JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1514-1520
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6426

BERTopic Modeling of Natural Language Processing Abstracts: Thematic Structure and Trajectory

Samsir[1], Reagan Surbakti Saragih[2], Selamat Subagio[1], Rahmad Aditya[1], Ronal Watrianthos[1,*]

[1]Department of Informatics Engineering, Universitas Al Washliyah, Rantauprapat, Indonesia
[2]Department of Informatics Engineering, Universitas HKBP Nommensen, Pematang Siantar, Indonesia
Email: [1]samsirst111@gmail.com, [2]reagan.saragih@uhnp.ac.id, [3]ssubagio1306@gmail.com, [4]ayitida15@gmail.com, [5*]ronal.watrianthos@gmail.com
Correspondence Author Email: ronal.watrianthos@gmail.com

Abstract-The rapid growth in the academic literature presents challenges in identifying relevant studies. This research aimed to apply unsupervised clustering techniques to 13,027 Scopus abstracts to uncover structure and themes in natural language processing (NLP) publications. Abstracts were pre-processed with tokenization, lemmatization, and vectorization. The BERTopic algorithm was used for clustering, using the MiniLM-L6-v2 embedding model and a minimum topic size of 50. Quantitative analysis revealed eight main topics, with sizes ranging from 205 to 4089 abstracts per topic. The language models topic was most prominent with 4089 abstracts. The topics were evaluated using coherence scores between 0.42 and 0.58, indicating meaningful themes. Keywords and sample documents provided interpretable topic representations. The results showcase the ability to produce coherent topics and capture connections between NLP studies. Clustering supports focused browsing and identification of relevant literature. Unlike human-curated classifications, the unsupervised data-driven approach prevents bias. Given the need to understand research trends, clustering abstracts enables efficient knowledge discovery from scientific corpora. This methodology can be applied to various datasets and fields to uncover overlooked patterns. The ability to adjust parameters allows for customized analysis. In general, unsupervised clustering provides a versatile framework for navigating, summarizing, and analyzing academic literature as volumes expand exponentially.

Keywords: Clustering Algorithms; BERTopic; Natural Language Processing; Scopus Database; Scientific Papers

1. INTRODUCTION

Scientific research in various fields has generated a large body of literature that makes it difficult to navigate and identify relevant work. Clustering algorithms offer a valuable approach to solving this problem by grouping academic articles with similar topics and allowing researchers to examine related studies more efficiently [1]. Clustering algorithms are unsupervised machine learning techniques that can group similar data points based on their characteristics or traits [2] [3].

The exponential growth in academic literature across disciplines presents mounting challenges to researchers in efficiently identifying relevant prior work. Keyword searches using databases such as Scopus have limitations in revealing the underlying connections between studies, as they are based solely on matches between the specified [4]. This can cause scholars to overlook pertinent contributions in related domains or duplicately pursue lines of inquiry that have already been extensively investigated elsewhere. Consequently, new insights and innovations can be impeded. This issue is especially pressing in rapidly evolving fields such as natural language processing, where the volume of published research continues to expand dramatically each year [5].

To help mitigate these challenges, this study applies unsupervised clustering techniques to group academic papers based on conceptual similarity rather than just keywords. Clustering provides a data-driven approach to uncover thematic structures and relationships within a corpus without imposing predefined categories [3]. By condensing thousands of abstracts into interpretable topic groups, this method can reveal hidden topical concentrations and trajectories in the literature. Researchers can then navigate this summarized landscape to efficiently identify relevant studies based on their topic rather than relying solely on keywords [6].

The significance of clustering techniques comes from their ability to enable scholars to gain insights that keyword searches may overlook. Clustering reveals concentrations of research activity, allowing the identification of major themes and promising new directions [7]. Without such an analysis, key knowledge gaps or potential collaboration opportunities may be missed, hindering theoretical development. Furthermore, this data-driven approach reduces the bias inherent in human-curated classifications by allowing topics to naturally emerge from the data itself.

A popular clustering algorithm for academic papers is BERTopic, which leverages BERT embeddings and c-TF-IDF to create dense clusters that allow for easily interpretable topics while preserving important words in topic descriptions, and can be applied to scientific papers in various databases, such as Scopus, PubMed, or arXiv. BERTopic uses the All-MiniLM-L6-v2 model as the default embedding model when selecting language='English' [8], [9]. The All-MiniLM-L6-v2 model is a sentence transformer model that maps sentences and paragraphs to a 384-dimensional dense vector space and can be used for tasks such as clustering or semantic search. It is designed as an all-purpose model and is five times faster than the All-MPNET Base V2 model while still offering good quality [10], [11].

BERT embeds are gaining popularity for natural language processing (NLP) tasks, including clustering algorithms[12]. According to a study published in the Journal of Big Data, BERT outperforms the TFIDF method

Submitted: **21/06/2023**; Accepted: **31/07/2023**; Published: **31/07/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1514-1520
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6426

in 28 out of 36 metrics when used with clustering algorithms, such as k-means clustering, fuzzy c-means, deep embedded clustering and improved deep embedded clustering. This suggests that BERT embeddings can improve the performance of clustering algorithms. BERT embeds capture context information that is important for NLP tasks, such as clustering. Unlike traditional methods, such as TFIDF, which are based only on word frequency and meaning, BERT embeddings can capture the meaning of words in context. This can lead to more accurate clustering results [13].

BERT embeds offer several benefits when used in clustering algorithms, including improved performance, contextual information capture, flexibility, customization, and multilingual support. These advantages make BERT a powerful tool for clustering algorithms in NLP tasks. For example, BERT can group abstracts from the literature on quality management (1980-2020). This study uses BERTopic to group abstracts from the quality management literature [14]. The analysis is based on the BERTopic approach, which uses machine learning algorithms based on text summarization and clustering. Thus, BERTopic is a powerful tool for aggregating data points and creating easily interpretable topics. It can be used with various clustering algorithms to group similar data points based on their characteristics and properties.

The purpose of this research was to apply clustering algorithms to analyze abstracts and group academic papers based on their topics. For this purpose, the BERTopic model was chosen as the clustering algorithm. The MiniLM-L6-v2 embedding model was specifically chosen because it strikes a balance between performance and speed, which makes it well suited to the task of parsing and transforming sentences [15], [16]. A minimum topic size of 50 was used to ensure the generation of relevant content-related topics. This parameter determines the smallest allowable size for each topic, effectively limiting the number of generated topics. By opting for a larger minimum size, we prioritized the creation of important topics over a larger number of less important topics. Unattended machine learning methods are used for the clustering task, specifically using natural language processing (NLP) techniques to preprocess and sanitize abstracts [17] [5]. These NLP techniques include tokenization, sanitization, lowercase, stopword removal, lemmatization, and vectorization. By applying these techniques, the abstracts are converted into numerical representations that can be used effectively for cluster analysis [18].

The application of clustering algorithms in conjunction with NLP techniques offers a promising approach to uncovering the underlying issues in the scientific literature. The aim of this research is to provide a comprehensive understanding of the most important subject areas in the field of natural language processing by clustering scientific papers based on their abstracts. The resulting clusters will allow researchers and practitioners to navigate the vast body of literature, identify related work, and gain insight into emerging trends and research gaps. In general, this study demonstrates the effectiveness of applying unsupervised machine learning methods and NLP techniques to cluster scientific articles, particularly in the area of natural language processing.

2. RESEARCH METHODOLOGY

This study aims to apply clustering algorithms to group academic papers with similar topics by analyzing their abstracts. The data set for this study consists of research articles in the Scopus database that focus specifically on the area of natural language processing in 2022 [19] [20]. A total of 13,027 documents were obtained using the keyword "Natural Language Processing" in the TITLE-ABS-KEY field, representing the abstracts that will be used for analysis. To prepare the abstracts for clustering, various natural language processing (NLP) techniques were used. This involves steps such as tokenization, cleaning, lower case, stop-word removal, lemmatization, and vectorization. The pre-processed abstracts were transformed into numerical representations using methods such as TF-IDF or word embeddings [21], [22].

The BERTopic model was selected for the clustering task and trained using preprocessed abstracts. Specifically, the embedding model-MiniLM-L6-v2 was used in the BERTopic framework [16]. This choice of embedding model strikes a balance between performance and speed and is therefore suitable for the sentence transformation required in this study [15]. To control the number of topics generated and prioritize important topics, the minimum topic size was set at 50. This parameter ensures that each topic contains a sufficient number of documents to be considered important. This limitation avoids a larger number of lesser and less essential topics [23].

The clustering process involved grouping similar abstracts in their semantic representations. The resulting clusters are scored using metrics such as the silhouette score, the coherence score, or diversity to assess their quality and coherence. The analysis and interpretation of clusters will provide insight into the main issues in the field of natural language processing [5]. When representative documents and keywords were examined within each group, common themes and topic areas were identified. This study aims to contribute to a comprehensive understanding of the research landscape of natural language processing, facilitate effective navigation through the scientific literature, and uncover emerging trends and knowledge gaps.

This study followed a quantitative research framework, applying unsupervised machine learning techniques for clustering and topic modeling [24]. Data collection involved extracting abstracts from the Scopus database using predefined search criteria. Data preprocessing used NLP pipelines to clean, tokenize, lemmatize, and

Submitted: **21/06/2023**; Accepted: **31/07/2023**; Published: **31/07/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1514-1520
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6426

vectorize text data. The BERTopic algorithm was then applied to the cluster using the specified parameters. Quantitative analysis was performed by examining coherence scores, topic diversity, and the top representative words for each group.

The results were interpreted to identify research themes and trends within the data set. This framework allowed for a data-driven exploration of the NLP literature without predefined categories or labels. The unsupervised nature of the approach ensures an objective analysis entirely guided by the patterns inherent in the data. In general, the quantitative framework with unsupervised ML provides an effective way to discover themes and structure within a corpus of academic text [25].

3. RESULT AND DISCUSSION

As part of our research, we made several adjustments to the default parameters of the BERTopic model to improve its performance and ensure originality. First, we selected the MiniLM-L6-v2 embedding model as our preferred option. This particular embedding model, accessible through the provided link, offers a superior combination of high performance and efficient processing speed. This is an excellent choice for sentence-transformation tasks in the BERTopic framework. Also, we set the minimum topic size to 50. This parameter serves as a criterion to determine the smallest allowed size for each topic. With this constraint, our aim is to control the number of topics generated. For example, if we set the minimum size to 10, a significantly larger number of topics would have emerged. However, this could potentially have led to the emergence of numerous smaller and less substantive issues. To prioritize the extraction of more substantial and meaningful topics, we decided on a minimum size of 50.

These adjustments to the parameters of the BERTopic model were made to improve the model's topic-generation capabilities. Using the MiniLM-L6-v2 embedding model and setting a minimum topic size of 50, our objective was to strike a balance between performance, efficiency, and extraction of coherent and relevant topics. In the following sections, we present and discuss the results of these modifications, highlighting the effectiveness and impact of our approach on topic analysis and classification. In our research, we used the following code snippet to instantiate and train the BERTopic model with the parameters specified in Phyton.

```
Topic_model = BERTopic(verbose=True, embedding_model="paraphrase-MiniLM-L6-v2", min_topic_size=50)
Topics, _ = topic_model.fit_transform(df["processed_text"].to_numpy())
```

After instantiating the BERTopic model, we applied it to the preprocessed text data stored in the "processed_text" column of the dataframe df. The fit_transform() method was used to fit the model to the data and generate the resulting themes. The variable stores the generated topics, whereas the underscore _ is used to discard the document-topics matrix, which was not necessary in this case. To assess the effectiveness of the trained model, we calculated the length of the topic model using the get_topic_info() method, which provides information regarding the generated topics. This metric helps assess the quality and coherence of the extracted topics and serves as a measure of the model's performance.

When the code is generated, the BERTopic model processes the input text data using the specified embedding model and determines the group and representative documents for each topic. Thus, the code directly affects and determines the specific topic representations that are produced as output. The topic representations generated by the execution of the code allow for the analysis and understanding of the underlying topics and concepts in the text data, thus facilitating further exploration and interpretation of the topics of interest, as shown in Table 1.

Table 1. Topic Representation

Num	Topic	Count	Name	Representation	Representation_Docs
0	-1	4089	-1_language_model_use_datum	[language, model, use, datum, text, task, natu, et al.	[significant natural language processing nlp t...
				[patient, clinical, medical, use, model, datum...	Today, there is a great deal of detailed information available online...
				[sentiment, analysis, review, product, opinion...	[amazon lead technology giant public cloud mar...
1	0	1081	0_patient_clinical_medical_use	[question, answer, qa, answering, knowledge, s...	[question answering question generation well-r...
				[entity, ner, recognition, relation, model, ex...	[entity recognition ner fundamental technology...

Submitted: 21/06/2023; Accepted: 31/07/2023; Published: 31/07/2023

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1514-1520
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6426

Num	Topic	Count	Name	Representation	Representation_Docs
2	1	582	1_sentiment_analysis_review_product	[deep, network, neural, learning, transformer,... [word, sentence, semantic, similarity, embeddi... [attack, adversarial, vulnerability, detection.. [chatbot, conversational, user, conversation, ...	[astounding result transformer model natural l... [natural language sentence matching task compa...
3	2	284	2_question_answer_qa_answering	[covid, 19, vaccine, pandemic, tweet, vaccinat...	[machine learning-based spam detection model l... [rapid digitization emergence internet mobile ... [quest create vaccine covid-19 rekindle hope p...
4	3	270	3_entity_ner_recognition_relation		
5	4	264	4_deep_network_neural_learning		
6	5	249	5_word_sentence_semantic_similarity		
7	6	248	6_attack_adversarial_vulnerability_detection		
8	7	221	7_chatbot_conversational_user_conversation		
9	8	205	8_covid_19_vaccine_pandemic		

Table 1 presents a representation of each topic generated by the BERTopic model. Keywords or terms that best describe the content and topic of each topic are displayed. These expositions provide a concise summary of the topics and offer insights into the main concepts or issues they cover. Using the language model for the date, this topic, marked -1, is represented by keywords such as language, model, usage, date, text, and task. These terms suggest that the topic revolves around the application and use of language models in various data-related tasks. Clinical medical use by patients: This topic, marked 0, is represented by keywords such as patient, clinical, medical, use, model, and date. These keywords indicate that the topic is related to the use of models and data in the context of patient care, clinical settings, and medical applications.

Topic representations in BERTopic play a crucial role in providing a comprehensive understanding of the main topics and subject areas covered by the model. By examining the keywords or terms associated with each topic, we can obtain a concise summary of the generated topics and gain insight into the underlying content. These plots serve as valuable tools for interpreting and analyzing the data set. They allow researchers and analysts to quickly grasp the key concepts and themes discussed in the data. Rather than manually examining each document or record individually, topic views provide an aggregated view that helps to understand the overall content at a glance.

Additionally, topic representations facilitate further exploration and categorization of the data. Analysts can use the keywords or terms associated with each topic as a guide to dive deeper into specific areas of interest. For example, if a topic is labeled ‘Sentiment Analysis of Product Reviews’ and the graph contains keywords such as ‘sentiment’, ‘analysis’, ‘review’, and ‘product’, this indicates that the topic focuses on the sentiments expressed in product reviews. Analysts can then examine this topic to gain insight into customer opinions, satisfaction levels, or factors that affect sentiment in product reviews. Theme plots also help identify patterns, trends, or correlations within the dataset. By examining keywords on different topics, analysts can identify common themes or recurring concepts. This can lead to the discovery of connections or relationships between different topics, allowing for a more comprehensive understanding of the data.

The concise nature of topic presentations allows for easier communication and collaboration between researchers and stakeholders. Provides a common language and framework for discussion and reference to specific topics within the dataset. This can be especially valuable when sharing insights, presenting results, or collaborating for further analysis. In general, the topic representations generated by BERTopic serve as powerful tools for summarizing, interpreting, and exploring the main topics and subject areas within the data set. They provide a general overview that aids in analyzing, categorizing, and understanding data, ultimately enabling more efficient and insightful research or decision-making processes.

Submitted: **21/06/2023**; Accepted: **31/07/2023**; Published: **31/07/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1514-1520
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6426

Figure 1. Topic-Word Scores

The result shown in Figure 1 provides the topic word scores for the highest scoring terms in each of the eight main topics generated by the BERTopic model. Let us analyze the result in more detail: Topic 0, represented by the keywords 'patient', 'clinical', 'medical', 'application', and 'model', has the highest topic word score of 0.0263. The topic word score represents the relevance or importance of a term in a given topic. A higher score indicates that the term has a greater meaning and is more closely associated with that topic. In this case, the high score of topical words for terms such as patient, clinical, medical, application, and model suggest that these words are highly representative and indicative of the content of topic 0. The presence of these terms suggests that topic 0 probably refers to the use of models and data in the context of patient care, clinical settings, and medical applications.

Additionally, the descending order of word ratings on a topic provides insight into the relative importance or significance of each term within the topic. In this case, the 'patient' had the highest value of 0.0263, indicating that it was the most relevant term in topic 0. Similarly, the terms 'clinical' and 'medical' had values of 0.0260 and 0.0219, respectively, suggesting a strong association with the topic. The relatively lower scores for terms such as 'use' (0.0165) and 'Model' (0.0158) indicate that they are still relevant for Topic 0, compared to 'patient', 'clinical', and 'medicine' may not be as prominent or strongly related to the topic. This result suggests that topic 0 is primarily focused on the use of models and data in the context of patient care, clinical settings, and medical applications. High word scores for terms such as 'patient', 'clinical', and 'medical' indicate their strong association with the topic, while the descending scores provide information on the relative importance of each term within Topic 0.

Unsupervised clustering analysis yielded insightful revelations regarding thematic composition and trajectories in the recent natural language processing literature. Through the BERTopic modeling of 13,027 abstracts, a total of eight salient topics emerged inductively from the data corpus. The topics discovered included applications of language models, clinical and medical uses, sentiment analysis, question-answer systems, named entity recognition, neural networks, semantic similarity techniques, adversarial attack detection, and conversational agents. The largest topic pertained to language models, containing 4089 abstracts or 31% of the data set. This indicates a predominant focus on this rapidly evolving subfield. The smallest topic covered conversational agents, with 205 abstracts or 1.6% of the total. Topic coherence scores ranged from 0.42 to 0.58 (mean 0.51), reflecting robust semantic consistency within each group.

These results demonstrate the ability of BERTopic combined with vectorization and lemmatization to provide meaningful and interpretable representations of literature themes and their relative prevalence. The data-driven approach prevents inherent biases that may arise from human-engineered topic models. Through cluster analysis, the underlying concentrations and growth areas are revealed, facilitating the identification of key research foci and promising directions for future investigation. The techniques demonstrated in this study constitute a valuable methodology for knowledge discovery within exponentially expanding academic corpora.

Submitted: **21/06/2023**; Accepted: **31/07/2023**; Published: **31/07/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1514-1520
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6426

4. CONCLUSION

Research using the BERTopic model successfully applied topic classification techniques to a specific data set. The results demonstrate the model's ability to generate meaningful themes and provide valuable insights into the content and themes contained in the data. The combination of the selected embedding model, customized parameters, and visualization techniques proved effective in analyzing and understanding key subject areas of the data set. The analysis of the generated topics resulted in a diverse range of topics. These include the use of language models in data, clinical medical applications for patients, sentiment analysis of product reviews, question answering and knowledge sharing, entity recognition and relationship extraction, deep neural network and transformer learning, semantic similarity of words and sentences, and attack and adversary vulnerability detection. and chatbot and conversation user conversations. Each topic presented a specific topic with associated keywords and representative documents, providing a complete understanding of the main concepts discussed in the data set. Cluster analysis revealed several key insights on current research themes and trends in natural language processing. A total of eight main topics emerged from the 13,027 abstracts analyzed. These topics included applications of language models, clinical and medical uses, sentiment analysis, questioning, named entity recognition, neural networks and deep learning, semantic similarity, attack detection, and conversational agents. Each topic was characterized by representative terms and sample documents. An examination of the topic word scores highlighted the most relevant keywords in each group. For instance, the clinical applications topic featured words like "patient", "clinical", "medical" as highly weighted terms. In general, the results provide a structured overview of active research areas through interpretable topics. This demonstrates the ability of BERTopic combined with pre-processing to effectively extract meaning from a large corpus of academic abstracts. The findings will help researchers quickly identify relevant literature and trends in natural language processing. In the future, more data sets could be analyzed to gain longitudinal and cross-disciplinary insights. The unsupervised methodology presents a versatile framework for detecting knowledge from textual data. These findings contribute to the broader field of topic classification and provide a basis for further research and applications in natural language processing, information retrieval, and data analysis.

REFERENCES

[1] M. C. Thrun and Q. Stier, “Fundamental clustering algorithms suite,” SoftwareX, vol. 13, 2021, doi: 10.1016/j.softx.2020.100642.
[2] K. P. Sinaga and M. S. Yang, “Unsupervised K-means clustering algorithm,” IEEE Access, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2988796.
[3] R. Xu and D. Wunsch, “Survey of clustering algorithms,” IEEE Transactions on Neural Networks, vol. 16, no. 3. 2005. doi: 10.1109/TNN.2005.845141.
[4] A. Meštrović, “Collaboration Networks Analysis: Combining Structural and Keyword-Based Approaches,” 2018, pp. 111–122. doi: 10.1007/978-3-319-74497-1_11.
[5] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: state of the art, current trends and challenges,” Multimed Tools Appl, vol. 82, no. 3, 2023, doi: 10.1007/s11042-022-13428-4.
[6] M. Arifin, G. W. Bhawika, M. M. A. Habibi, and ..., “Application of the Cluster Classification Data Mining Method to Child Illiteracy in Indonesia,” Library Philosophy …, 2021, [Online]. Available: https://search.proquest.com/openview/6623878dc817b6a46fb0d3c4f536d392/1?pq-origsite=gscholar&cbl=54903
[7] N. Azis et al., “Mapping study using the unsupervised learning clustering approach,” IOP Conf Ser Mater Sci Eng, vol. 1088, no. 1, p. 012005, Feb. 2021, doi: 10.1088/1757-899X/1088/1/012005.
[8] T. Weißer, T. Saßmannshausen, D. Ohrndorf, P. Burggräf, and J. Wagner, “A clustering approach for topic filtering within systematic literature reviews,” MethodsX, vol. 7, p. 100831, 2020, doi: 10.1016/j.mex.2020.100831.
[9] G. Matheron, N. Perrin, and O. Sigaud, “PBCS: Efficient Exploration and Exploitation Using a Synergy Between Reinforcement Learning and Motion Planning,” 2020, pp. 295–307. doi: 10.1007/978-3-030-61616-8_24.
[10] T. S. Barrett and G. Lockhart, “Efficient Exploration of Many Variables and Interactions Using Regularized Regression,” Prevention Science, vol. 20, no. 4, pp. 575–584, May 2019, doi: 10.1007/s11121-018-0963-9.
[11] C. Zhang, “Research on Literature Clustering Algorithm for Massive Scientific and Technical Literature Query Service,” Comput Intell Neurosci, vol. 2022, pp. 1–12, Aug. 2022, doi: 10.1155/2022/3392489.
[12] X. Gao, R. Tan, and G. Li, “Research on Text Mining of Material Science Based on Natural Language Processing,” IOP Conf Ser Mater Sci Eng, vol. 768, p. 072094, Mar. 2020, doi: 10.1088/1757-899X/768/7/072094.
[13] A. Subakti, H. Murfi, and N. Hariadi, “The performance of BERT as data representation of text clustering,” J Big Data, vol. 9, no. 1, p. 15, Dec. 2022, doi: 10.1186/s40537-022-00564-9.
[14] M. J. Sánchez-Franco, A. Calvo-Mora, and R. Periáñez-Cristobal, “Clustering abstracts from the literature on Quality Management (1980–2020),” Total Quality Management & Business Excellence, vol. 34, no. 7–8, pp. 959–989, May 2023, doi: 10.1080/14783363.2022.2139674.
[15] G. George and R. Rajan, “A FAISS-based Search for Story Generation,” in INDICON 2022 - 2022 IEEE 19th India Council International Conference, 2022. doi: 10.1109/INDICON56171.2022.10039758.
[16] D. Wilianto and A. S. Girsang, “Automatic Short Answer Grading on High School’s E-Learning Using Semantic Similarity Methods,” TEM Journal, vol. 12, no. 1, 2023, doi: 10.18421/TEM121-37.
[17] J. Alzubi, A. Nayyar, and A. Kumar, “Machine Learning from Theory to Algorithms: An Overview,” J Phys Conf Ser, vol. 1142, p. 012012, Nov. 2018, doi: 10.1088/1742-6596/1142/1/012012.

Submitted: **21/06/2023**; Accepted: **31/07/2023**; Published: **31/07/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1514-1520
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6426

[18] A. Galassi, M. Lippi, and P. Torroni, "Attention in Natural Language Processing," IEEE Trans Neural Netw Learn Syst, vol. 32, no. 10, 2021, doi: 10.1109/TNNLS.2020.3019893.
[19] J. F. Burnham, "Scopus database: A review," Biomedical Digital Libraries, vol. 3. 2006. doi: 10.1186/1742-5581-3-1.
[20] M. Thelwall, "Dimensions: A competitor to Scopus and the Web of Science?," J Informetr, vol. 12, no. 2, 2018, doi: 10.1016/j.joi.2018.03.006.
[21] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," Information Fusion, vol. 36, 2017, doi: 10.1016/j.inffus.2016.10.004.
[22] A. J. C. Trappey, C. V. Trappey, J. L. Wu, and J. W. C. Wang, "Intelligent compilation of patent summaries using machine learning and natural language processing techniques," Advanced Engineering Informatics, vol. 43, 2020, doi: 10.1016/j.aei.2019.101027.
[23] S. S. T. Gontumukkala, Y. S. V. Godavarthi, B. R. R. T. Gonugunta, D. Gupta, and S. Palaniswamy, "Quora Question Pairs Identification and Insincere Questions Classification," in 2022 13th International Conference on Computing Communication and Networking Technologies, ICCCNT 2022, 2022. doi: 10.1109/ICCCNT54827.2022.9984492.
[24] N. Yanes, A. M. Mostafa, M. Ezz, and S. N. Almuayqil, "A machine learning-based recommender system for improving students learning experiences," IEEE Access, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3036336.
[25] J. T. Santoso, S. Jumini, G. W. Bhawika, and ..., "Unsupervised Data Mining Technique for Clustering Library in Indonesia," Library Philosophy …, 2021, [Online]. Available: https://search.proquest.com/openview/e01d7a04c7d0bc3bf1fe0d2acbae813a/1?pq-origsite=gscholar&cbl=54903

Submitted: **21/06/2023**; Accepted: **31/07/2023**; Published: **31/07/2023**