IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 2, February 2024

DOI: 10.17148/IJARCCE.2024.13214

Natural Language Processing in Scientific Literature Mining: Advancements, Applications, and Challenges

Dr. Jaynesh H. Desai[1]

Assistant Professor, Bhagwan Mahavir College of Computer Application, Bhagwan Mahavir University, Surat, Gujarat, India[1]

Abstract: Natural Language Processing (NLP) techniques have revolutionized the extraction of valuable information from vast repositories of scientific literature. This paper aims to provide an in-depth analysis of the applications, methodologies, and challenges associated with leveraging NLP in scientific literature mining. It explores the advancements in NLP algorithms, their application in knowledge discovery, text summarization, entity recognition, and sentiment analysis within the context of scientific literature. Additionally, this paper addresses the challenges, such as domain-specific language complexities, data scarcity, and ethical considerations, while proposing potential solutions to further enhance the efficacy of NLP in scientific literature mining. The evolution of the Exposome concept revolutionised the research in exposure assessment and epidemiology by introducing the need for a more holistic approach on the exploration of the relationship between the environment and disease. At the same time, further and more dramatic changes have also occurred on the working environment, adding to the already existing dynamic nature of it. Natural Language Processing (NLP) refers to a collection of methods for identifying, reading, extracting and untimely transforming large collections of language. In this work, we aim to give an overview of how NLP has successfully been applied thus far in Exposome research. Methods: We conduct a literature search on PubMed, Scopus and Web of Science for scientific articles published between 2011 and 2021.

Keywords: Text Mining, Information Extraction, Scientific Document Analysis, Named Entity Recognition (NER),Topic Modeling, Sentiment Analysis, Machine Learning in NLP, Data Mining, Bioinformatics, Knowledge Discovery, natural language processing; exposure research; exposome; machine learning.

I. INTRODUCTION

Overview of Natural Language Processing (NLP) in scientific literature mining - Natural Language Processing (NLP) is a branch of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language. In the context of scientific literature mining, NLP involves the application of computational techniques and algorithms to analyse, process, and derive meaningful information from a vast amount of textual data present in scientific documents, research papers, articles, and journals.NLP techniques encompass various methodologies such as text mining, information retrieval, machine learning, and linguistics. These techniques aid in extracting valuable information, identifying patterns, and uncovering hidden relationships within scientific texts, enabling researchers and scientists to access, analyse, and utilize the wealth of knowledge available in these repositories more efficiently.Natural Language Processing (NLP) stands as a crucial research field within Artificial Intelligence (AI), aiming to equip computers with the ability to comprehend natural language, both spoken and written, akin to human understanding [1]. This interdisciplinary domain integrates computational linguistics, statistical analysis, machine learning, and deep learning to achieve its objectives [2]. The term "Exposome," introduced by [3], delineates a research area focused on systematically measuring a person's lifelong exposures, encompassing factors such as occupational, physical environment, and socio-economic elements, to understand their impact on health outcomes [3]. Despite its conceptual significance, the integration of the term "Exposome" into exposure research is incomplete, often replaced by the more generic term "exposure research" [4]. Concurrently, there is a rising trend in applying text mining and NLP techniques in various exposure-related studies. While extensive surveys exist for NLP and its subtasks [5–7], a comprehensive review of NLP and text mining applications in occupational and environmental exposure research is absent. This review addresses this gap by presenting an overview of tools utilizing NLP and text mining techniques in this specific research domain. Employing a hybrid approach that combines traditional and automated review methods, including the use of RobotAnalyst [8], a recently developed web-based software integrating text mining and machine learning algorithms, we scrutinize papers from PubMed, Scopus, and WoS databases. The aim is to answer key research questions and provide insights into the current landscape of NLP applications in the realm of occupational and environmental exposure research.

- What are the most common text mining and NLP approaches used in exposure assessment research?

DOI: 10.17148/IJARCCE.2024.13214

- What resources are used for this task?
- What are the most common NLP methods used?
- What are the main challenges and future directions of research?

IMPORTANCE OF NLP IN EXTRACTING INSIGHTS FROM VAST SCIENTIFIC REPOSITORIES

Scientific literature comprises an immense volume of information across diverse domains, including medicine, biology, physics, engineering, and more. This wealth of knowledge is pivotal for advancements in research, innovation, and problem-solving across various fields.However, the sheer volume and complexity of scientific literature pose significant challenges for researchers in accessing, comprehending, and extracting relevant information. NLP plays a crucial role in addressing these challenges by automating tasks such as information extraction, summarization, categorization, and semantic analysis.

NLP techniques facilitate:

- Efficient search and retrieval of relevant literature.
- Summarization and abstraction of lengthy texts for quick comprehension.
- Extraction of key entities, relationships, and trends from vast amounts of data.
- Cross-referencing and analysis of connections between different research papers or disciplines.
- Enabling data-driven insights and discoveries by uncovering hidden patterns or correlations within the literature.

By utilizing NLP tools and methodologies, researchers can navigate through extensive scientific literature more effectively, saving time, and uncovering valuable insights that could potentially lead to groundbreaking discoveries or advancements in various scientific domains.In summary, the introduction highlights how NLP serves as a vital tool in scientific literature mining, aiding researchers in overcoming the challenges posed by vast repositories of information, and enabling them to extract, analyse, and utilize knowledge more efficiently, ultimately contributing to advancements in scientific research and innovation.

NLP Techniques in Scientific Literature Mining

NLP techniques have been instrumental in unlocking valuable insights from scientific literature by employing various methodologies tailored for processing textual data inherent in these documents.

- ***Text preprocessing methods for scientific documents*** - serve as the foundation of NLP in scientific literature mining. This involves a series of steps such as tokenization, stemming, and lemmatization to standardize the text. Tokenization breaks down text into smaller units like words or phrases, while stemming and lemmatization reduce words to their root forms, aiding in normalization and better analysis. Additionally, techniques like stop-word removal, where common words with little semantic value are eliminated, contribute to refining the dataset for analysis.
- ***Entity recognition and extraction*** - involve identifying and extracting essential elements from text, such as names of people, organizations, locations, dates, and scientific terms. Named Entity Recognition (NER) techniques leverage machine learning algorithms to identify and classify these entities, allowing for the extraction of specific information critical for understanding relationships and contextualizing scientific findings.
- ***Topic modelling and document clustering*** - are pivotal for organizing and summarizing scientific documents. Topic modelling algorithms, such as Latent Dirichlet Allocation (LDA), help in uncovering latent topics within a collection of documents, enabling researchers to understand prevalent themes or subjects across scientific literature. Document clustering techniques group similar documents together, aiding in efficient categorization, exploration, and retrieval of relevant information based on similarities in content or context.

Sentiment analysis and opinion mining- are emerging techniques in scientific literature mining. While more prevalent in social media or consumer reviews, these methods are increasingly applied in understanding the tone, attitudes, or opinions expressed in scientific texts. This analysis can uncover subjective insights, preferences, or evaluations present in research articles, contributing to a nuanced understanding of the scientific community's perspectives on certain topics or findings.In essence, these NLP techniques collectively enable researchers to preprocess, extract, organize, and analyse scientific text efficiently, facilitating comprehensive exploration and utilization of vast scientific repositories for knowledge extraction and discovery

Example: Entity Recognition in Biomedical Research

Biomedical research involves an enormous amount of scientific literature, making it challenging for researchers to identify and extract crucial information efficiently. NLP techniques play a vital role in simplifying this process by recognizing and extracting essential entities such as genes, proteins, diseases, and their interactions from a multitude of research articles.For instance, consider a scenario where researchers are exploring the relationship between a specific gene, let's say "BRCA1," and breast cancer. Utilizing NLP techniques like Named Entity Recognition (NER) and information extraction, a system can automatically scan through numerous scientific articles and identify instances where "BRCA1" is mentioned in relation to breast cancer.

***Named Entity Recognition (NER)*:**
NER algorithms trained on biomedical text recognize mentions of genes, proteins, diseases, etc., within the scientific literature.In this case, the system identifies "BRCA1" as a gene entity mentioned in various articles.

- Information Extraction:Once "BRCA1" is identified, further extraction techniques can gather additional information related to its interactions, functions, mutations, or associations specifically concerning breast cancer.The system can identify sentences or paragraphs discussing "BRCA1 mutations and breast cancer risk" or "BRCA1's role in specific breast cancer treatments."
- By employing these NLP techniques, researchers can:Quickly gather information from a vast number of articles about the relationship between "BRCA1" and breast cancer.Create structured databases or summaries detailing the various contexts in which "BRCA1" is discussed in relation to breast cancer, aiding in comprehensive analysis and literature review.Facilitate the discovery of potential new insights or connections between "BRCA1" and breast cancer, contributing to ongoing research in the field.This real-world example demonstrates how NLP techniques, specifically entity recognition and extraction, streamline the process of mining biomedical literature, enabling researchers to efficiently extract relevant information crucial for advancing our understanding of complex relationships between genes, diseases, and treatments.

Applications of NLP in Scientific Literature Mining:

1. ***Knowledge discovery and information retrieval-***

Explanation: NLP facilitates efficient discovery and retrieval of knowledge from extensive scientific literature by enabling precise search capabilities and information extraction.
Example: Imagine a researcher interested in exploring the efficacy of a particular drug for treating a specific disease. NLP-powered search engines can swiftly retrieve relevant scientific articles discussing the drug's effectiveness, its side effects, dosage, patient outcomes, and comparative studies, helping the researcher gain insights for further analysis.

2. ***Summarization and abstraction techniques-***

Explanation: NLP techniques aid in condensing lengthy scientific documents into concise summaries, capturing the essential information while retaining the context and key findings.
Example: Consider a research paper discussing a groundbreaking study on climate change impacts. NLP-driven summarization tools can generate condensed summaries, highlighting the main hypotheses, methodologies, significant findings, and implications, allowing other researchers to grasp the essence of the study quickly.

3. ***Relationship extraction and network analysis-***

Explanation: NLP enables the extraction and analysis of relationships between entities, helping researchers comprehend complex networks and connections within scientific literature.Example: In the field of genetics, NLP can identify and analyse relationships between genes, proteins, diseases, and their interactions from a multitude of articles. For instance, NLP can identify the relationships between specific genes and their associations with various diseases, facilitating the understanding of genetic pathways or disease mechanisms.

4. ***Biomedical and life sciences applications*** -

Explanation: NLP plays a pivotal role in biomedical research by aiding in the analysis of vast amounts of scientific literature in areas like genomics, drug discovery, clinical trials, and disease research.Example: In biomedical studies, NLP techniques can be used to extract information about drug interactions, adverse effects, or patient outcomes from clinical trial reports. This enables researchers and healthcare professionals to access critical information for evidence-based decision-making in patient care and drug development.In summary, NLP applications in scientific literature mining encompass diverse areas, ranging from knowledge discovery and summarization to relationship extraction and specific domain applications like biomedical research. These applications showcase how NLP significantly contributes to efficiently accessing, processing, and deriving valuable insights from extensive scientific repositories, fostering advancements and discoveries across various domains of science and research.

ADVANCEMENTS AND INNOVATIONS

Deep learning models in NLP for literature mining - Deep learning models, particularly neural networks with multiple layers, have revolutionized NLP by significantly improving the performance of various tasks in literature mining. Techniques like Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and more notably, Transformer models (e.g., BERT, GPT) have shown exceptional capabilities in understanding contextual nuances, semantic relationships, and language patterns within scientific texts.These models excel in:

- Semantic Understanding: Capturing context and meaning from textual data, allowing for more accurate entity recognition, relationship extraction, and summarization.
- Representation Learning: Learning rich representations of text, enabling better information retrieval and knowledge discovery by understanding latent patterns and connections.

DOI: 10.17148/IJARCCE.2024.13214

For instance, Transformer-based models like BERT have been applied to biomedical literature mining, enabling more accurate entity recognition, thereby aiding in the extraction of critical biomedical entities such as genes, proteins, and diseases.

Transfer learning approaches for domain adaptation - Transfer learning, a technique where models trained on one domain are fine-tuned for another, addresses challenges posed by domain-specific language in scientific literature. Pre-trained language models, initially trained on large corpora, are fine-tuned using domain-specific data to adapt them to a particular field of study.
This approach offers:
Improved Performance: By leveraging pre-trained models' general language understanding and adapting them to specific scientific domains, it enhances performance in tasks like entity recognition, summarization, and relationship extraction.
Reduced Data Dependency: Enables effective learning even with limited annotated data in specialized domains, making it feasible for smaller research communities or less studied fields.For example, a pre-trained language model like GPT-3, fine-tuned on a dataset of environmental science papers, can better understand and process texts in the environmental science domain, aiding in more accurate extraction of environmental concepts or relationships.

Multimodal approaches integrating text and other data sources- Integrating multiple data modalities (e.g., text, images, graphs) has gained traction in NLP for literature mining. Combining textual information with other data sources such as images, graphs, or tables from scientific documents allows for a more comprehensive understanding and analysis of the content.Benefits include:

- Enhanced Context: Integration of diverse data types provides a richer context, improving information extraction and comprehension.
- Improved Inference: Leveraging complementary information from multiple modalities enables more robust inference and knowledge synthesis.

For instance, in biological research, combining textual information from research articles with graphical representations of protein-protein interactions enhances the understanding of complex biological networks and pathways.In essence, these advancements showcase how leveraging deep learning, transfer learning, and multimodal approaches in NLP for literature mining significantly enhances the comprehension, analysis, and utilization of scientific literature across various domains, enabling more accurate extraction of knowledge and fostering innovation in research and discovery.

CHALLENGES AND LIMITATIONS

1. ***Domain-specific language nuances and jargon****-*
Explanation: Scientific literature often contains specialized terminology, domain-specific jargon, abbreviations, and complex linguistic structures that may pose challenges for standard NLP models.
Impact: Such nuances can lead to difficulties in accurately interpreting or recognizing specialized entities, concepts, or relationships, impacting the effectiveness of information extraction and analysis.
Example: In fields like medicine or genomics, where terminology is highly specialized, NLP models may struggle with identifying and contextualizing domain-specific terms or acronyms, leading to errors in entity recognition or relationship extraction.

2. ***Limited labelled datasets and data sparsity*** *–*
Explanation: Developing NLP models requires large, well-labelled datasets for training, but scientific literature in specific domains might have limited annotated or labelled data.
Impact: Insufficient labelled data hampers the training of accurate domain-specific NLP models, affecting their ability to generalize and perform effectively in specialized scientific domains.
Example: In niche research areas with limited published literature or annotations, creating comprehensive labelled datasets becomes challenging, limiting the development and performance of NLP models tailored for those domains.

3. Ethical considerations in mining sensitive information-
Explanation: Scientific literature often contains sensitive or confidential information, such as patient data, proprietary research findings, or potentially harmful content.
Impact: Ethical concerns arise regarding privacy, consent, and responsible handling of sensitive data while mining scientific texts, necessitating measures to prevent misuse or unauthorized access to sensitive information.
Example: Extracting and analysing patient-related data from medical literature for research purposes must comply with ethical standards and data privacy regulations to safeguard individuals' confidentiality and prevent unauthorized use.

Addressing these challenges involves various strategies:

➢ Developing specialized NLP models or adapting existing ones to comprehend domain-specific language nuances.

➢ Exploring semi-supervised or unsupervised learning methods to mitigate the reliance on labelled datasets.

DOI: 10.17148/IJARCCE.2024.13214

➢ Implementing robust data anonymization and access control mechanisms to ensure ethical handling of sensitive information in scientific texts.

In conclusion, while NLP offers tremendous potential in mining scientific literature, challenges related to domain-specific language complexities, data availability, and ethical considerations underline the need for specialized approaches, responsible data practices, and ongoing research to enhance the effectiveness and ethical usage of NLP techniques in scientific literature mining.

FUTURE DIRECTIONS AND POTENTIAL SOLUTIONS

- ***Enhanced domain adaptation strategies -***

Explanation: To overcome limitations arising from domain-specific language nuances and the scarcity of labelled data, enhanced domain adaptation strategies aim to improve the adaptability of NLP models to specialized domains.

Solutions: Develop techniques that facilitate more effective transfer learning, enabling NLP models pre-trained on general corpora to adapt and fine-tune more efficiently to specific scientific domains. This includes domain-adversarial training, where models learn to distinguish between domain-specific and generic language, leading to improved adaptation.

Example: Advancements in unsupervised domain adaptation methods allow models to transfer knowledge from high-resource domains to low-resource domains more effectively, improving the performance of NLP models in understanding domain-specific contexts within scientific literature.

- ***Incorporation of context-aware models -***

Explanation: Context-aware models aim to enhance the understanding of linguistic nuances and contextual relationships present in scientific texts, addressing issues related to complex language structures and varied context usage.

Solutions: Develop NLP models that dynamically adjust their understanding based on contextual cues within scientific literature. This involves leveraging attention mechanisms, contextual embeddings, or memory-augmented networks to capture and utilize context more effectively.

Example: Context-aware models, like Transformer architectures with attention mechanisms, can better capture long-range dependencies and contextual information within scientific texts, aiding in more accurate entity recognition, relationship extraction, and summarization.

- ***Collaboration for data sharing and standardization -***

Explanation: The limited availability of labelled datasets in specialized scientific domains hinders the development of robust NLP models. Collaborative efforts for data sharing and standardization aim to mitigate data scarcity issues.

Solutions: Encourage collaboration among research institutions, publishers, and stakeholders to create standardized annotated datasets. Facilitate open-access repositories for labelled scientific data to encourage sharing, enabling the development and evaluation of NLP models across diverse scientific domains.

Example: Initiatives like PubMed Central in biomedicine or arXiv in various scientific fields promote open access to scientific literature. Expanding such repositories to include annotated datasets can significantly benefit NLP research in scientific literature mining.

These future directions and potential solutions emphasize the need for continuous advancements in NLP methodologies, fostering collaboration among researchers, and embracing responsible data practices to overcome existing challenges and enhance the effectiveness of NLP techniques in scientific literature mining. Through these initiatives, the field can progress towards more accurate, adaptable, and ethically sound applications in extracting knowledge from scientific texts.

Case Studies and Practical Implementations

Highlighting successful applications of NLP in specific scientific domains

1) **Biomedical and Life Sciences:**

Applications: NLP in biomedical research involves tasks like information extraction, entity recognition, and relationship extraction from vast amounts of biomedical literature.

Example: **Clinical Named Entity Recognition** - NLP models can accurately identify and extract named entities from clinical text, such as diseases, symptoms, treatments, and medications. For instance, tools like MetaMap leverage NLP techniques to extract clinical concepts from electronic health records (EHRs), enabling researchers and healthcare professionals to analyse patient data efficiently.

2) **Environmental Sciences:**

Applications: NLP aids in analysing climate-related literature, extracting climate change impacts, identifying mitigation strategies, and synthesizing complex environmental research.

Example: **Climate Change Text Mining** - NLP tools process vast amounts of climate science literature to identify trends, patterns, and climate change impacts. For example, Climate Pulse uses NLP to extract information from climate-related documents, aiding policymakers and researchers in understanding the implications of climate change on various ecosystems and human activities.

DOI: 10.17148/IJARCCE.2024.13214

3) **Computer Science and AI Research:**

Applications: NLP assists in analysing research papers, summarizing methodologies, tracking trends, and facilitating information retrieval in computer science and AI.

Example: **Semantic Scholar** - This tool uses NLP to analyse and summarize scientific papers in computer science and AI. Semantic Scholar aids researchers in discovering relevant papers, extracting key information, and visualizing research trends, enhancing the efficiency of literature review and staying updated with advancements in the field.

4) **Material Science and Engineering:**

Applications: NLP techniques help extract materials properties, synthesis methods, and relationships between materials from scientific articles in materials science.

Example: **Materials Discovery** - NLP tools extract materials-related data from scientific literature, aiding material scientists in identifying new materials or properties. For instance, Materials Graph Network (MEGNet) uses NLP-driven methods to predict materials properties based on information extracted from materials science literature, contributing to the discovery of novel materials for various applications.

In these scientific domains, NLP plays a pivotal role in facilitating information extraction, knowledge synthesis, and trend analysis from extensive scientific literature. The applications mentioned demonstrate how NLP techniques are leveraged to extract valuable insights, accelerate research processes, and aid decision-making across diverse scientific disciplines, showcasing the versatility and impact of NLP in scientific literature mining.

Real-world examples demonstrating the impact of NLP in literature mining-

- **PubMed and MEDLINE (Biomedical Literature):**

Real-world Impact: PubMed, a widely used database of biomedical literature, uses NLP for indexing and information retrieval. MEDLINE, a subset of PubMed, employs NLP algorithms to index and search millions of biomedical articles, aiding researchers in accessing relevant scientific literature efficiently.

Effect: Researchers worldwide rely on PubMed and MEDLINE for literature searches in medicine, biology, and related fields. NLP-driven indexing and search capabilities enhance the accessibility of scientific literature, facilitating advancements in medical research, diagnoses, and treatment development.

- **CORD-19 (COVID-19 Research Literature):**

Real-world Impact: During the COVID-19 pandemic, the CORD-19 dataset was created by applying NLP techniques to compile research articles related to the virus. NLP tools extracted, categorized, and made this vast collection of articles accessible to researchers worldwide.

Effect: This dataset accelerated COVID-19 research, aiding in understanding the virus, vaccine development, epidemiology, and treatment strategies. NLP-driven efforts facilitated quick access to critical information during a global health crisis.

- **Semantic Scholar and ArXiv (Scientific Papers):**

Real-world Impact: Platforms like Semantic Scholar and arXiv use NLP to analyse and categorize scientific papers, extract key information, and provide enhanced search functionalities to researchers.

Effect: Researchers benefit from these platforms as NLP techniques help them discover relevant papers efficiently, access summarized information, and explore emerging trends across scientific disciplines, leading to informed decision-making and innovative research discoveries.

- **Climate Change Research and Analysis:**

Real-world Impact: NLP tools are employed in analyzing climate change-related literature, extracting information, and synthesizing complex findings, aiding in understanding climate impacts and mitigation strategies.

Effect: Climate scientists and policymakers benefit from NLP-driven analysis, enabling them to comprehend climate change patterns, assess environmental risks, and formulate evidence-based policies for mitigating climate-related challenges. These real-world examples demonstrate the practical impact of NLP in literature mining across diverse fields. NLP-driven tools and datasets have significantly improved access to information, accelerated research processes, and facilitated knowledge discovery, thereby contributing to advancements and decision-making in various scientific domains.

II. REVIEW METHODOLOGY

This literature review involved a comprehensive search across three scientific literature databases to identify relevant articles for the study. The initial search yielded a total of 6420 articles, and after removing duplicates, 5957 articles were retained for pre-screening. Figure 1 illustrates the selection process, outlining the queries used on PubMed, Scopus, and Web of Science platforms. The queries included terms such as "natural language processing," "text mining," "text-mining," "text and data mining," "ontology," "lexic*," "corpus," and "corpora," combined with terms related to exposure research, such as "exposome," "exposure," "socioexposome," and "risk factor" in conjunction with terms related to work,

DOI: 10.17148/IJARCCE.2024.13214

occupation, or environment. The pre-screening process consisted of two steps. Initially, to streamline the workload, RobotAnalyst [8], a web-based software system employing text mining and machine learning methods, identified 998 full papers based on their relevance. The system utilized an iterative classification process, making decisions based on abstract content. Subsequently, manual screening of titles and abstracts was conducted, adhering to inclusion and exclusion criteria. These criteria, provided by two occupational exposure experts, aimed to select studies pertinent to occupational exposure research. Following this process, 80 papers specifically focusing on text mining and/or natural language processing in exposure research were identified. Further scrutiny involved a review of the full papers to assess their relevance to occupational exposure and the utilization of NLP or text mining methods. Out of the identified papers, 40 were obtained and thoroughly reviewed, ultimately resulting in 37 articles that met the defined inclusion and exclusion criteria for this study.

Inclusion criteria:

- Original work;
- Study exposures concerning humans;
- Study occupational and/or environmental exposures of humans, such as airborne agents (e.g., particulates or substances and biological agents (viruses)), stressors, psycho-social and physical (e.g., muscle-skeletal) exposures as well as workplace accidents;
- Have their full texts available;
- Are written in English;
- Focus on text mining or natural language processing and their texts containing a method, experiments and result section.

Exclusion criteria:

- Studied animal or plant exposures;
- Studied drug, nutrition or dietary exposures on humans;
- Written in another language than English;
- Commentaries, opinion papers or editorials.

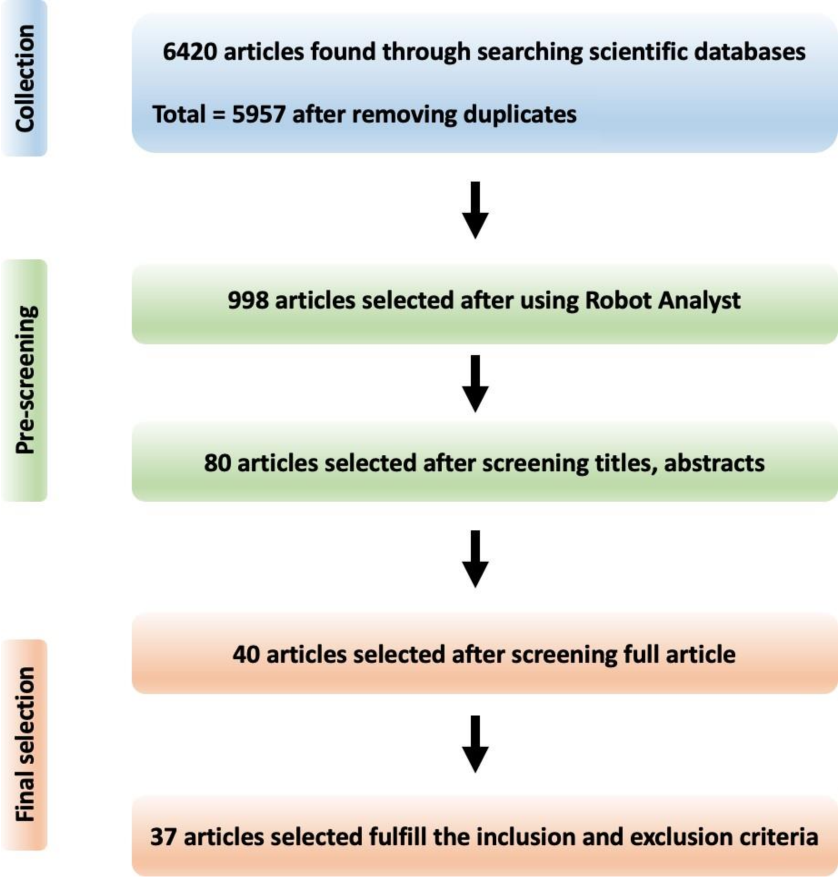
Collection
6420 articles found through searching scientific databases
Total = 5957 after removing duplicates
Pre-screening
998 articles selected after using Robot Analyst
80 articles selected after screening titles, abstracts
Final selection
40 articles selected after screening full article
37 articles selected fulfill the inclusion and exclusion criteria

Figure 1. Overview of article selection process used in this narrative literature review.

DOI: 10.17148/IJARCCE.2024.13214

III. RESULTS

In this section, we provide a condensed summary of the literature review findings, emphasizing the utilized resources, computational methods, and existing Natural Language Processing (NLP) tools. Figure 2 visually represents the annual publication trends, revealing a noticeable upward trajectory over time. Additionally, Table 1 categorizes each paper based on the NLP tools employed, the resources utilized, and the computational methods applied.

The graph in Figure 2 illustrates a consistent increase in the number of publications each year, indicating a growing interest in the field. Table 1 serves as a comprehensive categorization, offering insights into the diversity of NLP tools, resources, and computational techniques used across the reviewed papers. Furthermore, we provide a succinct overview of literature reviews and qualitative research within this domain, shedding light on the broader context and existing qualitative insights in the area under scrutiny.

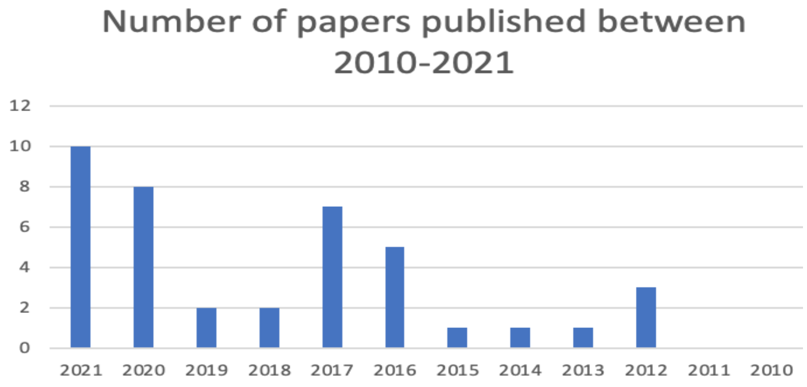

Figure 2. Number of NLP papers applied to occupational exposure research published each year from 2010 to 2021.

Table 1. A categorisation of each paper based on tools used, resources and computational methods.

Tool used	Papers
NLTK	[9–13],[14]
Other	[9,15–18],[19–21],[13,22–24],[25–27]
Not declared	[15,28–32],[33–37],[38–40]

Resources	Papers
Scientific literature	[12,15,28,29,41,42],[14,22,23,31],[24,43,44],[19–21,33,34,45], [35,36,46,47]
Existing Database	[13,30,35,37,45]
Twitter	[11,18,39]
EHR	[9,21,48]

Computational Method	Papers
Machine learning	[9,15,28,41],[10,12,28,29],[13,17,30,42,48],[11,14,18,22,31],[23–25,27]
Knowledge based	[19–21,43,44]
Database creation and fusion	[27,33–35,45,46], [36,37,47]
Rule-based algorithms	[27,40]

A. Resources

There are different types of resources used, where the most common resource is the existing scientific literature (see Figure 3). Other data sources include databases, social media platforms, electronic health records and accident reports (see Table 1).

B. Computational Methods

Overall, there are four main categories of computational approaches used which include machine learning, knowledge-based approaches, and database creation and fusion approaches. Figure 4 shows the split of computational approaches found in this review.

C. Existing NLP tools

There are a number of different existing NLP preprocessing tools used (see Figure 5), where NLTK [49] is the most commonly used for preprocessing textual data. Given the vast number of different NLP tools used in other studies, we have summarised the tools as '*Other*'. However, it has to be noted that a large amount of studies did not declare the type of text mining tool that was used in their work.

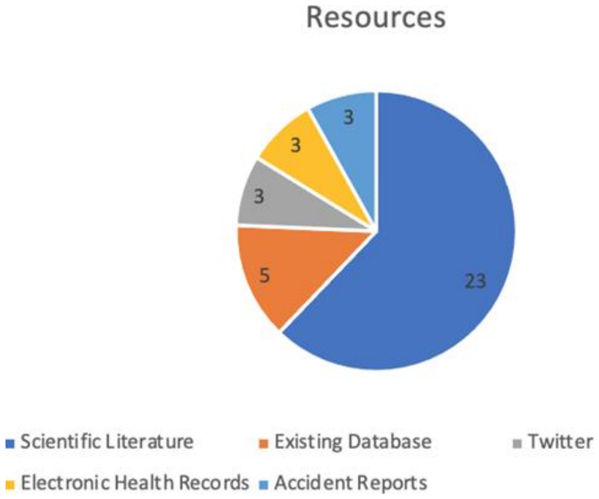

Figure 3. A chart showing the different types of resources used in the selected articles.

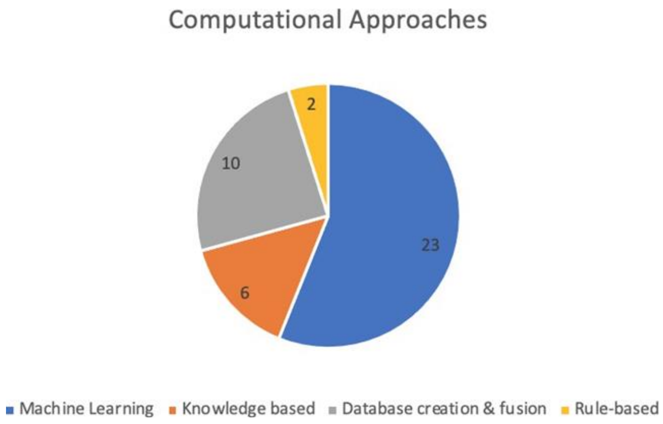

Figure 4. A chart showing the computational methods utilised in the selected articles.

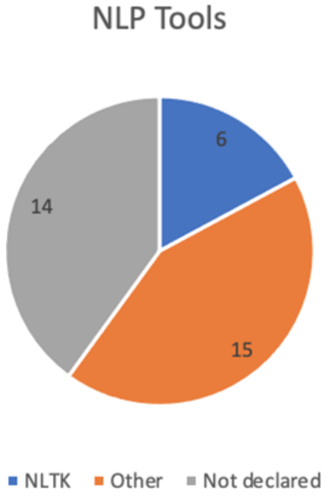

Figure 5. A chart showing a summary of the different types of NLP tools in each article.

IV. CONCLUSION

Recapitulation of the significance of NLP in scientific literature mining-

- **Efficient Information Retrieval**: NLP techniques enable precise search capabilities, facilitating the rapid retrieval of relevant scientific literature from expansive databases. Researchers can access pertinent information swiftly, streamlining their literature review processes and aiding in knowledge discovery.
- **Accurate Entity Recognition and Extraction**: NLP models excel in identifying and extracting essential entities, such as genes, proteins, diseases, or materials, from scientific texts. This capability helps in comprehending relationships, patterns, and trends, empowering researchers to derive insights and make connections critical for scientific advancements.
- **Enhanced Summarization and Abstraction**: NLP tools condense lengthy scientific documents into concise summaries, capturing the essence of research findings. These summaries aid in quick comprehension, enabling researchers to navigate complex literature more effectively and extract essential information efficiently.
- **Facilitation of Cross-Disciplinary Insights**: By analysing scientific texts from various disciplines, NLP fosters interdisciplinary connections. It enables researchers to synthesize information from diverse fields, uncover hidden correlations, and derive interdisciplinary insights crucial for innovation and problem-solving.
- **Accelerated Research Processes**: The automation of tasks such as information extraction, literature review, and trend analysis through NLP expedites research processes. Researchers can focus on higher-level analysis and innovation rather than spending extensive time sifting through large volumes of literature manually.
- **Contribution to Decision-Making and Policy Formulation**: NLP aids policymakers, healthcare professionals, and scientists in making informed decisions by providing access to synthesized information. It supports evidence-based decision-making in various fields, ranging from healthcare to environmental policy.
- **Global Accessibility to Knowledge**: Platforms utilizing NLP, like PubMed or Semantic Scholar, enhance access to scientific literature worldwide. These tools democratize knowledge, enabling researchers globally to access, analyse, and contribute to the collective scientific knowledge pool.

In conclusion, the significance of NLP in scientific literature mining is profound. It empowers researchers, enhances accessibility to information, accelerates discoveries, and fosters interdisciplinary collaboration. As NLP methodologies continue to evolve, their impact on scientific research and innovation is poised to further revolutionize how we extract, understand, and utilize information from scientific texts, driving progress across diverse scientific domains.

Future prospects and challenges to be addressed for further advancements

Future Prospects:

- **Refinement of Domain-Specific Models**: Developing NLP models specifically tailored to different scientific domains to enhance accuracy in recognizing domain-specific jargon, terminology, and relationships.
- **Integration of Multimodal Information**: Advancing NLP techniques to effectively analyse and synthesize information from diverse modalities (text, images, graphs) in scientific literature for a more comprehensive understanding.
- **Continued Development of Contextual Understanding**: Improving NLP models' ability to understand nuanced context, discourse, and subtle meaning shifts in scientific texts, enabling more accurate information extraction and interpretation.
- **Ethical AI in Literature Mining**: Ensuring responsible usage of NLP in literature mining, adhering to ethical standards, privacy regulations, and addressing biases in algorithms to prevent misuse or unethical practices.

Challenges to Address[50]:

- **Domain-Specific Data Scarcity**: Access to labelled datasets for training NLP models in specialized scientific domains remains limited, hindering the development of robust domain-specific models.
- **Complexity of Scientific Language**: Coping with the intricacies of domain-specific terminology, jargon, and linguistic nuances within scientific literature that pose challenges for accurate information extraction and comprehension.
- **Ethical and Privacy Concerns**: Managing sensitive information within scientific texts ethically, ensuring data privacy, and establishing guidelines for responsible data usage and sharing.
- **Generalization across Diverse Domains**: Developing NLP models that can generalize effectively across different scientific disciplines while also catering to specific domain requirements.
- **Interdisciplinary Knowledge Integration**: Addressing challenges in integrating information across multiple scientific disciplines, considering variations in terminologies and methodologies used.
- **Continued Model Interpretability**: Enhancing the interpretability of NLP models to ensure transparency and reliability in decision-making processes based on information extracted from scientific texts.

Addressing these challenges will pave the way for future advancements in NLP for scientific literature mining. Collaborative efforts among researchers, investments in domain-specific datasets, ethical guidelines, and continuous innovation in NLP methodologies are key to unlocking the full potential of NLP in advancing scientific research and innovation.

REFERENCES

[1]. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.

[2]. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Prentice Hall: Hoboken, NJ, USA, 2002.

[3]. Wild, C.P. The exposome: From concept to utility. *Int. J. Epidemiol.* **2012**, *41*, 24–32. [CrossRef] [PubMed]

[4]. Haddad, N.; Andrianou, X.D.; Makris, K.C. A scoping review on the characteristics of human exposome studies. *Curr. Pollut. Rep.* **2019**, *5*, 378–393. [CrossRef]

[5]. Kreimeyer, K.; Foster, M.; Pandey, A.; Arya, N.; Halford, G.; Jones, S.F.; Forshee, R.; Walderhaug, M.; Botsis, T. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J. Biomed. Inform.* **2017**, *73*, 14–29. [CrossRef] [PubMed]

[6]. Chowdhury, G.G. Natural language processing. *Annu. Rev. Inf. Sci. Technol.* **2003**, *37*, 51–89. [CrossRef]

[7]. Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; Huang, X. Pre-trained models for natural language processing: A survey. *Sci. China Technol. Sci.* **2020**, *63*, 1872–1897. [CrossRef]

[8]. Przybyła, P.; Brockmeier, A.J.; Kontonatsios, G.; Le Pogam, M.A.; McNaught, J.; von Elm, E.; Nolan, K.; Ananiadou, S. Prioritising references for systematic reviews with RobotAnalyst: A user study. *Res. Synth. Methods* **2018**, *9*, 470–488. [CrossRef]

[9]. Balasubramanian, V.; Vivekanandhan, S.; Mahadevan, V. Pandemic tele-smart: A contactless tele-health system for efficient monitoring of remotely located COVID-19 quarantine wards in India using near-field communication and natural language processing system. Med. Biol. Eng. Comput. 2021, 60, 61–79 [CrossRef]

[10]. Dong, T.; Yang, Q.; Ebadi, N.; Luo, X.R.; Rad, P. Identifying Incident Causal Factors to Improve Aviation Transportation Safety: Proposing a Deep Learning Approach. J. Adv. Transp. 2021, 2021, 5540046. [CrossRef]

[11]. Medina Sada, D.; Mengel, S.; Gittner, L.S.; Khan, H.; Rodriguez, M.A.P.; Vadapalli, R. A Preliminary Investigation with Twitter to Augment CVD Exposome Research. In Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, Austin, TX, USA, 5–8 December 2017; pp. 169–178.

[12]. Lee, S.W.; Kwon, J.H.; Lee, B.; Kim, E.J. Scientific Literature Information Extraction Using Text Mining Techniques for Human Health Risk Assessment of Electromagnetic Fields. Sens. Mater. 2020, 32, 149–157. [CrossRef]

[13]. Lamurias, A.; Jesus, S.; Neveu, V.; Salek, R.M.; Couto, F.M. Information Retrieval using Machine Learning for Biomarker Curation in the Exposome-Explorer. bioRxiv 2020, 6, 689264. . [CrossRef]

[14]. Larsson, K.; Baker, S.; Silins, I.; Guo, Y.; Stenius, U.; Korhonen, A.; Berglund, M. Text mining for improved exposure assessment.

[15]. PLoS ONE 2017, 12, e0173132. [CrossRef] [PubMed]

[16]. Tewari, S.; Toledo Margalef, P.; Kareem, A.; Abdul-Hussein, A.; White, M.; Wazana, A.; Davidge, S.T.; Delrieux, C.; Connor, K.L. Mining Early Life Risk and Resiliency Factors and Their Influences in Human Populations from PubMed: A Machine Learning Approach to Discover DOHaD Evidence. J. Pers. Med. 2021, 11, 1064. [CrossRef] [PubMed]

[17]. Varghese, A.; Cawley, M.; Hong, T. Supervised clustering for automated document classification and prioritization: A case study using toxicological abstracts. Environ. Syst. Decis. 2018, 38, 398–414. [CrossRef]

[18]. Li, J.; Wang, J.; Xu, N.; Hu, Y.; Cui, C. Importance degree research of safety risk management processes of urban rail transit based on text mining method. Information 2018, 9, 26. [CrossRef]

[19]. Leroy, G.; Harber, P.; Revere, D. Public sharing of medical advice using social media: An analysis of Twitter. Grey J. (TGJ) 2016,

[20]. 12, 104–113.

[21]. Karystianis, G.; Buchan, I.; Nenadic, G. Mining characteristics of epidemiological studies from Medline: A case study in obesity.

[22]. J. Biomed. Semant. 2014, 5, 22. [CrossRef]

[23]. Karystianis, G.; Thayer, K.; Wolfe, M.; Tsafnat, G. Evaluation of a rule-based method for epidemiological document classification towards the automation of systematic reviews. J. Biomed. Inform. 2017, 70, 27–34. [CrossRef]

[24]. Fan, J.w.; Li, J.; Lussier, Y.A. Semantic modeling for exposomics with exploratory evaluation in clinical context. J. Healthc. Eng.

[25]. 2017, 2017, 3818302 . [CrossRef]

DOI: 10.17148/IJARCCE.2024.13214

[26]. Ali, I.; Guo, Y.; Silins, I.; Högberg, J.; Stenius, U.; Korhonen, A. Grouping chemicals for health risk assessment: A text mining-based case study of polychlorinated biphenyls (PCBs). Toxicol. Lett. 2016, 241, 32–37. [CrossRef]

[27]. Davis, A.P.; Wiegers, T.C.; Johnson, R.J.; Lay, J.M.; Lennon-Hopkins, K.; Saraceni-Richards, C.; Sciaky, D.; Murphy, C.G.; Mattingly,

[28]. C.J. Text mining effectively scores and ranks the literature for improving chemical-gene-disease curation at the comparative toxicogenomics database. PLoS ONE 2013, 8, e58201. [CrossRef]

[29]. Vishnyakova, D.; Pasche, E.; Gobeill, J.; Gaudinat, A.; Lovis, C.; Ruch, P. Classification and prioritization of biomedical literature for the comparative toxicogenomics database. In Proceedings of the MIE, Pisa, Italy, 26–29 August 2012; pp. 210–214.

[30]. Lu, Y.; Xu, H.; Peterson, N.B.; Dai, Q.; Jiang, M.; Denny, J.C.; Liu, M. Extracting epidemiologic exposure and outcome terms from literature using machine learning approaches. Int. J. Data Min. Bioinform. 2012, 6, 447–459. [CrossRef] [PubMed]

[31]. Giummarra, M.J.; Lau, G.; Gabbe, B.J. Evaluation of text mining to reduce screening workload for injury-focused systematic reviews. Inj. Prev. 2020, 26, 55–60. [CrossRef]

[32]. Warth, B.; Spangler, S.; Fang, M.; Johnson, C.H.; Forsberg, E.M.; Granados, A.; Martin, R.L.; Domingo, X.; Huan, T.; Rinehart, D.; et al. Exposing the Exposome with Global Metabolomics and Cognitive Computing. bioRxiv 2017, 145722. [CrossRef]

[33]. Berrang-Ford, L.; Sietsma, A.J.; Callaghan, M.; Minx, J.C.; Scheelbeek, P.F.; Haddaway, N.R.; Haines, A.; Dangour, A.D. Systematic mapping of global research on climate and health: A machine learning review. Lancet Planet. Health 2021, 5, e514–e525. [CrossRef]

[34]. Minet, E.; Haswell, L.E.; Corke, S.; Banerjee, A.; Baxter, A.; Verrastro, I.; e Lima, F.D.A.; Jaunky, T.; Santopietro, S.; Breheny, D.; et al. Application of text mining to develop AOP-based mucus hypersecretion genesets and confirmation with in vitro and clinical samples. Sci. Rep. 2021, 11, 6091. [CrossRef]

[35]. Taboureau, O.; El M'Selmi, W.; Audouze, K. Integrative systems toxicology to predict human biological systems affected by exposure to environmental chemicals. Toxicol. Appl. Pharmacol. 2020, 405, 115210. [CrossRef]

[36]. Russ, D.E.; Ho, K.Y.; Colt, J.S.; Armenti, K.R.; Baris, D.; Chow, W.H.; Davis, F.; Johnson, A.; Purdue, M.P.; Karagas, M.R.; et al. Computer-based coding of free-text job descriptions to efficiently identify occupations in epidemiological studies. Occup. Environ. Med. 2016, 73, 417–424. [CrossRef]

[37]. Semenza, J.C.; Herbst, S.; Rechenburg, A.; Suk, J.E.; Höser, C.; Schreiber, C.; Kistemann, T. Climate change impact assessment of food-and waterborne diseases. Crit. Rev. Environ. Sci. Technol. 2012, 42, 857–890. [CrossRef]

[38]. Zhao, F.; Li, L.; Chen, Y.; Huang, Y.; Keerthisinghe, T.P.; Chow, A.; Dong, T.; Jia, S.; Xing, S.; Warth, B.; et al. Risk-Based Chemical Ranking and Generating a Prioritized Human Exposome Database. Environ. Health Perspect. 2021, 129, 047014. [CrossRef]

[39]. Dong, Z.; Fan, X.; Li, Y.; Wang, Z.; Chen, L.; Wang, Y.; Zhao, X.; Fan, W.; Wu, F. A Web-Based Database on Exposure to Persistent Organic Pollutants in China. Environ. Health Perspect. 2021, 129, 057701. [CrossRef]

[40]. Rugard, M.; Coumoul, X.; Carvaillo, J.C.; Barouki, R.; Audouze, K. Deciphering adverse outcome pathway network linked to bisphenol F using text mining and systems toxicology approaches. Toxicol. Sci. 2020, 173, 32–40. [CrossRef] [PubMed]

[41]. Barupal, D.K.; Fiehn, O. Generating the blood exposome database using a comprehensive text mining and database fusion approach. Environ. Health Perspect. 2019, 127, 097008. [CrossRef] [PubMed]

[42]. Wishart, D.; Arndt, D.; Pon, A.; Sajed, T.; Guo, A.C.; Djoumbou, Y.; Knox, C.; Wilson, M.; Liang, Y.; Grant, J.; et al. T3DB: The toxic exposome database. Nucleic Acids Res. 2015, 43, D928–D934. [CrossRef] [PubMed]

[43]. Zhang, H.; Hu, H.; Diller, M.; Hogan, W.R.; Prosperi, M.; Guo, Y.; Bian, J. Semantic Standards of External Exposome Data. Environ. Res. 2021, 197, 111185. [CrossRef]

[44]. Ekenga, C.C.; McElwain, C.A.; Sprague, N. Examining public perceptions about lead in school drinking water: A mixed-methods analysis of Twitter response to an environmental health hazard. Int. J. Environ. Res. Public Health 2018, 15, 162. [CrossRef]

[45]. Hollister, B.M.; Restrepo, N.A.; Farber-Eger, E.; Crawford, D.C.; Aldrich, M.C.; Non, A. Development and performance of text-mining algorithms to extract socioeconomic status from de-identified electronic health records. In Pacific Symposium on Biocomputing 2017; World Scientific: Singapore, 2017; pp. 230–241.

[46]. Hartmann, J.; Wuijts, S.; van der Hoek, J.P.; de Roda Husman, A.M. Use of literature mining for early identification of emerging contaminants in freshwater resources. Environ. Evid. 2019, 8, 33. [CrossRef]

[47]. Cawley, M.; Beardslee, R.; Beverly, B.; Hotchkiss, A.; Kirrane, E.; Sams II, R.; Varghese, A.; Wignall, J.; Cowden, J. Novel text analytics approach to identify relevant literature for human health risk assessments: A pilot study with health effects of in utero exposures. Environ. Int. 2020, 134, 105228. [CrossRef]

[48]. Jornod, F.; Rugard, M.; Tamisier, L.; Coumoul, X.; Andersen, H.R.; Barouki, R.; Audouze, K. AOP4EUpest: Mapping of pesticides in adverse outcome pathways using a text mining tool. Bioinformatics 2020, 36, 4379–4381. [CrossRef]

DOI: 10.17148/IJARCCE.2024.13214

[49]. Kiossogloua, P.; Bordaa, A.; Graya, K.; Martin-Sancheza, F.; Verspoora, K.; d Lopez-Camposa, G. Characterising the Scope of Exposome Research: A Generalisable Approach; IOS Press: Amsterdam, The Netherlands, 2017.

[50]. Davis, A.P.; Grondin, C.J.; Johnson, R.J.; Sciaky, D.; Wiegers, J.; Wiegers, T.C.; Mattingly, C.J. Comparative toxicogenomics database (CTD): Update 2021. Nucleic Acids Res. 2021, 49, D1138–D1143. [CrossRef]

[51]. Zgheib, E.; Kim, M.J.; Jornod, F.; Bernal, K.; Tomkiewicz, C.; Bortoli, S.; Coumoul, X.; Barouki, R.; De Jesus, K.; Grignard, E.; et al. Identification of non-validated endocrine disrupting chemical characterization methods by screening of the literature using artificial intelligence and by database exploration. Environ. Int. 2021, 154, 106574. [CrossRef]

[52]. Ayadi, A.; Auffan, M.; Rose, J. Ontology-based NLP information extraction to enrich nanomaterial environmental exposure database. Procedia Comput. Sci. 2020, 176, 360–369. [CrossRef]

[53]. Schwartz, K.L.; Achonu, C.; Buchan, S.A.; Brown, K.A.; Lee, B.; Whelan, M.; Wu, J.H.; Garber, G. Epidemiology, clinical characteristics, household transmission, and lethality of severe acute respiratory syndrome coronavirus-2 infection among healthcare workers in Ontario, Canada. PLoS ONE 2020, 15, e0244477. [CrossRef] [PubMed]

[54]. Loper, E.; Bird, S. Nltk: The natural language toolkit. arXiv 2002, arXiv:cs/0205028.

[55]. Rani, J.; Shah, A.R.; Ramachandran, S. pubmed. mineR: An R package with text-mining algorithms to analyse PubMed abstracts.

[56]. J. Biosci. 2015, 40, 671–682. [CrossRef] [PubMed]