.unesco.org/ark:/48223/pf0000379949_spa
- Balog, K., Azzopardi, L., & de Rijke, M. (2006). Formal models for expert finding in enterprise corpora. *Proceedings of the 29th annual international ACM SIGIR conference on*

Research and development in information retrieval, 43-50.

<https://doi.org/10.1145/1148170.1148181>

Blei, D. M. (s. f.). *Latent Dirichlet Allocation*. 30.

Boehm, B. (1986). A spiral model of software development and enhancement. *SIGSOFT Softw.*

Eng. Notes, 11(4), 14-24. <https://doi.org/10.1145/12944.12948>

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with

Subword Information. *arXiv:1607.04606 [cs]*. <http://arxiv.org/abs/1607.04606>

Charlin, L., & Zemel, R. S. (2013). *The Toronto Paper Matching System: An automated paper-*

reviewer assignment system. [https://www.semanticscholar.org/paper/The-Toronto-](https://www.semanticscholar.org/paper/The-Toronto-Paper-Matching-System%3A-An-automated-Charlin-Zemel/d5fa5bd2ae89296acbfd9766f50c909f2c42effe)

[Paper-Matching-System%3A-An-automated-Charlin-](https://www.semanticscholar.org/paper/The-Toronto-Paper-Matching-System%3A-An-automated-Charlin-Zemel/d5fa5bd2ae89296acbfd9766f50c909f2c42effe)

[Zemel/d5fa5bd2ae89296acbfd9766f50c909f2c42effe](https://www.semanticscholar.org/paper/The-Toronto-Paper-Matching-System%3A-An-automated-Charlin-Zemel/d5fa5bd2ae89296acbfd9766f50c909f2c42effe)

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio,

Y. (2014). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical*

Machine Translation (arXiv:1406.1078). arXiv. <http://arxiv.org/abs/1406.1078>

Cuevas, R. F., & Mestanza, M. (2002). *LA EVALUACIÓN CIENTÍFICA Y EL SISTEMA DE REVISIÓN*

POR PARES.

http://200.62.146.19/bibvirtualdata/publicaciones/consejo/boletin46/art_interes.pdf

de Givry, S., Schiex, T., Schutt, A., & Simonis, H. (2020). *Modelling the conference paper*

assignment problem. <https://hdl.handle.net/10468/11447>

Decoppet, G. O., & Martin, P. S. S. (2024). Systematic mapping of automated reviewer

recommendation solutions. *Journal of Computer Science and Technology*, 24(2), Article

2. <https://doi.org/10.24215/16666038.24.e16>

Decoppet, G. O., & San Martín, P. S. S. (2025). Hacia una metodología flexible al contexto para

la recomendación automatizada de revisores. *e-Ciencias de la Información*.

<https://doi.org/10.15517/eci.v15i1.62426>

- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. <https://doi.org/10.48550/arXiv.1810.04805>
- D'hondt, J., Vertommen, J., Verhaegen, P.-A., Cattrysse, D., & Duflou, J. R. (2010). Pairwise-adaptive dissimilarity measure for document clustering. *Information Sciences*, 180(12), 2341-2358. <https://doi.org/10.1016/j.ins.2010.02.021>
- Dong, Z., Shu, Z., Zhen, D., Jie, C., Yanping, Z., & Jie, T. (2020). *A Multi-Label Classification Method Using a Hierarchical and Transparent Representation for Paper-Reviewer Recommendation* | *ACM Transactions on Information Systems*. <https://doi.org/doi.org/10.1145/3361719>
- Duan, Z., Tan, S., Zhao, S., Wang, Q., Chen, J., & Zhang, Y. (2019). Reviewer assignment based on sentence pair modeling. *Neurocomputing*, 366, 97-108. <https://doi.org/10.1016/j.neucom.2019.06.074>
- Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., & Joulin, A. (2020). *Beyond English-Centric Multilingual Machine Translation* (arXiv:2010.11125). arXiv. <https://doi.org/10.48550/arXiv.2010.11125>
- Guo, L., Wu, J., Chang, W., Wu, J., & Li, J. (2018). K-Loop Free Assignment in Conference Review Systems. *2018 International Conference on Computing, Networking and Communications (ICNC)*, 542-547. <https://doi.org/10.1109/ICCNC.2018.8390343>
- hiiamsid. (s. f.). *Hiiamsid/sentence_similarity_spanish_es* · *Hugging Face* (Versión n.d.) [Software]. Hugging Face. Recuperado 28 de mayo de 2025, de https://huggingface.co/hiiamsid/sentence_similarity_spanish_es
- Hofmann, T. (2013). *Probabilistic Latent Semantic Analysis* (arXiv:1301.6705). arXiv. <https://doi.org/10.48550/arXiv.1301.6705>

- Ishag, M. I. M., Park, K. H., Lee, J. Y., & Ryu, K. H. (2019). A Pattern-Based Academic Reviewer Recommendation Combining Author-Paper and Diversity Metrics. *IEEE Access*, 7, 16460-16475. IEEE Access. <https://doi.org/10.1109/ACCESS.2019.2894680>
- Jin, J., Geng, Q., Mou, H., & Chen, C. (2019). Author–Subject–Topic model for reviewer recommendation. *Journal of Information Science*, 45(4), 554-570. <https://doi.org/10.1177/0165551518806116>
- Kalmukov, Y. (2020a). A Software Tool for Accuracy Evaluation of Calculated Similarity Factors between Papers and Reviewers. *2020 7th International Conference on Energy Efficiency and Agricultural Engineering (EE&AE)*, 1-5. <https://doi.org/10.1109/EEAE49144.2020.9279032>
- Kalmukov, Y. (2020b). Automatic Assignment of Reviewers to Papers Based on Vector Space Text Analysis Model. *Proceedings of the 21st International Conference on Computer Systems and Technologies*, 229-235. <https://doi.org/10.1145/3407982.3408026>
- Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology*, 51(1), 7-15. <https://doi.org/10.1016/j.infsof.2008.09.009>
- Kobren, A., Saha, B., & McCallum, A. (2019). Paper Matching with Local Fairness Constraints. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1247-1257. <https://doi.org/10.1145/3292500.3330899>
- Kreutz, C. K., Balog, K., & Schenkel, R. (2022). Diverse Reviewer Suggestion for Extending Conference Program Committees. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 79-86. <https://doi.org/10.1145/3486622.3493931>
- Kreutz, C. K., & Schenkel, R. (2022). RevASIDE: Assignment of Suitable Reviewer Sets for Publications from Fixed Candidate Pools. *The 23rd International Conference on*

Information Integration and Web Intelligence, 57-68.

<https://doi.org/10.1145/3487664.3487673>

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25.
<https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html>

Kusumawardani, R. P., & Khairunnisa, S. O. (2018). Author-Topic Modelling for Reviewer Assignment of Scientific Papers in Bahasa Indonesia. *2018 International Conference on Asian Language Processing (IALP)*, 351-356.
<https://doi.org/10.1109/IALP.2018.8629124>

Latypova, V. (2023). Reviewer Assignment Decision Support in an Academic Journal based on Multicriteria Assessment and Text Mining. *2023 IX International Conference on Information Technology and Nanotechnology (ITNT)*, 1-4.
<https://doi.org/10.1109/ITNT57377.2023.10139187>

Le, Q. V., & Mikolov, T. (2014). *Distributed Representations of Sentences and Documents* (arXiv:1405.4053). arXiv. <https://doi.org/10.48550/arXiv.1405.4053>

Li, B. (2013, octubre 20). *Distance Weighted Cosine Similarity Measure for Text Classification*.
https://doi.org/10.1007/978-3-642-41278-3_74

Li, K., Cao, Z., & Qu, D. (2017). Fair Reviewer Assignment Considering Academic Social Network. En L. Chen, C. S. Jensen, C. Shahabi, X. Yang, & X. Lian (Eds.), *Web and Big Data* (pp. 362-376). Springer International Publishing. https://doi.org/10.1007/978-3-319-63579-8_28

Lian, J. W., Mattei, N., Noble, R., & Walsh, T. (2018). The Conference Paper Assignment Problem: Using Order Weighted Averages to Assign Indivisible Goods. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Article 1.
<https://doi.org/10.1609/aaai.v32i1.11484>

- Lin, Y.-S., Jiang, J.-Y., & Lee, S.-J. (2014). A Similarity Measure for Text Classification and Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 26(7), 1575-1590. *IEEE Transactions on Knowledge and Data Engineering*.
<https://doi.org/10.1109/TKDE.2013.19>
- Liu, X., Suel, T., & Memon, N. (2014). A robust model for paper reviewer assignment. *Proceedings of the 8th ACM Conference on Recommender systems*, 25-32.
<https://doi.org/10.1145/2645710.2645749>
- Manovich, L. (2020). *Cultural Analytics*.
<https://direct.mit.edu/books/monograph/4966/Cultural-Analytics>
- Martínez, G. S. (2012). La revisión por pares y la selección de artículos para publicación. *Revista Colombiana de Psicología*, 21(1), 27-35.
- Medakene, A. N., Bouanane, K., & Eddoud, M. A. (2019). A New Approach for Computing the Matching Degree in the Paper-to-Reviewer Assignment Problem. *2019 International Conference on Theoretical and Applicative Aspects of Computer Science (ICTAACS)*, 1, 1-8. <https://doi.org/10.1109/ICTAACS48474.2019.8988127>
- Meñica, J. M. A. (2023). El diálogo que somos. *Estudios Filosóficos*, 72(211), Article 211.
<https://estudiosfilosoficos.dominicos.org/ojs/article/view/1637>
- Miao, F., & Holmes, W. (2023). *Guía para el uso de IA generativa en educación e investigación* | UNESCO. UNESCO. <https://www.unesco.org/es/articles/guia-para-el-uso-de-ia-generativa-en-educacion-e-investigacion>
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality* (arXiv:1310.4546). arXiv.
<https://doi.org/10.48550/arXiv.1310.4546>
- Mimno, D., & McCallum, A. (2007). Expertise modeling for matching papers with reviewers. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 500-509. <https://doi.org/10.1145/1281192.1281247>

- Mirzaei, M., Sander, J., & Stroulia, E. (2019). Multi-Aspect Review-Team Assignment using Latent Research Areas. *Information Processing & Management*, 56(3), 858-878.
<https://doi.org/10.1016/j.ipm.2019.01.007>
- Mittal, K., Jain, A., Vaisla, K. S., & Tayal, D. K. (2020). A Novel method for Reviewer Assignment Problem based on Reviewers' Profile and Fuzzy Graph Connectivity Measure. 2020 *International Conference on Intelligent Engineering and Management (ICIEM)*, 386-391. <https://doi.org/10.1109/ICIEM48762.2020.9160042>
- Mou, H., Geng, Q., Jin, J., & Chen, C. (2015). An Author Subject Topic Model for Expert Recommendation. En G. Zuccon, S. Geva, H. Joho, F. Scholer, A. Sun, & P. Zhang (Eds.), *Information Retrieval Technology* (pp. 83-95). Springer International Publishing.
https://doi.org/10.1007/978-3-319-28940-3_7
- Nguyen, J., Sánchez-Hernández, G., Agell, N., Rovira, X., & Angulo, C. (2018). A decision support tool using Order Weighted Averaging for conference review assignment. *Pattern Recognition Letters*, 105, 114-120. <https://doi.org/10.1016/j.patrec.2017.09.020>
- ONU. (2015). *Transformar nuestro mundo: La Agenda 2030 para el Desarrollo Sostenible*.
<https://www.unfpa.org/es/resources/transformar-nuestro-mundo-la-agenda-2030-para-el-desarrollo-sostenible-0>
- Ortiz Núñez, R. (2024). *La revisión por pares como herramienta de evaluación de la ciencia: Principales tensiones y algunas propuestas alternativas | e-Ciencias de la Información*.
<https://doi.org/10.15517/eci.v14i1.55921>
- Patil, A. H., & Mahalle, P. N. (2021a). Fair and Accurate Review in Publication Process: A learning-based Proactive Approach for Assigning Reviewers to Manuscripts. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(4), Article 4.
<https://doi.org/10.17762/turcomat.v12i4.455>
- Patil, A. H., & Mahalle, P. N. (2021b). Multi-label Reviewer Profile Building and Ranking based on Expertise, Recency, Authority and h-index: Vital Module of Reviewer Paper

Assignment. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(6), Article 6.

- Patil, A. H., Mahalle, P. N., Patil, V. H., & Bhavsar, S. A. (2023). PROACTIVE SYSTEM FOR REVIEWER PAPER ASSIGNMENT. *Journal of Data Acquisition and Processing*, 38(2), 1569-1609. <https://doi.org/10.5281/zenodo.776779>
- Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543. <https://doi.org/10.3115/v1/D14-1162>
- Pradhan, D. K., Chakraborty, J., Choudhary, P., & Nandi, S. (2020). An automated conflict of interest based greedy approach for conference paper assignment system. *Journal of Informetrics*, 14(2), 101022. <https://doi.org/10.1016/j.joi.2020.101022>
- Pradhan, T., Sahoo, S., Singh, U., & Pal, S. (2021). A proactive decision support system for reviewer recommendation in academia. *Expert Systems with Applications*, 169, 114331. <https://doi.org/10.1016/j.eswa.2020.114331>
- Protasiewicz, J., Pedrycz, W., Kozłowski, M., Dadas, S., Stanisławek, T., Kopacz, A., & Gałęzewska, M. (2016). A recommender system of reviewers and experts in reviewing problems. *Knowledge-Based Systems*, 106, 164-178. <https://doi.org/10.1016/j.knosys.2016.05.041>
- Radford, A., & Narasimhan, K. (2018). *Improving Language Understanding by Generative Pre-Training*. <https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035>
- Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (arXiv:1908.10084). arXiv. <https://doi.org/10.48550/arXiv.1908.10084>
- Rigaux, P. (2004). *An iterative rating method: Application to web-based conference management / Request PDF*. ResearchGate. <http://dx.doi.org/10.1145/967900.968234>

- Riggs, T., & Wilensky, R. (2001). *An algorithm for automated rating of reviewers* | *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries (world)*. Pages 381-387.
<https://doi.org/doi.org/10.1145/379437.379731>
- Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., & Gatford, M. (1994). *Okapi at TREC-3*. 0.
- Rodriguez, M. A., & Bollen, J. (2008). An Algorithm to Determine Peer-Reviewers. *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*, 319.
<https://doi.org/10.1145/1458082.1458127>
- Rodriguez, M., Bollen, J., & Sompel, H. (2005). The Convergence of Digital-Libraries and the Peer-Review Process. *Journal of Information Science*, 32.
<https://doi.org/10.1177/0165551506062327>
- Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 487-494.
- Salinas, M., Giorgi, D., Ponchio, F., & Cignoni, P. (2020). ReviewerNet: A visualization platform for the selection of academic reviewers. *Computers & Graphics*, 89, 77-87.
<https://doi.org/10.1016/j.cag.2020.04.006>
<https://www.teseopress.com/laapropiacioncreativadelatecnologiaeneducacion/chapter/capitulo-xii-del-dispositivo-hipermedial-dinamico-dhd/>
- Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404, 132306.
<https://doi.org/10.1016/j.physd.2019.132306>
- Stelmakh, I., Shah, N., & Singh, A. (2021). PeerReview4All: Fair and accurate reviewer assignment in peer review. *The Journal of Machine Learning Research*, 22(1), 163:7393-163:7458.
- Tan, M., Dai, Z., Ren, Y., Walsh, T., & Aleksandrov, M. (2021). Minimal-envy Conference Paper Assignment: Formulation and a Fast Iterative Algorithm. *2021 5th Asian Conference on*

Artificial Intelligence Technology (ACAIT), 667-674.

<https://doi.org/10.1109/ACAIT53529.2021.9731163>

Tan, Z.-Y., Cai, N., & Feng, Y.-J. (2019). A Parallel Model to Analyze Effectiveness of Peer Review for Academic Journals. *Proceedings of the 2019 The 2nd International Conference on Robotics, Control and Automation Engineering*, 285-290.

<https://doi.org/10.1145/3372047.3372099>

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., & Fan, A. (2020).

Multilingual Translation with Extensible Multilingual Pretraining and Finetuning

(arXiv:2008.00401). arXiv. <https://doi.org/10.48550/arXiv.2008.00401>

Team, N., Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K.,

Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood,

A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., ... Wang, J. (2022). *No*

Language Left Behind: Scaling Human-Centered Machine Translation

(arXiv:2207.04672). arXiv. <https://doi.org/10.48550/arXiv.2207.04672>

Thada, V., & Jaglan, V. (2013). Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm.

International Journal of Innovations in Engineering and Technology, 2, 202-205.

Thomas, H. (2012). *Tecnologías para la inclusión social en América Latina: De las tecnologías apropiadas a los sistemas tecnológicos sociales. Problemas conceptuales y soluciones estratégicas* (pp. 25-76).

Thomas, H., Becerra, L., & Bidinost, A. (2019). ¿Cómo funcionan las tecnologías? Alianzas socio-técnicas y procesos de construcción de funcionamiento en el análisis histórico. *Pasado Abierto*, 5(10), Article 10.

<https://fh.mdp.edu.ar/revistas/index.php/pasadoabierto/article/view/3639>

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., &

Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, 30.

https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Verón, E. (2015). Teoría de la mediatización: Una perspectiva semio-antropológica. *CIC.*

Cuadernos de Información y Comunicación, 20, 173-182.

https://doi.org/10.5209/rev_CIYC.2015.v20.50682

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). *Multilingual E5 Text Embeddings: A Technical Report* (arXiv:2402.05672). arXiv.

<http://arxiv.org/abs/2402.05672>

Xu, L., Zeng, D., Dai, J., & Gui, L. (2022). Combining Coverage with TMPS for Reviewer

Assignment. *2022 International Conference on Intelligent Education and Intelligent*

Research (IEIR), 107-113. <https://doi.org/10.1109/IEIR56323.2022.10050060>

Yang, C., Liu, T., Yi, W., Chen, X., & Niu, B. (2020). Identifying expertise through semantic

modeling: A modified BBPSO algorithm for the reviewer assignment problem. *Applied*

Soft Computing, 94, 106483. <https://doi.org/10.1016/j.asoc.2020.106483>

Yang, K.-H., Kuo, T.-L., Lee, H.-M., & Ho, J.-M. (2009). A Reviewer Recommendation System

Based on Collaborative Intelligence. *2009 IEEE/WIC/ACM International Joint*

Conference on Web Intelligence and Intelligent Agent Technology, 1, 564-567.

<https://doi.org/10.1109/WI-IAT.2009.94>

Yao, Y., Dong, B., Zhang, A., Zhang, Z., Xie, R., Liu, Z., Lin, L., Sun, M., & Wang, J. (2022). *Prompt*

Tuning for Discriminative Pre-trained Language Models (arXiv:2205.11166). arXiv.

<https://doi.org/10.48550/arXiv.2205.11166>

Yeşilçimen, A., & Yıldırım, E. A. (2019). An alternative polynomial-sized formulation and an

optimization based heuristic for the reviewer assignment problem. *European Journal*

of Operational Research, 276(2), 436-450. <https://doi.org/10.1016/j.ejor.2019.01.035>

Yong, Y., Yao, Z., & Zhao, Y. (2021). A Framework for Reviewer Recommendation Based on

Knowledge Graph and Rules Matching. *2021 IEEE International Conference on*

Information Communication and Software Engineering (ICICSE), 199-203.

<https://doi.org/10.1109/ICICSE52190.2021.9404099>

Zahrotun, L. (2016). Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method. *Computer Engineering and Applications Journal*, 5, 11-18. <https://doi.org/10.18495/comengapp.v5i1.160>

Anexo A

En la Tabla A1 se puede observar la cantidad total de artículos en diferentes idiomas, luego del proceso de depuración inicial:

Tabla 42

Cantidad total de artículos en diferentes idiomas, luego del proceso de depuración inicial

Idioma	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
Cantidad artículos	3356	3083	250	3	20	0	0

En la Tabla A2 se muestra el registro de reconocimientos que definen áreas de artículos en base a títulos y a diferentes idiomas.

Tabla 43

Estadísticas de reconocimiento de palabras que definen áreas de artículos científicos por búsqueda de títulos de cada área y asociado a los diferentes idiomas reconocidos en los recursos

Español

Palabras Claves

Idioma	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- palabras clave	2052	2029	16	0	7	0	0
palabras clave -salto línea-	7	7	0	0	0	0	0
-salto línea- palabras-clave	56	52	2	0	2	0	0
palabras clave	540	537	2	0	1	0	0
p a l a b r a s c l a v e	3	3	0	0	0	0	0
p a l a b r a s	2	2	0	0	0	0	0
-salto línea- clave:	25	25	0	0	0	0	0
palabras claves:	2	2	0	0	0	0	0
palabras-claves:	1	1	0	0	0	0	0
TOTAL	2688	2658	20	0	10	0	0

Resumen

Idioma	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- resumen -salto línea-	940	934	5	0	1	0	0
resumen -salto línea-	93	90	3	0	0	0	0
resumen -salto línea-	64	62	1	0	1	0	0
resumen -salto línea-	859	849	8	0	2	0	0
-salto línea- resumen:	577	569	4	0	4	0	0
r e s u m e n	10	10	0	0	0	0	0
resúmen -salto línea-	3	3	0	0	0	0	0

-salto línea- resumen	170	170	0	0	0	0	0
TOTAL	2716	2687	21	0	8	0	0

Introducción

Idioma	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- 1- introducción	3	3	0	0	0	0	0
-salto línea- 1. introducción	250	250	0	0	0	0	0
1- introducción	1	1	0	0	0	0	0
i- introducción	3	3	0	0	0	0	0
1. introducción	55	55	0	0	0	0	0
1. introducción	90	90	0	0	0	0	0
1. presentación	18	18	0	0	0	0	0
introducción1	14	14	0	0	0	0	0
introducción...	3	3	0	0	0	0	0
-salto línea- presentación -salto línea-	33	33	0	0	0	0	0
-presentación -salto línea-	1	1	0	0	0	0	0
-salto línea- presentación:	6	6	0	0	0	0	0
presentación: -salto línea-	3	3	0	0	0	0	0
-salto línea- introducción -salto línea-	664	663	1	0	0	0	0
-salto línea- introducción -salto línea-	574	573	0	0	1	0	0
introducción -salto línea-	102	102	0	0	0	0	0
introducción -salto línea-	289	289	0	0	0	0	0
introducción -salto línea-	173	173	0	0	0	0	0
-salto línea- introducción:	25	25	0	0	0	0	0
-salto línea- introducción	80	79	1	0	0	0	0
introduccion -salto línea-	6	6	0	0	0	0	0
introduccion -salto línea-	7	7	0	0	0	0	0
-salto línea- introduccion	1	1	0	0	0	0	0
-salto línea- fundamentación -salto línea-	3	3	0	0	0	0	0
fundamentación -salto línea-	1	1	0	0	0	0	0
fundamentación -salto línea-	6	6	0	0	0	0	0
-salto línea- fundamentación	14	14	0	0	0	0	0
-salto línea- fundamentacion	1	1	0	0	0	0	0
presentación -salto línea-	7	7	0	0	0	0	0
presentación -salto línea-	7	7	0	0	0	0	0
presentación -salto línea-	19	19	0	0	0	0	0
-salto línea- presentación	60	60	0	0	0	0	0
-salto línea- introducci'on	4	4	0	0	0	0	0
introducci'on	2	2	0	0	0	0	0
TOTAL	2525	2522	2	0	1	0	0

Bibliografía

Idioma	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- referencias bibliográficas	795	795	0	0	0	0	0
referencias bibliográficas -salto línea-	87	87	0	0	0	0	0
referencias bibliográficas -salto línea-	50	50	0	0	0	0	0
-salto línea- referencias bibliograficas	2	2	0	0	0	0	0
-salto línea- referencias bibliográficas	3	3	0	0	0	0	0

referencias bibliograficas -salto línea-	1	1	0	0	0	0	0
referencias bibliográficas -salto línea-	1	1	0	0	0	0	0
lista de referencia -salto línea-	3	3	0	0	0	0	0
lista de referencia -salto línea-	1	1	0	0	0	0	0
-salto línea- lista de referencia	10	10	0	0	0	0	0
-salto línea- trabajos citados	1	1	0	0	0	0	0
trabajos citados -salto línea-	1	0	1	0	0	0	0
-salto línea- referencias documentales	1	1	0	0	0	0	0
referencias documentales -salto línea-	1	1	0	0	0	0	0
-salto línea- notas bibliográficas	1	1	0	0	0	0	0
bibliografía referida -salto línea-	1	1	0	0	0	0	0
bibliografía citada -salto línea-	15	15	0	0	0	0	0
bibliografía referenciada -salto línea-	1	1	0	0	0	0	0
-salto línea- referenciasbibliográficas	1	1	0	0	0	0	0
-salto línea- referencias citadas	3	3	0	0	0	0	0
-salto línea- bibliografía utilizada	1	1	0	0	0	0	0
bibliografía utilizada -salto línea-	1	1	0	0	0	0	0
-salto línea- referencia bibliográfica	10	10	0	0	0	0	0
-salto línea- referencias bibliográficas	1	1	0	0	0	0	0
-salto línea- citas bibliográficas	1	1	0	0	0	0	0
-salto línea- bibliografía:	21	21	0	0	0	0	0
bibliografía	1	1	0	0	0	0	0
bibliografía: -salto línea-	1	1	0	0	0	0	0
-salto línea- referencias	1	1	0	0	0	0	0
bibliografía utilizada: -salto línea-	1	1	0	0	0	0	0
-salto línea- referencias -salto línea-	432	432	0	0	0	0	0
-salto línea- referencias -salto línea-	436	435	0	0	1	0	0
referencias -salto línea-	66	64	2	0	0	0	0
-salto línea- referencias	166	166	0	0	0	0	0
referencias -salto línea-	80	80	0	0	0	0	0
referencias -salto línea-	10	10	0	0	0	0	0
referencias: -salto línea-	1	1	0	0	0	0	0
-salto línea- referencia -salto línea-	1	1	0	0	0	0	0
-salto línea- referencia -salto línea-	2	2	0	0	0	0	0
-salto línea- bibliografía -salto línea-	227	227	0	0	0	0	0
-salto línea- bibliografía	288	288	0	0	0	0	0
-salto línea- bibliografía	5	5	0	0	0	0	0
bibliografía. -salto línea-	2	2	0	0	0	0	0
bibliografía -salto línea-	86	86	0	0	0	0	0
bibliografía -salto línea-	36	36	0	0	0	0	0
bibliografía -salto línea-	4	4	0	0	0	0	0
-salto línea- bibliografía	12	11	0	0	1	0	0
bibliografía -salto línea-	9	9	0	0	0	0	0
bibliografía -salto línea-	1	1	0	0	0	0	0
-salto línea- bibliografía	3	3	0	0	0	0	0
fuentes de consulta -salto línea-	1	1	0	0	0	0	0
fuentes consultadas -salto línea-	2	2	0	0	0	0	0
re(cid:22) erencias	1	1	0	0	0	0	0
referencia bibliográfica	1	1	0	0	0	0	0

referencias -salto línea-	1	1	0	0	0	0	0
bilbiografía	1	1	0	0	0	0	0
biblografía	1	1	0	0	0	0	0
biobliografía	1	1	0	0	0	0	0
r e f e r e n c i a s	4	4	0	0	0	0	0
b i b l i o g r a f	1	1	0	0	0	0	0
referencias bibliográficas	12	12	0	0	0	0	0
(cid:35)(cid:42)(cid:35)(cid:45)(cid:42)(cid:48) (cid:40)(cid:51)(cid:34)(cid:39)(cid:202)(cid:34)	2	2	0	0	0	0	0
(cid:35)(cid:74)(cid:67)(cid:77)(cid:74)(cid:80) (cid:72)(cid:83)(cid:66)(cid:71)(cid:205)(cid:66) (cid:1)(cid:90)(cid:1)(cid:69)(cid:80)(cid:68)(cid:86) (cid:78)(cid:70)(cid:79)(cid:85)(cid:80)(cid:84)	1	1	0	0	0	0	0
(cid:35)(cid:74)(cid:67)(cid:77)(cid:74)(cid:80) (cid:72)(cid:83)(cid:66)(cid:71)(cid:205)(cid:66)	1	1	0	0	0	0	0
-salto línea- notas -salto línea-	6	6	0	0	0	0	0
notas -salto línea-	10	8	0	0	2	0	0
-salto línea- fuentes	7	6	1	0	0	0	0
-salto línea- lecturas sugeridas	9	9	0	0	0	0	0
notasbibliografía1.	1	1	0	0	0	0	0
.bibliografía	1	1	0	0	0	0	0
. bibliografía	3	3	0	0	0	0	0
TOTAL	2952	2944	4	0	4	0	0

La inclusión de codificaciones en la búsqueda de palabras corresponde a formatos de visualización de los títulos de los artículos en cuestión.

Inglés

Palabras Claves

Palabras	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- keyword	2100	1913	172	3	12	0	0
keywords -salto línea-	3	2	1	0	0	0	0
-salto línea- key words	516	496	19	0	1	0	0
-salto línea- key words	99	95	4	0	0	0	0
-salto línea- key-words	17	15	1	0	1	0	0
k e y w o r d	5	3	2	0	0	0	0
keywords:	19	15	4	0	0	0	0
TOTAL	2759	2539	203	3	14	0	0

Resumen

Palabras	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- abstract	2539	2333	191	3	12	0	0
abstract -salto línea-	70	69	1	0	0	0	0
a b s t r a c t	11	5	6	0	0	0	0
-salto línea- summary	73	69	4	0	0	0	0
summary -salto línea-	8	6	1	0	1	0	0
summary -salto línea-	1	0	1	0	0	0	0
TOTAL	2702	2482	204	3	13	0	0

Introducción

Idioma	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- introduction -salto línea-	61	0	60	1	0	0	0
introduction -salto línea-	12	0	12	0	0	0	0
introduction -salto línea-	37	0	37	0	0	0	0
introduction -salto línea-	89	5	82	2	0	0	0
-salto línea- introduction:	11	7	4	0	0	0	0
-salto línea- introduction	34	30	4	0	0	0	0
TOTAL	244	42	199	3	0	0	0

Bibliografía

Palabras	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- references	221	4	217	0	0	0	0
references -salto línea-	5	0	5	0	0	0	0
references -salto línea-	6	1	5	0	0	0	0
r e f e r e n c e s	1	0	1	0	0	0	0
-salto línea- reference	11	9	2	0	0	0	0
-salto línea- bibliography	6	1	5	0	0	0	0
bibliographic -salto línea-	1	0	1	0	0	0	0
referencies -salto línea-	1	0	1	0	0	0	0
r e f e r e n c e	1	0	1	0	0	0	0
TOTAL	253	15	238	0	0	0	0

Francés

Palabras Claves

Palabras	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- mots clés	36	34	1	1	0	0	0
-salto línea- mots-clés	36	27	6	2	1	0	0
TOTAL	72	61	7	3	1	0	0

Resumen

Palabras	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- résumé	73	62	7	3	1	0	0
résumé -salto línea-	2	1	1	0	0	0	0
TOTAL	75	63	8	3	1	0	0

Introducción

Idioma	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- introduction -salto línea-	61	0	60	1	0	0	0
introduction -salto línea-	12	0	12	0	0	0	0
introduction -salto línea-	37	0	37	0	0	0	0
introduction -salto línea-	89	5	82	2	0	0	0
-salto línea- introduction:	11	7	4	0	0	0	0
-salto línea- introduction	34	30	4	0	0	0	0
TOTAL	244	42	199	3	0	0	0

Bibliografía

Palabras	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- références	3	0	0	3	0	0	0
-salto línea- bibliographie	1	1	0	0	0	0	0
TOTAL	4	1	0	3	0	0	0

Portugués

Palabras Claves

Palabras	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- palavras chave	94	90	1	0	3	0	0
-salto línea- palavras-chave	466	438	15	0	13	0	0
palavras-chave	7	6	0	0	1	0	0
palavras chave:	4	4	0	0	0	0	0
p a l a v r a s - c h a v e	2	2	0	0	0	0	0
TOTAL	573	540	16	0	17	0	0

Resumen

Palabras	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- resumo	565	532	16	0	17	0	0
resumo -salto línea-	22	22	0	0	0	0	0
resumo:	4	4	0	0	0	0	0
TOTAL	591	558	16	0	17	0	0

Introducción

Idioma	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- introdução -salto línea-	5	3	0	0	2	0	0
introdução -salto línea-	1	0	0	0	1	0	0
introdução -salto línea-	3	0	0	0	3	0	0
introdução -salto línea-	10	3	1	0	6	0	0
-salto línea- introdução:	2	2	0	0	0	0	0
-salto línea- introdução	2	2	0	0	0	0	0
TOTAL	23	10	1	0	12	0	0

Bibliografía

Palabras	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- referências	58	43	1	0	14	0	0
referências -salto línea-	3	3	0	0	0	0	0
referências -salto línea-	1	1	0	0	0	0	0
-salto línea- referència	1	1	0	0	0	0	0
referència -salto línea-	1	0	0	0	1	0	0
TOTAL	64	48	1	0	15	0	0

Palabras Total Español Inglés Francés Portugués Catalán Italiano

Catalán

Palabras Claves

Palabras	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- paraules clau	5	5	0	0	0	0	0
TOTAL	5	5	0	0	0	0	0

Resumen

Palabras	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- resum	1	1	0	0	0	0	0
-salto línea- resum:	3	3	0	0	0	0	0
-salto línea- resum -salto línea-	2	2	0	0	0	0	0
TOTAL	6	6	0	0	0	0	0

Introducción

Idioma	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- introducció	1519	1516	2	0	1	0	0
TOTAL	1519	1516	2	0	1	0	0

Bibliografía

Palabras	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- notes	56	5	51	0	0	0	0
notes -salto línea-	3	2	1	0	0	0	0
TOTAL	59	7	52	0	0	0	0

Italiano

Palabras Claves

Palabras	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- palore chiave	0	12	0	0	0	0	0
parole chiave -salto línea-	0	9	0	0	0	0	0
TOTAL	0	21	0	0	0	0	0

Resumen

Palabras	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
TOTAL	0	0	0	0	0	0	0

Introducción

Idioma	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
TOTAL	0	0	0	0	0	0	0

Bibliografía

Palabras	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano
-salto línea- bibliografía	9	0	0	0	0	0	0
bibliografía -salto línea-	12	0	0	0	0	0	0
TOTAL	21	0	0	0	1	0	0

En la Tabla A3 se puede observar el resumen del análisis de los resultados obtenidos previamente.

Tabla 44

Resumen para análisis de los resultados obtenidos en los diferentes idiomas

	Idioma	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano		
	Cantidad artículos	3356	3083	250	3	20	0	0		
	Palabras	Total	Español	Inglés	Francés	Portugués	Catalán	Italiano	Promedio por tipo	Promedio general
Español	Palabras Claves	2688	2658	20	0	10	0	0	0,86	0,88
	Resumen	2716	2687	21	0	8	0	0	0,87	
	Introducción	2525	2522	2	0	1	0	0	0,82	
	Bibliografía	2952	2944	4	0	4	0	0	0,95	
Inglés	Palabras Claves	2759	2539	203	3	14	0	0	0,81	0,84
	Resumen	2702	2482	204	3	13	0	0	0,82	
	Introducción	244	42	199	3	0	0	0	0,80	
	Bibliografía	253	15	238	0	0	0	0	0,95	
Francés	Palabras Claves	72	61	7	3	1	0	0	1	1,00
	Resumen	75	63	8	3	1	0	0	1	
	Introducción	244	42	199	3	0	0	0	1	
	Bibliografía	4	1	0	3	0	0	0	1	
Portugués	Palabras Claves	573	540	16	0	17	0	0	0,85	0,76
	Resumen	591	558	16	0	17	0	0	0,85	
	Introducción	23	10	1	0	12	0	0	0,6	
	Bibliografía	64	48	1	0	15	0	0	0,75	

Catalán	Palabras Claves	5	5	0	0	0	0	0	No hay datos	No hay datos
	Resumen	6	6	0	0	0	0	0	No hay datos	
	Introducción	1519	1516	2	0	1	0	0	No hay datos	
	Bibliografía	59	7	52	0	0	0	0	No hay datos	
Italiano	Palabras Claves	0	21	0	0	0	0	0	No hay datos	No hay datos
	Resumen	0	0	0	0	0	0	0	No hay datos	
	Introducción	0	0	0	0	0	0	0	No hay datos	

Anexo B

Tabla 45

Información del proceso de detección de ubicación de diferentes áreas promedio con definición de umbral para corte y muestra de artículos fuera de umbral

Definición de umbral de búsqueda de resumen	0,12	Se incluye desde este valor hasta el inicio
---	------	---

promedio resumen	0,028
------------------	-------

cont.	id_recurso	% ubicación documento	ubic.	largo doc.
1	219216	0,87	61057	70215
2	113764	0,506	16067	31780
3	186094	0,435	21463	49287
4	189930	0,975	55397	56840
5	158550	0,259	23144	89303
6	240275	0,88	34377	39061
7	69418	0,3	12888	42907
8	76087	0,45	39339	87509
9	206330	0,144	15798	109922
10	66584	0,72	46532	64651
11	157920	0,697	38012	54567
12	185599	0,617	39873	64659
13	235336	0,128	5513	43030
14	125801	0,237	8060	34072
15	188031	0,996	19548	19626
16	166249	0,771	42695	55356
17	223479	0,768	51747	67419
18	145825	0,818	8654	10585
19	195462	0,975	39987	41005
20	64834	0,903	19130	21178
21	200558	0,641	46073	71879
22	140280	0,257	18719	72761

Definición de umbral de búsqueda de palabras claves	0,25	Se incluye desde este valor hasta el inicio
---	------	---

promedio palabras claves	0,052
--------------------------	-------

cont.	id_recurso	% ubicación documento	ubic.	largo doc.
1	85074	0,793	53049	66864

2	93392	0,993	47220	47559
3	79390	0,981	52755	53778
4	112700	0,899	63581	70741
5	106002	0,995	46771	46996
6	15575	0,997	125815	126194
7	39474	0,996	48860	49061
8	226977	0,939	53616	57109
9	176146	0,937	31717	33856
10	82260	0,98	59491	60699
11	179291	0,565	53957	95544
12	220420	0,584	50246	86000
13	52217	0,994	46170	46435
14	45272	0,981	55442	56513
15	155287	0,974	47258	48511
16	214739	0,403	23976	59487
17	113147	0,346	1723	4985
18	52909	0,994	43357	43599
19	95791	0,919	82839	90134
20	168537	0,731	32917	45014
21	14660	0,994	43664	43913
22	93774	0,995	69593	69966
23	70831	0,997	55384	55542
24	75814	0,953	270641	284130
25	210038	0,995	55183	55448
26	185693	0,953	65656	68894
27	192591	0,994	51615	51909
28	9014	0,476	25733	54109
29	210243	0,977	74718	76474
30	9937	0,993	47353	47683
31	155949	0,548	40645	74153
32	188822	0,996	54512	54718
33	27922	0,731	21178	28961
34	125603	0,288	11541	40022
35	51615	0,988	69956	70830
36	51702	0,988	69956	70830
37	92026	0,518	33857	65417
38	177604	0,975	55773	57229
39	3784	0,93	32763	35238
40	210997	0,572	19513	34111
41	65744	0,968	50587	52270
42	45589	0,965	57645	59761
43	83373	0,631	12343	19571
44	199010	0,264	11008	41694
45	60169	0,961	36007	37469
46	106430	0,997	76594	76856
47	34550	0,935	40161	42934
48	210970	0,908	33846	37261
49	162445	0,662	36295	54818

50	70832	0,998	68456	68562
51	116744	0,262	13589	51895

Definición de umbral de búsqueda de introducción	0,25
--	------

Se incluye desde este valor hasta el inicio

promedio ubicación de búsqueda de introducción	10,34
--	-------

cont.	id_recurso	% ubicación documento	ubic.	largo doc.
1	20530	0.52145107686995550988	24163	46338
2	225416	0.92200879426058779398	15936	17284
3	224190	0.52357046651157090889	4276	8167
4	181594	0.32465270022781078652	7268	22387
5	24689	0.51071550338512494704	20971	41062
6	221957	0.51852833941000997520	15644	30170
7	213843	0.54025892011367226964	5133	9501
8	162445	0.68632565945492352011	37623	54818
9	211280	0.65359487934370918971	39517	60461
10	65107	0.76706947884610576782	32328	39969

promedio ubicación de búsqueda de bibliografía	0,6
--	-----

Se incluye desde este valor hasta el final

promedio palabras claves	0,863
--------------------------	-------

cont.	id_recurso	% ubicación documento	ubic.	largo doc.
1	178213	0,562	55410	98564
2	164266	0,586	42835	73051
3	222380	0,597	33850	56718
4	172979	0,59	28640	48504
5	152980	0,564	24964	44275
6	33517	0,546	34387	63036
7	203090	0,593	39520	66672
8	195947	0,593	32163	54198
9	95791	0,588	53003	90134
10	122665	0,565	35896	63532
11	220313	0,596	44013	73857
12	85812	0,541	50970	94299
13	193501	0,514	43755	85182
14	226101	0,6	43346	72301
15	207569	0,58	83968	144764
16	132462	0,516	36794	71289
17	211552	0,547	31923	58387
18	46317	0,512	41363	80789
19	194792	0,551	24764	44907
20	21000	0,592	31899	53846

21	70541	0,537	27135	50569
22	176514	0,522	38692	74149
23	60177	0,549	42439	77318
24	223509	0,585	27901	47658
25	1293	0,502	275737	549788

Anexo C

Al efectuar la primera instancia de desambiguación se consiguieron 757 perfiles confirmados y la distribución sobre la discriminación por perfiles encontrados por autor es la que se observa en la Tabla C1:

Tabla 46

Perfiles confirmados luego de primera etapa de búsqueda de perfiles

Cantidad de devoluciones en búsqueda de nombres de investigadores	Perfiles Confirmados
1	623
2	51
3	22
4	5
5	8
6	8
7	2
8	5
9	1
10	32
Total perfiles confirmados	757

Dentro del proceso para encontrar coincidencias, se probó hacer una simplificación y modificación de nombres de artículos quitando caracteres especiales principalmente y se consiguieron 235 coincidencias más, como se muestra en la Tabla C2:

Tabla 47

Devolución de perfiles previa y posterior a la modificación de títulos

Trabajo sobre artículos	Cantidad
Títulos iguales	522
Modificación de títulos	235
Total	757

En la segunda instancia no hubo resultados positivos y en la tercera se consiguieron 24 perfiles más, quedando al final el resultado que se observa en la Tabla C3.

Tabla 48*Perfiles confirmados por cada etapa de procesamiento*

Total de perfiles confirmados	Cantidad
etapa 1	757
etapa 2	0
etapa 3	24
Total	781

Anexo D

Manuscrito a revisar: Nombre de manuscrito

=====

Nombres: Candidato a revisor 1 en orden alfabético

ID conicet: Valor

Lugar de Trabajo: Valor

Título: Valor

Grado: Valor

Citas	Citas desde 2018	h-index	h-index desde 2018	i10-index	i10-index desde 2018
Valor	Valor	Valor	Valor	Valor	Valor

Citas anuales	
año	citas
Año x	Valor
Año x + 1	Valor

.....	
Año x + 1	Valor

Producción científica a incluir

id_recurso	similitud	orden	fecha	titulo_x	revista
Valor	Valor	Valor	Valor	Valor	Valor
Valor	Valor	Valor	Valor	Valor	Valor

.....					
Valor	Valor	Valor	Valor	Valor	Valor

Revisiones anteriores

Manuscrito	Fecha solicitud	Aceptación	Fecha aceptación	Fecha de entrega	Evaluación	Observación
Valor	Valor	Valor	Valor	Valor	Valor	Valor
Valor	Valor	Valor	Valor	Valor	Valor	Valor

.....						
Valor	Valor	Valor	Valor	Valor	Valor	Valor

Nombres: Candidato a revisor 2 en orden alfabético

ID conicet: Valor

Lugar de Trabajo: Valor

Título: Valor

Grado: Valor

Citas	Citas desde 2018	h-index	h-index desde 2018	i10-index	i10-index desde 2018
Valor	Valor	Valor	Valor	Valor	Valor

Citas anuales	
año	citas
Año x	Valor
Año x + 1	Valor

Año x + 1	Valor

Producción científica a incluir

id_recurso	similitud	orden	fecha	titulo_x	revista
Valor	Valor	Valor	Valor	Valor	Valor
Valor	Valor	Valor	Valor	Valor	Valor

Valor	Valor	Valor	Valor	Valor	Valor

Revisiones anteriores

Manuscrito	Fecha solicitud	Aceptación	Fecha aceptación	Fecha de entrega	Evaluación	Observación
Valor	Valor	Valor	Valor	Valor	Valor	Valor
Valor	Valor	Valor	Valor	Valor	Valor	Valor

Valor	Valor	Valor	Valor	Valor	Valor	Valor

=====

=====

Nombres: Candidato a revisor n en orden alfabético

ID conicet: Valor

Lugar de Trabajo: Valor

Título: Valor

Grado: Valor

Citas	Citas desde 2018	h-index	h-index desde 2018	i10-index	i10-index desde 2018
Valor	Valor	Valor	Valor	Valor	Valor

Citas anuales	
año	citas
Año x	Valor
Año x + 1	Valor

Año x + 1	Valor

Producción científica a incluir

id_recurso	similitud	orden	fecha	titulo_x	revista
------------	-----------	-------	-------	----------	---------

Valor	Valor	Valor	Valor	Valor	Valor
Valor	Valor	Valor	Valor	Valor	Valor

Valor	Valor	Valor	Valor	Valor	Valor
-------	-------	-------	-------	-------	-------

Revisiones anteriores

Manuscrito	Fecha solicitud	Aceptación	Fecha aceptación	Fecha de entrega	Evaluación	Observación
Valor	Valor	Valor	Valor	Valor	Valor	Valor
Valor	Valor	Valor	Valor	Valor	Valor	Valor

Valor	Valor	Valor	Valor	Valor	Valor	Valor
-------	-------	-------	-------	-------	-------	-------

=====

Informe final priorizado

orden	puntaje	autor	importante
1	Valor	Candidato a revisor	Si hay posible conflicto de intereses se informa acá
2	Valor	Candidato a revisor	Si hay posible conflicto de intereses se informa acá

i	Valor	Candidato a revisor	Si hay posible conflicto de intereses se informa acá
---	-------	---------------------	--

Investigadores fuera del informe

autor	puntaje	motivo
Candidato a revisor	Valor	Descripción del/los motivos
Candidato a revisor	Valor	Descripción del/los motivos

Candidato a revisor	Valor	Descripción del/los motivos
---------------------	-------	-----------------------------