Böhmecke-Schwafert, Moritz; Dörries, Colin

Measuring Innovation in Mauritius' ICT Sector Using Unsupervised Machine Learning: A Web Mining and Topic Modeling Approach

Journal of the Knowledge Economy

Provided in Cooperation with:
Springer Nature

Suggested Citation: Böhmecke-Schwafert, Moritz; Dörries, Colin (2023) : Measuring Innovation in Mauritius' ICT Sector Using Unsupervised Machine Learning: A Web Mining and Topic Modeling Approach, Journal of the Knowledge Economy, ISSN 1868-7873, Springer US, New York, NY, pp. 1-34,
https://doi.org/10.1007/s13132-023-01587-0

This Version is available at:
https://hdl.handle.net/10419/309804

Journal of the Knowledge Economy
https://doi.org/10.1007/s13132-023-01587-0

Measuring Innovation in Mauritius' ICT Sector Using Unsupervised Machine Learning: A Web Mining and Topic Modeling Approach

Moritz Böhmecke-Schwafert[1] · **Colin Dörries**[1]

Received: 20 March 2023 / Accepted: 28 October 2023

Abstract
Measuring innovation accurately and efficiently is crucial for policymakers to encourage innovation activity. However, the established indicator landscape lacks timeliness and accuracy. In this study, we focus on the country of Mauritius that is transforming its economy towards the information and communication technology (ICT) sector. We seek to extend the knowledge base on innovation activity and the status quo of innovation in Mauritius by applying an unsupervised machine learning approach. Building on previous work on new experimental innovation indicators, we combine recent advances in web mining and topic modeling and address the following research questions: *What are potential areas of innovation activity in the ICT sector of Mauritius?* Furthermore, *do web mining and topic modeling provide sufficient indicators to understand innovation activities in emerging countries?* To answer these questions, we apply the natural language processing (NLP) technique of Latent Dirichlet Allocation (LDA) to ICT companies' website text data. We then generate topic models from the scraped text data. As a result, we derive seven categories that describe the innovation activities of ICT firms in Mauritius. Albeit the model approach fulfills the requirements for innovation indicators as suggested in the Oslo Manual, it needs to be combined with additional metrics for innovation, for example, with traditional indicators such as patents, to unfold its potential. Furthermore, our approach carries methodological implications and is intended to be reproduced in similar contexts of scarce or unavailable data or where traditional metrics have demonstrated insufficient explanatory power.

Keywords Innovation · Indicators · Developing countries · Natural language processing · Emerging countries · ICT sector · Topic modeling · Web mining

✉ Moritz Böhmecke-Schwafert
moritz.boehmecke-schwafert@tu-berlin.de

Colin Dörries
colin.doerries@googlemail.com

[1] Department of Economics and Management, Chair of Innovation Economics, Technical University of Berlin, Berlin, Germany

Published online: 25 November 2023

JEL Classification O30 · O33 · C81 · C88

Introduction

Innovation is the primary driver of economic growth (Aghion et al., 2014; Aghion & Howitt, 1992) and an opportunity for developing and emerging countries (emerging countries hereafter) to leapfrog development stages and catch-up with industrialized countries (Fagerberg & Verspagen, 2007; Fagerberg et al., 2010). Therefore, measuring innovation accurately and efficiently is a salient quest for policymakers to encourage further innovative activity (Dziallas & Blind, 2019; OECD, 2018) and conduct mission-oriented research and innovation policy (Cantner & Vannuccini, 2018). However, the established indicator landscape lacks timeliness and accuracy (Dziallas & Blind, 2019; Kleinknecht et al., 2002). Traditional indicators are primarily based on input and output measures, such as R&D expenditures, analyses of intellectual property protection (e.g., patents), surveys, literature-based output indicators, and other indicators based on science and technology resources (Becheikh et al., 2006; Dziallas & Blind, 2019, Böhmecke-Schwafert & García-Moreno, 2023).

In the specific case of emerging countries, traditional indicators also tend to underestimate firms' innovations (Cirera & Muzi, 2020; Marins, 2008; Marins, 2008; Zawislak & Marins, 2008) and might not capture the phenomenon that Schumpeter (1934) described as a commercialization of novel combinations of knowledge and resources. Moreover, data is often unavailable, expensive, and challenging to obtain because of limited infrastructure or the underlying institutional frameworks (Casadella & Tahi, 2022; Kleinknecht, 1993). Established innovation questionnaires with broad recognition, such as the Community Innovation Surveys (CIS), have contributed to the provision of economy-wide indicators pertaining to innovation inputs and outputs at the firm level for the European Union since 1993 (Arundel & Smith, 2014). Nevertheless, a significant gap exists in the availability of similar instruments for other regions, notably Sub-Saharan Africa. It is evident that the development of conventional innovation indicators for emerging economies remains limited in scope and suffers from issues of data quality and lack of data in general (Böhmecke-Schwafert & García-Moreno, 2023). Therefore, new datasets and methodologies need to be identified and explored to understand the phenomenon of innovation in these contexts.

In this study, we focus on the Sub-Saharan country of Mauritius that is on the verge of becoming a knowledge and digital economy by transforming the information and communication technology (ICT) sector into the major driving force of the economy (Oolun et al., 2012). Considering the increasing economic importance of the ICT sector as a key enabler of innovation (OECD, 2017), the government has published a roadmap in 2018, setting out plans for Mauritius to boost the digital economy with the target of a technology-driven society (MTCI, 2018). However, this strategic focus has not yet translated into significant innovative activity according to traditional innovation measures: for example, data on patent

filings suggest only shallow patenting activity with filings per year in the low two digits (WIPO, 2021). Additionally, indices such as the World Economic Forum's Global Competitiveness Index or the Global Innovation Index calculated anually by the World Intellectual Property Organization (WIPO) consistently rank Mauritius very low regarding traditional innovation indicators such as overall R&D expenditures or patent applications (Schwab, 2019; Dutta et al., 2019). At the same time, the country is ranked highest of all Sub-Saharan Africa regarding several economic and institutional metrics (Schwab, 2019; Dutta et al., 2019). As previously discussed, gaining a comprehensive understanding of innovation in emerging countries like Mauritius has historically faced significant constraints (Chapman & Boothroyd, 1988; Zawislak & Marins, 2008; Marins, 2008; Cirera & Muzi, 2020, Böhmecke-Schwafert & García-Moreno, 2023). Further, there is usually a higher prevalence of incremental innovation and more challenging business environments in emerging countries, which makes it additionally difficult to distinguish when a product or process is genuinely new or significantly improved, and thus, no traditional data is reflecting this information (Cirera & Muzu, 2020). Therefore, this paper seeks to extend the knowledge base and measurement of innovative activity in emerging countries such as Mauritius and explore an alternative approach based on unsupervised machine learning that can serve as best practice for future research on innovation systems of emerging countries. For this, we developed the following two research questions:

RQ1: What are potential areas of innovation activity in the ICT sector of Mauritius?
RQ2: Do web mining and topic modeling provide sufficient indicators to understand innovation activities in emerging countries?

To address our research questions, we extend previous work on novel innovation indicators from Kinne and Axenbeck (2020), Kinne and Lenz (2021), and Böhmecke-Schwafert and García-Moreno (2023) by a country-level analysis: we combine advances in web mining with topic modeling, more specifically with Latent Dirichlet Allocation (LDA), harnessing unstructured data from ICT companies' websites. Our methodological framework not only serves as a robust solution for addressing the challenges associated with data scarcity and quality, usually present in emerging countries, but also exhibits versatility in its application, making it readily adaptable to other socioeconomic contexts where conventional metrics may fall short in providing in-depth explanations. The efficiency of data extraction through web mining, coupled with the interpretive power of topic modeling, positions our methodology to offer sectoral insights in other areas than innovation measurement (e.g., in management studies). Furthermore, it serves as a valuable tool for conducting robustness checks on alternative innovation indicators as well as traditional innovation indicators, thus augmenting the comprehensiveness of innovation research by triangulation. Moreover, we present a lean and comprehensively described methodology committed to the open science principles: we utilize free and accessible data sources. For example, any list of websites can be analyzed with this approach. Further, all tools and software used for the analytical

approach are well documented in the "Methodology" section to be reproduced and are entirely based on open source. Lastly, we provide all components of the study in an open GitHub repository[1] (GitHub, 2023a).

The paper is structured as follows: In the following subchapter, we provide the study's theoretical background and introduce the properties of innovation indicators. Moreover, we present the ICT sector of Mauritius and its evolution as the primary subject of investigation. Besides, we briefly discuss different approaches to natural language processing techniques. In the third chapter, we introduce the methodological approach. The fourth chapter describes the result, followed by a discussion and empirical data and academic literature in Mauritius in the fifth section. The sixth chapter discusses the limitations of the study and presents avenues for future research highlighting the potential of the analytical approach to serve as a blueprint for future innovation studies in emerging countries and beyond. Finally, the seventh section summarizes the main results.

Background

The following subchapters provide an overview of innovation indicators in the context of emerging countries, their data sources and guidelines, and the current state of research on innovation indicators in emerging countries. The subsequent subchapters describe the evolution of Mauritius to an ICT-driven economy and present the current state of ICT innovation in the country.

The Measurement of Innovation Phenomena

An Introduction to Innovation Indicators

According to the *Oslo Manual 2018* from the OECD (2018), "an innovation indicator is a statistical summary measure of an innovation phenomenon observed in a population or a sample thereof for a specified time or place" (OECD, 2018, p. 214). In other words, innovation indicators serve the statistical analysis of innovation, innovative behavior, and other inventive activities. Inventive activities can be understood as measures taken by a company to develop an innovation. These measures can be developmental, financial, or commercial activities (OECD, 2018).

The evaluation of innovation is not a trivial task. It is discussed extensively in scientific literature, and various approaches and methodologies for analyzing innovation are proposed. Dziallas and Blind (2019) conducted a comprehensive systematic literature review on innovation indicators published between 1985 and 2015. Their review identifies 82 unique indicators to evaluate innovations.

Becheikh et al. (2006) distinguish between *direct* and *indirect* indicators. Indicators are, for example, patents (Hagedoorn & Cloodt, 2003) and R&D spending

[1] GitHub repository of this paper: https://github.com/MoritzBS/innovation_indicators_through_topic_models.

(Flor & Oltra, 2004). Direct indicators are, for example, the number of new products (Cooper & Kleinschmidt, 1993) and the share of ideas that can be commercialized (Dewangan & Godse, 2014; Dziallas & Blind, 2019).

Data Sources and Guidelines for Innovation Indicators

The OECD's (2018) *Oslo Manual* provides guidelines for constructing innovation indicators. Innovation indicators can consist of multiple data sources. Even data that was not explicitly designed to support the statistical analysis of innovation can be harnessed for innovation indicators (OECD, 2018). Hagedoorn and Cloodt (2003) support the idea of multiple data sources. They propose a composite indicator that can help assess companies' innovative performance. In this case, the indicators range from R&D spending, patent counts, and patent citations to new product announcements (Hagedoorn & Cloodt, 2003). Other data sources include surveys, institutional and administrative data, and the internet. However, Freeman and Soete (2009) argue that the relevance of "traditional" or "classical" indicators should be revised constantly. Those indicators may no longer have the same importance and meaningfulness today and may lead to misleading conclusions (Freeman & Soete, 2009). Kleinknecht et al. (2002) argue that the two widely used indicators, *R&D spending* and *patent applications*, have more shortcomings than assumed (e.g., diverging patent valuations, time lags). Therefore, a variety of sources for the construction of indicators is essential. According to the *Oslo Manual*, the abundance of data and the increasing availability, for example, can lead to novel innovation indicators based on data from the world wide web. Additionally, data collection methods can be improved through automation. This functions as a critical factor for extending the variety of data sourcing options (OECD, 2018).

The *Oslo Manual* summarizes a set of four properties that innovation indicators should fulfill (OECD, 2018, p. 215): First, *relevance* summarizes that the innovation indicator should serve the need of actual and potential users, e.g., policymakers or other decision-makers such as companies or investors. Second, the innovation indicator should represent the innovation phenomenon in an unbiased way and provide a certain degree of *accuracy and validity*. Third, an innovation indicator requires *reliability and precision*, which means that the results of the indicator should be reproducible with a high signal-to-noise ratio. Lastly, the *Oslo Manual* highlights the relevance of the *timeliness* of innovation indicators. Their construction should be available on a sufficiently timely basis to be useful for decision-making. These four properties will later be the basis for the evaluation of our methodological approach.

Evaluation of Innovation in Emerging Countries

Innovation is equally important for emerging countries as it is for industrialized countries and provides promising avenues for countries to escape the low-income and development trap (Fagerberg et al., 2010). Effective mission-oriented research and innovation policy relies on the accurate measurement of innovation (Cantner & Vannuccini, 2018). Since the *Oslo Manual* guidelines mainly focus on OECD member countries, different requirements and attributes from emerging

countries might not be covered comprehensively. Nevertheless, the authors claim that "recommendations are relevant to both developed and developing countries so that the manual provides effective global guidance" (OECD, 2018, p. 29).

Traditional innovation indicators such as patents or R&D expenditures are limited in explaining the phenomenon of innovation in emerging countries (Fagerberg et al., 2010; Marins, 2008, Böhmecke-Schwafert & García-Moreno, 2023). Marins (2008) analyzed the accuracy of traditional innovation indicators in emerging countries, focusing on Latin American countries. Their study concluded that traditional indicators are not appropriate, and hence, they suggest a new set of indicators that instead focus on the process of how companies innovate (Marins, 2008). Fagerberg et al. (2010) argue that financial and opportunity costs for filing intellectual property such as patents in emerging countries might be too high for inventors; hence, many innovation activities remain unrecognized (Fagerberg et al., 2010). Therefore, researchers and policymakers should rather focus on the measurement of capabilities such as science, research, and innovation openness (e.g., scientific publications), investment openness (e.g., FDIs), or the quality of ICT infrastructure (e.g., penetration of broadband connections) among others (Fagerberg et al., 2010). Still, data for these capabilities might be difficult to obtain in emerging countries. Moreover, Fu et al. (2018) collected data with a unique innovation survey of 501 firms in Ghana. Their methodological approach relied on the structural model as proposed by Crépon et al. (1998) to analyze innovation, which implies that productivity is a function of innovation output, and the latter is a function of research investment. The survey captured the innovation activities of firms and their differences in absorptive capacity, formality, or size (Fu et al., 2018). In addition, Goedhuys et al. (2014) analyzed the influence and importance of different knowledge sources (R&D expenditures, or the education of managers) of firms in five emerging countries using data from survey and face-to-face interviews. Their results reveal strong sector-specific differences. While the food processing firms with higher educated management and foreign ownership are more productive, the more productive firms in the textile sector rely on their own R&D and the imports of new machinery and equipment (Goedhuys et al., 2014). In summary, the literature on innovation in emerging countries is primarily based on input and output measures considered traditional innovation indicators that carry several limitations (Becheikh et al., 2006). Kinne and Axenbeck (2020) summarize four shortcomings of traditional indicators to measure innovation in their study on firm-level innovation in Germany: namely *granularity*, *timeliness*, *costs*, and *coverage*. In the following, we briefly describe these shortcomings in the context of emerging countries:

- *Granularity*: Traditional indicators suffer from insufficient sectoral and technological granularity. Our analysis focuses on the ICT sector mainly, and sector-specific indicators for innovation and R&D are rare for emerging countries.
- *Timeliness*: Traditional indicators such as patents are time-lagged and depict the state of the technology system as it was months or even years previously. For dynamically developing emerging countries, the timeliness of innovation indicators is essential.

- *Cost*: Traditional indicators often involve high data collection costs, primarily when conducted on a large scale. Particularly, institutions in emerging countries often do not have the funds to finance these data collection processes.
- *Coverage*: Traditional indicators often cover only a fraction of the overall firm population. For example, innovation data based on questionnaires could underlie several biases (e.g., selection and self-reporting bias) which might be even more problematic in rural emerging countries with less accessibility.

In conclusion, traditional indicators do not fulfill at least two of the properties of innovation indicators according to the *Oslo Manual*, namely *relevance* and *timeliness*. In the following, we want to focus on the ICT sector of the emerging country of Mauritius that is subject to our analysis. Although the country has excellent framework conditions, metrics of traditional innovation indicators are underperforming.

The ICT Sector in Mauritius

The Evolution of Mauritius Towards a "Cyber Island"

In the last decade, the ICT sector evolved to be one of the five pillars of the Mauritian economy along with the manufacturing, tourism, agriculture, and financial service sectors (Soyjaudah et al., 2002; Seechurn et al., 2013). For centuries, one of the most important economic resources of the Mauritius economy has been sugar cane and the processing of this resource in various forms. In the early 2000s, the tourism sector superseded the agricultural sector. It developed to be an essential foundation of the Mauritius economy, being the strongest foreign exchange–earning sector in the late 1990s (Zafar, 2006) and accounting for 23.8% of the GDP in 2017 (Turner, 2018). In the early 2000s, the Mauritius government launched a strategy to turn Mauritius into a "Cyber Island" and prepare the ICT sector as the future's dominating driving force.

The first step towards Mauritius being an ICT economy was establishing the Ministry of Information and Communication Technology (MICT), the first-ever ministry dedicated solely to the ICT sector in 1997 (Joseph & Troester, 2013). The foundation was a crucial measure that became a turning point in paving Mauritius's transformation into a knowledge-based economy. The ministry then elaborated the National ICT Strategic Plan in 1998, responding to the rapidly evolving ICT sector's structural changes, namely the evolution in technology, markets, and users' demands (Oolun et al., 2012). Since 2003, the government pursues a dynamic framework to progress the establishment of ICT as a central pillar of the economy of Mauritius, and the initiated policies turned the ICT sector into a regional strength (Soyjaudah et al., 2002). As a result, the telecommunication sector has been liberalized in 2003 and initiated the ICT sector's growth (OECD, 2006). Other policies focused on providing a consistent and stable regulatory framework, implementing appropriate legislation, and providing technical infrastructure and building capacity (Joseph & Troester, 2013). The evolution of the countries' ICT sector is reflected in the ICT Development Index, an indicator from the International Telecommunication

Union that tracks the digital divide of countries and measures their progress towards becoming information societies. The index provides evidence that the undertaken measures for the transition to an ICT economy were deemed successful in the international comparison: With an ICT Development Index of 6.02 in 2016, Mauritius ranked 72nd place out of 176 countries—by far the highest ranking among African countries (Government of Mauritius, 2020). The government of Mauritius is actively fostering an ICT innovation cluster in the country. Through the government-owned company Business Parks of Mauritius Ltd (BMPL), the technology park "Cyber-City" has been set up in Ebène where several renowned international firms in software development and BPO settled (OECD, 2006; POWC, 2022).

Institutional quality is hypothesized to have a positive effect on innovation (Rodriguez-Pose & di Cataldo, 2015; Dunning et al., 2008), and empirical evidence has shown the positive role of institutions particularly for innovation systems of emerging countries (e.g., Rodríguez-Pose & Zhang, 2020). Mauritius is considered to have a robust institutional framework and financial system compared to the countries' peer groups (Schwab, 2019). According to the "Global Competitiveness Report 2019," an annual ranking of countries' competitiveness in various dimensions, Mauritius ranks 29th in institutional quality and even 27th in the quality of the financial system. Figure 1 provides an overview of Mauritius' scores and the corresponding ranks in all 13 dimensions of the report for the year 2019. The circle at each dimension illustrates the score of Mauritius (from 0 to 100, the higher the better); the triangle depicts the average score of peer countries, and the rectangle implies the average of the sub-Saharan country cluster. Overall, the country ranks 52nd above all 13 dimensions and by far spearheads the region of sub-Saharan Africa (Schwab, 2019). Moreover, Mauritius is characterized by solid *business*

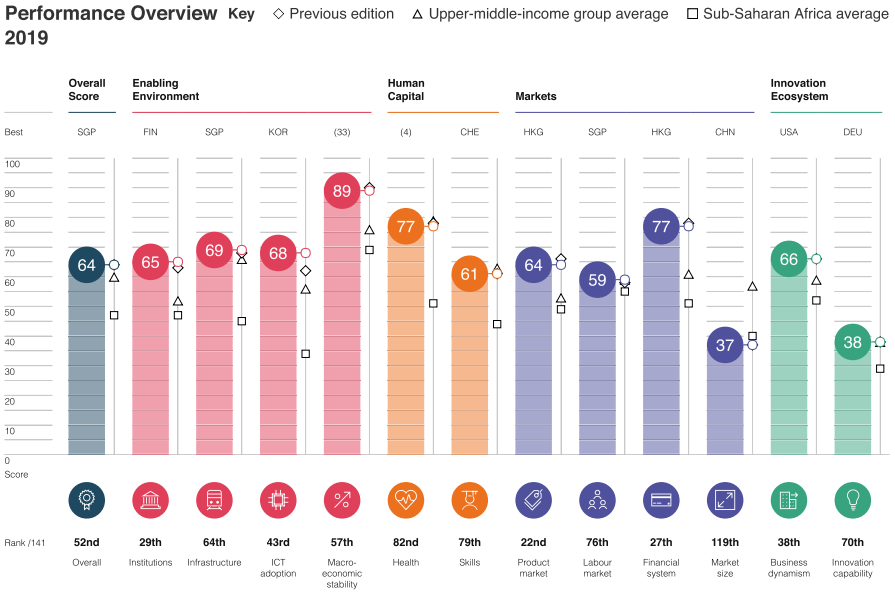

Fig. 1 Country Profile of the Global Competitiveness of Mauritius, Source: Schwab (2019), p. 382

dynamism driven by lean administrative requirements, while it ranks comparably low (70th) regarding the *innovation capability* (Schwab, 2019).

These framework conditions attracted multinational companies (MNCs) to open subsidiaries in Mauritius, and the country developed into a hub for business process optimization (BPO) firms. Apart from the stable institutional framework, other framework conditions such as only three hours of time difference to Central Europe and the bilingual capabilities of the population (English and French) lay a solid foundation for Mauritius being a predestinated BPO hub for European companies.

The ICT industry develops further to being a key sector through innovation and creativity in the future, and several government initiatives have been launched in the previous years (GESCI, 2017; Lim & De Meester, 2016; OECD, 2006). The development of the number of ICT and BPO companies in Mauritius is shown in Fig. 2: Having only 90 companies in 2005, the number of companies more than quadrupled until 2010 to 400 companies increased to 750 companies in 2016. In 2019, the ICT sector contributed an estimated 5.8% of the value-added of the GDP (Government of Mauritius, 2020).

Comparing the value added to the GDP, a worker in the ICT sector adds almost double the value of an agriculture worker (Central Intelligence Agency, 2020). The employment in large establishments in the ICT sector (employing ten or more persons) is estimated to sum up to 16,157 employees in 2018, with a steady increase in the previous years (Government of Mauritius, 2020). It is noteworthy that the workforce is evenly distributed among the genders, with 8544 males and 7613 females working in the ICT sector (Government of Mauritius, 2020). Therefore, the ICT sector is considered a catalyst for growth due to the employment effects. After removing price effects, the sector's growth rate was 5.1% in 2019 (Government of Mauritius, 2020). Nevertheless, the sector's annual growth rate was continuously decreasing since its peak of 8.2% in 2000 (The World Bank, 2020).

According to the Global Competitiveness Report, mobile phone penetration in Mauritius remains very high, as it is relatively usual for sub-Saharan countries, and the country ranks 11th worldwide. However, more strikingly, Mauritius ranks even 15th in terms of internet subscription penetration leading to an overall rank of 43 in ICT adoption (Schwab, 2019).

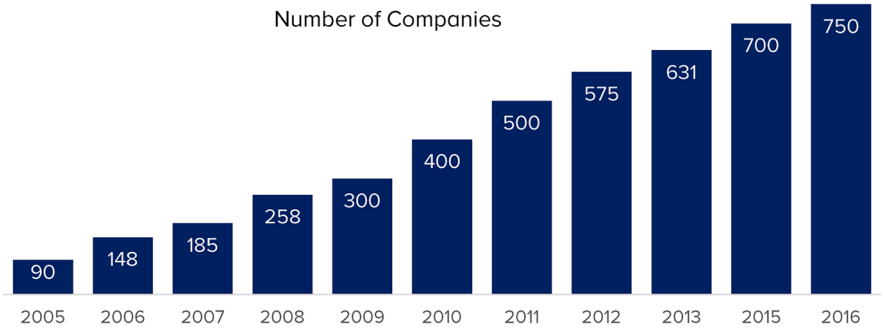

Fig. 2 Development of the number of ICT and BPO companies in Mauritius from 2005 to 2016; Source: Board of Investment (2018)

Table 1 Selected indicators on IP and the corresponding international rank of Mauritius based on the Global Innovation Index (GII) Global Competitiveness Report (GCR)

Indicator	Value	Rank	Source
Patents by origin/bn PPP$ GDP	0.5	82	GII (Dutta et al., 2020)
Patent applications/million pop	0.97	67	GCR (Schwab, 2019)
R&D expenditures % GDP	0.2	95	GCR (Schwab, 2019)
Scientific publications H Index	67.7	121	GCR (Schwab, 2019)

In the annual country ranking on the capacity and success of innovation published by the WIPO, Mauritius ranks overall 52 worldwide (Dutta et al., 2020) and the best in the region of sub-Saharan Africa. The substantial improvement in comparison to 2019 (rank 82) is mainly driven by performance improvement in several categories, namely *institutions* (rank 22), *investment environment* (rank 9), and *education* (rank 36). Further, it is also a consequence of higher data availability and accuracy (Dutta et al., 2020). However, output measurement of innovation (e.g., patent) falls short in Mauritius, as described in the following.

Innovation in the ICT Sector of Mauritius

From an output perspective on the measurement of innovation, it is challenging to develop an estimate for Mauritius' ICT sector due to the low patenting activity in general in the country. Table 1 summarizes selected key indicators for R&D activities and innovation with their corresponding global rank from the Global Innovation Index 2018 (Dutta et al., 2020) and the Global Competitiveness Report 2019 by the World Economic Forum (Schwab, 2019). From 2018 to 2020, Mauritius has substantially improved its rankings. Compared to its relative rank in other evaluation dimensions, Mauritius ranked particularly low as 113th concerning *patents by origin* in 2018. In 2020, this figure significantly improved, and Mauritius climbed to rank 82nd globally.

The low patenting activity over all sectors is further revealed in the detailed analysis of applied and granted patents within the last decade. Figure 3 shows applications and patent grants between 2010 and 2019 based on data from the WIPO (2021).

Applied and granted patents are differentiated depending on whether the applicant was resident, non-resident, or whether a Mauritius company has applied for a patent abroad. The figure shows that the number of patents filed abroad is much higher than patents filed under Mauritius' IP jurisdiction. The number of patent applications filed in the Mauritius market reveals a shallow patenting activity that ranges from only 18 applied patents in 2011 (respectively zero in 2010) to 38 in 2016. The vast majority of patents (i.e., more than 90% in this period) is applied by non-residents (WIPO, 2021) and hence is not necessarily captured by an innovation indicator. Figures on granted patents during this period are even lower and are—additionally—exclusively attributed to non-residents, aggregating to a total of 39 patents. The reasons for this low patenting activities are subject to future research;

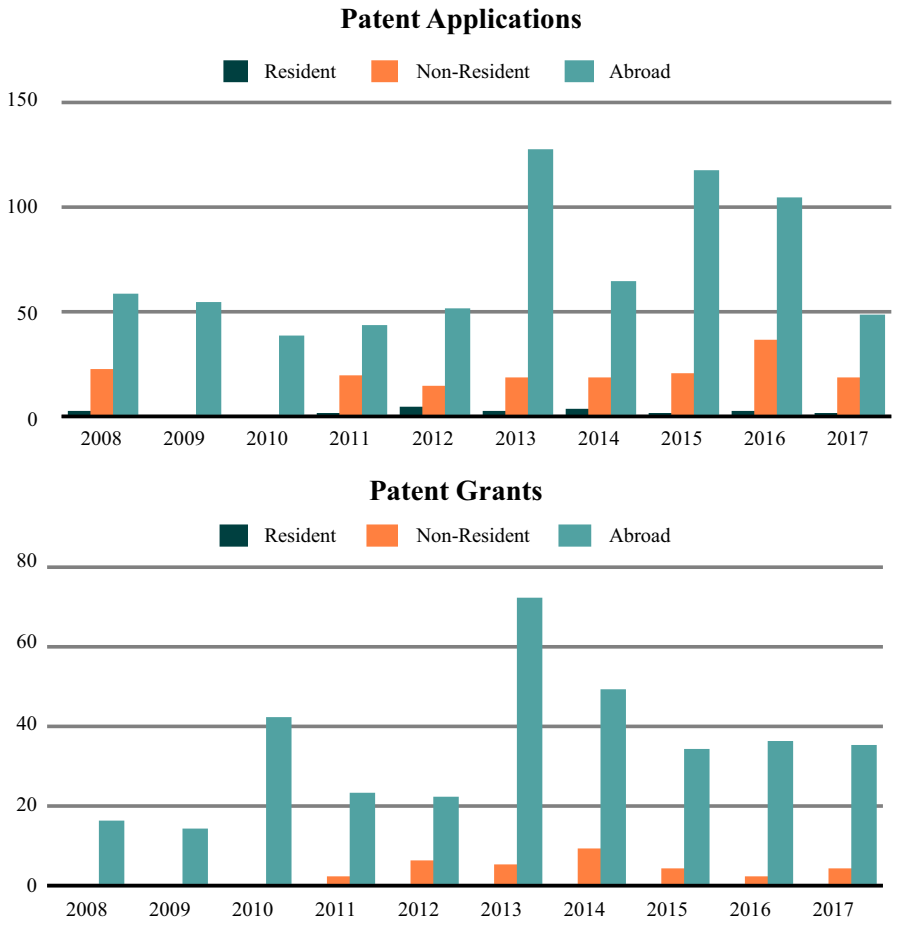

Fig. 3 Patent applications and patent grants from residents and non-residents in Mauritius and Mauritius enterprises abroad between 2010 and 2019. Source: WIPO (2021)

however, we hypothesize that one major reason is that Mauritius has not signed the international Patent Cooperation Treaty (PCT) from the World Intellectual Property Organization that facilitates the filing of international patents being simultaneously protected in all 152 PCT contracting countries if granted. Furthermore, we used the IP analytics tool IPlytics[2] (2022) to further analyze the sectoral distribution of filed patents by Mauritius companies. In the period from 1990 to 2017, the majority of patents is attributed to the "Electrical Engineering" industry (41.2%), which includes, e.g., *computer technology*, *IT management methods*, *digital communication*, and *telecommunication*. The second-largest represented industry is "mechanical

[2] IPlytics is an IP intelligence tool that enables the analysis of patents, SEPs, standards, literature per country (IPlytics, 2022).

engineering" with 31.8% followed by chemistry (14.2%) (IPlytics, 2022). Overall, these data do not provide much insight about sectoral innovation activities.

From an input perspective, the innovative activity can be measured by analyzing aggregated data on R&D expenditures. The Global Competitiveness Report estimates R&D expenditures as a percentage of the GDP and reveals for Mauritius a shallow value of 0.2, which leads to a global rank of 95 in this dimension (Schwab, 2019). Moreover, figures in scientific publications such as the H-Index (that should be cautiously interpreted as its value highly depends on the size of a country) rank Mauritius very low at 121st. Data on co-authorships of scientific publications of IPlytics (2022) suggests that Mauritius's peer-reviewed academic research lacks international cooperation.

Although Mauritius joined the WIPO convention in 1976, the country has not ratified the PCT as described above. The Ministry of Technology, Communication, and Innovation of Mauritius has published a strategic plan Digital Mauritius 2030 (MTCI, 2018) to foster the ICT sector in the country, planning the protection of IPR as an *integral part of an enabling legal and regulatory framework* (MTCI, 2018, 35). Moreover, a SWOT analysis of the ministry highlighted an inadequate IPR regulatory and legal framework as a country's weakness for innovation in emerging technologies (MTCI, 2018). From the synthesis of these various data, we conclude that innovation in an emerging country such as Mauritius is challenging to measure using standard input and output factors.

Methodology

In the following section, we describe the underlying methodological approach. Since a major contribution of this study is a lean methodology to explore innovation activities, we put a particular emphasis on this section. First, we provide background on the analytical approach of text data analysis and topic models. Next, we provide a comprehensive four-step description of the methodology. Lastly, we present the calculation of coherence scores to determine the optimal number of topics.

Text Data Analytics with Machine Learning

In this study, we construct an indicator for potential innovation activity in the Mauritian ICT sector based on unsupervised machine learning and the analysis of unstructured data using natural language processing (NLP) techniques. We collect the unstructured data by scraping company website texts.

Machine learning approaches for text data analysis are manifold. This paper focuses on using NLP, which deals with the understanding of natural (human) language. The "statistical revolution" enabled the breakthrough in NLP. Before, most of the methods for processing linguistics were rule-based systems, which only have a limited ability to understand natural language. A different approach based on statistical methods enabled by machine learning cleared the way for a new era of computer-based NLP today (Chowdhury, 2003).

Typical tasks of NLP are natural language understanding (or interpretation), natural language generation, and speech recognition (Chowdhury, 2003; Rajman & Besançon, 1998). In this paper, we solely focus on natural language understanding as the other two tasks are not substantial for the research questions. In this subfield of NLP, a variety of activities exists. Particularly for the analysis of websites, typical tasks include the extraction of entities, the categorization of content, the clustering of content, and relationships (Sun et al., 2017). For our purpose, clustering content is highly relevant because it identifies main topics or discovers new topics within the text corpora (Sun et al., 2017; Wallach et al., 2009).

Topic Modeling

Topic models are used to discover patterns in text documents and are a type of unsupervised learning. The goal is to process natural language to find topics within text corpora. It is applied in search engines, user recommendation systems, customer service automation, and other related fields (Doll, 2018).

Their statistical language models characterize topic models. There exist multiple topic modeling approaches such as *latent semantic analysis* (LSA) (Landauer & Dumais, 1997), *latent semantic indexing* (LSI) (Dumais et al., 1994), or *probabilistic LSI* (pLSI) (Hofmann, 1999), which use singular value decomposition on the document term matrix based on linear algebra. In addition, there is *non-negative matrix factorization* (NMF) (Xu et al., 2003) which is also based on linear algebra. Furthermore, there is Latent Dirichlet Allocation (LDA) (Blei et al., 2003) which is heavily used in the NLP community and based on probabilistic graphical models (Kapadia 2019).

In this study, we focus on LDA, a generative probabilistic model which is appropriate for text classification, dimensionality reduction, and collaborative filtering (Blei et al., 2003). LDA is based on a three-level hierarchical Bayesian model, which analyzes collections of discrete information such as text corpora. In LDA, each item of a text collection is modeled as a finite combination over an underlying set of topic probabilities. In general, LDA tries to solve modeling text corpora or other discrete data (Blei et al., 2003). It can therefore be used to discover topics in text documents automatically. An LDA model represents text documents as mixtures of topics based on words within the document that occurs with a certain probability (Blei et al., 2003; Tokunaga et al., 2008).

Our methodological approach is modular and individual steps can be reproduced for similar studies on emerging countries. Our approach is an adaption of the firm innovation prediction model proposed by Kinne and Lenz (2021) for the context of emerging countries. The sequence of steps is depicted in Fig. 4. The first step requires a curated list of company websites. In the following steps, unstructured text data is scraped from websites. After data wrangling and cleaning, we decontextualized the data and applied topic modeling. Lastly, we re-contextualize the results by developing a coding structure for the identified topics. In the following, we describe our methodology in four steps.

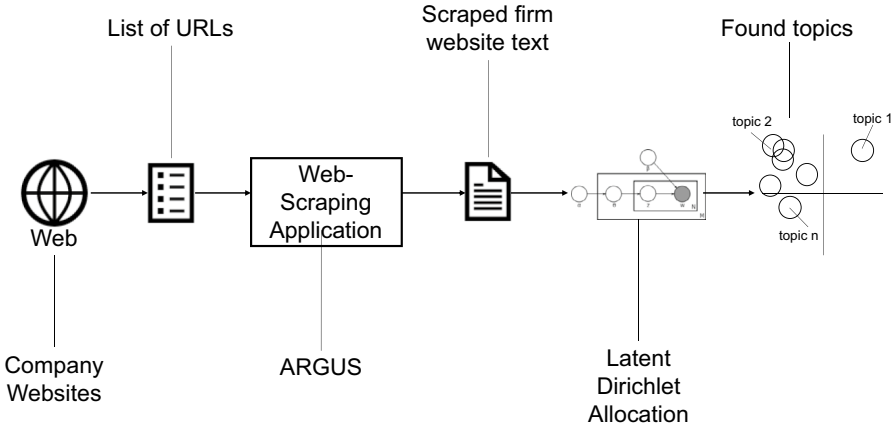

Fig. 4 Proposed approach adapted from Kinne and Lenz (2021)

4-Step Methodological Approach

Step 1—Data Collection

In the first step, we constructed a list of URLs for relevant ICT firms from Mauritius as an input for the web scraping (see Fig. 4). This step can be performed by simply contacting institutions or conducting online research using different keywords.

Our dataset consisted of two company directories: the first directory was the outcome of an extensive online search using different keywords (e.g., "ICT" AND "Mauritius"). We identified a company directory with websites of 347 Mauritian ICT companies provided by the National Computer Board (NCB), a parastatal organization under the Ministry of Technology, Communication, and Innovation (MTCI) of Mauritius (National Computer Board, 2018). The NCB directories' purpose is to provide comprehensive information on Mauritian ICT companies with the focus of attracting international investors and businesses looking for outsourcing opportunities. The directory was created in 2018 and provides information about the listed companies' activities, expertise, products, and services, as well as contact details. The NCB categorized the companies according to their offered products and services based on data provided by operators and members from local industry associations (National Computer Board, 2018). The most prevalent categories are *consultancy services* (152 firms), *hardware and infrastructure* (132), and *software development* (124 firms). Figure 7 in the appendix depicts the detailed categorization of the NCB (2018).

The second company directory is based on a list of 270 companies and their contact details compiled by the Mauritius Research and Innovation Council (MRIC) in July 2019 as part of another scientific project.[3] The MRIC received different ICT

[3] We want to express our gratitude to the MRIC who provided the list as part of the project *The role of Innovation and IP in the ICT sector of Mauritius*.

company lists from the NCB and the Economic Development Board, a statutory body in Mauritius. Since the list did not contain any websites, we used the email contact details and search engines to identify the URLs of the listed companies. After merging both datasets, eliminating duplicates, and curating the data, the total dataset consisted of 427 unique websites of ICT firms.

Step 2—Web Scraping

In the next step, the collected URLs are scraped with the open-source tool *ARGUS*. ARGUS has been developed for the study from Kinne and Axenbeck (2020) and was since then applied in other innovation studies to scrape web data of companies, also in an emerging country context (see, e.g., Kinne & Lenz, 2021, Böhmecke-Schwafert & García-Moreno, 2023). ARGUS was developed to meet the complex requirements of web mining and is based on the Scrapy Python framework (GitHub, 2023b; Kinne & Axenbeck, 2020). The tool is designed to easily scrape a wide variety of web contents from any website, producing output in a structured and consistent format. ARGUS is scalable; it can scrape tens of millions of web pages from millions of different websites in a reasonable time frame (GitHub, 2023b; Kinne & Axenbeck, 2020).

Additionally, ARGUS is easy to use, enabling researchers to use the web-scraping tool without the need for extensive knowledge of web-scraping technology featuring a graphical user interface (GUI) that allows researchers to control web scraping without using command lines. The GUI also provides options such as parallel processing of scraper processes, scraping limit, and preferential downloading of shorter hyperlinks. ARGUS can be downloaded from GitHub (see GitHub, 2023b) and requires the installation of additional Python packages. Users must also install cURL and add it to system environment variables. Finally, users must provide a list of website URLs and IDs to run the ARGUS scraper, which is processed and saved in an output file with structured information such as website content, metadata, URLs, and timestamps (GitHub, 2023b).

ARGUS can automatically scrape raw text data from websites using a CSV table of URLs as an input. During the scraping process, all text data is being downloaded from the website. However, *ARGUS* does not scrape websites randomly but starts at the main page and then continues to lower-level sub-pages. The parameter *scrape limit* was set to 100 because this threshold offers a balanced trade-off between general descriptions of firms and precise information. Kinne and Axenbeck (2020) showed that a *scrape limit* of 50 is sufficient to scrape more than 60% of a website. However, to scrape 90% of a website's content, the limit must be increased to 250 (Kinne & Axenbeck, 2020).

The basis of the web scraping process is the data set of 427 URLs as described in the "4-Step Methodological Approach" section. *ARGUS* allows to define a preferred scraping language to scrape, which is relevant for Mauritius, since the country is bilingual in French and English. Therefore, two scraping processes were initiated, and a total of 427 bodies of website texts were extracted. The purpose of conducting two separate scraping processes was to achieve a higher data quality in the scraping output, because the text of the selected language was preferred considering the

restrictions of the *scrape limit.* In summary, 351 company websites were in English (82% of the total dataset), whereas 76 URLs were in French (18% of the total dataset). *ARGUS* generates a dataset with the scraped text corpora as well as metadata as output. In total, *ARGUS* could retrieve data from 325 company websites (76%). The other 102 websites have not been responsive. This could have various reasons, for example, technical issues of the web scraper *ARGUS*: while only a very small fraction of websites is written in JavaScript, we observed that *ARGUS* runs into technical issues when scraping JavaScript websites. However, we assume that most of the share of unresponsive websites is a result of these firms not operating anymore or under different names. After two separate scraping runs, the output was merged into one data set that contained all web scraping results.

Step 3—De-contextualization

In the third step, we applied NLP techniques to extract information from the unstructured text data. For similar studies, it is also possible to perform the following steps with other types of unstructured data, such as patent counts, related research papers, or government reports. The objective is to identify reoccurring topics within text corpora. For this study, we applied the NLP algorithm of Latent Dirichlet Allocation (LDA). Blei et al. (2003) developed LDA as a generative probabilistic model which can be used for text classification, dimensionality reduction, and collaborative filtering. This can be understood as a type of unsupervised learning method to identify clusters of thematically related word groups while not requiring a pre-trained model (Blei et al,, 2003). LDA envisions each document as a blend of topics, with a particular distribution of topics unique to each document. This distribution of topics is the result of two steps: first, the choice of topics for the document, and second, the selection of words from the chosen topics (Blei et al., 2003; Linton et al., 2017). LDA assigns to each document a probability distribution over topics, signifying that documents can encompass multiple topics in varying proportions (Blei et al., 2003). Simultaneously, LDA assigns to each topic a probability distribution over words in the vocabulary. This distribution characterizes the likelihood of particular words occurring in the context of that topic.

Standard preprocessing and data cleaning steps in NLP are necessary to enhance the quality and consistency of textual data, enabling accurate and meaningful analysis (Yogish et al., 2019). We followed the standard preprocessing and data cleaning steps for NLP analyses (see, e.g., Yogish et al., 2019). These steps include tokenization, stop word and punctuation removal, and lowercasing (see, e.g., Linton et al., 2017) that are described in the following. Tokenization describes the activity of breaking down a document into meaningful words (Yogish et al., 2019). The removal of stop words cleans the corpora of non-significant words that do not carry any relevant information in the context of this study (e.g., "the," "she," "it"). According to Schofield et al. (2017), the removal of frequent stop words can improve topic coherence and classification accuracy. However, since the ICT sector spans across diverse applications, we decided to not add more "topic-specific" stop words, to avoid any risk of author-induced biases. We used the widely spread Python library *Natural Language Toolkit* (NLTK) for the preprocessing steps and the inherent stop words

corpus for the exclusion of stop words. The library contains more than 2400 stop words and is available in English and French (Yogish et al., 2019). Lastly, removing punctuation marks such as periods, commas, and exclamation points and ensuring lowercasing reduces further noise in the text data, focusing the analysis on essential words and phrases (Yogish et al., 2019).

After the preprocessing steps, the final text corpus comprised 46,863 words. Due to the bilingualism of Mauritius, we retrieved website data in both French and English language. However, the majority was in English. Many French websites have a high prevalence of English keywords such as "cloud computing," "business process optimization," or "AI." After randomly selecting French websites and cross-checking our assumptions, we concluded that most Mauritian ICT companies use the same English terminology on their web pages, and thus, we decided against translating the French websites into English to avoid any risk of author-induced biases.

In the last step, the number of topics for the topic model is determined. While this number is not arbitrary, there is no universal approach to solving this specific trade-off, and various researchers have assessed the importance of setting the number of topics (Panichella, 2021; Panichella & Poshyvanyk, 2013; Agrawal et al., 2018). On the one side, if we include only too few topics, their description might be too general. On the other side, if we determine too many topics, their specificness complicates the analysis and interpretation process. Therefore, we use two different types of coherence scores to identify the optimal number of topics, the coherence score *C_Umass* by Mimno et al. (2011) and the coherence score *C_V* by Röder et al. (2015). The coherence score describes how the topics and their tokens in the topic model describe the inherent information of the unstructured data (Röder et al., 2015; Kapadia, 2019). The result is a topic model with a specific number of topics that need qualitative assessment and coding while synthesizing the most prevalent tokens per topic.

Step 4—Re-contextualization

In the last step, we visually analyzed the results from the intertopic distance map (ITDM) generated with the Python library PyLDAvis. These ITDMs serve as a graphical representation within a two-dimensional space, providing a visualization of topics. In this representation, each topic is denoted by a circular element, and the relative size of these circles is proportionate to the cumulative frequency of words attributed to each topic (Sievert & Shirley, 2015). The positioning of topics within the ITDM is a direct reflection of their semantic content, specifically the thematic commonalities shared among them. Consequently, topics that exhibit closer proximity within this spatial representation, share a more substantial overlap. PyLDAvis generates an interactive HTML file that users can browse and explore, such as the 30 most salient terms behind each topic. The visualizations enabled us to gain an initial overview of the relationships and thematical distances between topics within the corpus (Sievert & Shirley, 2015).

Thereby, it provided the starting point for the following iterative qualitative coding process which was partially adapted from Miles et al. (1994). This iterative process is a collaborative effort among the co-authors to ensure a thorough examination of the topics and to reduce the risk of bias. In the first iteration, each co-author focused on

the top five tokens within each topic. This initial coding process aimed to provide a preliminary characterization of the topics based on the most salient keywords. Subsequently, a second review was conducted individually, expanding the analysis to the top ten tokens within each topic. This broader scope allowed for a more comprehensive understanding of the topics. In the final phase of coding, another more detailed review was individually performed based on the top 30 tokens for each topic. Following these individual coding stages, the co-authors engaged in an in-depth discussion with the aim of converging on the final titles for each topic model. It involved a comprehensive review of the coding results, a synthesis of the insights gained from each iteration, and a consensus-building process to determine the most appropriate and descriptive title for each identified topic. This qualitative analysis approach not only provided depth and richness to the topic modeling results but also facilitated a holistic understanding of the underlying themes and patterns within the text. After the finalization of the coding, we further assess the robustness of the categories, based on a sectoral categorization of ICT companies in Mauritius by the National Computer Board (2012).

Calculation of the Optimal Number of Topics

Figure 5 shows the results from both topic coherence score calculations for different amounts of topics. The plots show that both coherence scores increase naturally with the number of topics. However, the slope decreases, and for large numbers of topics, keywords often repeat, which reduces the explanatory power of the model. *C_Umass* on the left-hand side in Fig. 5 suggests that the coherence score fluctuates heavily with different local maxima (e.g., at 5 and 7 topics) and then increases to an optimal amount of 12 topics. In the following, the slope of the coherence flattens with an increasing number of topics. Looking at the coherence score *C_V* on the right-hand side, we observe a local maximum at seven topics and coherence then decreases again for 12 topics in contrast to *C_Umass*. Subsequently, it rises again to a global maximum at 20 topics. However, with this amount of topics, the model would not have much explanatory power, and the risk overfitting increases (Mimno et al., 2011; Röder et al., 2015). Since 12 topics are not optimal for both scores,

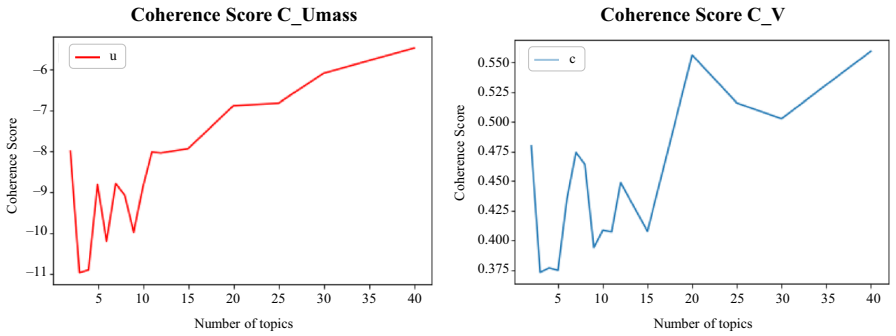

Fig. 5 Comparison of the topic model coherence scores C_Umass and C_V

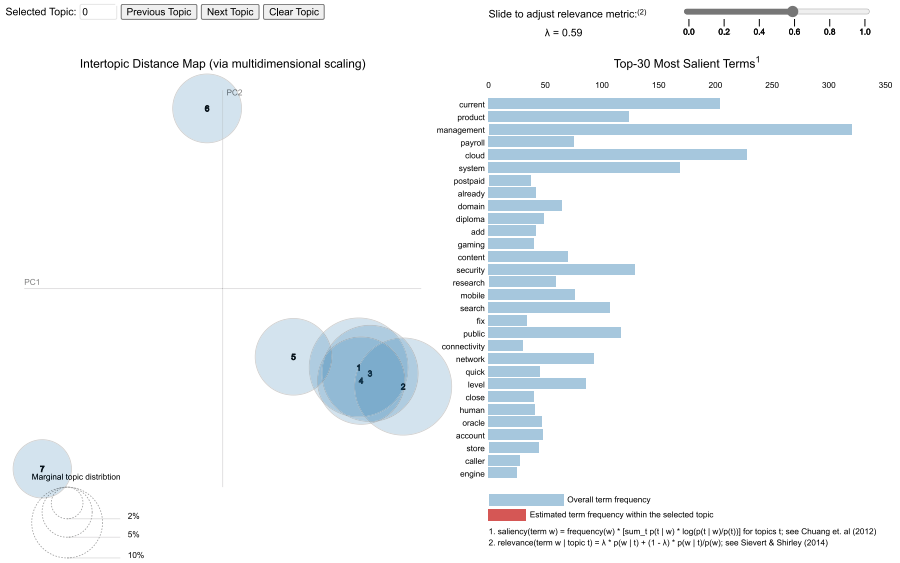

Fig. 6 Intertopic distance map of the seven topics and the 30 most salient words of topic 1. An interactive HTML file in the GitHub (2023a) repository of this paper shows all seven topic models and their top words

while both scores show local maxima for seven topics, we conclude from the analysis of coherence scores *C_Umass* and *C_V* an optimal amount of seven topics.

Results

In the following sections, we describe the results of the topic modeling approach. First, we discuss the ITDM, a visualization of the identified topics, and second, we present the topic models and their coding.

Inter-topic Distance Map

We calculated the ITDM for the ideal amount of seven topics using PyLDAvis. This Python-based visualization tool allows us to analyze the semantic distance of topics and the frequency of the underlying tokens in a two-dimensional space (Sievert & Shirley, 2015). The ITDM for our corpus is shown in Fig. 6. The interactive visualization of the models is provided as an HTML file in the GitHub (2023a) repository of this paper.

On the left-hand side, the seven topics are depicted in a two-dimensional space. The first dimension of the two-dimensional space labeled principal component 1 (PC1) represents the primary dimension that captures the most significant source of variation among the topics in the ITDM. The second dimension, labeled principal component 2 (PC2), on the other hand, is the second primary dimension orthogonal to PC1. It captures the second most substantial source of variation among the topics.

The size of the circles represents the topics' share on all tokens. PC1 and PC2 are often used in data analysis and visualization to reduce complex data to a few key dimensions that can be easily understood and visualized.

Using the interactive HTML output, we can switch between these seven topics and see the 30 most prevalent words within each topic (see the list on the right-hand side). The ITDM shows that topics 1, 3, and 4 are very close and almost completely overlap. This indicates that they are thematically related and share a certain amount of similar tokens. Topics 2 and 5 are only slightly more distant and are overlapping most of the previously mentioned topics. Hence, we can consider topics 1–5 a cluster with a high degree of similarity among their tokens. In comparison to that, topics 6 and 7 each are far more distant, implying a certain uniqueness from the other topic cluster. The relative prevalence of tokens from topics 1 and 2 is the largest in this model, as indicated by the circle size.

Summary of the Topic Model Coding

In the next step, we coded the seven topics according to the procedure described above. The results are category titles for each topic model that describes an innovation category of the Mauritius ICT sector. Following Miles et al. (1994), the collaborative process involved the individual coding of topic models by each co-author, subsequently followed by an exhaustive consolidation of our findings through comprehensive deliberations. Notably, the coding of topics encountered spirited debates among the co-authors, particularly due to the substantial overlap observed among topic models 1 to 5. To address these instances of overlap, a rigorous analysis was conducted, focusing on the examination of the 30 most salient terms and comparing them against each other from overlapping topics. This scrutiny facilitated the identification of commonalities and distinctions, enabling a more nuanced understanding of their relationships. Table 2 depicts the seven topic models with their tokens' prevalence, the ten most prevalent tokens, and the coded category title.

The first topic is coded as *BPO and management services* and represents the relative majority of tokens with a share of 19.5% tokens. The most prevalent tokens such as *management*, *system*, *customer*, *human*, and *service* clearly relate to the categories' final title.

The second topic consisted of 18.8% of tokens and is titled *cloud.* Although the most prevalent token *cloud* mainly determines this category, other keywords such as *innovation, public* (public cloud), and *technology* point to this technology category.

Security is the title of the third topic model that comprised 18.5% of the tokens. This interpretation is mainly driven by the most prevalent token *security* followed by other tokens such as *global* or *industry.*

The fourth topic (15.2% token share) was highly debated and comprised tokens such as *business*, *consulting*, *financial*, *or trust.* Based on these keywords, the category is entitled *consulting*. However, it also contains tokens that we could not interpret in a similar context (e.g., *dating* or *spoon*).

Table 2 Coding summary of the seven topic models

	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**
Category	**BPO and management services**	**Cloud**	**Security**	**Consulting**	**Web development**	**Mobile connectivity**	**Generic**
Share of tokens	19.5%	18.8%	18.5%	15.2%	11.7%	9.4%	6.8%
Top-10 most salient terms	Management	Cloud	Security	Business	Domain	Postpaid	Current
	System	Business	Store	Consult	Content	Gaming	Product
	Payroll	Diploma	Oracle	Trust	Search	Fix	Already
	Service	Level	Industry	Provide	Engine	Mobile	Add
	Customer	Research	Contact	Inventory	Server	Connectivity	Cisco
	Human	Technology	Global	Evolution	Translation	Quick	Kitchen
	Marketing	Innovation	Travel	Dating	Virtual	Account	Electronics
	Public	Support	Analytics	Spoon	Browser	Network	Gardening
	Design	Information	Category	Financial	German	Caller	Sound
	Productivity	Public	Brand	First	Talent	Close	Leisure

The fifth topic with only 11.7% token share is coded as *web development*. The context of web development and the internet was unanimously agreed upon among the co-authors based on tokens such as *domain*, *search*, *engine*, *server*, and *browser.*

Another topic where less debate among the co-authors occurred was the sixth topic, with a token share of 9.4%. The tokens *postpaid*, *gaming*, *mobile*, *connectivity*, and *network* determined the category title *mobile connectivity*.

Finally, the seventh topic model only accounted for 6.8% of the tokens. It was entitled as generic because the co-authors could not interpret the most prevalent 30 tokens to a meaningful category (see tokens in column 8 of Table 2).

In the following sub-chapter, we will evaluate these findings based on other research on the Mauritius ICT sector and assess the value of the approach and resulting categories as an indicator of innovation.

Discussion and Implications

The main result of our explorative analysis of company website data is a categorization of potential innovation activity in the ICT sector of Mauritius to answer the first research question: *What are the potential areas of innovation activity in the ICT sector of Mauritius?* In the following, we discuss the implications of these findings along with previous research findings and data on ICT innovation in Mauritius.

The generated topic model suggests that the ICT innovation landscape in Mauritius is composed of the following seven categories: *BPO and management services*, *cloud*, *security*, *consulting*, *web development*, and *mobile connectivity*. Besides, we also coded a generic category that we could not specify due to a high degree of heterogeneity of its tokens.

The first and most significant category *BPO and management services* is in line with our expectations and previous research on the ICT sector in Mauritius, assuming the sector is driven by BPO and outsourcing of multinational companies (MNC) (Joseph & Troester, 2013). Despite its geographic isolation, the country has several competitive advantages that make it a predestined location for MNC's BPO activities such as IBM, Accenture, HP, Huawei, or Microsoft: a stable institutional framework, little time difference to major markets in Central Europe (only three hours), bilingual capabilities of the population (English and French), a high degree of gender equality of the workforce, or high ICT literacy (Schwab, 2019; Oolun et al., 2012). The government of Mauritius actively fostered an ICT BPO cluster in the early 2000s and built a technology and science park called "Cyber-City" in the city of Ebène. These science parks can play an important role in innovation and entrepreneurship in regional innovation systems of emerging economies (Chang et al., 2012). "Cybercity" consists of multiple-story office buildings and is equipped with high bandwidth fiber optic cables. Since its opening in 2004, "Cyber-City" has attracted several renowned international firms (e.g., Accenture) in software development and BPO firms growth (OECD, 2006; POWC, 2022). However, a recent study by Diez-Vial and Fernández-Olmos (2017) indicates younger firms benefit the most from being located in these science parks with increasing innovative capacity and higher business growth than established firms (Diez-Vial & Fernández-Olmos, 2017).

We assume that BPO companies' innovative activities predominantly focus on process innovation due to the nature of their business to remain competitive in the global BPO market. This assumption is supported by an econometric analysis of firm-level data from Bertschek et al. (2017) who analyzed a data set comprising 1452 firms from the German manufacturing and services sectors, however from the outsourcing firms' perspective. The authors provided evidence for a significant positive relationship between IT outsourcing and process innovation with differences depending on the sector (Bertschek et al., 2017). Process innovations, incremental or radical, are not necessarily labeled as innovations or in some cases cannot be appropriated. Moreover, studies have shown that firms more often choose secrecy for process innovation, over formal appropriation (e.g., with patents), a phenomenon less observed for product innovations (see, e.g., Biswas & McHardy, 2012). Since non-appropriated process innovations do not appear in any statistics (e.g., the patent office statistic), it has been hard to measure these innovative activities in Mauritius. However, with the *Industrial Property Act 2019*, the Government of Mauritius proclaimed a bill that introduces utility models to protect intellectual property with less stringent requirements. The introduction of utility models significantly increases the incentives to protect new processes as the costs and likelihood of failure decrease compared to patent applications (Heikkilä & Lorenz, 2017). Therefore, we expect more measurable innovative activity from the first and largest category of ICT innovation activities.

The second category, entitled *cloud*, is highly related to the former, as many BPO services rely on cloud computing technology. Overall the Mauritius economy consists of many small and medium enterprises (SMEs) that benefit from cloud computing, for example, that makes software as a service (SaaS) easily accessible or helps SMEs to scale and react to seasonal demand due to its flexibility to cost structure and scalability (Antoo et al., 2015). Hence, it is essential that firms in the ICT sector also significantly focus on providing cloud services for the SMEs in the national economy. In addition, the Mauritian government actively fostered the adoption and started to provide an affordable package for SMEs in 2013 (Antoo et al., 2015). Besides, cloud computing is recognized as one of the major emerging technological trends by the Mauritius government in their strategic plan for the ICT sector called *Digital Mauritius 2030* (MTCI, 2018). Therefore, it has significant opportunities to provide cloud-based services for national and international markets (MTCI, 2018). Furthermore, cloud computing can undoubtedly act as an infrastructural resource, spurring innovation, particularly in process innovation. However, product innovation of software provided as a service can potentially allow Mauritius firms to tap international markets, particularly with the reputation and relationships built upon BPO activities.

The third category, *security*, refers to the innovation activities of Mauritius ICT firms regarding cybersecurity. The Mauritius government launched several strategic imperatives in the last two decades to foster the cybersecurity competency of the economy, e.g., with Mauritius' National Cyber Security Strategy for 2014–2017 and the National Cybercrime Strategy 2017–2020 (Government of Mauritius, 2014;

MTCI, 2018). Consequently, the country ranks persistently high in the Global Cybersecurity Index spearheading the cluster of Africa and 14th worldwide in the year 2020 (ITU, 2021). Regarding innovation, these ICT firms within the category are most likely providing security services that could either be product or process innovation.

The fourth category, *consulting*, comprises consulting services performed by ICT firms in Mauritius. Besides, there are several MNCs such as Accenture, Deloitte, or KPMG active in Mauritius. These ICT firms support other companies to transform their processes using digital technology and are likely an indirect enabler of innovation. This category is also associated with the importance of the financial sector in the country and positively contributes to GDP growth (Joseph & Troester, 2013). However, we consider the consulting sector less R&D intensive and thus rather expect process innovation and only a few product innovations from ICT companies in consulting.

The fifth category is clustering companies in the sphere of *web development*, another pillar of the Mauritius ICT sector. Unfortunately, there is no data or literature on the potential role of Mauritian web development firms in international markets. However, ICT firms in web development are potentially vital service providers for the Mauritius economy's central pillars, such as the tourism sector, financial service sector, and the ICT sector itself. Nevertheless, much innovative activity concerning product or process innovation is not expected to be associated with web development companies.

The sixth category, *mobile connectivity*, refers to companies engaging in mobile ecosystems. Mobile phone penetration in sub-Saharan Africa is generally high, but Mauritius even ranks 11th worldwide according to the Global Competitiveness Index (Schwab, 2019). Besides, the government of Mauritius is fostering and engaging in the development of mobile phone applications, for example, applications such as *Smart Traffic* or *Emergency Alert* for national disaster relief (Ministry of Technology Communication and Innovation, 2018). Furthermore, given the high mobile phone subscription penetration, high ICT literacy, full 4G LTE coverage across the country already in 2017, and resilient subsea optical fiber connectivity to all key markets and continents (MTCI, 2018), the mobile connectivity category captures potential innovators of product and process innovation.

The presented categorization of the ICT sector is based on data scraped from company websites and analyzed with an unsupervised learning approach. As discussed above, we consider this categorization as an indication of innovative activity within the ICT sector. We performed robustness checks by replicating our approach on different data sets: patent descriptions and titles and scientific publications. However, the databases were too small to produce any significant result. Therefore, we considered the categorization by the National Computer Board (see Fig. 7 in the appendix) for a further robustness check. The results for BPO and management services are also reflected in the NCB's sectoral categorization (e.g., *BPO general*). Also, the NCB's categorization of *consultancy services*, *web design & development*, and *mobile & gaming* (see Fig. 7 in the appendix) is related to our identified topics, *consulting*, *web development*, and *mobile connectivity* and their inherent tokens.

While our topic models and the recent strategic imperatives of the Mauritius government indicate sectoral activity in the area of *security*, the NCB categorized only seven firms in the context of *security services*. In summary, several connections could be identified to consider our results as robust. However, one major difference is the occurrence of cloud-related terms in the topic model which is not reflected in NCB's sectoral categorization. We assume that this could be a consequence of a time lag since the categorization of the NCB.

In the second research question, we sought to assess the feasibility of this approach for emerging countries such as Mauritius to address the research question: *Do web mining and topic modeling provide sufficient indicators to understand innovation activities in emerging countries?* For this, we assessed our results against the properties of innovation indicators as described in the *Oslo Manual* (OECD, 2018), using these as evaluation dimensions for our approach, namely *relevance*, *accuracy/validity*, *reliability/precision*, and *timeliness*.

The results of the assessment are depicted in Table 3. We focused on the strengths and weaknesses of our approach in regard to the criteria. First, for the dimension of *relevance*, we conclude that the proposed approach satisfies the requirements of serving the needs of actual and potential users as defined by the *Oslo Manual* (OECD, 2018). Policymakers and scholars can apply the approach at low costs and at high frequency to understand innovation activity. The methodology is lean and modular; hence, it can be easily adapted and, for example, if company descriptions are already available, the web scraping module might not be necessary. The result is an indicator of potential sectoral innovation activity. However, the major weakness is that to provide an accurate insight into the innovativeness of firms in a country, the results need to be triangulated with data and metrics from traditional methods (e.g., surveys) and traditional innovation indicators (e.g., patent statistics). The approach can be established on both ends of the research process in combination with traditional methodologies and data. First, the categorization can be integrated into the design of mixed-method studies that are more specific and targeted by narrowing down the phenomenon of innovative activity on a sectoral level (e.g., a starting point for further surveys and hypothesis-building), particularly when data is scarce such as in emerging countries. Second, the approach can be applied to cross-validate results from qualitative (e.g., survey) or quantitative studies (time series patent analysis).

Regarding the second dimension *accuracy/validity*, the approach has limitations in providing an unbiased representation of innovation phenomena. Topic modeling carries the burden of any explorative approaches based on unsupervised machine learning algorithms that naturally fall short in one gold standard metric of accuracy. While a detailed discussion of evaluation metrics for unsupervised learning techniques goes beyond the scope of this paper, we focused on an evaluation and discussion of the result along with empirical data from academic papers, governmental reports, and other databases (see above). The evaluation of the results along with empirical data from academic papers, governmental reports, and other databases provides an accurate overview of potential innovation activity. However, due to the nature of topic models without a test set, no high accuracy can be guaranteed.

Table 3 Evaluation of the methodological approach based on the OECD (2018) criteria for innovation indicators

Dimension	OECD criteria	Features of the model approach	
		Strengths	*Weaknesses*
Relevance	Serve the need of actual and potential users	Policymakers and scholars can cost-efficiently apply the approach that is documented as a methodological blueprint. Moreover, the approach can be adapted to the framework conditions due to its modular design	The result is an indicator of potential sectoral innovation activity that needs to be triangulated with data and metrics from traditional methods (e.g., surveys) and innovation indicators (e.g., patent statistics)
Accuracy/validity	Provide an unbiased representation of innovation phenomena	The evaluation of the results along with empirical data from academic papers, governmental reports, and other databases provides an accurate overview of potential innovation activity	Due to the nature of topic models without a test set, accuracy cannot be evaluated, particularly for self-reported website data. Furthermore, the results underly the limitations of qualitative coding
Reliability/precision	Results of measurement should be identical when repeated. High signal-to-noise ratio	The approach can be easily replicated, and the topic modeling module produces the same results under the condition of similar input data	The web scraping module of the methodological approach does not necessarily produce identical results due to dynamically changing websites
Timeliness	Available on a sufficiently timely basis to be useful for decision-making	Once the model is set up, it can generate the indicator at a very high frequency and even provide time series data of potential innovation activity	The approach requires input data (e.g., company descriptions for the topic modeling) without a dynamic source at hand (e.g., Crunchbase)

Furthermore, the results underly the limitations of qualitative coding (see section "Limitations and Future Research"). With more data available (e.g., for larger countries than Mauritius), we expect an increase in accuracy. Thus, we conclude that this dimension is partially satisfied with further potential to reproduce similar models and compare their metrics.

For the third evaluation category, *reliability/precision* that requires the results of measurement to be identical when repeated, we conclude that our methodological approach fulfills the criteria of the *Oslo Manual's* (OECD, 2018) metrics for innovation indicators. The presented modular approach is designed to be easily repeated and reproduced at high frequency. The underlying topic modeling algorithm produces the same results if the input data are similar. However, the web scraping module of the methodological approach does not necessarily produce identical results due to dynamically changing websites.

Fourth, we consider the requirement of *timeliness* as satisfyingly fulfilled, since the indicator for innovation activity is available on a sufficiently timely basis and thus useful for decision-making. As described above, scholars can reproduce the approach at a very high frequency, possibly with parallel computing processes and depending on the performance capabilities several times a day. Hence, the presented method has significant advantages in comparison to traditional innovation indicators in terms of *timeliness* (e.g., patent counts that are characterized by time lags).

In summary, the approach offers a method to analyze a sizeable sector-specific number of firm websites from an emerging market and extract their main topics to create an indicator for innovation activity in the sector. The method can be arbitrarily repeated at low costs and large scale. Thus, we conclude that the approach presents a satisfying alternative and a complementary indicator according to the criteria of the *Oslo Manual* for innovation indicators to gain insight into potential innovation activity. Moreover, for the analysis of innovation in emerging countries with little data availability, the method appears to be an excellent tool for understanding potential innovation areas within a sector as a first step in the research process. This can then be essential for building hypotheses and a decision basis for investing in further cost-intensive and time-intensive manual data collection, for example conducting qualitative expert interviews or setting up surveys.

Limitations and Future Research

This study is subject to a few limitations that are partially inherent to novel methodologies. Also, these limitations are a consequence of limited data for emerging countries that can be addressed by future research and data collection initiatives.

First, researchers need input data to construct a sound data basis of company URLs (e.g., business registers) from which text data can be scraped, and this can be a challenging process for emerging countries. We combined company directories provided by two different parastatal organizations organization under the aegis of the Ministry of Technology, Communication, and Innovation (MTCI) of

Mauritius and used search engine queries to construct a list of company URLs based on these lists. While these directories underly a selection bias to correctly classify ICT companies, we consider them as an official ICT business register of Mauritius. However, these granular data are often lacking in other emerging countries. There might be only a general business register for a country available. In this case, researchers need to manually identify the companies to include in the analysis. Usually, this could be done by defining NACE codes that are relevant to the analysis. NACE is the industry standard classification system used in the European Union and also used in many business registers (see Robledo et al., 2012). If the business register is classified with NACE codes, the codes could then be applied to identify the relevant companies. Furthermore, future research could use additional data sources such as the start-up directory Crunchbase and filter for ICT companies located in Mauritius. ICT NACE codes could be matched to the categories of Crunchbase for the identification of relevant companies. Another data source to cross-validate the data are patent descriptions filed by inventors from the country. However, since the patenting activity was relatively low in Mauritius in the previous years, our robustness checks analyzing patent titles and descriptions did not produce significant results. Moreover, to harness the existing data input of company URLs from the directories, future research could scrape websites from the directories at various points in time and conduct a time series analysis addressing different research questions other than what we did in this paper, for example with the aim to investigate the effects of specific policy schemes on innovation activity and sectoral composition.

Second, during the scraping process using the web scraper *ARGUS*, we acknowledged a significant share (24%) of unresponsive websites. *ARGUS* has technical issues with scraping websites written in JavaScript. While only a small fraction of websites is written in JavaScript, we recommend combining *ARGUS* with other web scrapers to receive a more saturated result in future research. The researcher should focus on the phenomenon of unresponsive websites (if replicable) over time and investigate whether, for example, firms cease to operate or change their names more often in the context of emerging countries or within the ICT sector.

Third, qualitative coding underlies the usual limitations of qualitative research. For example, the co-authors' personal experience and knowledge influence observations and conclusions. We addressed this issue by following a rigorous coding process as suggested by Miles et al. (1994). However, future research should consider including independent experts who are not co-authors in the assessment of the coding to decrease the risk of biases.

Fourth, the proposed indicator for innovation activity needs to be combined with other metrics, for example, traditional innovation indicators, to increase the explanatory power of innovation activities. We did this partially by analyzing our results along with governmental reports, academic literature, and empirical data (few data on patents) on the Mauritian ICT sector. We also conducted robustness checks with existing sectoral categorizations. Future research designs could consider triangulating the results of alternative approaches to measuring innovation with our approach.

Besides, the robustness of traditional indicators can be increased by applying our methodology, in particular, if the traditional indicators are exclusively based on qualitative data such as from surveys or expert interviews, being more costly to obtain and underlying several limitations (e.g., interviewee biases, or non-representative responses for surveys).

Overall, our methodological framework serves as a robust solution and with a comprehensive documentation in the "Methodology" section as a blueprint for addressing the challenges associated with data scarcity and quality, usually present in emerging countries. However, the approach is versatile and modular by design; thus, it can be simply transferred and applied in any other socioeconomic contexts in a granular, timely, and cost-efficient manner. Even in countries where data quality is high (e.g., patent data in countries such as Germany), conventional metrics may fall short in providing in-depth explanations. Besides, the proposed methodology can be more efficient in case of time and budget constraints.

Conclusions

The paper's contributions are twofold. First, we provide an indicative overview of the composition of innovative activity in the ICT sector of Mauritius using a lean unsupervised machine learning approach with web-scraped data from company websites. We generated a topic model with seven topics and coded it qualitatively. The results suggest that innovative activity in the Mauritius ICT sector is focused in the following areas: *BPO and management services*, *cloud*, *security*, *consulting*, *web development*, and *mobile connectivity*. Moreover, we discussed these categories considering their relevance for process and product innovation and other empirical data and academic literature. This discussion concludes that Mauritian ICT firms' innovation activities most likely produce process innovation that traditional innovation indicators would not always capture. This argument is in line with the low patenting activity in Mauritius in the last years.

Second, we present a novel methodological approach that can serve as a blueprint for future innovation studies in emerging countries, as the modular design can be simply reproduced and any freely accessible website data can be harnessed. The approach partially fulfills the criteria of innovation indicators as proposed in the *Oslo Manual*. While it has weaknesses regarding accuracy (bias of website data, qualitative coding) and relevance (necessity to triangulate the results with other data or indicators), it offers a time- and cost-efficient method to analyze a sizeable sector-specific number of firms' websites in a reliable and reproducible process. The resulting topic model allows for hypothesis building and thus, provides a starting point for deductive research based on, for example, qualitative analyses such as expert interviews or surveys. The proposed methodology outperforms traditional innovation indicator in terms of *timeliness*. Compared to traditional indicators characterized by time lags (e.g., patent counts), the approach allows reproduction with different datasets at a very high frequency.

AppendixFunding Open Access funding enabled and organized by Projekt DEAL.

Data Availability Statement A full documentation of the data, analysis, and deployed programming code is available in the GitHub repository of the manuscript: https://github.com/MoritzBS/innovation_indicators_through_topic_models.

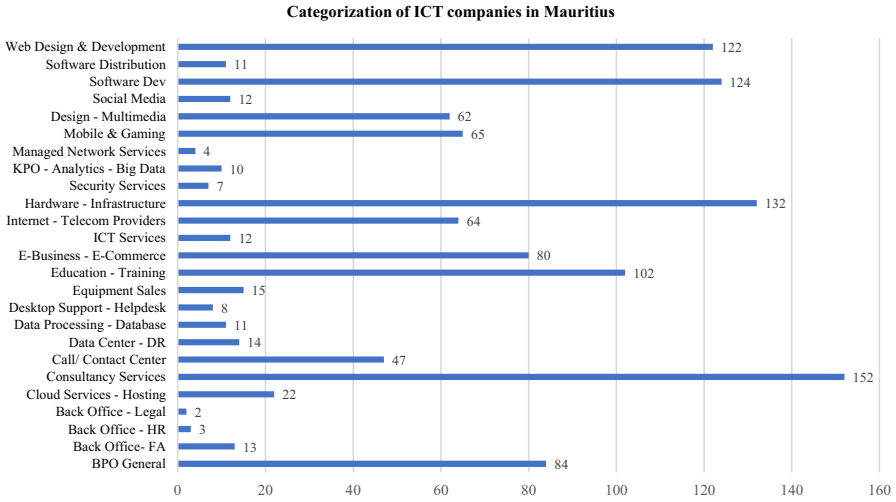

Fig. 7 Number of companies by provided service. Based on the National Computer Board (2018)

Declarations

Ethical Conduct The authors confirm that they comply with the ethical guidelines of the journal.

Conflict of Interest The authors declare no competing interests.

References

Aghion, P., Akcigit, U., & Howitt, P. (2014). What do we learn from schumpeterian growth theory? In *Handbook of Economic Growth*, 2:515–63. Elsevier B.V. https://doi.org/10.1016/B978-0-444-53540-5.00001-X

Aghion, P., & Howitt, P. (1992). A model of growth through creative destruction. *Econometrica, 60*(2), 323–351. https://doi.org/10.3386/w3223

Agrawal, A., Wei, Fu., & Menzies, T. (2018). What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology, 98*(June), 74–88. https://doi.org/10.1016/J.INFSOF.2018.02.005

Antoo, M., Cadersaib, Z., & Gobin, B. (2015). PEST framework for analysing cloud computing adoption by Mauritian SMEs. *Lecture Notes on Software Engineering, 3*(2), 107–112. https://doi.org/10.7763/lnse.2015.v3.175

Arundel, A., & Smith, K. (2014). History of the community innovation survey. *Handbook of Innovation Indicators and Measurement*, 60–87. https://doi.org/10.4337/9780857933652.00011

Becheikh, N., Landry, R., & Amara, N. (2006). Lessons from innovation empirical studies in the manufacturing sector. *Technovation, 26*(5–6), 644–664. https://doi.org/10.1016/j.technovation.2005.06.016T4-Asystematicreviewoftheliteraturefrom1993-2003

Bertschek, I., Erdsiek, D., & Trenz, M. (2017). IT outsourcing—A source of innovation? Microeconometric evidence for Germany. *Managerial and Decision Economics, 38*(7), 941–954. https://doi.org/10.1002/MDE.2835

Biswas, T., & McHardy, J. P. (2012). Secrecy versus patents : Process innovations and the role of uncertainty. *Sheffield Economic Research Paper Series*. Sheffield: Univ. of Sheffield, Dep. of Economics.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022. http://dl.acm.org/ft_gateway.cfm?id=944937&type=pdf

Board, National Computer. (2012). Directory of ICT companies in Mauritius. Edited by NCB. Mauritius. 2012. http://ictexport.govmu.org/English/Documents/DirectoryofICTcompaniesinMauritius.pdf

Board of Investment. (2018). *Industry review 2016 ICT/BPO*. https://www.tralac.org/images/docs/11093/mauritius-industry-review-2016-ict-bpo.pdf

Böhmecke-Schwafert, M., & García-Moreno, E. (2023). 2023: Exploring blockchain-based innovations for economic and sustainable development in the global south: A mixed-method approach based on web mining and topic modeling. *Technological Forecasting and Social Change, 191*, 122446. https://doi.org/10.1016/j.techfore.2023.122446

Casadella, V., & Tahi, S. (2022). National innovation systems in low-income and middle-income countries: Re-evaluation of indicators and lessons for a learning economy in Senegal. *Journal of the Knowledge Economy*. https://doi.org/10.1007/s13132-022-00945-8

Cantner, U., & Vannuccini, S. (2018). Elements of a Schumpeterian catalytic research and innovation policy. *Industrial and Corporate Change, 27*(5), 833–850. https://doi.org/10.1093/icc/dty028

Central Intelligence Agency. (2020). *The world factbook*. CIA.Gov. https://www.cia.gov/library/publications/the-world-factbook/geos/mp.html

Chang, Y. C., Chen, M. H., Lin, Y. P., et al. (2012). Measuring regional innovation and entrepreneurship capabilities. *Journal of the Knowledge Economy, 3*, 90–108. https://doi.org/10.1007/s13132-011-0081-4

Chapman, D. W., & Boothroyd, R. A. (1988). Threats to data quality in developing country settings. *Comparative Education Review, 32*(4), 416–429. https://doi.org/10.1086/446794

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology, 37*(1), 51–89. https://doi.org/10.1002/aris.1440370103M4

Cirera, X., & Muzi, S. (2020). Measuring innovation using firm-level surveys: Evidence from developing countries. *Research Policy, 49*(3), 103912.

Cooper, R. G., & Kleinschmidt, E. J. (1993). New-product success in the chemical industry. *Industrial Marketing Management, 22*(2), 85–99. https://doi.org/10.1016/0019-8501(93)90034-5

Crépon, B., Duguet, E., & Mairesse, J. (1998). *Research, innovation, and productivity: An econometric analysis at the firm level*. No. 6696. https://papers.ssrn.com/abstract=122293

Dewangan, V., & Godse, M. (2014). Towards a holistic enterprise innovation performance measurement system. *Technovation, 34*(9), 536–545. https://doi.org/10.1016/j.technovation.2014.04.002

Diez-Vial, I., & Fernández-Olmos, M. (2017). The effect of science and technology parks on a firm's performance: A dynamic approach over time. *Journal of Evolutionary Economics, 27*(3), 413–434. https://doi.org/10.1007/s00191-016-0481-5

Doll, T. (2018). *LDA topic modeling: An explanation*. https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd

Dumais, S. T., et al. (1994). Latent semantic indexing (LSI) and TREC-2. *Nist Special Publication Sp*, 105.

Dunning, J. H., Changsu, K., & Donghyun, P. (2008). Old wine in new bottles: A comparison of emerging-market TNCs today and developed-country TNCs thirty years ago. In *The Rise of Transnational Corporations from Emerging Markets*, edited by Karl. P. Sauvant. Edward Elgar Publishing. https://ideas.repec.org/h/elg/eechap/13036_8.html

Dutta, S., Lanvin, B., & Wunsch-Vincent, S. (2019). *Global innovation index 2019 - Creating healthy lives - The future of medical innovation*. Ithaca, Fontainebleau, and Geneva.

Dutta, S., Lanvin, B., & Wunsch-Vincent, S. (2020). *Global innovation index 2020 - Who will finance innovation*? Ithaca, Fontainebleau, and Geneva.

Dziallas, M., & Blind, K. (2019). Innovation indicators throughout the innovation process. *Technovation, 80–81*(July), 3–29. https://doi.org/10.1016/j.technovation.2018.05.005T4-Anextensiveliteratureanalysis

Fagerberg, J., Srholec, M., & Verspagen, B. (2010). Innovation and economic development. *Handbook of the Economics of Innovation, 2*(1), 833–872. https://doi.org/10.1016/S0169-7218(10)02004-6

Fagerberg, J., & Verspagen, B. (2007). Innovation, growth and economic development: Have the conditions for catch-up changed? *International Journal of Technological Learning, Innovation and Development, 1*(1), 13–33. https://doi.org/10.1504/IJTLID.2007.015017

Flor, M. L., & Oltra, M. J. (2004). Identification of innovating firms through technological innovation indicators. *Research Policy, 33*(2), 323–336. https://doi.org/10.1016/j.respol.2003.09.009T4-AnapplicationtotheSpanishceramictileindustry

Freeman, C., & Soete, L. (2009). Developing science, technology and innovation indicators. *Research Policy, 38*(4), 583–589. https://doi.org/10.1016/j.respol.2009.01.018T4-Whatwecanlearnfromthepast

Fu, X., Mohnen, P., & Zanello, G. (2018). Innovation and productivity in formal and informal firms in Ghana. *Technological Forecasting and Social Change, 131*(June), 315–325. https://doi.org/10.1016/J.TECHFORE.2017.08.009

GESCI. (2017). *Assessment of knowledge society development in Mauritius*.

GitHub. (2023a). *GitHub repository for the manuscript*. https://github.com/MoritzBS/innovation_indicators_through_topic_models

GitHub. (2023b). *ARGUS web-scraper*. Online Documentation. https://github.com/datawizard1337/ARGUS

Goedhuys, M., Janz, N., & Mohneny, P. (2014). Knowledge-based productivity in 'low-tech' industries: Evidence from firms in developing countries. *Industrial and Corporate Change, 23*(1), 1–23. https://doi.org/10.1093/ICC/DTT006

Government of Mauritius. (2014). *National cyber security strategy*. https://www.itu.int/en/ITU-D/Cybersecurity/Documents/National_Strategies_Repository/Mauritius_2014_NationalCyberSecurityStrategy-2014-EN.pdf

Government of Mauritius. (2020). *Economic and social indicators information and communication technologies (ICT) statistics*. https://statsmauritius.govmu.org/Pages/Statistics/By_Subject/ICT/SB_ICT.aspx

Hagedoorn, J., & Cloodt, M. (2003). Measuring innovative performance. *Research Policy, 32*(8), 1365–1379. https://doi.org/10.1016/S0048-7333(02)00137-3T4-Isthereanadvantageinusingmultipleindicators?

Heikkilä, J., & Lorenz, A. (2017). Need for speed? Exploring the relative importance of patents and utility models among German firms. *Economics of Innovation and New Technology,* 27(1), 80–105. https://papers.ssrn.com/abstract=2956271

Hofmann, T. (1999). Probabilistic latent semantic analysis. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*.

IPlytics. (2022). IP Intelligence Tool. *All registered patents and published research papers from Mauritius*. https://www.iplytics.com/

ITU. (2021). *Global cybersecurity index 2020*. https://www.itu.int/en/myitu/Publications/2021/06/28/13/22/Global-Cybersecurity-Index-2020

Joseph, A., & Troester, B. (2013). C*an the Mauritian miracle continue? - The role of financial and ICT services as prospective growth drivers* (No. April). http://finance-and-trade.htw-berlin.de/fileadmin/working_paper_series/wp_01_2013_Joseph_Troester_Can_the_Mauritian_Miracle_continue.pdf

Kapadia, S. (2019). *Topic modeling in python: Latent Dirichlet Allocation (LDA)*. https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0

Kinne, J., & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics*, 18–033.

Kinne, J., & Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. *Edited by Wonjoon Kim. PLOS ONE, 16*(4), e0249071. https://doi.org/10.1371/journal.pone.0249071

Kleinknecht, A. (1993). Why do we need new innovation output indicators? An introduction. In *New Concepts in Innovation Output Measurement*, 1–9. Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-22892-8_1

Kleinknecht, A., van Montfort, K., & Brouwer, E. (2002). The non-trivial choice between innovation indicators. *Economics of Innovation and New Technology, 11*(2), 109–121. https://doi.org/10.1080/10438590210899

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem. *Psychological Review, 104*(2), 211.

Lim, A. H., & Bart De Meester. (2016). *WTO domestic regulation and services trade : Putting Principles into Practice*.

Linton, M., Teo, E. G. S., Bommes, E., Chen, C. Y., & Härdle, W. K. (2017). Dynamic topic modelling for cryptocurrency community forums. In *Applied Quantitative Finance*, 355–72. Springer.

Marins, L. (2008). The challenge of measuring innovation in emerging economies' Firms. *MERIT Working Papers*, 044. https://ideas.repec.org/p/unm/unumer/2008044.html

Miles, M. B., Huberman, A. M., & Saldaña, J. (1994). *Qualitative data analysis a methods sourcebook edition*. SAGE PUBN.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). *Optimizing semantic coherence in topic models*. https://www.aclweb.org/anthology/D11-1024.pdf

Ministry of Technology Communication and Innovation. (2018). Digital government transformation strategy 2018–2022. http://cib.govmu.org/English/Documents/DGTS/DigitalGovernmentTransformationStrategy2018-2022.pdf

MTCI. (2018). Digital Mauritius 2030 strategic plan. http://mtci.govmu.org/English/Documents/2018/LaunchingDigitalTransformationStrategy191218/DM203017December2018at12.30hrs.pdf

National Computer Board. (2018). *National computer board - ICT industry in Mauritius.* http://ictexport.govmu.org/English/For_Buyers/ICTIndustryinMauritius/Pages/default.aspx

OECD, (2006). *African economic outlook: Mauritius*. https://doi.org/10.1787/440280862401

OECD. (2018). Oslo manual 2018: Guidelines for collecting, reporting and using data on innovation, 4th edition. The measurement of scientific, technological and innovation activities, *OECD Publishing, Paris/Eurostat, Luxembourg, 2018.* https://doi.org/10.1787/9789264065659-es

OECD. (2017). *"ICT and innovation", in OECD Science, Technology and Industry Scoreboard 2017: The digital transformation*. OECD Publishing.

Oolun, K., Ramgolam, S., & Dorasami, V. (2012). The making of a digital nation: Toward i-Mauritius. *The Global Information Technology Report 2012: Living in a Hyperconnected World*, 161–68. http://reports.weforum.org/global-information-technology-2012/

Panichella, A. (2021). A systematic comparison of search-based approaches for LDA hyperparameter tuning. *Information and Software Technology, 130*(February), 106411. https://doi.org/10.1016/J.INFSOF.2020.106411

Panichella, A., & Poshyvanyk, D. (2013). How to effectively use topic models for software engineering tasks? An approach based on genetic algorithms. In *2013 35th International Conference on Software Engineering (ICSE)*. https://www.academia.edu/16506119/How_to_effectively_use_topic_models_for_software_engineering_tasks_An_approach_based_on_Genetic_Algorithms

POWC. (2022). List of BPO companies. Public Officers' Welfare Council (POWC). https://powc.govmu.org/Documents/Companies/ListOfBPOCompaniespdf.pdf

Rajman, M., & Besançon, R. (1998). Text mining: Natural language techniques and text mining applications. In *Data Mining and Reverse Engineering*, 50–64. Springer US. https://doi.org/10.1007/978-0-387-35300-5_3

Robledo, J. C., Mas, M., & Perez, J. (2012). ICT sector definition. Transition from NACE Rev. 1.1 to NACE Rev. 2: A methodological note. Publications Office. https://doi.org/10.2791/40232

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. Edited by Xueqi Cheng, Hang Li, Evgeniy Gabrilovich, and Jie Tang. WSDM'15. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. Jan. 31-Feb. 6, 2015, Shanghai, China. New York, NY: ACM Association for Computing Machinery. https://doi.org/10.1145/2684822.2685324

Rodriguez-Pose, A., & di Cataldo, M. (2015). Quality of government and innovative performance in the regions of Europe. *Journal of Economic Geography, 15*(4), 673–706. https://doi.org/10.1093/JEG/LBU023

Rodríguez-Pose, A., & Zhang, M. (2020). The cost of weak institutions for innovation in China. *Technological Forecasting and Social Change, 153*(April), 119937. https://doi.org/10.1016/J.TECHFORE.2020.119937

Schofield, A., Magnusson, M., Thompson, L., & Mimno, D. (2017). Pre-processing for latent Dirichlet allocation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 432–436. Valencia.

Schumpeter, J. A. (1934). *The theory of economic development. An inquiry into profits, capital, credit, interest, and the business cycle*. *Half-Title: Harvard Economic Studies*. Harvard University Press.

Schwab, K. (2019). *The Global Competitiveness Report 2019*.

Seechurn, R. K., Ramtohul, L., Googoolye, K., Vaghjee-Rajiah, T., & Neeliah, H. (2013). A tale of five sectors in Mauritius: Agriculture, textile/EPZ, tourism, financial services and ICT/BPO. *An International HRD Conference, Mauritius 2013: Excellence in HRD for Sustainable Growth.*

Sievert, C., & Shirley, K. (2015). PyLDAvis. *Python Library for Interactive Topic Model Visualization*. https://github.com/bmabey/pyLDAvis

Soyjaudah, K. M. S., Oolun, M. K., Jahmeerbacus, I., & Govinda, S. (2002). ICT development in Mauritius. In *IEEE AFRICON. 6th Africon Conference in Africa,* 53–58. IEEE. https://doi.org/10.1109/AFRCON.2002.1146805

Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion, 36*(July), 10–25. https://doi.org/10.1016/j.inffus.2016.10.004

The World Bank. (2020). *Country profile Mauritius*. https://data.worldbank.org/country/mauritius

Tokunaga, T., Ortega, A., Masada, T., Kiyasu, S., & Miyahara, S. (Eds.). (2008). Comparing LDA with PLSI as a dimensionality reduction method in document clustering. In *Berlin, Heidelberg: Springer Berlin Heidelberg*. https://link.springer.com/chapter/10.1007/978-3-540-78159-2_2

Turner, R. (2018). *Travel and tourism: Economic impact 2018 - Mauritius*. https://www.wttc.org/-/media/files/reports/economic-impact-research/countries-2018/mauritius2018.pdf

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. *ACM. Doi, 10*(1145/1553374), 1553515.

WIPO. (2021). *Statistical country profiles - Mauritius*. https://www.wipo.int/ipstats/en/statistics/country_profile/profile.jsp?code=MU

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*.

Yogish, D., Manjunath, T. N., & Hegadi, R. S. (2019). Review on natural language processing trends and techniques using NLTK. *Communications in Computer and Information Science, 1037*, 589–606. https://doi.org/10.1007/978-981-13-9187-3_53/COVER

Zafar, A. (2006). Mauritius: An economic success story. http://siteresources.worldbank.org/AFRICAEXT/Resources/258643-1271798012256/YAC_chpt_5.pdf

Zawislak, P. A., & Marins, L. M. (2008). Strenghtening innovation in developing countries. *Journal of Technology Management & Innovation, 2*(4), 11.