PROBLEMS: EASY TO SAY BUT DIFFICULT TO WRITE

Giordano, Vito (1,3);
Consoloni, Marco (1,3);
Chiarello, Filippo (1,3);
Fantoni, Gualtiero (2,3)

1: School of Engineering, Department of Energy, Systems, Land and Construction Engineering, University of Pisa, Italy;
2: School of Engineering, Department of Civil and Industrial Engineering, University of Pisa, Italy;;
3: B4DS - Business Engineering for Data Science lab, University of Pisa, Italy

ABSTRACT

Patents are an invaluable source of data that can be beneficial for Engineering Design (ED). Patenting is one of the main means for disclosing the inventive process. For this reason, the description of the problem solved should also be included in any patents.

The ED literature lacks a proper definition of a problem, resulting in a fragmented scenario. Prior studies have employed Text Mining (TM) to extract problems from patents. We argue that TM can assist ED researchers in understanding how problems are articulated in text. Based on the literature, we propose two hypotheses: (1) problem-related text exhibits a negative sentiment polarity compared to other sections of patents; (2) problem-related keywords identified in the literature are predominantly used to describe problems rather than other aspects.

We analyse Japanese patents to validate our hypotheses, since they explicit Problem and Solution in the abstract. Finally, we compare our results with a set of problem-related sentences extracted from USPTO patents.

Our study reveals a higher positive sentiment in problem-related sentences compared to solution-related ones and highlights the inadequacy of using problem-related keywords alone to differentiate between the two.

Keywords: Patent Analysis, Text mining, Knowledge management, Semantic data processing, Conceptual design

Contact:
Giordano, Vito
University of Pisa
Italy
vito.giordano@phd.unipi.it

Cite this article: Giordano, V., Consoloni, M., Chiarello, F., Fantoni, G. (2023) 'Problems: Easy to Say But Difficult to Write', in *Proceedings of the International Conference on Engineering Design (ICED23),* Bordeaux, France, 24-28 July 2023. DOI:10.1017/pds.2023.295

1 INTRODUCTION

Patents are an invaluable source of data that can be beneficial for engineering design, solving complex problems, and evaluating potential technological advancement. Manually studying patents by experts is a laborious process, and it has many issues. Bonaccorsi et al. (2020) demonstrate as this process is time consuming and bias prone. Moreover, the increasing number of patenting activity turns the manually analysis into a probative task. In this scenario, the automatic patent analysis becomes a fundamental support for companies, researchers, and policy makers. (Guarino et al., 2022).

Patenting is a major way of safeguarding an invention and is the most significant output of disclosing the inventive process. The inventive process is a complex problem-solving task, which starts with the framing of the problems to be solved. World Intellectual Property Organization (WIPO) (Organization, 2004) states that a patent should include the description of a technical problem solved by the invention to be compliant with the patentability criteria. However, a proper definition of a problem is not provided by international guidelines, and various definitions exist in engineering design literature, leading to a fragmented scenario where it is difficult to provide a formal definition of an engineering problem. As demonstrated by Giordano et al. (2022), the massive analysis of problems enables designers, inventors and researchers to (1) map the technical prior art, (2) generate new ideas for the conceptual design phase, and (3) study the technological evolution. Previous studies in literature rely on text mining techniques for identifying problems in patent texts. In this paper, we argue that text mining techniques can help designers, inventors, and researchers to provide a clear definition of what a problem is. For this reason, the main research question we address is "*Can text mining help engineering design community in defining how a problem is expressed in patent text?*". To provide an answer at this research question, we formulate two hypotheses: (**HP1**) the text related to problems in patents has a negative sentiment respect to other patent parts; (**HP2**) the problem-related keywords from background works are mostly used in patents for describing a problem-related sentences respect to other parts.

We use text mining techniques to validate or reject our hypothesis. The analysis is performed on Japanese patents since they explicitly subdive the abstract in two parts: "Problem to be solved", and "Solution to the problem". Our results show that problem-related sentences have higher positive sentiment compared to solution-related ones. Moreover, we discover that relying solely on problem-related keywords from background works may result in poor differentiation between problem and solution-related sentences.

2 BACKGROUND PAPERS ON TECHNICAL PROBLEM

2.1 What problem is in engineering design

A different perspective on problems can be found in TRIZ theory, where problem is suggested to be rephrased referring to the two major principles of TRIZ: contradiction and ideality. As for the contradiction, Zanni-Merk et al. (2009) define a problem as a contradiction between two or more system parameters. When the first parameter is improved, the second one reduces. For a correct formulation of a contradiction a third element is necessary: an action parameter (Guarino et al., 2022). For those readers more familiar with Axiomatic design we can say that the Action parameter in TRIZ is behave like the Design Parameter in Axiomatic (Suh, 1998). When a Design parameter is connected with two requirements the design is coupled and a contradiction exists. Actually, if a design action on the specific design parameter improves the satisfaction of one requirement while reducing the satisfaction of the other one. Concerning the ideality, a problem can be defined as the loss of ideality of a device in terms of reduced performance, emergence of undesired side effects and/or excessive consumption of resources (Becattini et al., 2011). From the description above, the concepts of effectiveness and efficiency stand behind the formulation of a negative effect. While effectiveness refers to the intended or expected result, the efficiency introduces the amount of resources needed to achieve this result. Given that the resources are finite, increasing efficiency can be considered an additional normative assumption (and also a broad empirical generalization from the history of technology), as it is reflected in the notion of ideality. The variable of the time, necessary to perform a certain function, is itself an element of efficiency and the higher time the lower the perceived quality of the product, therefore it must be considered in the analysis, but it is usually converted into money and thus included in the definition of effectiveness. Those that are defined problems to be solved in

https://doi.org/10.1017/pds.2023.295 Published online by Cambridge University Press

patents can be defined alternatively as “drawback”, “disadvantage” and “failure” in other documents (tenders, specifications, FMECA Sheets, and so on). In the engineering design lexicon these terms are often used as synonyms. Even if they are not perfectly overlapping, a drawback is defined as “undesirable feature” or “hindrance” (Chiarello et al., 2017), while a failure has a more complex definition in both dictionaries and standard technical lexicons. In particular, part of the definition of failure refers to: (1) a non-functioning behaviour; (2) a reduced performance (output) with respect to something due, required, or expected; (3) an excessive use of necessary resources with respect to specifications or expectations. In other words, the notion of failure is implicitly normative: a failure can be defined with respect to something that is expected from an engineering perspective (Cascini et al, 2013). Therefore, a clear distinction between problems, failures and drawbacks is hard to be made both in theory and in practice. Moreover, when they are formulated in natural language terms the boundaries among them are even more blurred. In the next section, we provide an overview of the NLP approaches used for identifying problem concepts in patents. Summing up, we propose that all useful definitions of problems, disadvantages, drawbacks and failures can be collapsed into three categories as follows: (1) the wanted output (desired effect of the system) obtained is not enough; (2) too many unwanted outputs (undesired effects of the system) are produced; (3) too many resources (time included) are needed to achieve a desired effect (it implies less efficiency). Therefore, we can expect that problems in patents explicit at least one of the three conditions (1-3) and be formulated according to one of the previous forms.

2.2 Extracting problems with text mining

Jeong and Kim (2014) classify the NLP systems to automatically extract problems from patents in: keyword-based, grammar-based, and machine learning methods. Keyword-based methods are the simplest approach to identify problems. They consist of lists of known keywords to map mentions of terms within texts. Tiwana and Horowitz (2009) extract problem-related sentences using a list of words, such as "need", "demand", and "objection". Similarly, Liang and Tan (2007) develop a list of terms related to problems (e.g. "effectiveness", "efficiency", "goal", and "important"). Grammar-based methods rely on regular expressions and morphosyntactic information to create knowledge-based systems able to describe problem concepts. Among these methods, the most famous is the Subject-Action-Object (SAO) approach, where, the problem statement is represented by the action-object (AO) item, and the solution is the subject (S) (Moehrle et al. 2005). Kim and Yoon (2021) map the problems outlined in the prior art of patents based on SAO approach. Souili et al. (2015) mix keyword and grammar-based methods to extract problems defining a list of keywords to use in conjunction with morphosyntactic rules. Similarly, Jeong and Kim (2014) list several phrases (i.e., words or multiwords) combined with grammar-based system for a more precise extraction of problems. Machine learning models use the textual representation in vectors of real numbers for recognizing problems in textual data. Giordano et al. (2022) apply machine learning models on patents for identifying problems, solutions and advantages using Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). Chikkamath et al. (2021) affirm that problems in patents have a negative sentiment and develop a machine learning-based sentiment analysis for delineating negative, neutral, and positive sentences. Chiarello et al. (2017) propose a system for performing a technical sentiment analysis, using an enormous gazette composed of more than 6,000 positive items and more than 14,000 negative phrases. Keyword-based method is a fast and accurate approach, but the creation of the keywords list usually requires a huge manual effort and field-specific knowledge (Puccetti et al., 2023). Differently, grammar-based approaches allow us to recognize the problems in a more abstract way than the first method, but with a lower precision ((Puccetti et al., 2023). However, most grammatical systems in literature rely on Part-of-Speech and dependency parsing rules, which have a too specific level of detail to determine the constituents of a problem. Machine learning is the state-of-the-art in many fields of NLP and reaches high performance both in terms of precision and recall, but it requires a large amount of effort to collect and annotate large datasets. Moreover, machine learning is a black box system which does not permit to understand which the atoms of problems are (Puccetti et al., 2023). For these reasons, in our paper, we rely on sentiment analysis and keywords-based approach. Table 1 lists the keywords used by background works.

https://doi.org/10.1017/pds.2023.295 Published online by Cambridge University Press

3 DATA AND METHOD

Based on the literature review of Section 2, we formulated two hypotheses (HPs) of how a problem is expressed in natural language: (**HP1**) the text related to problems in patents has a negative sentiment respect to other patent parts; (**HP2**) the problem-related keywords from background works are mostly used in patents for describing a problem-related sentences respect to other parts.

Table 1. Keywords to express problem in patents listed by background works.

References	Keywords
Jeong and Kim (2014)	problem, drawback, matter, trouble, defect, weak, flaw, fault, shortcoming, demerit, fail, wrong, error, harm, complain, disadvantage, bad, too, low, loss, slow, complex, complicate, frustrate, difficult, hard to, restrict, limit, disable, uneasy, uneasiness, unpleasant, inconvenient, uncomfortable, discomfort, usability, throughput, cost, expensive, pros and cons, object, need, desire, require, demand, awkward, danger, pervert, fussy, fastidious, refractor, stress, distress, hurt, painful, pain, suffer, anxiety, strain, burden, tens, injury, stuck, undermine, ruin
Liang and Tan (2007)	difficult, effectiveness, efficiency, goal, important, improved, increase, issue, limit, needed, overhead, performance, problem, reduced, resolve, shorten, simplify, suffer, superior, weakness
Souili et al. (2015)	blemish, break, bug, cause, crack, damage, defect, deform, degrade, deprive, destroy, deteriorate, disadvantage, disparate, hamper, harm, hinder, impair, smash, spoil, stain, trouble, weaken, fail, worsen, complication, deficiency, deformity, degradation, deprivation, destruction, deterioration, detriment, difficulty, drawback, drawbacks, failure, flaw, hampers, impairing, imperfection, instability, limitation, prejudice, problem, spoiling, weakness, however, if
Tiwana and Horowitz (2009)	need, advantage, solution, invention, demand, want, motive, desire, complaint, objection, problem, benefit, reward, answer, result, resolution, solvent, innovation, excogitation, conception, design, creation

In order to validate the hypothesis above, Japanese patent abstracts were chosen since they have a mandatory structure which is composed of two sections: “Problem to be solved” and “Solution to the problem” (Kim and Choi, 2007). The validation process was performed using NLP techniques. We selected 100,000 patent abstracts, where the “Problem to be solved” textual part was extracted. Since to validate the HPs, we needed a textual part for comparing the results, we also selected “Solution to the problem” parts from the same abstracts of Japanese patents. To validate HP1, we analysed the sentiment analysis of these sentences, and we compared the polarity value of problems with those of solutions. Moreover, the validation of HP2 was performed, building a keywords-based system developed based on the phrases suggested by previous background works (listed in Table 1). In this case, we compared the distribution of these keywords in “Problem to be solved” and “Solution to the problem” parts. Furthermore, in order to assess the possibility of generalizing our approach, we compared the insights resulted from Japanese patents with those obtained from the patents of the United States Patent and Trademark Office (USPTO) to be sure that our results are office independent. In this section, we firstly describe the data used (in Section 3.1), then we detailed describe the main methodological steps (in Section 3.2).

3.1 Japanese and USPTO Patents set

Japanese patent abstracts have a mandatory structure which is composed of two sections: “Problem to be solved” and “Solution to the problem” (Kim and Choi, 2007). We selected 100,000 patent abstracts from Japanese database and subdivided “Problem to be solved” part from “Solution to the problem” one. USPTO makes publicly available the patent full text for advancing the research in innovation. Based on the five IP offices (IP5)[1], the part of patents that is called “description” also contains the “Technical Problem” and “Solutions to the Problem”. Chikkamath et al. (2021) use the USPTO data for creating a dataset contains a collection of patent paragraphs (60,000 patents) referred to (A) *Technical Problem*; (B) *Solutions to the Problem*; and (C) *Advantageous Effects of the Invention*. We relied on the dataset of Chikkamath et al. (2021) to compare the Japanese problems with those of USPTO in order to be sure that our insights are not influenced by the office where a patent is submitted. For the rest of paper, we call the textual part of the “Problem to be solved” and “Technical

[1] https://www.fiveipoffices.org/activities/globaldossier/CAF. Accessed on Nov. 30, 2022.

https://doi.org/10.1017/pds.2023.295 Published online by Cambridge University Press

Problem" as "**problem**", and the "Solutions to the Problem" as "**solution**", to simplify the reading process.

3.2 Methodology for analysing problems

We adopted two main methods for analysing problem-related sentences. First, we performed a sentiment analysis of texts related to problem and solution for validating HP1. Second, we developed a NLP system for recognizing problem-related keywords into patents and validating the HP2. Melluso et al. (2023) use a similar approach to study the cognitive aspects of affordance and biases. In literature various approaches of sentiment analysis exist, based on keywords or machine learning system. Among these works, we relied on the system developed by Chiarello et al. (2017), which uses a lexicon of positive and negative phrases extracted from patent documents. The system is suitable for analysing technical documents such as patents, since it was built on patent documents, which have complex syntactic structure and use a technical-juridical jargon. The system of Chiarello et al. (2017) is composed of more than 6,000 positive and 14,000 negative phrases. Moreover, to obtain a final score of textual polarity, we used three other lexicons: negators, amplifiers, and deamplifiers[2]. The first is a list of terms reversing the intent of a positive or negative word; the amplifiers lexicon increases the intensity of a positive or negative word; the deamplifiers lexicon decreases the intensity of a positive or negative word. Finally, the polarity score was calculated using the method developed by Hu and Liu (2004), which combine the lexicons of positive and negative phrases, with those of negators, amplifiers, and deamplifiers. A high polarity score indicates a high level of positive sentiment and vice versa. To analyse if text related to problems in patents has a negative sentiment respect to solution, we compared the distribution polarity of problems with the one of solutions.

Table 2. Keywords we added for the NLP system.

Category	Keywords
Wanted output obtained is not enough	avoid, without, anxiety, incontinence, incomplete
Too many unwanted outputs	less, reduce, undermine, ugly, dent, chip, cheap, unfinish, miss, poor, overlapping, incorrect, discoloration, gap, improper, prevent, scratch, unsafe, misshapen, unnecessary, unstable, fragile
Too many resources are needed	decay, decrease, depletion, diminish, diminution, discharge, dissipate, effort, energy, expenditure, wear, denial, discard, dismiss, garbage, junk, litter, material, money, refuse, resource, rubbish, scrap, slump, swill, trash, delay, downtime, duration, extent, halt, hold, lag, late, long, on hold, period, rest, retard, span, stop, suspension, term, time, wait
General	albeit, although, consumption, even, excess, in spite of, nevertheless, though, too much, waste, theory, ideality

Concerning the HP2, we built a keywords-based system creating a lexicon composed of phrases suggested by Jeong and Kim (2014), Liang and Tan (2007), Souili et al. (2015) and Tiwana and Horowitz (2009). Moreover, we added some other elements missed by previous works, which are connected to the definition of problem based on the literature review of Section 2.1. Moreover, we enriched the lexicons with words included in Table 2 and expressing the following concepts: (1) the wanted output obtained is not enough; (2) too many unwanted outputs are produced; (3) too many resources are needed to achieve a desired effect. Moreover, we added some general phrase which should introduce a problem according to the tripartite definition. In table 2 for brevity, we just highlighted one of the many forms the keyword can assume in the text, the NLP system will take care of its stemming and expansion. We merged the keywords of Table 1 and Table 2, reaching a lexicon composed of 185 phrases. The lexicon was used for building a NLP system able to map the mention of a problem-related term in patents. For each phrase, we analysed the occurrence in problems and solutions.

[2] The negators, amplifiers, and deamplifiers lexicons are available on https://github.com/trinker/qdapDictionaries. Accessed on Nov. 30, 2022.

https://doi.org/10.1017/pds.2023.295 Published online by Cambridge University Press

4 RESULTS AND DISCUSSION

4.1 Sentiment analysis results

Figure 1 shows the distribution of the sentiment polarity of both problems and solutions related descriptions in Japanese patents. The median of polarity for problem and solution is 0.295 and 0.205, respectively. Since the distributions are normal (after the Shapiro-Wilkinson test), we performed the Wilcoxon-Mann-Whitney test for analysing if the distributions of problem and solution are different. The p-value of Wilcoxon-Mann-Whitney test is 2.2×10^{-16}, which confirms that the distributions are different, and the polarity value of problem is higher than the solution one. This result is highly counterintuitive and need a deeper analysis and investigation.

Figure 1. Distribution of sentiment polarity in problem and solution-related texts.

As one may expected, some problem descriptions would reach high values of negative sentiment polarity. The negative sentiment of problems is understandably related with the occurrences of negative-labelled words of the sentiment dictionary which is used to carry out the sentiment analysis. For instance, the sentence *"To suppress noise generation."* (**JP3620966B2**) gets a high value of negative sentiment because the words "*suppress*" and "*noise*" are negative-labelled words and strongly define an issue. In the same way, the description *"To solve such a problem that when a print system is created on a cloud platform not strictly secured in consistency, unnecessary data is printed due to occurrence of inconsistency"* (**JP2011170804A**) has a high negative sentiment because it explicitly mentions the negative-labelled word *"problem"* and it contains the words *"unnecessary"* and *"inconsistency"*. Contrary to what one may expect, it occurs that some problem descriptions record high positive sentiment. Although it may seem unreasonable at the first sight, problem descriptions can score high sentiment polarity when: (1) problems are kept hidden or given for granted by implicitly suggesting that state-of-the-art technologies are not capable of carrying out a certain function; (2) problems are expressed in terms of enhancements of features or functioning of already-developed technologies without directly specifying the underlying problems. In the first case, we note that the pattern "**capable of + ing-form verb**" occurs in most problem descriptions fallen in this class. For instance, problem descriptions, which implicitly suggest that state-of-the-art technologies are not capable of carrying out a certain function, instead of mentioning failures or missing functions directly, are: "*To provide a coordinate input technology capable of improving the operability of coordinate inputting."* (**JP4185825B2**); *"To provide a light source device and a luminaire capable of extracting light from a light source more efficiently."* (**JP6653450B2**). Similarly, the common pattern **"(high OR highly) + adjective + noun"** matches problems which fallen in the second class, as for example: *"To provide a dielectric ceramic composition having a high dielectric constant and high reliability*" (**JP5835012B2**); *"To provide a highly reliable semiconductor device having an insulating film …"* (**JP6681930B2**). So far, we have explained why problem descriptions can record positive sentiment unexpectedly. Similarly, we observe that certain solution descriptions record high value of negative sentiment. It is possible to spot some text regularities that explain the unexpected negative sentiments of solutions. Solution descriptions with negative sentiment typically occur when two different writing strategies are being used. The first one occurs when a solution description explicitly and briefly mentions the problems is addressing. This writing strategy is used to improve the understandability of the solution proposed providing its context, as in the following cases: *"In an error propagation path estimating…"* (**JP2001337143A**)*, "If a discharge failure is detected and the discharge failure occurs when the number of printed sheets is equal to 50 or more…"* (**JP6213346B2**). The second pattern that increases the negative sentiment of solutions occurs when names of components/devices mentioned in solutions are composed of a negative-labelled word, such as: *"A leak test method includes..."* (**JP5742824B2**), *"Disclosed is the noise suppressing device…"* (**JP4162604B2**). The more this kind of components are

https://doi.org/10.1017/pds.2023.295 Published online by Cambridge University Press

cited in a certain solution description, the higher the negative sentiment of that description is computed. The examples we provided do not offer a comprehensive analysis of all possible cases that may occur in problem description. However, the analysis allows us to reject the first hypothesis, i.e., text related to problems in patents has a negative sentiment respect to other patent parts (HP1).

4.2 Problem lexicon results

In Table 3, we compare the occurrences of problem-related keywords (column A) in problem (column B) and solution descriptions (column C) in percentage value. For instance, from Table 3, we can see how the keyword *reduce* appears in around 13% of problem descriptions (13,722 on 100,000 patents), while it appears in around 5% of the solutions. In the column D, the difference between the keyword occurrence in problem and solution. Finally, for each keyword, we tested if there was a statistical difference between the distribution in problems and the one in solution descriptions with the Wilcoxon-Mann-Whitney test. We report the p-value of each test in the column E. Table 3 shows the top-10 keywords with highest difference both for positive and negative values.

Table 3. Keywords with the highest absolute difference values.

(A) Keywords	(B) % in Problems	(C) % in Solutions	(D) Difference (B - C)	(E) p-value
reduce	13.722	5.445	8.277	0.0e+00
prevent	10.720	3.917	6.803	0.0e+00
without	9.0320	2.656	6.376	0.0e+00
even	8.407	2.636	5.771	0.0e+00
cost	3.781	0.466	3.315	0.0e+00
performance	3.560	0.674	2.886	0.0e+00
efficiency	3.193	0.557	2.636	0.0e+00
problem	2.377	0.305	2.072	0.0e+00
deteriorate	2.552	0.560	1.992	2.7e-286
consumption	2.166	0.692	1.474	2.3e-171
		…		
stop	0.421	1.141	-0.720	2.8e-74
limit	0.725	1.590	-0.865	1.9e-72
if	2.319	3.420	-1.101	6.1e-45
unfinish	0.454	1.735	-1.281	7.5e-166
period	1.117	2.683	-1.566	1.2e-143
discharge	1.976	3.818	-1.842	2.5e-130
less	0.947	3.354	-2.407	2.8e-297
time	6.801	10.122	-3.321	3.8e-131
material	3.828	7.960	-4.132	0.0e+00
low	6.374	10.851	-4.477	1.1e-240

From Table 3, we observe that some words tend to occur much more in problem descriptions rather than in solution and vice versa. Therefore, these words are able to discriminate a text as a problem or a solution. For instance, the verb "**simplify"** tends to occur much more in problem than in solution descriptions. In fact, the expression "*To simplify*" is generally used to introduce the description of problems as follow: *"To simplify the removal of a jam …"* (**JP2004004400A**), *"To simplify the construction of a shifter …"* (**JPS6137497B2**). Similarly, the expression **"in spite of"** is often used in problem descriptions to mention specific problems addressed by patents as follows: *"…in spite of an error generated on a bit string…*" (**JPH08214301A**), and *"…in spite of a simple structure of the fuel filter…"* (**JPH09310648A**).The lexicon developed by previous background works also contains words that occur much more in solution than in problem descriptions and that is the case, for example, of the word **"low"** proposed by Jeong and Kim (2014). In fact, this word is typically contained in solution descriptions such as: *"The imaging apparatus of high quality and low cost…"* (**JP2007104242A**), and *"This electron emission element is provided with a low voltage side..."* (**JP3561176B2**). However, **"low"** also appears in problem descriptions, but we cannot use this keyword for recognizing problem-related sentences using NLP in Japanese patent abstracts for the reasons explained above. Another,

https://doi.org/10.1017/pds.2023.295 Published online by Cambridge University Press

remarkable example is provided by the word "**dent".** As a matter of fact, on the one hand this word may indicates defects of technologies in problem descriptions but, on the other hand, it refers also to specific features or components of patented solutions, such as: *"Groove-shaped recesses 33 having bottom faces dented in a V shape…"* (**JP2004001338A**), and *"The formed bonding pad 45 is cylindrical and has a recessed dent 47."* (**JP2007035911A**). From our analysis, we observe that from 185 keywords we listed with the literature review (see Table 1 and Table 2), only 97 have a positive difference, i.e., they occur more in problem than in solution. This insight leads us demonstrate that it is not possible to rely on the problem-related keywords developed in the background works for describing a problem-related sentences in Japanese patent abstracts respect to other textual parts (HP2).

4.3 Japanese versus USPTO patents

In this section, we firstly compare the sentiment polarity of the Japanese patent abstracts with the one of USPTO patents, for both problem and solution descriptions. In the case of USPTO, we collected about 60,000 problems and solutions from the Description section of patents. Second, we analyse the distribution of problem-related keywords in Japanese and USPTO patents. Concerning the sentiment polarity, Table 5 provides a summary of the distribution of sentiment polarity for problems and solutions, in the case of USPTO patents. We can observe how in Table 5, problems have a negative polarity (with a median of -0.082) than solutions (median equals to 0.000). The USPTO has an opposite behaviours respect to Japanese patents, where the polarity values of problems is more positive than those of solutions.

Table 4. Statistics for problem and solution descriptions in USPTO patents.

	Min.	1st Quartile	Median	3rd Quartile	Max.
Problem	-13.432	-0.275	-0.082	0.101	8.481
Solution	-10.492	-0.127	0.000	0.009	15.228

Regarding the distribution of keywords, Figure 2 shows how the problem-related keywords are distributed in Japanese and USPTO patents. We plot on x-axis the percentage of problem descriptions which contain the keywords-related problem, while the percentage of solutions is shown on y-axis. We show in Figure 2 the line of equal distribution, where there are keywords equal distributed between problems and solutions. On the bottom of the equal distribution line, we can find keywords that occurs more in problems than in the solutions (in blue colour), vice-versa on the top (in red colour). We can observe as in the case of USPTO, that keywords are concentrated on the bottom of the line, and they have higher values on x-axis than the Japanese case. Indeed, there are 130 keywords (out of 185) that appear more often in problems than in solutions (97 for Japanese patents, as described in Section 4.2).

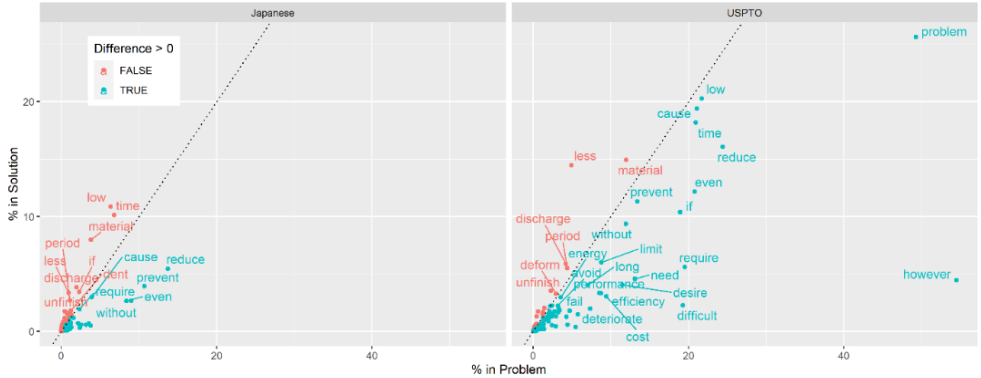

Figure 2. Distribution of keywords in Japanese and USPTO patents.

In the keywords comparison, we can have 4 cases: (1) keywords occur more in solutions than problems for Japanese patents, while they occur more in problems than solutions for USPTO patents; (2) keywords occur more in problems than solutions for Japanese patents, while they occur more in solutions than problems for USPTO patents; (3) keywords occur more in solutions than problems for both Japanese and USPTO patents; (4) keywords occur more in problems than solutions for both Japanese and USPTO patents. Table 5 shows some examples of keywords for each class with the number of keywords which fallen in the class, where we highlight in grey negative values of difference. From Table 5, **case 1** includes 29 keywords, which widely occur in USPTO patents for

https://doi.org/10.1017/pds.2023.295 Published online by Cambridge University Press

describing a problem. For instance, the keyword "**however**" occurs in more than 50% problems of USPTO, instead it occurs in only 0.031% of Japanese problem descriptions. These results allow us to observe how USPTO problems have a different structure if compared with the Japanese ones from the natural language point of view.

Table 5. Comparison of keywords between Japanese and USPTO patents.

#	Size	Keywords	Japanese			USPTO		
			% in Problem	% in Solution	Difference	% in Problem	% in Solution	Difference
1	29	however	0.031	0.247	**-0.216**	54.491	4.450	**50.041**
		if	2.319	3.420	**-1.101**	18.915	10.364	**8.551**
		time	6.801	10.122	**-3.321**	20.930	18.169	**2.761**
		low	6.374	10.851	**-4.477**	21.698	20.281	**1.417**
		although	0.107	0.130	**-0.023**	5.761	1.465	**4.297**
2	6	simplify	1.170	0.213	**0.957**	1.236	1.476	**-0.241**
		term	0.233	0.147	**0.086**	1.255	1.666	**-0.410**
		unnecessary	0.402	0.291	**0.111**	1.275	1.417	**-0.142**
		deficiency	0.040	0.029	**0.011**	0.099	0.137	**-0.038**
		blemish	0.003	0.001	**0.002**	0.003	0.008	**-0.005**
3	24	less	0.947	3.354	**-2.407**	4.898	14.463	**-9.565**
		material	3.828	7.960	**-4.132**	11.956	14.920	**-2.964**
		hold	0.252	0.807	**-0.555**	0.644	1.718	**-1.074**
		period	1.117	2.683	**-1.566**	4.379	5.503	**-1.124**
		gap	0.096	0.560	**-0.464**	0.445	1.280	**-0.835**
4	87	problem	2.377	0.305	**2.072**	49.260	25.635	**23.625**
		difficult	0.493	0.148	**0.345**	19.286	2.255	**17.030**
		require	2.346	1.918	**0.428**	19.528	5.587	**13.941**
		need	0.979	0.859	**0.120**	13.081	4.564	**8.518**
		desire	1.279	1.244	**0.035**	11.472	4.033	**7.438**

5 CONCLUSION

Patent must mandatorily include the description of the technical problems which are solved by the invention. For an engineering design perspective, the definition of problem is blurred. A clear definition of problem is needed to automatically extract problem form patents and analyse technical problems with text mining techniques. The massive analysis of problems enables to (1) map the technical prior art, (2) generate new ideas for the conceptual design phase and (3) study the technological evolution. For these reasons, previous studies rely on three main approaches of text mining: keywords-based, rule-based and machine learning methods. Sentiment analysis for identifying problems is also used in literature (Chiarello et al., 2017). We argue that text mining techniques can help designers, inventors, and researchers to provide a clear definition of what a problem is. We formulated two hypotheses of how a problem is expressed in natural language. We use the Japanese patents which explicitly split the abstract in two parts: "Problem to be solved", and "Solution to the problem". To test our hypothesis, we rely on sentiment analysis and keywords-based methods, since these techniques provide clear way of expressing problems in patents. Japanese problems, in opposite to what literature argued, have a more positive sentiment polarity than solution, leading to reject HP1. Moreover, our results suggest as the problem-related keywords identified by background works cannot be used for extracting problems from Japanese patents. Finally, the results obtained with Japanese patents are compared with the one of USPTO. The paper demonstrates that USPTO patents follow an opposite behaviour respect to Japanese ones. This work advanced our understanding of the ways of expressing problems and solutions in patents and it shows that these theoretical elements share hidden patterns that need to be further investigated. Further studies could focus on context-based NLP techniques to study in deeper details when problem-related keywords and negative-labelled words are used to specify solutions (instead of problems). The inherent relation between problems and solutions could be leveraged in future research to develop machine learning systems capable of 1) linking solutions to their corresponding problems and vice versa and 2) automatically generate solutions from input problems with Natural Language Generation.

https://doi.org/10.1017/pds.2023.295 Published online by Cambridge University Press

ACKNOWLEDGMENTS

This research has been partly funded by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme.

REFERENCES

Becattini, N., Cascini, G., & Rotini, F. (2011), "Correlations between the evolution of contradictions and the law of identity increase", *Procedia Engineering*, 9, 236-250. https://doi.org/10.1016/j.proeng.2011.03.115

Bonaccorsi, A., Apreda, R., & Fantoni, G. (2020), "Expert biases in technology foresight. Why they are a problem and how to mitigate them", *Technological Forecasting and Social Change*, 151, 119855. https://doi.org/10.1016/j.techfore.2019.119855

Cascini, G., Fantoni, G., & Montagna, F. (2013), "Situating needs and requirements in the FBS framework", *Design Studies*, 34(5), 636-662. https://doi.org/10.1016/j.destud.2012.12.001

Chiarello, F., Fantoni, G., & Bonaccorsi, A. (2017), "Product description in terms of advantages and drawbacks: Exploiting patent information in novel ways", In DS 87-6 *Proceedings of the 21st International Conference on Engineering Design* (ICED 17) Vol 6: Design Information and Knowledge, Vancouver, Canada, 21-25.08. 2017 (pp. 101-110).

Chikkamath, R., Parmar, V. R., Hewel, C., & Endres, M. (2021), "Patent Sentiment Analysis to Highlight Patent Paragraphs", *arXiv* preprint arXiv:2111.09741. https://doi.org/10.48550/arXiv.2111.09741

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018), "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv* preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Giordano, V., Puccetti, G., Chiarello, F., Pavanello, T., Fantoni, G. (2022), "Unveiling the Inventive Process from Patents by Extracting Problems, Solutions and Advantages with Natural Language Processing", Available at SSRN. http://dx.doi.org/10.2139/ssrn.4223458

Guarino, G., Samet, A., & Cavallucci, D. (2022), "PaTRIZ: A framework for mining TRIZ contradictions in patents", *Expert Systems with Applications*, 117942. https://doi.org/10.1016/j.eswa.2022.117942

Hu, M., & Liu, B. (2004), "Mining and summarizing customer reviews", In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177).

Jeong, C., & Kim, K. (2014), "Creating patents on the new technology using analogy-based patent mining", *Expert Systems with Applications*, 41(8), 3605-3614. https://doi.org/10.1016/j.eswa.2013.11.045

Kim, J. H., & Choi, K. S. (2007), "Patent document categorization based on semantic structural information", *Information processing & management*, *43*(5), 1200-1215. https://doi.org/10.1016/j.ipm.2007.02.002

Kim, S., & Yoon, B. (2021), "Patent infringement analysis using a text mining technique based on SAO structure", *Computers in Industry*, 125, 103379. https://doi.org/10.1016/j.compind.2020.103379

Liang, Y., & Tan, R. (2007), "A text-mining-based patent analysis in product innovative process", Trends in computer aided innovation, 89-96. https://doi.org/10.1007/978-0-387-75456-7_9

Melluso, N., Fantoni, G., & Martini, A. (2023), "The ABC (Affordance-Bias-Cognition) Reasoning of Product Use Interaction: A Text Mining Approach", In *Design Computing and Cognition'22* (pp. 51–66). Springer International Publishing. https://doi.org/10.1007/978-3-031-20418-0_4

Moehrle, M. G., Walter, L., Geritz, A., & Müller, S. (2005), "Patent-based inventor profiles as a basis for human resource decisions in research and development", *R&d Management*, 35(5), 513-524. https://doi.org/10.1111/j.1467-9310.2005.00408.x

Organization, W. I. P. (2004), "WIPO Intellectual Property Handbook: Policy, Law and Use", 489. *World Intellectual Property Organization*.

Puccetti, G., Giordano, V., Spada, I., Chiarello, F., & Fantoni, G. (2023), "Technology identification from patent texts: A novel named entity recognition method", *Technological Forecasting and Social Change*, 186, 122160. https://doi.org/10.1016/j.techfore.2022.122160

Souili, A., Cavallucci, D., Rousselot, F., & Zanni, C. (2015), "Starting from patents to find inputs to the problem graph model of IDM-TRIZ", *Procedia Engineering*, 131, 150-161. https://doi.org/10.1016/j.proeng.2015.12.365

Suh, N. P. (1998). Axiomatic design theory for systems, "Research in engineering design", 10(4), 189-209, https://doi.org/10.1007/s001639870001

Tiwana, S., & Horowitz, E. (2009), "Extracting problem solved concepts from patent documents", In *Proceedings of the 2nd international workshop on Patent information retrieval* (pp. 43-48). https://doi.org/10.1145/1651343.1651356

Zanni-Merk, C., Cavallucci, D., & Rousselot, F. (2009), "An ontological basis for computer aided innovation", *Computers in Industry*, 60(8), 563-574. https://doi.org/10.1016/j.compind.2009.05.012