Article

Explore actual sustainable energy topics by using Yake!, Krovetz, GSDMM and short text summaries

Boris Chigarev *

Posted Date: 11 July 2023

doi: 10.20944/preprints202307.0677.v1

Keywords: sustainable energy topics, short text summaries, bibliometric records, Yake!, Krovetz stemmer, GSDMM algorithm

Article

Explore Actual Sustainable Energy Topics by Using Yake!, Krovetz, GSDMM and Short Text Summaries

Boris Chigarev

Abstract: This paper considers the reasonability of using GSDMM as a method for clustering short texts - titles and abstracts of publications 2021-2023 of MDPI journals on sustainable energy topics. The paper proposes an approach to identifying relevant research topics based on the use of the Python script Yake!, the Krovetz streamer, the GSDMM algorithm, and short text annotations. It is emphasized that researchers prefer to rely on specific publications rather than keywords when searching for information on a topic of interest. The bibliometric records of sustainable energy publications from 2021 to 2023 in Sustainability (Energy Sustainability section — 1,926 search results) and Energies (Sustainable Energy section — 994 search results) were used as data for analysis. The GSDMM algorithm was used to cluster the texts, and Yake! to extract the keywords for the GSDMM algorithm's vocabulary. This article summarizes topics found in 9 clusters, including general sustainable energy issues, biomass recycling, region-specific issues, building energy use, economic growth and investment, numerical flow modeling, generation cost optimization, heat to energy conversion, weather-related risks, and false data attacks.

Keywords: sustainable energy topics; short text summaries; bibliometric records; Yake!, Krovetz stemmer; GSDMM algorithm

Introduction

Disclaimer: This paper is largely the personal opinion of the author and has no claim to be a comprehensive study of the subject matter of the title.

The motivation for this study

For researchers, publications are both the final stage of their own research and the main source of information on the topic under study. Even when assessing emerging and promising research topics, researchers tend to rely on specific publications rather than on a set of key terms describing a topic of interest. However, the identification of research topics using keywords is more commonly used in bibliometric analyses, e.g., by well powerful software such as VOSviewer [1] or Bibliometrix [2].

From what I have observed of my colleagues, they are more likely to search for a specific article, such as taken from the reference list of a publication that has piqued their interest, than to pick up or select author keywords from the same publication. Moreover, they often use the titles of the publications they are interested in as the query text, given the way modern search engines like Google work. Of course, in databases such as WoS and Scopus, it is difficult to do without the use of keywords.

In their current work, researchers use Mendeley Reference Manager or Zotero more often than VOSviewer or Bibliometrix. This means that the main object they search for and analyze is the article.

These observations encourage me to favour clustering of publications according to the bibliometric records collected, rather than clustering of keywords when identifying relevant research topics.

The citation rate of publications is the second aspect of the choice of approaches for the analysis of bibliometric records. The following should be noted:

- When looking at articles from the last two years or papers from recent conferences, it is difficult to assess the actuality of a topic presented in them on the basis of the citation rate; it has not yet been established.

 doi:10.20944/preprints202307.0677.v1

- Citations are well represented in subscription-based abstract databases, but not in open publisher databases such as ScienceDirect or MDPI. There, citations are available on the pages of specific publications. Thus, by selecting a small list of publications on a topic of interest, additional information can be obtained from open sources, but not all information can be exported.

The citation rate can be seen as the attention paid by the "readers" of a publication. However, equally important is the attention given by the "authors" of publications, that is, identifying topics that attract the attention of authors from leading journals or staff of strong universities, or topics of publications by authors recognized as experts in the field of research. Such publications may not yet be highly cited, but are of great interest in identifying promising research topics. In addition, a community of authors may not yet be established in an emerging field, resulting in a low citation rate for an important publication.

Brief rationale for the novelty of the proposed research

As this paper analyses bibliometric data of publications from two MDPI journals for the years 2021-2023, queries are made on all publications from the same publisher for the same years to broaden the context.

Search date: 2021-2023 (up to 27-06-2023)
The query data are in Keywords (search results) format
topic modeling (3,613), Topic modeling is presented well
short text summarizer (12), not much, yet present
VOSviewer (607)
Bibliometrix (149)
KeyBERT (27)
BERTopic (7)
Krovetz (0)
Yake! (8), but only in authors' names, no actual publications
GSDMM (0)

The search could certainly be extended to publications by other publishers, but the fact that the combined use of GSDMM+ Krovetz+ Yake! methods has not been sufficiently explored seems obvious.

The objective of this study

The purpose of this study is to demonstrate the reasonableness of using GSDMM as a method for clustering short texts — titles and abstracts of publications 2021–2023 of MDPI journals on the topic of sustainable energy.

Materials and methods

Bibliometric records of publications for the years 2021-2023 in the journals Sustainability (Section Energy Sustainability - 1,926 Search Results) and Energies (Section Sustainable Energy - 994 Search Results), up to 19-05-2023, were used as the data for analysis. Information (Article types = research-article, review-article) were exported as text files which were transformed to TSV format and merged. A total of 2920 entries were used. The records were homogeneous both by topic (Sustainable Energy) and by publication type (research-article, review-article), what is important when analyzing the texts.

The main work was done using title and abstract texts merged into a single paragraph.

Brief justification for choosing GSDMM for text clustering in this study

This paper focuses on methods that do not use pre-trained language models. Methods based on pre-trained language models deserve a separate study, fortunately both KeyBERT [3] and BERTopic [4] work well and are easy to handle.

K-means [5,6] and LDA [7,8] are most commonly used in record clustering and topic modelling.

Comparing GSDMM and LDA, the GSDMM is better suited to short texts as it assumes that there is one topic in the text [9,10].

The advantages of GSDMM over K-means are analyzed in [11] and in the article by the authors of GSDMM algorithm [12].

I used an implementation of this method written in Rust [https://github.com/rwalk/gsdmm-rust]

Dictionary selection rationale for the algorithm GSDMM

The choice was made taking into account the non-use of pre-trained language models.

Author keywords available in records exported from MDPI could be used as a dictionary, but then we would go beyond using the title and abstract fields. It is good that in both ScienceDirect and MDPI author keywords are available in exported records, but then this approach may not be suitable for analysing conference proceedings in which author keywords are often missing. It is the conference proceedings which can provide good material for identifying new research topics.

Personal experience shows that noun phrases are also good candidates for forming a list of key terms. But the application of modern morphological analysis is also based on pre-formed linguistic models. The application of n-gram-based language models works well when the text is well-preprocessed (proper stop-word list generated, lemmatization performed, ranking criterion selected).

For routine work, it is a good idea to have a program that runs reliably, takes into account the above requirements, and is stably maintained. In my opinion (based on personal experience), such requirements are satisfied by Yake! [13]. To learn more about how this program is designed, to see tutorials and even to use it as an online service, please visit http://yake.inesctec.pt/.

I used this program with the parameters: -n 2 -t 2000 -v.

Why I use bigrams — most of the keywords consist of two terms. 2000 meaningful words and phrases is enough to describe a narrow subject area. Increasing the parameter t leads to a significant increase in execution time, which is not desirable in the spirit of trial-and-error experiments.

Why I don't use individual words - some individual words, most often adjectives, may get low ratings, but be meaningful in word combinations, e.g., big data, clean energy. However, in the total list of 2,000 terms, 1,554 are bigrams.

Before using Yake!, the title and annotation texts underwent some cleaning, e.g., Naïve → Naive; Piauí → Piaui; Xi'an → Xian. The list of words that are candidates for checking can be obtained using GREP and a regular expression like [^a-zA-Z0-9\ \.\,\;\:\-_\?\–\~\+\<\>\°\%\@\[\]\(\)\/\—\=*\−\€], from the resulting list, you can select the words you want to replace, then use a similar regular expression again to remove unnecessary characters, e.g., ®. I prefer such "manual" tests over the use of commonly available scripts, because I can see exactly what is changed, what is deleted, and the list of substitutions may vary depending on the task, for example, the program being applied may interrupt, considering one or another character unacceptable.

The resulting list of words and phrases was reduced to a list of unique words, which were further used as a dictionary for GSDMM.

Certainly, it is possible to use bigrams as dictionary terms, but then they should also be marked as bigrams in the title and annotation texts. This is easily accomplished by replacing the spaces in the bigrams with underscores. Such work goes beyond the scope of this publication, written in the spirit of proof of concept, and requires a separate analysis of the effect of dictionary compilation on the results of clustering using the GSDMM algorithm.

Krovetz stemmer [14] was applied to the dictionary and title and annotation texts.

In fact, Krovetz is a morphological analyzer that reduces morphological variants to a root form. For example, 'elephants'->'elephant', 'amplification'->'amplify', and 'european'->'europe' supplemented with a few rules for words not included in the replacement dictionaries.

The choice of this stemming method is due to its "softness"; it changes the spelling of words minimally, compared to Porter's stemming method. Its second advantage comes from the ability to edit dict_supplement.txt, and direct_conflations.txt files. [https://github.com/diazf/kstem].

After the above procedures, the GSDMM dictionary included 739 terms. By applying deduplication, we finally obtained 617 terms used in the GSDMM dictionary.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2023 doi:10.20944/preprints202307.0677.v1

The topics of the clusters were disclosed by means of a brief annotation of the publications found by a number of the most frequent terms for a given cluster. The GSDMM generates both a list of terms and their occurrence in the cluster and a list of publications related to that cluster. Having a list of records related to a given cluster, it is possible to describe the cluster topics using the methods offered by VOSviewer or Bibliometrix. The results are illustrative, but they are not included here, because this study focuses on describing the topics with a set of articles, not with key terms.

Research results

Preparing the dictionary for the GSDMM

Table 1 shows the 30 terms generated by the Yake! script with the highest score out of a total of 2,000 terms, including 1,554 bigrams and 446 single words.

Table 1. The 30 terms with the highest scores.

keyword	score	keyword	score	keyword	score
Renewable Energy	5.37E-07	Energy Management	5.68E-06	power generation	1.12E-05
energy system	1.27E-06	sustainable energy	6.48E-06	energy performance	1.15E-05
energy	1.48E-06	System	6.55E-06	energy generation	1.19E-05
energy consumption	1.90E-06	power	7.68E-06	Energy Power	1.22E-05
power system	2.01E-06	energy production	8.40E-06	energy demand	1.25E-05
energy storage	2.38E-06	energy transition	8.45E-06	storage system	1.26E-05
energy efficiency	2.77E-06	wind power	9.02E-06	sustainable development	1.29E-05
energy sources	3.59E-06	thermal energy	9.22E-06	energy resources	1.31E-05
solar energy	3.97E-06	case study	9.78E-06	systems	1.38E-05
wind energy	5.01E-06	Building Energy	1.04E-05	power plants	1.43E-05

Bigrams dominate the table. They are well suited as keywords to describe the general subject matter of the corpus of texts compiled from the headings and annotations of the 2,920 original records.

Note that the terms obtained with Yake! were converted to lower case, subjected to Krovetz stemming, a list of biterms was converted into a list of words, followed by deduplication. For example, after stemming, the plural word was translated into the singular and became the same as the singular word. Before applying the GSDMM algorithm, the header and annotation texts were subjected to a similar procedure, except for word deduplication.

Note: A list of terms generated by Yake! can also be used as a filter for each of the articles to obtain a list of generated keywords for each article. These keywords can be used in a similar way to "indexed keywords" in Scopus.

Table 2 shows examples of terms included in the dictionary file used by the GSDMM algorithm.

Table 2. 100 terms from the dictionary used by the GSDMM algorithm.

Term	Term	Term	Term	Term
renewable	demand	electric	wave	economic
energy	development	flow	intensity	global
system	resources	integrated	future	neural
consumption	systems	battery	china	network
power	plants	distribution	quality	optimal
storage	sector	distributed	fuel	reduce
efficiency	hybrid	results	cell	planning

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2023 doi:10.20944/preprints202307.0677.v1

sources	green	security	models	research
solar	supply	vehicles	market	simulation
wind	model	poverty	optimization	operation
management	based	turbines	sustainability	primary
sustainable	analysis	clean	hydrogen	data
production	carbon	technologies	high	european
transition	emissions	cost	communities	union
thermal	photovoltaic	industry	saving	gas
case	potential	ghg	smart	paper
study	control	show	maximum	environmental
building	grid	method	heat	impact
generation	policy	recovery	current	conversion
performance	electrical	proposed	strategy	total

As in the case of Table 1, these terms describe the topic of sustainable energy well. The clustering of publications according to the occurrence of these terms can be considered adequate for the task at hand.

Note: In text analysis, it is difficult to give an unambiguously "right" solution. The choice of methods to analyze and preprocess texts and the criteria for comparison are essentially subjective. For example, we can choose bigrams and trigrams generated by Yake! as a dictionary, combine the words contained in them by underlining, do the same for the source texts, and then apply the GSDMM algorithm.

Results of text clustering using the GSDMM algorithm

Records clustering with the GSDMM algorithm was carried out using the resulting dictionary and prepared texts.

The following parameters were used for the GSDMM algorithm: -a 0.1 -b 0.1 -m 1000 -k 10.

Selecting the parameters: -a 0.1 -b 0.1 and k 10 are recommended parameters, e.g. [https://github.com/jrmazarura/GPM, https://github.com/rwalk/gsdmm] [15]. A large number of iterations (-m 1000) was chosen simply to observe the convergence of the clustering results. In our case, a fast convergence to 9 clusters was observed.

A Rust implementation of the GSDMM algorithm [https://github.com/rwalk/gsdmm-rust] was used. The algorithm is fast, so there were no problems with choosing a large number of iterations.

Selection of publications that describe the theme of the cluster

The terms with the highest GSDMM score for a given cluster can be used to select publications that describe the topic of a particular cluster. However, the research has shown that this approach is not very clear due to the frequent occurrence of words common to all clusters, such as energy, research, result, analysis, target, implement, develop, approach. For this reason, in this paper the following approach was implemented to select the terms, which were used to search the publications out of 2920 exported records:

- In the first step, a list of terms occurring in the lists of all clusters was compiled consistently using INNER JOIN. There were 210 of these terms.

- Then the terms common to all clusters were removed from the term list of each cluster using the EXCEPT operator.

Table 3 shows examples of the described choice of terms for the 9 clusters.

Table 3. The 20 most frequent terms in each cluster, excluding terms occurring in all clusters.

Cl-1	Cl-2	Cl-3	Cl-4	Cl-5	Cl-6	Cl-7	Cl-8	Cl-9	
development	187 temperature	66 development	175 building	150 development	93 numerical	124 load	104 temperature	20 data	153
data	104 property	62 case	148 case	78 policy	91 temperature	114 problem	100 fluid	12 operation	152
important	96 biomass	58 plant	134 temperature	61 data	82 characteristic	83 operation	92 operation	11 voltage	148
case	92 material	57 policy	119 residential	54 china	75 fluid	83 schedule	72 net	10 load	137
article	87 combustion	53 future	115 data	53 country	63 case	61 case	71 plant	9 network	135
social	87 chemical	51 photovoltaic	113 city	38 finding	63 mass	60 objective	71 rankine	9 strategy	109
challenge	84 alternative	48 scenario	109 comfort	38 relationship	59 distribution	58 distribution	66 organic	8 photovoltaic	108
policy	83 surface	46 country	108 annual	35 growth	57 large	58 strategy	66 recovery	8 problem	102
framework	79 engine	45 costs	103 type	35 positive	52 turbine	57 effectiveness	64 refrigeration	8 wind	100
country	77 promising	45 wind	103 material	34 empirical	49 surface	56 network	62 turbine	7 converter	95
review	77 oil	44 fossil	99 environment	33 panel	48 important	50 distribute	60 configuration	6 accuracy	94
transition	74 organic	43 data	95 strategy	33 negative	47 development	47 wind	59 collector	5 machine	93

doi:10.20944/preprints202307.0677.v1

europe	70 solid	39 review	93 condition	30 evidence	45 dynamics	47 management	57 mass	5 neural	86
future	70 biodiesel	38 sector	91 construction	30 role	42 material	46 uncertainty	54 mathematics	5 controller	85
finding	68 type	37 strategy	90 load	28 green	41 data	44 voltage	52 capture	4 case	83
strategy	68 fossil	34 resources	89 saving	27 important	40 operation	43 battery	51 costs	4 error	83
related	67 density	33 transition	86 urban	27 industry	39 speed	43 dispatch	49 desalination	4 fault	80
sector	67 biofuel	32 important	83 important	26 economy	35 behavior	42 photovoltaic	47 generator	4 development	79
plan	64 cell	31 plan	76 alternative	25 regress	34 average	41 ieee	46 location	4 learning	79
literature	63 hydrogen	31 project	74 development	25 affect	33 wind	40 scenario	45 parametric	4 distribution	78

doi:10.20944/preprints202307.0677.v1

 doi:10.20944/preprints202307.0677.v1

It should be emphasized that exactly the terms occurring in all clusters are excluded. Terms which occur in only a few clusters are retained. Examples are: development, temperature, operation, policy, etc. As has been repeatedly noted, when analyzing texts, options are possible, the main thing is that the article specifies which one is chosen. In this case, the publications were selected according to the most common terms in Table 3. The second criterion was the selection of those terms that, when used together, allowed us to find at least 2–4 articles in the publication records of a given cluster that contained the maximum number of terms from Table 3.

The publications selected in this way have been used as information for the description of one of the actual tasks of the cluster topics.

Description of the topics of the 9 clusters, with the texts of short abstracts of the articles selected for each cluster

Once entries are selected for each cluster, their titles, abstracts and DOIs are available.

Using DOI it is convenient to generate bibliographic references in the required format using services like https://zbib.org/.

To avoid repeating the information given in the bibliographic reference, only a brief abstract was used to describe the subject of the publication. This procedure can be done using a suitable automatic text summarization service or scripts, such as, https://miso-belica.github.io/sumy/.

Cluster 1. General focus on sustainable energy issues

The aim of the study [16] is to identify, map and assess the maturity and impact level of the specific energy-oriented economy and other SMART management concepts and social, technological, finance (economical), environmental, and communication (S.T.F.E.C.) trends which arose from the dynamic development and spread of the Industry 4.0 revolution on processes of effective competitiveness and the creation of modern enterprises. The authors aim to search for answers to three specific research questions, concluding that recently, special attention is paid to such issues as co-creation and co-production, energy-oriented and circular economy, eco-energy, and sustainability. Researched data allows us to conclude that openness to social, environmental, and technological trends and issues, with an approach based on sustainable and eco-energy-oriented development, play an increasingly important role.

Open innovations (OI) are playing an increasingly important role in the innovative development (RI) of SMEs. This has led to a need to analyze the impact of OI on innovative development serving the implementation of the assumptions of sustainable development, the positive effect of which is to reduce the negative impact on the environment thanks to a more rational use of both natural and produced resources (e.g., energy). Moreover, the results show that SMEs cooperating with the environment are more developed in terms of sustainable innovative development than those that base their development on their own internal resources (no cooperation). Hence, it follows that OIs have a positive impact on sustainable innovative development [17].

Inclusion of a sustainability framework and managing ESG-related risks have become part of the overall strategy of most companies within the energy industry. The current research focuses on the systematic literature on the M&A deals in the energy industry through the lens of sustainability by applying the SALSA methodology. Further, we applied a SWOT analysis of M&A in the energy industry from the perspective of sustainable development [18].

The present research [19] prioritized the most important direction of energy power transformation — energy sector digitalization — and its contribution to the achievement of the sustainable development goals focused on climate change mitigation and responsible consumption and production. The authors evaluated the economic efficiency of the implementation of digital tools due to the decrease in energy production costs.

Design/methodology/approach: The study [20] performed bibliometric analysis to investigate sustainable energy research between 1980 and 2022 using a sample of 1498 research papers from the Web of Science (WoS) databases, with only published articles on sustainable energy. Findings: A bibliometric analysis reveals trends in sustainable energy research publications, showing sustainable

 doi:10.20944/preprints202307.0677.v1

energy as an emerging topic and trends in sustainability and energy research. The fact that the keywords "sustainable energy", "renewable energy", "sustainability", and "sustainable development" are frequently included in the literature shows that interdisciplinary academic studies in these fields are of great importance.

Cluster 2. Highlighting thermal biomass processing

The experimental results [21] show that once the biomass is made into a briquette, when the reaction temperature is 900 °C, the sulphur release ratio for TB was reduced from 34.7% to 4.3% and for WS was reduced from 12.4% to 1.6%. When the reaction temperature increases to 1000 °C, the sulphur release ratio for TB was reduced from 73.4% to 30.4%, for WS it was reduced from 58.4% to 10.2%.

The prediction and pre-evaluation of the thermal properties and combustion-related problems (e.g., emissions and ash-related problems) are critical to reducing emissions and improving combustion efficiency during the agricultural crop residues combustion process. This study [22] integrated the higher heating value (HHV) model, specific heat model, and fuel indices as a new systematic approach to characterize the agricultural crop residues. The specific heat of flue gas during the combustion process was estimated from the concentrations of C, H, O, S, and ash content under various excess air (EA) ratios and flue gas temperatures.

The produced syngas in the two different tests was characterized and compared in terms of composition (H_2, CH_4, CO, CO_2, O_2) and fate of contaminants such as volatile organic compounds (VOCs), tar and metals. The aim of this work [23] is to show which advantages and disadvantages there are in choosing the most suitable material and to optimize the biomass gasification process by reducing the undesirable effects, such as heavy metal production, bed agglomeration and tar production, which are harmful when syngas is used in internal combustion engines (ICE).

Cluster 3. Highlighting the problems of specific regions

This paper [24] estimates the life-cycle land-use requirement for PV development in Vietnam, to provide the scientific-based evidence for policy makers on the quantity of land required, so that the land budget can be suitably allocated. Regarding the life-cycle land use, the land occupation is 241.85 m2a and land transformation is 16.17 m2 per MWh. Most of the required land area is for the installation of the PV infrastructure, while the indirect land use of the background process is inconsiderable.

The research [25] analyses how the future of the biogas business in three case study countries is developing until 2030. The study is based on experts' views within the biogas business branch in Germany, The Netherlands, and Finland. To be able to show the full potential of biogas technology for society, stable and predictable energy policy and cross-sector co-operation are needed.

The article [26] explores the viability of renewable energy using the strengths, weaknesses, opportunities and threats (SWOT) analysis approach on the key renewable potential in the country (South Africa). Several opportunities favor switching to renewable energy, and these include regional integration, global awareness on climate change and the continuous electricity demand. Some threats hindering the renewable energy sector in the country include land ownership, corruption and erratic climatic conditions.

Sustainability indicators are biodiversity loss, fossil-fuel use, mineral depletion, energy use, carbon emissions and eutrophication. The most promising results arise from shifting consumption of meat and animal-based products to a more plant-based diet, and transitioning to organic agriculture. Net-zero sustainability goals and a reduction in eutrophication are achieved by 75% downshift of animal products and the upscaling of organic agriculture (German agriculture) [27].

 doi:10.20944/preprints202307.0677.v1

Cluster 4. Building energy consumption issues

The paper [28] examines how to achieve an appropriate model for integrating photovoltaics on the rooftop of residential buildings in Hail city to provide alternative energy sources. The results show a significant area of rooftop suitable for PV system in residential buildings in Hail city, which exceeds 9 million square meters. There is a significant amount of energy produced from the use of all residential rooftops in Hail, and there is also a significant reduction in the amount of CO_2 emissions.

This study [29] investigates the spatial patterns in the surface urban heat island (SUHI) over the study site and develops its relationships to socioeconomic, demographic, and buildings' characteristics. Numerous studies have focused primarily on the influence of biophysical and meteorological factors on variations in land surface temperatures (LSTs); however, very little attention has been paid to examining the influence of socioeconomic, demographic, and building factors on SUHIs within a city. Linear regression and multiple regression correlations are further run to examine selected factors' variance on SUHI.

One of the most important reasons for the high consumption of electrical energy in RBs is the big difference between indoor and outdoor temperatures. In this paper [30], a heat exchanger was designed and tested experimentally to reduce this temperature difference by using a domestic ground water tank (GWT) as a sink/source (water-cooled condensers instead of air-cooling). The proposed system resulted in a reduction in energy consumption by 28% of the electrical energy needed in the conventional system and an increase in COP by 39%.

This research [31] is concerned with focusing on the indirect effect of solar photovoltaic rooftop panels (shading effect) on the roof surface to see whether this effect is worth studying and calculating the total electrical load in the residential sector. The results highlight that renewable energy is very important in our times due to climate change and the increased demand for electricity by the residential sector, which is stimulated to find multiple ways to decrease and adapt to this change, and the aim of this paper helps to encourage to use solar energy by identifying the indirect effect of solar panels on building's rooftops.

Cluster 5. Economic Growth, Renewable Energy, and Investment Issues

Situated within the current trend of declining foreign direct investment flows (FDI), the study [32] examines the role of ESG factors in attracting FDI and enabling progress toward SDGs. We econometrically examine the linkages between ESG and FDI inflows for a sample of 161 counties. Results suggest that FDI inflows to the full sample of countries are positively attracted by good governance in a destination country. Sustainability reporting attracts FDI to commodity exporting countries.

The present research paper [33] intends to investigate the relationship between economic growth and sustainable financial development on the use of energy from renewable sources in both the short and long run in the context of China. In the case of the short run, there is a positive relationship between economic and financial development and the use of energy from renewable sources in the context of all of China. While in the case of long-term effects, the results indicate the adverse impact of financial development on the use of energy from renewable sources in the western regions of China.

This study [34] applies the technology acceptance model and incorporates environmental concerns, value propositions, and government policies as variables to explore the behavioral intentions of Taiwan's Generation Z toward using electric motorcycles. The study revealed that: (1) consumers' perceived usefulness and perceived ease of use positively influence their attitudes toward using electric motorcycles; (2) consumers' environmental concerns do not influence their attitudes toward using electric motorcycles; and (3) consumers' attitudes toward using electric motorcycles, value propositions, and government policies positively influence their behavioral intentions toward using electric motorcycles.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 July 2023 doi:10.20944/preprints202307.0677.v1

Cluster 6. Numerical flow modeling issues

The present study [35] deals with the numerical simulation of mixed convective heat transfer from an unconfined heated square cylinder using nanofluids (Al_2O_3-water) for Reynolds number (Re) 10–150, Richardson number (Ri) 0–1, and nanoparticles volume fractions (φ) 0–5%. Minor variations in flow and thermal characteristics are observed between the two approaches for the range of nanoparticle volume fractions considered. The local and mean Nusselt numbers increase with Reynolds number, Richardson number, and nanoparticle volume fraction.

This paper [36] studies the characteristics of the hydraulic PTO (power-take-off) experimental verification. The performances of the hydraulic PTO in start-up processes with different initial temperatures and in long term operation are assessed. The efficiency of hydraulic PTO degrades when it starts at low temperatures. An improved numerical model of the hydraulic PTO system is proposed.

The numerical simulation of the external flow field and structure coupling of the aerodynamic heat problem is performed [37]. Numerical results indicate that the average temperature and maximum temperature of the optical dome for inner and outer walls exhibit an "M" shape with time, with two high-temperature cusps and one low-temperature cusp. Therefore, this study provides a reliable reference for the preliminary design and parameter research of optical domes of hypersonic aircraft.

Cluster 7. Generation cost optimization planning models and generation deviations

Aiming at the impact of the uncertainty of source load on the optimal scheduling in an integrated energy system (IES), based on hybrid resolution modeling and hybrid instruction cycle scheduling technology, three-time scales of day-ahead, intra-day rolling and real-time feedback optimization scheduling models are established [38]. Then, the chaotic gravitational search algorithm (CGSA) is used to solve the problem, and the composite coordination optimization operation strategy of IES with mixed time scales based on CGSA is proposed.

The proposed SA-SOPF has objective to find a day-ahead base-solution that minimizes the generation cost and expectation of deviations in generation and node voltage set-points during real-time operation. The results also depict that the proposed SA-OPF formulation can reduce the expectation in voltage and generation deviation more than 60% in real-time operation with an additional day-ahead scheduling cost of 4.68% only for 14-bus system. These results are strong indicators of possibility of achieving the day-ahead solution which lead to lower real-time deviation with minimal cost increase [39].

This work [40] addresses this problem by proposing an operational planning approach to determine the optimal allocation of WT units, PV systems, and hybrid energy storage systems (HESS) in smart grids. The proposed operational planning approach is formulated as a nested optimization problem that guarantees the optimal planning and operation of the RES and HESS simultaneously.

The paper [41] proposes a power flow optimization strategy model of a distribution network with non-fixed weighting factors of source, load and storage. The k-means algorithm is used to cluster the equivalent load curve in different periods, and then the fuzzy comprehensive evaluation method is used to determine the weighting factor of the optimization model in each period. Then, the particle swarm optimization algorithm is used to solve the multi-objective power flow optimization model, and the optimal strategy and objective function values of each unit output in the operation period are obtained.

Cluster 8. The smallest cluster, only 31 entries. Thermal-to-electrical energy conversion problems

This paper [42] proposes a new combined multi-cooling and power generation system (CMCP) driven by solar energy. The PTC mathematical model is used to calculate the heat transfer fluid outlet temperature and the performance of the CMCP system on a specific day of the year. A 1D model of an ejector with a constant area is adopted to evaluate the ejector performance.

 doi:10.20944/preprints202307.0677.v1

Converting waste heat into electricity has captured the interest of scientists for years because of its enormous potential to improve energy efficiency and to lessen environmental impacts. We concluded that the combination of the Kalina Cycle and Organic Rankine Cycle can be efficiently used for the recovery of waste heat energy [43].

In the experiments [44], the heat source temperature as well as the mass flow rates of the working fluid and cooling water were controlled. As for the influences of key parameters, with the increase in heat source temperature from 130 °C to 160 °C, the involved heat has a small increase, while the net work increases from 0.44 kW to 0.55 kW, and the cycle efficiency greatly increases from 6.71% to 8.72% at a mass flow rate of working fluid 25 g/s. As for the mass flow rate of cooling water, it has a similar impact on the cycle performances.

Cluster 9. Weather-related risk issues and false data injection attacks

In recent years, demand for electric energy has steadily increased; therefore, the integration of renewable energy sources (RES) at a large scale into power systems is a major concern. Clouds' influence on solar irradiation forecasting, data categorization per month for successive years due to the similarity of patterns of solar irradiation per month during the year, and relative seasonal similarity of windspeed patterns have not been taken into consideration in previous work. Moreover, the similarity of patterns of solar irradiation per month during the year and the relative seasonal similarity of windspeed patterns in a timeseries measurements dataset for several successive years demonstrates that they contribute to very high one-day-ahead windspeed and solar irradiation forecasting performance [45].

This paper [46] proposes distributed mitigation layers for the false data injection attacks (FDIA) on voltages and currents of DGUs in meshed DC microgrids. The proposed control strategy is based on integrating two layers for cyber-attack detection and mitigation to immune the primary and the secondary control loops of each DGU.

Remark:

The publications listed above characterize a particular class of problems within a given cluster of publications. Publications were selected by a set of frequently occurring terms for a given cluster (usually four). The appearance of four terms in the bibliometric record of a publication (without a list of references) significantly narrows the sample, especially since the number of records in the cluster is small (several hundred). The sample was generated by querying a sequence of terms (e.g., (.*building)(.*environment)(.*city)(.*temperature).*$), which further reduced the sample size and narrowed the topics of the selected publications. A different set of terms could be compiled to produce a different list of publications, but it is unlikely that there will be publications whose records contain four terms that are not frequently encountered. Another approach for sampling can be a query based not on the sequence of occurring terms, but on their co-occurrence (AND operator). The approach outlined above is not so much focused on any "correct" description of clusters, but rather on identifying tasks that cannot be "turned a blind eye," and for which there are enough publications in the collected data for a more detailed review. Naturally, an analytical review of a specifically chosen problem would require a follow-up, more specialized search of publications.

Conclusion

Applying the approach proposed in this paper to identify current research subjects using Yake!, Krovetz, GSDMM, and short text summaries to analyze bibliometric records of MDPI journals on sustainable energy issues, the following key topics of current research were identified::

1. Assessing the maturity and impact of the energy-based economy, SMART management concepts and social, technological, financial, economic, environmental and communication (S.T.F.E.C.) trends arising from the Industry 4.0 revolution on the competitiveness and creation

 doi:10.20944/preprints202307.0677.v1

of modern enterprises. Research priorities are the digitalization of the energy sector and its contribution to climate change mitigation and responsible consumption and production.

2. A study to reduce sulphur emissions from the production and use of biomass briquettes. A study to characterize crop residues and determine the most suitable conditions for biomass gasification. A study to minimize the formation of heavy metals and tar when using syngas in internal combustion engines.
3. Life cycle assessment of land use requirements for photovoltaic development. Energy policy and cross-sectoral cooperation studies. Issues of renewable energy viability in developing countries, regional integration and electricity demand. Assessment of sustainable development indicators, including biodiversity and mineral depletion.
4. Identification of areas suitable for the integration of residential rooftop photovoltaic systems. Study spatial patterns of urban surface heat islands and their relationship to socio-economic, demographic and building characteristics. Development of heat exchangers to reduce temperature differences between indoors and outdoors. Investigation of the role of renewable energy sources in the domestic sphere.
5. Study the role of environmental, social and governance factors to attract foreign direct investment in resource exporting countries. Research on the relationship between economic growth and the financing of renewable energy sources. A behavioral intention study on electric vehicles.
6. Numerical simulation of mixed convective heat transfer using nanofluids in a heated reservoir. Experimental study of power take-off in a hydraulic system during start-up and long-term operation. Research and design of optical domes for hypersonic aircraft.
7. Studies the impact of source load uncertainty on optimal scheduling in integrated power systems using hybrid resolution modelling and instruction cycle scheduling technology. Issues of operational planning and simultaneous operation of RES and HEPS. Modelling strategies for optimizing power flows in distribution networks.
8. Study of combined multi-cooling and solar thermal power generation systems for waste heat recovery.
9. Issues in the integration of renewable energy into energy systems. A study of the effect of clouds on solar radiation forecasting, predicting wind speed and solar radiation one day in advance. Analysis of cyber-attack detection and mitigation to protect primary and secondary control loops in power grids.

To do

To use Yake! and KeyBERT jointly to form the vocabulary for the GSDMM algorithm.

Utilize a dictionary containing bigrams and trigrams, and respectively prepare clustered texts by selecting compound terms from the dictionary using underscore.

References

1. Van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010;84:523–38. https://doi.org/10.1007/s11192-009-0146-3.
2. Aria M, Cuccurullo C. bibliometrix : An R-tool for comprehensive science mapping analysis. Journal of Informetrics 2017;11:959–75. https://doi.org/10.1016/j.joi.2017.08.007.
3. Grootendorst M. MaartenGr/KeyBERT: BibTeX 2021. https://doi.org/10.5281/ZENODO.4461265.

4. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure 2022. https://doi.org/10.48550/ARXIV.2203.05794.
5. Wadnare RJ, Sherekar DrSS, Thakare DrVM. Development of Text Clustering Method with K-Means for Analysis of Text Data. IJSRCSEIT 2021:143–51. https://doi.org/10.32628/CSEIT217237.
6. Duo J, Zhang P, Hao L. A K-means Text Clustering Algorithm Based on Subject Feature Vector. JWE 2021. https://doi.org/10.13052/jwe1540-9589.20612.
7. Uthirapathy SE, Sandanam D. Topic Modelling and Opinion Analysis On Climate Change Twitter Data Using LDA And BERT Model. Procedia Computer Science 2023;218:908–17. https://doi.org/10.1016/j.procs.2023.01.071.
8. Gupta RK, Agarwalla R, Naik BH, Evuri JR, Thapa A, Singh TD. Prediction of research trends using LDA based topic modeling. Global Transitions Proceedings 2022;3:298–304. https://doi.org/10.1016/j.gltp.2022.03.015.
9. Mazarura J, De Waal A. A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text. 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), Stellenbosch, South Africa: IEEE; 2016, p. 1–6. https://doi.org/10.1109/RoboMech.2016.7813155.
10. Weisser C, Gerloff C, Thielmann A, Python A, Reuter A, Kneib T, et al. Pseudo-document simulation for comparing LDA, GSDMM and GPM topic models on short and sparse text using Twitter data. Comput Stat 2023;38:647–74. https://doi.org/10.1007/s00180-022-01246-z.
11. Zhao Y, Liang S, Ren Z, Ma J, Yilmaz E, De Rijke M. Explainable User Clustering in Short Text Streams. Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, Pisa Italy: ACM; 2016, p. 155–64. https://doi.org/10.1145/2911451.2911522.
12. Yin J, Wang J. A Text Clustering Algorithm Using an Online Clustering Scheme for Initialization. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA: ACM; 2016, p. 1995–2004. https://doi.org/10.1145/2939672.2939841.
13. Campos R, Mangaravite V, Pasquali A, Jorge A, Nunes C, Jatowt A. YAKE! Keyword extraction from single documents using multiple local features. Information Sciences 2020;509:257–89. https://doi.org/10.1016/j.ins.2019.09.013.
14. Krovetz R. Viewing morphology as an inference process. Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '93, Pittsburgh, Pennsylvania, United States: ACM Press; 1993, p. 191–202. https://doi.org/10.1145/160688.160718.
15. Yin J, Wang J. A dirichlet multinomial mixture model-based approach for short text clustering. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, New York New York USA: ACM; 2014, p. 233–42. https://doi.org/10.1145/2623330.2623715.
16. Adamik A, Nowicki M, Puksas A. Energy Oriented Concepts and Other SMART WORLD Trends as Game Changers of Co-Production—Reality or Future? Energies 2022;15:4112. https://doi.org/10.3390/en15114112.
17. Stanisławski R. Characteristics of Open Innovation among Polish SMEs in the Context of Sustainable Innovative Development Focused on the Rational Use of Resources (Energy). Energies 2022;15:6775. https://doi.org/10.3390/en15186775.
18. Andriuškevičius K, Štreimikienė D. Energy M&A Market in the Baltic States Analyzed through the Lens of Sustainable Development. Energies 2022;15:7907. https://doi.org/10.3390/en15217907.
19. Galkovskaya V, Volos M. Economic Efficiency of the Implementation of Digital Technologies in Energy Power. Sustainability 2022;14:15382. https://doi.org/10.3390/su142215382.
20. Kemeç A, Altınay AT. Sustainable Energy Research Trend: A Bibliometric Analysis Using VOSviewer, RStudio Bibliometrix, and CiteSpace Software Tools. Sustainability 2023;15:3618. https://doi.org/10.3390/su15043618.
21. Qi J, Li H, Wang Q, Han K. Combustion Characteristics, Kinetics, SO2 and NO Release of Low-Grade Biomass Materials and Briquettes. Energies 2021;14:2655. https://doi.org/10.3390/en14092655.
22. Qian X, Xue J, Yang Y, Lee SW. Thermal Properties and Combustion-Related Problems Prediction of Agricultural Crop Residues. Energies 2021;14:4619. https://doi.org/10.3390/en14154619.
23. Vincenti B, Gallucci F, Paris E, Carnevale M, Palma A, Salerno M, et al. Syngas Quality in Fluidized Bed Gasification of Biomass: Comparison between Olivine and K-Feldspar as Bed Materials. Sustainability 2023;15:2600. https://doi.org/10.3390/su15032600.

24. Sanseverino ER, Cellura M, Luu LQ, Cusenza MA, Nguyen Quang N, Nguyen NH. Life-Cycle Land-Use Requirement for PV in Vietnam. Energies 2021;14:861. https://doi.org/10.3390/en14040861.
25. Winquist E, Van Galen M, Zielonka S, Rikkonen P, Oudendag D, Zhou L, et al. Expert Views on the Future Development of Biogas Business Branch in Germany, The Netherlands, and Finland until 2030. Sustainability 2021;13:1148. https://doi.org/10.3390/su13031148.
26. Uhunamure SE, Shale K. A SWOT Analysis Approach for a Sustainable Transition to Renewable Energy in South Africa. Sustainability 2021;13:3933. https://doi.org/10.3390/su13073933.
27. Ahrens F, Land J, Krumdieck S. Decarbonization of Nitrogen Fertilizer: A Transition Engineering Desk Study for Agriculture in Germany. Sustainability 2022;14:8564. https://doi.org/10.3390/su14148564.
28. Abdelhafez MHH, Touahmia M, Noaime E, Albaqawy GA, Elkhayat K, Achour B, et al. Integrating Solar Photovoltaics in Residential Buildings: Towards Zero Energy Buildings in Hail City, KSA. Sustainability 2021;13:1845. https://doi.org/10.3390/su13041845.
29. Sidiqui P, Tariq MAUR, Ng AWM. An Investigation to Identify the Effectiveness of Socioeconomic, Demographic, and Buildings' Characteristics on Surface Urban Heat Island Patterns. Sustainability 2022;14:2777. https://doi.org/10.3390/su14052777.
30. Almasri RA, Abu-Hamdeh NH, Alajlan A, Alresheedi Y. Utilizing a Domestic Water Tank to Make the Air Conditioning System in Residential Buildings More Sustainable in Hot Regions. Sustainability 2022;14:15456. https://doi.org/10.3390/su142215456.
31. Albatayneh A, Albadaineh R, Juaidi A, Abdallah R, Zabalo A, Manzano-Agugliaro F. Enhancing the Energy Efficiency of Buildings by Shading with PV Panels in Semi-Arid Climate Zone. Sustainability 2022;14:17040. https://doi.org/10.3390/su142417040.
32. Chipalkatti N, Le QV, Rishi M. Sustainability and Society: Do Environmental, Social, and Governance Factors Matter for Foreign Direct Investment? Energies 2021;14:6039. https://doi.org/10.3390/en14196039.
33. Guan D, Comite U, Sial MS, Salman A, Zhang B, Gunnlaugsson SB, et al. The Impact of Renewable Energy Sources on Financial Development, and Economic Growth: The Empirical Evidence from an Emerging Economy. Energies 2021;14:8033. https://doi.org/10.3390/en14238033.
34. Zhang X, Chang M. Applying the Extended Technology Acceptance Model to Explore Taiwan's Generation Z's Behavioral Intentions toward Using Electric Motorcycles. Sustainability 2023;15:3787. https://doi.org/10.3390/su15043787.
35. Rajpoot RS, Dhinakaran Shanmugam, Alam MdM. Numerical Analysis of Mixed Convective Heat Transfer from a Square Cylinder Utilizing Nanofluids with Multi-Phase Modelling Approach. Energies 2021;14:5485. https://doi.org/10.3390/en14175485.
36. Niu Y, Gu X, Yue X, Zheng Y, He P, Chen Q. Research on Thermodynamic Characteristics of Hydraulic Power Take-Off System in Wave Energy Converter. Energies 2022;15:1373. https://doi.org/10.3390/en15041373.
37. Wang Z, Zhang A, Pan J, Lu W, Sun Y. Fluid-Thermal Interaction Simulation of a Hypersonic Aircraft Optical Dome. Energies 2022;15:8619. https://doi.org/10.3390/en15228619.
38. Liu Z, Guo F, Liu J, Lin X, Li A, Zhang Z, et al. A Compound Coordinated Optimal Operation Strategy of Day-Ahead-Rolling-Realtime in Integrated Energy System. Energies 2023;16:500. https://doi.org/10.3390/en16010500.
39. Pareek P, Nguyen HD. State-Aware Stochastic Optimal Power Flow. Sustainability 2021;13:7577. https://doi.org/10.3390/su13147577.
40. Ali A, Shaaban MF, Sindi HF. Optimal Operational Planning of RES and HESS in Smart Grids Considering Demand Response and DSTATCOM Functionality of the Interfacing Inverters. Sustainability 2022;14:13209. https://doi.org/10.3390/su142013209.
41. Zheng F, Meng X, Wang L, Zhang N. Power Flow Optimization Strategy of Distribution Network with Source and Load Storage Considering Period Clustering. Sustainability 2023;15:4515. https://doi.org/10.3390/su15054515.
42. Hammemi R, Elakhdar M, Tashtoush B, Nehdi E. Multi-Objective Optimization of a Solar Combined Power Generation and Multi-Cooling System Using CO2 as a Refrigerant. Energies 2023;16:1585. https://doi.org/10.3390/en16041585.
43. Öksel C, Koç A. Modeling of a Combined Kalina and Organic Rankine Cycle System for Waste Heat Recovery from Biogas Engine. Sustainability 2022;14:7135. https://doi.org/10.3390/su14127135.

44. Gao Y, Song Q, Su W, Lin X, Sun Z, Huang Z, et al. Experimentally Identifying the Influences of Key Parameters for an Organic Rankine Cycle Using R123. Sustainability 2023;15:814. https://doi.org/10.3390/su15010814.
45. Blazakis K, Katsigiannis Y, Stavrakakis G. One-Day-Ahead Solar Irradiation and Windspeed Forecasting with Advanced Deep Learning Techniques. Energies 2022;15:4361. https://doi.org/10.3390/en15124361.
46. EL-Ebiary A, Mokhtar M, Mansour A, Awad F, Marei M, Attia M. Distributed Mitigation Layers for Voltages and Currents Cyber-Attacks on DC Microgrids Interfacing Converters. Energies 2022;15:9426. https://doi.org/10.3390/en15249426.