sustainability

Article

Applying Text Mining, Clustering Analysis, and Latent Dirichlet Allocation Techniques for Topic Classification of Environmental Education Journals

I-Cheng Chang [1], Tai-Kuei Yu [2], Yu-Jie Chang [3] and Tai-Yi Yu [4],*

[1] Department of Environmental Engineering, National Ilan University, Yilan 260, Taiwan; icchang@niu.edu.tw
[2] Department of Business Administration, National Quemoy University, Kinmen 892, Taiwan; yutk2012@nqu.edu.tw
[3] Department of Earth and Life Science, University of Taipei, Taipei 100, Taiwan; yjchang@utaipei.edu.tw
[4] Department of Risk Management and Insurance, Ming Chuan University, Taipei 111, Taiwan
* Correspondence: yti@mail.mcu.edu.tw

Abstract: Facing the big data wave, this study applied artificial intelligence to cite knowledge and find a feasible process to play a crucial role in supplying innovative value in environmental education. Intelligence agents of artificial intelligence and natural language processing (NLP) are two key areas leading the trend in artificial intelligence; this research adopted NLP to analyze the research topics of environmental education research journals in the Web of Science (WoS) database during 2011–2020 and interpret the categories and characteristics of abstracts for environmental education papers. The corpus data were selected from abstracts and keywords of research journal papers, which were analyzed with text mining, cluster analysis, latent Dirichlet allocation (LDA), and co-word analysis methods. The decisions regarding the classification of feature words were determined and reviewed by domain experts, and the associated TF-IDF weights were calculated for the following cluster analysis, which involved a combination of hierarchical clustering and K-means analysis. The hierarchical clustering and LDA decided the number of required categories as seven, and the K-means cluster analysis classified the overall documents into seven categories. This study utilized co-word analysis to check the suitability of the K-means classification, analyzed the terms with high TF-IDF wights for distinct K-means groups, and examined the terms for different topics with the LDA technique. A comparison of the results demonstrated that most categories that were recognized with K-means and LDA methods were the same and shared similar words; however, two categories had slight differences. The involvement of field experts assisted with the consistency and correctness of the classified topics and documents.

Keywords: environmental education; cluster analysis; natural language processing; latent Dirichlet allocation

Citation: Chang, I.-C.; Yu, T.-K.; Chang, Y.-J.; Yu, T.-Y. Applying Text Mining, Clustering Analysis, and Latent Dirichlet Allocation Techniques for Topic Classification of Environmental Education Journals. *Sustainability* **2021**, *13*, 10856. https://doi.org/10.3390/su131910856

Academic Editor: Pedro Vega-Marcote

Received: 16 August 2021
Accepted: 24 September 2021
Published: 29 September 2021

1. Introduction

The booming development of Internet technology and the digital economy has pushed today's society into the tide of the big data era. Facing the big data wave, the usage of applying artificial intelligence techniques to extract knowledge and wisdom and analyzing the massive and messy data provide the key technologies for innovative business value and business models. Combining the big data and environmental education issues, most scholars are committed to educational learning issues, such as "e-classrooms," "learning models," "smart learning environments," and "distance cooperative learning" [1,2]. In artificial intelligence, two key areas of intelligent agents and natural language processing (NLP) lead the way [3,4]. Intelligent agents automatically process huge amounts of data through computer programs, assist with sorting and filtering data streams, and automatically organize them into manageable high-value information. The NLP, which is another key technical field of artificial intelligence, is applied as a useful technique to explore the topic classification of environmental education research.

The words or languages that are spoken and written by human beings are termed natural language, which involves a combination of words and grammar. Nowadays, with the rapid development of information and networks, the number of texts is growing exponentially. It has become a crucial issue to classify and manage the vast number of texts and topics of documents; however, we relied on domain specialists in the past. The time and human resources that are required for domain experts to classify, deconstruct, and analyze massive texts or documents are unable to overcome the burden of the current situation and future research work. Therefore, the NLP must manage a large amount of network text and digital text. With the assistance of domain knowledge in specific fields, theory of data statistics, and the NLP program, massive text data could be analyzed to extract the knowledge and rules of the text.

With the development of text mining technology, massive amounts of text data can explore the existing relevant rules in the massive texts, disassemble and combine the texts, and analyze hidden knowledge and rules in the texts [5]. The importance and relevance of keywords in journal papers are commonly used as key clues to discover hidden knowledge in text mining, and word frequency is one key indicator that is used to demonstrate the importance of keywords. A document–term matrix (DTM) [6,7], which represents the relationship between dominant keywords and documents, is also applied as a useful technique for automatic document classification. The application of NLP to automatic document classification in environmental education rarely appears in academic research. Combining at least three professional fields, namely, information, statistics, and environmental education fields, to quickly and objectively provide and verify representative visual analysis results is an important issue facing the era of big data and artificial intelligence. This research sampled the abstracts, title, and keywords of the SSCI (Social Science Citation Index) journals entitled with environmental education in ten years, uses the supervised word segmentation processes of domain experts to establish the TF-IDF (term frequency–inverse document frequency) [8–11] weights and the document–term matrix, and performs a K-means cluster analysis [12–14] and the LDA to analyze the characteristics, trends, and categories of research topics, provide interrelations among feature words with the co-word analysis, and supplies suggestions on the automatic topic classification of environmental education journal papers.

This study focused on environmental education and applied text mining and multivariate technologies to classify topics of journal papers, compare the categorized results with two different techniques, and perform auditing procedures with domain experts to confirm the consistency of topic classification on environmental education papers. The exploration, comparison, and confirmation of LDA and hierarchical K-means clustering on the topic classification provided dominant features for mass autoclassification techniques with text mining technology.

1.1. Text Mining

There is increasing availability of artificial intelligence models, machine learning, and statistical and data mining techniques and procedures that can identify similarities in data/documents and develop decision rules and predictive models that are easy to use in the context of data/documents that are compiled from the environmental education sector. Text mining (TM) [15,16] is a novel artificial intelligence technology in the era of big data that has significant differences from traditional data mining (DM). Traditional DM focuses on processing structured data, and those data can be used as input data and provide a fixed output with the DM algorithm. However, text mining is usually applied to extract unstructured data, such as news, online forums, and social networking media, such as Facebook, Twitter, and Line. The word content is composed of natural language, which means that DM calculations cannot be applied directly. The original model must be modified to a structural form with an indirect method, and then treated with follow-up calculations, classifications, and analyses using DM, mechanical learning, statistics, and computer linguistics. Sumathy and Chidambaram [17] presented the major challenging issues in text mining as follows: (1) the complexity of natural language itself, (2) the intermediate form, (3) multilingual text refining, (4) domain knowledge integration, and (5) personalized autonomous mining.

Text mining can obtain valuable knowledge and wisdom from huge amounts of unstructured text messages. Early TM technology was used for file classification, such as the proper classification of library documents [18], which costs a lot of money and human resources when performing item-by-item coding, reviewing, and classification. However, various types of text messages are increasing rapidly, especially the rapid development of online social media brought about by Web 2.0, such as Facebook, Twitter, or Line, which are accumulating large amounts of electronic text messages [19–21].

Therefore, to dig deep into the key information of huge texts, the automation of TM technology is urgently needed [15,22]. The TM technology has evolved from information retrieval, information extraction, and NLP, to now combining DM, machine learning (ML), and statistics [16,23]. The TM technique could provide personalized document recommendations by analyzing customer information and characteristics, extracting personal preferences, and recommending articles based on their data [24,25]. For the procedure after the word segmentation, the participation of domain experts is very important for the selection and labeling of keywords [26–28] since a high quality of segmentation could not be achieved without utilizing domain-specific knowledge.

The TF-IDF has been the most commonly adopted term weighting technique for text mining and document-processing tasks [29], where this statistical method evaluates the importance of specific words to massive documents [30,31]. Zhang et al. [32] compared three term-weighting schemes, namely, TF-IDF, LSI (latent semantic indexing), and multi-words, for text representation. Chen et al. [8] undertook comparative studies on two term-weighting schemes, namely, TF-IDF and TF-IGM (term frequency and inverse gravity moment), and treated the TF-IDF weighting scheme as a practical benchmark. Kim et al. [9] applied TF-IDF weights and affinity propagation (AP) clustering on patent documents in Korean electric car companies and verified the proposed methodology. Qaiser and Ali [33] applied TF-IDF to examine the relevance of keywords to documents in a corpus and compared the strengths and weaknesses of the TD-IDF scheme.

1.2. Topic Modeling

The purpose of the value-added processing of document classification is to provide users with the ability to search documents by subject or abstract through classified topics, without being restricted by terms of the documents and specialists. The use of automated document classification technology is an important issue regarding quickly and effectively assisting manual classification to cope with the surge in demand for the field of information services and knowledge management. In addition, it is necessary to understand the main idea of the document before assigning it to a specific category within the task of past document classification. Therefore, topic classification is a fairly high-level knowledge processing task that needed the participation and contribution of specialists in the past. Few academic studies with text mining techniques were performed in environmental education; text mining technologies were mostly applied to studies on document classification and topic modeling. Calvo et al. [34] demonstrated the machine learning process and automatic document classification techniques for managing large numbers of news articles or Web page descriptions and classifying them into predefined categories with a naïve Bayes algorithm, lightening the load on domain experts. Hung [35] conducted text mining to investigate the longitudinal trends of e-learning research with 689 journal articles and proceedings, which were separated into 4 groups/15 clusters based on abstract analysis. Zawacki-Richter and Naidu [36] analyzed the titles and abstracts of 515 full papers and demonstrated the trends in distance education research. Zawacki-Richter and Latchem [37] analyzed abstracts and titles of 3674 full papers in *Computers & Education* in 1976–2016 with a text-mining tool and revealed four distinct stages.

The hierarchical K-means clustering technique is one of the most commonly used clustering techniques for document clustering and top classifications [38–40], which is a multivariate method for classifying different target clusters while ensuring that data objects in the same cluster have minimal similarities and other data objects across different clusters have maximal similarities. Lakshmi and Baskar [41] applied a new initial centroid selection for a K-means document clustering algorithm to improve the performance of text document clustering. Christy et al. [42] proposed an unsupervised learning algorithm for text document clustering by adopting a keyword weighting function to cluster a BBC news collection with simple K-means and hierarchical clustering algorithms.

In topic modeling, Do et al. [43] identified research trends for freshwater exotic species in South Korea using text-mining methods in conjunction with bibliometric analysis, where they included 245 articles from research articles and abstracts of conference proceedings and found the major research topic focused on rainbow trout and Nile tilapia, especially the physiological and embryological conditions associated with these species. Bohr and Dunlap [44] employed topic modeling to identify key themes and trends within the environmental sociology field; they identified 25 central topics and examined their prevalence over time, co-occurrence, impact, and prestige. Marín et al. [45] presented a thematic analysis in the field of educational technology in higher education, which involved a separation into three themes: universities, education, and technologies.

The latent Dirichlet allocation (LDA) is an extended application of the hierarchical Bayesian model [46]. Its basic concept involves treating a document as a collection of words, each document as a combination of multiple topics, and each topic as consisting of several words. The LDA has the advantages of supervised learning, flexible extension, greatly improved calculation speed, and recognized effectiveness and values [47,48]. The LDA was successfully applied as a dimensionality reduction technique for many classification problems, such as speech recognition [49], face recognition [50], and topic modeling [51,52]. Moro et al. [53] analyzed 219 journal papers for trends in business intelligence applications for the banking industry in 2002–2013, the LDA modeling results categorized 19 topics without clear categories, where the first terms of the top five categories were credit, predict, neural network, retention, and fraud. Paek and Kim [54] examined the current impact and predicted future impacts of artificial intelligence, and performed topic modeling with the LDA. Zhu and Liu [55] applied the LDA to classify the themes for disaster-related social media data during Typhoon Mangkhut and identified four topics: general response, urban transportation, typhoon status and impact, and animals and humorous news. Hwang et al. [56] utilized the LDA topic modeling and term co-occurrence network analysis to analyze the awareness differences between various social groups regarding the Sustainable Development Goal 13.3 and identified twenty topics for distinct social groups.

1.3. Co-Word Analysis

Co-word analysis is part of the content analysis method, which can construct the internal structure of the subject field and present the relevance among distinct topics using the relevance of the feature keywords. With the relevance strength of keywords in the documents, the characteristics of a specific subject field can be obtained. Ding [57] applied information retrieval as a research topic and discussed the trends and differences of SCI and SSCI journal databases in different research periods. Hui and Fong [58] used co-word analysis and conceptual clustering to construct the relevance of documents and effectively search and retrieve related documents. Van den Besselaar and Heimeriks [59] constructed the trend of information technology development with co-word analysis and categorized the development of information technology into four subject areas to understand the features and differences of distinct topics. An and Wu [60] used three visual analysis methods, namely, a cluster tree, strategy diagram, and social network maps, to demonstrate the classification of medical information with co-word analysis.

Dai and Zhang [61] applied the NoteExpress, Bibexcel, and Ucinet software and the analytical methods of co-word analysis and multi-dimensional scaling to explore the spatial structure of keywords in the environmental crisis management literature. Corrales-Garay et al. [62] performed a descriptive analysis and a co-word analysis to analyze entrepreneurship through open data and found that open data sources, innovation, and business models were critical factors. Soler-Costa et al. [63] analyzed 471 documents contained in the Web of Science with co-word analysis on technological pedagogical content knowledge (TPACK) and identified two main lines: "framework–framework–TPACK" and "technology–pedagogy–beliefs." Corell-Almuzara et al. [64] applied the scientific mapping of the literature and co-word analysis to analyze the influence of COVID-19 on education with the Web of Science database, where the major themes were mental health, organic chemistry, general public, first year undergraduate, and upper division undergraduate in 2020, and autism spectrum disorder, adoption, internet, and intervention in 2021.

2. Materials and Methods

Most text data are in unstructured or semi-structured forms, and since these words could not be classified or indexed manually, this study applied a data pre-processing procedure. Data pre-processing (NLTK package in Python, Figure 1) first extracts, converts, and cleans up the text data. After parsing the content of the document, the document can be retrieved, consolidated, and converted into a "corpus," and then the noise is cleaned up for spaces, punctuation marks, numbers, English letters, and so on. After the standard operating procedure, the NLP algorithm performs word segmentation to convert the corpus into a "structured" data style. In the program, word segmentation is regarded as the crucial key factor affecting text mining works. This study applied the open platform Python software, Gensim module (this module contains several topic-modeling-related techniques, such as LDA, fastText, word2vec, and doc2vec), and other accessible statistical modules and statistical tools (such as IBM SPSS and Matlab) as the TM tools as part of the research scope.

This study built the vocabulary lists, extracted feature keywords with domain experts, constructed the DTM, calculated the TF-IDF (Equation (1)) weights, performed topic classification and co-word analysis, verified the TM results with domain experts, and extracted the important information and knowledge from the unstructured documents. Figure 1 shows the text-mining process of this study. In selecting the feature word list, the process of constructing TF-IDF weight, the auditing themes, and the method for verifying the TM results, this study invited three experts with domain knowledge to determine and verify the consistency and correctness of the feature keywords because of the large number of total keywords and the massive documents.

TF indicates the occurrence frequency of a specific word in all documents, where a specific word with a high TF value shows a high degree of importance. On the other hand, DF represents the number of times a specific word appears in the document. Generally, words with a high DF value are of low importance because they represent specific words appearing in most documents. The IDF, which is the reciprocal of DF, can therefore measure the importance of specific words in a document.

$$TF - IDF_{ij} = TF_{ij} \times IDF_i \tag{1}$$

TF_{ij} : frequency that keyword t_j occurs in document d_i;

IDF_i: inverse value of the document frequency (df_i), where df_i is the document frequency of keyword t_i. Sometimes, the IDF_i factor has many variants, such as log (N/df_i + 1) and log(N/df_i) + 1.

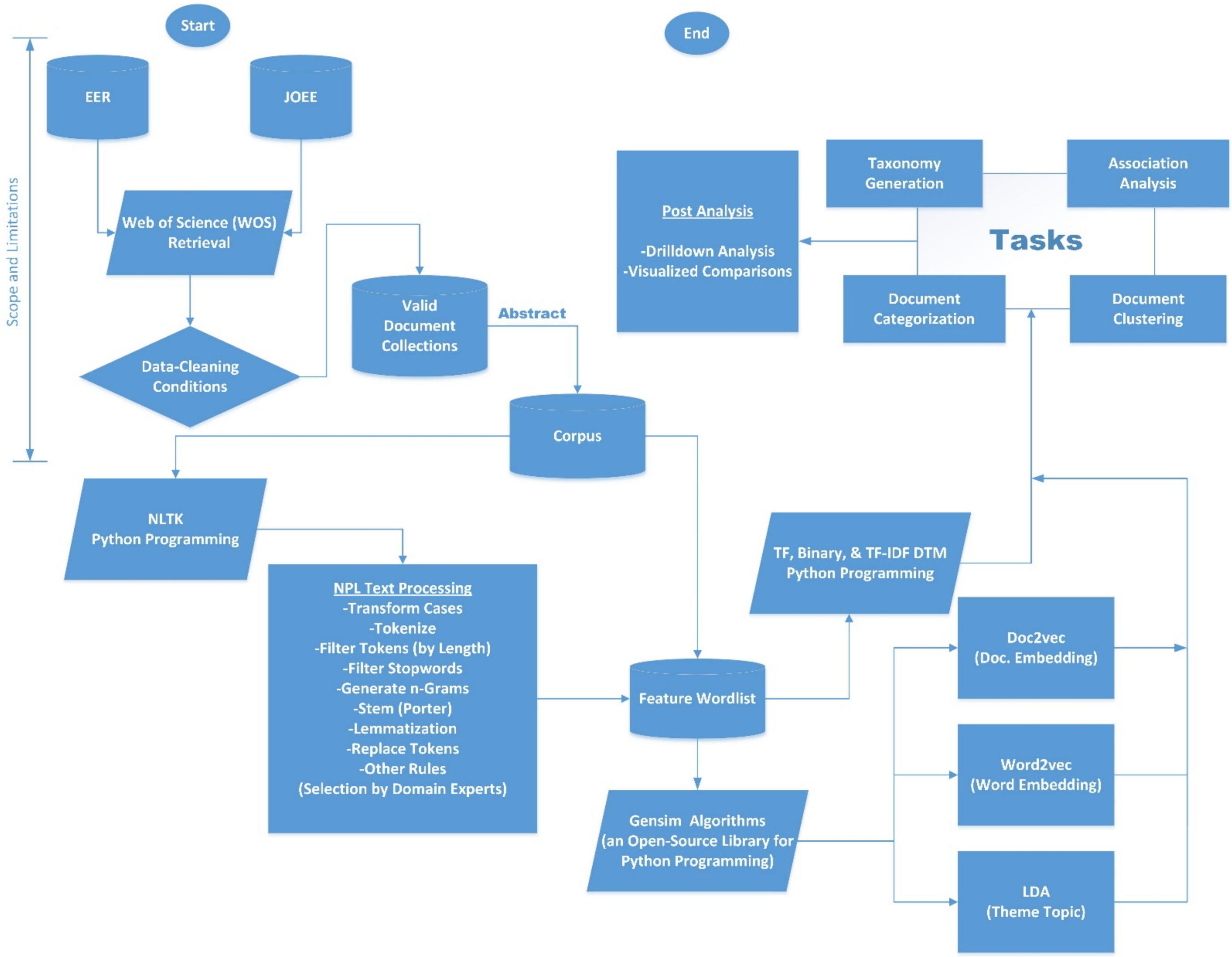

Figure 1. Flow chart of this study.

The research objects of this study were limited to documents published in two SSCI indexed journals that mention environmental education in their title, namely, *Environmental Education Research* (EER) and *Journal of Environmental Education* (JOEE), during 2011–2020. The abstracts and keywords of the published journal articles were used as data sources for the document collections. This study then used Python programming and data cleaning procedures on the Web of Science database to extract the abovementioned documents to form the non-structured documents, and the keywords and abstract of each research paper were defined as the corpus. However, the keywords corpus was not involved in the processing of word segmentation. This study used Python programming to visualize the word clouds for the pre-processing and post-processing results.

The word segmentation stage is a crucial stage that affects the TM results. This study used the self-defined filter, which identified special text symbols and keywords, to clean up the word database (corpus), and then applied appropriate NPL algorithms to perform pre-segmentation. Based on the literature's suggestions and conventions, this study introduced Jieba word-segmentation algorithms. In addition, the keywords in the segmentation words must appear in at least 10% of abstracts, and the structured DTM data was then decided using the style of the abstracts.

Based on the calculation of TF-IDF weights, the DTM analysis was achieved through Python programming, and the processes of taxonomy generation and document classification with hierarchical clustering and LDA methods were also performed. In the post-analysis of the text mining results, this study performed co-word analysis and visualization auditing processing. This procedure (Figure 1) introduced an alternative method for reviewing e-articles with a big data analytical procedure to perform document classification, topic search induction, knowledge extraction of domain information, research context analysis, and find other electronic files of related research. The upper-right corner of Figure 1 illustrates the four related tasks used in this study—document clustering, document categorization, association analysis, and taxonomy generation—and a comprehensive comparison was achieved and demonstrated through the post-analysis process. The determining factor for applying technical tools and principles to perform the four aspects properly is the content of the input data entering the tasks. The aforementioned four aspects did not possess the relationship of interdependence, but displayed a subordinate dependency, as shown in Figure 1.

3. Results and Discussion

3.1. Analysis and Diagram for Text Frequencies

The numbers of published journal papers of EER and JOEE were 669 and 208, respectively, from 2011 to 2020 (Table 1). The maximum yearly numbers of papers for EER were 109 in 2020 and 35 in 2020 for JOEE. The minimum yearly numbers of papers for EER were 40 in 2014 and 12 in 2015 for JOEE. The top 30 words in terms of text frequency and document frequency from the abstracts are listed in Table 2. Based on the text frequency, the top 10 terms regarding text frequency were student (934), environmental education (769), sustainability (608), learn (589), change (536), teacher (499), program (486), school (484), behavior (431), and nature (420); meanwhile, environmental education (333), student (330), learn (274), change (241), relation (231), school (223), experience (222), social (220), sustainability (214), and understand (214) were the top 10 terms regarding document frequency.

Table 1. The numbers of published papers during 2011–2020.

Journals/Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Sum.
EER	43	42	45	40	61	52	80	100	97	109	669
JOEE	13	13	15	15	12	22	27	29	28	35	209
Sum	56	55	60	55	73	74	107	129	125	144	878

EER: *Environmental Education Research*, JOEE: *Journal of Environmental Education*.

Table 2. Text frequency and document frequency for the top 30 words from the abstracts.

	Text Frequency		Document Frequency	
	Words	Counts	Words	Counts
1	Student	934	Environmental education	333
2	Environmental education	769	Student	330
3	Sustainability	608	Learn	274
4	Learn	589	Change	241
5	Change	536	Relation	231
6	Teacher	499	School	223
7	Program	486	Experience	222
8	School	484	Social	220
9	Behavior	431	Sustainability	214
10	Nature	420	Understand	214
11	Climate	399	Practice	211
12	Sustainable development	384	Issue	207
13	Social	369	Social	220
14	Experience	362	Program	200
15	Knowledge	354	Action	191
16	Relation	353	Knowledge	184
17	Action	351	Behavior	182

Table 2. *Cont.*

	Text Frequency		Document Frequency	
	Words	**Counts**	**Words**	**Counts**
18	Practice	339	Theory	181
19	Child	334	Approach	178
20	Understand	320	Teacher	177
21	Climate change	299	Nature	177
22	Issue	298	Environment	171
23	Concept	295	Concept	169
24	Education sustainable development	294	Gignificance	162
25	Environment	284	Group	140
26	Theory	270	Interview	137
27	Approach	269	Human	136
28	Attitude	245	Value	131
29	Community	240	Process	131
30	Human	239	People	128

According to the mission of the *Environmental Education Research* journal, this journal attempts to provide advanced research-based and scholarly understanding of environmental and sustainability education; meanwhile, the *Journal of Environmental Education* aims to provide pedagogical research in environmental and sustainability education, both formally and informally, from early childhood to higher and vocational education. The top-ranking words, which revealed a certain degree of priorities, in terms of high text frequency and document frequency values were environmental education, sustainability, student, teacher, learn, school, and understand; however, the relationships between and characteristics of feature words could not be realized and interpreted without topic classification in such a situation.

3.2. Topic Classification with Hierarchical K-Means Clustering

According to the flow chart (Figure 1), three domain experts cited 510 feature wordlists, where these words included general aspects of environmental education, sustainable education, sustainability, and sustainable development. Based on the cumulative frequency for word count and TF-IDF weights (Figure 2), the top 1% of the words accounted for 6.7% of the cumulative word count and 7.3% of the cumulative TF-IDF weights; the top 10% of the words accounted for 36.9% of the cumulative word count and 37.6% of the cumulative TF-IDF weights; the top 20% of the words accounted for 54.5% of the cumulative word frequency and 54.7% of the cumulative TF-IDF weights; the top 30% of the words accounted for 65.6% of the cumulative word frequency and 66.1% of the cumulative TF-IDF weights. This result demonstrated that the TF-IDF weight was more statistically effective than the word count.

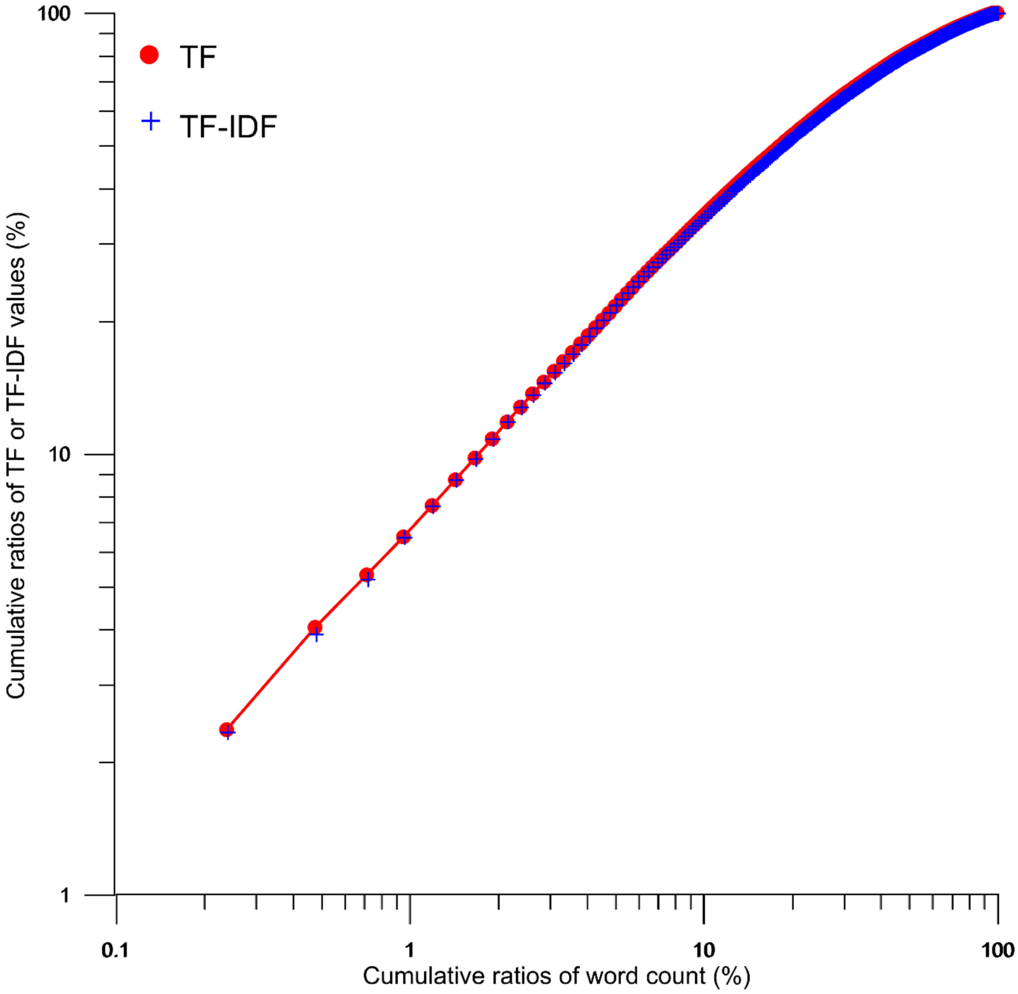

Figure 2. Relationship between cumulative ratios of words, TF, and TF-IDF weights.

For the hierarchical clustering, this study selected the Euclidean distance as the distance measure and Ward's method as the between-group linkage method. The relationship between the decreasing ratios of the total residual and the number of clusters (Figure 3) demonstrated that when the number of clusters was 6, 7, and 8, the decreasing ratios in the total residuals were 1.349, 1.230, and 0.996%, respectively. The average explained ratio of each variable was 0.196% for 510 feature words. With the increase in the number of clusters from 7 to 8, the increase in the explained ratio was less than 1.0%, and the decreasing slope of the explained ratio was low. This study then decided that the number of clusters to use in the K-means approach was seven. After the results of the K-means clustering, each category listed feature words with high TF-IDF weights, numbers of articles, and those ratios. The processes of hierarchical clustering and K-means clustering were achieved with IBM SPSS software at this stage. The optimal number of topics could be determined with statistical indicators (performance, perplexity, and coherence), the elbow method (relationship between the number of clusters and the cost function), subjective judgment, topic interpretability, and topic separation [65–68] in topic modeling, and topic interpretability was the holistic factor.

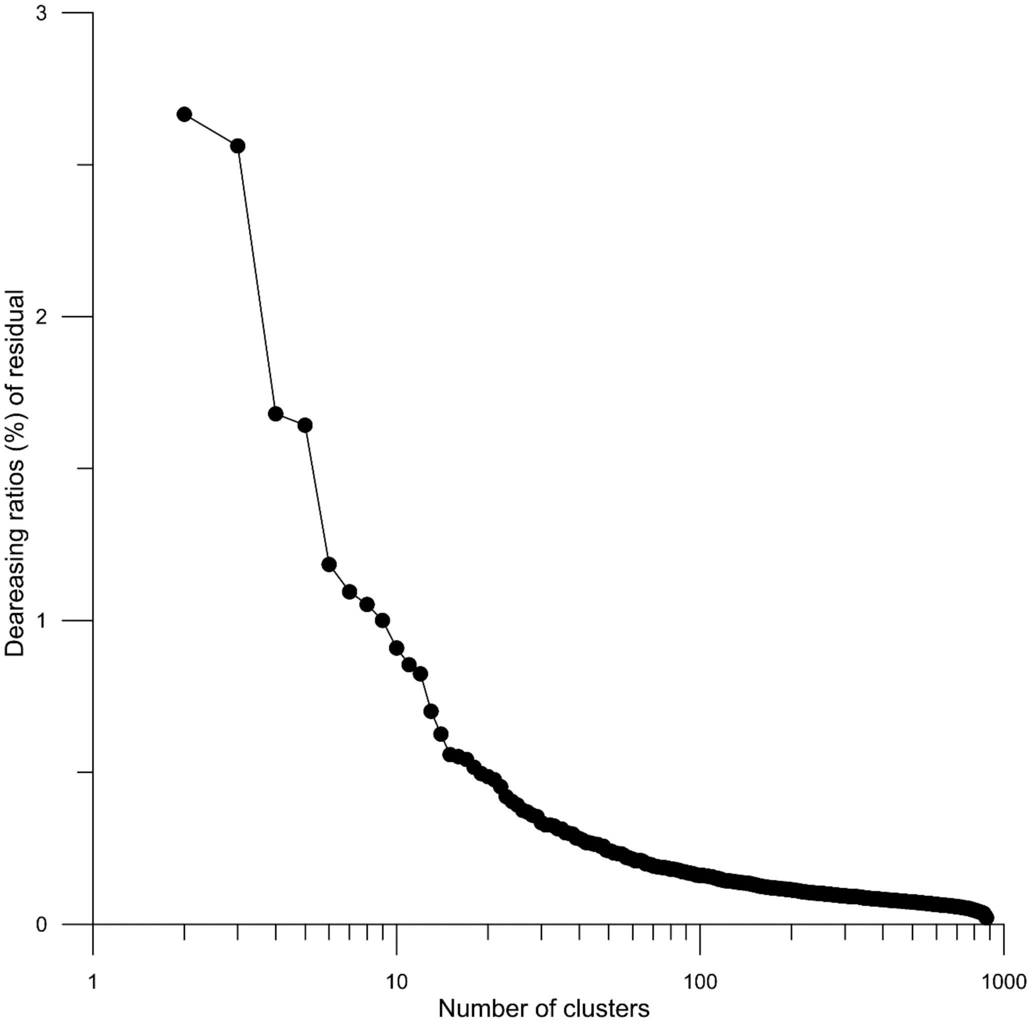

Figure 3. Relationship between the number of clusters and the decreasing ratios of the residuals.

Regarding the objectives of the two SSCI journals targeted by this study, their topics involved at least environmental education, sustainable education, vocational education, or specific environmental education for various learners (such as children, college students, and adults). Table 3 demonstrates the numbers and ratios of classification results with hierarchical K-means clustering. The document ratios of distinct topics were 7.97% for cluster K1 (climate change), 10.93% for cluster K2 (sustainability), 19.82% for cluster K3 (environmental education focused on education for students, teachers, and schools), 31.89% for cluster K4 (environmental education focused on programs, learning, and practice), 11.39% for cluster K5 (environmental ethics), 10.82% for cluster K6 (behavior), and 7.18% for cluster K7 (sustainable development).

The top fifteen feature words with high TF-IDF weights for the first cluster (K1), termed as "climate change," were climate, change, climate change, student, action, science, teacher, social, learn, understand, climate change education, issue, behavior, knowledge, and environmental education. The purpose of these abstracts revealed related issues and actions about climate change and environmental education about climate change to the student.

The top fifteen feature words with high TF-IDF weights for the second cluster (K2), termed as "sustainability," were sustainability, sustainability education, environmental sustainability, learn, environmental sustainability education, practice, student, teacher, understand, approach, change, pedagogy, university, issue, and social. These abstracts revealed learning methods, practice, issues, pedagogy of sustainability, or sustainability education in university.

Table 3. The numbers and ratios of distinct topic categories that were found using the K-means clustering method.

Clusters	Terms with High TF-IDF Weights	Number of Documents (%)	Topic
K1	Climate, change, climate change, student, action, science, teacher, social, learn, understand, climate change education, issue, behavior, knowledge, environmental education, school, young, human, individual, child	70 (7.97%)	Climate change
K2	Sustainability, sustainability education, environmental sustainability, learn, environmental sustainability education, practice, student, teacher, understand, approach, change, pedagogy, university, issue, social, action, curriculum, theory, concept	96 (10.93%)	Sustainability
K3	Student, teacher, school, environmental education, learn, program, knowledge, environment, course, experience, attitude, model, science, eco, relation, understand, change, interview, issue, behavior, action, university	174 (19.82%)	Environmental education
K4	Environmental education, learn, program, practice, social, theory, experience, place, community, relation, educator, concept, action, issue, human, environment, eco, approach	280 (31.89%)	Environmental education
K5	Nature, child, relation, human, experience, environmental education, school, environment, learn, understand, concept, eco, participant, program, value, connectedness, young, student, outdoor, action, attitude	100 (11.39%)	Environmental ethics
K6	Behavior, attitude, knowledge, environmental behavior, program, conservation, pro-environmental, learn, structured interview, intention, environmental education, significance, pro-environmental behavior, model, change, experience, participant, level, child, influence, action, environment	95 (10.82%)	Behavior
K7	Sustainable development, teacher, school, policy, development education, sustainable development education, sustainability, concept, environmental education, approach, practice, student, social, issue, implement, learn, relation, global, discourse	63 (7.18%)	Sustainable development

The top fifteen feature words with high TF-IDF weights for the third cluster (K3), termed as "environmental education focused on education for students, teachers, and schools," were student, teacher, school, environmental education, learn, program, knowledge, environment, course, experience, attitude, model, science, eco, and relation. These abstracts represented classical environmental education for students, teachers, and schools, as well as revealed knowledge, courses, experience, and models of environmental education.

The top fifteen feature words with high TF-IDF weights for the fourth cluster (K4), termed as "environmental education focused on program, learning, and practice," were environmental education, learn, program, practice, social, theory, experience, place, community, relation, educator, concept, action, issue, and human. These abstracts represented programs, learning, practice, place-based education (PBE), and social aspects of environmental education.

The top fifteen feature words with high TF-IDF weights for the fifth cluster (K5), termed as "environmental ethics," were nature, child, relation, human, experience, environmental education, school, environment, learn, understand, concept, eco, participant, program, and value. These abstracts represented several elements in environmental ethics, such as human (child, young, student), nature, experience, outdoor, and the relationship between human and nature.

The top fifteen feature words with high TF-IDF weights for the sixth cluster (K6), termed as "behavior," were behavior, attitude, knowledge, environmental behavior, program, conservation, pro-environmental, learn, structured interview, intention, environmental education, significance, pro-environmental behavior, model, and change. These abstracts represented several elements in pro-environmental behavior, such as attitude, knowledge, conservation, intention, and action, as well as some feature words related to statistical models and questionnaires, such as influence, level, significance, and significance.

The top fifteen feature words with high TF-IDF weights for the seventh cluster (K7), which could be classified and termed as "sustainable development," were sustainable development, teacher, school, policy, development education, sustainable development education, sustainability, concept, environmental education, approach, practice, student, social, issue, and implement. These abstracts represented several aspects in sustainable development, such as policy, disclosure, social impact, and global issue, as well as some feature words related to sustainable development education, such as student, teacher, learn, and environmental education.

3.3. Topic Modeling with the LDA

In topic modeling with the LDA method, the number of topics is a crucial factor. The coherence score of a topic is a metric that is generally used to evaluate topic models by measuring the degree of the semantic similarity scores of the words [69,70], and the coherence score helps to determine the optimal numbers of topics. This study utilized the coherence model from Gensim to calculate the coherence value [71], and a higher value of the coherence score for a topic model represents better coherence [72]. According to the relationship between coherence value and the number of topics (Figure 4), the coherence value increased as the topic numbers increased, displaying the highest value at seven topics (coherence value = 0.3643), which was determined the optimal number of topics. Since the optimal number of topics with the LDA approach was the same as that found with the hierarchical K-means method, it allowed for suitable comparison throughout the overall study.

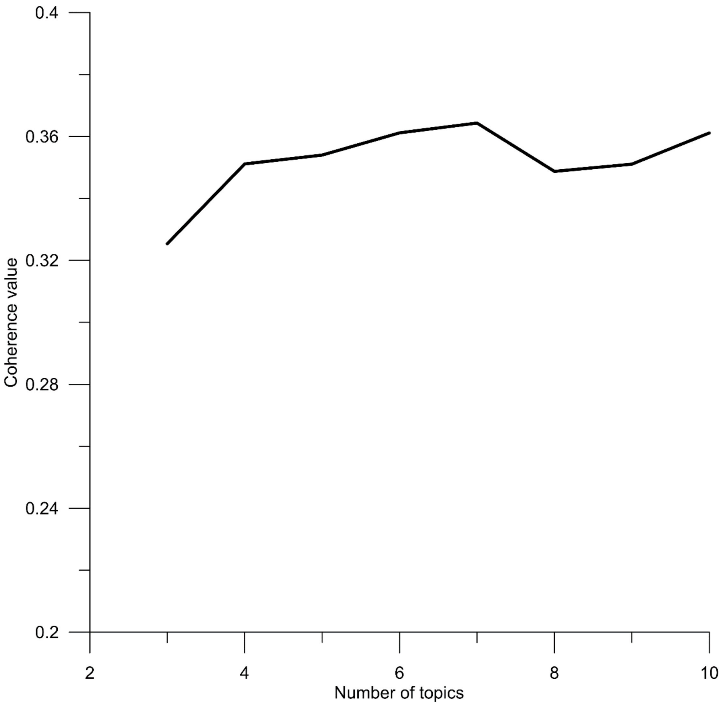

Figure 4. Relationship between the number of topics and the coherence value that was found using the LDA approach.

With the assistance of hierarchical K-means clustering, each document could be classified into a unique topic, and the best LDA topic presented the highest predicted confidence value [10]. Confirming possible overlapping over several topics with the LDA method for a specific document, this study adopted the confidence values for distinct topics for the top 60 articles. Figure 5 illustrates the confidence values that were calculated with the LDA model for different topics. Regarding the first document, the confidence values were 0.161 for L1 (Li represents the ith topic found using the LDA approach), 0.068 for L2, 0.003 for L3, 0.003 for L4, 0.370 for L5, 0.392 for L6, and 0.004 for L7. The maximum confidence value appeared for L6, and the first document was categorized as being on the sixth topic. Regarding the second document, the confidence values were 0.070 for L1, 0.004 for L2, 0.003 for L3, 0.405 for L4, 0.003 for L5, 0.510 for L6, and 0.004 for L7. The second document was then categorized as being on the sixth topic. The maximum confidence values for the third, fourth, and fifth documents were 0.977 for L4, 0.978 for L3, and 0.677 for L5, respectively. The types of the LDA topics for these documents were then determined. However, the difference in confidence values for the highest and the second-highest levels was small for some documents. The confidence values was 0.370 for L5 and 0.392 for L6 for the first document, 0.510 for L3 and 0.425 for L6 for the ninth document, and 0.504 for L5 and 0.460 for L6 for the fifteenth document; these results demonstrated that a few documents could be classified into two similar topics.

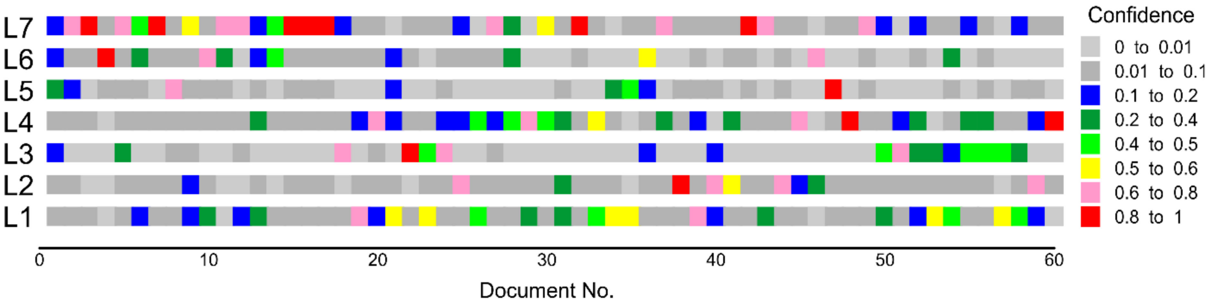

Figure 5. Confidence levels of distinct documents that were found using the LDA method.

Table 4 presents 13–16 representative keywords per topic that were found using the LDA model, where this study accordingly assigned the names of the topic to them. The first topic (L1), which could be termed as "behavior," was the same topic as K6, where the terms with high TF values were student, behavior, attitude, program, school, knowledge, value, model, environment, literacy, positive, change, and measure. The purpose of these terms was to reveal related issues about several elements in pro-environmental behavior about the student, such as behavior, attitude, knowledge, value, model, literacy, and measure.

The second topic (L2) could be termed as "environmental education focusing on education for students, teachers, and schools," was the same topic as K3, where the terms with high TF values were teacher, knowledge, student, system, animal, conservation, science, group, resource, participant, concept, educator, eco, and interview. The purpose of these terms was to present related issues about classical environmental education for students and teachers, as well as revealed knowledge, conservation, resources, and science.

The third topic (L3) could be termed as "sustainability/sustainable development," were the same topics as K2 and K7, where the terms with high TF values were sustainability, sustainable development, teacher, school, curriculum, university, approach, social, student, issue, concept, implementation, practice, and policy. These terms presented related issues for teachers, schools, and students, as well as revealed approaches, curriculums, issues, concepts, and the practice of sustainable education and sustainable development.

The fourth topic (L4) could be termed as "environmental education," which was the same topic as K4, where the terms with high TF values were environmental education, community, practice, policy, theory, social, field, outcome, program, concept, educator, environment, and issue. These abstracts represented community, practice, policy, theory, field, and social aspects of environmental education.

The fifth topic (L5) could be termed as "climate change," which was the same topic as K1, where the terms with high TF values were climate change, action, young, social, climate, identity, program, individual, engagement, community, influence, science, model, and change. These terms revealed related issues, action, social aspect, program, influence, and engagement regarding climate change.

The sixth topic (L6) could be termed as "environmental education," which was the same topic as K4, where the terms with high TF values were learning, place, student, course, project, environmental sustainability education, experience, science, theory, process, approach, pedagogy, local, and field. These abstracts represented local and field programs, learning, course, pedagogy, experience, theory, place-based education (PBE), and environmental sustainability education.

The seventh topic (L7) could be termed as "environmental ethics," which was the same topic as K5, where the terms with high TF values were nature, human, relation, experience, child, eco, environmental education, pedagogy, practice, environment, life, ethic, concept, and culture. These terms included several elements in environmental ethics, such as nature, human, relation, eco, environment, ethics, concept, and culture.

Table 4. The distinct topic categories that were found using the LDA method.

Clusters	Terms with High TF Rankings	Topic
L1	Student, behavior, attitude, program, school, knowledge, value, model, environment, literacy, positive, change, measure	Behavior
L2	Teacher, knowledge, student, system, animal, conservation, science, group, resource, participant, concept, educator, eco, interview	Environmental education
L3	Sustainability, sustainable development, teacher, school, curriculum, university, approach, social, student, issue, concept, implement, practice, policy	Sustainability/sustainable development
L4	Environmental education, community, practice, policy, theory, social, field, outcome, concept, program, educator, environment, issue	Environmental education
L5	Climate change, action, young, social, climate, identity, program, individual, engagement, community, influence, science, model, change	Climate change
L6	Learn, place, student, course, project, environmental sustainability education, experience, science, theory, process, approach, pedagogy, local, field	Environmental education
L7	Nature, human, relation, experience, child, eco, environmental education, pedagogy, practice, environment, life, ethics, concept, culture	Environmental ethics

Comparing the classification results from the two distinct approaches, namely, the hierarchical K-means and the LDA techniques, this study presented some differences as follows: (1) Most topics that were categorized with two distinct techniques were the same, while a few topics were different. The topic entitled "sustainability/sustainable development" was categorized as one topic in the LDA method but as two separate topics in the hierarchical K-means. The topic entitled "environmental education focused on program, learning, and practice" was categorized as one topic in the hierarchical K-means (K4) method but as two separate topics (L4 and L6) in the LDA method. The L4 category emphasized the community, practice, policy, theory, and social aspects of environmental education, while the L6 category focused on local and field programs, courses, and experiences, and place-based education. (2) The hierarchical K-means method presented more detailed categorized results than the LDA technique. The topics entitled "sustainability" and "sustainable development" were categorized as two topics in the hierarchical K-means but were combined as one topic using the LDA method. (3) Based on the confidence value of the LDA approach, most documents could be easily categorized as discussing a unique topic, while a few documents (the probability was less than 5%) could be classified into two or more topics. In contrast, the K-means clustering approach identified a unique topic for each document. More document auditing processes are needed to confirm the correctness and consistency of the topic classification for these two distinct techniques.

3.4. Co-Word Analysis

As the previous section mentioned, this study undertook a co-word analysis of the classification result from the hierarchical K-means clustering method and represented the relationships between feature words with the top fifty TF-IDF weights. Figure 6 demonstrates the results of the co-word analysis for different topics, namely, (1) climate change; (2) sustainability; (3) environmental education focused on education for students, teachers, and schools; and (4) environmental education focused on program, learning, and practice. Figure 6a presents the strong relationships between the terms climate change, climate, and change; media relations and action, climate change, climate, and change; media relations and understand, climate change, climate, and change; and media relations and student, climate change, climate, and change. Figure 6b presents the strong relationships between sustainability and practice; sustainability and learning; sustainability and student; sustainability and understand; sustainability and approach; media relations and learning, student, sustainability, and practice; media relations and sustainability and teacher; sustainability and theory; sustainability and issue; sustainability and action; sustainability and university; sustainability and change; and sustainability and social. Figure 6c presents strong relationships between student, teacher, and school, and between student and learning. The media relations linked student and several terms, such as experience, environment, program, and relation. Figure 6d presents more complicated relations than Figure 6c and represents strong relationships between environmental education and several terms, such as practice, program, relation, learning, experience, social, theory, and concept. The media relations linked social and learning, learning and relationship, practice and theory, environmental education and issue, learning and context, educator and learning, environmental education and educator, environmental education and change, and so on.

Figure 7 visually represents the association strength degrees with a web diagram for the different feature words, and represents the results of the co-word analysis for various feature words, namely, (1) unspecific words, (2) environmental education, (3) sustainability, and (4) student. Figure 7a describes several complicated relationships between sets of feature words, presenting strong relationships between learning, student, and change; experience, student, and environment; and understand and environment. Figure 7a also presents medium relationships between teacher and learning, student and university, learning and action, learning and approach, university and learning, nature and environment, relation and environment, teacher and learning, understand and environment, and so on. Because no specific feature words were set in Figure 7a, the co-word analysis provided many dominant relationships between the feature words, but could not supply a clear relationship between feature words over a specific topic. Without the topic classification, the co-word analysis could not represent obvious relationships between feature words. Figure 7b presents strong relationships between environmental education and feature words such as issue, experience, practice, learning, student, relation, program, nature, and theory; while presenting medium relationships between environmental education and feature words such as environment, field, teacher, school, social, action, and change. Figure 7c presents strong relationships between sustainability and feature words, such as issue, practice, relation, social, theory, understand, action, approach, and change; while presenting medium relationships between sustainability and feature words such as experience, pedagogy, program, teacher, environment, field, teacher, theory, university, and concept. Figure 7d presents strong relationships between student and feature words such as environmental education, learning, program, school, teacher, understand, action, and change; while presenting medium relationships between student features such as issue, nature, relation, social, theory, value, and attitude.

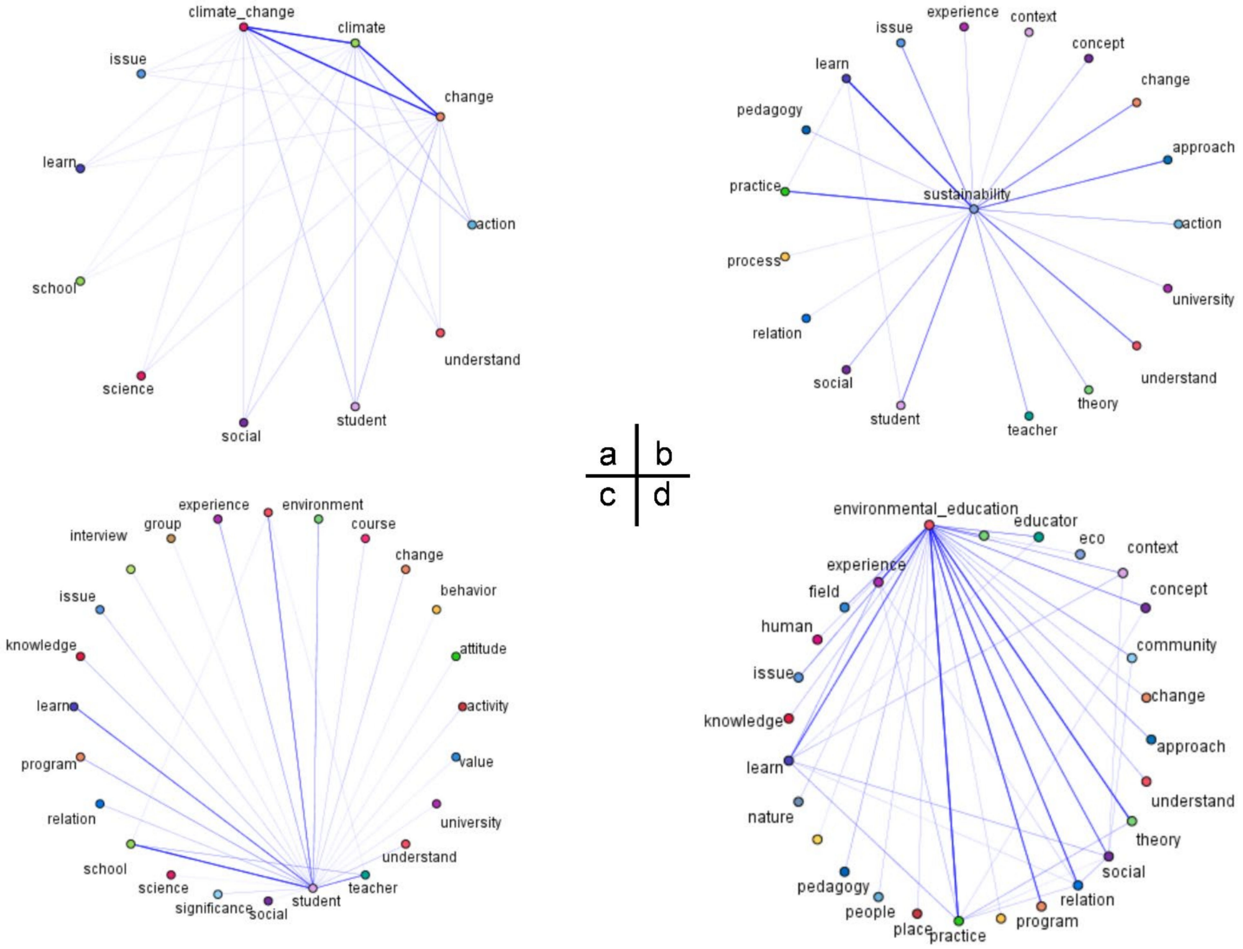

Figure 6. Web charts of co-words for distinct K-means-grouped documents: (**a**) climate change; (**b**) sustainability; (**c**) environmental education focused on education for students, teachers, and schools; (**d**) environmental education focused on program, learning, and practice.

This study constructed a co-word analysis with web charts for terms with the top fifty TF-IDF weights over specific feature nouns and categorized topics to present the relationship between feature nouns. Co-word analysis identifies the degrees of interrelated relationships for feature words and determines dominant feature wordlists for given texts terms in specific topics (subjective) or words; as such, co-word analysis is an important tool to establish feature words to provide a knowledge base for text mining. Co-word analysis uses the co-occurrence feature of vocabulary to divide the specific topic into several subclusters.

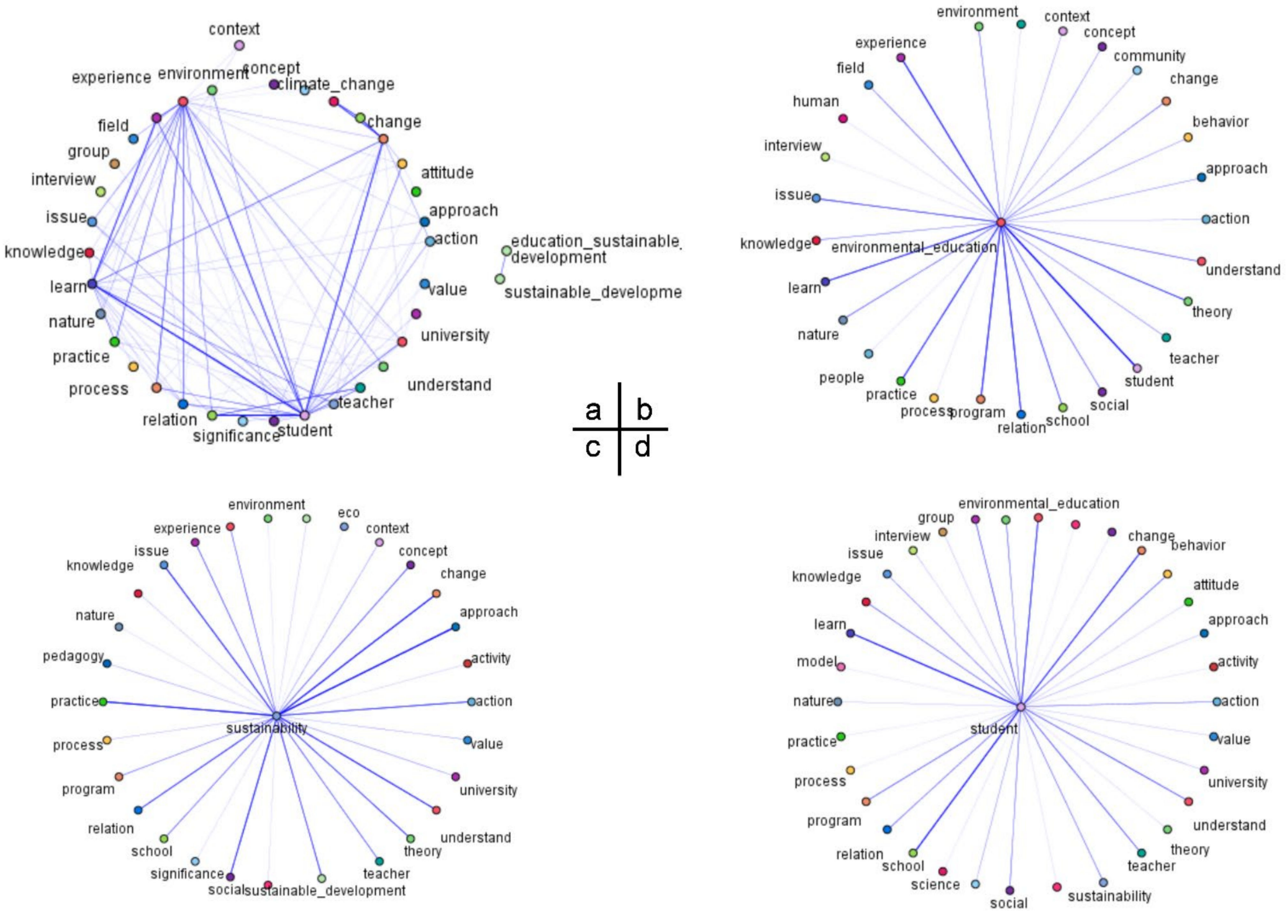

Figure 7. Web charts of co-words for (**a**) all words, (**b**) environmental education, (**c**) sustainability, and (**d**) student.

4. Conclusions

Based on AI methods, this study applied an ordered standard operating procedure (SOP) for text mining techniques on two SSCI journals that consider environmental education in 2011–2020, which mainly included confirming the corpus, drawing word clouds, extracting feature wordlists, performing texts in the corpus (abstracts in selected journal papers), constructing structural DTMs with domain specialists, and undertaking topic classification with the LDA and hierarchical K-means methods. In other words, TM revealed the relationships of feature words with advanced NLP technology and integrated statistical tools to confirm and obtain reasonable AI analysis results. According to the text content of the abstract corpus in this study, 500 feature words were retrieved within the research scope after performing NLP word segmentation. Based on the DTM in this study, the hierarchical K-means analysis summarized the abstract corpus of all documents into seven topic categories: (1) climate change (8.0%); (2) sustainability (10.9%); (3) environmental education focused on education for students, teachers, and schools (19.8%); (4) environmental education focused on program, learning, and practice (31.9%); (5) environmental ethics (11.4%); (6) behavior (10.8%); and (7) sustainable development (7.2%). The classification of topics and the detailed degree of classification for the hierarchical K-means method were better than the LDA approach in this study. Most topics that were categorized with the hierarchical K-means and the LDA methods were the same, but two topics were different.

The above classification results of journal papers could reasonably diagnose the suitability of classification and determine levels of relationships between feature words with the co-word analysis. These results demonstrated that the TM methodology should be effectively applied to topic classification issues in smart article reviewing, taxonomy grouping, automatic document classification, knowledge extraction, and so on. The analytical results of word segmentation, the decision of feature words, and the consideration of various topics with domain experts were key factors in this study. The contribution of this research was the provision of an alternative framework for automatic document classification in the field of environmental education, compare the differences between different topic classification methods, and emphasize the importance of introducing domain experts in environmental education at the current stage. The main purpose of this study was to cause subsequent related discussions about text mining on environmental education. The correctness and consistency of the classification in topic modeling and documents with various AI schemes will be an interesting topic in the near future.

Author Contributions: Conceptualization, T.-K.Y., Y.-J.C. and T.-Y.Y.; investigation, I.-C.C. and T.-Y.Y.; supervision, T.-K.Y., Y.-J.C.; visualization and coding, I.-C.C. and T.-Y.Y.; writing—original draft, I.-C.C. and T.-Y.Y.; writing—review and editing, T.-K.Y.; formal analysis, T.-K.Y., I.-C.C. All authors have read and agreed to the published version of the manuscript.

Funding: The authors express their gratitude to the Ministry of Science and Technology, Taiwan (MOST 107-2511-H-130-004) for funding this project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kivunja, C. Innovative methodologies for 21st century learning, teaching and assessment: A convenience sampling investigation into the use of social media technologies in higher education. *Int. J. Higher. Educ.* **2015**, *4*, 1–26. [CrossRef]
2. Chen, N.S.; Cheng, I.L.; Chew, S.W. Evolution is not enough: Revolutionizing current learning environments to smart learning environments. *Int. J. Artif. Intell. Educ.* **2016**, *26*, 561–581.
3. Hirschberg, J.; Manning, C.D. Advances in natural language processing. *Science* **2015**, *349*, 261–266. [CrossRef]
4. Lucas, C.J. *American Higher Education: A History*; Palgrave Macmillan: New York, NY, USA, 2006.
5. Delen, D.; Crossland, M.D. Seeding the survey and analysis of research literature with text mining. *Expert Syst. Appl.* **2008**, *34*, 1707–1720. [CrossRef]
6. Valls, F.; Redondo, E.; Fonseca, D.; Torres-Kompen, R.; Villagrasa, S.; Martí, N. Urban data and urban design: A data mining approach to architecture education. *Telematematics Inform.* **2018**, *35*, 1039–1052. [CrossRef]
7. Martí-Parreño, J.; Méndez-Ibáñez, E.; Alonso-Arroyo, A. The use of gamification in education: A bibliometric and text mining analysis. *J. Comput. Assist. Learn.* **2016**, *32*, 663–676. [CrossRef]
8. Chen, K.; Zhang, Z.; Long, J.; Zhang, H. Turning from TF-IDF to TF-IGM for term weighting in text classification. *Expert Syst. Appl.* **2016**, *66*, 245–260. [CrossRef]
9. Kim, G.; Lee, J.; Jang, D.; Park, S. Technology clusters exploration for patent portfolio through patent abstract analysis. *Sustainability* **2016**, *8*, 1252. [CrossRef]
10. Kim, D.; Seo, D.; Cho, S.; Kang, P. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. *Inf. Sci.* **2019**, *477*, 15–29. [CrossRef]
11. Miao, R.; Wang, Y.; Li, S. Analyzing urban spatial patterns and functional zones using sina Weibo POI data: A case study of Beijing. *Sustainability* **2021**, *13*, 647. [CrossRef]
12. Abualigah, L.M.; Khader, A.T.; Al-Betar, M.A.; Alomari, O.A. Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. *Expert Syst. Appl.* **2017**, *84*, 24–36. [CrossRef]
13. Naeem, S.; Wumaier, A. Study and implementing K-means clustering algorithm on English text and techniques to find the optimal value of K. *Int. J. Comput. Appl.* **2018**, *182*, 7–14. [CrossRef]
14. Salloum, S.A.; Al-Emran, M.; Monem, A.A.; Shaalan, K. Using text mining techniques for extracting information from research articles. In *Intelligent Natural Language Processing: Trends and Applications*; Springer: Cham, Switzerland, 2018; pp. 373–397.
15. Liu, B.; Cao, S.G.; He, W. Distributed data mining for e-business. *Inf. Technol. Manag.* **2011**, *12*, 67–79. [CrossRef]
16. Chen, Y.L.; Liu, Y.H.; Ho, W.L. A text mining approach to assist the general public in the retrieval of legal documents. *J. Am. Soc. Inf. Sci. Technol.* **2013**, *64*, 280–290. [CrossRef]
17. Sumathy, K.L.; Chidambaram, M. Text mining: Concepts, applications, tools and issues-an overview. *Int. J. Comput. Appl.* **2013**, *80*, 29–32.
18. Miner, G.; Elder, I.V.J.; Fast, A.; Hill, T.; Nisbet, R.; Delen, D. *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*; Academic Press: Waltham, MA, USA, 2012.
19. Gunter, B.; Koteyko, N.; Atanasova, D. Sentiment analysis: A market-relevant and reliable measure of public feeling? *Int. J. Mark. Res.* **2014**, *56*, 231–247. [CrossRef]

20. Salloum, S.A.; Al-Emran, M.; Monem, A.A.; Shaalan, K. A survey of text mining in social media: Facebook and twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J.* **2017**, *2*, 127–133. [CrossRef]
21. Bae, J.H.; Son, J.E.; Song, M. Analysis of twitter for 2012 South Korea presidential election by text mining techniques. *J. Intell. Inf. Syst.* **2013**, *19*, 141–156.
22. He, W.; Zha, S.; Li, L. Social media competitive analysis and text mining: A case study in the pizza industry. *Int. J. Inf. Manag.* **2013**, *33*, 464–472. [CrossRef]
23. Salton, G.; Allan, J.; Buckley, C. Automatic structuring and retrieval of large text files. *Commun. ACM* **1994**, *37*, 97–108. [CrossRef]
24. Lai, C.H.; Liu, D.R. Integrating knowledge flow mining and collaborative filtering to support document recommendation. *J. Syst. Softw.* **2009**, *82*, 2023–2037. [CrossRef]
25. Lavie, T.; Sela, M.; Oppenheim, I.; Inbar, O.; Meyer, J. User attitudes towards news content personalization. *Int. J. Hum.-Comput. Stud.* **2010**, *68*, 483–495. [CrossRef]
26. Tseng, Y.H.; Lin, C.J.; Lin, Y.I. Text mining techniques for patent analysis. *Inf. Process. Manag.* **2007**, *43*, 1216–1247. [CrossRef]
27. Jun, S.; Park, S.; Jang, D. A technology valuation model using quantitative patent analysis: A case study of technology transfer in big data marketing. *Emerg. Mark. Financ. Trade* **2015**, *51*, 963–974. [CrossRef]
28. Goularte, F.B.; Nassar, S.M.; Fileto, R.; Saggion, H. A text summarization method based on fuzzy rules and applicable to automated assessment. *Expert. Syst. Appl.* **2019**, *115*, 264–275. [CrossRef]
29. Kim, S.W.; Gil, J.M. Research paper classification systems based on TF-IDF and LDA schemes. *Hum.-Centric Comput. Inf. Sci.* **2019**, *9*, 30. [CrossRef]
30. Khan, R.; Qian, Y.; Naeem, S. Extractive based text summarization using K-meanss and TF-IDF. *Int. J. Inf. Eng. Elect. Bus.* **2019**, *3*, 33–44.
31. Chen, X.; Zou, D.; Xie, H. Fifty years of British Journal of Educational Technology: A topic modeling based bibliometric perspective. *Br. J. Educ. Tech.* **2020**, *51*, 692–708. [CrossRef]
32. Zhang, W.; Yoshida, T.; Tang, X. A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Syst. Appl.* **2011**, *38*, 2758–2765. [CrossRef]
33. Qaiser, S.; Ali, R. Text mining: Use of TF-IDF to examine the relevance of words to documents. *Int. J. Comput. Appl.* **2018**, *181*, 25–29. [CrossRef]
34. Calvo, R.A.; Lee, J.M.; Li, X. Managing content with automatic document classification. *J. Digit. Inf.* **2004**, *5*. Available online: https://journals.tdl.org/jodi/index.php/jodi/issue/view/22 (accessed on 18 September 2021).
35. Hung, J.L. Trends of e-learning research from 2000 to 2008: Use of text mining and bibliometrics. *Br. J. Educ. Tech.* **2012**, *43*, 5–16. [CrossRef]
36. Zawacki-Richter, O.; Naidu, S. Mapping research trends from 35 years of publications in Distance Education. *Distance Educ.* **2016**, *37*, 245–269. [CrossRef]
37. Zawacki-Richter, O.; Latchem, C. Exploring four decades of research in Computers & Education. *Comput. Educ.* **2018**, *122*, 136–152.
38. Nguyen, H.; Bui, X.N.; Tran, Q.H.; Mai, N.L. A new soft computing model for estimating and controlling blast-produced ground vibration based on hierarchical K-means clustering and cubist algorithms. *Appl. Soft. Comput.* **2019**, *77*, 376–386. [CrossRef]
39. Moussa, M.; Măndoiu, I.I. Single cell RNA-seq data clustering using TF-IDF based methods. *BMC Genom.* **2018**, *19*, 31–45. [CrossRef] [PubMed]
40. Luo, N.C. Massive data mining algorithm for web text based on clustering algorithm. *J. Adv. Comput. Intell. Intell. Inform.* **2019**, *23*, 362–365. [CrossRef]
41. Lakshmi, R.; Baskar, S. DIC-DOC-K-meanss: Dissimilarity-based Initial Centroid selection for DOCument clustering using K-meanss for improving the effectiveness of text document clustering. *J. Inf. Sci.* **2019**, *45*, 818–832. [CrossRef]
42. Christy, A.; Gandhi, G.M.; Vaithyasubramanian, S. Clustering of text documents with keyword weighting function. *Int. J. Intell. Enterp.* **2019**, *6*, 19–31. [CrossRef]
43. Do, Y.; Ko, E.J.; Kim, Y.M.; Kim, H.G.; Joo, G.J.; Kim, J.Y.; Kim, H.W. Using text-mining method to identify research trends of freshwater exotic species in Korea. *Korean J. Ecol. Environ.* **2015**, *48*, 195–202. [CrossRef]
44. Bohr, J.; Dunlap, R.E. Key topics in environmental sociology, 1990–2014: Results from a computational text analysis. *Environ. Sociol.* **2018**, *4*, 181–195. [CrossRef]
45. Marín, V.I.; Duart, J.M.; Galvis, A.H.; Zawacki-Richter, O. Thematic analysis of the international journal of educational Technology in Higher Education (ETHE) between 2004 and 2017. *Int. J. Educ. Technol. High. Educ.* **2018**, *15*, 8. [CrossRef]
46. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
47. Fernandes, S.; Bala, J. Performance analysis of PCA-based and LDA-based algorithms for face recognition. *Int. J. Signal. ProcessSyst.* **2013**, *1*, 1–6. [CrossRef]
48. Séaghdha, D.O.; Korhonen, A. Probabilistic distributional semantics with latent variable models. *Comput. Linguist.* **2014**, *40*, 587–631. [CrossRef]
49. Kolossa, D.; Zeiler, S.; Saeidi, R.; Astudillo, R.F. Noise-adaptive LDA: A new approach for speech recognition under observation uncertainty. *IEEE Signal Process. Lett.* **2013**, *20*, 1018–1021. [CrossRef]
50. Yu, H.; Yang, J. A direct LDA algorithm for high-dimensional data—with application to face recognition. *Pattern Recognit.* **2001**, *34*, 2067–2070. [CrossRef]
51. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y.; Zhao, L. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimed. Tools Appl.* **2019**, *78*, 15169–15211. [CrossRef]

52. Maier, D.; Waldherr, A.; Miltner, P.; Wiedemann, G.; Niekler, A.; Keinert, A.; Pfetsch, B.; Heyer, G.; Reber, U.; Häussler, T.; et al. Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Commun. Methods Meas.* **2018**, *12*, 93–118. [CrossRef]
53. Moro, S.; Cortez, P.; Rita, P. Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Syst. Appl.* **2015**, *42*, 1314–1324. [CrossRef]
54. Paek, S.; Kim, N. Analysis of worldwide research trends on the impact of artificial intelligence in education. *Sustainability* **2021**, *13*, 7941. [CrossRef]
55. Zhu, H.; Liu, K. Temporal, spatial, and socioeconomic dynamics in social media thematic emphases during Typhoon Mangkhut. *Sustainability* **2021**, *13*, 7435. [CrossRef]
56. Hwang, H.; An, S.; Lee, E.; Han, S.; Lee, C.H. Cross-societal analysis of climate change awareness and its relation to SDG 13: A knowledge synthesis from text mining. *Sustainability* **2021**, *13*, 5596. [CrossRef]
57. Ding, Y.; Chowdhury, G.G.; Foo, S. Bibliometric cartography of information retrieval research by using co-word analysis. *Inf. Process. Manag.* **2001**, *37*, 817–842. [CrossRef]
58. Hui, S.C.; Fong, A.C.M. Document retrieval from a citation database using conceptual clustering and co-word analysis. *Online Inf. Rev.* **2004**, *28*, 22–32. [CrossRef]
59. Van den Besselaar, P.; Heimeriks, G. Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics* **2006**, *68*, 377–393. [CrossRef]
60. An, X.Y.; Wu, Q.Q. Co-word analysis of the trends in stem cells field based on subject heading weighting. *Scientometrics* **2011**, *88*, 133–144. [CrossRef]
61. Dai, S.; Zhang, W. Knowledge map of environmental crisis management based on keywords network and co-word analysis, 2005–2018. *J. Clean. Prod.* **2020**, *262*, 121168. [CrossRef]
62. Corrales-Garay, D.; Mora-Valentín, E.M.; Ortiz-de-Urbina-Criado, M. entrepreneurship through open data: An opportunity for sustainable development. *Sustainability* **2020**, *12*, 5148. [CrossRef]
63. Soler-Costa, R.; Moreno-Guerrero, A.J.; López-Belmonte, J.; Marín-Marín, J.A. Co-word analysis and academic performance of the term TPACK in web of science. *Sustainability* **2021**, *13*, 1481. [CrossRef]
64. Corell-Almuzara, A.; López-Belmonte, J.; Marín-Marín, J.A.; Moreno-Guerrero, A.J. COVID-19 in the field of education: State of the art. *Sustainability* **2021**, *13*, 5452. [CrossRef]
65. Kodinariya, T.M.; Makwana, P.R. Review on determining number of cluster in K-means clustering. *Int. J. Adv. Res. Comput. Sci. Manag. Stud.* **2013**, *1*, 90–95.
66. Bholowalia, P.; Kumar, A. EBK-means: A clustering technique based on elbow method and k-means in WSN. *Int. J. Comput. Appl.* **2014**, *105*, 17–24.
67. Guo, L.; Vargo, C.J.; Pan, Z.; Ding, W.; Ishwar, P. Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journal. Mass Commun. Q.* **2016**, *93*(2), 332–359. [CrossRef]
68. Shahbazi, Z.; Byun, Y.C. Analysis of domain-independent unsupervised text segmentation using LDA topic modeling over social media contents. *Int. J. Adv. Sci. Technol.* **2020**, *29*, 5993–6014.
69. Dahal, B.; Kumar, S.A.; Li, Z. Topic modeling and sentiment analysis of global climate change tweets. *Soc. Netw. Anal. Min.* **2019**, *9*, 24. [CrossRef]
70. Xue, J.; Chen, J.; Chen, C.; Zheng, C.; Li, S.; Zhu, T. Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet allocation for topic modeling on Twitter. *PLoS ONE* **2020**, *15*, e0239441.
71. Gensim. Models.coherencemodel—Topic Coherence Pipeline. Available online: https://radimrehurek.com/gensim/models/coherencemodel.html (accessed on 18 September 2021).
72. Mohammed, S.H.; Al-augby, S. Lsa & lda topic modeling classification: Comparison study on e-books. *Indones. J. Electr. Eng. Comput. Sci.* **2020**, *19*, 353–362.