2019 – EPO Academic Research Programme

Understanding the Business Value of SMEs' Patent Portfolio: An Artificial Intelligence based approach.

FINAL REPORT

Pisa (IT), 31/03/2022

FOREWORD

This report presents the findings from the project, titled **"Understanding the Business Value of SMEs' Patent Portfolio: An Artificial Intelligence based approach"**, and granted during the 2019 EPO Academic Research Programme. The project is conducted by a joint research team from Scuola Superiore Sant'Anna and ISTI- National Research Council, both institutions based in Pisa (IT).

The project's primary goal is to assess and forecast the commercial value of SMEs' patents measuring the proximity between their portfolio and their business model. With the use of artificial intelligence methodologies, the project aims to:

i. Identify the closeness of firms' business model development from their technological footprint (patents).

ii. Predict the success likelihood of a specific business model applied to a given patent.

iii. Suggest alternative business models more in line with the patent portfolio characteristics.

The project relies on original and relatively rare data regarding company business models disclosed directly by SMEs extracted from funding applications, submitted during the period 2014 to 2019, to the Horizon 2020 SME Instrument (SMEi) program. By leveraging team's members ' expertise in IP strategic management, business model development, big data, and machine learning we matched SMEi's proposals with applicants' patent portfolio data to perform an innovative analysis about the value of the patents in conjunction with detailed information on business models and commercialization strategies.

The project is organized in two parts. In the first part, starting from the patent analysis carried out to assess the company's technological background, we analyzed firms' patent portfolio and business models characteristics. Then, we defined a set of proximity indexes to measure and assess the consistency of the companies' technological capabilities with their new business models.

In the second part, by leveraging such indexes, we will implement a set of machine learning/artificial intelligence methodologies and, subsequently, we will make them openly available to the community. Overall, the project will address the following two research questions:

1. Is the proximity between patent portfolios and new business models related to successful commercialization of the embedded technology?

2. Based on the observed past matches between patents and business plans, is it possible to educate an AI-based tool to predict the characteristics of future potential commercial applications of new patents?

The key contribution of this project is three-fold. First, it enhances the knowledge regarding the relationship between technological assets (i.e., patent portfolio value) and the breakthrough, market-creating innovation. Second, the methodology implemented in the project can be used, at the managerial level, to formulate firm objectives, strategic technological trajectories, and potential alternative business strategies to discover new business opportunities. Last, this research will help policymakers in (i) distinguishing technological preconditions for the implementation of policies for the market-creations and breakthrough innovations; (ii) describing modes of commercialization of innovative established SMEs and start-ups; and (iii) assessing the importance of patented technologies for SMEs business venturing.

Contents

EXECUTIVE ABSTRACT 6
THE PROJECT 7
The role of SMEs in innovation development 7
The roots of the project 7
Firms' technological capabilities and business model 9
Artificial Intelligence and patents- the state of the art 9
Research question 12
Contributions 13
WORK PACKAGE 1 15
Task 1.1 Dataset Description 15
Description of the SME Instrument Programme 15
The Dataset 16
Task 1.2 Business application characterization (business models) 18
Business models identification 18
Analytic description of the topics 24
Business Models Analysis 37
Task 1.3 Firms' technological capabilities identification (patents) 42
Analytic description of the patents' topics 42
WORK PACKAGE 2 50
How proximity affect successful innovation? An empirical analysis 50
An empirical model combining the power of AI and economic analysis 50
LSI-based similarities 52
Computing Transformer-based similarities 54
Theory and Hypotheses 56
Corporate Coherence 56
The moderating effect of slack 57
Exploring proposal success: empirical model specification and variables 57
Dependent variables measuring proposal success 58
Independent variables 59
Empirical results 60
Summary and discussions 63
WORK PACKAGE 3 66
Dissemination of the results 66

References 67
Appendix 75

EXECUTIVE ABSTRACT

Small and Medium Enterprises (SMEs) account for 99,8% of the overall EU28 companies and their role for economic growth, creation of employment and innovation development has been recognized and supported through innovation policies. Moreover, their smaller scale, their leaner structures and their less institutionalized routines and processes allow them to build and improve their invention portfolio.

In order to support SMEs, the EU deployed the SME instrument (SMEi). This innovative funding scheme by EU Horizon 2020 is focused on the contribution into solving societal challenges through SME' smart growth and job creation. SMEi supports high-risk projects with highly innovative potentials with the capability to generate an impact on EU and global markets.

Then, the aim of this research project is to assess and estimate the potential commercial value of SMEs' patent portfolio by developing an original Artificial Intelligence (AI) methodology to define new and better solutions to develop SMEs' entrepreneurial activities. Then, we investigate the relation among SMEs' business models and their technological assets, we estimate the existence and intensity of a relationship between a business model strategy and specific technological capabilities asset and, lastly, we identify the ability of a business model to be successful according to SMEs' technological footprint.

We focus on the role of SMEi applicants' corporate coherence, defined as the degree of similarity between each firm's proposal and the cluster of technologies already present in their technologies' portfolio, to determine a positive evaluation of each firm's SMEi proposal.

The research project is subdivided into three Work Packages, as follows: WP1 defines the SMEi data collection, the SMEs' business models' characteristics and their technological capabilities identifications through a unsupervised and semi-supervised approach; WP2 estimates SMEs' business models, their technological capabilities proximity and their business model success prediction; while, WP3 refers to the dissemination of the results through an analytics dashboard.

Overall, our findings show that corporate coherence is positively correlated with firms' evaluation score assigned to the SME-I's proposal. This can be translated as the ability of firms into leveraging their core technological knowledge and competencies by submitting a proposal which is adherent to their activities' technological trajectory. Not only a SME-I proposal is positively evaluated when

firms both signal corporate coherence and ability to mitigate innovation risks, but also when firms operate in an innovative and dynamic environment.

THE PROJECT

The role of SMEs in innovation development

Nowadays, the intellectual value for most inventions is embedded in the distinctive combination of breakthrough ideas and business model application. In this versatile innovation environment, SMEs require to build and improve an ever-evolving invention portfolio to optimize the resources as well as to develop the ability to survive or to disrupt the market (Del Sarto et al. 2019). SMEs are usually focused on specific technologies, embedded in patents, to produce inventions and often employ a highly specialized workforce. With a lean organizational structure, the decision process is faster, and their reaction to technological opportunities is timelier. According to their size constraint, SMEs operate at a smaller scale using focused and specialized resources, capabilities, and processes addressing a specific market, consumers, and industry domain (Levy & Powell, 2000; Meister, 2017). Because of their limited scale and focused operations, most of them do not have a significant number of slack resources (e.g., financial and human resources) or the appropriate capabilities to develop the routines and processes to be agile (Neirotti & Raguseo, 2017). Having this level of agility can also be considered an asset, especially in a high-tech dominated environment because they can quickly react to the external changes (Levy & Powell, 2000; Levy, Powell, & Yetton, 2001). This level of agility allows SMEs to lower the costs of reconfiguring their operational model with leaner structures and lesser institutionalized routines and processes to be changed. SMEs are also crucial because they represent the main engine of economic growth, creation of employment, and innovation development (Georghiou and Roessner 2000, Lanhan 2016; Lerner 2009; Solow 1956). The EU is no exception: SMEs play a fundamental role in the EU28 economy, accounting for 99.8% of the overall number of EU companies and contributing to employment and growth (Muller et al., 2017). At policy level, the role of small businesses for economic growth and innovation has been recognized by governments all over the world (Autio, 2016; Muller et al., 2016). The EU confirms the global trend of supporting theinnovation policy of SMEs' research and development (R&D) activities improvement (APRE, 2016; Fresco et al., 2015).

The roots of the project

In line with the public effort made by the EU, in this study, we consider, as a research setting, the innovative funding scheme of the SME Instrument (SMEi) by Horizon 2020 in the eighth EU

Research and Innovation Framework Programme (Di Minin et al. 2016). The SMEi's goal is to enable the flourishing of a business ecosystem in which SMEs' smart growth and job creation can contribute to solving societal challenges (EASME 2016). It addresses the financial needs of EU innovative SMEs oriented toward internationalization and growth, committed to implementing high-risk ideas with highly innovative potential. It selects contracts and provides coaching to the most innovative EU SMEs (EASME 2017) by supporting projects with European relevance that are potentially able to introduce disruptive innovations and change the business world. The SMEi targets EU Innovation Champions that show the capability of generating an impact on the EU and global market through the implementation of business models and activities that exploit their technological capabilities and innovative potentials. Working closely with European and Italian institutions (i.e., EASME; Netval), Sant'Anna School of Advanced Studies acquired high-level expertise and provided a valuable contribution to the small business research fields. Sant'Anna engaged in research on SMEs innovation policies, focusing on the most recent EU Framework Programme for Research and Innovation, Horizon 2020, and its support to small businesses by exploring different aspects of enterprise behavior such as:

i. Whether firms interested in the SME Instrument are accessing other sources of funding, in particular, venture capital.
ii. Whether firms that access venture capital present different characteristics from those that are instead oriented towards the SME Instrument.
iii. Illustrate the process of SMEs' adaptation and their innovative responses to the evolution of the digital platform economy.
iv. Explore how new and original business strategies are emerging in European SMEs operating through digital platforms/digital markets.

Through these studies, the team drew an in-depth profile of EU innovative SMEs considering two levels of analysis. On a first level, we examine the factors enabling SMEs business success based on market indicators; on a second level, we explore SMEs' successful managerial practices and innovation strategies. These are, nowadays, essential assets to implement the proposed research study adding the development of a level of analysis regarding the SMEs patents portfolio. Accordingly, working closely with the Italian National Research Council, we have integrated the use of AI-based approach to investigate new relationships and overlaps between these levels. Our knowledge in terms of ML/AI methodologies and innovation management allow us to pursue new interactions between SMEs' innovation capabilities and the effectiveness of their business models implementation.

Firms' technological capabilities and business model

The literature considers technological capabilities as an unobservable construct involving several measurable elements and indicators such as R&D intensity and patents (Coombs and Bierly, 2006). Exploring the technological capabilities of a company allows us to understand the opportunities that a firm can pursue in the virtue of its organizational learning skills. Technological capabilities offer insights on the technological domains in which the company already conducted research activities and acquired R&D expertise. Indeed, technological capabilities represent the firm's identity, strategies, and real technological assets' status (Aharonson and Schilling, 2016; Lee et al., 2009). Firm's technological capabilities can also be quantified through the information included in the patent document (see Squicciarini et al. 2013 for an overview of patent indicators). For instance, technological patent classification can be used to assess the scope of the firm's technological capabilities (Lerner, 1994) or the firm's technological specialization (Granstrand et al., 1997). Similarly, both backward and forward citations provide the basis for measuring critical features of the firms' technological capabilities and sources of knowledge, such as the extent of knowledge recombination (Gompers et al., 2005; Trajtenberg et al., 1997), the breadth of technological impact (Galasso et al., 2011), and the closeness to basic science (Cassiman et al., 2008; Narin et al., 1997). However, considering only technological capabilities (measured by patents) is not sufficient for a comprehensive understanding of the firm's competitive advantage. Indeed, the literature finds that it is hard to infer about commercial and business activities relying only on patent data since the conventional analyses of patents usually collect information on a company's current technology assets, rather than identifying new business opportunities (Lee et al., 2009). Also, the literature points to the importance of business models design evaluation as a complementary factor to consider in the assessment of firms' commercial and business activities. Business models enable firms to create and share value (Di Minin et al., 2016; Zott and Amit, 2010) and allow the exploitation of a company's technological capabilities on the market. In line with the most recent studies, in the project we implement the use of AI/ML methodologies to define new and better solutions helping SMEs in developing faster and leaner entrepreneurial activities.

Artificial Intelligence and patents- the state of the art

In recent years a new research area has emerged. This research area is called patent mining and it has been recognized as an essential task at government level. Public patent authorities in the United States, United Kingdom, China, and Japan have invested various resources in improving the performance of creating valuable patent analysis results for various patent analysis tasks. Patent

analysis is a non-trivial task; it is necessary to have a certain degree of expertise in different research domains, including data mining, information retrieval, domain-specific technologies, and business intelligence.

The main patent analysis issue can be summarized in five different tasks:

1. Patent Retrieval: it is a subdomain of information retrieval, in which the basic elements to search are patent documents. Against the recent advances, the task of patent retrieval remains challenging from multiple perspectives:
    a. *Low Readability:* people may use rhetorical structures and ambiguous terms to defend their invention in order to obtain broader protection.
    b. *Lengthy Query:* people often use the whole patent document as a query to perform searching.
    c. *High Recall:* missing one powerfully relevant document in patent retrieval is unacceptable because of the enormous cost of a patent lawsuit.

    To tackle this task, the classical techniques used are the Query Generation (Bhatia et al 2012, Kim et al. 2011, Mahdabi et al. 2011, Trappey et al 2012) and Query Expansion (Al-Shboul et al. 2011, Ganguly et al. 2011, Hristidis et al 2010, Magdy et al 2011, Mahdabi et al 2012, Tannebaum et al 2012); some new techniques use context-dependent methods, such as Biomedical text mining methods (Alves et al. 2017, Lou et al. 2018) or Chemical text mining methods (Akhondi et al. 2019).
2. Patent Classification: Since 1960, automatic classification has been identified as an interesting problem in text mining and natural language processing. Nowadays, researchers have devised many excellent algorithms to address this task, but it is still a non-trivial problem in the domain of patent mining due to the complexity of patent documents and patent classification criteria. There are several challenges to tackle this task:
    i. Patent documents often involve sophisticated structures, verbose pages, and rhetorical descriptions.
    ii. The hierarchical structure of the patent classification schema is quite complex.
    iii. The huge volume of patent documents, and the increasing variety of patent topics, exacerbates the difficulty of automatic patent classification.

    The major focus along this research direction includes utilizing different types of information to perform classification (Kim et al. 2007, Teodoro et al. 2010); and testing the performance

of different classification algorithms on patent documents (Chen et al. 2012, Fall et al 2003, Tikk et al. 2007, Bergeaud et al. 2017).

3. Patent Valuation: the evaluation of the importance/quality of patent documents is an important process, which aims to assist internal decision making for patent protection strategies. To tackle this issue, researchers often rely on two types of approaches: 1) unsupervised exploration (Erdi et al. 2012, Jin et al. 2011, Lee et al. 2012, Messeni Petruzzelli et al. 2014, Van Zeebroeck 2011); 2) supervised evaluation (Erdi et al. 2012, Hu et al. 2012; Jin et al. 2011; Lupu et al. 2010, Van Zeebroeck 2011, Ploskas et al. 2019, Oh et al. 2014, Ponta et al. 2019, Dong et al. 2018).
4. Patent Visualization: The complex structure of patent documents often prevents the analysts from quickly understanding the core idea of patents. To resolve this issue, it would be helpful to visualize patent documents; this task is often referred to as patent visualization, an application of information visualization. The major techniques can be grouped into three categories:
    i. Structured data visualization: including patent number, filing date, issued date, and assignees, which can be utilized to generate a patent graph by employing data mining techniques (Tang et al. 2012; Yang et al. 2008, Yeap et al. 2003);
    ii. Unstructured data visualization: consisting of textual content of patent documents, such as abstract, descriptions of the invention, and major claims, which can be used to generate a patent map by employing text mining techniques (Honghua et al. 2009, Lee et al. 2009);
    iii. Hybrid visualization (Carrier 2012, Suh et al. 2009; Tang et al. 2012, Yang et al. 2010, Yeh et al 2018).
5. Cross-Language Mining: Patent documents are quite sensitive to regions, i.e., patents from different regions might be described by different languages. However, in reality, patent analysts prefer to receive localized patent information, even if multiple languages describe them. Also, international patent documents are required to be written by the language accepted worldwide, which is often referred to as patent globalization. In such cases, cross-language patent mining is needed to support patent localization/globalization. The primary task is cross-language information retrieval, which enables us to retrieve information from other languages using a query written in the language that we are familiar with. In general, a cross-language patent retrieval system can be constructed using two techniques: 1) machine translation

(Chechev et al. 2012, Fujii et al 2009, Goto et al. 2011, Jochim et al. 2010, Magdy et al. 2011); 2) semantic correspondence (Kondo et al 2011, Li et al 2007, Jin et al. 2010).

Research question

Using entrepreneurial projects and patent portfolio as proxies, we aim at assessing and estimating the potential commercial value of SMEs and start-ups' patent portfolio responding to the following research questions:

1. Is the proximity between patent portfolio and business model related to technology commercialization success?
2. Based on the observed past matches between patents and business models, is it possible to educate an AI-based model to predict the characteristics of the most likely successful business model with respect to the analyzed patent portfolio?

With the use of artificial intelligence methodologies, firstly we investigate if firms develop their business models according to their technological assets (patents). Secondly, we estimate the success likelihood of a specific business model applied to a given patent; and lastly, we identify alternative SMEs' business models more suitable with their patent portfolio. Since it is challenging to infer information about commercial and business activities relying only on patent data, linking patents to entrepreneurial projects proposals, offered by Horizon 2020's SME Instrument, provides a unique opportunity to evaluate SMEs' strategies to pursue competitive advantage. In this project, to be developed under EPO's auspices, we propose to analyze, for the first time, patent data in conjunction with detailed information on the content of SMEs' business models. With the proposed research questions, we are investigating:

i. Whether European SMEs design their business model strategies building on their accumulated technological assets (i.e., patents portfolio).
ii. The existence and intensity of a relationship between the deployment of a particular business model strategy and specific technological capabilities asset.
iii. The ability of a given business model to be successful according to the firm's technological footprint.

To measure the proximity among business model strategies and technological capabilities, we are developing an original AI-tool. We use patents as a proxy for SMEs' technological footprint, and entrepreneurial projects proposals, submitted to SMEi, as a proxy for business model application. Furthermore, the project aims to unfold the relations between firms' technological competencies, the business model, and the technology-commercialization as output. Once those relations are identified

and measured, leveraging ML/AI methodologies, we devise an algorithmic approach that forecasts the likelihood of success of a given business model according to the company's technological background.

Contributions

The new use of the AI/ML methodologies allows us to detect the value of SMEs' intellectual property combined with their business model application. We are developing a tool that helps SMEs, in this versatile innovation environment, to better exploit their portfolio of inventions and to optimize resources. Focusing on EU SMEs' we contribute, from the scientific and managerial perspective, in developing a methodology suitable for further research on patents' value (Aharonson and Schilling 2016, Gerken and Moehrle 2012). To guarantee the dissemination and replicability of our methods, we will publish in "open access" the complete set of algorithms, making them available to the scientific community. This research project entails several implications for practitioners, academics, and policymakers according to the following three dimensions:

1. Demonstration of the incremental use of the ML/AI methodologies in analyzing firms' patent portfolio. Through this project, we construct a comprehensive database with patent data and entrepreneurial projects matches, and standardized secondary info on SME Instrument participants. Therefore, we develop a tool for measuring the proximity between firms' business models with their technological footprint (patent portfolio). Moreover, with this innovative tool, we will be able to match the most likely successful business model with a specific technological ability (patent). At the end of the project, we will publish an "open access" scientific paper with attached the complete set of algorithms, to allow the use of our tool in other research projects.
2. Managerial implications: we have a unique opportunity of empirically investigating the links between SMEs' technological capabilities and business model design. This study has significant managerial implications related to business strategy designs that incorporate strategically important technological perspectives based on the company patent portfolio. With this project, we are able to identify business strategy profiles and effective entrepreneurship models of European SMEs. Managers can use the methodology we propose to formulate firm objectives, strategic technological trajectories, and potential alternative business strategies to discover new business opportunities. Companies, especially SMEs, may have an imprecise business strategy to follow, given the mixed domain interests. It can be challenging to keep ideas on track aligning the technological capability with the business

model design and the technology development plan. To minimize the effort, it is essential that there is an easy way to match the more suitable business model to follow for the owned technological capability. This knowledge provides an advantage, especially to SMEs, which usually suffer from a financial and professional resources shortage. The implemented methodology opens doors to combining unique solutions in different domains to ensure better and more comprehensive IP protection for new applications.

3. Policy implications: our research describes new and alternative entrepreneurial strategies to consider in identifying the priorities of the policies in terms of innovation-enhancement and entrepreneurial-development. This study is the touchstone to assess the coherence and consistency of companies' technology assets with their expressed business modeling. The study proves useful tools for policymakers in:
    i. Distinguishing technological preconditions for the implementation of policies for the market-creations and breakthrough innovations.
    ii. Describing modes of commercialization of innovative established SMEs and start-ups.
    iii. Assessing the importance of patented technologies for SMEs business venturing.

Table 1 briefly summarizes the three main work packages of the project.

Table 1 Work Packages description

Work Package	**Activities**
WP1: Data Collection and Enrichment	**Task 1.1 ->** SMEi data collection and enrichment **Task 1.2 ->** Business application characterization **Task 1.3 ->** Firms' technological capabilities identifications
WP2: Data Analysis	**Task 2.1 ->** SMEs business model and technological capabilities proximity estimation **Task 2.2 - >** Econometric analysis for relationship estimation **Task 2.3 - >** Business model success prediction
WP3: Reporting and Dissemination	**Task 3.1 ->** Report writing **Task 3.2 ->** OA dissemination and software deployment (dashboard)

WORK PACKAGE 1

Task 1.1 Dataset Description

Description of the SME Instrument Programme

According to the project description, we aim to explore the relationships between technology-based European SMEs' technological footprints (as the results of the SME's technology capabilities resulting from its patent portfolio) and their business models. In this vein, we built our project by relying on different sources of primary and unstructured data and secondary data.

First, to explore SMEs' business models, we analyzed their grant applications to the SME Instrument Programme. This Programme aims to offer SMEs EU funding and support to develop their innovation projects and reach the market. The Programme is intended for small and medium-sized enterprises that count less than 250 employees and an annual turnover of no more than €50 million and/or a balance sheet of no more than €43 million. The project started in 2014 and it is organized in different cut-offs per year during which SMEs can apply for the grants. The fundings offered by the SME Instrument cover two different phases of innovation projects developments:

i. Concept and feasibility assessment phase (phase 1).
ii. Innovation project (phase 2).

Concerning Phase 1, funding is conceived to help SMEs explore and assess the technical feasibility and commercial potential of breakthrough innovation in a given industry (e.g., a risk assessment, market study, intellectual property management of a new product, service, or a new application of existing technologies). The amount offered is a lump sum of €50 000 for a six-month project. Generally, SMEs invest this amount in elaborating a structured business plan.

As regards Phase 2, funding is intended to sustain innovation projects underpinned by a strategic business plan and a feasibility assessment. In this case, the amount recognized by the EU Commission is usually between €500.000 and €2.5 million for projects lasting around 1-2 years. During this period, SMEs are requested to reach the market with a new idea (product, process, service) or to develop a business innovation plan which includes a detailed commercialization strategy and a plan on how to attract private investors[1].

[1] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/sme_en.htm

The Dataset

Overall, we collected data from 2014 to 2019, including the cut-offs from 1 to 7 because, starting from cut-off 8 (October 2019), the programme has been restructured by the European Commission who launched the EIC Accelerator Programme[2]. As reported in Table 2, our dataset is composed of 72.973 proposals of which 46.073 are applications for Phase 1 while 26.900 are applications for Phase 2. The number of applications per year is reported in Table 2. Overall, we collected proposals from 24.800 SMEs.

Table 2 Description of the proposals included in the dataset

Proposals in the Dataset			
	Tot	**Phase 1**	**Phase 2**
2014-2019	72.973	46.073	26.900
2014	5.516	4.307	1.209
2015	11.154	7.528	3.626
2016	12.057	7.968	4.089
2017	15.395	9.002	6.393
2018	14.575	8.492	6.083
2019	14.276	8.776	5.500
Main list	5.150	3.928	1.222
Below available budget	15.402	5.105	10.297
Below Threshold	51.391	36.381	15.010
Ineligible	1.030	659	371

Table 3 shows the distribution of the proposal per year according to the evaluation received by the EU Commission.

[2] https://ec.europa.eu/easme/en/eic-accelerator

Table 3 Proposals evaluations per year

Phase 1						
	2014	**2015**	**2016**	**2017**	**2018**	**2019**
Main list	259	574	698	616	936	845
Below available budget	120	671	681	1.107	1.132	1.394
Below Threshold	3.688	6.207	6.473	7.109	6.367	6.537
Ineligible	240	76	116	170	57	-
Phase 2						
Main list	74	144	202	249	317	236
Below available budget	178	1.241	1.748	2.873	2.141	2.116
Below Threshold	870	2.171	2.061	3.197	3.563	3.148
Ineligible	87	70	78	74	62	-

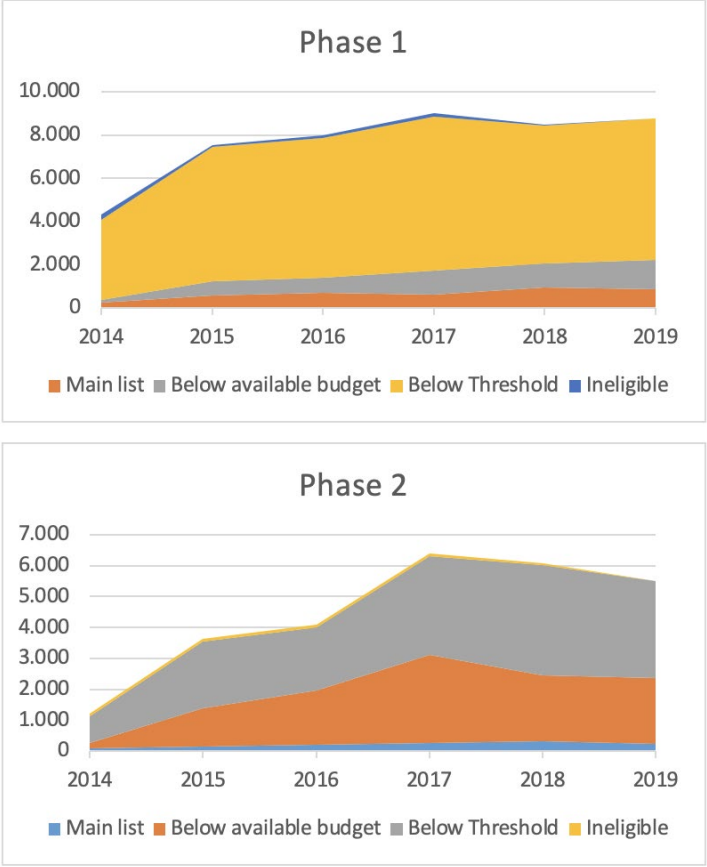


Figure 1 Distribution of the proposals

Concerning the identification of the SMEs technological footprint we finally identified, among the 24.800 SMEs who applied for a grant, those with at least one international registered patent reaching a final number of 24.339 SMEs included in our dataset. By using the EPO's online database PATSTAT we finally downloaded 191.086 patents as shown in Table 4.

Table 4 Number of SMEs and Patents included in the dataset

Number of SMEs and Patents in the Dataset	
Smes	24.800
Smes with at least one registered patent	24.339
Patents	191.086

Task 1.2 Business application characterization (business models)

In recent years, topic modeling has become one of the most useful tools for exploring textual corpora's latent structure. Latent Dirichlet Allocation[3], or LDA, models are statistical machine learning models for clustering words into topics and documents into a mixture of topics.

LDA uses a Bayesian inference model that associates each document with a probability distribution over topics, where topics are probability distributions over words. We used LDA to extract latent information from proposals and patents (Task 1.3) by evaluating different preprocessing and analysis strategies.

Business models identification

Our dataset contains 59,252 abstracts of the proposals with 110,683 distinct words[4]. A first characterization of the dataset is given by the number of words and the length of the words (number of characters) that make up the abstracts. In Figure 2 the histograms show the abstract length distribution and the word length distribution respectively.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, null (3/1/2003), 993–1022.

[4] This preliminary analysis has been conducted using the information we had available in January 2020. The remaining of the information updated and collected during the following months will validate the model here implemented.

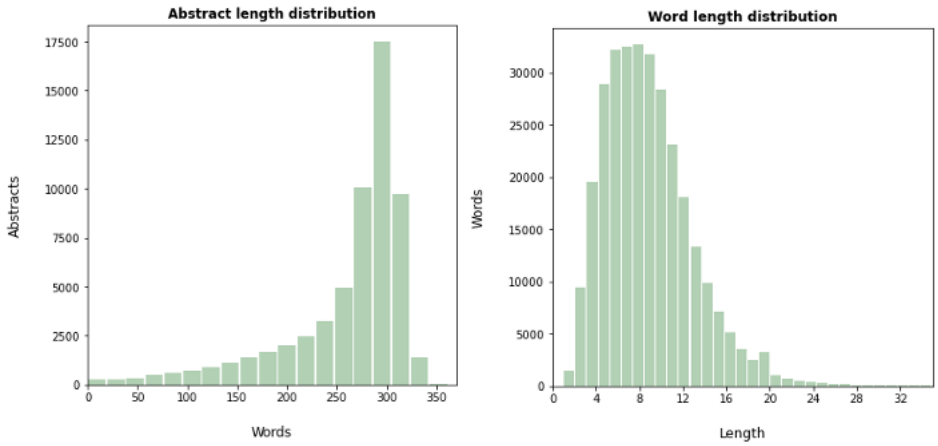


Figure 2 The histogram on the left shows the distribution of abstract length and the one on the right shows the distribution of word length

We can immediately notice that most of the abstracts are around 300 words, abstracts with a small number of words will also be dealt with in the pre-processing stage. In addition, the second histogram shows the presence of words longer than 18 to be attributed to errors in the conversion of the abstract from pdf to text or web addresses that will have to be removed in the pre-processing phase.

Cleaning the text helps us obtain a high-quality output of the model, removing all text irrelevant for analysis and getting the basic form of words. We removed the insignificant words by looking at their part-of-speech tags. We applied three different types of part-of-speech filtering on the raw data to generate three sets of input data per LDA model and evaluate which one is the best performing. The abstracts were tagged and lemmatized using the spaCy library[5] to create three lemmatized datasets with:

- "Noun": identifies any class of people, places, things, or concepts.
- "Noun, Adj": adds to the previous one the adjective to describe nouns.
- "Noun, Adj, Verb, Adv": enriches the dataset "Noun and Adj" with verbs and adverbs to describe events and actions.

Before training the models, we removed the most common words (words appear in more than 20%, including stop words) and less common words (words appear in less than 200) from the datasets, yielding 1,826 distinct words for the "Noun" dataset, 2,499 for the "Noun, Adj" and 3,182 for the "Noun, Adj, Verb, Adv".

[5] Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

After the cleaning process, the average number of significant words was reduced to 75, 100, 125 respectively for the three data sets, whereas the distribution of word length does not change compared to that obtained on the raw data (Figure 3). We also removed abstracts that contain a small number of significant words (greater than or equal to 4).

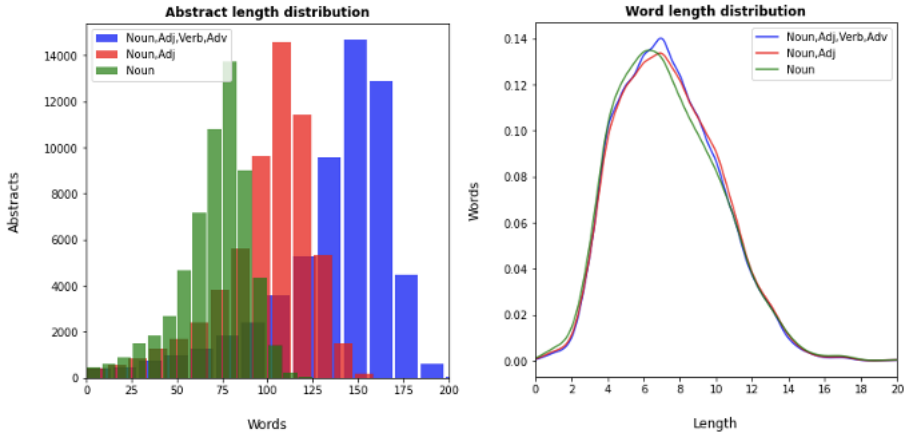


Figure 3 The histograms on the left shows the distribution of abstract length and the one on the right shows the distribution of word length for the three preprocessed dataset ("Noun", "Noun,Adj" and "Noun,Adj,Verb,Adv").

The language also consists of sequences of 2 or 3 individual words that provide a unique meaning, such as 'machine learning', 'internet of things'. The single word cannot convey the details properly, then we enriched the datasets with bigrams (2 consecutive words) and trigrams (3 consecutive words) using the Gensim topic modeling framework[6].

Table 5 shows the top 20 non-random combinations of two words that go together regularly based on statistical measure the pointwise mutual information[7]. Only bigrams with their pointwise mutual information score greater than 0.15 were accepted.

The LDA algorithm results depend on the input dataset, the number k of topics and the concentration parameters alpha and beta. The latter parameters are used to draw the probability distribution of the document on topics, Dir(alpha), and the probability distribution of a topic on words, Dir(beta). LDA uses these probability distributions to infer the words related to a given topic and the topics discussed in a given document. The best probability distributions are found during the algorithm training

[6] ŘEHŮŘEK, Radim and Petr SOJKA. Software Framework for Topic Modelling with Large Corpora. In Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. Valletta, Malta: University of Malta, 2010. p. 46--50. ISBN 2-9517408-6-7.

[7] T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean, Distributed representations of words and phrases and their compositionality, Neural information processing systems

process based on the Bayesian inference model which allows the model to improve as it continues to display new documents. Starting from this point, we adopted two different methodologies: 1) unsupervised approach and 2) semi-supervised approach. We describe the two methodologies below.

Unsupervised approach

To find the optimal parameters, we performed the grid search algorithm on 30% of the datasets (validation set), calculating the degree of semantic similarity between words within a topic and coherence score. The grid search exhaustively considers all parameter combinations in a subset of the dataset, k, alpha, and beta, and trains an LDA model using them as inputs (Table 6). Finally, it outputs the settings that achieved the highest score in the validation procedure. The highest coherence score was obtained by training a model with k=9, alpha="0.41", beta=0.8 and using the Noun dataset as input.

The best parameters were used to perform the LDA model, and the output is shown by using the interactive visualization pyLDAvis tool which produces a plot for better understanding and interpreting individual topics and the relationships between the topics.

Table 5 Top twenty bigram collocations generated by Gensim library on the three datasets with pointwise mutual information score upper than 0.15 and with frequency lower than 200 (https://arxiv.org/abs/1310.4546).

RANK	"NOUN"	"NOUN, ADJ"	"NOUN,ADJ,VERB, ADV"
1	feasibility study	feasibility study	feasibility study
2	state art	real time	real time
3	supply chain	business plan	business plan
4	feasibility assessment	business model	business model
5	value chain	end user	end user
6	health care	long term	cost effective
7	sme instrument	energy consumption	long term
8	wind turbine	state art	energy consumption
9	climate change	main objective	state art
10	return investment	large scale	main objective
11	proof concept	renewable energy	large scale
12	machine learning	raw material	renewable energy
13	internet thing	artificial intelligence	supply chain

14	greenhouse gas	supply chain	raw material
15	hardware software	big datum	artificial intelligence
16	side effect	easy use	big datum
17	fuel consumption	environmental impact	add value
18	decision support	clinical trial	environmental impact
19	value proposition	medical device	sme instrument

The tool shows the top most "relevant" terms for the specific topic. The parameter λ controls the relevance metric to highlight terms according to their probability or "distinctiveness" within the topic. For each model, for each topic, we divide the keywords according to their relevance[8] I.

Table 6 Subset of the parameters used by Grid Search and the best parameters in bold.

PARAMETER	RANGE
K	[1,2,3,4,5,6,7,8,**9**,10,11,12,13,14]
ALPHA	[0.01, 0.21, **0.41**, 0.6, 0.8, 1.0,"asymmetric","symmetric"]
BETA	[0.01, 0.21, 0.41, 0.6**, 0.8**, 1.0]
DATASET	[**"Noun","**Noun,Adj"," Noun,Adj,Verb,Adv"]

Let ϕkw denote the probability of term w ∈ {1, ..., V} for a topic k ∈ {1,..., K}, where V denotes the number of terms in the vocabulary, and let pw denotes the marginal probability of term w in the corpus, the relevance of term w to the topic k given a weight parameter λ (where $0 \leq \lambda \leq 1$) is defined as:

$$r(w,k|\lambda) = \lambda\, log(\phi_{kw}) + (1-\lambda)\, log\,(\frac{\phi_{kw}}{p_w}) \quad (1)$$

where λ determines the weight given to the probability of term w under topic k relative to its lift (measuring both on the log scale). λ=1 results in the familiar ranking of terms in decreasing their topic-specific probability; λ=0 ranks terms solely by their lift. We selected the top 10 keywords for

[8] Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63-70).

each $\lambda \in [0.0, 0.2, 0.4, 0.6, 0.8]$ to assign a label to each topic based on a global interpretation by looking for distinct and almost self-explained groups.

Semi-supervised approach

Together with the unsupervised approach, we modeled a hierarchical semi-supervised framework. This model aims to cluster, in a recursive way, the proposals in two hierarchical levels. We aimed to divide the proposals into macro-topic, and successively, in more fine-grained clusters based on the sub-topics at the second level. The difference between the two approaches explored lies in selecting the datasets and the topic modeling algorithm's input parameters. It is necessary to underline that the imposition of constraints contravenes the topic modeling algorithm's standard use. We choose a priori the dataset filtered based on semantically full lemmas, i.e., nouns, verbs, adverbs, and adjectives to perform the hierarchical model. The second constraint we impose concerns the range of the number of topics passed as input to the Latent Dirichlet Allocation algorithm, i.e., the parameter k. To introduce an adequate range, we performed a preliminary study on the full dataset. During this phase, we applied LDA with the number of topics $k \in [2,15]$ to obtain preliminary information on both the optimal number and the upper limit to be imposed later.

We note that with k=10, the most frequent keywords are disconnected from each other, and the topic's interpretation is challenging. For this reason, we evaluate that the best clustering is obtained with k=9. We focused on the nine topics by repeating the manual interpretation horizontally, searching for consistency of business models' orientation. In this phase, we assigned six main labels, i.e., digital_services, industrial applications, energy, recycling, health and food_production, which are further annotated with the application field. Based on these observations, we fix the value of k and we re-apply LDA to the dataset. Based on these observations, we set the value of k to 6 and perform Grid Search to find the best Alpha and Beta parameters (*Alpha* in the range [0.01,0.91] plus "asymmetric" and "symmetric"; *Beta:* range [0.01,0.91] plus "symmetric"). Then, we apply LDA to the dataset by forcing the clustering into 6 groups and setting the values *Alpha*'= 'a'ym' and *Beta* = 0.90. After obtaining the first level clusters, we applied Grid Search (k in the range [2,15]; *Alpha* in the range [0.01,0.91] plus "asymmetric" and "symmetric"; *Beta* in the range [0.01,0.91] plus "symmetric") to find the optimal number of subdivisions for each of the six clusters. Following this procedure, we obtained 6 groups of well-describable and separate proposals at the first level of clustering.

Moving on to the second level analysis, in most cases, Grid Search tends to divide each group into 6 additional clusters. This further subdivision allowed us to obtain very detailed and well-divided sub-

clusters. However, this methodology leads to results that are difficult to label generically. In smaller first-level clusters, the sub-clustering led to identifying groups composed of less than 10 proposals. In light of the results obtained, we believe that a too narrow division of the proposals cannot identify business models. Therefore, we have decided to follow the unsupervised approach.

Analytic description of the topics

As shown in Figure 4, the LDA analysis performed using the abstracts of the proposals in the dataset divides the proposals into 9 topics, which are different per size and positioning. Using the information retrieved by the most relevant keywords identified for each different grade of the relevance of **λ**, the team has performed a qualitative analysis of the contents to label the topics identified.
Two members of the team have interpreted the most **relevant** words that characterize each topic and labeled them as follows (detailed description of each topic will follow below):

1. Water Management.
2. Health, Diagnosis and Treatments.
3. Health, Rehabilitation and Medical Devices.
4. Digital solutions for cyber security and surveillance.
5. Recycling and Circular economy.
6. Digital solutions for transportation and mobility.
7. Food production, agriculture, fishing, and livestock.
8. Energy production, storage, and distribution.
9. Digital solutions for e-commerce, business platforms and content sharing.

An important dimension that emerged from the positioning of the topics on the plot and the consequent qualitative analysis of the contents is that the 9 topics are placed on the plot according to companies' business models. Indeed, as shown in Figure 4, topics are spread into the plot according to the business models applied by companies in implementing the inventions proposed.
The first area, placed on the upper central-left part of the plot, is the *digital platform based* and grouped Topic 4, 6, and 9. In this area are mainly grouped the inventions dedicated to the digital arena. Companies operating in this area propose inventions for both companies and end-users needs. The proposals intend to introduce innovations as digital applications that can be integrated into existing digital environments such as platforms or clouds. Accordingly, in this part of the plot are placed proposals regarding digital solutions for cyber security and surveillance, transportation and mobility, e-commerce, business platforms, and content sharing.

The second identified area is placed at the bottom of the left side of the plot, and it collects *manufacturing-oriented* proposals, and more specific proposals of topics 5 and 7. Companies involved in this area mainly deal with implementing innovation and inventions both for products and processes at the industrial level. Proposals in this area respond to a B2B logic and address manufacturing companies' needs to introduce innovations in the value chain. Here are placed inventions regarding food production and the circular economy.

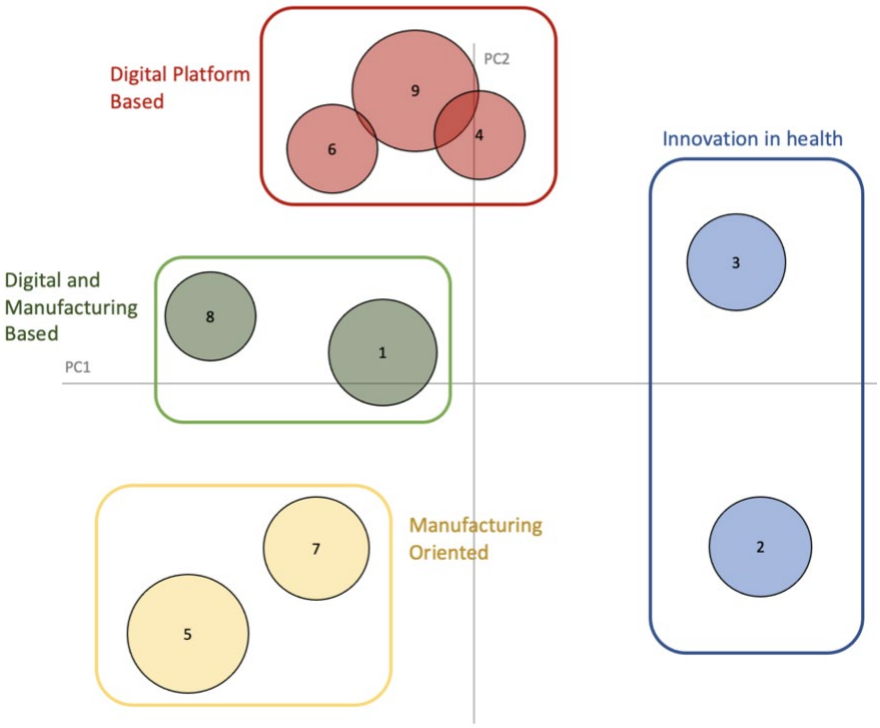


Figure 4 Plot of the topics

It is important to notice that proposals in energy production and water management (Topic 1 and 8) represent a *trait-d'union* between the two main areas identified because the proposed inventions in these two topics show similar distinctive peculiarities. Indeed, on the one hand they propose innovation at the industrial process level, but on the other hand they suggest the implementation of digital solutions operating in digital environments for companies and end-users.

The third area, on the right part of the plot, regards *innovation in health* both at diagnostic and biomedical level. Companies that are positioned in this field respond to peculiar logics according to the inner characteristics of the medical sector. Invention and innovative solutions proposed are

intended for companies, operators such as doctors and nurses and end-users intended as patients and their families.

In the following sub-sections, we illustrate a more detailed description of each topic identified. More specifically, we organized the structure of each sub-section as follows. First, we report the most frequent keywords per each degree of relevance (**λ**). Second, we provide a description of S'Es' innovative solutions, their finalities, and the industrial sectors they are meant for.

Topic 1: Water Management

As shown in Table 7 from a first analysis of the most frequent ten keywords (according to their level of relevance), it seems that the cluster is characterized only by technical terms regarding procedural aspects of the proposals. Thus, to offer a more comprehensive understanding of the topic's nature, we report in Table 8, the keywords from 11 to 20.

According to Table 8, it emerges that Topic 1 groups all the proposals regarding water management. This sector represents a very exclusive niche in which innovation is conceived to offer better and more rationalized water use in agriculture and public use. Proposals in this topic aim at offering, on the one hand, innovative solutions for farmers to operationalize in terms of digitalization of their irrigation systems connecting them to digital applications that offer information regarding the water and the rain. On the other hand, they provide innovative solutions to local communities for public water management in terms of drinking water to avoid leakage, waste of water and energy, and wastewater to prevent floods and contaminations.

Topic 2: Health, Diagnosis and Treatments

Topic 2 is characterized by the keywords listed in Table 9, and it groups the proposals regarding Health, Diagnosis, and Treatments field. The cluster mainly deals with the health sector, and more specifically, cancer diagnosis and treatment. It is well known that the health sector, especially all the activities related to the oncology branch, represents a significant field of interest for the application of innovative solutions in terms of diagnosis and treatment of the disease.

Companies located in the topic mainly deal with proposals tailored explicitly to the health sector that needs integrating innovation in both processes and products with the application of platform-based solutions for information handling and sharing and data storage. Within this vein, these companies are equidistant in terms of business and market orientation from both groups of the sample, the companies web and platform-based, and the companies' processes and products innovation-oriented.

Topic 3: Health, Rehabilitation and Medical Devices

Health, Rehabilitation and Medical Devices topic mainly deals with the healthcare sector, specifically to the innovation proposed in the field of rehabilitation and medical devices, as shown by the keyword listed in Table 10.

As for Topic 2, this sector represents a significant field of interest for the application of innovative solutions in terms of patient care and smart medical devices for the cure and the rehabilitation of people injured or affected by chronic diseases. Moreover, companies mainly deal with proposals tailored explicitly to the health sector's needs integrating innovation in both processes and products with the application of platform-based solutions for information handling and sharing and data storage. Within this vein, these companies are equidistant in terms of business and market orientation from both groups of the sample, the companies web and platform-based, and the companies' processes and products innovation-oriented.

Topic 4: Digital solutions for cyber security and surveillance

Topic 4, as shown by the list of keywords in Table 11, is characterized by proposals about innovative digital solutions for cybersecurity and surveillance. Proposals in the topic promote the digital platforms' creation to improve the extensive use of the internet, apps, and mobile related solutions for surveillance and cybersecurity. For instance, the topic points to the creation of platforms to use apps and mobiles to allow consumers, customers, and users to share confidential data and information for sensitive activities such as financial and banking online operations. The topic comprises the activities strictly related to data and cybersecurity. These tools aim to increase protection regarding sensitive data storage against cyber-attacks. Indeed, cybersecurity is tightly linked to data security, and these solutions aim to secure people's and companies' assets by providing an increased level of protection for digital transaction systems against cyber-attacks. These tools focus on improving cyber-operators' detection and protection capabilities against threats, frauds, and incidents. Moreover, the inventions proposed in the topic suggest using remoting controlling system from surveillance cameras and drones for the surveillance of building, sensitive targets as airports, ports, and railway stations.

Similarly, to Topic 6 and 9, it is characterized by the presence of small and medium companies oriented to develop software, apps, and mobile tools to operate in online platforms, to build the application of digital technologies built on data, software, and cloud-based platforms to speed up the delivery of the commercial and financial products and services. The topic shows that companies implement their activities by relying on the presence of broad ecosystems of partners. They establish technological and service collaborations to secure the highest level of integration and interoperability of their products and services on digital marketplaces.

Topic 5: Recycling and Circular economy

As shown in Table 12, Topic 5 groups proposals regarding inventions in recycling and circular economy. The topic mainly deals with innovative industrial applications that impact the recycling sector in promoting the circular economy for environmental protection. The proposed industrial solutions aim to reduce the traditional industry material's footprint, such as rubber, plastic, and chemical products. The topic aims at promoting innovative solutions both at the products and processes level, on the one hand, to offer more smart and integrated solutions in the waste management value chain, and on the other hand, to push the production of innovative recyclable materials.

Companies positioned in Topic 5 encourage the implementation of innovative solutions and smart industrial applications in the waste management sector, covering not only the end of the products' life but also promoting innovative solutions in product materials such as biodegradable packaging. All in all, the waste management sector is characterized by high demand for innovative solutions and industrial implementations to maximize the use of the raw materials to optimize the waste management process and to reduce the waste footprint augmenting the percentage of recycled materials and to promote smart solutions for the circular economy. As for Topic 7, it is characterized by the presence of small and medium companies oriented to developing practical industrial applications to integrate into existing value chains or to implement new production lines to offer the integration of smart solutions.

Topic 6: Digital solutions for transportation and mobility

Topic 6, labeled Digital solutions for transportation and mobility, deals with digital solutions concerning urban and transportation needs, safety and security issues, smart cities, and travelers' necessities. As shown in Table 13, the topic aims to develop digital integrated solutions and tools for platforms that concern urban transportations, sensors for traffic monitoring, solutions for smart cities, parking services, and transportation safety and security to reduce incidents. Some tools aim to implement public transportation use by suggesting the implementation of digital applications and platforms for promoting best practices, commercial road transportations, collecting data, sharing information, and promoting a new driving behavior.

Topic 6, similarly to Topic 4 and 9, is characterized by the presence of small and medium companies oriented to develop software, apps, and mobile tools to operate in online platforms. Also, proposals are based on the application of digital technologies built on data, software, and cloud-based platforms to improve the urban living conditions in terms of traffic and mobility, transportation safety and

security and to monitor habits and behavior to promote best practices and alternative solutions. These companies deal with data collection regarding people's behavior and sensitive information storage. Topic 6 shows that companies implement their activities relying on the presence of broad ecosystems of partners, with which they establish technological and service collaborations to secure the highest level of integration and interoperability of their services in the digital environment. Moreover, these companies rely on an integrated network of data collections from different sources, such as sensors and access to people's smartphone information.

Topic 7: Food production, agriculture, fishing, and livestock

Food production, agriculture, fishing, and livestock topic mainly deals with the food production sector, agriculture, fishing, and livestock, as shown by the keywords in Table 14. The proposed industrial solutions aim at innovating the industrial processes related to food production, agricultural techniques such as the irrigation of the fields, the spread of pest control products, the seeding, and the livestock management. The topic aims at promoting innovative solutions both at the processes level introducing innovation and digitalization in a very traditional sector.

Companies positioned in Topic 7 are close to those located in Topic 5. They encourage the implementation of innovative solutions and smart applications in the food production value chain by promoting integrated management of the production chain, reducing risks for public health, and protecting economic operators from unnecessary losses. The implementation of smart solutions in this sector is mainly regarding the processes and the systems' control and the manufacturing processes involved in transforming the raw materials (fish, vegetables, and meat). It is characterized by small and medium companies oriented to develop practical industrial applications to integrate into existing production chains or productive systems that offer the integration of smart solutions.

Topic 8: Energy production, storage, and distribution

Energy production, storage and distribution topic mainly deals with industrial innovations in the energy sector. The topic (Table 15) shows peculiarities regarding the applications of the proposed inventions that are placed between the manufacturing-oriented and the platform-based topics. Indeed, closely to Topic 5 and 7, it encourages the implementation of innovative tools and smart industrial applications in the energy sector in terms of production, distribution, and consumption. The proposed industrial solutions regard both traditional and sustainable sources of energy production, as demonstrated by the presence of words like wind, solar, heat, renewable, gas, offshore, fuel, among others. The applications of this topic regard both traditional and renewable sources of energy. Indeed, solutions related to energy sectors both for conventional or renewable sources are characterized by

the continuous implementation of innovation and new inventions in industrial processes to reduce the impact on the environment and climate change. On the other hand, closely to Topic 4, 6, and 9, it aims to propose implementing digital applications to control energy consumption, monitor the distribution grids, and rationalize energy storage.

All in all, the energy sector is characterized by high demand for innovative solutions and industrial implementations to maximize the use of energy sources, optimize distribution channels, and rationalize consumption. It is characterized by small and medium companies oriented to develop practical industrial applications to integrate into existing value chains for the digital transformation or to implement in new production lines to offer the integration of smart solutions. However, also to develop digital solutions to integrate into platform-based environments.

Topic 9: Digital solutions for e-commerce, business platforms and content sharing

Digital solutions for e-commerce, business platforms, and content sharing topic deals with the idea of incrementing digital services for daily commercial and financial activities, on apps and mobile devices. The fields concerned are the digital applications of mobile-based integrated solutions in business activities such as e-commerce apps and websites, digital marketing tools for the consumers and retailers, and business instruments for the digitalization of payments and financial transactions.

Topic 9 (Table 16) promotes the digital platforms' creation to improve the extensive use of the internet, apps, and mobile related solutions for daily life activities. For instance, the topic points to the creation of platforms to use apps and mobiles on the one hand to allow consumers, customers, and users to buy or perform financial and banking activities online; on the other hand, to develop networks for data collection, monitoring, decision-making, and process optimization available for retailers and banks for marketing purpose.

Moreover, it is characterized by proposals about innovative digital solutions for data sharing and information processing. The digital services and solutions proposed in this topic are mainly conceived for education gaming and content sharing. It focuses on digital services and solutions offered to students in online education, learning, and gaming. These solutions generally aim to improve technologies in daily life activities, digital learning, online gaming platforms, education content sharing, and student support.

Topic 9 promotes digital platforms' creation to collect and share data, making them available to end-users. On the one hand, these implementations aim to increase public services efficiency, such as e-learning and encourage sharing common experiences in learning, and gaming, mainly among students

and young people. On the other hand, these solutions aim to facilitate a student's learning experience, connect people, and help users share content.

Topic 9 is characterized by the presence of small and medium companies oriented to develop software, apps, and mobile tools to operate in online platforms, to build the application of digital technologies built on data, software, and cloud-based platforms to speed up the delivery of the commercial and financial products and services. These companies also deal with digital solutions to collect data regarding consumer behavior to create sources for promotion, valuation, and monetization of innovative solutions in crucial sectors for daily based activities. Topic 9 shows that companies implement their activities by relying on the presence of broad ecosystems of partners, with which they establish technological and service collaborations to secure the highest level of integration and interoperability of their products and services on digital marketplaces.

Table 7 Topic 1 – List of the most relevant keywords

λ	**Keywords**									
0.0	Drinking	Drinking_water	Precision agriculture	Irrigation	Viability	Leakage	Soil mosture	Flood	Risk mitigation	Ipr
0.2	Viability	Irrigation	Ipr	Drinking	Drinking_water	Wine	Leakage	Flood	Precision agriculture	Conduct feasibility study
0.4	Business	Market	Viability	Water	Innovation	Development	Irrigation	Commercialization	Wine	Ipr
0.6	Market	Businesses	Project	Viability	Water	Innovation	Product	System	Commercialization	Feasibility
0.8	Feasibility	Market	Study	Business	Plan	Feasibility study	Phase	Project	Strategy	Solution
1	Market	Feasibility	Study	Business	Phase	Project	Plan	Feasibility study	Solution	Product

Table 8 Topic–1 - keywords from 11 to 20

λ	**Keywords**								
0.0	Study	feasibilityfeasibility_study	viability	feasibility assessment	sme_instrument	irrigation	phase_sme_instrument	drinking	drinking_water
0.2	feasibility	study	phase	plan	businefeasibility_study	strategy	assessment	viability	instrument
0.4	feasibility	studyfeasibility_study	market	business	phase	plan	project	strategy	assessment
0.6	market	feasibility	study	business	phase	plan	projectfeasibility_study	solution	system
0.8	market	feasibility	study	business	project	phase	planfeasibility_study	solution	product

1	market	feasibility	study	business	project	phase	planfeasibility_stud y	solution	product

Table 9 Topic–2 - List of the most relevant keywords

λ	**Keywords**									
0.0	Cancer	Drug	Biomarker	Tissue	Vaccine	Efficacy	Tumor	Side effect	Wound	Antibody
0.2	Treatment	Cancer	Cell	Drug	Biomarker	Disease	Therapy	Trial	Diagnosis	Blood
0.4	Treatment	Cancer	Patient	Cell	Disease	Drug	Test	Therapy	Biomarker	Blood
0.6	Treatment	Cancer	Cell	Drug	Patient	Disease	Test	Biomarker	Therapy	Trial
0.8	Treatment	Patient	Cancer	Cell	Drug	Disease	Test	Therapy	Market	Project
1	Treatment	Patient	Cancer	Cell	Disease	Drug	Market	Test	Project	Year

Table 10 Topic–3 - List of the most relevant keywords

λ	**Keywords**									
0.0	Child	Rehabilitation	Doctor	Disability	Parent	Sleep	Baby	Caregiver	Exercise	Obesity
0.2	Health	People	Patient	Care	Hospital	Healthcare	Population	Heart	Injury	Disorder
0.4	health	patient	people	care	device	life	child	hospital	healtcare	population
0.6	health	People	Patient	care	device	system	year	life	child	disease
0.8	health	patient	people	care	device	life	system	solution	market	hospital
1	health	patient	people	care	device	system	solution	market	year	life

Table 11 Topic–4 - List of the most relevant keywords

λ	**Keywords**									
0.0	security	threat	attack	camera	resolution	surveillance	radio	authentication	cyber	crime
0.2	device	security	sensor	detection	image	measurement	hardware	threat	camera	smartphone
0.4	system	device	technology	application	sensor	software	datum	detection	communication	analysis
0.6	device	system	technology	security	market	solution	sensor	application	datum	software
0.8	device	system	technology	security	market	solution	sensor	application	datum	software
1	device	system	technology	security	market	solution	sensor	application	datum	software

Table 12 Topic–5 - List of the most relevant keywords

λ	**Keywords**									
0.0	metal	turbine	recycling	wood	wind_turbine	biogas	coating	wastewater	steel	disposal
0.2	material	waste	water	plant	oil	heat	gas	construction	chemical	plastic
0.4	material	technology	production	process	waste	plant	water	gas	emission	fuel
0.6	material	technology	production	process	water	waste	plant	market	cost	project
0.8	material	production	technology	process	waste	market	plant	cost	industry	energy

1	material	technology	production	process	water	waste	market	project	plant	cost

Table 13 Topic–6 - List of the most relevant keywords

λ	**Keywords**									
0.0	vehicle	car	road	satellite	parking	ship	passenger	aircraft	ship	drone
0.2	vehicle	transport	car	road	traffic	driver	transportation	mobility	satellite	accident
0.4	vehicle	transport	safety	car	road	city	maintenance	space	driver	engine
0.6	vehicle	system	solution	time	transport	car	safety	road	city	maintenance
0.8	system	vehicle	transport	car	solution	time	market	cost	service	safety
1	system	vehicle	transport	car	solution	time	market	cost	service	safety

Table 14 Topic–7 - List of the most relevant keywords

λ	**Keywords**									
0.0	food	packaging	fish	robot	ingredient	shelf	meat	milk	traceability	fruit
0.2	product	production	food	consumer	chain	packaging	producer	supply-chain	value_chain	fish
0.4	product	production	food	industry	market	chain	quality	consumer	machine	manufacturing
0.6	product	food	production	market	industry	chain	technology	consumer	quality	process
0.8	product	market	food	production	industry	technology	quality	chain	consumer	process
1	product	market	food	production	industry	technology	quality	consumer	chain	solution

Table 15 Topic–8 - List of the most relevant keywords

λ	**Keywords**									
0.0	energy	power	building	electricity	building	storage	battery	grid	lighting	utility
0.2	energy	power	building	consumption	electricity	battery	storage	grid	utility	installation
0.4	energy	system	power	building	cost	electricity	storage	consumption	battery	efficiency
0.6	energy	system	power	building	consumption	cost	solution	market	electricity	efficiency
0.8	energy	system	power	building	consumption	solution	market	cost	efficiency	storage
1	energy	system	power	market	building	solution	cost	consumption	electricity	technology

Table 16 Topic–9 - List of the most relevant keywords

λ	**Keywords**									
0.0	language	payment	student	advertising	transaction	music	bank	audience	website	credit
0.2	service	platform	datum	user	information	content	web	app	enterprise	language
0.4	service	platform	business	company	datum	customer	model	information	tool	software
0.6	service	platform	business	market	user	company	datum	customer	solution	model
0.8	service	platform	market	business	user	company	datum	customer	solution	project
1	service	platform	market	business	user	company	datum	solution	customer	project

Business Models Analysis

Proposals were assigned to the most likely business model based on the multinomial distribution on topic (business models) produced by LDA. Digital solutions for e-commerce and the circular economy are the most used business models in the proposals as shown in Figure 5.

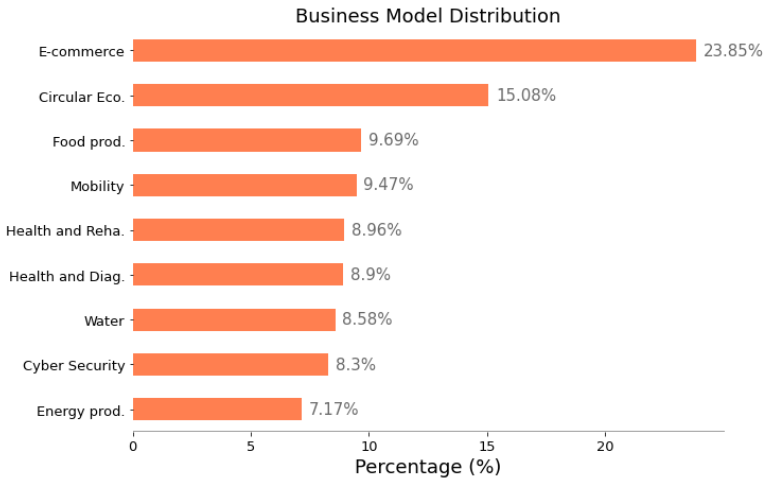


Figure 5 Business model frequency distribution (P(bm))

The proposals, evaluated by the H2020 SME commission, were classified into:

- **Main list**: proposals that have received funding
- **Below the available budget (**reserve list), if the available budget is too small to fund all proposals that have achieved the qualification score in the evaluation cycle, some proposals may be placed on a reserve list
- **Below the threshold**: proposals that receive an evaluation score below the threshold
- **Inadmissible/inadmissible/withdrawn:** inadmissible proposals

We have grouped the 58.900 proposals into two categories: accepted (Main and reserve list) and rejected (below threshold), the inadmissible proposals were discarded.

We can estimate the probability that a proposal will be accepted at 28% (p (A) in Figure 5)) and at 72% the probability that it will be rejected.

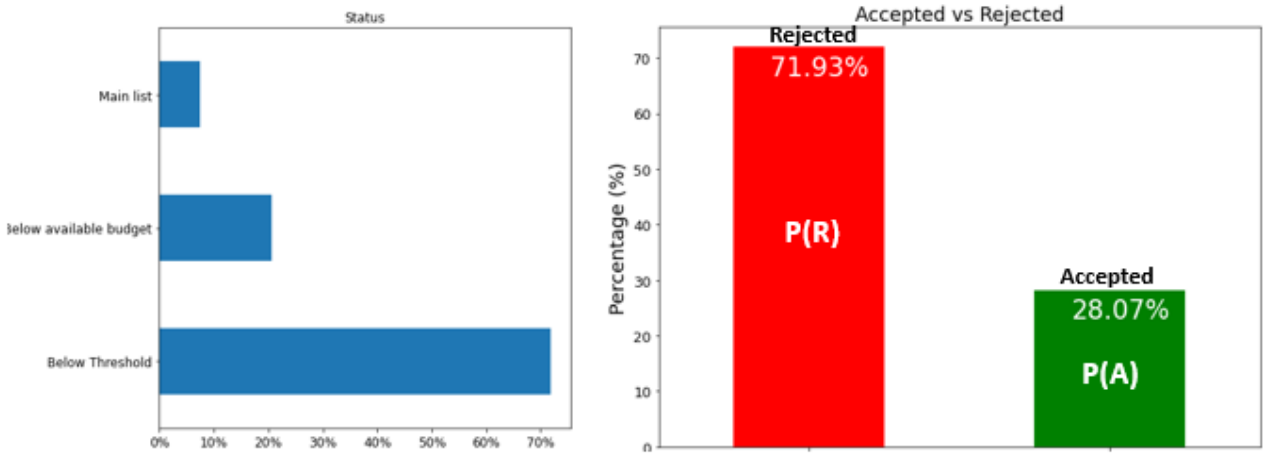


Figure 6 Proposals Evaluation Results Distribution

Figure 7 shows how the distribution of business models on rejected and accepted proposals changes. We can see that if the proposal was accepted it is more likely that a business model of e-commerce, circular economy, health or food production was used. On the other hand, if the proposal is rejected, the most likely business models are e-commerce, circular economy, mobility or water management.

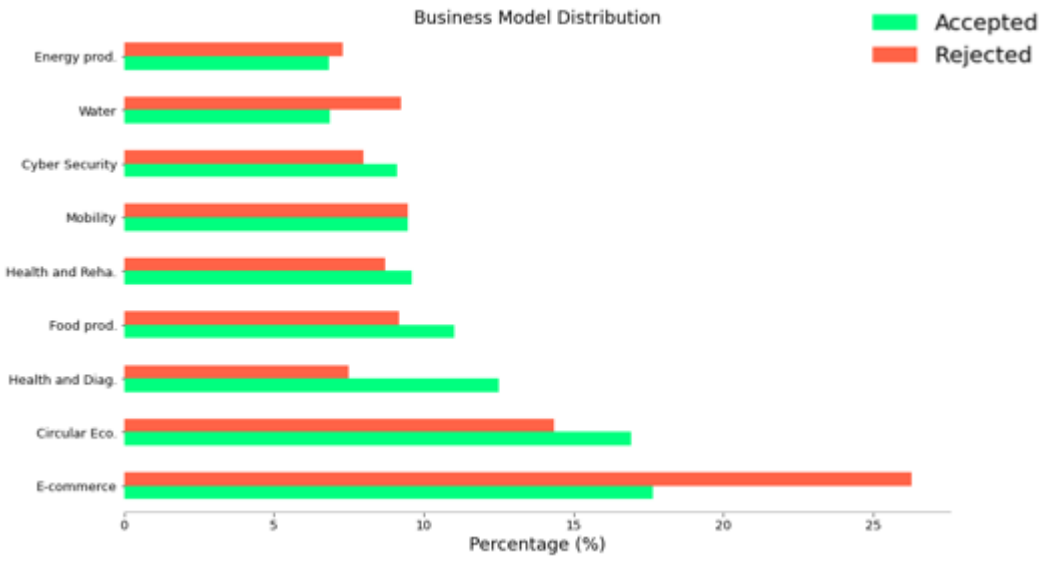


Figure 7 Business model frequency distribution in accepted proposals (p(bm|A), the green bars) and in Rejected Proposals (p(bm|A), the red bars)

The distribution above can be interpreted as the conditional probability of whether the proposal has been accepted what is the probability that a specific business model will be used (p (bm | A)) or if the proposal has been rejected what is the probability that a business specific model (p (bm |

R)) is used. But we are interested in answering the inverse question, given a specific business model what the probability of success (p (A | bm)) is?

We calculated Bayes' formula to "invert" the conditional probabilities and answer this question p (A | bm) = p (A) p (bm | A) / p (bm). The results, in Figure 8, show that using the Health and Diagnosis business model for the proposal is nearly 20% more likely to be accepted than the e-commerce model.

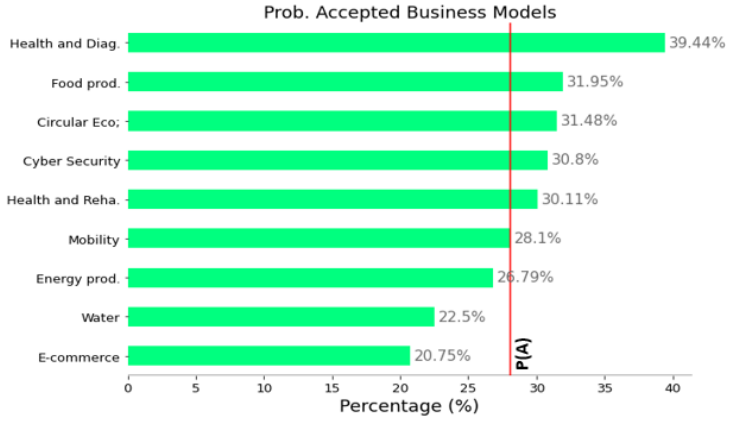


Figure 8 Probability of success of business models. The vertical line shows the probability of being accepted

The next analysis shows when the relevant words highlighted by the LDA model to identify the business models (Figure 9) can influence the evaluation of the proposal. Also, in this case we used the bayes theorem:

$$P(A|Wr,t) = \frac{P(Wr,t|A)\,P(A)}{P(Wr)} = \frac{P(Wr,t|A)\,P(A)}{P(Wr,t|A)\,P(A) + P(Wr,t|R)\,P(R)} \quad (1)$$

$$P(R|Wr,t) = \frac{P(Wr,t|R)\,P(R)}{P(Wr,t)} = \frac{P(Wr,t|R)\,P(R)}{P(Wr,t|R)\,P(R) + P(Wr,t|A)\,P(A)} \quad (2)$$

Wr,t is a relevant word in Figure 9 for a specific topic $t \in \{Topic\ 1-9\}$

P(Wr,t | A) is the probability that an accepted document contains Wr,t ($Wr \in table\ 3-12\ and\ t \in \{Topic\ 1-9\}$)

P(Wr,t | not A) or *P(Wr,t | R)* is the probability that an not accepted document contains Wr,t ($Wr \in table\ 3-12\ and\ t \in \{Topic\ 1-9\}$)

P(A) is the probability that a document will be accepted

P(not A) or P*(R)* is the probability that a document is not accepted (rejected)

P(A|Wr,t) is the probability that an a word Wr of the topic t is contained in the accepted proposals and *P(R|Wr,t)* is the probability that a word Wr of the topic t is contained in the rejected proposals. This probability was calculated for each topic using a Baysian classifier trained on proposals related to topics in order to determine which of the relevant words for the business models negatively or positively influence the evaluation.

The bar charts in Figure 9 shows for each topic the ratios are known as likelihood ratios:

$$\frac{P(A|Wr,t)}{P(R|Wr,t)} \; if \; P(A|Wr,t) \geq P(R|Wr,t) \; or \; \frac{P(R|Wr,t)}{P(A|Wr,t)} \; if \; P(A|Wr,t) < P(R|Wr,t)$$

they can be useful to compare different feature-outcome relationships and determine which features it found most effective for distinguishing the accepted or rejected proposals.

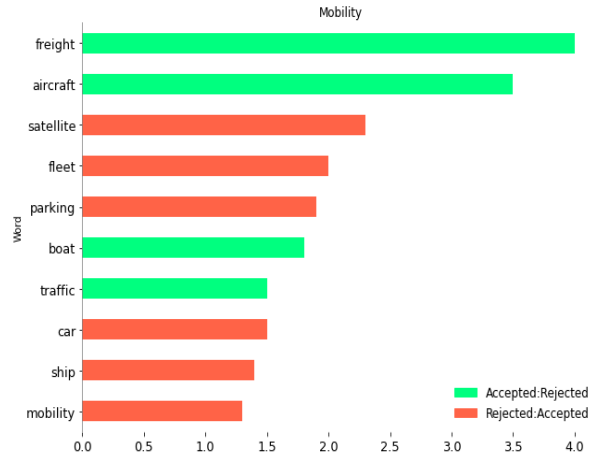
In the Digital solutions for transport and mobility business model the word " freight " occurs 4 times more often in accepted proposals than in rejected ones. In the Health, Diagnosis and Treatment business model, the word "infection" appears 3.5 times more in rejected proposals than accepted ones. In the Health, Diagnosis and Treatments business model, the word "infection" appears 3.5 times more in rejected proposals than accepted ones, ‘side effects’ are important characteristics to accept.

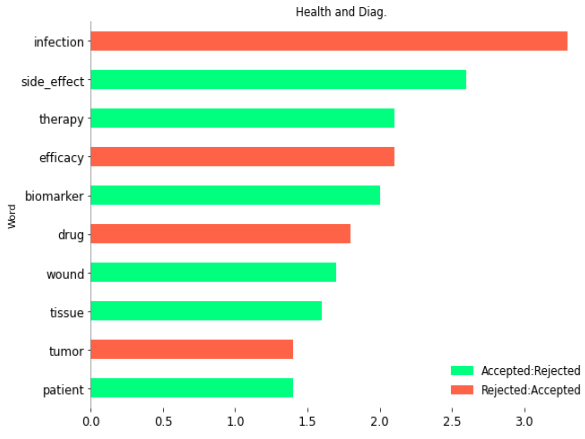
"Sludge" and "wood" are more effective to be accepted in recycling and circular economy than the "turbine" which is 4 times in rejected proposals than accepted ones.

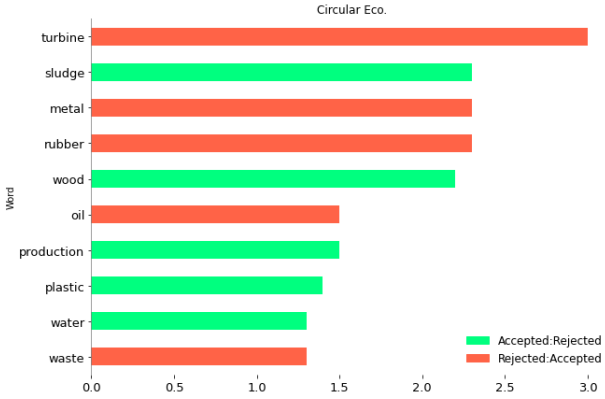
"Teacher" appears 7 times more often in accepted digital solutions for e-commerce, corporate platforms and content sharing.

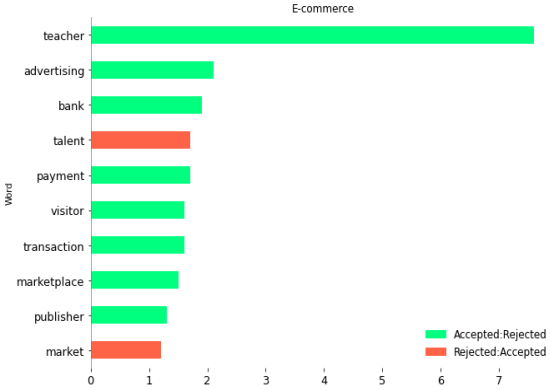
"Voltage" and "grid" are important for the business model Energy production, storage and distribution, "Drinking water" for Water Management, and "supply chain" and "fish" for Food production, agriculture, fishing and livestock,

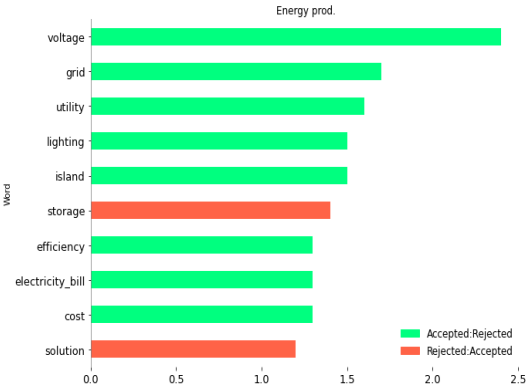
In Health, Rehabilitation and Medical Devices, the words "parent" and "child" negatively affect the decision compared to "child" and "hospital" which positively affect the decision. Finally, "authentication" and "breach" are more often in accepted and "treatment" and "camera" are more often in rejected proposals for Digital solutions for cyber security and surveillance.

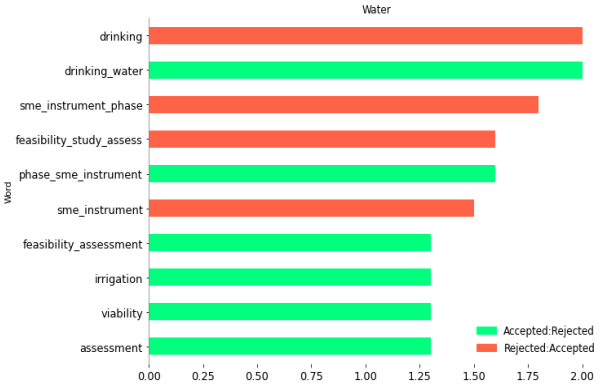

Mobility
freight
aircraft
satellite
fleet
parking
boat
traffic
car
ship
mobility
Word
Accepted:Rejected
Rejected:Accepted
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

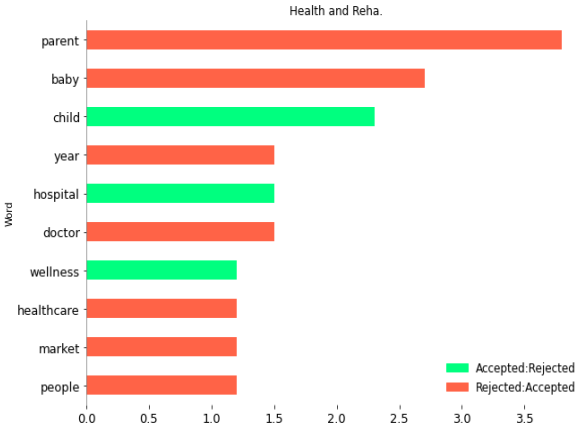

Health and Diag.
infection
side_effect
therapy
efficacy
biomarker
drug
wound
tissue
tumor
patient
Word
Accepted:Rejected
Rejected:Accepted
0.0
0.5
1.0
1.5
2.0
2.5
3.0

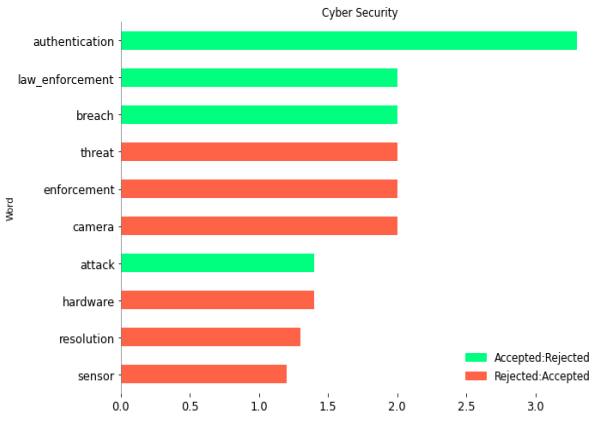

Circular Eco.
turbine
sludge
metal
rubber
wood
oil
production
plastic
water
waste
Word
Accepted:Rejected
Rejected:Accepted
0.0
0.5
1.0
1.5
2.0
2.5
3.0


E-commerce
teacher
advertising
bank
talent
payment
visitor
transaction
marketplace
publisher
market
Word
Accepted:Rejected
Rejected:Accepted
0
1
2
3
4
5
6
7


Energy prod.
voltage
grid
utility
lighting
island
storage
efficiency
electricity_bill
cost
solution
Word
Accepted:Rejected
Rejected:Accepted
0.0
0.5
1.0
1.5
2.0
2.5


Water
drinking
drinking_water
sme_instrument_phase
feasibility_study_assess
phase_sme_instrument
sme_instrument
feasibility_assessment
irrigation
viability
assessment
Word
Accepted:Rejected
Rejected:Accepted
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00


Health and Reha.
parent
baby
child
year
hospital
doctor
wellness
healthcare
market
people
Word
Accepted:Rejected
Rejected:Accepted
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5


Cyber Security
authentication
law_enforcement
breach
threat
enforcement
camera
attack
hardware
resolution
sensor
Word
Accepted:Rejected
Rejected:Accepted
0.0
0.5
1.0
1.5
2.0
2.5
3.0

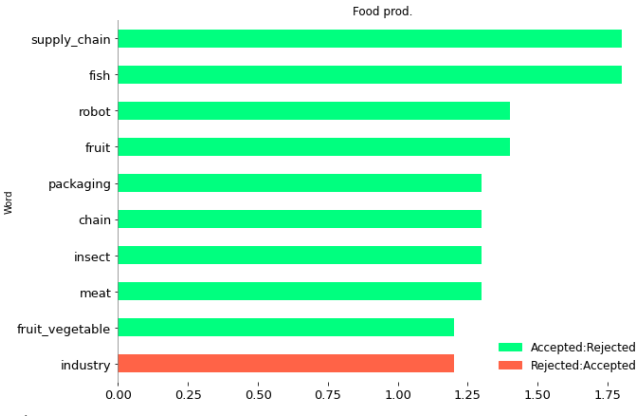


Figure 9 Bar charts shows for each topic the likelihood ratios between P(A|Wr,t)/P(R|Wr,t) in green and P(R|Wr,t)/P(A|Wr,t) in red.

Task 1.3 Firms' technological capabilities identification (patents)

Following the methodology implemented in Section "*Business models identification"* we conducted the analysis of the SMEs' patent portfolios. The patents allow us to identify the most relevant sectors in which European SMEs are directing their innovation efforts.

Analytic description of the patents' topics

As shown in Figure 10, the LDA analysis performed using the abstracts of the patents of the companies of our dataset divides the proposals into 6 topics which are different per size and positioning. Using the information retrieved by the most relevant keywords identified for each different grade of the relevance of **λ**, the team has been able to perform a qualitative analysis of the contents to label the topics identified.

Two members of the team have interpreted the most frequent words that characterize each topic and labeled them as follows (detailed description of each topic will follow below):

1. Semiconductors, lasers, and radiations.
2. Cancer diagnostic and treatment.
3. Information and data collection, management, and storage.
4. Energy, fluids, and gases containment, control and distribution.
5. Chemicals applications in waste and production processes.
6. Housing construction materials and tools.

A consequent qualitative analysis conducted on the contents of the six topics according to their positioning on the plot mainly emerged three main areas of innovation. As shown in Figure 10, the first area, placed on the lower-right part of the plot, can be identified as the one grouping inventions regarding semiconductors, tools, and hardware for information handling. This

information might be in the form of data or images. The latter element justifies the proximity of Topic 1 with Topic 3 since the former encompasses inventions in terms of semiconductors, lasers, and radiations also implemented for scanning and imagining. On the other hand, Topic 3 groups information system technologies that deal with data and servers.

The second area, the one placed in the upper-right part of the plot, groups Topic 4 and Topic 6. These two topics mainly group inventions related to manufacturing processes. The innovations proposed in the two topics deal with industrial applications and the use of raw materials. On the one hand, Topic 4 comprises inventions related to energy, fluids, and gases and their distribution and control structures. On the other hand, Topic 6 promotes inventions in terms of construction, housing-building, and manufacturing machinery.

The left side of the plot is occupied by two separate topics, which for different reasons are related to chemical applications. Indeed, the lower part plots Topic 2, which is related to chemical compositions and pharmaceutical applications for cancer diagnostic and treatment. Conversely, on the upper part, Topic 5 covers chemical innovations for production processes and waste management.

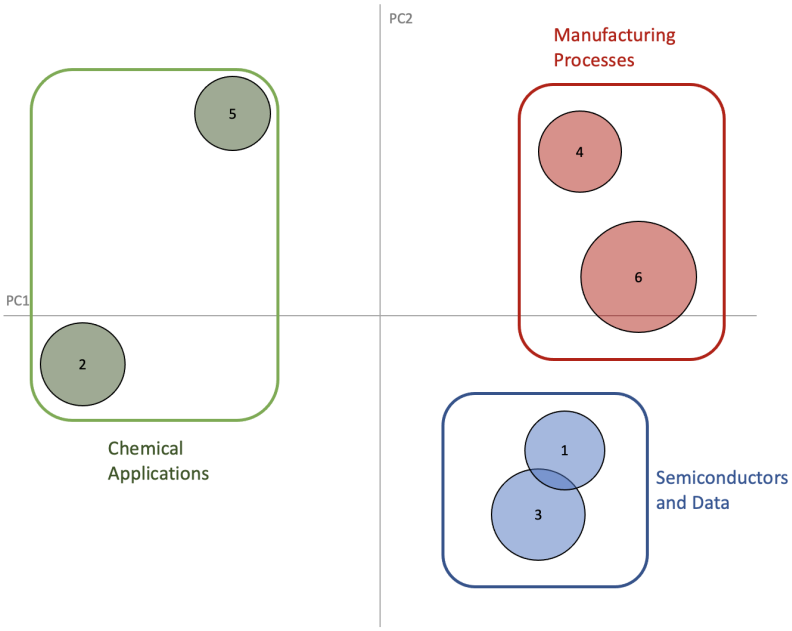


Figure 10 Plot of the p'tents' topics

In the following sub-sections, we illustrate a more detailed description of each topic identified. More specifically, we organized the structure of each sub-section as follows. First, we report the

most frequent keywords per each degree of relevance (λ). Second, we provide a description of the patents of the SMEs, their finalities and the industrial sectors they are meant for.

Topic 1: Semiconductors, lasers, and radiations

Topic 1 is characterized by the most frequent ten keywords (according to their relevance level) listed in Table 17. SMEs' patents that are grouped in this topic mainly regard inventions and technological applications in the field of radiations, lasers, and semiconductors. On the one hand, these technologies mainly regard industrial applications in the field of electrical components for sensors; on the other hand, these technologies refer to imaging techniques that use lasers and radiations.

Topic 2: Cancer diagnostic and treatment

Topic 2 mainly deals with the cancer treatment issue. Most frequent keywords are presented in Table 18 and mainly regard innovations important for detecting and treating cancer. On the one hand, patents of this topic introduce pharmaceutical compositions, formulae, and kits to detect cancer in individuals. On the other, the patents grouped in this topic also refer to the therapy and treatment of this specific disease.

Topic 3: Information and data collection, management, and storage

Topic 3, which mainly groups SMEs' patents that present in their abstracts the most frequent keywords of Table 19, inventions related to data collection, management, and storage. Patents clustered in this topic regard innovations in data handling systems, data storage (memories), graphical interfaces, data storage solutions, and data analysis processes. These inventions usually regard user involvement in the innovation processes of the companies.

Topic 4: Energy, fluids and gases containment, control and distribution

Topic 4 groups inventions and devices for the containment and distribution of energy, fluids and gases. This topic is mainly characterized by the most frequent words listed in Table 20. The technologies included to the topic are also related with the adequate tools for the distribution networks control, especially regarding the supply chain side.

Topic 5: Chemicals applications in waste and production processes

Topic 5 mainly encompasses chemical innovations. Most frequent keywords are presented in Table 21 and regard chemical applications for products and industrial processes. The proposed technologies deal with a wide plethora of applications from food production and conservation, waste management and treatment, and surfaces' treatment (e.g., coating).

Topic 6: Housing construction materials and tools

Topic 6 is featured by the most frequent ten keywords (according to their level of relevance) listed in Table 22. Topic 6 mainly groups inventions related to construction materials and tools. Patents clustered in this topic mainly cover innovations in terms of tools and materials utilized in the construction sectors and dedicated to the building. In fact, tools that are important to assembly, support the surfaces, or housing construction structure mainly characterize this topic.

Table 17 Topic 1 - List of the most relevant keywords

λ	**Keywords**									
0.0	Wavelength	Image	Substrate	Electrode	Light	Detector	Radiation	Laser	Semiconductor	Ray
0.2	Semiconductor	Image	Substrate	Electrode	Light	Detector	Radiation	Surface	Laser	Semiconductor
0.4	Film	Image	Substrate	Electrode	Light	Camera	Detector	Radiation	Laser	Conductor
0.6	Camera	Image	Film	Substrate	Electrode	Light	Detector	Radiation	Laser	Beam
0.8	Film	Image	Display	Substrate	Electrode	Light	Detector	Radiation	Laser	Beam
1	Sensor	Image	Film	Substrate	Electrode	Light	Detector	Radiation	Beam	Laser

Table 18 Topic 2 - List of the most relevant keywords

λ	**Keywords**									
0.0	Cancer	Protein	Formula	Disease	Antibody	Prevention	Kit	Pharmaceutical_composition	Therapy	Peptide
0.2	Therapy	Pharmaceutical_ composition	Treatment	Cancer	Protein	Formula	Disease	Antibody	Prevention	Kit
0.4	Therapy	Pharmaceutical_ composition	Treatment	Cancer	Protein	Formula	Composition	Disease	Tissue	Antibody
0.6	Kit	Prevention	Treatment	Pharmaceutical_ composition	Cancer	Protein	Formula	Tissue	Disease	Agent
0.8	Prevention	Kit	Treatment	Pharmaceutical_ composition	Cancer	Protein	Formula	Tissue	Disease	Agent

1	Patient	Kit	Treatment	Pharmaceutical_ composition	Cancer	Protein	Formula	Tissue	Disease	Agent

Table 19 Topic 3 - List of the most relevant keywords

λ	**Keywords**									
0.0	Datum	Signal	User	Information	Data	Network	Server	Computer	Processor	Memory
0.2	Datum	Signal	Sensor	User	Value	Processing	Information	Network	Data	Server
0.4	Datum	Signal	Sensor	Unit	User	Value	Processing	Information	Network	Data
0.6	Datum	Signal	Sensor	Unit	User	Value	Processing	Information	Network	Data
0.8	Datum	Signal	Sensor	Unit	User	Value	Processing	Communication	Network	Data
1	Datum	Signal	Sensor	Unit	User	Value	Processing	Information	Network	Data

Table 20 Topic 4 - List of the most relevant keywords

λ	**Keywords**									
0.0	Battery	Flow	Fluid	Air	Valve	Supply	Power	Battery	Inlet	Pump
0.2	Battery	Flow	Pressure	Fluid	Gas	Water	Fluid	Air	Valve	Energy
0.4	Battery	Pressure	Flow	Fluid	Gas	Water	Power	Valve	Temperature	Energy
0.6	Temperature	Pressure	Flow	Fluid	Gas	Water	Power	Air	Valve	Energy
0.8	Temperature	Pressure	Flow	Fluid	Gas	Water	Power	Air	Valve	Energy

1	Temperature	Pressure	Flow	Fluid	Gas	Water	Power	Air	Valve	Energy

Table 21 Topic 5 - List of the most relevant keywords

λ	**Keywords**									
0.0	Particle	Mixture	Polymer	Waste	Reactor	Carbon	Catalyst	Nanoparticle	Solvent	Biomass
0.2	Material	Process	Product	Reactor	Mixture	Solution	Metal	Reaction	Polymer	Fiber
0.4	Material	Process	Product	Reactor	Mixture	Solution	Metal	Reaction	Polymer	Water
0.6	Material	Process	Product	Reaction	Solution	Mixture	Polymer	Metal	Water	Coating
0.8	Material	Process	Product	Production	Particle	Mixture	Reaction	Polymer	Metal	Water
1	Material	Process	Product	Water	Mixture	Metal	Polymer	Surface	Temperature	Waste

Table 22 Topic 6 - List of the most relevant keywords

λ	**Keywords**									
0.0	Wing	Support	Plate	Rotor	Rotation	Shaft	Ring	Wheel	Wire	Stator
0.2	Assembly	Wire	Wheel	Support	Plate	Movement	Actuator	Rotor	Rotation	Shaft
0.4	Rotation	Assembly	Rotor	Housing	Support	Structure	Surface	Movement	Plate	Rotor
0.6	Assembly	Rotation	Vehicle	Rotor	Support	Structure	Surface	Housing	Container	Shaft
0.8	Assembly	Body	Rotor	Support	Surface	Structure	Plate	Movement	Housing	Vehicle

1	Assembly	Rotor	Actuator	Support	Surface	Structure	Plate	Movement	Housing	Vehicle

WORK PACKAGE 2

How proximity affect successful innovation? An empirical analysis

An empirical model combining the power of AI and economic analysis

Big data and business intelligence play a significant role in the future of businesses, and the technology can be an invaluable partner for all the organizations. Using data can drive better decision making, but it is not sufficient to rely on numbers alone, but it is necessary to further understand the meaning behind big data. If strategy could be data-driven, measurement and evaluation have not been replaced by human judgement and intuition and, in the same way, algorithmic prediction always require human expertise. Even if data has some superiorities to humans in terms of observational and analytic abilities, the humans have to define what is important and what should be measured, because data is not going to do it without human intervention. As shown in the Figure 11 below, humanity and human judgement are needed together with data and predictive modeling approaches, because models are only an approximation of reality and must be complemented by a proper knowledge of the themes under research. In this spirit, in order to exploit and characterize the link between SMEs technologies' characteristics and the successful application to the SME Instrument, we rely on big data and artificial intelligence approaches, combined with microlevel data, through the lens of human judgement, focusing on firms' technological competencies and positioning.

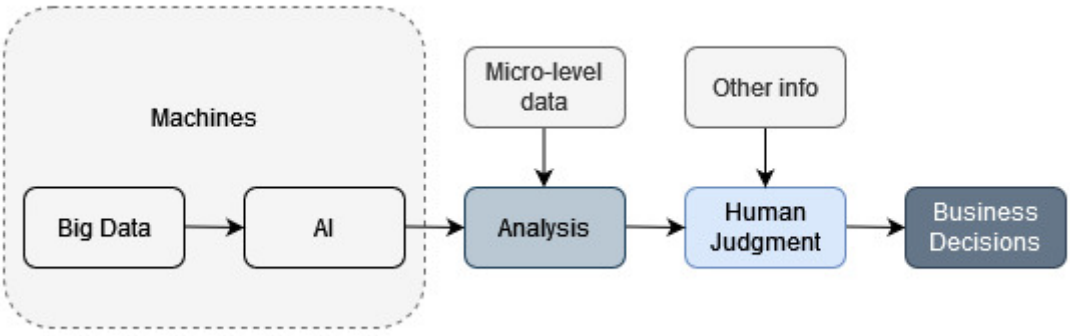


Figure 11-AI and economic analysis

Starting from technology as intermediate tool of devices and knowledge which mediates input and output between process and product technology (Anderson, 1986), literature emphasizes the concept of proximity as communication resource for enterprise competitiveness (Albino, 2007). In fact, the

authors rely on different dimensions of proximity as drivers that affect the competitive success of the firms, by allowing them to obtain internal and external knowledge resources and improve learning efficiencies, and then promote organizational innovative ability.

If this is true, the innovation capabilities of a firm depend on the degree of proximity of different dimensions, and this translates, within the SME-I framework, into a different likelihood of a successfully technology commercialization of SMEI applicants, passing through what we defined as **corporate coherence** (Piscitello, 2004).

Corporate coherence refers to the degree of similarity between each firm's proposal and the cluster of technologies in which it is patenting, that among all the clusters of patents of the firm, presents the highest similarity measure. Thus, starting from the work of Pugliese et al. (Pugliese, 2019), we define corporate coherence on the basis of the distance between each proposal and the patents each firm has, to better capture the technological characteristics embedded in the firm competencies and technologies' portfolio. We decided to not rely on IPC classification of patents because technology classes tend to display a substantial overlap leading to technologically very similar patent in different classes (McNamee, 2013) and because such classification scheme define new technologies on the basis of already existing technologies or their combination, so that if technology change a specific technology could be classified in a different category (Kay, 2014).

In our idea, we do not want to focus on the positioning each firm has, because we are assuming that all firms applying to SME-I are competing in the same competitive space, but we instead rely on Pugliese et al. (Pugliese, 2019), that started from combinative capabilities theory and tested the hypothesis that the performance of a firm is related not only to diversification but also with the coherence of its technological capabilities.

We base our corporate coherent measure on the recombinant perspective on innovation, well established in the economic and managerial literature.

According to this view (Schumpeter, 1939; Arthur, 2007) innovation emerges from the process of searching and recombining existing knowledge elements via combinative capabilities and through the spillovers generated by inventions based on similar technologies. There is a large consensus over the fact that successful diversification strategies cannot be based on randomly assembling technologies, but it is not easy to identify a consistent measure of the degree of coherence that goes into the technological portfolio of the firm and its association with the real business area of activity.

Of course, different approaches have been adopted to explain both the hidden mechanisms within the boundary of the firms and those thar require the acquisition of the external technologies, we privilege

a more data-driven approach that results in a synthetic measure of the distance between the real business area cluster and the most similar pattern of technologies within the global landscape of firm's patented inventions.

Thus, we define the corporate coherence on the basis of the natural language processing (NLP) of patents' text, by explicitly individuating the key findings of inventions and consequently, quantifying the structural dissimilarities among the business area of each company and its basket of patents, in ideas space.

Even though patent data has been used in large-scale empirical analysis (Griliches, 1990), there are some limitations related to the fact that not all the inventions are eligible for protection under the current intellectual property legislation and not all the firms have the same incentive to protect their capital.

Consequently, even if patent data offer a partial view of innovation of the firm, for our purposes the advantages of using patents as primary data source overweight the drawbacks. In fact, for each firm that apply to the SME-I we consider the whole set of patents available at the time they apply, and by analyzing the abstracts of each of them, we are able to decompose the patent portfolio into their constituent technologies and to individuate all the technological areas in which each firm is active and invests to grow and compete. Before turning into the description of our measure of corporate coherence (Corp_Coherence), we briefly discuss LSI-based similarity measures.

LSI-based similarities

Latent Semantic Indexing (LSI) (Dumais S. , 1991; Laham, 1998; Dumais S. , 2004), (also known as Latent Semantic Analysis), is a Natural Language Processing indexing method identifying patterns between terms and concepts in a collection of unstructured documents. This technique is commonly used to associate *concepts* in the space of much lower dimension than a space of *terms* in several tasks, e.g., computational linguistics and information retrieval. It is based on the distributional hypothesis that terms with close meanings will occur in similar documents. Hence, LSI considers two documents semantically close if they have many terms in common, while its documents are semantically distant if they have few terms in common.

The method analyses a set of documents and projects it in a so-called lower-dimensional "latent" space across relevant concepts by leveraging Singular Value Decomposition (SVD).

The SVD is a matrix factorization technique that represents a matrix in the product of matrices:

$$M_{mxn} = U_{mxn}\, \Sigma_{nxn}\, V^{*}$$

where M_{mxn} is the matrix, $U_{mxn}$is the distribution of terms across the different contexts, Σ_{nxn} diagonal matrix with - non-negtive - real numbers that represent the diagonal matrix of the association among the concepts, and $V^*$is the transpose of V_{mxn} that is the distribution of contexts across the different documents.

First, LSI employs the Bag of Word (BoW) (Rui, 2016) model, which provides a term-document matrix that represents the occurrence of corpus terms in a document, with rows representing terms and columns documents. Accordingly, LSI represents the matrix with the distribution of the terms within the set of documents, thus, in a matrix *(m, n)* where *m* is the number of unique terms and *n* is the corpus cardinality. The element *aij* represents the frequency of the term *i* in document *j*. The SVD model is applied to the matrix to create both term and document vector spaces by approximating each single term frequency.

We build an LSI-based algorithm to understand the relationship between concepts in patents and those in the related proposals.

To this end, first, we build a similarity matrix where the corpus is composed of a'l PMI's patents. Then, we use this matrix to compute a similarity matrix for all proposals for a PMI against its patents active at the time of the proposal. The similarity measure used is cosine between two vectors and ranges in [-1, 1], where the higher the score, the greater the similarity.

The obtained similarity score allows us to get a general overview of how terms in PMI's proposals relate with those in PMI's patents. Furthermore, by merging obtained information with data from eCORDA database, e.g., proposal status, it is possible to study if and how similarity patterns between PMIs' patents and proposals impact the proposals' success.

For instance, the histogram in Figure 12 shows the distribution of similarity scores based on the PMIs proposal status, i.e., below threshold, below available budget, and main list.

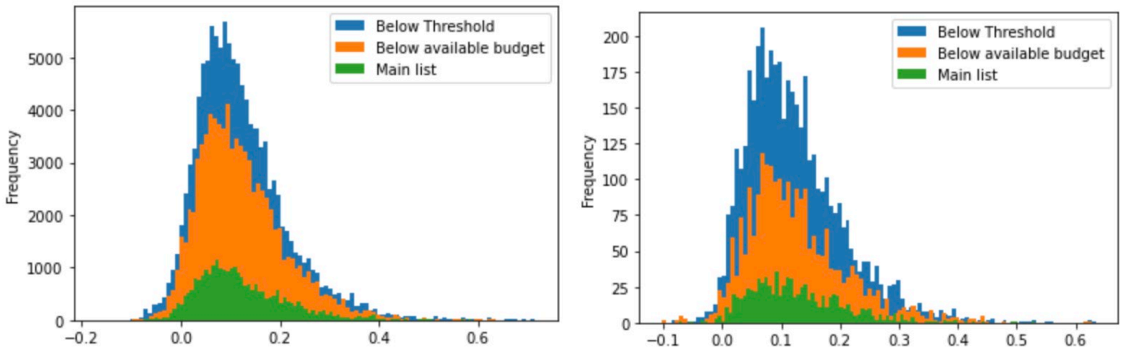


Figure 12-(left) Histogram of similarity scores based on PMI's proposal status. (right) Histogram of similarity scores based on PMI's proposal status (order 0).

Computing Transformer-based similarities

The LSI model has some limitations in capturing semantics from text, e.g., it cannot effectively handle nonlinear dependencies, polysemic words and word ordering.

In recent years, so-called Neural Language Models based on transformer architectures have outperformed statistical approaches in several NLP Tasks. These models are pre-trained in large corpora to solve multiple NLP tasks (such as next word or sentence prediction, question answering, sentiment analysis, paraphrasing) in order to generate a rich textual representation that can be transfer universally to wide variety of NLP tasks such as semantic similarity, clustering, paraphrase detection and text classification.

In our second experiment on semantic similarity, we used the universal-sentence-encoder-large model based on a Transformer (USE-T) [Cert et al., 2018] to encode abstract patents and the related proposals abstracts into 512-dimensional vectors and compare them using the cosine similarity method. The USE-T is designed to be as general purpose as possible. This is accomplished by using multi-task learning with shared encoder parameters whereby the single encoding model is used to feed multiple downstream tasks. The final goal is to learn a model that can return encodings representing a variety of natural language relationships, including semantic similarity measurement and relatedness.

The model architecture is composed of a stack of N = 6 identical Encoder layers (as shown in Figure 13). Each layer has two sub-layers: multi-head self-attention layer and the position wise fully connected feed-forward network. There is a residual connection around each of the two sub-layers, followed by layer normalization. The Positional Encoding that provides to the model information about position of the tokens in the input sequence.

The multi-head self-attention sub-layer consists of several attention layers running in parallel. Each attention layer performs a scaled dot-product between a specific word and other words within a sentence in order to learn very long-term dependencies between words and weigh their relevance to each other to generate output encodings.

The feed-forward neural network further processes each output encoding individually and these output encodings are then passed to the next encoder as its input.

Finally, the final output encodings (context aware word representations) are converted to a fixed length sentence encoding vector by computing the element-wise sum of the representations at each word position.

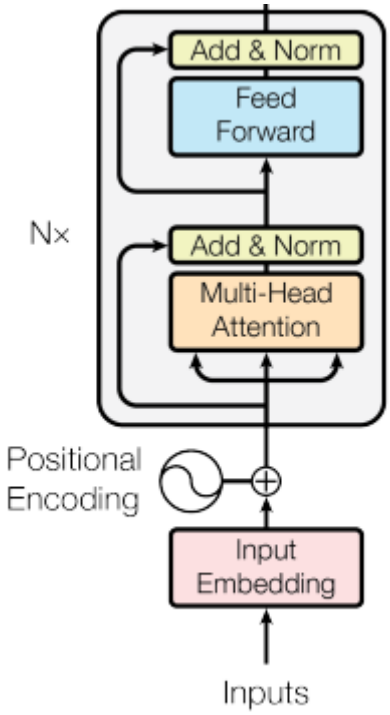


Figure 13-USE-T model uses the encoder sub-graph from the original the transformer architecture proposed by Vaswani et al., 2017

Theory and Hypotheses

Corporate Coherence

In our research setting, we consider a set of explanatory variables, able to explain the successful application to the SME-I.

The first variable of interest is the Corporate Coherence (Corp_ Coherence) measure, designed to capture the adherence of each proposal to the technological portfolio of a firm, with respect to their patents.

We rely on the definition of relatedness in (Zaccaria, 2014) and we change the perspective from simply measuring the breath in scope of business activities to measuring the distance between the topic of the proposal submitted to the SME-I and the business areas in which each firm invests, patents in and diversifies. Thus, we transpose the concept of relatedness at firm's level, by embracing the concept embedded in coherent diversification theory. In order to combine the general structure of technologies portfolio, for each firm in the database we preliminary consider all the patent clusters in which it is patenting, to uncover the whole structure of the existing network of technologies. We then computed the similarity index among the topic cluster of the proposal and the whole technological portfolio at the moment each firm apply to the SME-I, by assuming that a firm presenting a proposal in a topic covered by technological fields in which it actively innovates could benefit in terms of performance, i.e., the successful application to the SME-I. This synthetic index will be our measure of corporate coherence (Corp_ Coherence).

In what we follow, we test the hypothesis that the performance of a firm is related to the coherence of its technological capabilities and their adherence to the cluster topic of the proposal, by trying to understand the link among technological intangible capabilities and the connected proposal's content. In particular, if a firm tends to invest in a topic in which it holds competencies and capabilities and in which it performs R&D, it could be reasonable to expect that benefits are enhanced in case of higher coherence of the fields in which their research activities concentrate. Since we are interested in verifying the effect of Corp_ Coherence on the successful application to SME-I, we test the following hypothesis:

$$Hypothesis\ 1: Corp\_Coherence\ positively\ affects\ firms'\ successfull\ application\ to\ SME\ Instrument$$

The moderating effect of slack

In addition to the moderating role of size, financial resources have to be considered as they can influence the diversification strategies a firm can follow, thus affecting technological and innovation trajectories.

More specifically, in our study we include "financial slack", as the difference between total resources and total necessary payments (Cyert, 1963). Even though different alternative definitions are available, it is important to highlight that financial slack protects from risk via buffering of a firm's technical core (Bourgeois, 1983) and creation of new opportunities (Kim, 2008). Financial slack provides additional resources for exploring new external opportunities, undertaking new investments and absorbs shocks due to external turbulences (Wiklund, 2011; Zona, 2012), and in this framework, its management allows to grow (Rezende, 2020) and survive. In contrast to these positive effects, others argued that financial slack is the result of inefficiency, leading to a sub-optimal investment choice: the more the excess of slack a firm has, the more opportunities it can miss. Since in this study we are interested in exploiting the performances' implication of Corp_ Coherence on the successful application to the SME-I, we have to generalize the possible effect of slack in this sense. Thus, firms with high Corp_ Coherence are able to withdraw a proposal in a way coherent to their technology portfolio and to attenuate the potential negative risks associated to the project activities; moreover, given the risk mitigation properties of financial slack, firms that combine high Corp_ Coherence and high financial slack could significantly reduce or reset the risks connected to the implementation of new innovative activities.

Therefore, we formulate the following hypothesis:

$$Hypothesis\ 2: The\ positive\ effect\ of\ Corp\_Coherence\ on\ the\ firm's\ performance\ became\ stronger\ when\ the\ firm\ has\ high\ slack$$

Exploring proposal success: empirical model specification and variables

To test our hypotheses, we built our work on 72973 SME Instrument applications, during the period 2014-2019. We retry information on 24800 SME-I applicants, from eCORDA proposal database, containing applicants' data and project data for all the evaluated project (both funded and not funded) that apply to the SME Instrument. To analyze proposal success, we further combined patent and standardized financial accounts, obtained from Bureau van Dijk Orbis databases (both Orbis and Orbis Intellectual Property- OrbisIP). More precisely, we retry 191086 patents, over the period 1999-2019 for all the applicants, by considering for each company all patents up to the year of submission.

We then matched SME-I and ORBIS Amadeus BvD database, using BvD Identification number. Afterward, we included Regional Innovation Scoreboard (RSI) indicators for innovation, using NUTS as query for the successful matching, based on NUTS 1 or NUTS 2, depending on differences in regional data availability.

The final sample includes 2682 companies applying to the SME-Instrument.

Dependent variables measuring proposal success

Each proposal subjected to the evaluation process receives an overall score, ranging from 0 to a maximum of 15 status. According to the score achieved, the project proposals are ranked and assigned a certain status. Four different statuses can be distinguished: (a) below threshold: a project proposal is assessed below a certain threshold and rejected; (b) below available budget: a project is assessed above threshold, but it receives a seal of excellence; (c) main list: a project is included in the main list and selected for funding; (d) inadmissible, ineligible or withdrawn. To gain insights into the factors distinguish successful from non-successful application, we run different regression model on the proposal assessment outcome. The first model is a non-linear probability model, based on the binary classification of project status: m=1 for "main list" and "below available budget", and m=0 for "below threshold". We consider the status "below threshold" (m=0) as our reference category and specify the model in terms of

$$P(y_s = m|x) = \frac{\exp\,(x\beta_{m|0}\,)}{\sum_{j=0}^{j=1}\exp\,(x\beta_{j|0})}$$

To estimate the probability of observing a different status for a given set of explanatory variables x. In addition, we estimate a linear regression model based on the evaluation scores of the form:

$$y_s = x\beta + \varepsilon$$

in order to relate the experts' evaluation scores y_s to the same set of explanatory factors in the previous equation. In this case, the β_s denote the ordinary least square (OLS) coefficients and ε is the corresponding error term. The explanatory variables in x reflect both project and company information, that are detailed in the next section.

Independent variables

The first variable of interest is the Corporate Coherence (Corp_ Coherence) index, measured through the dissimilarity index between each proposal and the technological portfolio of the firm, as previously described. For each firm, we considered all the patents in the year of application.

Specifically, the index was computed on the basis of text-based analysis of both proposals and patents' abstracts. With respect to the proposal, the abstract is in fact useful in order to retrieve information related to the domain where firms apply and underscores the technological competencies they leverage; regarding patent data, we focused on abstracts' because they are able to "communicate the technical description in a concise and straightforward manner, avoiding unnecessary words that may increase noise in the extraction process" (Tshitoyan, 2019) .

For financial slack, some previous researchers distinguished between available and potential slack. As in (Gral, 2014; Rafailov, 2017) available financial slack was measured by two indicators: current ratio, expressed as the ratio between current activities and current liabilities, and the working capital ratio, computed as the level of working capital divided by sales. Potential slack is measured through debt-to-equity ratio, as a measures of financial leverage. Since we are dealing with the identification of factors affecting the successful application to a public funding program, we decide to focus on available slack, as variable that signals risk behavior of the firm, computed as current activities divided by current liabilities.

Moreover, given the specific goal of the call to support the innovativeness of SMEs, we account for the level of innovation of the country of the applicant, via *SMEs introducing product or process innovations as percentage of SMEs* indicator, that is able to capture the technological innovation intensity through the number of SMEs that introduced a new product or a new process in the market, respect to the total number of SMEs.

Next to those variables to control for size effects, compute project size, which measures the amount (in log) of funding requested in the project proposal, and firm size, which is represented by the logarithm of total assets reported to the year in which each firm applies to the SME-I. Moreover, we control for the number of previous attempts before the most recent attempt, with the construction of a tailored variable.

In addition, since eCORDA database accounts for both Phase 1 and Phase 2, that have different success rates[9](SMES Innovation Report), we compute a Boolean variable with value 1 if SMEs apply to the Phase 2 of SME-I, and zero otherwise.

More precisely Phase 1 support SMEs in exploring and assessing the technical feasibility and commercial potential of a breakthrough innovation in their industry (thanks to a lump sum of €50.000,00) while Phase 2 offer support for innovation projects underpinned by a strategic business plan and feasibility, and then encourage SMEs to engage in R&D and testing activities, through a flexible grant. Given the different configurations of the two Phases, it is important to control for the duration of each proposal, that it is usually shorter for projects of Phase 1 respect to those submitted to Phase 2 (usually between 12 and 24 months), so that we add Proposal Duration, as control variable. Finally, we include a time trend variable, in order to control for different years of the call (from 2014 to 2019), and a factor variable able to capture the different topics in which each firm competes. More precisely, its levels (from 1 to 9) were identified through the semantic analysis of the abstract of each proposal submitted, as described in the Section 1.2-Business application characterization (business models). An overview of all the variable is given in the following Table.

Empirical results

Due to the high positive correlation between the Phase each firm applies and the project size we were not able to estimate a model including both variables. We therefore estimated two model variants; the first with Phase variable, and the last refers to the model with project size variable. The model variants can be interpreted separately (correlation values are shown in the appendix).

Table 23 reports the estimation result for our proposal success model based on both nonlinear probability model and multiple linear regression model, as previously mentioned. The first four columns refer to the multivariate probit estimation and the last four contain results of the OLS regression models based on expert evaluation scores as outcome variables. As mentioned before, we estimated four different variants of each specification: Model 1 and Model 3 contains the results of the standard model, with Phase and project size variables, separately considered, and with no moderation effect; Model 2 and Model 4 introduce the moderating effect of financial slack on Corp_

[9] The average success rate for SME-I submissions is 4.7 % in Phase 2 and 8.6 % in Phase 1. However, if success rate is computed per project proposal rather than per submission, around 11.5 % of Phase 2 and 16.6 % of Phase 1 proposals are ultimately successful.

Coherence. The project status “below threshold” is the reference for interpreting the coefficient of all regression specifications.

Turning to the estimation results, we can observe a positive relationship between the corporate coherence (Corp_ Coherence) index and the successfully application to the SME-I. Results are confirmed both with probit and OLS specifications. Interestingly, the coefficients are slightly higher for OLS regression (ranging from a minimum of 0.9823 of Model 4 to a maximum of 1.083 of Model 1) respect to probit parametrization (coefficients range from 0.596 of Model 4 to 0.696 of Model 1). Moreover, all effects are significant except for Model 4 of probit specification, when the moderating effect of available financial slack is introduced together with project size dimension. These conclusions in general confirm the validity of both Hypothesis 1, predicting a positive effect of Corp_ Coherence on the successful application to the SME-I. In NLP model, when we estimate the probability of being in the STATUS m=1 (that means being in the “Main list” or “below available budget”), we see that, while keeping all other variables constant at their mean values, the effect on proposal success increases by a maximum of 69.68% (when accounting for different Phases). Moreover, when considering the effect of Corp_ Coherence on the expert evaluation scores for proposal, we can see that an increase in the coherence between proposal abstract and technological portfolio, summarized by an increase in the Corp_ Coherence measure, will translate in an increase of the evaluation score assigned to the proposal of 1.083 and 1.002 (Model 1 and Model 3, respectively). This result is conceptually important: firms can obtain a higher experts’ evaluation score through a more technological coherent proposal, signaling the experts that they are able to leverage all the core technological knowledge and competences they have. Therefore, for a successful application to the SME-I firms have to design and submit a proposal adherent to the technology trajectory of the firm’s activities. This conclusion also suggests that from a managerial point of view, firms should balance the benefits of technological diversification with its disadvantages, by choosing the direction carefully and extending their activities into technology fields that share a common knowledge and competences base.

Hypothesis 2 regards the moderating effect of financial slack on Corp_ Coherence and predicts that the positive effect of Corp_ Coherence on firm' s innovation performance (successful application to the SME-I) became stronger when the firm has high slack. According to the results, the interaction term between Corp_ Coherence and available financial slack is positive, meaning that higher levels of coherence and higher level of financial slack are associated with a more positive outcome. Interestingly, this effect is significant only when considering their conjoint effect on the expert

evaluation scores, and not the probability of being in the STATUS m=1 ("Main list" or "below available budget"), partially supporting the Hypothesis 2. This result highlights that, even if higher level of coherence and higher level of slack are not able to significantly increase the probability of get a successful application to the SME-I, firms that simultaneously signal higher coherence in their technology trajectory and higher ability to mitigate innovation risks, are evaluated in a more positive way by the experts, in terms of quality and efficiency of implementation. In all cases, the effect of available slack is positively associated to both outcome variables, evidencing how for small and medium sized firms' higher levels of financial slack do not indicate inefficiency but are considered as a safe strategy in order to face less downside risk from poor performances of their investments.

Regarding other control variables, we can see that the successful application to the SME-I, in terms of both probability of being successful and higher evaluation scores, is positively and significant associated to the innovation intensity level of the country of origin of the firm, represented by the ratio between SMEs that have introduced a new product or a new process respect to the total of SMEs. This is in line with the scope of the program to support innovation in small and medium-sized enterprises (SMEs), and to develop and capitalize on their innovativeness potential, by filling the gap in funding for early-stage high-risk projects. Thus, if a proposal is submitted by a firm that operates in a more innovative and dynamic environment (higher rate of innovation intensity), it will probably have a higher innovativeness potential, therefore it will maximize the probability of a successful application (higher probability of being successful and to have a higher evaluation score). Moreover, when considering the effect of project related characteristics, we can argue that project size, in terms of amount of funding requested is always positive related to the outcome variable, even if this conclusion must be interpreted in the lens of Phase variable. SME-I in fact supports different phases of innovation processes, from concept and feasibility assessment to innovation implementation and commercialization. Since proposals submitted to the Phase 1 refer to feasibility studies, with a fixed lump-sum of €50000, while Phase 2 directly support innovations, in terms of product, process or service to lunch on the market and of their commercialization strategies (with a varying grant, between a minimum of €500,00 and a maximum of €2.5 million), it seems reasonable to expect that proposals with higher level of funding request, applying for Phase 2, will have higher innovativeness contents and of quality, that easier results into higher scores and therefore, in higher probability of being successful. In addition, with respect to the Prev_attempts variable, that account for the number of times each firm apply to the SME-I, we find a positive relationship with the outcome variables. More interestingly, this variable significant impact on the expert evaluation scores, meaning that

having applied more times to the same instrument translate into a better ability of designing a clear and more readable proposal, gaining an overall higher evaluation score but this does not automatically translate in a higher probability of success. Finally, the time dummy indicates a yearly increase in application from 2014 to 2016, with a drop in 2017, followed by a new increase in 2018. This is due to the reconfiguration of the instrument, active from 2018, when the SME-I was included in the EIC Pilot, with a more bottom-up approach, deleting the requirement of thematic topics and introducing some new features.

Summary and discussions

We exploited the link between SMEs technologies' characteristics and the application of SME-I by relying on big data and artificial intelligence which were combined with microlevel data.

We built our analysis by focusing on the role of corporate coherence, defined as the degree of similarity between each firm's proposal and the cluster in which it is patenting, within the SME-I framework.

By applying Latent Semantic Indexing, we decomposed the patent portfolio for each firm applying to the SME-I and we established the technological areas in which the firm is active, invests and competes. Then, corporate coherence was measured as dissimilarity index between each SME-I proposal and the technological portfolio of the firm.

Then, we both run non-linear probability model on the proposal assessment outcome to estimate the probability of successful application to SME-I and multiple linear regression model to estimate the experts' evaluation scores.

The empirical results shows that corporate coherence is positively correlated with firms' evaluation score assigned to the SME-I's proposal. This can be translated as the ability of firms into leveraging their core technological knowledge and competencies by submitting a proposal which is adherent to their activities' technological trajectory. Not only a SME-I proposal is positively evaluated when firms both signal corporate coherence and ability to mitigate innovation risks, but also when firms operate in an innovative and dynamic environment.

Like the other studies of this kind, the study is affected by some limitations that, however, offer insights for developing new research trends. First, the study is performed on a predeterminate sample of European SMEs mainly oriented toward the development of innovations. Second, we recognized

that other content analysis tools and different similarity indices exist. These limitations, on the one hand, might affect the generalization of the results. On the other hand, they can offer the stimulus to perform further research by comparin' SMEs' strategies in low-medium-tech industrial sectors or by analyzing the strategies adopted in different countries and linking them to the local innovation ecosystems and the national innovation policy.

Table 23-Regression results: nonlinear probability and Ordinary Least Square regression

Model Type	**NLP: Probit**				**OLS Regression**			
Dep.var.	**PROJ_STATUS (0,1)**				**SCORE**			
	Model 1	***Model 2***	***Model 3***	***Model 4***	***Model 1***	***Model 2***	***Model 3***	***Model 4***
Corp_Coherence	**0.6968491** *	**0.6799355** *	**0.6165548** *	0.5961870	**1.0839125** **	**1.0678.249** **	**1.0029255** **	**0.9823674** **
	(0.36)	(0.36)	(0.36)	(0.36)	(0.49)	(0.49)	(0.49)	(0.49)
Prev_attemps	0.2301960	0.2267652	0.2366509	0.2332608	**0.9111058** ***	**0.8914157** ***	**0.9184567** ***	**0.8987192** ***
	(0.20)	(0.20)	(0.20)	(0.20)	(0.16)	(0.16)	(0.16)	(0.16)
Firm_size	0.0126838	0.0142255	0.0108120	0.0123770	-0.0336348	0.0318238	0.0336263	0.0317761
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)
2.Phase	0.1438018		0.1421130		**0.7671934** ***		**0.7557607** ***	
	(0.14)		(0.14)		(0.21)		(0.21)	
Innov_intensity	**0.6676735** ***	**0.6560463** ***	**0.6723234** ***	**0.6603801** ***	**1.1735835** ***	**1.1367483** ***	**1.1726452** ***	**1.1356436** ***
	(0.20)	(0.20)	(0.21)	(0.21)	(0.29)	(0.29)	(0.30)	(0.29)
Prop_Duration	-0.0109792	**0.0200506** ***	0.0106993	**0.0201383** ***	**-0.0249197** **	**0.0368427** ***	**0.0242174** **	**0.0367634** ***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Proj_Size		**0.0971043** **		**0.0989211** **		**0.2942813** ***		**0.2949414** ***
		(0.04)		(0.04)		(0.06)		(0.06)
Av_Slack			0.0000492	0.0000522			**0.0000391** *	**0.0000410** *
			(0.00)	(0.00)			**(0.00)**	**(0.00)**
Corp_Coherence ##Av_Slack			0.0006104	0.0006474			**0.0005377** *	**0.0005611** *
			(0.00)	(0.00)			(0.00)	(0.00)
_cons	-1.2093593 **	2.2564567 ***	1.1898065 **	2.2556241 ***	9.7493.280 ***	6.5473520 ***	9.7414644 ***	6.5351478 ***
	(0.51)	(0.66)	(0.51)	(0.66)	(0.63)	(0.90)	(0.63)	(0.90)
Dummy Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dummy Topic_cluster	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. of Obs.	2.138	2.138	2.125	2.125	2.125	2.125	2.112	2.112
Pseudo R2/R2	***0.1453***	***0.1468***	***0.1470***	***0.1486***	***0.0957***	***0.1013***	***0.0968***	***0.1026***
IC	2558.9	2554.5	2558.2	2553.6	8758.2	8745.1	8722.9	8709.3

WORK PACKAGE 3

Dissemination of the results

In recent years, increasing attention is paid to the exploration and dissemination of both data and results. To this end, we are still working on several initiatives. On the one hand we are working on finalizing different working papers as results of the methodological exercises conducted. These working papers will allow us to participate to international conferences (e.g., Accademy of Management, R&D Management Conference, Strategic Management Society Conference) to collect feedbacks from the community and the experts in the field. Thus, we will submit the papers to international top-tier journals focused on innovation and IP strategies.

On the other hand, we are working on making available to the community an analytics dashboard. The platform represents an interactive tool to explore and characterize the topics identified by the procedure in Section "*Business models identification"*. The dashboard allows browsing the results obtained with topic modeling, displaying the intertopic distance map and the more salient terms. Moreover, each topic is characterized by its most relevant words, shown in word clouds. Besides, the platform has an extensive section displaying several statistics through more than 30 analytic charts. A user can choose whether to view aggregated statistics, thus related to all SME's proposals, or single, by topic. The analyses include, but are not limited to, the ratio of the number of SMEs to submissions, funded submissions, and under budget submissions, the relationship between proposals and the number of participants and the status, and the relationship between acceptance rate and finance rate. Finally, the dashboard shows statistics related to SME's topics, such as the distribution of proposals, and the possible overlaps between different topics.

On the one hand, we believe that this platform can summarize the approach presented by us in detail. On the other hand, we believe that visualizations can help understand and observe multiple types of relationships.

References

Aharonson, B. a. (2016, 02). Mapping the Technological Landscape: Measuring Technology Distance, Technological Footprints, and Technology Evolution. Research Policy, 45, 81-96. doi:10.1016/j.respol.2015.08.001

Aharonson, B. S., & Schilling, M. A. (2016). "Mapping the technological landscape: Measuring

Akhondi, S. A., Rey, H., Schwörer, M., Maier, M., Toomey, J., Nau, H., ... & Doornenbal, M. (2019). Automatic identification of relevant chemical compounds from patents. Database, 2019.

Al-Shboul B. and Myaeng S. (2011) "Query phrase expansion using wikipedia in patent class search." Information Retrieval Technology, pages 115–126.

Albino, V. a. (2007, 12). Proximity as a communication resource for competitiveness: A rationale for technology clusters. International Journal of Learning and Intellectual Capital, 4. doi:10.1504/IJLIC.2007.016337

Alves, T., Rodrigues, R., Costa, H., & Rocha, M. (2017, June). Development of Text Mining Tools for Information Retrieval from Patents. In International Conference on Practical Applications of Computational Biology & Bioinformatics (pp. 66-73). Springer, Cham.

Anderson, P. a. (1986, 09). Technological Discontinuities and Organizational Environments. Administrative Science Quarterly, {}. doi:10.2307/2392832

APRE (2016). "PMI e Ingegneria Finanziaria in Horizon 2020". Agenzia per la Promozione della Ricerca Europea, Rome.

Arthur, W. B. (2007). The structure of invention. Research Policy, 36(2), 274-287. doi:https://doi.org/10.1016/j.respol.2006.11.005

Autio, E., 2016. Entrepreneurship support in Europe: Trends and challenges for EU policy. *London, England: Imperial College Business School*.

Bergeaud, A., Potiron, Y., & Raimbault, J. (2017). Classifying patents based on their semantic content. PloS one, 12(4), e0176310.

Bhatia, S., He, B., He, Q., & Spangler, S. (2012, October). A scalable approach for performing proximal search for verbose patent search queries. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 2603-2606). ACM.

Bottazzi, G. a. (2005, 02). Growth and Diversification Patterns of the Worldwide Pharmaceutical Industry. Review of Industrial Organization, 26, 195-216. doi:10.1007/s11151-004-7296-5

Bourgeois, L. J. (1983). Organizational Slack and Political Behavior Among Top Management Teams. Academy of Management Proceedings, 43-47.

Carrier, M. A. (2012). A roadmap to the smartphone patent wars and FRAND licensing. CPI Antitrust Chronicle, 2.

Cassiman, B., R.Veugelers, and P. Zuniga (2008). “In search of performance effects of (in)direct industry science links”, Industrial and Corporate Change, 17(4): 611–646.

Chang, C.-H. a.-S.-J. (2014). Determinants of absorptive capacity: contrasting manufacturing vs services enterprises. R&D Management, 44(5), 466-483.

Chechev, M., Gonzàlez, M., Màrquez, L., & España-Bonet, C. (2012, April). The patents retrieval prototype in the MOLTO project. In Proceedings of the 21st International Conference on World Wide Web (pp. 231-234). ACM.

Chen, Y. L., & Chang, Y. C. (2012). A three-phase method for patent classification. Information Processing & Management, 48(6), 1017-1030.

Coombs, J. E., & Bierly, P. E. (2006). “Measuring technological capability and performance”. R&D Management, 36(4), 421-438.

Cyert, R. a. (1963). Behavioral Theory of the Firm. Englewood Cliffs: Wiley.

Del Sarto, N., Di Minin, A., Ferrigno, G., & Piccaluga, A. (2019). Born global and well educated: start-up survival through fuzzy set analysis. Small Business Economics, 1-19.

Di Minin, A., De Marco, C.E., & Karaulova, M. (2016).“The SME Instrument - So Far So Good? Expectations, Reality and Lessons to Learn”. Berkeley Roundtable on the International Economy, 2016-4.

Dong, H. R., Chen, D. Z., & Huang, M. H. (2018, December). Do Long-term Patents Have a Higher Citation Impact?. In 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 1518-1522). IEEE.

Dumais, S. (1991). Improving the retrieval of information from external sources. Behavior Research Methods, Instruments, & Computers, 23, 229-236. doi:10.3758/BF03203370

Dumais, S. (2004). Latent Semantic Analysis. Annual Review of Information Science and Technology, 189-230.

EASME (2016). “Catalysing European Innovation. EASME’s report of the first two years of implementation of the SME Instrument 2014-2015”. Brussels.

EASME (2017). “Accelerating Innovation in Europe. Horizon 2020 SME Instrument Impact Report 2017 Edition”. Brussels.

Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P., & Zalányi, L. (2013). Prediction of emerging technologies based on analysis of the US patent citation network. Scientometrics, 95(1), 225-242.

Fall, C. J., Törcsvári, A., Benzineb, K., & Karetka, G. (2003, April). Automated categorization in the international patent classification. In Acm Sigir Forum (Vol. 37, No. 1, pp. 10-25). ACM.

Fresco, L.O., Martinuzzi, A., Wiman, A. (2015). "COMMITMENT and COHERENCE essential ingredients for success in science and innovation. Ex-Post-Evaluation of the 7th EU Framework Programme (2007-2013)". High Level Expert Group.

Fujii, A., Utiyama, M., Yamamoto, M., & Utsuro, T. (2009, July). Evaluating effects of machine translation accuracy on cross-lingual patent retrieval. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (pp. 674-675). ACM.

Galasso, A., Schankerman, M., and C. J. Serrano (2011). "Trading and Enforcing Patent Rights". NBER Working Papers No. 17367, National Bureau of Economic Research, Inc., www.nber.org/papers/w17367 (Revised August 2012).

Ganguly, D., Leveling, J., Magdy, W., & Jones, G. J. (2011, October). Patent query reduction using pseudo relevance feedback. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 1953-1956). ACM.

Georghiou, L., & Roessner, D. (2000). "Evaluating technology programs: tools and methods." Research policy, 29(4-5), 657-678.

Gompers, P., Lerner, J., and D. Scharfstein (2005). "Entrepreneurial Spawning: Public Corporations and the Genesis of New Ventures, 1986 to 1999." The Journal of Finance, 60(2), 577-614.

Goto, I., Chow, K. P., Lu, B., Sumita, E., & Tsou, B. K. (2013, June). Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. In NTCIR.

Gral, B. (2014). How Financial Slack Affects Corporate Performance. Springer Gabler, Wiesbaden. doi:https://doi.org/10.1007/978-3-658-04552-4

Granstrand, O., Patel, P., and K. Pavitt (1997). "Multi-technology corporations: why they have "distributed" rather than "distinctive core" competencies", California management review, 39(4), 8-25

Griliches, Z. (1990, 12). Patent Statistics as Economic Indicators: A Survey. Journal of Economic Literature, 28(4). Tratto da http://www.jstor.org/stable/2727442

Honghua, Q., & Xiang, Y. (2009). Research on a method for building up a patent map based on k-means clustering algorithm. Science Research Management, 2(9), 70-76.

Hristidis, V., Ruiz, E., Hernández, A., Farfán, F., & Varadarajan, R. (2010, October). Patentssearcher: a novel portal to search and explore patents. In Proceedings of the 3rd international workshop on Patent information retrieval (pp. 33-38). ACM.

Hsiang, J.-S. L. (2020). Patent classification by fine-tuning BERT language model. World Patent Information, 61.

Hu, P., Huang, M., Xu, P., Li, W., Usadi, A. K., & Zhu, X. (2012, October). Finding nuggets in IP portfolios: core patent mining through textual temporal analysis. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 1819-1823). ACM.

Jin, X., Spangler, S., Chen, Y., Cai, K., Ma, R., Zhang, L., ... & Han, J. (2011, December). Patent maintenance recommendation with patent information network model. In 2011 IEEE 11th International Conference on Data Mining (pp. 280-289). IEEE.

Jin, Y. (2010, August). A hybrid-strategy method combining semantic analysis with rule-based MT for patent machine translation. In Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010) (pp. 1-4). IEEE.

Jochim, C., Lioma, C., Schütze, H., Koch, S., & Ertl, T. (2010, October). Preliminary study into query translation for patent retrieval. In Proceedings of the 3rd international workshop on Patent information retrieval (pp. 57-66). ACM.

Kay, L. a. (2014). Patent Overlay Mapping: Visualizing Technological Distance. Journal of the Association for Information Science and Technology, 65. doi:10.1002/asi.23146

Kesting, P., & Günzel-Jensen, F. (2015). “SMEs and new ventures need business model sophistication”. Business Horizons, 58(3), 285-293.

Kim, H. a. (2008, 06). Ownership Structure and the Relationship Between Financial Slack and R&D Investments: Evidence from Korean Firms. Organization Science, 19, 404-418. doi:10.1287/orsc.1080.0360

Kim, J. H., & Choi, K. S. (2007). Patent document categorization based on semantic structural information. Information processing & management, 43(5), 1200-1215.

Kim, Y., Seo, J., & Croft, W. B. (2011, July). Automatic boolean query suggestion for professional search. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (pp. 825-834). ACM.

Kondo, S., Komachi, M., Matsumoto, Y., Sudoh, K., Duh, K., & Tsukada, H. (2011, December). Learning of Linear Ordering Problems and its Application to JE Patent Translation in NTCIR-9 PatentMT. In NTCIR.

Lanahan, L. (2016). “Multilevel public funding for small business innovation: a review of US state SBIR match programs.” The Journal of Technology Transfer, 41(2), 220-249.

Lee, C., Cho, Y., Seol, H., & Park, Y. (2012). A stochastic patent citation analysis approach to assessing future technological impacts. Technological Forecasting and Social Change, 79(1), 16-29.

Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. Technovation, 29(6-7), 481-497.

Lee, S., Yoon, B., Lee, C., & Park, J. (2009). “Business planning based on technological capabilities: Patent analysis for technology-driven roadmapping”. Technological Forecasting and Social Change, 76(6), 769-786.

Lerner, J. (1994), “The Importance of Patent Scope: An Empirical Analysis”, RAND Journal of Economics, 25(2): 319-333.

Lerner, J. (2009), "The empirical impact of intellectual property rights on innovation: Puzzles and clues." American Economic Review, 99(2): 343-48.

Levy, M., & Powell, P. (2000). Information systems strategy for small and medium sized enterprises: An organizational perspective. Journal of Strategic Information Systems, 9, 63–84.

Levy, M., Powell, P., & Yetton, P. (2001). SMEs: Aligning IS and the strategic context. Journal of Information Technology, 16, 133–144.

Li, Y., & Shawe-Taylor, J. (2007). Advanced learning algorithms for cross-language patent retrieval and classification. Information processing & management, 43(5), 1183-1199.

Lin, Y.-M. C.-H.-J. (2013). Does technological diversification matter to firm performance? The moderating role of organizational slack. Journal of Business Research, 66(10), 1970-1975.

Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Wang, J., & Lin, H. (2018). A neural network approach to chemical and gene/protein entity recognition in patents. Journal of cheminformatics, 10(1), 65.

Lupu, M., Piroi, F., & Hanbury, A. (2010, October). Aspects and analysis of patent test collections. In Proceedings of the 3rd international workshop on Patent information retrieval (pp. 17-22). ACM.

Magdy, W., & Jones, G. J. (2011, October). A study on query expansion methods for patent retrieval. In Proceedings of the 4th workshop on Patent information retrieval (pp. 19-24). ACM.

Magdy, W., & Jones, G. J. (2011, October). An efficient method for using machine translation technologies in cross-language patent search. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 1925-1928). ACM.

Mahdabi, P., Andersson, L., Keikha, M., & Crestani, F. (2012, August). Automatic refinement of patent queries using concept importance predictors. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (pp. 505-514). ACM.

Mahdabi, P., Keikha, M., Gerani, S., Landoni, M., & Crestani, F. (2011, June). Building queries for prior-art search. In Information Retrieval Facility Conference (pp. 3-15). Springer, Berlin, Heidelberg.

McNamee, R. C. (2013). Can't see the forest for the leaves: Similarity and distance measures for hierarchical taxonomies with a patent classification example. Research Policy, 42(4), 855-873.

Meister, D.B. (2017). Entrepreneurial firms and information systems capabilities. Twenty-Third Americas Conference on Information Systems.

Moehrle, M. G., & Gerken, J. M. (2012). "Measuring textual patent similarity on the basis of combined concepts: design decisions and their consequences." Scientometrics, 91(3), 805-826.

Muller, P., Devnani, S., Julius, J., Gagliardi, D., & Marzocchi, C. (2016). "Annual report on European SMEs". European Commission.

Muller, P., Julius, J., Herr, D., Koch, L., Peycheva, V., & McKiernan, S. (2017). Annual Report on European SMEs 2016/2017: Focus on self-employment. European Commission.

Narin, F., K.S. Hamilton, and D. Olivastro (1997), "The increasing linkage between U.S. technology and public science". Research Policy, 26: 317-330.

Neirotti, P., & Raguseo, E. (2017). Flexible work practices and the firm's need for external orientation: An empirical study of SMEs. Journal of Enterprise Information Management, 30, 922–943.

Oh, S., Lei, Z., Lee, W. C., & Yen, J. (2014, May). Patent evaluation based on technological trajectory revealed in relevant prior patents. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 545-556). Springer, Cham.

Petruzzelli, A. M., Rotolo, D., & Albino, V. (2015). Determinants of patent citations in biotechnology: An analysis of patent influence across the industrial and organizational boundaries. Technological Forecasting and Social Change, 91, 208-221.

Ploskas, N., Zhang, T., Sahinidis, N. V., Castillo, F., & Sankaranarayanan, K. (2019). Evaluating and ranking patents with multiple criteria: How many criteria are required to find the most promising patents?. Computers & Chemical Engineering, 123, 317-330.

Ponta, L., Puliga, G., Oneto, L., & Manzini, R. (2019, April). Innovation Capability of Firms: A Big Data Approach with Patents. In INNS Big Data and Deep Learning conference (pp. 169-179). Springer, Cham.

Pugliese, E. A. (2019, 10). Coherent diversification in corporate technological portfolios. PLOS ONE, 14(10), 1-22.

Piscitello, L. (2004). Corporate diversification, coherence and economic performance. Industrial and Corporate Change, 757–787.

Rafailov, D. (2017). Financial Slack and Performance of Bulgarian Firms. Journal of Finance and Bank Management, 5.

Rezende, J. (2020). Financial slack as driver of brazilian firms' growth. Revista de Administração da UFSM, 13.

Schumpeter, J. A. (1939). Business Cycles: A Theoretical, Historical and Statistical Analysis of the Capitalist Process. New York Toronto London: McGraw-Hill Book Company.

Solow, R. M. (1956). "A contribution to the theory of economic growth." The quarterly journal of economics, 70(1), 65-94.

Squicciarini, M., Dernis, H., & Criscuolo, C. (2013)."Measuring patent quality".

Suh, J. H., & Park, S. C. (2009). Service-oriented technology roadmap (SoTRM) using patent map for R&D strategy of service industry. Expert Systems with Applications, 36(3), 6754-6772.

Tang, J., Wang, B., Yang, Y., Hu, P., Zhao, Y., Yan, X., ... & Usadi, A. K. (2012, August). PatentMiner: topic-driven patent analysis and mining. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1366-1374). ACM.

Tannebaum, W., & Rauber, A. (2012, July). Analyzing query logs of uspto examiners to identify useful query terms in patent documents for query expansion in patent searching: a preliminary study. In Information Retrieval Facility Conference (pp. 127-136). Springer, Berlin, Heidelberg.

technology distance, technological footprints, and technology evolution." Research Policy, 45(1), 81-96.

Teodoro, D., Gobeill, J., Pasche, E., Vishnyakova, D., Ruch, P., & Lovis, C. (2010). Automatic Prior Art Searching and Patent Encoding at CLEF-IP'10. In CLEF (Notebook Papers/LABs/Workshops).

Tikk, D., Biró, G., & Törcsvári, A. (2008). A hierarchical online classifier for patent categorization. Emerging Technologies of Text Mining: Techniques and Applications.

Trajtenberg, M, Jaffe, A., and R. Henderson (1997). "University versus Corporate Patents: A Window on the Basicness of Inventions", Economics of Innovation and New Technology, 5(1): 19-50.

Trappey, A. J., Trappey, C. V., Wu, C. Y., Fan, C. Y., & Lin, Y. L. (2012, May). Intelligent recommendation methodology and system for patent search. In Proceedings of the 2012 IEEE 16th

International Conference on Computer Supported Cooperative Work in Design (CSCWD) (pp. 172-178). IEEE.

Tshitoyan, V. a. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. Nature, 95-98.

Van Zeebroeck, N. (2011). The puzzle of patent value indicators. Economics of innovation and new technology, 20(1), 33-62.

Wiklund, S. W. (2011, 07). The Importance of Slack for New Organizations Facing 'Tough' Environments. Journal of Management Studies, 48, 1071-1097. Tratto da https://ideas.repec.org/a/bla/jomstd/v48y2011ip1071-1097.html

Yang, Y. Y., Akers, L., Yang, C. B., Klose, T., & Pavlek, S. (2010). Enhancing patent landscape analysis with visualization output. World Patent Information, 32(3), 203-220.

Yang, Y., Akers, L., Klose, T., & Yang, C. B. (2008). Text mining and visualization tools–impressions of emerging capabilities. World Patent Information, 30(4), 280-293.

Yeap, T., Loo, G. H., & Pang, S. (2003, July). Computational patent mapping: intelligent agents for nanotechnology. In Proceedings International Conference on MEMS, NANO and Smart Systems (pp. 274-278). IEEE.

Yeh, H. Y., Lo, C. W., Chang, K. S., & Chen, S. H. (2018). Using hot patents to explore technological evolution: a case from the orthopaedic field. The Electronic Library, 36(1), 159-171.

Zaccaria, A. A. (2014, 12). How the Taxonomy of Products Drives the Economic Development of Countries. PLOS ONE, 9, 1-17.

Zona, F. (2012). Corporate Investing as a Response to Economic Downturn: Prospect Theory, the Behavioural Agency Model and the Role of Financial Slack. British Journal of Management, 23(S1), S42-S57.

Zott, C., & Amit, R. (2010). "Business model design: an activity system perspective". Long range planning, 43(2-3), 216-226.

Appendix

Table A1-Pairwise correlations matrix

Variables	(PRJ_STATUS)	(SCORE)	(Corp_Coherence)	(Av_Slack)	(Prev_attempts)	(Firm_size)	(Phase)	(Proj_Size)	(Innov_Intensity)	(Prop_Duration)
PRJ_STATUS	1.000									
SCORE	0.638***	1.000								
Corp_Coherence	0.026	0.028	1.000							
Av_Slack	-0.024	0.002	-0.010	1.000						
Prev_attempts	0.029	0.074***	-0.027	-0.003	1.000					
Firm_size	-0.018	-0.037*	0.015	-0.029	0.009	1.000				
Phase	-0.017	0.101***	0.040**	0.022	0.002	-0.038*	1.000			
Proj_Size	0.000	0.117***	0.037*	0.022	0.008	-0.041**	0.979***	1.000		
Innov_Intensity	0.072***	0.102***	0.003	0.036*	0.002	0.002	0.077***	0.085***	1.000	
Prop_Duration	-0.024	0.081***	0.023	0.006	0.007	-0.039*	0.908***	0.904***	0.069***	1.000

***** p<0.01, ** p<0.05, * p<0.1**