Improving Publishing: Extracting Keywords and Clustering Topics

Yosua Setyawan Soekamto[1,*], Indra Maryati[2], Christian[3], Edwin Kurniawan[4]

[1,2,3,4] *Department of Information Systems, Universitas Ciputra Surabaya, Surabaya, 60219, Indonesia*

(Received: March 28, 2024; Revised: April 20, 2024; Accepted: April 26, 2024; Available online: May 17, 2024)

Abstract

Humans, by nature, are inclined to share knowledge across various platforms, such as educational institutions, media outlets, and specialized research publications like journals and conferences. The consistent oversight and evaluation of these publications by ranking bodies serve to maintain the integrity and quality of scholarly discourse on a global scale. However, there has been a decline in the proliferation of such publications in recent times, partly attributed to ethical misconduct within specific segments of the scholarly community. Despite implementing systems such as the Open Journal System (OJS), publishers grapple with the formidable task of managing editorial and review processes. Compounding the multifaceted nature of scholarly content, manual review procedures often lead to considerable time investment. Thus, a pressing need exists for advanced technological solutions to streamline the article selection process, empowering publishers to prioritize articles for review based on topical relevance. This study advocates adopting a comprehensive framework integrating advanced text analysis techniques such as keyword extraction, topic clustering, and summarization algorithms. These tools can be implemented and integrated by connecting with the database of the existing system. By leveraging these tools with the expertise of editorial and review teams, publishers can significantly expedite the initial assessment of submitted articles. Given the rapid technological advancements, publishers must embrace robust systems that enhance efficiency and effectiveness, particularly in reviewer assignments and article prioritization. This research employs the neural network approach of BERT and K-Means clustering to perform keyword extraction and topic clustering. Furthermore, using BERT facilitates accurate semantic understanding and context-aware representation of textual data. Additionally, BERT's pre-trained models enable its fine-tuning capability to allow customization to specific domains or tasks. By harnessing the power of BERT, publishers can gain deeper insights into the content of scholarly articles, leading to more informed decision-making and improved publication outcomes.

Keywords: Keyword Extraction, Topic Clustering K-Means, BERT

1. Introduction

In their journey through life, humans undergo physical growth, development, and knowledge enrichment. As they grow and evolve, humans strive for survival and the continuation of future generations. Beyond mere survival, humans aim to cultivate their potential by expanding their understanding of the world. This progression of human knowledge significantly influences the course of history. Humanity has traversed various stages of the industrial revolution, each marking pivotal advancements crucial for survival. Concurrently, humans are rapidly advancing their knowledge with the evolution of industries. Enabled by technology, individuals can now work more efficiently [1].

Humans are social creatures; therefore, they share their knowledge with others. This knowledge-sharing occurs through various means, such as teaching in the education sector, broadcasting through mass media, and publication in specialized research outlets like journals and conferences. One aim of disseminating knowledge is to educate the next generation about existing knowledge and the latest advancements. Another objective is to share information about recent successful and unsuccessful research. This enables future research to build upon successes and avoid repeating similar mistakes [2], [3].

Approximately 45,000 journals and conferences are registered by the SCOPUS journal and conference ranking association [4]. Additionally, there are various topics and sub-focus areas based on existing fields of study. The purpose of this ranking association is to categorize journal or conference rankings based on their impact on society, particularly in research. One of the methods applied is to examine the number of citations from publications. Simply put, this citation count can measure the quality and impact of a research report on other studies. Over time, the number of journal

***Corresponding author: Yosua Setyawan Soekamto (yosua.soekamto@ciputra.ac.id)**

DOI: https://doi.org/10.47738/jads.v5i2.199

and conference publishers continues to increase. The growth in journals and conferences is regularly recorded and reviewed by ranking associations to maintain publication quality worldwide. Over time, there has also been a reduction in journals or conferences for various reasons. One reason for this reduction is that the publications from these journals or conferences engage in misconduct, categorizing them as predatory journals or conferences [5], [6], [7]. Furthermore, some publishers may cease publishing publications altogether.

Journal publishers have adopted an integrated recording system. The Open Journal System (OJS) is an open-source system developed by the Public Knowledge Project (PKP). PKP, founded in 1998, released OJS in 2002, which was subsequently adopted by various journal publishers across universities. As of 2022, PKP has recorded those 30,000 journal publishers are using OJS [8].

In publishing a journal or conference, publishers need to gather several research reports (papers), conduct reviews of these papers, and edit the formatting. It's common for publishers to be overwhelmed due to a lack of editorial and reviewer teams. The shortage of team members can result in delays in the publication process. This problem could lead to a publisher being perceived poorly by the community, causing authors to hesitate to submit their papers to that publisher. Some publishers engage in misconduct by not correctly conducting the review process, leading them to be classified as predatory journals or conferences.

The reviewer team manually conducts the review and selection process, making it essential to have diverse reviewers with varied expertise, even within the same field of study [9]. It's challenging for a single reviewer to master all topics; hence, the review process generally takes a long time. Many publishers face this issue, and one solution is to invite voluntary authors or hire reviewers with a robust educational reputation. However, the review process still requires a relatively longer time because reviewers need to read the authors' manuscripts and grasp their content. Therefore, a system is needed within the publisher to assist the reviewer team in understanding the authors' manuscripts.

This research proposes that publishers require a system integrating keyword extraction, topic extraction, and topic summarization with the entire reviewer and editorial team. Additionally, a module is needed to visualize the relationship between the reviewed manuscripts and previously published papers. This visualization module lets publishers quickly determine suitable reviewers for each manuscript's topic. In this study, it is proposed to first develop a keyword extraction module. Researchers believe that this keyword extraction can serve as a good starting point because, with this module, publishers can skim through the topics of the accepted manuscripts. Subsequently, these keywords can be further utilized for topic extraction and summarization. This research will apply information extraction methods using natural language processing (NLP) rules for tokenization and neural networks to calculate the similarity of keywords extracted from the manuscripts.

With the rapid advancement of technology in the industrial sector, publishers must have a robust system to facilitate and expedite the publication process, particularly for reviewer teams. This research is expected to inspire researchers, especially those participating in publication. It is hoped that most publisher systems can incorporate features that assist the publication process.

2. Literature Review

2.1. Journal Metrics

Journal Metric Systems typically involve various quantitative and qualitative measures used to evaluate the impact and quality of scholarly journals. These metrics help researchers, institutions, and publishers assess the significance and influence of academic publications within specific fields or disciplines. The Journal Metric Systems are listed below:

1) Impact Factor (IF): this is one of the most well-known metrics. It's calculated based on the average number of citations received by articles published in a journal within a specific timeframe. It's often used to gauge a journal's relative importance within its field. This metric Can help researchers identify high-impact journals for publication but focuses solely on citations within a specific timeframe (usually two years), which may not accurately reflect long-term impact [10].

2) Citation Count: beyond the Impact Factor, examining the total number of citations a journal receives can provide insight into its influence and relevance within the academic community. Citation count reflects the cumulative influence of a journal's articles over time but is also vulnerable to bias from outliers or highly-cited articles [11].
3) H-index: this metric evaluates the productivity and impact of a researcher's, institution's, or journal's scholarly output. It considers the number of publications and the number of citations those publications receive. This metric helps distinguish between highly cited articles and overall research productivity and can be influenced by the researcher's career stage, publication history, or field of study [12], [13].
4) Altmetrics: these metrics measure the impact of scholarly works through non-traditional sources such as social media mentions, downloads, and views. They provide a broader perspective on disseminating and engaging research outputs and capture a broader range of scholarly impact beyond traditional citations. The downside of altmetrics is that they are susceptible to manipulation through artificial boosting of social media mentions or downloads and lack standardization and consistency across different altmetric providers [14].
5) Eigenfactor Score: similar to the Impact Factor, the Eigenfactor Score evaluates a journal's importance based on the number of incoming citations, but it also considers the significance of the journals that cite it. The Eigenfactor Score may be influenced by journal size or publication frequency and limited to journals indexed in specific databases like Web of Science or SCOPUS [15].
6) Journal Rankings: some systems rank journals within specific subject categories based on various metrics, such as the ones mentioned above. For example, SCImago Journal & Country Rank and Journal Citation Reports (JCR). Journal rankings can inform decisions about where to publish research based on the journal's reputation and visibility [10], [16].

2.2. Keyword Extraction

Keyword extraction is identifying the essential words or phrases from text that represent the context of the text content. It is part of information retrieval, natural language processing, and computational linguistics. The frequency-based method was used to find the keyword terms at the beginning of developing keyword extraction techniques. The researchers then used a statistical approach to reduce the bias of the frequency-based method, and then they proposed the Term Frequency-Inverse Document Frequency (TF-IDF) weighting method [17], [18]. After that, researchers imbued graph-based algorithms into keyword extraction techniques; they proposed methods like TextRank and LexRank to represent the text as a graph. Each graph node represents the word, and the edge represents the semantic relation between words (nodes). The essential words (nodes) are weighted by the strength of the relationship and the sum of the importance scores of its neighboring nodes [19], [20], [21]. In the machine learning era, researchers used supervised and unsupervised techniques to classify words or phrases as keywords or non-keywords. Those techniques reduce the bias of frequent word features. In recent years, researchers have explored neural networks (NNs) approaches to keyword extraction. The deep learning techniques such as recurrent neural networks (RNNs) and transformers method have shown promising results for keyword extraction [22], [23].

2.3. Bidirectional Encoder Representation from Transformers (BERT)

Traditional NLP models, such as word embeddings and language models, often treat words in isolation and do not capture the contextual nuances of language well. BERT was developed to overcome this limitation by providing deep bidirectional language representations, allowing the model to understand words in the context of the entire sentence or document. BERT was designed to pre-train deep bidirectional representations of language, which could be fine-tuned on downstream NLP tasks, leading to better performance with less task-specific data. BERT's pre-training on large-scale unlabeled text data, such as Wikipedia and web text, helps alleviate the need for large amounts of task-specific labeled data, making it more applicable to various languages and domains. Traditional RNNs struggle to capture long-range dependencies in text due to vanishing gradient problems and sequential processing. BERT's attention mechanism and Transformer architecture enable it to capture long-range dependencies in text more effectively, leading to a better understanding of context and semantics. BERT was developed to address the limitations of existing NLP models, provide a better contextual understanding of language, enable transfer learning across tasks and domains, and advance state-of-the-art language understanding benchmarks. Its success has profoundly impacted the field of NLP, leading to further research and advancements in language representation learning. [18], [24], [25], [26], [27].

2.4. K-means Clustering

K-means clustering is a popular unsupervised machine learning algorithm for partitioning data into distinct clusters based on similarity. The main objective of K-means clustering is to partition a dataset into K clusters, where each data point belongs to the cluster with the nearest mean (centroid). Each data point is assigned to the cluster with the closest centroid based on a distance metric, typically Euclidean distance. After all data points are assigned to clusters, the centroids are updated by computing the mean of all data points assigned to each cluster. The assignment and update steps are repeated iteratively until convergence, where the centroids no longer change significantly or a maximum number of iterations is reached [28], [29], [30], [31], [32].

The number of clusters K is a hyperparameter that needs to be specified before running the algorithm. Various methods, such as the elbow method or silhouette score, can be used to determine an optimal value for K based on the structure of the data. K-means clustering is computationally efficient and scalable to large datasets, making it suitable for applications with many data points. However, the algorithm's performance may degrade with high-dimensional or sparse data, as Euclidean distance becomes less meaningful in higher dimensions. The algorithm is sensitive to the initial placement of centroids, and different initializations may lead to different clustering results. Outliers and noise in the data can affect the clustering performance, as K-means seeks to minimize the sum of squared distances from data points to centroids. K-means clustering is a versatile and widely used algorithm for partitioning data into clusters, with applications across different domains and industries. It provides an efficient and interpretable way to explore data structure and discover underlying patterns, such as customer segmentation, image compression, document clustering, and anomaly detection [28], [29], [30], [31], [32].

3. Method

3.1. Proposed Method

Our method aims to streamline the publication review process by integrating seamlessly with the OJS commonly used by publishers. By automating key aspects of the review workflow and providing actionable insights, this research seeks to reduce manual effort and expedite the review timeline. The primary system proposed by the researchers can be seen here Figure 1. This system can be integrated with OJS because publishers can self-manage independently and connected by the database. As depicted in the figure 1, this system has several features and workflows specifically designed to assist in the publication review process. The system is inspired by the principles of business intelligence, particularly in terms of decision-making. Drawing inspiration from business intelligence principles, our method is designed to empower publishers with comprehensive data analytics capabilities. By enabling publishers to understand current publication dynamics, predict future trends, and take proactive measures, this research aims to enhance decision-making processes and optimize editorial workflows.

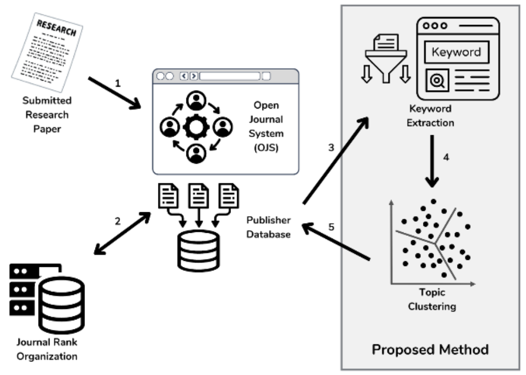


Figure 1. Proposed System Architecture

The current system architecture commonly involves using OJS. In the illustration Figure 1, the system flow is depicted by numbers 1 and 2. The flow starts with the authors submitting their research papers to the open journal system managed by the publisher. Periodically, the data from the publisher is processed by the journal metrics system organization, enabling the author's publications to be indexed and accessed worldwide. This research focuses on keyword extraction and topic clustering using Neural Networks and K-means clustering. Keyword extraction is part of text-mining and information-extraction methods. Previous research has been conducted on keyword extraction using rule-based and text-mining methods, namely NLP [1], [2].

One of the key components of our method is advanced neural network-based keyword extraction. By leveraging state-of-the-art techniques such as the BERT-based KeyBERT library, this research aims to improve the efficiency of keyword extraction, enabling publishers to extract relevant keywords from research papers. The BERT method was chosen in this study because it does not require a labeled dataset and is capable of. Therefore, KeyBERT is used in this study. KeyBERT is a Python library for keyword extraction using BERT embeddings, and it provides pre-trained BERT models. The KeyBERT chosen because it's easy to use and integrated with other NLP tools and libraries in Python, such as Spacy and NLTK. KeyBERT also provides customizable parameters such as the number of keywords to extract and the model to use [3], [4], [5].

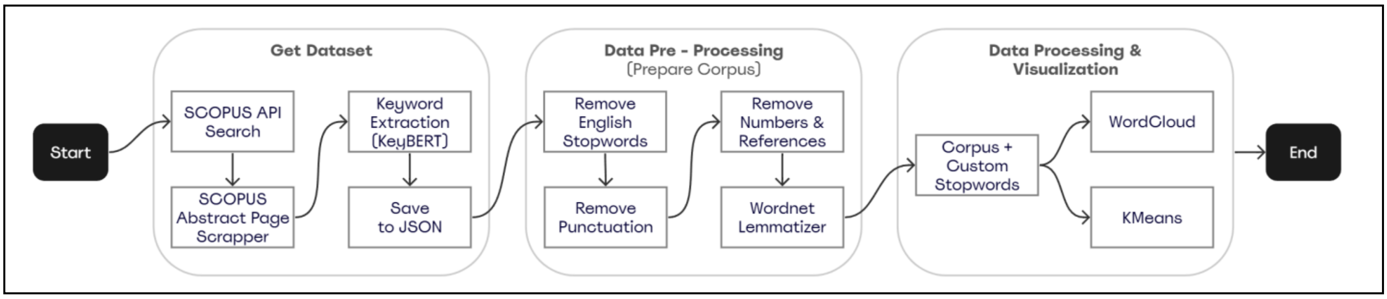


Figure 2. Proposed Method

3.2. Exploratory Data Analysis (EDA)

In this study, abstract datasets obtained from SCOPUS using the SCOPUS API. Reinforcement Learning domain publication chosen in this research, and there are 29,679 articles on Reinforcement Learning, determined based on the search results from the SCOPUS API. In Figure 2, the dataset retrieval is depicted in the “Get Dataset” section. Several processes are conducted in dataset retrieval, such as searching the articles using SCOPUS API and retrieving the article's metadata. The metadata is stored in JSON format, and then the keyword extraction conducted for all 29,679 abstracts (articles). Finally, the raw extracted keyword is stored in JSON format. The dataset consists of articles from 2012 to 2022 with a distribution shown in figure 3. To provide an overview of publications, EDA was also conducted to observe the distribution of authors as shown in figure 4.

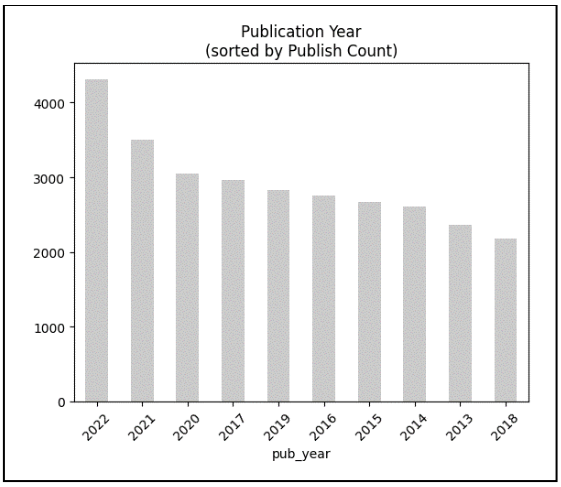


Figure 3. Publication Count by Year

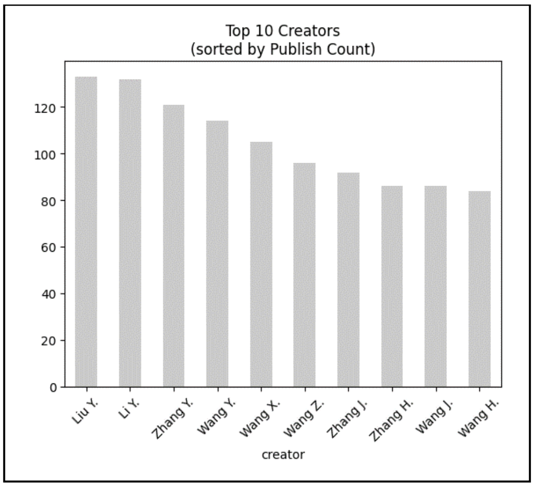


Figure 4. Top 10 Authors by Publication Count

This research focused on the reinforcement learning domain due to its prominence and relevance in contemporary research. Reinforcement learning is a rapidly evolving field within artificial intelligence, with numerous applications across various domains such as robotics, gaming, and optimization. By selecting this domain, the research aimed to analyze the trends and patterns within a specific and well-defined area of research. This approach allows for a more targeted examination of the dataset, providing insights into the characteristics and dynamics of reinforcement learning publications over the past decade. Additionally, focusing on a single domain facilitates more meaningful comparisons and interpretations of the data, enabling us to draw more robust conclusions about the research landscape in this field.

3.3. Data Pre-Processing

After performing keyword extraction, the research process proceeds with data pre-processing, a crucial phase aimed at refining the extracted keywords for subsequent analysis. This phase encompasses a series of meticulous steps to ensure the accuracy, consistency, and relevance of the dataset. The first step in data pre-processing involves thorough keyword cleaning, which entails the removal of extraneous elements that could potentially distort the analysis results. To achieve this, a systematic approach is adopted, beginning with the elimination of stopwords. Since the dataset comprises, abstracts sourced from SCOPUS, predominantly in English, the removal of English stopwords is prioritized to filter out commonly occurring but non-informative words. Subsequently, punctuation marks are stripped from the keywords to standardize their format and facilitate uniformity across the dataset. This process aims to ensure that the keywords are devoid of any extraneous characters or affixes that may obscure their underlying meaning.

Following the cleansing phase, the Wordnet Lemmatizer is employed to further refine the keywords by reducing them to their base or root form. This process, known as lemmatization, helps to consolidate variations of words with similar meanings, thereby promoting consistency and coherence within the text corpus. In terms of implementation, the NLTK Python library is utilized to streamline the data pre-processing tasks, including the removal of English stopwords, tokenization, and stemming of the keywords. By leveraging these techniques, the objective is to optimize the quality and relevance of the dataset, laying a robust foundation for subsequent analysis, particularly in the context of topic clustering.

3.4. Topic Clustering using K-means

In this study, topic clustering employed using the K-means method as the final step in our analysis. Topic clustering plays a crucial role in facilitating the review process for publishers by enabling the grouping of manuscripts based on previously identified topics. Publishers often undergo a lengthy process of searching for and determining the suitability of manuscript topics with reviewers. This streamlines the reviewer selection process and ensures a more tailored approach to manuscript evaluation.

The K-means clustering method is chosen for its effectiveness in partitioning data into distinct clusters. To determine the optimal number of clusters (K), the elbow method analysis utilized, as shown in Figure 5 and figure 6. The elbow method analysis results indicate that the suitable K is 3. However, this value can be re-evaluated based on the sought or processed research domain. For illustrative purposes, the year 2022 dataset chosen. Remember that a raw abstract dataset means the abstract text is used directly. Meanwhile, the number of texts generated by KeyBERT will be greater than the raw abstract's number.

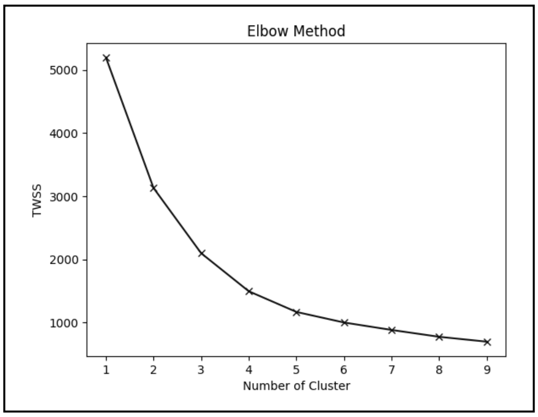


Figure 5. Finding K using the raw abstract dataset

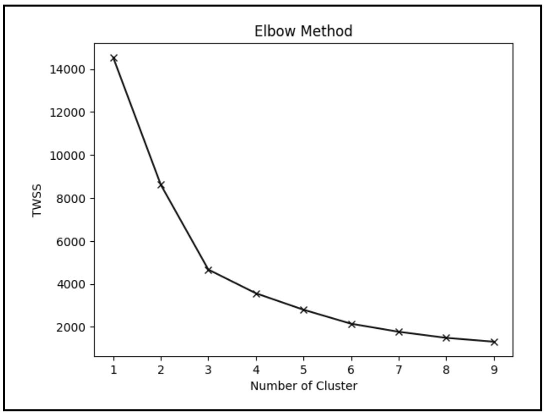


Figure 6. Finding K using KeyBERT

After determining the number of clusters, cluster creation is done using K-means. This research utilizes the Gensim Python library, precisely the Word2Vec method, to create a vector model representing the words and the Sklearn

Python library for K-means clustering. Once the keywords are transformed into vectors, the K-means algorithm iteratively assigns them to the nearest cluster centroid. The centroids, initially randomly initialized, represent the center points of the clusters. After assigning all keywords to clusters, the centroids are updated based on the mean of the vectors belonging to each cluster. This process continues until the centroids stabilize and the clusters converge, indicating that the algorithm has reached an optimal solution. The result is shown in Figure 7. In this clustering visualization, the positive and negative numbers in axes were not differentiated. In this research only a topic clustering conducted, not sentiment or else, so negative or positive numbers mean nothing in this visualization. The only important is the distance between the keywords and the centroid or most relevant topic.

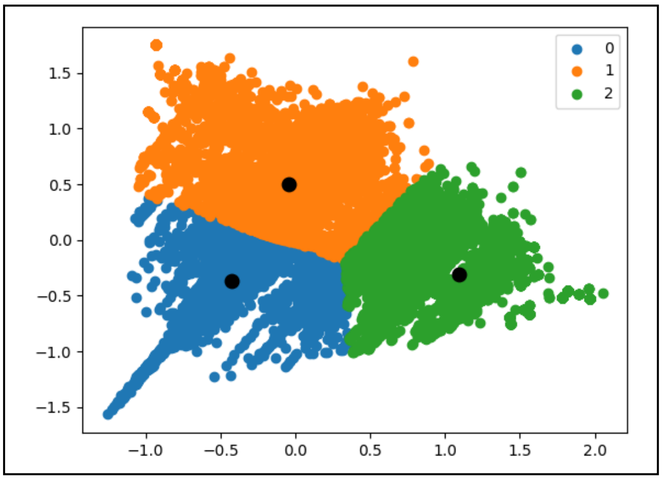
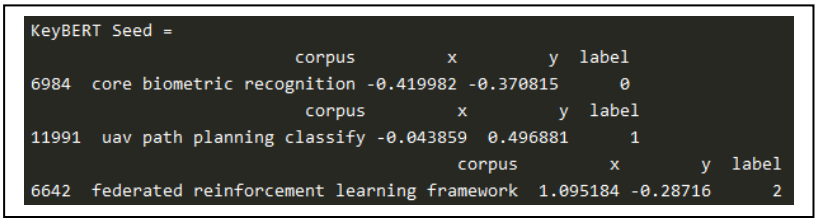


Figure 7. K-means result for year 2022 dataset

To evaluate the effectiveness of the topic clustering, identification of the closest keywords to the centroid for each cluster are needed. These keywords offer valuable insights into the main themes encapsulated within each cluster, as demonstrated in Figure 8. In Figure 8., there are 3 keywords representing the topic of reinforcement learning discussion in the 2022 dataset. It can be understood that recently, the topic of reinforcement learning has been extensively discussed the application of reinforcement learning in classification and recognition, as well as the development of reinforcement learning frameworks.



```
KeyBERT Seed =
                          corpus         x         y  label
6984  core biometric recognition -0.419982 -0.370815      0
                           corpus         x         y  label
11991  uav path planning classify -0.043859  0.496881      1
                                          corpus         x        y  label
6642  federated reinforcement learning framework  1.095184 -0.28716      2
```

Figure 8. The example of the closest keyword to the centroid

3.5. Data Visualization using Word Cloud

Alongside the clustering process, data visualization is also conducted to observe the top keywords that emerge based on the searched domain. This data visualization aims to assist researchers in analyzing research outcomes. The word cloud result is shown in Figure 9. A word cloud is a visual representation of text data where the size of each word corresponds to its frequency or importance within the text. It is a popular tool for quickly and visually summarizing the most significant terms in a large body of text: the corpus from KeyBERT and K-means. This study uses the Wordcloud Python library, with a maximum length of 4 words.

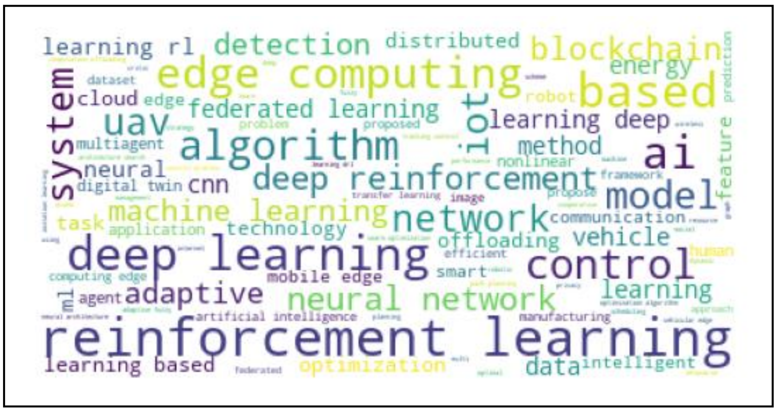


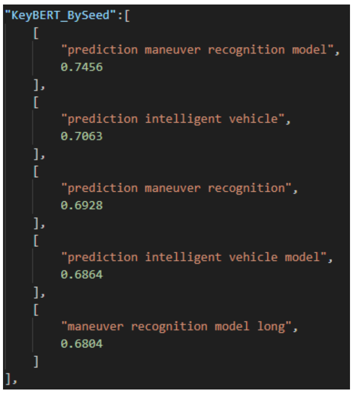
Figure 9. The generated word cloud from K-means clustering

4. Result and Discussion

4.1. Exploratory Data Analysis

The results of this study indicate that the proposed method can be used. The proposed keyword extraction using BERT has proven capable of generating keywords from the included dataset documents. Implementing BERT in this research involves using KeyBERT and adding hyperparameters, such as the document domain, into the KeyBERT function. An example domain used in this study is reinforcement learning. KeyBERT has two important hyperparameters: "seeds" and "candidates." The "candidates"-hyperparameter indicates that KeyBERT will generate keywords from a list of candidates, not from the dataset corpus used. The "seeds" hyperparameter indicates that KeyBERT will use the specified domain, reinforcement learning, and then use that domain as guidance to search for keywords from the dataset corpus used. In this study, the "seeds"-hyperparameter is used.

The pre-trained model used in KeyBERT is the Sentence Transformer "all-miniLM-L6-v2." This pre-trained model is designed for general-purpose use. In this study, a maximum of 5 keywords, each up to 4 words long, is searched for in each document, resulting in a maximum of 20 keywords per document. The top-n keywords produced by KeyBERT are obtained from the highest cosine similarity of each keyword, and then the top-n highest values are selected. From the top-5 keywords generated, an average cosine similarity value of 0.7 is obtained, indicating that KeyBERT can effectively filter keywords, as shown in figure 10. This value signifies a high degree of semantic similarity between the extracted keywords and the original documents. The resulting word cloud analysis further corroborates the relevance of the extracted keywords to the reinforcement learning domain, as depicted in figure 10. This is evidenced by the word cloud results that align with the domain and sub-topic of reinforcement learning. For instance, the processed documents discuss topics such as intelligent vehicle modeling and the application of reinforcement learning methods in predictive modeling.



```
"KeyBERT_BySeed":[
    [
        "prediction maneuver recognition model",
        0.7456
    ],
    [
        "prediction intelligent vehicle",
        0.7063
    ],
    [
        "prediction maneuver recognition",
        0.6928
    ],
    [
        "prediction intelligent vehicle model",
        0.6864
    ],
    [
        "maneuver recognition model long",
        0.6804
    ]
],
```

Figure 10. Keywords and cosine similarity

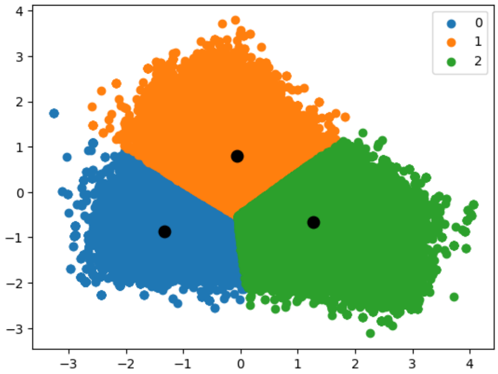
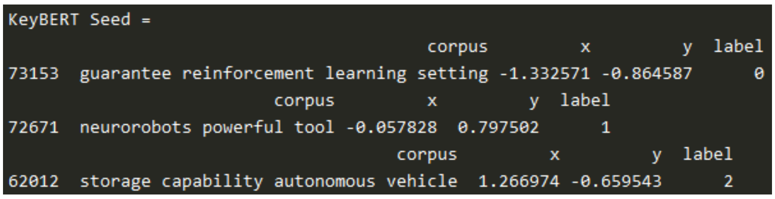


Figure 11. Cluster results for all dataset



```
KeyBERT Seed =
                                           corpus         x         y  label
73153  guarantee reinforcement learning setting -1.332571 -0.864587      0
                           corpus         x         y  label
72671  neurorobots powerful tool -0.057828  0.797502      1
                                       corpus         x         y  label
62012  storage capability autonomous vehicle  1.266974 -0.659543      2
```

Figure 12. The closest keyword to the centroid for all dataset

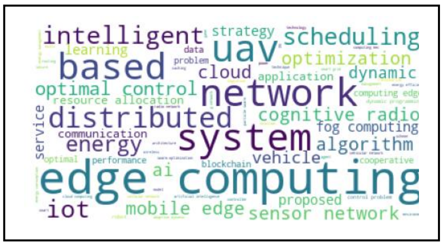


Figure 13. Word cloud for cluster-1

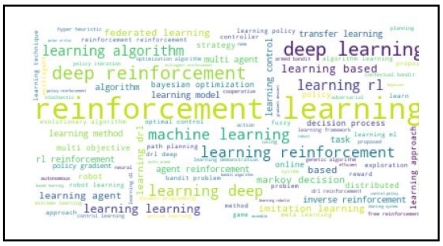


Figure 14. Word cloud for cluster-2

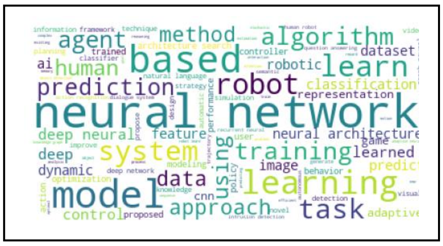
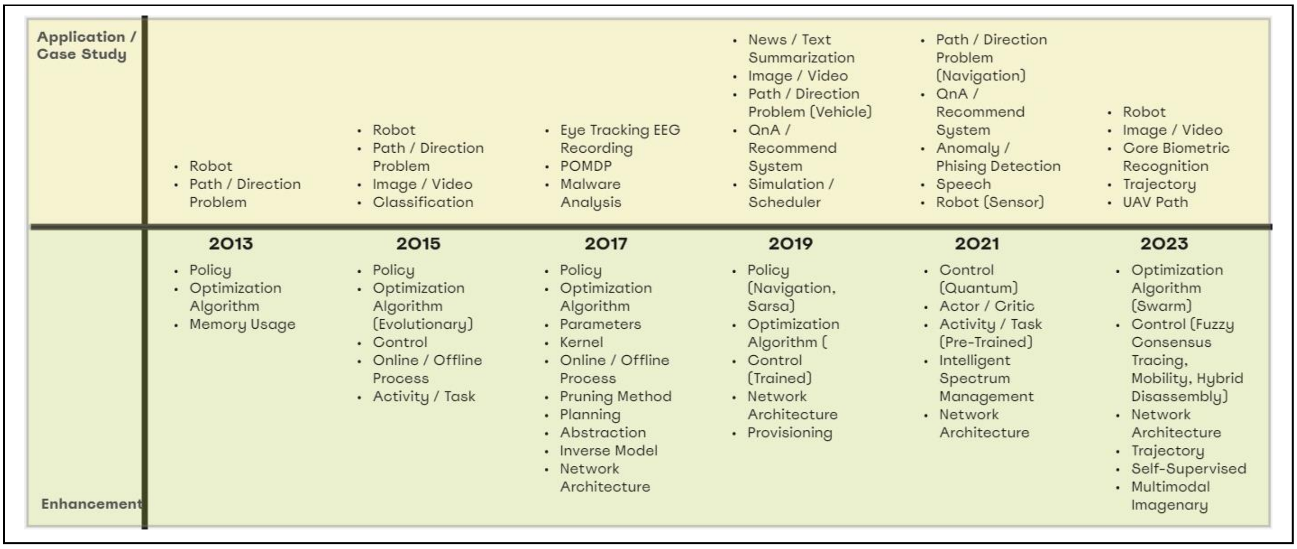


Figure 15. Word cloud for cluster-3



	2013	2015	2017	2019	2021	2023
Application / Case Study	• Robot • Path / Direction Problem	• Robot • Path / Direction Problem • Image / Video Classification	• Eye Tracking EEG Recording • POMDP • Malware Analysis	• News / Text Summarization • Image / Video • Path / Direction Problem (Vehicle) • QnA / Recommend System • Simulation / Scheduler	• Path / Direction Problem (Navigation) • QnA / Recommend System • Anomaly / Phising Detection • Speech • Robot (Sensor)	• Robot • Image / Video • Core Biometric Recognition • Trajectory • UAV Path
Enhancement	• Policy • Optimization Algorithm • Memory Usage	• Policy • Optimization Algorithm (Evolutionary) • Control • Online / Offline Process • Activity / Task	• Policy • Optimization Algorithm • Parameters • Kernel • Online / Offline Process • Pruning Method • Planning • Abstraction • Inverse Model • Network Architecture	• Policy (Navigation, Sarsa) • Optimization Algorithm (• Control (Trained) • Network Architecture • Provisioning	• Control (Quantum) • Actor / Critic • Activity / Task (Pre-Trained) • Intelligent Spectrum Management • Network Architecture	• Optimization Algorithm (Swarm) • Control (Fuzzy Consensus Tracing, Mobility, Hybrid Disassembly) • Network Architecture • Trajectory • Self-Supervised • Multimodal Imagenary

Figure 16. Topic progression visualization

K-means clustering results with the full dataset from 2012 to 2022 indicate that keywords are well-clustered. The visualization of the clusters generated also clearly separates the three clusters, as seen in figure 11. Furthermore, the results of the nearest keyword to the centroid also show good separation, as observed in Figure 12. To give more visualization, each cluster presented by using a word cloud to understand the results of each cluster, as shown in figure 13, figure 14, and figure 15 [6].

In figure 13, the key focus lies on implementing reinforcement learning, with examples including UAV and blockchain, as well as optimizing methods like edge computing and memory load distribution for navigation prediction in moving robots. Similarly, figure 14 emphasizes optimizing reinforcement learning algorithms, encompassing areas such as transfer learning, planning, multi-agent systems, and policy optimization within reinforcement learning techniques. As reinforcement learning involves exploring environments, ongoing discussions persist regarding refining these methods. figure 15, on the other hand, highlights broader reinforcement learning topics such as neural networks, training processes, and dynamic programming methods. These discussions often delve into how agents interact with their environment, leading to an array of research topics surrounding neural networks, training methodologies, and frameworks.

The researchers performed additional visualization to depict the progression of topics in reinforcement learning over the years, based on the clusters obtained from the dataset spanning from 2012 to 2022. This visualization, represented in Figure 16., illustrates the evolution of various aspects within the reinforcement learning domain, such as enhancements and applications. As seen in figure 16, the enhancement section depicts the progression from its inception in 2012, where development initially focused on policy optimization. This evolved into optimization within online training methods and supervised learning, and further advanced to the development of network architectures and self-supervised learning. Similarly, in the application section, starting from 2012, reinforcement learning was utilized for robotics and path prediction, which then expanded to include malware analysis, text summarization, and recommended systems. Subsequently, it progressed to more advanced applications such as anomaly detection, biometrics, and unmanned aerial vehicles (UAVs).

5. Conclusion

In conclusion, this study demonstrates the feasibility of the proposed system, showcasing the effectiveness of the BERT keyword extraction method for topic clustering. By employing this system, publishers can efficiently identify suitable reviewers for manuscript topics, contributing to the streamlining of the publication review process in the era of artificial intelligence (AI). The findings underscore the importance of leveraging AI to automate tasks and computational processes, enabling reviewers to dedicate more attention to ensuring the quality and reliability of publications. However, it is acknowledged that there is room for improvement in the data cleaning processes. Specifically, addressing issues such as the presence of author names and reference numbers in extracted keywords, as well as enhancing text encoding for a more robust corpus, are essential steps forward. Recommendations include implementing normalization and data transformation techniques to enhance text encoding quality, along with integrating user feedback mechanisms into keyword extraction processes for iterative improvements. These enhancements are crucial for refining the accuracy and relevance of extracted keywords.

Building upon the satisfactory outcomes of this study, the next pivotal step involves the development of an intelligent topic summarization system. Given the advancements in AI-driven text summarization, it is proposed that publishing systems incorporate this feature to expedite and enhance the publication process. By integrating text summarization capabilities, publishers can achieve faster and more accurate dissemination of research findings, further optimizing the scholarly communication landscape. Such a system has the potential to enhance the efficiency of the publication process by providing concise summaries of research findings, making them more accessible to readers. However, challenges such as the need for advanced natural language processing algorithms and concerns about summarization accuracy should be carefully considered in the design and implementation of such a system. Additionally, outlining potential avenues for future research, such as exploring alternative keyword extraction methods or investigating the impact of different data cleaning techniques on clustering accuracy, would help to contextualize the findings of the study and highlight areas for further investigation.

6. Declarations

6.1. Author Contributions

Conceptualization: Y.S.S., I.M., C., and E.K.; Methodology: I.M.; Software: Y.S.S.; Validation: Y.S.S., I.M., C., and E.K.; Formal Analysis: Y.S.S., I.M., C., and E.K.; Investigation: Y.S.S.; Resources: I.M.; Data Curation: I.M.; Writing Original Draft Preparation: Y.S.S., I.M., C., and E.K.; Writing Review and Editing: I.M., C., and Y.S.S.; Visualization: Y.S.S.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The articles of this study are derived from research funded by the Department of Information Systems of Universitas Ciputra Surabaya.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. M. Lerner, Individuals as Producers of Their Own Development. New York: Taylor & Francis, 2021. doi: https://doi.org/10.4324/9781003089407.

[2] M. Denis, McQuail's Mass Communication Theory. 6th ed. London: SAGE Publication, 2010.

[3] A. Jarin, "Influence of UNESCO in the Development of Lifelong Learning," Open J Soc Sci, vol. 8, pp. 103–112, 2020, doi: 10.4236/jss.2020.83010.

[4] V. K. Singh, P. Singh, M. Karmakar, J. Leta, and P. Mayr, "The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis," Scientometrics , vol. 126, pp. 5113–5142, 2021, doi: https://doi.org/10.1007/s11192-021-03948-5.

[5] J. Beall, "Predatory publishers are corrupting open access," *Nature,* vol. 489, p. 179, 2012, doi: https://doi.org/10.1038/489179a.

[6] S. A. Elmore and E. H. Weston, "Predatory Journals: What They Are and How to Avoid Them," *Toxicologi Pathology,* vol. 48, no. 4, pp. 607–610, 2020, doi: 10.1177/0192623320920209.

[7] T. F. Frandsen, "Authors publishing repeatedly in predatory journals: An analysis of Scopus articles," *Learned Publishing,* vol. 35, no. 4, pp. 598–604, Oct. 2022, doi: 10.1002/leap.1489.

[8] N. M. Wanjiku, "Publishing with Open Journal Systems (OJS): A Librarian's Perspective," Serials Review, vol. 46, no. 1, pp. 21–25, 2020, doi: https://doi.org/10.1080/00987913.2020.1732717.

[9] S. Pandey, S. P. Mahapatra, D. Pandey, and S. K. Pandey, "Fundamentals of Journal Impact Factor and Indexing Database Metrics," *The Indian Practitioner,* vol. 73, pp. 38-44, 2020.

[10] I. Tahamtan and L. Bornmann, "What Do Citation Counts Measure? An Updated Review of Studies on Citations in Scientific Documents," *Scientometrics* , vol. 121, pp. 1635–1684, 2019, doi: https://doi.org/10.1007/s11192-019-03243-4

[11] B. Cronin and L. Meho, "Using the h-index to rank influential information scientists," *Journal of the American Society for Information Science and Technology,* vol. 57, no. 9, pp. 1275–1278, Jul. 2006, doi: 10.1002/asi.20354.

[12] J. E. Hirsch, "An index to quantify an individual's scientific research output," Proceedings of the National Academy of Sciences, vol. 102, no. 46, pp. 16569–16572, Nov. 2005. doi:10.1073/pnas.0507655102

[13] J. Priem, D. Taraborelli, and P. Groth, "altmetrics: a manifesto", *Cameron Neylon Science and Technology Facilities Council,* vol. 13, no. 7, pp. 1–16, 2011.

[14] C. T. Bergstrom, J. D. West, and M. A. Wiseman, "The EigenfactorTM metrics," *Journal of Neuroscience,* vol. 28, no. 45. pp. 11433–11434, Nov. 05, 2008. doi: 10.1523/JNEUROSCI.0003-08.2008.

[15] S. Ali and S. Bano, "Visualization of Journal ranking using Scimago: An Analytical tool," *Library Philosophy and Practice,* pp. 1–12, 2021.

[16] A. Jalilifard, V. F. Caridá, A. F. Mansano, R. S. Cristo, and F. P. C. da Fonseca, "Semantic Sensitive TF-IDF to Determine Word Relevance in Documents," Jan. 2020, doi: 10.1007/978-981-33-6977-1.

[17] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.05794

[18] A. Kazemi, V. Pérez-Rosas, and R. Mihalcea, "Biased TextRank: Unsupervised Graph-Based Content Extraction," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.01026

[19] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer, "Variations of the Similarity Function of TextRank for Automated Summarization," Feb. 2016, [Online]. Available: http://arxiv.org/abs/1602.03606

[20] E. G¨unes and R. R. Dragomir, “LexRank: Graph-based Lexical Centrality as Salience in Text Summarization,” *Journal of Artificial Intelligence Research*, vol. 22, no. 1, pp. 457–479, 2004, doi: https://doi.org/10.1613/jair.1523.

[21] P. Pęzik, A. Mikołajczyk-Bareła, A. Wawrzyński, B. Nitoń, and M. Ogrodniczuk, “Keyword Extraction from Short Texts with a Text-To-Text Transfer Transformer,” Sep. 2022, [Online]. Available: http://arxiv.org/abs/2209.14008

[22] C. Raffel et al., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.10683

[23] Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[25] A. Priyanshu and S. Vijay, “AdaptKeyBERT: An Attention-Based approach towards Few-Shot & Zero-Shot Domain Adaptation of KeyBERT,” Nov. 2022, [Online]. Available: http://arxiv.org/abs/2211.07499

[26] A. Vaswani et al., “Attention Is All You Need,” Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.03762

[27] A. E. Widjaja, A. Fransisko, C. A. Haryani, and Hery, “Text Mining Application with K-Means Clustering to Identify Sentiments and Popular Topics: a Case Study of the three Largest Online Marketplaces in Indonesia,” *Journal of Applied Data Sciences*, vol. 4, no. 4, pp. 441–453, Dec. 2023, doi: 10.47738/jads.v4i4.134.

[28] M. Jordan, J. Kleinberg, and B. Schölkopf, Pattern Recognition and Machine Learning. Springer, 2006.

[29] T. Hastie, R. Tibshirani, and J. Friedman, Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction, 2nd ed. California: Springer, 2008.

[30] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data Clustering: A Review,” 2000.

[31] G. Steinbach, M. Kumar, and V. Karypis, "A Comparison of Document Clustering Techniques. New York, USA:," ACM Press/Addison-Wesley Publishing Co., 1804.

[32] Z. H. Amur, Y. K. Hooi, G. M. Soomro, H. Bhanbhro, S. Karyem, and N. Sohu, “Unlocking the Potential of Keyword Extraction: The Need for Access to High-Quality Datasets,” 2023, doi: 10.3390/app.

[33] Maarten Grootendorst, “KeyBERT Project,” GitHub - MIT. Accessed: Apr. 14, 2024. [Online]. Available: https://github.com/MaartenGr/KeyBERT?tab=readme-ov-file

[34] J. Ha and D. Kim, “Exploring acceptance of autonomous vehicle policies using KeyBERT and SNA: Targeting engineering students,” 2023. doi: https://doi.org/10.48550/arXiv.2307.09014.

[35] H. Gao, Y. Qin, C. Hu, Y. Liu, and K. Li, “An Interacting Multiple Model for Trajectory Prediction of Intelligent Vehicles in Typical Road Traffic Scenario,” *IEEE Trans Neural Netw Learn Syst*, vol. 34, no. 9, pp. 6468–6479, Sep. 2023, doi: 10.1109/TNNLS.2021.3136866.