Futures 117 (2020) 102511

Contents lists available at ScienceDirect

Futures

journal homepage: www.elsevier.com/locate/futures

Emerging technologies in the renewable energy sector: A comparison of expert review with a text mining software

Alberto Moro[a,*], Geraldine Joanny[b], Christian Moretti[c]

[a] European Commission, Joint Research Centre, Ispra, Italy
[b] European Commission, Joint Research Centre, Brussels, Belgium
[c] Utrecht University, Copernicus Institute of Sustainable Development, Utrecht, Netherlands

ARTICLE INFO

Keywords:
Bibliometrics
Text mining
Emerging technologies
Renewable energy
Quantitative and qualitative
Policy support

ABSTRACT

This paper compares the results from quantitative text mining to qualitative expert reviews to identify emerging technologies in the fields of solar photovoltaics (PV), wind power, ocean and tidal energy, hydropower. The text mining analysis is based on the software “Tools for Innovation Monitoring” (TIM). The TIM software extracts a set of relevant keywords from a corpus of pertinent scientific publications. TIM outputs are compared to those extracted by the software VOSviewer, showing agreement. The top 300 ranked keywords are the optimum trade-off between retrieved technologies and analyst efforts. The emerging technologies identified by the experts can be retrieved in the top 300 keywords with a probability of 65 %, 25 %, depending on the technology sector and the algorithm adopted. The more salient keywords tend to correspond to technologies with an established and univocal jargon such as: "dye sensitised solar cells" or "vertical axis wind turbines". Two methods are here used and compared: the frequency of author keywords and the term frequency-inverse document frequency (TF-IDF) algorithm. The comparison of their performances is not showing a general prevalence of one method against the other, but rather a different suitability to different technology sectors.

1. Introduction

1.1. Relevance, scope and structure of this paper

This paper compares qualitative technology assessments performed by experts with quantitative bibliometric and text mining analysis of relevant terms. Semi-automated software findings are quantitatively compared with qualitative cognitive expert reviews by adopting specific indicators described in our previous work (Moro, Boelman, Joanny, & Garcia, 2018). This work should be considered propaedeutic to readers interested in understanding in depth our methodology. In this paper some methodological implementation is presented. Results are discussed and used to suggest methodologies and approaches to the analysts interested in technology mapping by software.

Policy makers are targeting new low carbon energy technologies to both reduce greenhouse gas emissions and support economic growth (European Commission, 2017a). In this context, policy makers would benefit from the use of text mining and analysis software allowing to identify relevant emerging technologies. The study presented here is focused on a set of emerging technologies in different renewable energy sectors: photovoltaics, wind power, ocean and tidal energy, hydropower. No generalisation of the

[*] Corresponding author at: Joint Research Centre, via Enrico Fermi 2749, Ispra (Va), 21027, Italy.
E-mail addresses: alberto.moro@ec.europa.eu (A. Moro), geraldine.joanny@ec.europa.eu (G. Joanny), c.moretti@uu.nl (C. Moretti).

https://doi.org/10.1016/j.futures.2020.102511
Received 30 May 2019; Received in revised form 21 September 2019; Accepted 3 January 2020
Available online 13 January 2020

numerical results we found can be made for other technological sectors.

The Joint Research Centre (JRC) of the European Commission is working on bibliometric analysis to complement expert consultation for the detection and monitoring of emerging technologies (Joanny et al., 2015). In the current literature, it is recognised that focused expert reviews are more suitable than text mining methods in identifying weak signals (Amanatidou et al., 2012); so results from the JRC bibliometric and text mining software are tested against the results of expert reviews using a set of indicators.

Emerging technologies targeted in this study are deemed to be still far away from commercial deployment, with a Technology Readiness Level (TRL) not exceeding TRL 4 (De Rose et al., 2017).

1.2. *Software-based analysis*

Information on scientific and patent production can provide quantitative evidence to research and development (R&D) policies (Huang, Zhang, Guo, Zhu, & Porter, 2012). Bibliometric analysis, counting activity levels and identifying patterns in R&D bibliographic records plus patent analyses (Leydesdorff, 2015; Zhang, Yan, & Guan, 2014), can help to identify emerging technologies.

JRC has developed a text mining system for tracking the evolution of established and emerging technologies called Tools for Innovation Monitoring (TIM). TIM can retrieve bibliometric data from several sources: Scopus database of peer-reviewed scientific journals (Elsevier, 2017); Cordis, the database of European Union (EU) research projects (European Commission, 2017b), and Patstat, a worldwide database of patents (European Patent Office, 2017).

TIM can extract from the targeted data base a set of documents (data set) according to a specifically designed search string (see Section 2.2.1). Normally, similar tools are used to visualise networks of concepts and researchers operating in a sector by looking at the co-occurrence of terms and co-authoring (Ciano, Pozzi, Rossi, & Strozzi, 2019).

The use of this software, here, is different; we aim to identify emerging technologies in a specific sector. Therefore, the analysis focuses only on the keywords associated to the retrieved documents (also "extracted" keywords). It should be clarified that a keyword, in general, is not coinciding with the name of a technology but it could be univocally linked to it. A semantic post processing of the extracted keywords can help in revealing if the retrieved documents are dealing with emerging technologies.

Keywords can be ranked or prioritised according to different algorithms. In this paper two different keyword prioritisation (or ranking) algorithm (see Section 2.2.2) are used and compared.

It is not easy to assess and compare the efficacy of these algorithms in high-ranking keywords connected to emerging technologies. In a previous work we developed a set of indicators expressly designed to this scope (Moro et al., 2018).

These indicators (details in Section 2.3) associate a quantitative value to each keyword or keyword list extracted by the software in comparison with a set of emerging technologies identified by experts.

With some implementation (detailed in Section 2.4), these indicators are here used to compare the efficacy of the various algorithms and software (see Section 3.5) in prioritising keywords related to emerging technologies in the considered fields.

2. Methodology

The main approach followed in this paper was presented in detail in our previous work (Moro et al., 2018) and is here only summarised. Please refer to our previous paper for a full understanding of background, methodology and drawbacks. Summarising, the methodology relies on the fact that expert reviews, well established in horizon scanning, are considered one of the best methods to identify weak signals (Amanatidou et al., 2012). The retrieval of emerging technology related keywords by the software is benchmarked against marker keywords proposed by panels of experts for several energy sectors: photovoltaics (PV), wind power, ocean energy and hydropower.

The qualitative analysis from the experts is compared versus the quantitative analysis from the TIM software by using the indicators reported in Sections 2.3 and 2.4.

Two different TIM algorithms are tested and compared, one based on the count of the frequency of the author keywords of the publications and the other one based on the TF-IDF algorithm of terms for the whole dataset (Dillon, 1983; Salton & Buckley, 1988).

Quantitative results (Section 3) are discussed (Section 4) and general findings proposed to the scientific community. To extend the validity of the findings, a small inter-comparison exercise is performed (Section 3.5) between the outputs from TIM and those extracted using VOSviewer from Leiden University (Centre for Science & Technology Studies, L. U., 2019).

2.1. *Qualitative cognitive analysis by expert review*

International experts gathered at the Joint Research Centre and performed four different technology review exercises from 2016 to 2018. The experts were required to identify and assess emerging technologies in their specific research sectors. For the energy sectors *photovoltaics* (Moro et al., 2017) and *tidal and ocean* (Magagna et al., 2018) the full reports about such consultations are publicly available. For *wind energy* and *hydropower* we anticipate here some of the findings that are still pending publication.

The keywords characterising the technologies identified by the experts (marker keywords) are used to benchmark the keywords retrieved from bibliometric and text mining analysis by checking to what extent the keyword lists generated by the TIM software match with the marker keywords from the experts.

Table 1
Search strings used in the TIM software.

Aspect captured	String	Documents retrieved
"Future" aspects	ti_abs_key:(("future" OR "emerging" OR "innovative" OR "disruptive" OR "visionary" OR "exploratory" OR "unexpected" OR "new" OR "novel") NOT emergency) AND …	
Photovoltaic technologies	…AND ("solar power" OR "photovoltaic" OR "solar cell")	32691
Wind power technologies	…AND ("wind power" OR "wind energy" OR "wind turbine")	14500
Ocean and tidal techs.	…AND ("ocean energy" OR "tidal energy" OR "wave energy")	1966
Hydropower technologies	…AND ("hydropower" OR "hydro power")	2695

2.2. *Keyword extraction by the TIM tool*

TIM can extract words from titles, abstracts and keyword fields from the documents in the dataset and normalises them with a clumping algorithm (e.g. grouping different spellings) benefiting from the use of the term frequency-inverse document frequency approach. The extracted terms are then prioritised using several available algorithms. In this paper "cleaned author keywords" (AK) and "relevant keywords" are tested. The first ranks keywords on the base of their frequency while the latter ranks terms using a computation that includes term frequency-inverse document frequency (TF-IDF).

2.2.1. *Search string design*

Each dataset is extracted by TIM using a two-part Boolean search string (Table 1). The first part (first line of Table 1) is designed to capture future emerging and other innovative or exploratory aspects. The second part of the string (connected by the operator AND) is targeting the specific energy sector: photovoltaics, wind power, ocean energy or hydropower (lines 2–5 of Table 1 are alternatively used).

The bibliometric searches were performed on 15th March 2019 by using the TIM 2019 version. The timeframe adopted for the search is 7 years (from 2012 to 2018), which is the value maximizing the recall according to the study (Moro et al., 2018). The PV sector provides, in output of this search, the largest collection of documents (32691), followed by wind power (14500 documents), hydropower (2695) and ocean energy (1966), as reported in the last column of Table 1. For all energy sector datasets, most of the retrieved documents are scientific publications from Scopus (including reviews, conference proceedings, book chapters and articles); patents represent only 10 % of the dataset.

2.2.2. *Keyword extraction process and algorithms*

The base of the extraction process is a custom made text mining dictionary developed by JRC and optimised for scientific text. The concepts in the dictionary may contain many versions of the same word (*Solar-Cell*, *solar cells*, *solar cell*…), so similar words are grouped into concepts according to the process known, in literature, as "clumping" (Porter & Zhang, 2015).

TIM extracts and ranks the keywords according two algorithms:

"*Cleaned author keywords*" (AK), a refinement of the author keywords where those are ranked by frequency (AK), alias the number of documents where the "author keyword" is found;

"Relevant Keywords", a modified version of the classic term frequency-inverse document frequency (TF-IDF), where there is a different position weight attributed to the terms coming from different parts of the document: weight 1 for the terms retrieved in the title, 0.5 for those retrieved in the abstract and 2 for those coming from the keyword fields. For simplicity this algorithm is identified, in this paper, as "TF-IDF".

2.3. *Indicators for the comparison between expert review and bibliometric software*

In this section we briefly recall some indicators useful to compare and benchmark findings from the TIM software against expert reviews. These indicators are also adopted to compare performances between TIM and VOSviewer. A full explanation is available in (Moro et al., 2018).

Four different *test sets* (Sanderson, 2010), describing four different technology sector (PV, wind, ocean, hydropower), are studied. For each sector there is a number of "*Marker Keywords*" (MK), defined by the experts, that the software should identify: 17 MKs for PV, 16 MKs for the wind sector, 9 for ocean and 8 for hydropower technologies.

$$r(FR) = n(MK \cap N) \tag{1}$$

r(FR) is the number n of MKs retrieved in the first ranked FR keywords. We consider significant a FR value of "300" on the base of our experience (Moro et al., 2018) and relevant literature (Porter & Zhang, 2015). Confirmation of the importance of this choice is discussed in Section 4.3.

In percent, Eq. (1) becomes:

$$Recallrate(FR) = (\frac{r(FR)}{MK} * 100) \tag{2}$$

This can be also defined recall rate at a fixed ranking or RR(FR).

Rank(MK_i) is the value of the rank of a specific marker keyword in a list of keywords produced by the software. To assess the overall performance of the software in high ranking all the MKs the function (3) could be ideally adopted:

$$SumRank(MK) = \sum_{i=1}^{n(MK)} Rank(MK_i) \tag{3}$$

However, this indicator can be calculated only if the software retrieves all the MKs. Considering that the software could not identify some MKs, this calculation is done only on the subset (MK-W) corresponding to 68 % of the MKs (68 % equals to one standard deviation interval, that in empirical sciences is commonly considered an acceptable uncertainty):

$$SumRank(68\%) = SumRank(MK - W) = \sum_{i=1}^{n(MK-W)} Rank(MK_i). \tag{4}$$

2.4. Indicators implementation

High ranking keywords in a small dataset is easier than obtaining the same performances for huge datasets. To fairly compare search performances from the indicators above for different datasets (e.g. different technology sectors or different software) we performed a normalisation of the rank values to the number of documents of the data set (N.documents) according to formula (5):

$$NormalisedRankValue = \left(\frac{RankValue}{N.\ documents} * 1000\right). \tag{5}$$

This normalisation is adopted for both the keyword frequency and TF-IDF algorithm outputs.

For the same reason of comparability, the number W used in Eq. (4) is the real number allowing to calculate the sum rank of exactly 68 % of the MKs and is generally not an integer number.

2.4.1. Rank uncertainty and Average rank value for author keywords

The "Author keyword" algorithm ranks up keywords on the base of their frequency (number of documents) in the dataset. It can happen that different keywords, with different ranking, present the same frequency; this happens more often for low frequencies. Table 2 reports an example for the wind power dataset. The wind power panel of experts had identified the technology characterised by the keyword "vsc hvdc" (voltage source converter at high voltage direct current). The TIM software ranked it in position #175. However, this would not be the correct rank value to be used in the formulas because there is a set of keywords with the same frequency value (34), ranking from #174 to #181. There is no reason to think that (for the same frequency) one keyword is more relevant than another since the bibliometric software ranks them according to the order in which it retrieves the documents (TIM) or alphabetically (VOSviewer). The value of the rank of the keyword "vsc hvdc" could range between 174 (minimum rank value) to 181 (Maximum rank value), with a rectangular distribution.

We therefore define *Average rank value* (ARV) the arithmetical mean:

$$Average\ rank\ value = \left(\frac{Maximum\ rank\ value\ +\ minimum\ rank\ value}{2}\right). \tag{6}$$

Associated to the ARV there is the uncertainty:

$$Uncertainty\ (Average\ rank\ value) = \left(\frac{Maximum\ rank\ value\ -\ minimum\ rank\ value}{2}\right). \tag{7}$$

Variables affecting this uncertainty are independent and casual; its propagation, in sums, should be considered quadratically; so

Table 2
Uncertainty in the ranking of the author keywords. Example for the term "vsc hvdc" in the wind energy technologies dataset.

Rank	Author Keywords	Frequency
(…)	(…)	(…)
173	district energy (DE)	35
174	probabilistic	34
175	vsc hvdc	34
176	price	34
177	blade element momentum (BEM)	34
178	decomposition	34
179	oscillations	34
180	investment	34
181	maintenance	34
182	wind integration	33
(…)	(…)	(…)

Table 3
Indicators used to test the TIM software performances in high ranking the marker keywords (MKs) identified by the experts. a) Photovoltaic technologies and corresponding MKs [in brackets] from the experts; b) Technology readiness level (TRL); c) Normalised average rank value (NARV) of the author keywords (AK) and d) related uncertainty; e) normalised rank of the "relevant keywords" extracted with the TIM term frequency-inverse document frequency (TF-IDF) algorithm.

	a) Technologies identified by the PV expert panel and related marker keywords [in square brackets]	b) TRL	c) AK NARV	d) AK NARV Uncertainty	e) TF-IDF Normalised Rank
1	[Kesterite] thin film solar cells or [CZTS]	3 – 4	7.48	0.05	2.05
2	[Perovskite] thin film solar cells	4 – 5	0.40	0.00	0.21
3	[Organic solar cells] or [OSC]	5 – 6	0.24	0.00	0.18
4	Dye-Sensitized Solar cells or [DSSC]	5 – 6	0.09	0.00	0.06
5	[Intermediate band] solar cells or [IBSC]	2	8.70	0.11	4.50
6	[Plasmonic] solar cells	3 – 4	1.96	0.03	5.35
7	Low-cost manufacturing processes, [roll to roll] or [flexible substrate]	n.a.	13.64	0.28	11.13
8	Innovative [multi junction] solar cells	2 – 3	4.97	0.05	3.21
9	Thermo-photovoltaics or [photovoltaic thermal]	1 – 2	2.58	0.02	2.48
10	Innovative [III-V] compounds based solar cells	1 – 2	4.76	0.02	5.51
11	Photoelectrocatalytic devices, [photocatalysis]	2	4.85	0.05	13.55
12	[Ferroelectric] photovoltaics	1 – 2	15.48	0.24	26.74
13	[Multiple exciton generation] solar cells or [MEG]	2	30.65	1.10	17.99
14	[Hot carrier] solar cells	1 – 2	28.80	0.72	22.42
15	[Transparent conduct]ing materials or [Carrier selective contacts]	2	4.13	0.03	4.25
16	Solar cells from [semiconductor foils]	3	n.a.	n.a.	n.a.
17	New pv materials via [computational design]	n.a.	17.63	0.54	15.97
18	Recall rate (300)		65 %	–	59 %
19	Recall rate (600)		82 %	–	82 %
20	NormalisedSumRank(68 %)		47.8	0.25	47.59

the uncertainty of the SumRank function of Eq. (4), becomes:

$$Uncertainty\ (SumRank\,(x\%)) = \sqrt{\sum_{i=1}^{n(MK-W)} Unc.\ (Rank\,(MK_i))^2}. \tag{8}$$

These considerations on the uncertainty suggest the following adjustment in formulas (1) and (2) to avoid information losses: if the fixed rank threshold (e.g. FR = 300) falls into an interval of keywords with the same ARV, the FR value to be used is the minimum rank value of the interval.

3. Results

This section reports the results of the calculations performed according to the formulas in Sections 2.3 and 2.4 on the keywords extracted by the TIM adopting the search strings in Table 1. The considered energy technology sectors are photovoltaics, wind power, ocean and tidal energy, hydropower. The same calculations are done, for ocean energy, also by the VOS Viewer software and an inter-comparison test is done with the TIM (see Section 3.5).

3.1. Photovoltaics

The PV data set analysed in this paper is different from that of the previous work (Moro et al., 2018), both because the timeframe is here updated and because the number of marker keywords has been reduced, following the classification adopted by the experts in the report (Moro et al., 2017). This allows to have a consistent classification for all the energy sectors.

Table 3 presents the performances of the TIM software in high ranking the marker keywords identified by a panel of PV experts. The lower the rank, the better the performance. Values are dimensionless. Data in columns c), d) and e) are normalised according to Eq. (5) for the 32691 documents of the PV data set. The Normalised Average Rank Value (NARV) of the author keywords (AK) and related uncertainty are calculated following Eqs. (6) and (7). The recall rate described by Eq. (2) is shown for the first 300 (line 18) and 600 (line 19) keywords. The normalised sum rank (line 20) is calculated according to Eq. (4) and reported on columns c) and e), while its uncertainty is propagated according to the (8).

3.2. Wind power

Table 4 presents the performances of the TIM software in high ranking the marker keywords (column a) identified by a panel of experts in wind power. The lower the (dimensionless) rank value, the better the performance. Calculations in c), d) and e) are normalised according to Eq. (5) for the 14500 documents of the wind power data set. The Normalised Average Rank Value (NARV) of the Author keywords and related uncertainty are calculated following Eqs. (6) and (7). The Recall rate described by Eq. (2) is calculated both for the first 300 (line 17) and 600 (line 18) keywords. The normalised sum rank (line 19) is calculated according to

Table 4

Indicators used to test the TIM software performances in high ranking the marker keywords (MKs) identified by the experts. a) Wind power technologies and corresponding MKs [in brackets] from the experts; b) Technology readiness level (TRL); c) Normalised average rank value (NARV) of the retrieved author keywords (AK) and d) related uncertainty; e) normalised rank of the "relevant keywords" extracted with the TIM term frequency-inverse document frequency (TF-IDF) algorithm.

	a) Technologies identified by the wind power expert panel and related marker keywords [in square brackets]	b) TRL	c) AK NARV	d) AK NARV Uncertainty	e) TF-IDF Rank Normalised
1	[Airborne] wind energy or [AWE] or [kites]	3 – 4	56.8	3.9	15.2
2	Multiple [drones] or [fly generator].	2 – 3	n.a.	n.a.	407.5
3	Offshore floating wind concepts: [TLP], [Spar buoy], [semi-submersible].	4 – 9	34.3	1.7	15.4
4	Floating hybrid energy platforms: Wave energy converter [WEC] or floating vertical axis wind turbine [FVAWT].	1 – 5	18.0	0.3	15.0
5	Passive [load reduction]: [bend] twist coupling.	2 – 7	64.8	4.1	89.4
6	Active control systems: moving tips, [circulation control], individual pitch control [IPC].	2 – 7	29.0	1.1	12.2
7	Wind induced energy harvesting from aeroelastic phenomena: [flutter], [galloping], [VIV].	2 – 4	87.4	7.9	97.9
8	Unconventional power transmission for wind turbine rotors [hydraulic transmission], air transmission system	1 – 7	195.3	34.2	241.9
9	Multi rotor wind turbines or [multi unit].	2 – 6	140.4	20.6	288.8
10	[Diffuser augmented] wind turbine [DAWT] or [shroud] or [ducted turbine]	5 – 6	107.6	12.2	48.6
11	Other small wind turbine technologies [Vertical Axis] wind turbines, [Savonius]	5 – 9	1.9	0.0	1.1
12	Alternative support structures for wind turbines: [self-rising], [composite materials]	2 – 8	20.0	0.6	52.9
13	Modular HVDC generator [vsc hvdc]	3	12.2	0.2	8.3
14	Innovative blade manufacturing techniques and materials: [3D], [fabric]-based	2 – 3	305.3	75.8	208.3
15	high fidelity multi scale integrated models for complex wind inflow [mesoscale]	3	56.8	3.9	79.5
16	Wind energy [databases] and [big data] analysis	3	20.0	0.6	122.4
17	Recall rate (300)		31 %	–	38 %
18	Recall rate (600)		44 %	–	38 %
19	NormalisedSumRank(68 %)		390.6	10.3	423.8

Table 5

Indicators used to test the TIM software performances in high ranking the marker keywords (MKs) identified by the experts. a) Ocean and tidal technologies and corresponding MKs [in brackets] from the experts; b) Technology readiness level (TRL); c) Normalised average rank value (NARV) of the keywords extracted by term frequency (TF) and d) related uncertainty; e) normalised rank of the "relevant keywords" extracted with the TIM term frequency-inverse document frequency (TF-IDF) algorithm.

	a) Technologies identified by the ocean and tidal energy panel of experts and related marker keywords [in square brackets]	b) TRL	c) AK NARV	d) AK NARV Uncertainty	e) TF-IDF Rank Normalised
1	Rotor innovations for tidal energy turbines: [yawing], [contra-rotating].	5 – 7	1076	597.7	681.1
2	[Floating] tidal concepts	5 – 8	121.3	24.7	12.7
3	Third generation tidal energy converters: [sails], [kites], [flapping]	3 – 7	358.3	119.8	82.9
4	Novel approach to first generation converters: [multi point absorbers], [multiple OWC], oscillating surge wave energy converter [OSWEC], [double chamber].	2 – 5	1076	597.7	32.6
5	Novel wave energy converters: [flexible membranes], [water level carpet], [dielectric electro active polymer]	1 – 3	192.3	45.8	40.7
6	Innovative tidal and wave energy power take off: [Direct drive (DD)], [Mechanical Motion Rectifier], [dielectric elastomer], [magnetic screw/gear], [inertial sea wave energy converter (ISWEC)]	3 – 6	15.8	1.5	15.3
7	Control systems: [latching control], [reactive control], [prediction algorithm]	2 – 7	121.3	24.7	45.8
8	Mooring and station keeping systems: [synthetic], [weathervaning]	2 – 7	n.a.	n.a.	6672
9	Materials and components: [elastomers], [light weight], [friction reduction], [energy storage]	3 – 5	37.4	2.8	221.8
10	Recall rate (300)		56 %	–	67 %
11	Recall rate (600)		67 %	–	78 %
12	NormalisedSumRank(68 %)		975.5	246.0	256.5

Eq. (4) on the data sets of columns c) and e); its uncertainty is propagated quadratically according to the (8).

3.3. Ocean and tidal energy

Column a) of Table 5 reports the technologies identified by an international panel of experts on ocean and tidal technologies (Magagna et al., 2018) and the related marker keywords (in square bracket) adopted to benchmark the ability of the TIM in high ranking these concepts.

Columns c), d) and e) of Table 5 present the performances of the TIM software in high ranking the marker keywords from experts. The lower the value, the better the result. Values in c), d) and e) are normalised according to Eq. (5) for the 1966 documents of the ocean and tidal data set. The Normalised Average Rank Value (NARV) of the author keywords and related uncertainty are calculated following Eqs. (6) and (7). The Recall rate described by Eq. (2) is shown for the first 300 (line 10) and 600 (line 11) keywords. The normalised sum rank (line 12) is calculated according to Eq. (4) on columns c) and e), its uncertainty is propagated as from formula (8).

3.4. Hydropower

Columns c), d) and e) of Table 5 present the performances of the TIM software in high ranking the marker keywords defined from hydropower experts. The lower the value, the better the result. Calculations in c), d) and e) are normalised according to Eq. (5) for the 2720 documents of the hydropower dataset. The Normalised average rank value (NARV) of the author keywords and related uncertainty are calculated following Eqs. (6) and (7). The recall rate described by Eq. (2) is calculated both for the first 300 (line 9) and 600 (line 10) keywords. The normalised sum rank (line 11) is calculated according to Eq. (4), its uncertainty is propagated quadratically as from formula (8).

3.5. VOSviewer – TIM comparison

To test the validity of the calculations above and generalise the behaviour found for the TIM to other software, an inter-comparison is performed by using VOSviewer software on the same data set and comparing the TIM Vs VOSviewer indicators.

VOSviewer is a software tool for bibliometric analysis whose main functionality is to build bibliometric networks. Like TIM, it also offers text mining functionality.

The two software systems are set up for the same search string (Table 1) and the same reference corpus from Scopus. The two data sets are slightly different because data extraction was performed, for the VOSviewer, about one month later (17th April 2019) than for the TIM (15th March 2019).

The version of the VOSviewer we used relies indirectly on external data sets. Due to limitations in the number of meta-documents downloadable from Scopus (max 2000), this comparison was possible only for the ocean and tidal sector, where the number of documents retrieved by the defined search string is less than 2000. The technologies and marker keywords used as benchmark are the same defined by an international panel of experts (Magagna et al., 2018) and reported in Table 5.

Table 7 reports the quantitative values of the indicators described in Sections 2.3 and 2.4, calculated for the ocean energy technologies for both the TIM and the VOSviewer software. SumRank (68 %), which perform better when lower, is statistically consistent considering the uncertainties. The indicator Recallrate(600) has the same value, while the small difference in Recallrate (300) can be justified by the considerations reported in Section 4.5. Therefore, the two software systems perform equally versus the ocean expert analysis.

4. Discussion

The comparison exercise presented in this paper aims at providing analysts with practical insights on how to better use text mining software to identify emerging technologies.

4.1. On the expert consultations and TRL assessment

The indicators we defined (Moro et al., 2018) and adopted in this paper consider as reference the output of four expert review exercises. This assumption comes from the political relevance of the organised exercises; high level EU scientists were invited to our workshops and their outputs informed the officers designing the EU research policies. However, from a scientific point of view, we should consider that different expert groups with different background could assess differently the same technology. Particularly if the experts provide a quantitative assessment to something very complex, not well defined, or with a literature base not strong enough.

This is particularly true for the Technology Readiness Level (TRL). The TRL indicator is a quite new concept in the energy field; it can sometimes be not univocally defined or quantified, especially when referring to a technology in an early stage of development. Consequently, the TRL values appearing in this paper, even if agreed between the invited experts, have an intrinsic component of subjectivity and they are open to discussion.

Considering this, it is not possible to compare TRL of different technology sectors, assessed by different expert groups according to different hypotheses. However, the TRL values reported in this paper can make sense in relative terms and can be used for technology maturity comparisons inside a specific energy sector, since each panel of experts agreed on the assessed TRL values.

In order to compare TRL values from different energy sectors it is necessary to perform a “calibration” of the TRL scales used by the different expert groups. It is our intention to exploit further these aspects in our next paper.

4.2. Costs

Bibliometric software is here used to identify emerging technologies and its performances compared with expert review. The comparative analysis presented in this paper showed that the software presents some information losses when applied to emerging technologies belonging to the considered energy sectors. However it presents much lower costs.

The expert reviews we performed for the four energy sectors analysed required about 6–8 person months of work each one. In this time, the analysts performed a preliminary literature study, identified and invited the experts, planned and organised the meetings. The whole process took about 4 months and the logistic costs for the workshop organisation were about 20–30 kEuro each.

Conversely, for the bibliometric software method, to design a search string and screen the first 300 top ranked keywords for a technology sector an analyst needs less than two weeks of work (Moro et al., 2018).

So it is possible to say that, in cases when the software provides an acceptable loss of information (i.e. photovoltaics, ocean and tidal), the software retrieves most of the technologies identified by experts at a cost and in a time that are in the order of 10 % of an expert elicitation exercise (Moro et al., 2018).

4.3. The importance of analysing the top 300 keywords

In general, keywords produced by a text mining software are not automatically coinciding with the name of a technology; discriminating between keywords that can refer to a technology (marker keywords) and false positives requires a semantic/knowledge-based work. An analyst usually performs this post-processing of the keywords extracted by the software. The higher the number of keywords analysed by the human operator, the higher the probability to identify interesting technologies. However, the lists of keywords produced by the software can be very long (in this paper from 3300 to 130,000) and the working time of the analysts has a cost. On the basis of our experience and from the literature (Porter & Zhang, 2015) we preliminarily identified (Moro et al., 2018) a trade-off between a satisfactory number of technologies retrieved and a reasonable number of keywords analysed (working time of the analyst) in the 300 top high ranked keywords extracted by the software. This corresponds to one week of work of an analyst. This assumption was at the base of the definition of indicators cognate the recall rate at a fixed ranking described in Eq. (2). The efficacy of this “magic number” of 300 is confirmed by the results presented in Section 3.

In the sector of photovoltaics (Table 3), an analyst studying the first 300 ranked keywords from TIM would have a recall rate RR (300) varying between 65 % (considering the author keywords frequency algorithm) and 59 % (using the TF-IDF algorithm). By doubling the efforts (corresponding to the indicator RR(600)) the analyst could retrieve 82 % (for both the algorithms) of the MKs. This means that an increase of 100 % of the work would bring to the analyst a benefit of only 27%–40%.

For the wind energy dataset (Table 4) the RR(300) varies from 31 % (simple frequency) to 38 % (TF-IDF), while the RR(600) is 44 % – 38 % (benefit from +40 % to 0 %).

For the ocean energy technologies (Table 5) RR(300) of TIM is 56%–67% while the RR(600) is 67%–78% (benefit of 20 % – 17 %).

Table 6
Indicators used to test the TIM software performances in high ranking the marker keywords (MKs) identified by the experts. a) Hydropower technologies and corresponding MKs [in brackets] from the experts; b) Technology readiness level (TRL); c) Normalised average rank value (NARV) of the keywords extracted by term frequency (TF) and d) related uncertainty; e) normalised rank of the “relevant keywords” extracted with the TIM term frequency-inverse document frequency (TF-IDF) algorithm.

	a) Technologies identified by the hydropower panel of experts and related marker keywords [in square brackets]	b) TRL	c) AK NARV	d) AK NARV Uncertainty	e) TF-IDF Rank Normalised
1	Innovative flow control techniques to support a wide-range opearation of hydraulic turbines or [flow regulation]	2 – 3	399.8	123.0	255.5
2	Flow control techniques to suppress pump-turbine instabilities or [air injection]	3	1155	631.8	1584
3	[Numerical model]ling and controlling hydraulic turbines and pump-turbines operation by a digital avatar	3 – 4	72.4	7.4	526.1
4	Hydro generators with current controlled [rotor segments]	4 – 5	n.a.	n.a.	0.0
5	Integrating fast energy storage and hydropower or pumped-storage or [area control error (ACE)]	4	158.1	26.5	326.5
6	Low- and ultra-low head hydropower for onshore and [marine] (or [tidal]) applications	4 – 5	115.6	15.6	172.4
7	Novel water wheels: hydrostatic pressure machine and turbine [water wheel]	3 – 4	397.8	121.0	65.1
8	Hydropower development in conduit systems or [water supply] or [irrigation]	4	27.8	1.7	30.1
9	Recall rate (300)		38 %	–	25 %
10	Recall rate (600)		50 %	–	38 %
11	NormalisedSumRank(68 %)		947.6	149.3	1081

Table 7
Comparison between the term frequency (TF) of the author keywords retrieved by the two software TIM and VOSviewer; the comparison is performed for the same search string and data set describing the ocean and tidal energy sector.

Indicators	TF TIM	TF VOS
Number of documents	1762	1902
Number of keywords retrieved	3290	4570
Recallrate(300)	56 %	44 %
Recallrate(600)	56 %	56 %
SumRank(68 %)	1088	2272
Uncertainty	275	1049

In the last case of the hydropower sector (Table 6) the RR(300) is 38 % – 25 % while RR(600) is 50 % – 38 % (benefit: 33%–50%).

Concluding, starting from 300 as baseline, an analyst doubling the efforts to study +100 % of keywords would have an average benefit of 28 %. This result confirms the suitability of the choice of 300 as number of keywords in the output of a text mining software to be analysed, for identifying emerging technologies.

4.4. Better performances for technologies with an established specific jargon

Another useful indicator is SumRank(68 %), defined in Section 2.3. This sum rank indicator helps to benchmark the overall performances of the text mining tool for various datasets or software setting. The lower the sum rank value, the better the software performance.

Amongst the renewable energy datasets extracted by TIM and analysed here, photovoltaics has the best sum rank (about 48 both for the normal frequency and the TF-IDF algorithms), while the other technology sectors present much worse results: from 390 (AK) to 420 (TF-TDF) for wind, 1000 (AK) or 250 (TF-TDF) for ocean and about 1000 for both the algorithms in hydropower.

This can be justified by the fact that PV technologies have a quite established and recognised jargon, often with univocal keywords (no synonyms) that are coinciding with the name of the technology; in example: "kesterite", "perovskite", "plasmonic solar cells", "organic solar cells", "dye-sensitised solar cells".

Alternatively, it is more difficult to identify technologies that cannot be correctly framed by a specific jargon. In example: “load reduction”, “hydraulic transmission”, “composite materials” for wind power; "flow regulation" and “air-injection” for hydropower. This behaviour can be observed also by analysing the normalised rank values of keywords across different datasets, with “hot carrier solar cells” performing worse than “vertical axis wind turbines”.

4.5. Author keyword frequency Vs TF-IDF algorithm: suitability to different technology sectors

Two different methods are used here to rank keywords retrieved by the TIM. A simple frequency count of the author keywords (clumped as described in Section 1.2) and the TIM improved version of the TF-IDF algorithm (Section 2.2.2). Performance comparison between the two methods is not showing a general prevalence of one method versus the other for the four energy sector considered. This can be experienced by comparing results (for a single technology or for all the sector) in column c) with those in column e) for Tables 3–6. For some technologies and sectors the simple author keyword performs better while for others perform better the TF-IDF; in other cases they provide the same result.

An explanation of this apparently ambiguous behaviour can be found in the intimate nature of these two algorithms. The first one strongly relies on the ability of the authors in using keywords characterising well and univocally the technology they are focusing on in their papers (most of the documents come from Scopus). The TF-IDF algorithm is not relying only on the author keywords but also considers those present in the title and in the abstract, high ranking those considered more "rare" and downranking those considered "common". In this study we experienced three different behaviours.

For technologies univocally characterised by specific keywords, like in the PV sector, the two methods are equivalent (sum rank and recall rate presented in Table 3 are strongly overlapping), because the authors adopt keywords considered rare enough also by the TF-IDF algorithm.

For technology sectors where emerging technologies are not characterised by exclusive keywords but by more common terms, like "air injection" in hydropower (row 2 in Table 6), the frequency count of the author keywords performs better than the TF-IDF (see the good performances of the hydropower sector in lines 9, 10 and 11 of Table 6). This is also the case of innovations coming from other sectors, in a cross-fertilisation process. This is an important case, since "the convergence of scientific knowledge is always one of the major propellers of emerging technologies" (Zhou, Dong, Kong, & Liu, 2019). The TF-IDF performs inefficiently, in this case, because consider the keywords (that are new in the sector of interest but mainstream in other sectors) as "trivial", ranking them down.

A third behaviour can be found when technologies have very specific but not univocal terms, like in the ocean and tidal sector (Table 5). Here the TF-IDF based performs much better than the author keyword frequency because the authors, choosing a keyword amongst various possible synonyms, are *de facto* reducing its potential frequency and rank, while the TF-IDF can consider also the synonyms present e.g. in the abstract. As an example, the technology "Novel approach to first generation energy converters: [multi point absorbers], [multiple OWC], oscillating surge wave energy converter [OSWEC], [double chamber]" (row 4 in Table 5) presents

TF-IDF performances that are orders of magnitude better than in the simple author keyword frequency (TF-IDF ranks 32.6 Vs 1076 of the normalised author keyword frequency rank).

The suitability of the modified TF-IDF algorithm to the ocean energy sector can also justify the different average values of the indicators (although statistically consistent) for TIM and VOSviewer (Section 3.5). This is probably because TIM makes a preliminary use of the TF-IDF also for the creation of the author keyword dictionary (Section 1.2). As a result, VOSviewer considers the term "direct drive" (frequency = 3) separately from "direct-drive" (also with frequency = 3) and other synonyms, while TIM groups all them under the same term with frequency = 16.

5. Conclusions

Bibliometric software with text mining features can be used by technology analysts to identify emerging and promising technologies in specific sectors, using quantitative analysis.

This paper tests the use of Tools for Innovation Monitoring (TIM), a bibliometric software developed by the Joint Research Centre of the European Commission. TIM is used to identify emerging technologies in four energy technology sectors (photovoltaics, wind power, ocean and tidal energy, hydropower).

Results from TIM are compared with those produced, for the same sectors, by four panels of experts adopting qualitative cognitive analysis. The comparison between the quantitative semi-automated outputs of the software and the qualitative expert reviews is performed by using specially tailored indicators.

An inter-comparison is also performed between the TIM and the VOSViewer from Leiden University, showing outputs in line.

Different TIM performances can be experienced for the different technology sectors and the two different algorithms used by the software: the author keyword frequency (AK) and the term frequency-inverse document frequency (TF-IDF).

The TIM text mining software ranks up (in the first 300 positions of the lists it extracts) the keywords characterising the technologies identified by the experts with efficacy of 65 %–59 % (respectively for AK and TF-IDF) for the photovoltaics, 56 %–67 % for the ocean and tidal energy sector, 31 %–38 % for wind power and 38 %–25 % for hydropower.

An analyst studying the first 300 keywords in output of the software seems the optimum trade-off between retrieved technologies and effort, because by doubling (+100 %) the number of the keywords analysed the number of matches rises, in average, only by 28 %.

Technologies retrieved much more easily are those characterised by a well-established univocal jargon (like most of those in the PV sector) while emerging concepts with a not clear definition or characterised by common use terms are currently difficult to identify.

Comparing the methods, the TF-IDF based algorithm seems more effective to retrieve emerging technologies characterised by specific but not univocal (with synonyms) terms (like in the ocean and tidal sector). However, the simple frequency count of the author keywords is more effective in identifying innovation coming from other technology sectors (with apparently common terms).

The JRC is further developing the TIM Tools for Innovation Monitoring and expanding the analysis to other emerging energy technologies.[1]

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors gratefully acknowledge the contribution of Ivano Azzini, Sotirios Fragkiskos, Sergio Perani and David Shaw.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.futures.2020.102511.

References

Amanatidou, E., Butter, M., Carabias, V., Könnölä, T., Leis, M., Saritas, O., ... van Rij, V. (2012). On concepts and methods in horizon scanning: Lessons from initiating policy dialogues on emerging issues. *Science & Public Policy*. https://doi.org/10.1093/scipol/scs017.

Centre for Science and Technology Studies, L. U (2019). *Welcome to VOSviewer.* Retrieved fromhttp://www.vosviewer.com/.

Ciano, M. P., Pozzi, R., Rossi, T., & Strozzi, F. (2019). How IJPR has addressed “lean”: A literature review using bibliometric tools. *International Journal of Production Research*. https://doi.org/10.1080/00207543.2019.1566667.

De Rose, A., Buna, M., Strazza, C., Olivieri, N., Stevens, T., Leen, P., ... Daniel, T.-J. (2017). *Technology readiness level : Guidance principles for renewable energy technologies. Annexes.* Luxembourghttps://doi.org/10.2777/863818.

Dillon, M. (1983). Introduction to modern information retrieval. *Information Processing & Management, 19*(6), 402–403. https://doi.org/10.1016/0306-4573(83)90062-6.

Elsevier (2017). *Scopus*.

European Commission (2017a). *Building a European data economy (No. COM(2017) 9).* Retrieved fromhttps://ec.europa.eu/taxation_customs/sites/taxation/files/1_en_act_part1_v10_en.pdf.

European Commission (2017b). *CORDIS: Community research and development information source.* Retrieved January 27, 2018, fromhttp://cordis.europa.eu/.

[1] http://timanalytics.eu/.

European Patent Office (2017). *PATSTAT: Worldwide patent statistical database.* Retrieved January 27, 2018, fromhttps://www.epo.org/searching-for-patents/business/patstat.html#tab-1.

Huang, L., Zhang, Y., Guo, Y., Zhu, D., & Porter, A. L. (2012). Four dimensional Science and Technology planning: A new approach based on bibliometrics and technology roadmapping. *Technological Forecasting and Social Change*. https://doi.org/10.1016/j.techfore.2012.09.010.

Joanny, G., Agocs, A., Fragkiskos, S., Kasfikis, N., Le Goff, J.-M., & E. O (2015). Monitoring of technological development - detection of events in technology landscapes through scientometric network analysis. *Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference*.

Leydesdorff, L. (2015). Can technology life-cycles be indicated by diversity in patent classifications? The crucial role of variety. *Scientometrics*. https://doi.org/10.1007/s11192-015-1639-x.

Magagna, D., Margheritini, L., Alessi, A., Bannon, E., Boelman, E., Bould, D., ... Yeats, B. (2018). *Workshop on Identification of Future Emerging Technologies in the Ocean Energy Sector-27th. European Commission*https://doi.org/10.2760/23207.

Moro, A., Boelman, E., Joanny, G., & Garcia, J. L. (2018). A bibliometric-based technique to identify emerging photovoltaic technologies in a comparative assessment with expert review. *Renewable Energy, 123*. https://doi.org/10.1016/j.renene.2018.02.016.

Moro, A., Aycart, J., Bardizza, G., Bielewsky, M., Lopez-Garcia, J., Taylor, N., ... Garcia, L. J. (2017). *First Workshop on Identification of Future Emerging Technologies for Low Carbon Energy Supply*https://doi.org/10.2760/849373.

Porter, A. L., & Zhang, Y. (2015). Tech mining of science & technology information resources for future-oriented technology analyses. In J. C, & G. Glenn (Eds.). *Futures research methodology version 3.1* The Millennium Project. Retrieved from http//themp.org/.

Salton, G. A., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0.

Sanderson, M. (2010). *Test collection based evaluation of information retrieval systems. Foundations and trends® in information retrieval.* https://doi.org/10.1561/1500000009.

Zhang, J., Yan, Y., & Guan, J. (2014). Scientific relatedness in solar energy: A comparative study between the USA and China. *Scientometrics*. https://doi.org/10.1007/s11192-014-1487-0.

Zhou, Y., Dong, F., Kong, D., & Liu, Y. (2019). Unfolding the convergence process of scientific knowledge for the early identification of emerging technologies. *Technological Forecasting and Social Change, 144*(March), 205–220. https://doi.org/10.1016/j.techfore.2019.03.014.