Green Entrepreneurship: Identification, Emergence, and Legitimacy

A Thesis Submitted in Partial Fulfilment of the Requirements
for the Degree of Doctor of Philosophy in

Sustainable Development and Climate change

Doctoral Programme of National Interest

PhD SDC
SUSTAINABLE DEVELOPMENT
AND CLIMATE CHANGE

In the Curriculum
SOCIO-ECONOMIC RISK AND IMPACT

Le Masle Baptiste

Supervisor: Alessandra Colombelli

Co-supervisor: Stefano Bianchini

February, 2026

ABSTRACT

This thesis contributes to the field of green entrepreneurship, which has experienced substantial development in recent years. The field has attracted increasing interest from policymakers, society, and researchers as a means of fostering sustainable growth. Our work focuses on green startups, a widely used indicator of green entrepreneurship. However, research on green startups is hindered by significant methodological challenges, notably the limited availability of data and the difficulty of defining what constitutes a green startup. To address these challenges, we leverage recent advances in natural language processing (NLP) to introduce new methodological tools for identifying and studying green startups. We apply these methods to analyze the emergence of Italian startups and the ways in which they establish legitimacy.

The first chapter introduces a novel NLP-based methodology to identify green startups and compares it with two existing approaches used to capture the broader category of sustainable startups. Given that the operationalization of the green startup concept remains a subject of debate, we begin by developing a classification framework based on the Sustainable Development Goals (SDGs), a widely used and well-established reference in sustainability research. The SDGs comprise 17 goals defined by the United Nations to promote social, economic, and environmental sustainability. Focusing specifically on green startups, we apply NLP techniques to extract environment-related mentions and group them into 14 green thematic topics. Then, we apply the three approaches, each leveraging Dictionary, Latent Dirichlet Allocation (LDA) and BERTopic, to 10,939 websites of Italian innovative startups. We find that the three methods identify overlapping but distinct sets of green start-ups, primarily because each method captures different areas of environmental engagement. Finally, we relate the number of identified start-ups with regional SDG progress and spending from the National Recovery and Resilience Plan, and show that entrepreneurial activity and public policy align on the issues related to sustainable energy, sustainable practices, and air quality, but that public policy contributes more to water quality as well as disaster resilience.

In the second chapter, we focus on the emergence of the green startups identified in the first chapter. In particular, we explore the interplay between knowledge availability and green innovation in Italian provinces by extending the Knowledge Spillover Theory of Entrepreneurship (KSTE). We explore how green demand, the stock of knowledge, and its composition lead to the creation of green startups. Aligned with established literature, we show that knowledge stocks have a positive impact on green startup creation. Green demand also has a positive impact and amplifies the role of the local knowledge stocks. Those results suggest that green demand makes it easier for entrepreneurs to leverage knowledge spillovers and create green startups. Green demand softens the Knowledge Filter by increasing short-term expected returns on the innovation investment, ultimately allowing knowledge spillovers to be converted into startups more swiftly. Finally, green knowledge does not have a stronger impact than non-green knowledge, suggesting that green startups recombine green and non-green knowledge to innovate.

In the third chapter, we study how green and non-green entrepreneurs establish legitimacy and develop their networks by sharing narratives on social media. We focus on a subsample of 1,703 founders for whom we were able to scrape data and posts from their LinkedIn profiles. To study quantitatively entrepreneurial narratives, we introduce a new methodology leveraging Large Language Models to identify cognitive, normative, and pragmatic legitimacy claims in their posts. We find that those three types of legitimacy claims have an impact on online legitimacy, as measured by the number of reactions to their posts. Interestingly, green startups do not benefit from normative legitimacy claims, suggesting that their green status is already sufficient in establishing legitimacy. Furthermore, we highlight that cognitive legitimacy claims have a direct positive effect on network size for non-green firms, while all types of legitimacy claims positively affect network size indirectly by attracting attention. These results highlight the importance for green entrepreneurs of incorporating legitimacy claims into their communications.

ACKNOWLEDGEMENTS

I would like to express my profound and sincere gratitude to my supervisor Alessandra Colombelli for the wonderfull experience this doctoral programm has been for me. Thanks for your trust in my work and the autonomy you confered me. Thanks for your supervision and your guidance on my research.

Thank you very much to my co-supervisor Stefano Bianchini. I truly loved working with you on my research. Thanks for your scientific advices and your guidance, and also for your running tips.

I would like to express my gratitude for Chiara Ravetti and Matteo Tubiana. Thank you so much for your help and your positive and beneficial feedback.

Thanks to all the researchers who provided me feedback and guidance during my thesis, in particular Barabara Pernucci and Sando Montresor. Thanks for your interest. I hope one day we will have the opportunity to work together.

My thanks to Gianluca Orsatti e Claudia Ghisetti. I am honoured that they acepted to be members of the jury.

I will not start to thank my friends and family with whom I spent so many happy moments in my PhD journey, otherwise I would have to thank Suzanne, Leo, Ezo, Jeanne, Charlotte, Marco, Papa, Maman, Gaël, Vincent, William, Valentin, Anaelle, Nicolas, Pierre, Laura, Paul, Antoine, Aurore, toute la Kem's Production, Marco, Karen, Mamie, toute la famille Allenbach, Annick, Pablo, Adèle, Audrey, Zuling, Victor, Jade, Vitor, and more. But you deserve it, especially you, who are reading the thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Global introduction

1 GLOBAL INTRODUCTION

1.1 THE SUSTAINABILITY CRISES

January 2026. Greenhouse gas emissions from human activities will lead to an increase of 3 °C by the end of the century if left unchanged, leading to widespread and irreversible consequences such as extreme heatwaves, rising sea levels that threaten coastal cities, severe droughts and floods, large-scale ecosystem collapse, food and water insecurity, increased health risks, mass displacement of populations, and heightened social and economic instability worldwide. Similarly, inequalities are growing, populations are still discriminated against, poverty remains widespread, access to education and healthcare remains limited in many regions, placing severe strain on societies and undermining global cohesion and human well-being (Sachs, Lafortune, Fuller, & Iablonovski, 2025). The population is aging at an increasing pace, which might lead to economic slowdowns, pension crisis, problems of care end even endanger democracy (Grinin et al., 2023). International cooperation appears increasingly difficult with the increase of conflicts, misinformation, political instability, and polarization (de Groot et al., 2022; Iyengar et al., 2019; Jerit & Zhao, 2020; Levin et al., 2021; Rustad, 2025; Teruel, 2023).

The need for sustainability, as the practice of meeting the needs of the present without compromising the ability of future generations to meet their own needs, has never been greater. Admittedly, the task is enormous. The global nature of sustainability challenges requires global, international, and long-term commitment. Take, for example, global warming, one of the most tragic tragedies of the commons (Hardin, 1968). Nations, industries, and individuals benefit economically from activities that emit greenhouse gases, such as burning fossil fuels for energy, transportation, and industrial production. Each actor gains immediate advantages—economic growth, convenience, or profit—while the negative consequences of emissions are distributed globally and delayed over time. Because the costs of environmental degradation are shared by all, there is little incentive for any single actor to significantly reduce emissions on their own, especially if others continue to pollute. Further, widespread misinformation and intensive lobbying by powerful and wealthy actors undermine public understanding and obstruct meaningful policy action (Brulle et al., 2018) It calls for efficient institutions and effective governance, which, to date, have not demonstrated sufficient commitment or coordinated action to adequately address the issue.

Further complicating the matter, sustainability has a wide range of interrelated and sometimes conflicting dimensions. It is often considered to be built on three pillars: environmental, social, and economic sustainability. These pillars can come into conflict because progress in one area may create trade-offs in another, such as economic growth putting pressure on natural resources or social development requiring investments that challenge short-term economic priorities. A good example is the Yellow Vest contestation movement in France (Cepparulo & Giuriato, 2024). In 2018, the French government increased the carbon tax, designed to discourage the use of fossil fuels. However, the policy weighed especially heavily on low- and middle-income households in rural areas, who could not easily afford new means of transportation and often lacked access to reliable public transit. This imbalance sparked protests across the country, which at times turned violent and paralysed the country. Importantly, most protesters were not opposed to climate action itself; rather, they expressed concerns about social justice, inequality, and the perceived unfairness of how environmental policies were designed and implemented.

And the need for sustainability is pressing. The world's global warming is close to 1.5°C, putting major tipping points at risk of being crossed. Tipping points are critical thresholds that,

once exceeded, can trigger large, cascading, and often irreversible shifts in Earth systems, leading to profound environmental, social, and economic consequences (Lenton et al., 2023). In the cryosphere (frozen parts of the Earth's systems), melting of ice sheets would lead to dramatic sea-level rise, damage to ecosystems, the release of CO2, and altered weather patterns. In the biosphere, the dieback of forests would lead to an increase in CO2 emissions, impacting rainfall and biodiversity. The degradation of drylands and savannas would lead to biodiversity loss, desertification, and groundwater depletion. An increase in temperature in the oceans and the atmosphere would lead to modifications in water and air circulation, impacting ecosystems and weather patterns. Those tipping points would have irreversible consequences for society: for example, “the collapse of the Atlantic Ocean’s great overturning circulation combined with global warming could cause half of the global area for growing wheat and maize to be lost” (Lenton et al., 2023).

This brief state-of-the-art of sustainability should not be considered as an alarmist claim, but rather as a call to action. To limit the impact that climate change will have on Earth, our society, our family, our children, and their children, all of us have to act now at this level. Fortunately, an increasing number of such initiatives and individual contributions are already emerging across society. This thesis is a little contribution to the environmental sustainability scholarly literature.

1.2 TOWARD SUSTAINABILITY

Awareness and contributions to sustainability are increasing. The People’s Climate Vote (United Nations Development Programme, 2024) shows that 80% of the global population is calling for their country to strengthen its commitments to climate action and the majority of people are more worried about climate change than a year ago. The most worried countries are located in sub-Saharan Africa and in Latin America, showing that care about climate change is not necessarily correlated with economic development. However, although concern is rising and intentions to act are increasing —both to reduce contributions to and to adapt to climate change— these intentions often do not translate into actual behavior, creating an “intention-behavior gap” (Nguyen et al., 2019; Osberghaus et al., 2025; Shan et al., 2025)

Similarly, research on climate change is growing (Fu & Waltman, 2021) – from less than 2,000 to 15,000 publications between 2001 and 2018. Interestingly, the focus of this research has shifted from exploring and improving the understanding of climate systems to the development of climate technologies and policies to deal with climate change. This research was mostly driven by developed countries, with the USA producing approximately 73% of all climate change publications.

This reflects an increasing concern of governments. According to (Cepparulo & Giuriato, 2024), “*84 countries explicitly recognize a substantive right to a healthy environment in their constitutions*” (p.XXX), and such rights have a positive effect on environmental performance and a negative impact on greenhouse gas emissions. To keep the global warming at 1.5°C compared to pre-industrial time, the global emissions should be halved by 2030 , and reach net-zero by 2050. Still, global greenhouse gas emissions are still rising (Jones et al., 2023). Due to lock-ins and path dependency, governments often commit only to insufficient

incremental changes that fall short of the scale of sustainability challenges. As a result, the commitments of many countries remain largely symbolic rather than transformative.Although developed countries have begun reducing their territorial CO_2 emissions, European and North American countries remain among the highest per-capita emitters. Moreover, reductions in these regions are partly offset by rising emissions in developing economies, particularly China. In addition, developed countries often “offshore” emissions by importing carbon-intensive goods produced elsewhere. When emissions are adjusted for trade (i.e., measured on a consumption basis), the picture becomes clearer. In 2023, consumption-based emissions per capita were approximately 15.81 tonnes in the United States, 7.94 tonnes in European countries, and 7.63 tonnes in China. By contrast, low-income countries emitted only about 1.4 tonnes per capita (Ritchie & Ritchie, 2019). As a matter of comparison, the emissions to align with a 1.5°C level are approximately 2 tones per capita per year.

To solve global challenges, global cooperation is needed. Efforts to coordinate at the global level began in the late 20th century. In 1979, the First World Climate Conference marked the first major international recognition of climate change as a global issue. This momentum led to the creation of the Intergovernmental Panel on Climate Change (IPCC) in 1988, tasked with assessing scientific knowledge and informing policymakers. In 1992, the United Nations Framework Convention on Climate Change (UNFCCC) was adopted at the Rio Earth Summit. It recognized common but differentiated responsibilities between industrialized and less developed countries. Building on this, the Kyoto Protocol was adopted in 1997, introducing legally binding emission reduction targets for industrialized countries, though its impact was limited by incomplete participation and enforcement. In response to these shortcomings, a new approach emerged with the Paris Agreement in 2015, which aimed to involve all countries through nationally determined contributions (NDCs) and a shared long-term goal of limiting global warming to well below 2°C (preferably 1.5°C). The Paris Agreement emphasized transparency, periodic review, and progressive ambition rather than binding targets.

The 2030 Agenda for Sustainable Development, signed by all United Nations Member States on 25 September 2015, introduced the 17 Sustainable Development Goals (SDG), including social, environmental, economic, and governmental goals. Each of those goals is divided into targets and corresponding indicators to ease their operationalization and measurement. It offers a rich framework, which is widely used by researchers (Yamaguchi et al., 2023). Although there is a global commitment to the SDGs, with 190 out of 193 countries having presented national action plans for sustainable development, the SDGs are far off-track. None of the 17 goals and only 17% of the targets are currently on course to be achieved by 2030 (Sachs, Lafortune, Fuller, Iablonovski, et al., 2025).

Still, human society is increasingly prepared at starting the sustainable transition. Never before has it been so easy to access information about climate change. Thanks to scientific research institutions, international organizations, civil society initiatives, and the growing movement toward open and transparent science, reliable and evidence-based information is widely accessible. Major actors such as the United Nations, the IPCC, NASA, the World Meteorological Organization, and organizations like Carbon Brief or Our World in Data make climate data, analyses, tools, and projections freely available. This unprecedented access allows citizens, journalists, researchers, and policymakers alike to better understand climate change and base decisions on robust, shared knowledge. At the same time, this informational abundance comes with a downside, that is the digital environment also amplifies misinformation and deliberate disinformation. The real challenge, therefore, is no longer only producing and sharing knowledge, but strengthening the capacity to distinguish credible, evidence-based sources from unreliable ones.

For example, the French Government, with the website “Nos gestes climats”, offers an environmental footprint estimator, which shows which part of our consumption is the most emitting and suggests where we should allocate our efforts to have an efficient impact on our footprint. Diminishing meat consumption and diminishing flight and car usage can save tons of CO2 emissions. Similarly, the Shift Project, a think tank conducting studies on decarbonization to produce pragmatic propositions, published in 2020 the “Plan de Transition de l’Economie Française” (PTEF). The PTEF is a comprehensive roadmap outlining how France can transition to a low-carbon, resilient economy compatible with climate targets. It proposes sector-by-sector transformations—covering energy, transport, housing, industry, agriculture, and public services—based on physical constraints such as energy availability and carbon budgets. The PTEF emphasizes reducing energy demand, prioritizing sobriety, efficiency, and decarbonization, while maintaining social cohesion and economic stability. It also stresses the crucial role of innovation in increasing our carbon efficiency.

Growing global awareness, the multiplication of grassroots movements, scientific mobilization, technological innovation, and the sustained efforts of governments and international organizations to coordinate action all leave room for hope. The Global Tipping Points report (Lenton et al., 2025) stresses that our time is running out, while also highlighting that positive tipping points—mechanisms that can rapidly accelerate societal and technological change—could be harnessed to strengthen efforts to combat climate change. Enabling finance in the global south, and targeting capital flow to the high returns sustainable investments will further accelerate the process. Progress in renewable energy, electric vehicles, food consumption, and production could have a cascading impact.

1.3 NEW APPROACHES TO STUDY SUSTAINABILITY

This underlines the importance of leveraging the exponential growth of science (Bornmann & Mutz, 2014). Science first evolved around empirical observation and theoretical reasoning, forming the foundations for understanding natural phenomena through simplified laws and principles. As complexity grew, a third paradigm emerged based on computational models and simulations to study problems that could not be solved analytically. More recently, advances in artificial intelligence and data-intensive computing have given rise to a data-driven paradigm in which intelligent machines assist scientists in exploiting vast datasets to generate new knowledge (Bail & Bail, 2024; Tolle et al., 2011; Xu et al., 2021).

In the last decades, progress in Artificial Intelligence, and in particular Natural Language Processing (NLP) have opened new opportunities in economics (Gentzkow et al., 2019) and social sciences more in general (Bail & Bail, 2024). In very recent years, the publication of *“Attention Is All You Need”* (Vaswani et al., 2017), laid the ground for building language models capable of learning from vast quantities of text to understand textual data. Between 2018 and 2020, models such as GPT, BERT, and GPT-2 established the paradigm of large pretrained language models. By representing text as embeddings—high-dimensional vectors that capture semantic meaning—these models provided economists with tools to systematically analyze massive textual datasets that had previously been difficult to exploit.

In 2020, the release of GPT-3 demonstrated that Large Language Models could achieve broad, general-purpose language competence. Since then, new LLMs have been released frequently (i.e., ChatGPT in November 2022) and their performance has been improving at an accelerating rate. Important parts of this thesis are methodological contributions, leveraging those recent advancements.

However, new knowledge alone is not enough to reach sustainability; it must inform the design of innovative technologies, policies, and business models that reduce environmental harm while supporting social and economic well-being. The scientific literature has long recognized the importance of entrepreneurship in the innovation process (Acs & Audretsch, 1987, 1988). By extending the literature on entrepreneurship, and in particular green entrepreneurship, this thesis makes a modest contribution to support the transition toward environmental sustainability.

1.4 ENTREPRENEURSHIP AND GREEN ENTREPRENEURSHIP

The definition of entrepreneurship is well-known to be complex and a subject of debate. Historically, different schools of thought proposed a range of definitions. It has been defined as : (1) inborn trait, ability, and needs; (2) the process of resources and knowledge recombination to innovate; (3) the management of a new business for profit; (4) leadership mobilizing people and resources for change; (5) innovation within existing organizations (Cunningham & Lischeron, 1991). Recent research defined entrepreneurship broadly as “the act of generating and developing an idea for validation” (Prince et al., 2021). In this thesis, and in line with an important strand of the literature, we operationalize entrepreneurship with the creation of innovative young ventures, i.e. startups. With this operationalization, entrepreneurs generate a business idea by identifying a market opportunity, developing it by creating a company, and recombining knowledge to innovate, and then try to validate it by confronting it with the market.

Notwithstanding the lack of a consensual definition, academic research has produced meaningful results underlining the importance of entrepreneurship. Entrepreneurship is contributing at the macro level to economic growth, competitiveness, and job creation, and at the micro-level to firm performance (Wennekers & Thurik, 1999). Many studies explored the entrepreneurial process. It is a dynamic process, in which an entrepreneur, interacting with their environment, recognizes an opportunity, exploits it, and creates value, iterating each of those steps as the process goes on (van der Veen & Wakkee, 2004). The role and the interaction of the entrepreneur and its environment have been extensively studied at each of those steps.

An increasingly important strand of this research is the Knowledge Spillover Theory of Entrepreneurship (KSTE)(Ghio et al., 2015). The theory begins with the premise that incumbent firms generate knowledge but can commercialize only a portion of it; the remainder leaves the firm as knowledge spillovers. Recognizing a business opportunity, entrepreneurs attempt to grab those spillovers (Acs et al., 2009). To develop these spillovers into innovations, entrepreneurs face multiple challenges, including regulatory barriers, institutional constraints, and the inherent complexity of knowledge spillovers (Acs et al., 2013). They have to address these difficulties by leveraging their resources, such as absorptive capacity, networks, and capital, in order to commercialize their innovations. Ultimately, this process contributes to growth (Acs et al., 2012).

However, scholars have questioned the ability of entrepreneurship to face sustainability challenges. It led to the development of the branch of sustainable entrepreneurship. Although subject of debate, sustainable entrepreneurship can be defined as “a branch of

entrepreneurship that is concerned with the preservation of nature, life support, and community in the pursuit of perceived opportunities to bring into existence future products, processes, and services for gain, where gain is broadly understood to include both economic and noneconomic gains to individuals, the economy, and society" (Patzelt & Shepherd, 2011). This branch can be further subdivided into social entrepreneurship, which addresses social challenges, and green entrepreneurship, which focuses on environmental sustainability. Both subfields are difficult to define formally and operationalize due to their complexity, multidimensionality, and the lack of data inherent to new ventures. Focusing on green entrepreneurship, Chapter 1 introduces a green classification framework built from the SDGs, introduces a new automated method to classify green startups based on their website, and compares it to two established methods. This methodological contribution improves the understanding of the available tools to identify green startups in contexts of low data availability and computational resources.

Answering to the pressing sustainability crises, an increasing number of studies explore the differences between sustainable and mainstream entrepreneurship (Ike, 2025). For instance, specific education, institutional support, networks, sustainability awareness, and engagement shape the entrepreneurs' ability to recognize opportunities that address sustainability challenges. The sustainable entrepreneurial process can lead to long-term competitive advantages, driven by higher resource efficiency, better brand reputation, and access to new markets (Ghisetti et al., 2014). Ultimately, these new entrants are more likely to introduce radical, sustainability-oriented innovations.

Researchers explored how the KSTE applies to green startups. First, studies showed that a bigger knowledge base leads to more emergence of green startups (Giudici et al., 2019). Then, some scholars explored the role of the composition of the knowledge base, showing that both green and non-green knowledge contributes to the emergence of green startups (Cojoianu et al., 2020; Colombelli & Quatraro, 2019a). Chapter 2 contributes to this discussion by extending the KSTE framework to incorporate a green demand component, proxied by environmentally oriented behaviors that signal underlying demand for green products and services.

While the Knowledge Spillover Theory of Entrepreneurship and the sustainable entrepreneurship literature explain how opportunities emerge and how entrepreneurs mobilize knowledge and resources to exploit them, they do not fully account for a critical challenge faced by new ventures: gaining legitimacy. As young ventures, startups suffer from limited resources, low market recognition, weak networks, and other disadvantages collectively referred to as the "liability of newness" (Fisher, 2020; Stinchcombe, 1965). Startups, and green startups in particular, operate under conditions of high uncertainty, limited track records, and novel value propositions. As such, their ability to face the liability of newness and to reach a high level of growth depends not only on their ability to recombine knowledge and address market demand, but also on how external stakeholders perceive and evaluate them (Gordo Molina et al., 2022). Investors, customers, regulators, and partners must consider these ventures as appropriate, credible, and desirable actors before engaging with them (Suchman, 1995). This challenge is especially salient for green entrepreneurs, who often face multiple, complex, and challenging audiences (Castelló et al., 2016).

To build legitimacy, firms must engage in legitimacy work, that is, deliberate actions aimed at shaping stakeholders' perceptions. Researchers have examined multiple facets of legitimacy work, including building networks with legitimizing actors, gaining media attention, and crafting narratives (Riandita et al., 2022; Suddaby et al., 2017). Indeed, entrepreneurs are increasingly leveraging Social Networking Sites (SNS) to share narratives, interact, and build legitimacy with their networks (Zhao et al., 2023). Answering recent calls for research in this area (Audretsch & Lehmann, 2023), Chapter 3 introduces a new method that leverages LLM to study entrepreneurial narratives quantitatively and explores how legitimacy claims of entrepreneurs on SNS contribute to attracting attention and developing their networks.

1.5 OUTLINE OF THE THESIS

As discussed above, green entrepreneurship is increasingly seen as a lever in tackling the climate crisis, which makes it important to better understand what drives its emergence and success. However, the limited availability of data on new ventures makes this task particularly challenging. This thesis focuses on developing novel methodological tools to analyse the environmental orientation of entrepreneurial ventures.

1.5.1 Chapter 2

This methodological chapter aims to identify the most effective approach for detecting green startups under constraints of limited data and computational resources. Several approaches leveraging NLP methods applied to startups' websites have been proposed in the literature to identify sustainable startups. For example, (Mansouri & Momtaz, 2022) apply a Machine Learning (ML) dictionary method leveraging embeddings to construct three dictionaries — environmental, social, and governance—which they use to assign Environmental, Social, and Governance (ESG) scores to startups based on their website content. (Tiba et al., 2021) apply topic modeling using Latent Dirichlet Allocation (LDA) to extract topics from startups' websites and then manually classify which of these topics are sustainability-oriented. Although these methods have enabled valuable research, they are not fully automated, have not been systematically compared, and do not account for more recent methodological developments. Further, they have not been applied to the more selective set of green startups.

We fill those gaps in Chapter 1. First, we construct a green classification based on the SDGs framework. More specifically, we apply NLP embedding and clustering methods to the SDG targets with an environmental dimension, as identified in the Global Environment Outlook 6 (UN, 2019). This approach enables a more granular classification than methods that operate only at the goal level and allows us to focus exclusively on the green dimensions of the SDGs. The final classification is composed of 14 green topics.

We apply this framework to identify green startups within a sample of Italian firms by analyzing the content of their websites. To do so, we compare three different text analysis approaches: a keyword-based dictionary method, a traditional topic modeling technique (LDA), and a more recent method (BERTopic). The dictionary approach flags startups as green based on the presence of words closely related to 14 predefined green topics. The LDA and BERTopic methods instead detect themes directly from website texts and classify startups as green when these themes closely match the same 14 topics. While the three methods identify overlapping but not identical groups of green startups, they differ in performance. BERTopic achieves high precision and captures startups across all green areas. LDA reaches slightly higher precision but mainly identifies startups in the most common green categories. The dictionary method has lower precision but still detects startups across the full spectrum of green topics. Finally, we link the number of identified startups to regional SDG progress and spending under the National Recovery and Resilience Plan, showing that entrepreneurial

activity and public policy align on sustainable energy, sustainable development, and air quality, while public policy places greater emphasis on water quality and disaster resilience.

1.5.2 Chapter 3

In this chapter, we study the emergence of green startups through the lens of the Knowledge Spillover Theory of Entrepreneurship (KSTE). Startup success is strongly influenced by local knowledge factors, including the innovative ecosystems in which firms operate. This dependence may be particularly pronounced for green startups, which often face greater risks and uncertainty due to the complexity and novelty of sustainable innovations. Consequently, robust local ecosystems that provide the necessary knowledge, infrastructure, and market conditions are particularly important for green startups (York et al., 2018). Prior studies using KSTE show that entrepreneurs leverage knowledge spillovers to create innovative firms that drive economic growth (Colombelli et al., 2016). Subsequent research has examined the role of local knowledge stocks and their composition, highlighting the distinct effects of clean and dirty knowledge stocks on the creation of green startups (Colombelli & Quatraro, 2019b).

So far, the KSTE has primarily focused on supply-side factors, such as the availability of knowledge and the composition of the ecosystems, and comparatively neglected demand-side considerations. We posit that green demand plays a crucial role in driving green entrepreneurship. Because the innovations, products, and services generated by green entrepreneurial activities contribute to the provision of public goods, their consumption may be discouraged by the classic disincentives associated with public goods. Despite these theoretical constraints, we observe a growing demand for green entrepreneurial solutions, which parallels the expansion of green entrepreneurship.

To fill this gap, we expand the KSTE framework by incorporating demand-side factors relevant to the emergence of green startups, and we explore the interaction between this demand and the composition of the knowledge stock for green firm formation. To do so, we start by identifying green startups in Italian provinces by applying the BERTopic methodology developed in Chapter 1. Then, we approximate the demand for environmental sustainability with a proxy combining several indicators of environmentally sensitive behaviours using Principal Component Analysis (PCA). We measure the knowledge stock of each province using patent applications, and the share of green knowledge based on their Cooperative Patent Classification (CPC) code.

The results, consistent with the established literature, highlight the importance of knowledge stocks and underscore the role of demand-side factors. Specifically, we show that green demand—by increasing the short-term expected returns to innovation investment—facilitates firms' ability to leverage local knowledge in the creation of new ventures. Yet, contrary to expectations, green entrepreneurship depends more on the overall stock of knowledge than on environmentally specific knowledge. This suggests that green innovation emerges from recombining diverse knowledge bases rather than relying solely on green expertise.

1.5.3 Chapter 4

In this chapter, we examine how legitimacy contributes to entrepreneurial success. Specifically, we show that legitimacy helps entrepreneurs overcome the liability of newness by attracting attention and expanding their networks.

Legitimacy, defined as the perception that an entity is desirable for society, is beneficial for all ventures, and especially new ventures (Fisher, 2020; Zimmerman & Zeitz, 2002). To establish legitimacy, new firms must engage in legitimacy work aimed at persuading legitimating actors, the media, and consumers of their legitimacy. A substantial strand of the literature shows that firms establish legitimacy by sharing narratives, such as success stories about their ventures or posts highlighting their environmental and social engagement. Establishing legitimacy is particularly important for green new ventures, which are facing diverse and challenging audiences (O'Neil & Ucbasaran, 2016). Yet, due to methodological constraints, most of this research is qualitative and exploratory in nature and led to the identification of a wide range of dimensions of sustainability (Suddaby et al., 2017). Here, we contribute to the discussion on the relationship between sustainability and social networks. Prior research has largely focused on how strong ties with highly legitimate actors enhance a firm's legitimacy. We argue, instead, that legitimacy also plays an active role in facilitating the formation of new ties and the expansion of networks, as actors are more inclined to collaborate with organizations perceived as socially desirable.

We leverage the recent progress of NLP and LLMs to introduce a new method for identifying legitimacy claims in narratives. We focus on the three most fundamental dimensions of legitimacy (Stinchcombe, 1965): cognitive legitimacy, when a firm is perceived as understandable, taken-for-granted, and deemed necessary within a social context; normative legitimacy, when it aligns with societal norms and values (e.g., justice, human right, sustainability); and pragmatic legitimacy, when stakeholders believe it serves their interests. We apply it to the LinkedIn posts of 1,703 entrepreneurs to explore how cognitive, normative, and pragmatic legitimacy claims attract attention and contribute to increasing network size, and how this process differs between green and non-green firms.

Our results show that legitimacy claims have indeed a different impact on green and non-green firms. In particular, only cognitive and pragmatic legitimacy claims have a positive impact on online legitimacy, while non-green entrepreneurs also profit from normative legitimacy claims. This suggests that the environmental orientation of green startups inherently confers normative legitimacy, reducing the need for additional normative claims, and that their audiences have different expectations compared to non-green ones. Further, we show that the attention attracted by a post, as well as cognitive legitimacy in the case of non-green startups, affects positively the network size. Finally, by testing quantitatively legitimacy claims, we challenge results suggested by qualitative.

1.6 REFERENCES

Acs, Z. J., & Audretsch, D. B. (1987). Innovation, Market Structure, and Firm Size. *The Review of Economics and Statistics*, *69*(4), 567. https://doi.org/10.2307/1935950

Acs, Z. J., Audretsch, D. B., Braunerhjelm, P., & Carlsson, B. (2012). Growth and entrepreneurship. *Small Business Economics*, *39*(2), 289–300. https://doi.org/10.1007/s11187-010-9307-2

Acs, Z. J., Audretsch, D. B., & Lehmann, E. E. (2013). The knowledge spillover theory of entrepreneurship. *Small Business Economics*, *41*(4), 757–774. https://doi.org/10.1007/s11187-013-9505-9

Acs, Z. J., Braunerhjelm, P., Audretsch, D. B., & Carlsson, B. (2009). The knowledge spillover theory of entrepreneurship. *Small Business Economics*, *32*(1), 15–30. https://doi.org/10.1007/s11187-008-9157-3

Audretsch, D. B., & Lehmann, E. E. (2023). Narrative entrepreneurship: Bringing (his)story back to entrepreneurship: Narrative entrepreneurship: bringing (his)story back to entrepreneurship. *Small Business Economics*, *60*(4), 1593–1612. https://doi.org/10.1007/s11187-022-00661-2

Bail, C. A. (2024). Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, *121*(21), e2314021121. https://doi.org/10.1073/pnas.2314021121

Bornmann, L., & Mutz, R. (2014). *Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1402.4578

Castelló, I., Etter, M., & Årup Nielsen, F. (2016). Strategies of Legitimacy Through Social Media: The Networked Strategy. *Journal of Management Studies*, *53*(3), 402–432. https://doi.org/10.1111/joms.12145

Cepparulo, A., & Giuriato, L. (2024). Constitutionalizing the fight against climate change. Insights from France. *Environmental Science & Policy*, *157*, 103756. https://doi.org/10.1016/j.envsci.2024.103756

Cojoianu, T. F., Clark, G. L., Hoepner, A. G. F., Veneri, P., & Wójcik, D. (2020). Entrepreneurs for a low carbon world: How environmental knowledge and policy shape the creation and financing of green start-ups. *Research Policy*, *49*(6), 103988. https://doi.org/10.1016/j.respol.2020.103988

Colombelli, A., Krafft, J., & Vivarelli, M. (2016). To be born is not enough: The key role of innovative start-ups. *Small Business Economics*, *47*(2), 277–291. https://doi.org/10.1007/s11187-016-9716-y

Colombelli, A., & Quatraro, F. (2019). Green start-ups and local knowledge spillovers from clean and dirty technologies. *Small Business Economics*, *52*(4), 773–792. https://doi.org/10.1007/s11187-017-9934-y

De Groot, O. J., Bozzoli, C., Alamir, A., & Brück, T. (2022). The global economic burden of violent conflict. *Journal of Peace Research*, *59*(2), 259–276. https://doi.org/10.1177/00223433211046823

Fisher, G. (2020). The Complexities of New Venture Legitimacy. *Organization Theory*, *1*(2), 2631787720913881. https://doi.org/10.1177/2631787720913881

Fu, H.-Z., & Waltman, L. (2021). *A large-scale bibliometric analysis of global climate change research between 2001 and 2018* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2107.08214

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, *57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Ghisetti, C., & Rennings, K. (2014). Environmental innovations and profitability: How does it pay to be green? An empirical analysis on the German innovation survey. *Journal of Cleaner Production*, *75*, 106–117. https://doi.org/10.1016/j.jclepro.2014.03.097

Giudici, G., Guerini, M., & Rossi-Lamastra, C. (2019). The creation of cleantech startups at the local level: The role of knowledge availability and environmental awareness. *Small Business Economics*, *52*(4), 815–830. https://doi.org/10.1007/s11187-017-9936-9

Grinin, L., Grinin, A., & Korotayev, A. (2023). Global Aging: An Integral Problem of the Future. How to Turn a Problem into a Development Driver? In V. Sadovnichy, A. Akaev, I. Ilyin, S. Malkov, L. Grinin, & A. Korotayev (Eds.), *Reconsidering the Limits to Growth* (pp. 117–135). Springer International Publishing. https://doi.org/10.1007/978-3-031-34999-7_7

Hardin, G. (1968). The Tragedy of the Commons: The population problem has no technical solution; it requires a fundamental extension in morality. *Science*, *162*(3859), 1243–1248. https://doi.org/10.1126/science.162.3859.1243

Ike, D. (2025). Sustainable Entrepreneurship in Emergence: A Systematic Literature Review. *Sustainable Development*, *33*(S1), 1200–1214. https://doi.org/10.1002/sd.70058

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, *22*(1), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Jerit, J., & Zhao, Y. (2020). Political Misinformation. *Annual Review of Political Science*, *23*(1), 77–94. https://doi.org/10.1146/annurev-polisci-050718-032814

Jones, M. W., Peters, G. P., Gasser, T., Andrew, R. M., Schwingshackl, C., Gütschow, J., Houghton, R. A., Friedlingstein, P., Pongratz, J., & Le Quéré, C. (2023). National contributions to climate change due to historical emissions of carbon dioxide, methane,

and nitrous oxide since 1850. *Scientific Data*, *10*(1), 155. https://doi.org/10.1038/s41597-023-02041-1

Levin, S. A., Milner, H. V., & Perrings, C. (2021). The dynamics of political polarization. *Proceedings of the National Academy of Sciences*, *118*(50), e2116950118. https://doi.org/10.1073/pnas.2116950118

Mansouri, S., & Momtaz, P. P. (2022). Financing sustainable entrepreneurship: ESG measurement, valuation, and performance. *Journal of Business Venturing*, *37*(6), 106258. https://doi.org/10.1016/j.jbusvent.2022.106258

Nguyen, H. V., Nguyen, C. H., & Hoang, T. T. B. (2019). Green consumption: Closing the intention-behavior gap. *Sustainable Development*, *27*(1), 118–129. https://doi.org/10.1002/sd.1875

O'Neil, I., & Ucbasaran, D. (2016). Balancing "what matters to me" with "what matters to them": Exploring the legitimation process of environmental entrepreneurs. *Journal of Business Venturing*, *31*(2), 133–152. https://doi.org/10.1016/j.jbusvent.2015.12.001

Osberghaus, D., Botzen, W. J. W., & Kesternich, M. (2025). The intention-behavior gap in climate change adaptation: Evidence from longitudinal survey data. *Ecological Economics*, *231*, 108543. https://doi.org/10.1016/j.ecolecon.2025.108543

Patzelt, H., & Shepherd, D. A. (2011). Recognizing Opportunities for Sustainable Development. *Entrepreneurship Theory and Practice*, *35*(4), 631–652. https://doi.org/10.1111/j.1540-6520.2010.00386.x

Prince, S., Chapman, S., & Cassey, P. (2021). The definition of entrepreneurship: Is it less complex than we think? *International Journal of Entrepreneurial Behavior & Research*, *27*(9), 26–47. https://doi.org/10.1108/IJEBR-11-2019-0634

Riandita, A., Broström, A., Feldmann, A., & Cagliano, R. (2022). Legitimation work in sustainable entrepreneurship: Sustainability ventures' journey towards the establishment of major partnerships. *International Small Business Journal:*

Researching Entrepreneurship, *40*(7), 904–929. https://doi.org/10.1177/02662426211056799

Shan, L., Xu, Y., Jiao, X., Lu, Q., & Liu, X. (2025). Research on consumers' intention-behavior gap in sustainable diets: A moderating effects model incorporating face consciousness. *Sustainable Futures*, *10*, 101488. https://doi.org/10.1016/j.sftr.2025.101488

Suchman, M. C. (1995). Managing Legitimacy: Strategic and Institutional Approaches. *The Academy of Management Review*, *20*(3), 571. https://doi.org/10.2307/258788

Suddaby, R., Bitektine, A., & Haack, P. (2017). Legitimacy. *Academy of Management Annals*, *11*(1), 451–478. https://doi.org/10.5465/annals.2015.0101

Teruel, L. (2023). Increasing political polarization with disinformation: A comparative analysis of the European quality press. *El Profesional de La Información*, e320612. https://doi.org/10.3145/epi.2023.nov.12

Tiba, S., Van Rijnsoever, F. J., & Hekkert, M. P. (2021). Sustainability startups and where to find them: Investigating the share of sustainability startups across entrepreneurial ecosystems and the causal drivers of differences. *Journal of Cleaner Production*, *306*, 127054. https://doi.org/10.1016/j.jclepro.2021.127054

Tolle, K. M., Tansley, D. S. W., & Hey, A. J. G. (2011). The Fourth Paradigm: Data-Intensive Scientific Discovery [Point of View]. *Proceedings of the IEEE*, *99*(8), 1334–1337. https://doi.org/10.1109/JPROC.2011.2155130

Wennekers, S., & Thurik, R. (1999). Linking Entrepreneurship and Economic Growth. *Small Business Economics*, *13*(1), 27–56. https://doi.org/10.1023/A:1008063200484

Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., Wu, Y., Dong, F., Qiu, C.-W., Qiu, J., Hua, K., Su, W., Wu, J., Xu, H., Han, Y., Fu, C., Yin, Z., Liu, M., … Zhang, J. (2021). Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, *2*(4), 100179. https://doi.org/10.1016/j.xinn.2021.100179

Yamaguchi, N. U., Bernardino, E. G., Ferreira, M. E. C., De Lima, B. P., Pascotini, M. R., & Yamaguchi, M. U. (2023). Sustainable development goals: A bibliometric analysis of literature reviews. *Environmental Science and Pollution Research*, *30*(3), 5502–5515. https://doi.org/10.1007/s11356-022-24379-6

York, J. G., Vedula, S., & Lenox, M. J. (2018). It's Not Easy Building Green: The Impact of Public Policy, Private Actors, and Regional Logics on Voluntary Standards Adoption. *Academy of Management Journal*, *61*(4), 1492–1523. https://doi.org/10.5465/amj.2015.0769

Zhao, C., Liu, Z., & Zhang, C. (2023). Real or fictional? Digital entrepreneurial narratives and the acquisition of attentional resources in social entrepreneurship. *Journal of Innovation & Knowledge*, *8*(3), 100387. https://doi.org/10.1016/j.jik.2023.100387

Zimmerman, M. A., & Zeitz, G. J. (2002). Beyond Survival: Achieving New Venture Growth by Building Legitimacy. *The Academy of Management Review*, *27*(3), 414. https://doi.org/10.2307/4134387

CHAPTER 2

Identifying Green Start-ups: a Comparison of Three Natural Language Processing Algorithms

2 IDENTIFYING GREEN START-UPS: A COMPARISON OF THREE NATURAL LANGUAGE PROCESSING ALGORITHMS

2.1 ABSTRACT

Research on green entrepreneurship is growing, but there is a persistent challenge in accurately and efficiently identifying green start-ups. Natural language processing tools can analyze textual data without or with very limited human intervention. As such, they offer rapid, flexible, and cost-efficient ways of identifying green start-ups. In this article, we compare three approaches (Dictionary, Latent Dirichlet Allocation (LDA), and BERTopic) applied to 10,939 websites of Italian innovative start-ups from 2009 to 2023. We find that the three methods identify partially overlapping but distinct sets of green start-ups, primarily because each method captures different areas of environmental engagement. We show that these differences are relevant for analyzing green entrepreneurship and public policy. We relate the number of identified start-ups with regional SDG progress and spending from the National Recovery and Resilience Plan, and show that entrepreneurial activity and public policy align on the issues related to sustainable energy, sustainable development, and air quality, but that public policy contributes more to water quality as well as disaster resilience.

Keywords: Green firms; Artificial Intelligence (AI); Natural Language Processing (NLP); Sustainable Development Goals (SDG); innovative start-ups

2.2 INTRODUCTION

Green entrepreneurship is a growing field that attracts significant interest from policymakers, researchers, and organizations because of its potential to innovatively address environmental sustainability and climate-related issues (Dean & McMullen, 2007). Green entrepreneurs are key actors in developing environmentally sustainable economies and territories, thanks to their efforts to bring new green technologies to the market by creating green innovative start-ups (Colombelli & Quatraro, 2019; Horne & Fichter, 2022). However, there is no consensus on how to measure the sustainability orientation of start-ups, and identifying which start-ups are “green” remains challenging despite the growing interest in their role in the transition toward sustainable development. To date, there is no universally accepted definition of “green start-up” (Nikolaou et al., 2018), and comparisons between different identification methodologies remain limited.

So far, most academic research has relied on indirect identification of green firms, for example, through third-party assessment and financial indexes such as S&P500 Environmental, Social and Governance (ESG) leaders or MSCI ratings, or through green patents. These solutions typically require expensive datasets and are not easily applicable to start-ups, as they leave out green entrepreneurship that does not lead to patenting or listing on financial markets (Berg et al., 2022). Alternatively, other studies rely on questionnaires, surveys, and self-assessments, making the data collection process labor-intensive, prone to biases and errors, and difficult to scale and reproduce (Mrkajic et al., 2019).

As an alternative, researchers have proposed different solutions to identify sustainable start-ups based on Natural Language Processing (NLP) of companies’ documentation, such as Sustainability Reports or websites. With this approach, researchers can consider a broader set of firms in their identification effort. The earliest and conceptually simplest solution is the so-called “*dictionary approach*”, which classifies documents based on the presence of

keywords extracted from a predefined dictionary (Gorovaia & Makrominas, 2024; Stone et al., 1966).The key limitation of this rule-based method is that setting up the dictionary implies discretionary choices about the concepts linked to “green entrepreneurship”.

Recently, a few studies have combined NLP and Artificial Intelligence (AI) to run textual analyses without previously classified (“labelled”) data or established dictionaries. For instance, Mansouri & Momtaz (2022) used a Machine Learning (ML) dictionary approach to identify ESG start-ups, building dictionaries directly from the texts of companies’ websites. Similarly, Tiba et al. (2021) employed Latent Dirichlet Allocation (LDA), an unsupervised topic model, to conduct a semi-automated identification of sustainable start-ups.

Another key advancement in machine-learning NLP has been BERTopic (Grootendorst, 2022), an unsupervised topic model designed to better capture textual contexts. This method has been shown to outperform LDA and other unsupervised topic models (Egger & Yu, 2022; Umamaheswaran et al., 2023), but it has not been applied to green companies’ identifications so far. All these approaches are “neutral” in classifying start-ups since they recognize data patterns rather than relying on manual categorizations, which are inherently subjective and could introduce discretionary biases.

However, while these three AI-based methodologies hold great promise for identifying green start-ups, their full potential remains unclear without a direct comparison of their outcomes. A systematic evaluation of these empirical approaches is essential to determine their strengths, limitations, and specificities, and ultimately to help researchers choose the most effective approach for different types of analysis. Given the lack of a theoretical definition of “greenness”, the importance of empirical identification cannot be overstated. The present study contributes to the discussion about the empirical identification of green entrepreneurship through an in-depth comparison of three approaches: ML dictionary, LDA, and BERTopic - three highly promising unsupervised NLP methods leveraging machine learning.

From an application to 10,939 Italian startups, we find that the three methods identify partially overlapping but different sets of green start-ups, mainly because each method leads to the identification of green start-ups within different areas of environmental engagement. First, LDA is particularly fit for identifying energy start-ups, while BERTopic and ML dictionary detect start-ups from different and broader “shades of green”, such as those related to “Water Quality Management and Sustainable Use” or “Environmentally Sound Waste and Resource Management”. Second, from a technical perspective, BERTopic emerges as the most proficient method, since it captures the contextual meaning of words, while LDA and ML dictionary treat words as independent units (Egger & Yu, 2022; Grootendorst, 2022). Topic models can be useful tools for discovering latent topics in collections of documents. Recent studies have shown the feasibility of approach topic modelling as a clustering task. We present BERTopic, a topic model that extends this process by extracting coherent topic representations through the development of a class-based variation of TF-IDF. More specifically, BERTopic generates document embedding with pre-trained transformer-based language models, clusters these embeddings, and finally, generates topic representations with the class-based TF-IDF procedure. Compared to traditional topic models, the use of embeddings allows BERTopic to better understand semantic relationships between words,

leading to a better understanding of context and to the generation of richer topics (Egger & Yu, 2022; Grootendorst, 2022). This enhanced sophistication of BERTopic is especially valuable for exploring green entrepreneurship, an inherently complex domain with overlapping subtopics and interconnected themes that rely on continuously evolving language (e.g., decarbonization strategies, net-zero commitments, climate neutrality, climate-friendly technology).

Third, we show how such differences are significant for analyzing green entrepreneurship and public policy. We compare the number of identified green startups with public expenditures in SDG environmental dimensions within the Italian National Recovery and Resilience Plan (NRRP) and the SDG performance assessment of Italy by the Sustainable Development Solutions Network (SDSN), a UN-affiliated non-profit organization. While the three machine-learning NPL approaches agree on the two SDG environmental dimensions in which most green startups operate, there is substantial variation in SDG categories with fewer green startups, leading to different conclusions regarding the interplay of private and public forces to reach SDG goals.

The article is organized as follows. Section 2 positions our research within the literature on green entrepreneurship. Section 3 illustrates the data. Section 4 presents the AI-based classification methodologies, followed by their application in Section 5. Finally, Section 6 concludes.

2.3 BACKGROUND

Given the rise of environmental and climatic challenges worldwide and the critical need for sustainable innovations, the urge for a deeper understanding and better measurement of green entrepreneurship has never been greater. In this section, we present the relevant background on green entrepreneurship and the different measurement methods used in the literature.

2.3.1 Trends and definition of green entrepreneurship

Green entrepreneurship is a central policymaking theme (Lundmark & Audretsch, 2024; Rizzitello et al., 2025). OECD countries are implementing policies supporting green start-ups by developing green skills, incubators, and accelerators, facilitating access to green and sustainable finance, and implementing local public initiatives and public-private partnerships (OECD, 2022). At the European level, initiatives such as the Small Business Act or the Green Action Plan support the creation of green SMEs (Abdesselam et al., 2024).

In the economic sphere, its importance is also rising. Startup events such as VivaTechnology are dedicating spaces for showcasing green start-ups (VivaTechnology, 2025). Incubators specifically targeting climate technology are being created worldwide (OECD, 2022) and are assessed on their green performance (Fonseca & Chiappetta Jabbour, 2012). While early research argued that traditional investors rejected sustainable entrepreneurship, recent studies suggest that green start-ups are as attractive or even more than their “brown” counterparts (Hörisch, 2015; Mrkajic et al., 2019; Wöhler & Haase, 2022).

Academic research on green entrepreneurship started in the 1990s and has attracted increasing attention since the last decade (Anand et al., 2021; Olawumi & Chan, 2018; Sabando-Vera et al., 2025). The term “green entrepreneurship” has been used interchangeably with others such as “sustainable entrepreneurship”, “ecological entrepreneurship”, “environmental entrepreneurship”, “ecopreneurship”, and “enviropreneur” (Gast et al., 2017). We will use the term “green entrepreneurship” since it is the most frequent.

Green entrepreneurship can be defined in multiple ways depending on the context, either narrowly or broadly (Misztal & Kowalska, 2023). Narrow definitions view it as the creation of new business ideas with environmental sustainability as a fundamental principle (Kirkwood & Walton, 2010; Rodríguez-García et al., 2019). Other narrow definitions describe it as entrepreneurship leading to the creation of green goods and services (Shepherd & Patzelt, 2011); as entrepreneurship introducing new answers to demands in a market suffering from environmental inefficiencies (Dean & McMullen, 2007); or as initiatives that introduce green innovations (Bendig et al., 2022; Coll-Martínez et al., 2022). While all those definitions aim to identify the same construct, their differences may lead to the identification of subsets of green start-ups with substantial differences.

A broader definition conciliating those differences comes from the sustainability literature: Pacheco et al. (2010) define sustainable entrepreneurship as the entrepreneurial processes that contribute to sustainability goals. In this line, Horne & Fichter (2022) define sustainable start-ups as those with a better-expected impact on Sustainable Development Goals (UN, 2015) than the status quo. Various international organizations are developing sustainable taxonomies (IPSF Taxonomy Working Group, 2021), but the SDG is the most established one to date (UN, 2015).

However, some research has emphasized the importance of distinguishing “green” from the umbrella term “sustainable entrepreneurship”, which includes not only environmental issues but also social and economic ones (Halberstadt et al., 2024; Hörisch, 2015). Even though we avoid setting an ex-ante definition, we need a reputable and widely accepted corpus of text that contains the broadest possible range of topics capturing “greenness”. Such a corpus will feed the AI-based, unsupervised NLP algorithms to extract the definition from the data and identify green start-ups. For this reason, we focus on entrepreneurial processes that contribute to the *green* SDGs’ targets, as identified by the Global Environment Outlook 6 (UN, 2019), which selects 70 SDG targets related to environmental sustainability. Prior research used a less granular definition, characterizing the green dimension at the goal level of the SDG with four macro goals: water and sanitation, affordable and clean energy, climate action, and life below water (Gidron et al., 2023; Shi et al., 2019). As shown in our study, these macro-level goals hide other green topics discussed in more detail at the target level. Using the targets identified by GEO6 allows for the identification of these green topics, such as disaster resilience and climate-related risks.

2.3.2 Measuring green entrepreneurship and start-ups

The most used proxy for entrepreneurship is based on the start-up count (Cojoianu et al., 2020). However, when looking at green start-ups, there is no consensus method to identify them. Researchers have relied on third-party data providers (Cojoianu et al., 2024; Dong et al., 2022), questionnaires (Abdesselam et al., 2024; Chapman & Hottenrott, 2022; Hörisch, 2015), press articles (Gebhardt & Bachmann, 2023), or individual inspection of documents describing the start-ups (Hossnofsky et al., 2025; Jorzik et al., 2024). Those data are either prone to biases, time and financially expensive to acquire, and are hardly scalable and reproducible. Other researchers have used a sample of green firms, such as energy start-ups (Cojoianu et al., 2020; Colombelli & Quatraro, 2019) or start-ups with a green patent (Coll-

Martínez et al., 2022; Corradini, 2019; Mazaheri et al., 2024). Limitations of patent data for measuring innovation are well known (Griliches, 1979, 1990)[1], and focusing only on one green sector lacks representativeness.

Recently, studies have used NLP applied to websites to tag sustainable or ESG-related start-ups (Mansouri & Momtaz, 2022; Tiba et al., 2021). Websites are publicly available documents that reflect start-ups' engagement (Bottai et al., 2024; Tiba et al., 2021). While some websites use greenwashing tactics (Montgomery et al., 2023), start-ups are unlikely to engage extensively in such practices because of the liability of newness (Bruderl & Schussler, 1990; Stinchcombe, 1965), their focus on introducing their core business to the market, and the growing importance of trust and transparency (Rizvanović et al., 2023). Further, research has shown that greenwashing negatively affects new business performances (Neumann, 2021).

Identifying green start-ups based on the content of their websites is a text classification task. Three approaches are commonly used to classify text based on topics: supervised classification, unsupervised topic modelling, and dictionary approaches (Arseneau et al., 2022). Studies have explored supervised classification algorithms to identify sustainable start-ups: for example, Gidron et al. (2023) fine-tuned a BERT model on 4,247 labelled start-ups to identify sustainable ones. However, training such algorithms requires labelled data, which is rarely available and time-consuming to construct. In this paper, we present approaches that only require public data. Hence, we will not explore supervised classification, focusing on dictionary and unsupervised topic modelling approaches.

The oldest NLP approach used for text classification is the dictionary approach, relying on pre-established dictionaries to identify relevant keywords in texts (Guo et al., 2016). For identifying text written about sustainability, researchers have used external dictionaries (Gorovaia & Makrominas, 2024) or written tailored dictionaries to fit their classification framework, such as the SDGs (Horne et al., 2020; Hossnofsky et al., 2025). Researchers have recently implemented a Machine Learning (ML) dictionary approach to avoid subjectivity and generate specific dictionaries automatically. Such an approach has recently been used by Mansouri & Momtaz (2022) to build a dictionary for each ESG framework's dimensions. They initialized their dictionaries with keywords related to each ESG dimension and expanded them with words from the Corporate Social Responsibility (CSR) reports of start-ups that were semantically close.

In contrast, unsupervised topic modelling identifies latent topics in a corpus (such as a list of websites). The most used topic model is LDA (Blei et al., 2003), which utilizes the probabilistic distribution of words in texts to represent documents as a finite set of topics. A distribution of words characterizes these topics, and each document is defined as a mixture of topics, reflecting their relative prominence. Such topics offer valuable insight and can uncover hidden patterns in a corpus, which can be leveraged to categorize texts.

Recently, a more sophisticated unsupervised topic model has been developed: BERTopic (Grootendorst, 2022). BERTopic leverages pre-trained embedding models, trained on large text corpora, to associate words and sentences with high-dimensional vectors that capture their semantic meaning. The embeddings of similar words, such as "ecosystem" and "biodiversity," are close to each other, while the embeddings of unrelated words, such as "ecosystem" and "fuel," are far apart. The first embedding models were static, assigning a single vector to each word. More recent models, such as SentenceBERT (Reimers &

[1] Strong limits of patents are the difficulty to measure their intrinsic variability (some are hugely important while many others are small and incremental) and their limited coverage of innovation (not every type of innovation is easily patented).

Gurevych, 2019), generate different vectors depending on context. For example, the word "environment" would have very different embeddings in "work environment" versus "natural environment."

This embedding representation allows for performing mathematical operations on text. After representing the texts as embeddings, BERTopic clusters similar texts to identify meaningful topics. BERTopic has been compared to LDA to map climate business news and showed better results in terms of human interpretation and statistical indicators on similar tasks (Egger & Yu, 2022; Umamaheswaran et al., 2023). Its better representation of textual context leads to more nuanced insights and a deeper understanding of the underlying topics. However, it has never been used to identify green start-ups.

All these individual studies are promising, but AI-based NLP approaches have never been applied comparatively to identify green start-ups. Such a comparison is needed to understand the implications of using one method against the other. In this work, we fill this gap by executing three AI-based algorithms to identify green start-ups: an ML dictionary approach, an LDA- and a BERTopic-based approach to compare the different types of green firms identified[2].

2.4 DATA

The sample of start-ups we use to compare the three AI-based identification methods is the universe of innovative start-ups in Italy. In 2012, the Italian government launched the Italian Start-up Act, granting advantages to young, innovative, and highly skilled companies. To register as an innovative start-up at the Chamber of Commerce in Italy, a firm must be less than 5 years old, small (annual turnover <5 million, no dividends distribution, not listed on a stock market), and innovative (owning a patent or a license, spending in R&D investment > 15% of the revenues; or with 1/3 of the employees holding a PhD or 2/3 of the employees holding a master's degree).

As of May 2023, 26,892 Italian start-ups were registered as innovative start-ups. We used AIDA, an Amadeus-Bureau Van Dijk database, to access firm-level information for these companies, such as a link to their websites and their localization. Using this data, we implemented a scraping and translation procedure, detailed in Appendix A, to gather the full text of the websites of 10,939 start-ups.

To determine whether this subsample of web-scraped startups is representative of the population of innovative startups, we conducted several analyses. First, we explored the distribution by ATECO 2-digit code, the classification of economic activities used by the Italian National Statistical Institute (ISTAT). We plot the 10 most frequent ATECO codes for each

[2] Two other approaches are of interest: the traditional dictionary approach and the approach using Language Models (LM) fined-tuned for text classification. These approaches used respectively an external dictionary and labelled data for fine-tuning, hence they do not fit the scope of our analysis. Still, we implement in Appendix A a traditional dictionary approach for identifying green startups following Gorovaia & Makrominas (2024) and an LM approach using ClimateBert (Webersinke et al., 2022). The traditional dictionary method using reaches lower precision than the ML dictionary method, and the precision of ClimateBERT is on par with LDA and BERTopic.

sample and a category ‘others’ for readability in Fig 2.1. While the legal definition of innovative start-ups does not require activity in a specific technical field, both the registered population and subsample are empirically concentrated in software, ICT, and other knowledge-intensive services[3], and are very similar across the other sectors.

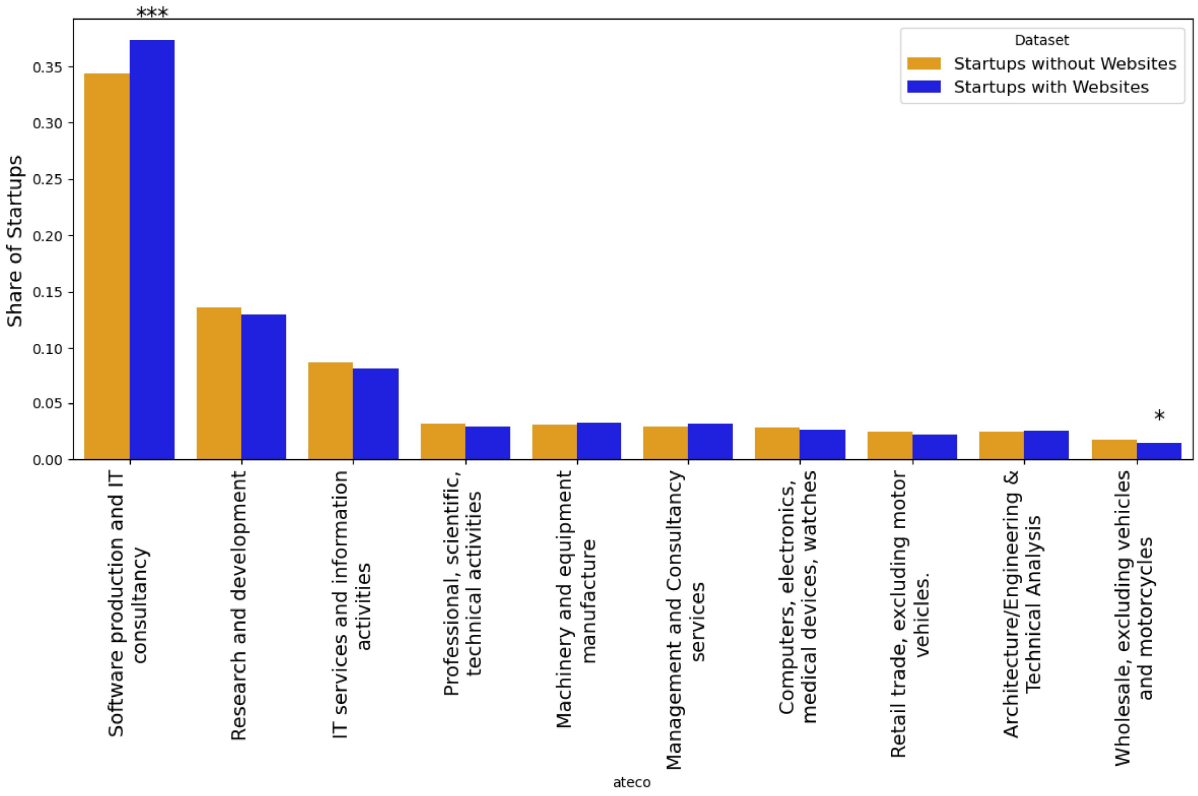


Fig 2.1 Comparison of areas of activities (measured by ATECO codes) between startups without and with websites. Only the top 10 ATECO are represented, covering three-fourths of the start-ups.

Further, we compare the population and subsample on a few economic characteristics. Table 2.1 shows that startups with websites are slightly younger, larger (as measured by assets or employees), and have a higher relative growth rate. Differences are statistically significant, but their magnitude is not economically worrisome.

Table 2.1 Comparison of statistics of startups without and with websites

Statistic	Startups without websites	Startups with websites
Age	6.88	6.25***
	(3.22)	(2.97)
Assets	132.55	158.91**
	(193.98)	(208.62)
Number of employees	1.77	2.21***
	(6.55)	(7.75)
Relative growth	0.02	0.03***

(0.06) (0.07)

In Fig. 2.2 and Fig. 2.3, we plot the geographical distribution of startups with and without websites in Italy, showing a very similar distribution.

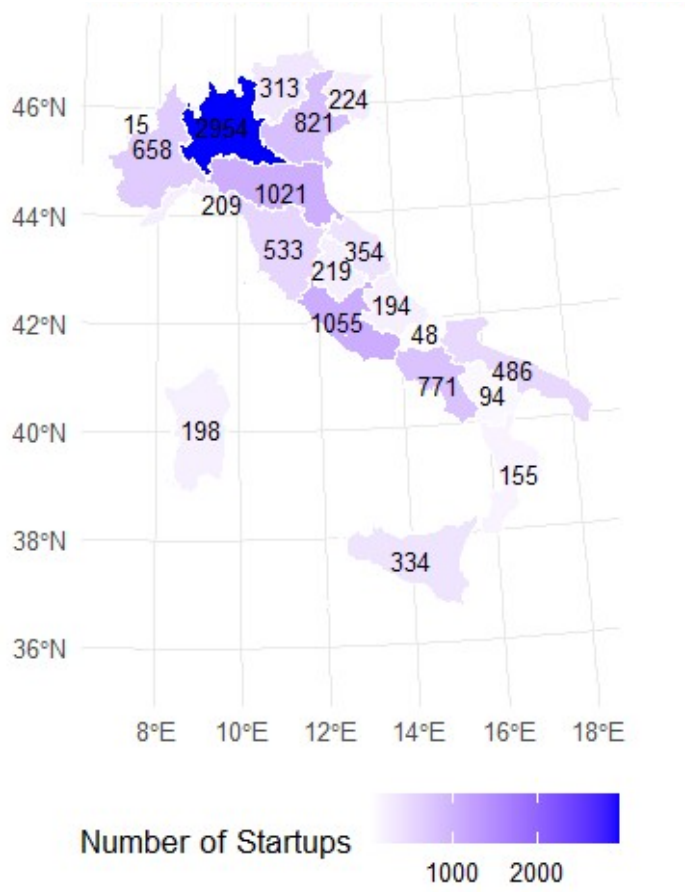


Fig 2.2 Number of startups with websites in Italian regions

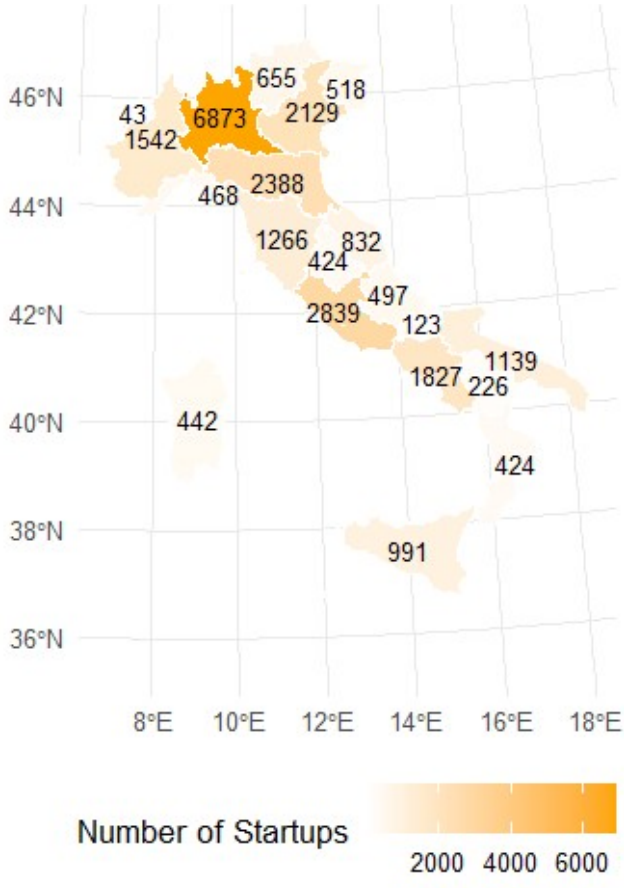


Fig 2. 3 Number of startups without websites in Italian regions

Finally, we plot in Fig. 2.4 the innovation criteria that startups with and without websites met. Startups without websites are slightly more likely to meet the high R&D criteria and less likely to meet the High Human Capital and Patent Owner criteria, but, again, the magnitude of the difference is not relevant for our methodological application.

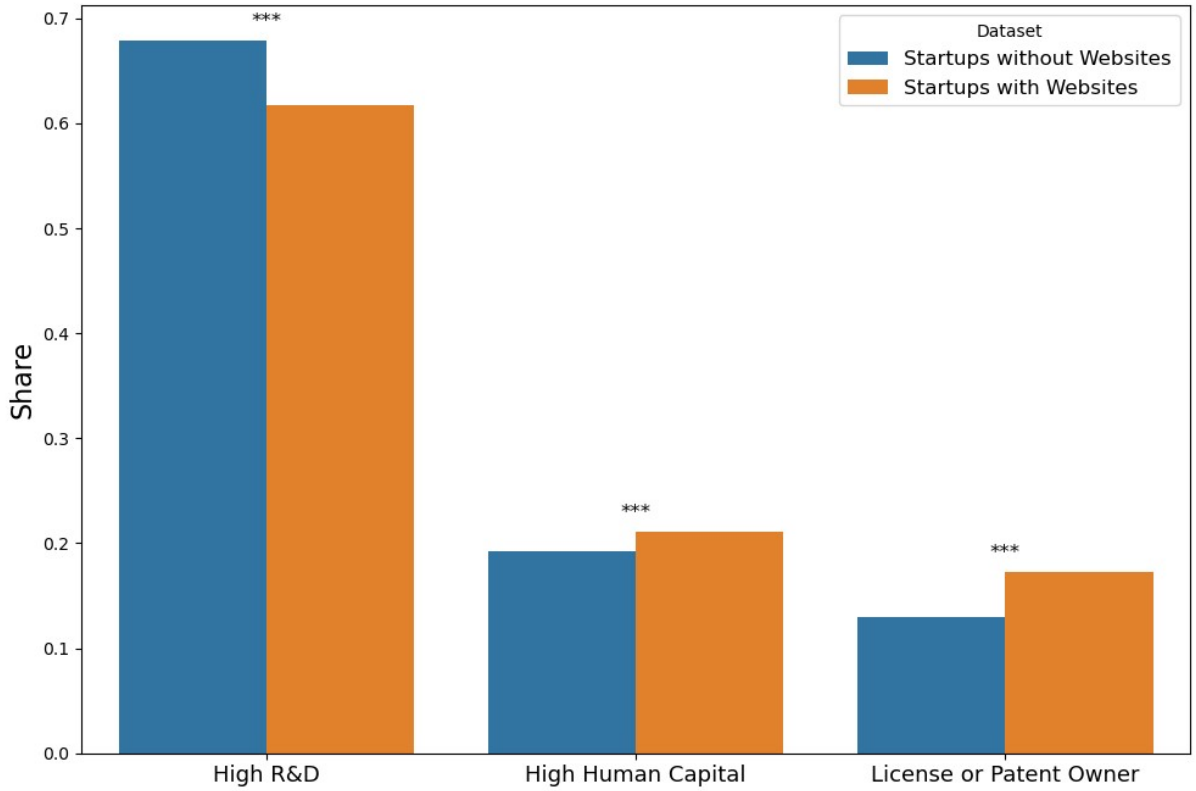


Fig 2.4 Barplots of innovation criteria met by the startups with and without websites

2.5 GREEN CLASSIFICATION FRAMEWORK AND METHODS

2.5.1 Green classification framework

Before introducing the three different algorithms to identify green startups, we define a green classification framework that articulates the concept of "greenness" sought within the data. While numerous sustainable taxonomies exist (IPSF Taxonomy Working Group, 2021), we rely on the SDGs framework, the most well-known and used in academic research (Hajikhani & Suominen, 2022; Mio et al., 2020; Van Zanten & Van Tulder, 2021). As the SDGs 17 goals with their 169 targets encompass economic, social, and environmental sustainability goals, we must refine them to identify green start-ups. In the Global Environment Outlook – GEO6 (UN, 2019), the UN identified 70 targets in 16 SDGs related to environmental sustainability. Yet, those targets remain multidimensional and encompass economic, social and environmental dimensions. Moreover, different SDG targets mention similar green goals. For instance, target 4.7 includes '*education on sustainable development*', while target 13.3 includes '*education on climate change*'.

First, we manually extract from each green SDG target any green “noun phrase”, defined as a group of words centered around a noun, which acts as a single unit within a sentence (for instance, “natural resources” or “renewable energy”), for a total of 220 noun phrases. Then, we use SentenceBERT (Reimers & Gurevych, 2019) all-MiniLM-L6-v2, a language model trained on English texts to represent the noun phrases as embeddings. Embeddings are mathematical representations of texts considering the contextual complexity of language. For instance, they can distinguish the meaning of the word ‘environment’ in ‘work environment’ versus ‘natural environment’. Then, we cluster the embeddings into 14 topics[4] using K-Means. Finally, we use ChatGPT 4.0 to assign green labels to the 14 green topics with the following prompt: “Here is a list of green noun phrases: {green noun phrases}. Please provide a concise topic label that effectively represents and captures the overall theme shared by all these noun phrases”. The different steps of this procedure are in Fig. 2.1 and details in the Appendix A.

In the three AI-based algorithms, we identify start-ups as “green” based on the 14 green topics (Table 2.1) using ML dictionary (Li et al., 2021), LDA (Blei et al., 2003) and BERTopic (Grootendorst, 2022).

[4] To define the appropriate number of topics in the list of green noun phrases, we clustered their embeddings with HDBSCAN, which automatically sets the number of topics to balance topics density (how close the embeddings are to another) and stability (its persistence over a range of density thresholds) and suggested 14 as the optimal number. We also explored using fewer or more topics, which did not appear to be appropriate since it led to the merging of two semantically different topics or the creation of two semantically similar topics.

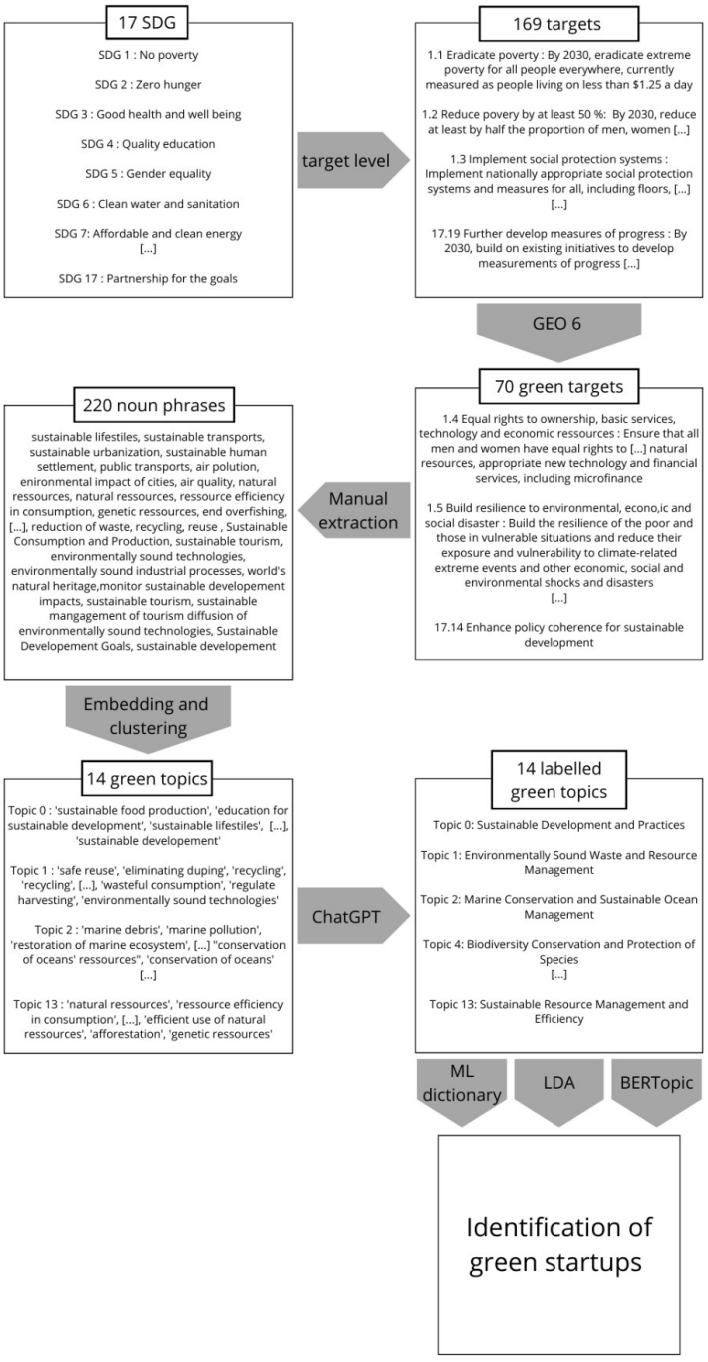


Fig 2.5 From SDG targets to green topics to green labels

Table 2.2 List of the 14 green topics' labels

	Green Topic Label
1	Air Quality and Pollution Management
2	Biodiversity Conservation and Protection of Species
3	Climate Change Adaptation and Mitigation Strategies
4	Disaster Resilience and Climate-related Risk Management
5	Environmentally Sound Waste and Resource Management
6	Forest Conservation and Sustainable Management
7	Freshwater Ecosystem Conservation and Restoration
8	Land and Ecosystem Conservation and Restoration
9	Marine Conservation and Sustainable Ocean Management
10	Sustainable Development and Practices
11	Sustainable Energy Services and Efficiency
12	Sustainable Fisheries Management and Illegal Fishing Prevention
13	Sustainable Resource Management and Efficiency
14	Water Quality Management and Sustainable Use

2.5.2 Machine Learning (ML) dictionary approach

To build a dictionary that is specific to our dataset and green framework, we developed a method close to Li et al. (2021), which has been applied by Mansouri & Momtaz (2022) to assign Environmental, Social, and Governance (ESG) scores to start-ups. For each of the dimensions of the ESG, Mansouri & Momtaz (2022) initialize a dictionary with the most used words in newspaper articles mentioning the ESG dimension as seed words. Then, they extend each dictionary with words from the CSR report that are semantically close to the seed words. Finally, they assign each startup an ESG score as the share of ESG words in their CSR report.

Following this procedure, we initialize a dictionary for each of the 14 green topics we derived in section 4.1 with the topics' labels as seed words. After standard preprocessing steps (stopword removal and lemmatization), we extract all groups of 1, 2, and 3 words (i.e. n-grams) found in start-up websites. Using sentenceBERT, we compute the embeddings of the seed words and the n-grams. Finally, we compute the cosine similarity between these two embeddings, which represents the semantic similarity of the words. We extend the dictionaries with all n-grams with cosine similarity to seed words higher than an arbitrary threshold of 0.5, identifying between 5,000 and 75,000 n-grams for each of the 14 dictionaries.

Finally, we compute a green dictionary score for each webpage as the number of green n-grams divided by the total number of n-grams in the text. We assign each start-up the highest green score of its webpages and identify the 15% of start-ups[5] with the highest score as green. Fig. 2.2 shows the distribution of the scores. All the steps of this methodology are presented in detail and compared to a traditional dictionary method in the Appendix A.

[5] This value is in line with Giudici et al. (2019), who identified a little less than 15% of the Italian startups as cleantech startups. Other thresholds are explored in the Appendix.

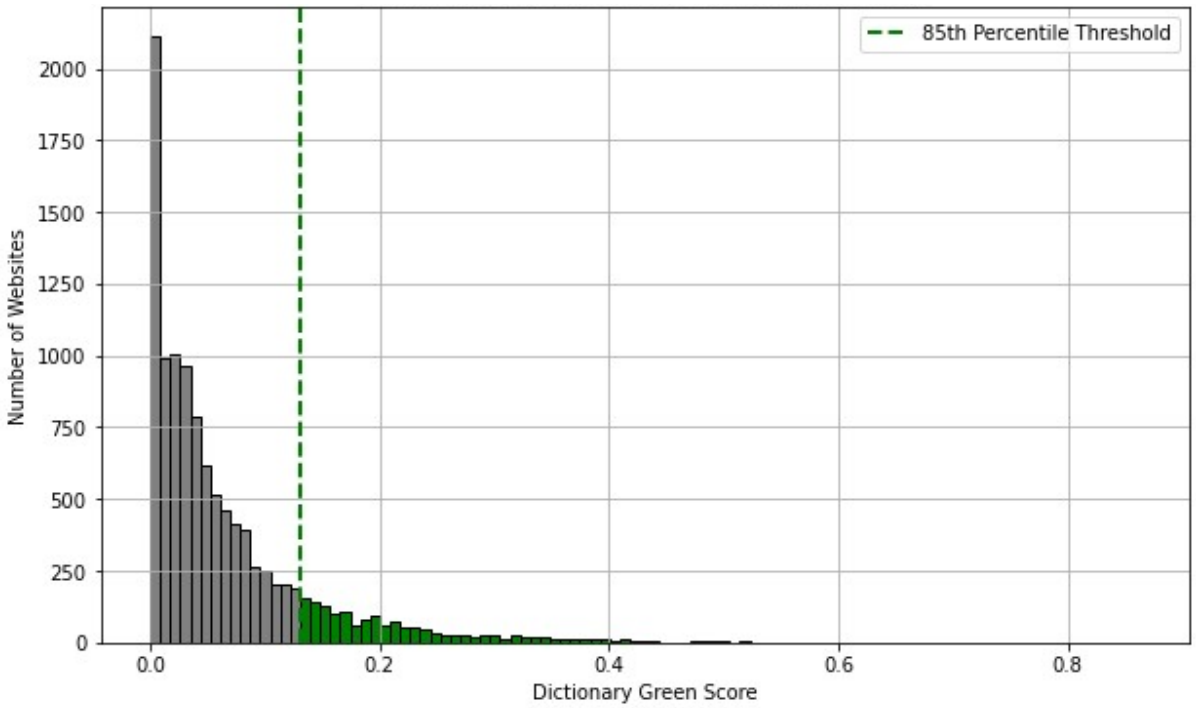


Fig 2.6 Distribution of websites' green score of the ML dictionary approach

2.5.3 Latent Dirichlet Allocation (LDA)

Second, we implement a well-established unsupervised topic model for natural language processing: Latent Dirichlet Allocation (LDA). It identifies which of the *k* topics are discussed in a corpus of text (Blei et al., 2003). Although it uses the bag-of-words representation of texts, which does not reflect textual context or embeddings, it is still widely used for corpus exploration and text classification (Yun & Geum, 2020).

Before running LDA on our clean corpus of text from the websites, we must set the number of topics *k.* The variety of start-ups prevents us from assuming the number of topics a priori. While researchers have set the number of topics *k* for optimizing statistical indicators, such indicators are uncorrelated with human judgment (Chang et al., 2009). Thus, we followed Tiba et al. (2021), trying different numbers of topics and finally setting it at 50, as more topics did not show noticeable improvement.

We classify the 50 LDA topics as green by computing their embeddings with SentenceBERT, comparing them to the 14 SDG green topics identified in section 4.1, and using cosine similarity as a criterion. We set the cosine similarity threshold at 0.5 and identify between 0 and 4 topics on each LDA run. Each webpage is assigned a green score based on its alignment with these topics, and websites are considered green if they contain at least one green webpage. We account for LDA's stochasticity by using a majority vote across 25 LDA iterations and assign to each start-up a green probability equal to its probability to be identified

as green on a single iteration. The top 15% of websites with the highest green scores are classified as green, identifying 1,520 start-ups (Fig. 2.7).

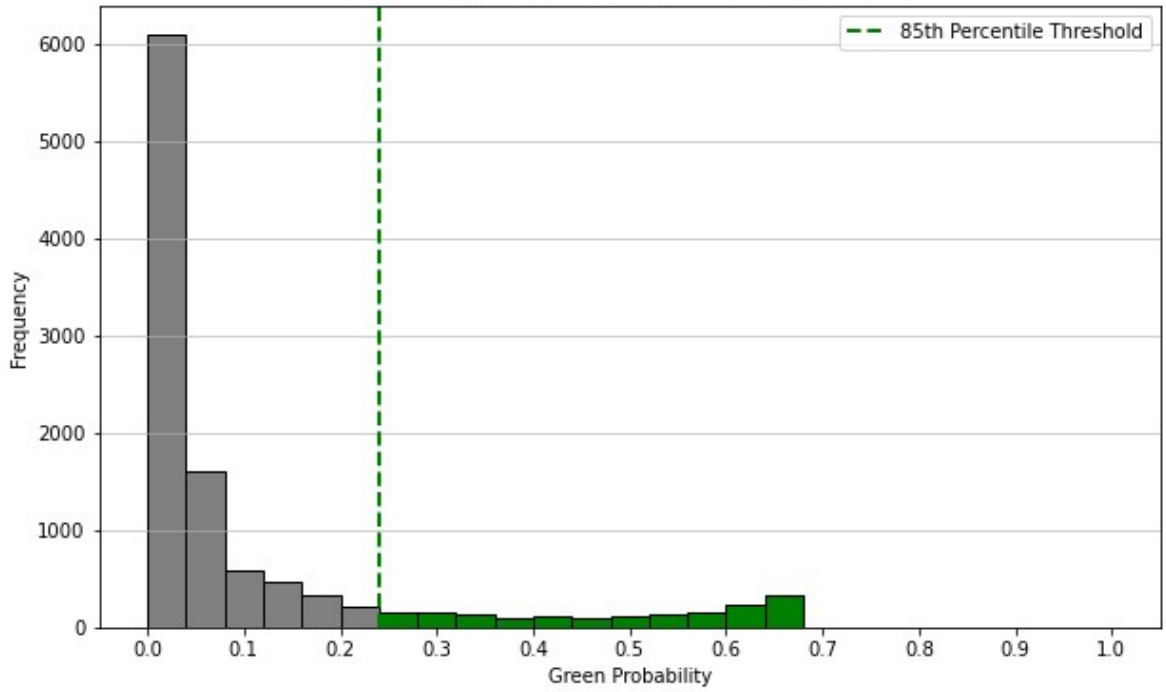


Fig 2. 7 Distribution of websites' green probabilities estimated by 25 iterations of LDA

Most websites are never identified as green

2.5.4 BERTopic

Third, we examine BERTopic, a state-of-the-art unsupervised topic model introduced by Grootendorst (2022), which leverages embeddings to identify topics in a corpus of text. It starts by embedding the texts, then clusters them into topics, and finally generates a topic representation for each cluster. This approach has already been compared in other contexts to unsupervised modelling techniques and has shown many advantages (Egger & al, 2022; Umamaheswaran & al, 2023), such as taking better account of the semantic context of the words, automatically determining the number of topics, and uncovering new insights from the data. While it has already been widely used in research, this paper is, to our knowledge, the first to apply it to identify green start-ups.

Our implementation of BERTopic can be divided into three steps. First, we embed the text of our webpages using SentenceBERT. Second, we reduce the dimensionality of the embeddings with UMAP (McInnes et al., 2020) and cluster them with HDBSCAN (McInnes et al., 2017). Third, we generate a topic representation for each cluster using a c-TF-IDF variant (Borčin & Jose, 2024). In the Appendix A, we discuss in detail the algorithms that we used for each step.

BERTopic assigns to each text its predominant topic. Then, we identify the green topics by computing the cosine similarity between each BERTopic topic representation and the SDG green topics from 4.1, assigning each webpage a green cosine score. At the website level,

each website is assigned the green cosine score of its highest-scoring webpage. To reduce BERTopic's stochasticity, 25 iterations are run, and websites in the top 15% of green cosine scores are identified as green in each run. We assign to each start-up a green probability, the probability of its website being identified as green in each run. Finally, we classify the top 15% of websites with the highest green probability as green, resulting in 1,524 green start-ups. (Fig. 2.8). The steps of the BERTopic methodology are described alongside those of the LDA methodology in Fig. 2.9[6].

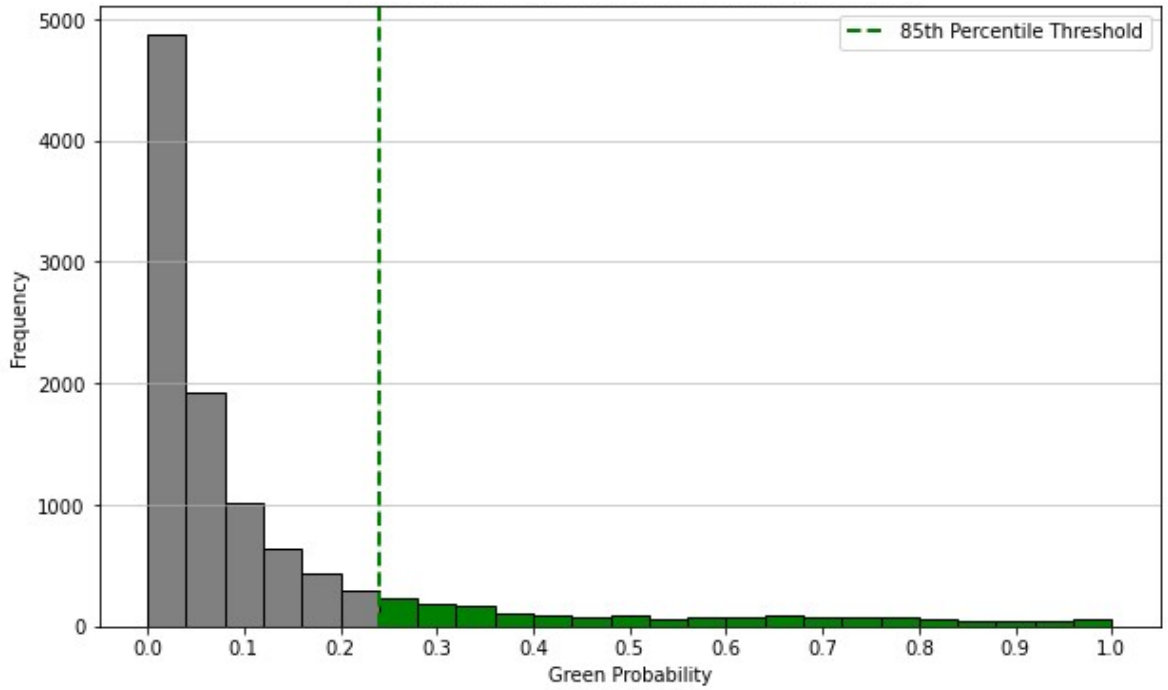


Fig 2.8 Distribution of websites found green by BERTopic for 25 iterations

More than 3500 websites are never identified as green, while very few websites have very high green probabilities.

[6] A similar figure describing the procedure to execute the ML dictionary approach in comparison with the traditional dictionary approach is in the Appendix (Fig. A.5).

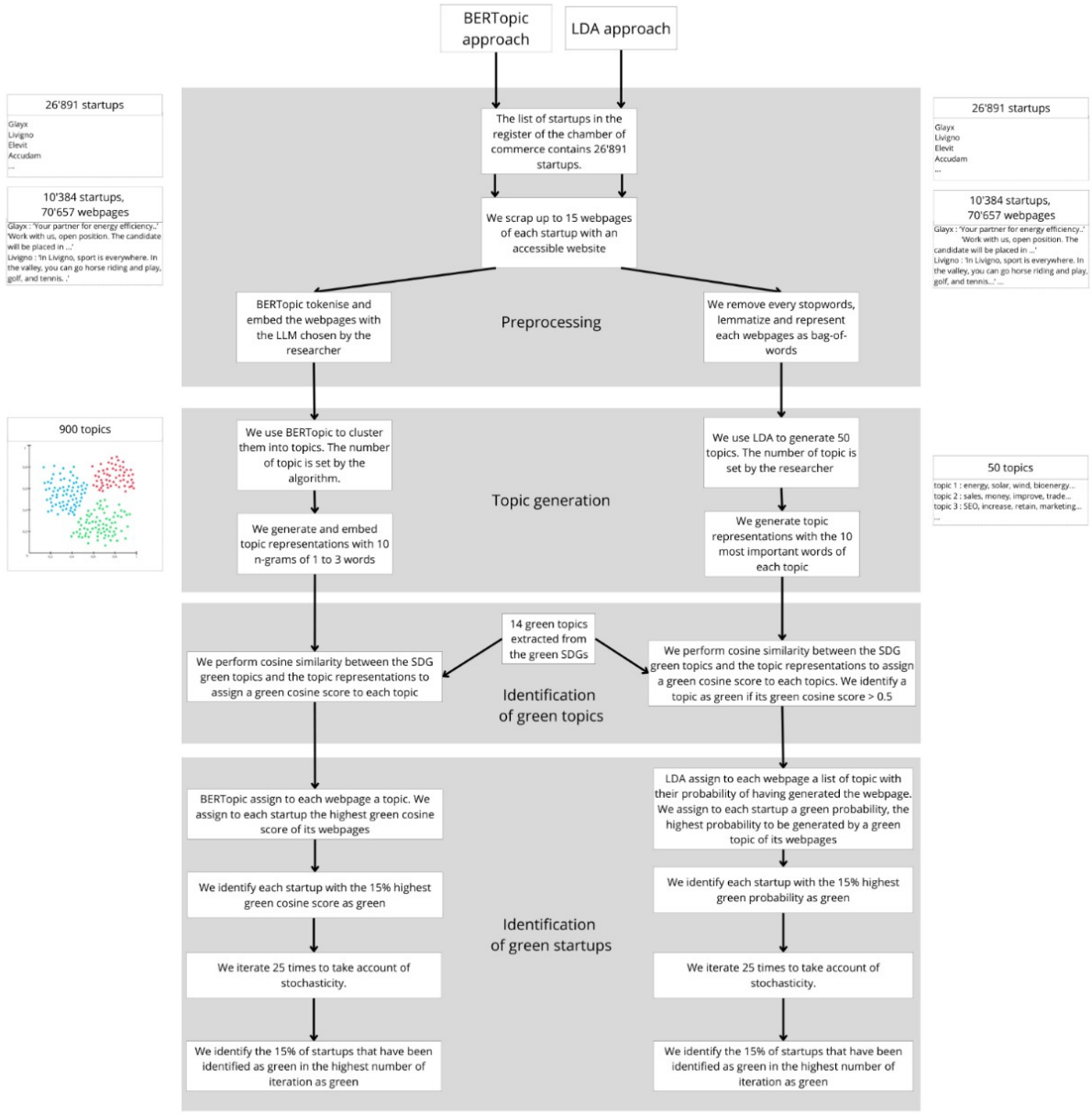


Fig 2.9 Summary of LDA and BERTopic to identify green startups

2.5.5 Evaluation of the approaches

We now present and compare the results of the ML dictionary, LDA, and BERTopic approaches in identifying the start-ups with the highest scores as “green” (different thresholds lead to similar results; see the Appendix A). The three approaches identify a similar number of green start-ups: the ML dictionary identifies 1,673 green start-ups, while LDA identifies 1,520, and BERTopic 1,524. The start-ups identified cover a wide range of activities, from green products or services, environmentally friendly processes, green certifications, or explicitly declaring their commitment to environmental sustainability (examples in the Appendix A). Our first result is that the three methodologies present a substantial overlap in identifying start-ups, sharing around 40% (624) of their green start-ups, but each identifies start-ups not detected by the other methods (Fig. 2.6).

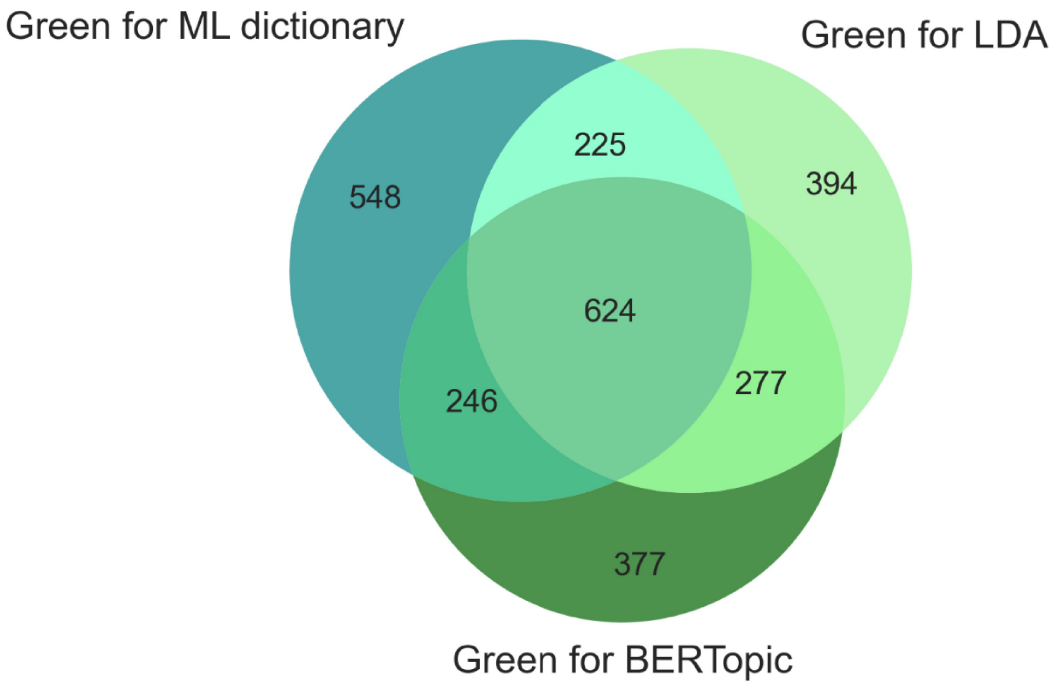


Fig. 2.6 Venn Diagram of the identification of green start-ups with ML dictionary, LDA, and BERTopic

Since all methods were developed without labelled data, we manually check websites to cross-validate the results. We manually examined 100 random websites classified as green: those identified by the intersection of all three methods, those identified by the intersection of two methods, and those identified exclusively by each individual method (details in Appendix A). To estimate the false negative rate, we conducted a similar examination on the start-ups that are never identified as green, which we define as the “baseline” for brevity. We use those estimations to calculate each method’s precision (true positives/ true+false positives), accuracy (correct predictions/total sample size), recall (true positives/true positives+false negatives), and F1-score (harmonic mean of precision and recall) (Table 2). LDA and BERTopic have high precision, respectively 93% and 91%, while the ML dictionary only fares 83%. The relatively low recall of each method, around 0.5 for the three approaches, could be improved by increasing the 15% threshold of green start-ups, but the gain is at the expense of precision.

We also calculate the performance indicators of different combinations of the three methods , considering the startups identified as green by: 1) any of the three algorithms (their union);2) at least two methods; and 3) all of the three methods (their intersection). As expected, identifying green start-ups with the union of the three methods leads to the highest recall (0.79) but low precision, 0.86, lower than BERTopic or LDA taken individually. The startups identified by at least two methods lead to a high precision (0.96) and a good recall of 0.55, both of which are higher than any of the individual methods. Finally, all startups identified

by the intersection of the three methods that we checked manually were true positives, leading to the maximum level of precision, but it is a relatively small sample, leading to the lowest recall (0.27) and F1-score (0.43). We show those results in Table 2.2

The comparison of method combinations further highlights the precision–recall trade-off. Using the union of all methods maximizes recall (0.79), making it suitable when minimizing missed cases is a priority, albeit at the cost of lower precision. In contrast, the intersection of all three methods achieves perfect precision in our sample but captures only a limited subset of start-ups (recall 0.27), making it appropriate for high-confidence classifications. The "at least two methods" approach provides a balanced compromise, combining high precision (0.96) with moderate recall (0.55), and thus represents a practical operational choice. Ultimately, the preferred approach depends on the research objective: stricter criteria are better suited for constructing highly reliable samples, whereas more inclusive definitions are preferable when maximizing coverage for downstream analysis.

Table 2.2 Precision, accuracy, recall, and F1-score of the different approaches

Indicator	ML dictionary	LDA	BERTopic	Startups identified by any of the 3 methods (union)	Startups identified by at least 2 methods	Startups identified by all 3 methods (intersection)
Precision	0.83	0.93	0.90	0.86	0.96	1.00
Recall	0.50	0.50	0.50	0.79	0.55	0.27
F1-Score	0.62	0.67	0.65	0.83	0.70	0.43
Accuracy	0.84	0.87	0.86	0.89	0.85	0.83

Methods with high precision (LDA, BERTopic, or the intersection of two or three methods) should be privileged by researchers for whom the identification of false positives (non-green start-ups tagged as green start-ups) is costly, for example in studies that focus on specific dynamics within groups of green firms where the erroneous presence of non-green businesses could introduce significant imprecisions. Conversely, a high recall should be privileged by researchers who are concerned about missing green start-ups (that is, a high rate of false negatives). All three methods have equal recall, taken individually, but naturally, this value can be increased by taking the union of more methodologies. The F1-score somewhat balances precision and recall and thus can be seen as an interesting indicator when both false positives and false negatives are important. Finally, accuracy can be misleading in unbalanced datasets such as ours, in which there are more grey or brown than green startups, but we reported it for the sake of completeness and comparability with other research.

2.5.6 Differences in green firms' characteristics

Afterwards, we leverage data from AIDA to compare the characteristics of the green start-ups for each method with the baseline in Table 2.3. We compare their age, total assets, number of employees, relative growth, and the word and page length of their websites. We inspect significant differences with an ANOVA and a series of Tukey's HSD tests (Appendix A). The start-ups identified as green by all three methods are older and have more webpages than the baseline, suggesting that start-ups with a green engagement communicate about it to capitalize on it (Amores-Salvadó et al., 2014; Pancić et al., 2023). The green start-ups typically have higher growth than the baseline start-ups, in line with recent research suggesting that going green is a competitive advantage (Colombelli et al., 2020, 2021; Hörisch, 2015; Mrkajic et al., 2019; Wöhler & Haase, 2022). The start-ups identified as green by the dictionary approach, in contrast to the two other green samples, have fewer words per webpage. Upon

manual inspection, the ML dictionary approach appears unfit for classifying short texts: many of the start-ups with a few words per webpage identified as green only by the ML dictionary approach are false positives. We also conducted a violin plot analysis (Appendix A), confirming very similar distributions for the green samples, with differences from the baseline for the variables with significant differences.

Table 2.3 Comparison of statistics of the start-ups identified as green by LDA, BERTopic, and the ML dictionary approach

Statistic	Baseline (non-green)	LDA	BERTopic	ML dictionary
Age	6.17	6.62***	6.54***	6.55***
	(2.93)	(3.12)	(3.08)	(3.14)
Assets	155.77	173.68**	168.59	170.22
	(205.02)	(224.44)	(222.38)	(222.61)
Number of employees	2.20	2.60	1.75	2.13
	(7.07)	(11.64)	(4.83)	(7.54)
Relative growth	1.96	2.67***	2.55**	2.44
	(6.04)	(8.27)	(8.10)	(7.78)
Number of webpages	4.65	6.79***	6.53***	7.14***
	(4.06)	(4.89)	(4.89)	(4.76)
Number of words	1912.70	2305.12***	2390.28***	2189.69
	(5470.01)	(2834.63)	(3098.40)	(2609.75)
Number of words per webpage	393.60	341.14	373.45	290.39***
	(1489.46)	(415.99)	(506.00)	(235.47)

*Notes: ***, **, and * indicates a significant difference from the baseline at the 0.01, 0.05 and 0.1 level for Tuckeys' HSD test. The baseline are the start-ups that are not identified as green by any of the techniques.*

To examine whether the start-ups identified differ depending on activities related to specific localization, for instance, start-ups working on marine-related topics that would be particularly well captured by one of our methods, we plotted the regional distribution of the baseline start-ups and the green ones as identified by our three approaches (Appendix A). All maps were very similar, with a higher concentration of baseline start-ups in the region around Roma and Milano. We also explored the time distribution of the creation of the start-ups, which is similar for every green sample. In recent years, a smaller share of the created start-ups is green.

As discussed in the Data section, to qualify as an innovative start-up in Italy, new ventures must either possess a patent or license; or allocate over 15% of their revenues to R&D activities; or ensure that at least one-third of their employees hold a PhD or that two-thirds have a master's degree. We explored this data (Appendix A) to show that baseline start-ups meet the high R&D criteria oftener than green start-ups, and the high human capital and license or patent ownership more rarely.

Our dataset also includes information on the prevalence of women, young people or foreigners (Appendix A). Here we note a new difference between the NLP methods: fewer green start-ups identified by the LDA approach have a prevalence of women or a prevalence of foreigners than the baseline. It suggests that the start-ups identified by LDA are in male-dominated sectors, such as energy or electronics (Colombelli et al., 2024).

Next, we inspect the distribution of activities carried out by the baseline and the green start-ups identified by the ML dictionary approach, LDA, and BERTopic. To categorize activities, we use the ATECO 2-digit code, the classification of economic activities used by ISTAT, the national institute of statistics of Italy. We plot the 10 most frequent ATECO codes for each sample and a category 'others' for readability (Fig. 2.1). The two most represented categories in the baseline are ATECO number 62 (*software production, informatic counselling and related activities*) and 72 (*scientific research and development*). While those two categories remain

the most important ones for green start-ups, the software production ATECO is much less prevalent.

To measure the diversity of ATECO between the start-ups, we compute their Shannon Index[7], which quantifies the diversity of categories in a population. The Shannon Index of the baseline ATECO is 2.57, while the one of the ML dictionary green start-ups is 2.70, LDA 2.80 and BERTopic 2.83. This suggests that the green start-ups as identified by the three approaches are more diverse than non-green start-ups, which is explained by the lower share of green start-ups with the ATECO producing software than the baseline.

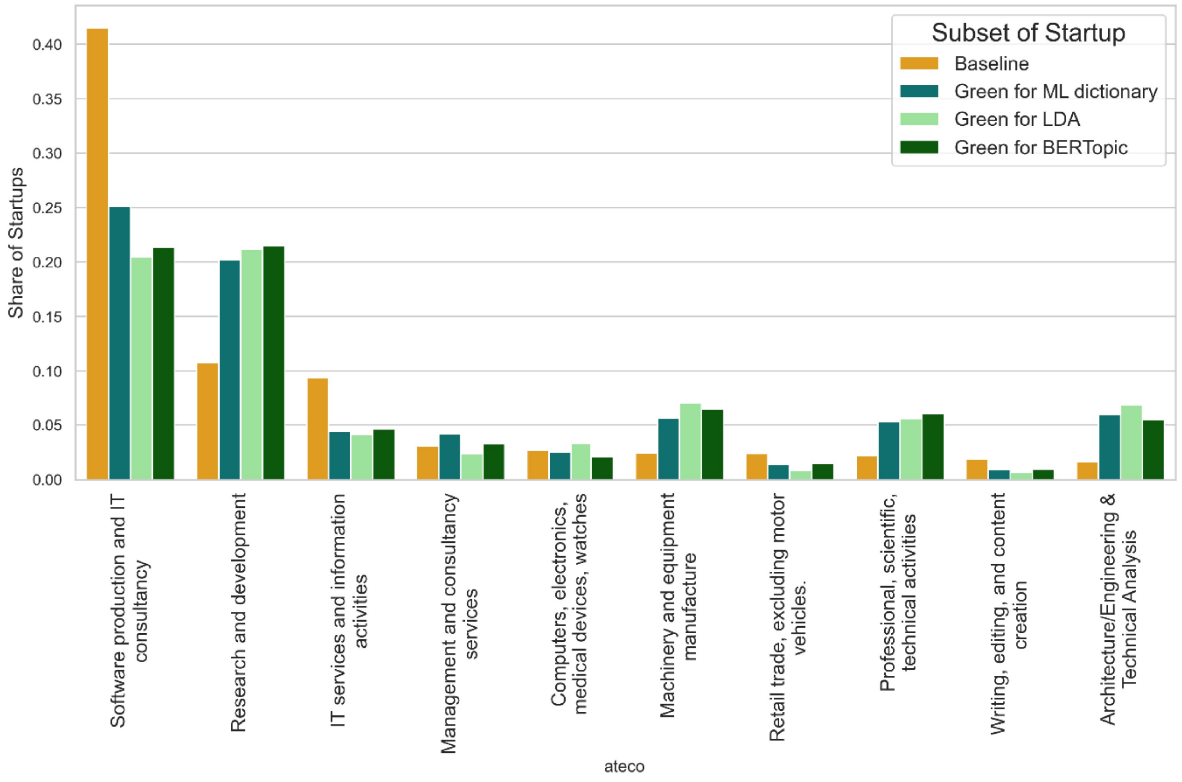


Fig. 2.7 Comparison of areas of activities (measured by ATECO codes) and Shanon index between the baseline and green start-ups identified by the ML dictionary approach, LDA and BERTopic

Only the top 10 ATECO are represented, covering three-fourths of the start-ups.

Our three algorithms allow us to identify which green topics have led to the identification of a start-up as green. We use this data to quantify the startups' contributions to each green topic in Fig. 2.8. To compute startups' contributions, we sum up the number of startup contributions

[7] Computed as $H' = -\sum_{i=1}^{S} p_i \ln(p_i)$, with *S* the number of categories and p_i the proportion of the total sample belonging to the *i*-th category.

to that topic, with each contribution weighted by the number of green topics the startup is involved in. This ensures that startups contributing to multiple green topics do not disproportionately influence the contribution of a single topic.

LDA identifies startup contributions solely in the most prominent green topics, whereas BERTopic detects startups' contributions across all green topics. This is in line with other research, which has identified BERTopic as a topic model able to uncover new, unexpected, and more specific insights from corpora of text. The specific dictionaries of the ML dictionary approach also allow it to identify green contributions in every green topic. The biggest difference between the two is between Air Quality and Pollution Management, for which BERTopic identifies many contributions, and Sustainable Resource Management and Efficiency, for which ML dictionary finds many contributions. To understand whether the strengths or weaknesses of the algorithm cause those differences, we opened 50 websites of each to estimate the rate of true positives. The startups identified with the tag Air Quality and Pollution Management by BERTopic, where we see a peak in green start-ups for this method, reach a rate of true positive of 0.94, and the startups with the ML dictionary tag Sustainable Resource Management and Efficiency, again showing a higher number of green start-ups identified compared to the other methods, reach a rate of true positives of 0.84. Those values are close to the rate of true positives of their respective methods, reassuring us that they are not anomalously capturing some extra false positives in these topics, and showing that each methods identify green startups with different types of engagement. Interestingly, the startups identified with the tag Sustainable Resource Management and Efficiency only by BERTopic have a low share of true positives (0.48), showing that although BERTopic is overall better than the ML dictionary approach, the latter still retains advantages for specific topics.

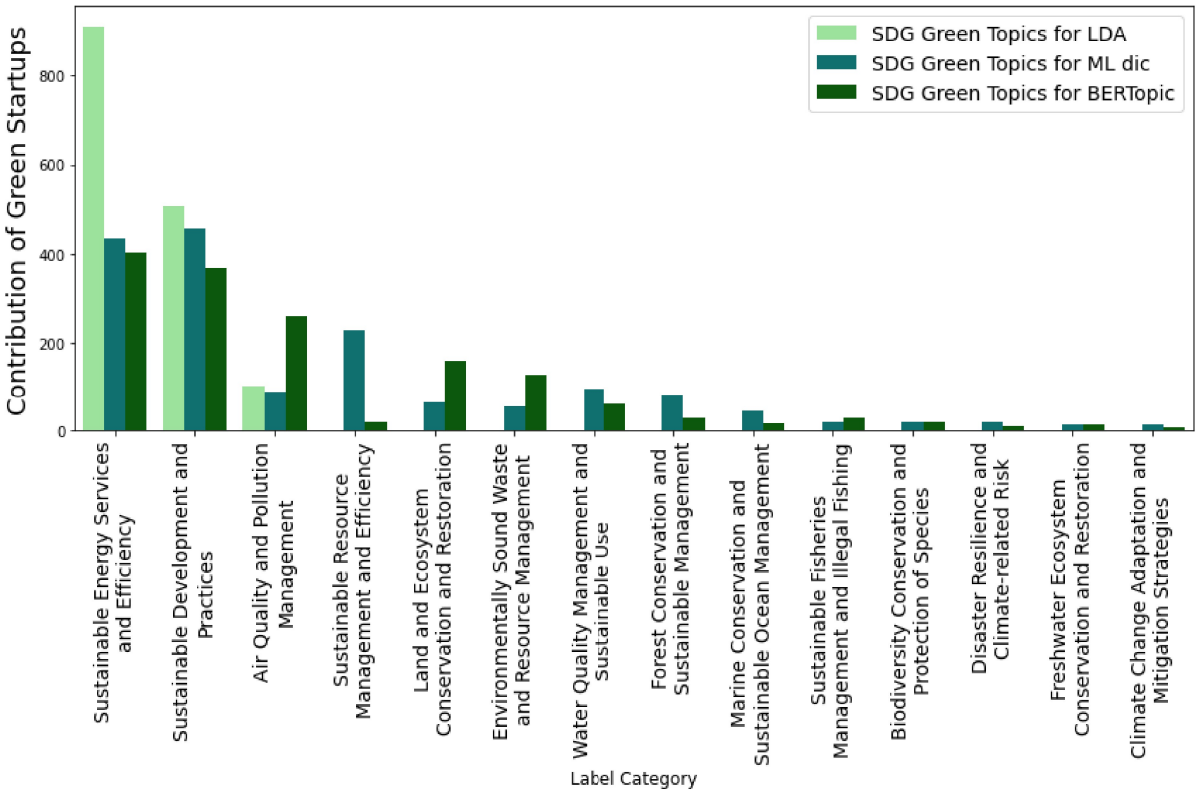


Fig. 2.8 Share of start-ups identified with each green topic

Each method identifies multiple topics in each start-up. The bars sum to more than one, as some start-ups have multiple labels

Finally, Table 2.4 summarizes the key technical characteristics of each implementation. There is a trade-off between number of preprocessing steps and the number of parameters needed to use the approach: while the dictionary approach has the most preprocessing steps, it has only two parameters; LDA offers a middle-ground, with less preprocessing steps but more parameters than ML dictionary, among which the number of topic to identify; BERTopic only needs to represent the text as embeddings, but one should decide the model it uses with the associated parameters at every step of the algorithm.

Table 2.4 Comparison of ML dictionary, LDA, and BERTopic approaches for the identification of green start-ups

Approach	ML Dictionary	LDA	BERTopic
Language restrictions	No restriction: monolingual or multilingual embedding models can be used	Restricted to monolingual corpora	No restriction: monolingual or multilingual embedding models can be used
Preprocessing steps	- Lemmatization - Stopwords removal - Building n-grams - Embedding	- Lemmatization - Stopwords removal - Bag-of-words	- Embedding
Hyperparameters to be set by the researcher	- Embedding model - Green topic cosine similarity threshold	- N. of topics - Alpha and Beta, respectively, for document-topic and topic-word sparsity - Topic-word matrix initialization and number of iterations - Green topic cosine similarity threshold	- Embedding, reduction, clustering, and topic representation model - We used 5 models with a total of 8 hyperparameters
Stochasticity	No	Low	High
Precision	83%	93%	91%
Weaknesses in the identification of green start-ups	Perform poorly for short texts	Identifies only well-represented categories	Identifies only one topic webpage per iteration[8]

[8] This is mitigated by the iteration process and by the fact that websites are composed of multiple webpages.

2.6 APPLICATION

To illustrate how the use of different methods can yield distinct insights into the role of green entrepreneurship, we provide a simple application comparing the start-ups' distribution within different green topics with (i) public spending and (ii) national performance in those areas, based on National Recovery and Resilience Plan (NRRP) investments and the Sustainable Development Report (Sachs et al., 2024) rating of SDG achievements in Italy. This exercise is not intended as a formal validation of the methodologies, but rather as an exploration of their differing implications in a real policy context. The NRRP is a public instrument to modernize Italy through sustainable infrastructure, digitalization, and social inclusivity, driving economic growth and the green transition. It consists of more than EUR 190 billion of investments from Next Generation Europe funds (Italia Domani, 2025). It can be seen as a proxy for public efforts towards specific green targets. ISTAT provides an association between each NRRP measure and a specific SDG indicator. Moreover, Sachs et al 2024 publishes data on countries' achievements for 125 of the 232 SDG indicators, assigning a color ranking to each indicator based on national progress toward the goals - red for major challenges remaining to achieve the indicators' goal; orange, for significant challenges; yellow, for some challenges; green, for achieved goals.

To associate NRRP investments and national performance for each green topic, we map their SDG indicators to our green topics (see Table 4 in the Appendix A). Unfortunately, progress on three green topics is not measured by any indicator in the Sustainable Development Report, but for all other 12 topics, we have corresponding public investments and performance ratings, which we can then compare to the distribution of green startups (Fig. 2.9).

Green Topic	BERTopic Green SU contribution	LDA Green SU contribution	ML Dic Green SU contribution	NRRP Investment (M€)	Italian Performance
Sustainable Energy Services and Efficiency	402	909	433	85'532	
Air Quality and Pollution Management	259	103	90	43'638	
Sustainable Development and Practices	366	507	457	38'219	
Water Quality Management and Sustainable Use	62	0	94	9'328	
Disaster Resilience and Climate-related Risk Management	8	0	19	5'800	
Environmentally Sound Waste and Resource Management	128	0	55	2'100	
Biodiversity Conservation and Protection of Species	20	0	18	1'039	
Sustainable Resource Management and Efficiency	20	0	229	650	
Marine Conservation and Sustainable Ocean Management	18	0	46	400	
Climate Change Adaptation and Mitigation Strategies	5	0	12	30	
Land and Ecosystem Conservation and Restoration	158	0	65	0	
Sustainable Fisheries Management and Illegal Fishing Prevention	29	0	18	0	
Forest Conservation and Sustainable Management	29	0	81	0	
Freshwater Ecosystem Conservation and Restoration	13	0	14	0	
Total	**1517**	**1519**	**1631**	**186'736**	

First quartile
Second quartile
Third quartile
Fourth quartile
Major challenges
Significant challenges
Some challenges
Goal reached

Fig. 2.9 Comparison of green topic Italian performance and entrepreneurship

The color of the entrepreneurship count and NRRP investments represent the quantile, and the colors of the national performance are based on Sachs et al., (2024).

From Fig. 2.9, it is evident that none of the green topics has fully achieved its goals. However, some of the areas with a few "goal achieved" indicators, namely '*Sustainable Energy Services and Efficiency*', '*Air Quality and Pollution Management*' and '*Sustainable Development and Practices*', are also topics with the highest public spending and the greatest number of green startups, for all three NPL methodologies. Therefore, any analysis that focuses on these areas can rely on either ML dictionary, LDA, or BERTopic and identify groups of green firms of comparable size.

For the topic of '*Water Quality Management and Sustainable Use*', an area that still receives significant public investments, the difference between the three methods becomes evident: LDA does not identify any green startups, while ML Dictionary and BERTopic still capture some entrepreneurial activity. Similarly, there are topics with relatively little public spending, but still quite a few green start-ups identified by these two methods (e.g., '*Environmentally Sound Waste and Resource Management'*, '*Sustainable Resource Management and Efficiency*', '*Land and Ecosystem Conservation and Restoration*'). These discrepancies could be crucial for mapping the green areas in which innovative private sector activities are developing despite a lack of public funding. Any in-depth analysis of the SDG performance in these areas that considers both the role of public finance and private local entrepreneurship should be aware of the different pictures that emerge, depending on how green start-ups are tagged.

Overall, this application highlights the differences between machine-learning NLP methods. For example, relying on LDA allows to capture a larger number of energy start-ups operating

in one of the areas that has received most public funding, so it could be appropriate to understand renewable energy transition dynamics; however, LDA suggests that no private activity exists on other “minor” green topics, and this could be problematic for a wider definition of sustainability at the intersection of private and public activity. In contrast, BERTopic and the ML dictionary reveal a significant number of start-ups operating in a broader range of green topics that have received little or no financing, and this could help understand how some progress has been made in these areas. All three methodologies are fundamentally grounded in the SDGs, which we have used to define the green topics; as a result, they can effectively inform any study of the political aspects of SDG implementation. Ultimately, though, the choice of a specific methodology captures different types of environmentally sustainable startups and can thus lead to different results.

2.7 CONCLUSION

This research presents and compares three machine-learning natural language processing models, ML dictionary, LDA, and BERTopic, for identifying green start-ups based on the text of their websites. For each method, our NLP-based labelling process allows us to find which websites display information related to the SDGs green topics. LDA reaches the highest precision (0.93), followed by BERTopic (0.90) and ML dictionary (0.83). This underlines the strength of topic models in extracting complex insights from texts without any predefined set of keywords or manual intervention. The ML dictionary, LDA, and BERTopic approaches identify green start-ups with similar characteristics, but with differences in their green engagement: LDA identified mostly start-ups in renewable energy and energy efficiency, and broad sustainable development practices, while the ML dictionary and especially BERTopic identified start-ups within 14 different topics. In this research, LDA appears appropriate for identifying well-represented topics in a corpus, reaching higher precision than BERTopic. Conversely, in line with recent research, BERTopic appears stronger at identifying topics that are more complex and less represented in the corpus. Although the ML dictionary approach has lower precision and struggles with short texts, it identifies more start-ups in specific topics.

Applying the three methodologies to assess green start-up activity in Italy relative to national SDG performance and NRRP investments highlights the strengths and limitations of each approach. The LDA method suggests that Italian green start-ups primarily concentrate in areas that also attract the largest NRRP investment. In contrast, BERTopic and ML dictionary present a more nuanced view, acknowledging the prominence of these few topics while also identifying additional areas that capture entrepreneurial interest, albeit with fewer start-ups.

Leveraging the textual content of websites allows for classifying start-ups based on public data, but is not free from limitations: this content is created by the company, which might not always reflect their actual impact and making it vulnerable to greenwashing (Montgomery et al., 2023).

Althoug this work introduce three fully automated pipeline for identifying green startups without labelled data and achieving good results, further work would be needed to optimize the parameter choices.

The methods presented in this paper could be extended to other fields. Future research could explore the use of the ML dictionary approach, LDA, and BERTopic to identify start-ups operating in specific industries.

2.8 DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work the author(s) used ChatGPT and Grammarly in order to improve the writing of the paper. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

2.9 REFERENCES

Abdesselam, R., Kedjar, M., & Renou-Maissant, P. (2024). What are the drivers of eco-innovation? Empirical evidence from French start-ups. Technological Forecasting and Social Change, 198, 122953. https://doi.org/10.1016/j.techfore.2023.122953

Amores-Salvadó, J., Castro, G. M., & Navas-López, J. E. (2014). Green corporate image: Moderating the connection between environmental product innovation and firm performance. Journal of Cleaner Production, 83, 356–365. https://doi.org/10.1016/j.jclepro.2014.07.059

Anand, A., Argade, P., Barkemeyer, R., & Salignac, F. (2021). Trends and patterns in sustainable entrepreneurship research: A bibliometric review and research agenda. Journal of Business Venturing, 36(3), 106092. https://doi.org/10.1016/j.jbusvent.2021.106092

Arseneau, D. M., Drexler, A., & Osada, M. (2022). Central Bank Communication about Climate Change.

https://www.federalreserve.gov/econres/feds/central-bank-communication-about-climate-change.htm

Bendig, D., Kleine-Stegemann, L., Schulz, C., & Eckardt, D. (2022). The effect of green startup investments on incumbents' green innovation output. Journal of Cleaner Production, 376, 134316. https://doi.org/10.1016/j.jclepro.2022.134316

Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate Confusion: The Divergence of ESG Ratings*. Review of Finance, 26(6), 1315–1344. https://doi.org/10.1093/rof/rfac033

Blei, D. M., Y. Ng, A., & I. Jordan, M. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.

Borčin, M., & Jose, J. M. (2024). Optimizing BERTopic: Analysis and Reproducibility Study of Parameter Influences on Topic Modeling. In N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, & I. Ounis (Eds.), Advances in Information Retrieval (pp. 147–160). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-56066-8_14

Bottai, C., Crosato, L., Domenech, J., Guerzoni, M., & Liberati, C. (2024). Scraping innovativeness from corporate websites: Empirical evidence on Italian manufacturing SMEs. Technological Forecasting and Social Change, 207, 123597. https://doi.org/10.1016/j.techfore.2024.123597

Bruderl, J., & Schussler, R. (1990). Organizational Mortality: The Liabilities of Newness and Adolescence. Administrative Science Quarterly, 35(3), 530–547. https://doi.org/10.2307/2393316

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., & Blei, D. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. Advances in Neural Information Processing Systems, 22. https://proceedings.neurips.cc/paper_files/paper/2009/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html

Chapman, G., & Hottenrott, H. (2022). Green start-ups and the role of founder personality. Journal of Business Venturing Insights, 17, e00316. https://doi.org/10.1016/j.jbvi.2022.e00316

Cojoianu, T. F., Clark, G. L., Hoepner, A. G. F., Veneri, P., & Wójcik, D. (2020). Entrepreneurs for a low carbon world: How environmental knowledge and policy shape the creation and financing of green start-ups. Research Policy, 49(6), 103988. https://doi.org/10.1016/j.respol.2020.103988

Cojoianu, T. F., Hoepner, A. G. F., Hu, X., Ramadan, M., Veneri, P., & Wójcik, D. (2024). Are cities venturing green? A global analysis of the impact of green entrepreneurship on city air pollution. Small Business Economics, 62(2), 523–540. https://doi.org/10.1007/s11187-023-00764-4

Coll-Martínez, E., Kedjar, M., & Renou-Maissant, P. (2022). (Green) Knowledge spillovers and regional environmental support: Do they matter for the entry of new green tech-based firms? The Annals of Regional Science, 69(1), 119–161. https://doi.org/10.1007/s00168-022-01111-3

Colombelli, A., D'Ambrosio, A., & Ravetti, C. (2024). Women in innovative start-ups and regional inclusiveness: 'Green' and socially-responsible companies. Regional Studies, 1–14. https://doi.org/10.1080/00343404.2024.2340999

Colombelli, A., Ghisetti, C., & Quatraro, F. (2020). Green technologies and firms' market value: A micro-econometric analysis of European firms. Industrial and Corporate Change, 29(3), 855–875. https://doi.org/10.1093/icc/dtaa003

Colombelli, A., Krafft, J., & Quatraro, F. (2021). Firms' growth, green gazelles and eco-innovation: Evidence from a sample of European firms. Small Business Economics, 56(4), 1721–1738.

Colombelli, A., & Quatraro, F. (2019). Green start-ups and local knowledge spillovers from clean and dirty technologies. Small Business Economics, 52(4), 773–792. https://doi.org/10.1007/s11187-017-9934-y

Corradini, C. (2019). Location determinants of green technological entry: Evidence from European regions. Small Business Economics, 52(4), 845–858. https://doi.org/10.1007/s11187-017-9938-7

Dean, T. J., & McMullen, J. S. (2007). Toward a theory of sustainable entrepreneurship: Reducing environmental degradation through entrepreneurial action. Journal of Business Venturing, 22(1), 50–76. https://doi.org/10.1016/j.jbusvent.2005.09.003

Dong, S., Gong, H., & Liu, T. (2022). Environmental technology spillovers and green start-up emergence: The moderating role of patent commercialization policy and patent enforcement. Environmental Science and Pollution Research, 29(46), 70070–70083. https://doi.org/10.1007/s11356-022-20791-0

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Frontiers in Sociology, 7. Scopus. https://doi.org/10.3389/fsoc.2022.886498

Fonseca, S. A., & Chiappetta Jabbour, C. J. (2012). Assessment of business incubators' green performance: A framework and its application to Brazilian cases. Technovation, 32(2), 122–132. https://doi.org/10.1016/j.technovation.2011.10.006

Gast, J., Gundolf, K., & Cesinger, B. (2017). Doing business in a green way: A systematic review of the ecological sustainability entrepreneurship

literature and future research directions. Journal of Cleaner Production, 147, 44–56. https://doi.org/10.1016/j.jclepro.2017.01.065

Gebhardt, L., & Bachmann, N. (2023). Entrepreneurial contributions to sustainability transitions—A longitudinal study of their representation and enactment through topic modeling and thematic analysis. Journal of Cleaner Production, 420, 138255. https://doi.org/10.1016/j.jclepro.2023.138255

Gidron, B., Bar, K., Finger Keren, M., Gafni, D., Hodara, Y., Krasnopolskaya, I., & Mannor, A. (2023). The Impact Tech Startup: Initial Findings on a New, SDG-Focused Organizational Category. Sustainability, 15(16), Article 16. https://doi.org/10.3390/su151612419

Giudici, G., Guerini, M., & Rossi-Lamastra, C. (2019). The creation of cleantech startups at the local level: The role of knowledge availability and environmental awareness. Small Business Economics, 52(4), 815–830. https://doi.org/10.1007/s11187-017-9936-9

Gorovaia, N., & Makrominas, M. (2024). Identifying greenwashing in corporate-social responsibility reports using natural-language processing. European Financial Management, n/a(n/a). https://doi.org/10.1111/eufm.12509

Griliches, Z. (1979). Issues in Assessing the Contribution of Research and Development to Productivity Growth. The Bell Journal of Economics, 10(1), 92–116. https://doi.org/10.2307/3003321

Griliches, Z. (1990). Patent statistics as economic indicators: A survey. NBER.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure (arXiv:2203.05794). arXiv. https://doi.org/10.48550/arXiv.2203.05794

Guo, L., Vargo, C., Pan, Z., Ding, W., & Ishwar, P. (2016). Big Social Data Analytics in Journalism and Mass Communication: Comparing Dictionary-Based Text Analysis and Unsupervised Topic Modeling. Journalism & Mass Communication Quarterly, 93. https://doi.org/10.1177/1077699016639231

Hajikhani, A., & Suominen, A. (2022). Mapping the sustainable development goals (SDGs) in science, technology and innovation: Application of machine learning in SDG-oriented artefact detection. Scientometrics, 127(11), 6661–6693. https://doi.org/10.1007/s11192-022-04358-x

Halberstadt, J., Schwab, A.-K., & Kraus, S. (2024). Cleaning the window of opportunity: Towards a typology of sustainability entrepreneurs. Journal of Business Research, 171, 114386. https://doi.org/10.1016/j.jbusres.2023.114386

Hörisch, J. (2015). Crowdfunding for environmental ventures: An empirical analysis of the influence of environmental orientation on the success of crowdfunding initiatives. Journal of Cleaner Production, 107, 636–645. https://doi.org/10.1016/j.jclepro.2015.05.046

Horne, J., & Fichter, K. (2022). Growing for sustainability: Enablers for the growth of impact startups – A conceptual framework, taxonomy, and systematic literature review. Journal of Cleaner Production, 349, 131163. https://doi.org/10.1016/j.jclepro.2022.131163

Horne, J., Recker, M., Michelfelder, I., Jay, J., & Kratzer, J. (2020). Exploring entrepreneurship related to the sustainable development goals—Mapping new venture activities with semi-automated content analysis. Journal of Cleaner Production, 242, 118052. https://doi.org/10.1016/j.jclepro.2019.118052

Hossnofsky, V., Herold, P.-P., Schlichte, F., & Junge, S. (2025). Green signals of new ventures: Investigating the impact of environmental orientation on funding and the moderating role of lead venture capitalists. Small Business Economics. https://doi.org/10.1007/s11187-025-01072-9

IPSF Taxonomy Working Group. (2021). Common ground taxonomy – Climate change mitigation instruction report.

Italia Domani. (2025). The NRRP's contribution to the 2030 Agenda implementation. https://www.italiadomani.gov.it:443/content/sogei-

ng/it/en/strumenti/il-contributo-del-pnrr-all-attuazione-dell-agenda-2030.html

Jorzik, P., Antonio, J. L., Kanbach, D. K., Kallmuenzer, A., & Kraus, S. (2024). Sowing the seeds for sustainability: A business model innovation perspective on artificial intelligence in green technology startups. Technological Forecasting and Social Change, 208, 123653. https://doi.org/10.1016/j.techfore.2024.123653

Kirkwood, J., & Walton, S. (2010). What motivates ecopreneurs to start businesses? International Journal of Entrepreneurial Behavior & Research, 16(3), 204–228. https://doi.org/10.1108/13552551011042799

Li, K., Mai, F., Shen, R., & Yan, X. (2021). Measuring Corporate Culture Using Machine Learning. The Review of Financial Studies, 34(7), 3265–3315. https://doi.org/10.1093/rfs/hhaa079

Lundmark, E., & Audretsch, D. B. (2024). Revisiting the Entrepreneurial Society framework: A constructive critique from a climate change perspective. International Small Business Journal, 42(4), 396–415. https://doi.org/10.1177/02662426231199417

Mansouri, S., & Momtaz, P. P. (2022). Financing sustainable entrepreneurship: ESG measurement, valuation, and performance.

Journal of Business Venturing, 37(6), 106258. https://doi.org/10.1016/j.jbusvent.2022.106258

Mazaheri, M., Bonnin Roca, J., Markus, A., Tur, E. M., & Walrave, B. (2024). Maturity assessment of green patent clusters: Methodological implications. Technological Forecasting and Social Change, 209, 123813. https://doi.org/10.1016/j.techfore.2024.123813

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. The Journal of Open Source Software, 2(11), 205. https://doi.org/10.21105/joss.00205

McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (arXiv:1802.03426). arXiv. https://doi.org/10.48550/arXiv.1802.03426

Mio, C., Panfilo, S., & Blundo, B. (2020). Sustainable development goals and the strategic role of business: A systematic literature review. Business Strategy and the Environment, 29(8), 3220–3245. https://doi.org/10.1002/bse.2568

Misztal, A., & Kowalska, M. (2023). Factors of green entrepreneurship in selected emerging markets in the European Union. Environment, Development and Sustainability. https://doi.org/10.1007/s10668-023-03811-y

Montgomery, A. W., Lyon, T. P., & Barg, J. (2023). No End in Sight? A Greenwash Review and Research Agenda. Organization & Environment, 10860266231168905. https://doi.org/10.1177/10860266231168905

Mrkajic, B., Murtinu, S., & Scalera, V. G. (2019). Is green the new gold? Venture capital and green entrepreneurship. Small Business Economics, 52(4), 929–950. https://doi.org/10.1007/s11187-017-9943-x

Neumann, T. (2021). Does it pay for new firms to be green? An empirical analysis of when and how different greening strategies affect the performance of new firms. Journal of Cleaner Production, 317, 128403. https://doi.org/10.1016/j.jclepro.2021.128403

Nigel, H., & Britt, H. (1966). A hierarchical grouping routine, IBM 360/65 FORTRAN IV program.

Nikolaou, I. E., Tasopoulou, K., & Tsagarakis, K. (2018). A Typology of Green Entrepreneurs Based on Institutional and Resource-based Views. The Journal of Entrepreneurship, 27(1), 111–132. https://doi.org/10.1177/0971355717738601

OECD. (2022, June 13). Policies to Support Green Entrepreneurship. OECD. https://www.oecd.org/en/publications/policies-to-support-green-entrepreneurship_e92b1946-en.html

Olawumi, T. O., & Chan, D. W. M. (2018). A scientometric review of global research on sustainability and sustainable development. Journal of Cleaner Production, 183, 231–250. https://doi.org/10.1016/j.jclepro.2018.02.162

Pacheco, D. F., Dean, T. J., & Payne, D. S. (2010). Escaping the green prison: Entrepreneurship and the creation of opportunities for sustainable development. Journal of Business Venturing, 25(5), 464–480. https://doi.org/10.1016/j.jbusvent.2009.07.006

Pancić, M., Serdarušić, H., & Ćućić, D. (2023). Green Marketing and Repurchase Intention: Stewardship of Green Advertisement, Brand Awareness, Brand Equity, Green Innovativeness, and Brand Innovativeness. Sustainability, 15(16), Article 16. https://doi.org/10.3390/su151612534

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (arXiv:1908.10084). arXiv. https://doi.org/10.48550/arXiv.1908.10084

Rizvanović, B., Zutshi, A., Grilo, A., & Nodehi, T. (2023). Linking the potentials of extended digital marketing impact and start-up growth: Developing a macro-dynamic framework of start-up growth drivers supported by digital marketing. Technological Forecasting and Social Change, 186, 122128. https://doi.org/10.1016/j.techfore.2022.122128

Rizzitello, E., Piazza, M., & Perrone, G. (2025). Unlocking green startup investments: How environmental policy pressures drive Venture Capital funding decisions. Technological Forecasting and Social Change, 217, 124158. https://doi.org/10.1016/j.techfore.2025.124158

Rodríguez-García, M., Guijarro-García, M., & Carrilero-Castillo, A. (2019). An Overview of Ecopreneurship, Eco-Innovation, and the Ecological Sector. Sustainability, 11(10), Article 10. https://doi.org/10.3390/su11102909

Sabando-Vera, D., Montalván-Burbano, N., Parrales-Guerrero, K., Yonfá-Medranda, M., & Plaza-Úbeda, J. A. (2025). Growing a greener future: A bibliometric analysis of green innovation in SMEs. Technological Forecasting and Social Change, 212, 123976. https://doi.org/10.1016/j.techfore.2025.123976

Sachs, J. D., Lafortune, G., & Fuller, G. (2024). The SDGs and the UN Summit of the Future. Sustainable Development Report 2024. Dublin: Dublin University Press. https://doi.org/10.25546/108572

Shepherd, D., & Patzelt, H. (2011). The New Field of Sustainable Entrepreneurship: Studying Entrepreneurial Action Linking “What Is to Be Sustained” With “What Is to Be Developed.” Entrepreneurship Theory and Practice, 35, 137–163. https://doi.org/10.1111/j.1540-6520.2010.00426.x

Shi, L., Han, L., Yang, F., & Gao, L. (2019). The Evolution of Sustainable Development Theory: Types, Goals, and Research Prospects. Sustainability, 11(24), Article 24. https://doi.org/10.3390/su11247158

Stinchcombe, A. L. (1965). Social structures and organisation: Vol. Handbook of Organization (James G. March, pp. 142–193). Rand McNally.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis (p. 651). M.I.T. Press.

Tiba, S., Van Rijnsoever, F. J., & Hekkert, M. P. (2021). Sustainability startups and where to find them: Investigating the share of sustainability startups across entrepreneurial ecosystems and the causal drivers of differences. Journal of Cleaner Production, 306, 127054. https://doi.org/10.1016/j.jclepro.2021.127054

Umamaheswaran, S., Dar, V., Sharma, E., & Kurian, J. S. (2023). Mapping Climate Themes From 2008-2021—An Analysis of Business News Using Topic Models. IEEE Access, 11, 26554–26565. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3256530

UN. (2015). Transforming our world: The 2030 Agenda for Sustainable Development | Department of Economic and Social Affairs. https://sdgs.un.org/2030agenda

UN. (2019, April 3). Global Environment Outlook 6. UNEP - UN Environment Programme. http://www.unep.org/resources/global-environment-outlook-6

Van Zanten, J. A., & Van Tulder, R. (2021). Towards nexus-based governance: Defining interactions between economic activities and Sustainable Development Goals (SDGs). International Journal of Sustainable Development & World Ecology, 28(3), 210–226. https://doi.org/10.1080/13504509.2020.1768452

VivaTechnology. (2025). 2025 Edition | Viva Technology. Https://Vivatechnology.Com. https://vivatechnology.com/

Webersinke, N., Kraus, M., Bingler, J., & Leippold, M. (2022). CLIMATEBERT: A Pretrained Language Model for Climate-Related Text. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4229146

Wöhler, J., & Haase, E. (2022). Exploring investment processes between traditional venture capital investors and sustainable start-ups. Journal of Cleaner Production, 377, 134318. https://doi.org/10.1016/j.jclepro.2022.134318

Yun, J., & Geum, Y. (2020). Automated classification of patents: A topic modeling approach. Computers & Industrial Engineering, 147, 106636. https://doi.org/10.1016/j.cie.2020.106636

CHAPTER 3

Knowledge spillovers, green entrepreneurship and the demand for sustainability: Evidence from Italian innovative startups

The following publication is extensively based on this chapter:

Colombelli, A., D'Ambrosio, A., Le Masle, B., Ravetti, C., & Tubiana, M. (2025). Knowledge spillovers, green entrepreneurship and the demand for sustainability: evidence from Italian innovative startups. *The Journal of Technology Transfer*, 1-26.

3 KNOWLEDGE SPILLOVERS, GREEN ENTREPRENEURSHIP AND THE DEMAND FOR SUSTAINABILITY: EVIDENCE FROM ITALIAN INNOVATIVE STARTUPS

3.1 ABSTRACT

Investigating the relationship between knowledge dynamics and the local demand for environmental sustainability can provide new insights into the forces behind the emergence of innovative green entrepreneurship. We explore the interplay between knowledge availability and green innovation in a territory within the Knowledge Spillover Theory of Entrepreneurship (KSTE) and expand the KSTE by integrating demand-side factors. We test how the creation of innovative green ventures correlates with the match between the size and composition of the knowledge stock and local green demand, proxied by measures of pro-environmental behaviours. We identify green startups from the Italian Registry of Innovative Startups with a novel AI-based methodology that performs unsupervised topic modelling on text from companies' websites. This machine-learning algorithm detects companies' alignment with environmental Sustainable Development Goals (SDGs). We then perform a province-level econometric analysis to examine the interaction between the (green) knowledge stocks, demand factors, and the creation of green startups. Our findings confirm that local demand for environmental sustainability is associated with innovative green startup creation and magnifies local knowledge stocks' role. Interestingly, we show that green entrepreneurship is more strongly related to the local knowledge stock's size than its "greenness".

Keywords: knowledge spillovers; innovative startups; green entrepreneurship; environmental sustainability; local green demand

Acknowledgements: We thank Stefano Bianchini for the valuable discussions on AI algorithms for natural language processing, which led to our choosen method to identify green startups, and Francesco Quatraro for providing province-level data on green patent applications. We are grateful to Ekaterina Prytkova and the participants at the Knowledge Dynamics, Industry Evolution, Economic Development (KID) 2024 Conference for comments and feedback on an early version of this work. We thank AI Link and the participants of the workshop "Networking event in honor of Distinguished Professor AI Link", as well as anonymous reviewers and guest editors of the special issue for helpful suggestions that improved earlier versions of this manuscript. All usual disclaimers apply.

3.2 INTRODUCTION

Green innovative startups are ascending in policymakers' priorities because they can bring to the market innovative technologies that simultaneously contribute to economic growth and promote a sustainable and eco-friendly future (Horne & Fichter, 2022). Startups are the primary drivers of "green" market innovation, introducing new ideas and inventions that might transform entire industries (Fichter et al., 2023). For this reason, green startups with a highly innovative profile are central to achieving the green transition (Abdesselam et al., 2024).

Classic entrepreneurial frameworks, such as the Knowledge Spillover Theory of Entrepreneurship (KSTE), have adeptly explained the emergence of new ventures through the lens of knowledge dissemination (Acs et al., 2009; Audretsch et al., 2020; Audretsch & Keilbach, 2007). In recent years, they have offered crucial insights into green entrepreneurship (Cojoianu et al., 2020; Colombelli & Quatraro, 2019; Giudici et al., 2019). Startup success heavily depends on local knowledge factors, such as the innovative ecosystem in which they operate. Such a dependency may be especially pronounced for green startups, which are often characterised by more significant entrepreneurial risks and uncertainties due to the complexity and novelty of sustainable innovations (Cillo et al., 2019; Del Río et al., 2016). Consequently, the literature suggests that robust local ecosystems providing the necessary knowledge, infrastructure, and market conditions are especially needed by green startups (York et al., 2018).

Recent KSTE literature has screened the channels that turn knowledge spillovers into actual innovations, acknowledging that there are no direct or automatic mechanisms (Audretsch et al., 2021; Audretsch & Belitski, 2020). Only a few entrepreneurial efforts are highly innovative and lead to the formation of high-impact, high-growth companies that drive aggregate economic growth (Colombelli et al., 2016; Shane, 2009). Innovativeness is particularly relevant in the context of green startups because it enables them to address complex environmental challenges with new solutions, and positions them as key players in local economic development.

The innovation literature has comparatively neglected demand-side considerations (Di Stefano et al., 2012; Kalcheva et al., 2018), except for studies about demand-side policies like public procurement within and outside the green domain (Caravella & Crespi, 2021; Costantini et al., 2015). Yet, the uncertainty about the returns of investment in green products and technologies is a crucial hindrance to translating green entrepreneurship into successful innovative businesses, and plays a pivotal role in specific industries (Caiazza et al., 2020; Horbach et al., 2013; Rennings, 2000). Hence, understanding the role of demand factors along with local knowledge spillovers in green entrepreneurship is essential to fully characterise this phenomenon and inform policymakers aiming to foster sustainable economic development.

We aim to expand the scope of the KSTE by incorporating both supply and demand side factors, to better understand how technological and market opportunities translate into entrepreneurial emergence and regional innovation. While the KSTE primarily focuses on supply-side factors (the availability of knowledge and supportive, innovative ecosystems), we expand the framework to include demand-side factors that we deem particularly relevant for the emergence of green startups. We also explore the interaction between local demand for sustainable innovations and the size and composition of the local knowledge stock for green firm formation. In doing so, we contribute to the rising debate on the interplay between entrepreneurial ecosystems and knowledge spillovers, pointing to the benefits of integrating demand as a contextual factor for entrepreneurial emergence (Morris et al., 2024; Qian, 2018).

An essential contribution of our paper lies in our original approach to identifying green startups. Existing methods are either based on patent or industrial classifications, focus on specific kinds of firms and technologies, or rely on manual inspections (Coll-Martínez et al., 2022; Colombelli & Quatraro, 2019; Corradini, 2019). We develop a flexible, scalable and time-efficient model that

employs advanced machine learning algorithms to analyse the text of startups' websites and identify topics, which are then classified as "green" by referencing the targets of environment-related Sustainable Development Goals (SDGs). We apply our AI-based identification method to the Italian Registry of Innovative Startups, which contains data on the population of Italy's most innovative new companies, to capture new ventures with a high potential for growth and positive local impacts.

We employ this measure to empirically test our arguments at the province level, studying how local factors contribute to the formation of green innovative startups. We account for the role of the knowledge base through the total patent application stock and the proportion of such patent applications that are green. To approximate the demand for environmental sustainability (from now on referred to as "green demand"), we rely on a proxy that combines several indicators of environmentally sensitive behaviours using Principal Component Analysis (PCA). We are aware that this behaviour-based approach may not capture every aspect of green demand—e.g., market conditions like prices and availability of green products, policy incentives, access to services and infrastructure, or other contextual factors (Brouhle & Khanna, 2012; Joshi & Rahman, 2015). Yet, it offers a theoretically grounded and empirically relevant measure. Our justification for this proxy stems from established theories of green consumption values, environmental attitudes, and pro-environmental behaviours. Research underscores that green demand is rooted in environmental values, social norms, and behavioural patterns, all of which correlate strongly with a greater willingness to pay for sustainable products and with environmentally-conscious purchasing decisions (Farrow et al., 2017; Haws et al., 2014; Lee et al., 2023; Nguyen et al., 2016; Paul et al., 2016). By integrating these behavioural dimensions, we expect to capture the core drivers of green demand and align closely with this field's theoretical and empirical evidence.

Our results indicate that the local green demand is pivotal for the emergence of innovative green startups in Italian provinces. In line with the KSTE, we confirm that the local stock of knowledge is highly relevant for green startup formation and further identify positive interaction effects of local knowledge stocks and high levels of green demand in the area. These results suggest synergistic effects of supply and demand factors and indicate that green demand boosts local knowledge spillovers. Interestingly, we find that the spillovers that matter for green entrepreneurship arise from larger, rather than "greener", knowledge stocks. These results suggest that the overall territorial innovation capacity matters more than specific specialisation patterns for green entrepreneurship, which aligns with the green innovation literature that points to a critical role of local recombinant capabilities.

The article is organised as follows. Section 2 presents the theoretical background and hypotheses tested. Section 3 describes the data, variables, and empirical model. Section 4 discusses the main results and section 5 concludes.

3.3 THEORETICAL BACKGROUND AND HYPOTHESIS DEVELOPMENT

To date, the KSTE is one of the key theoretical frameworks accounting for the creation of new innovative enterprises in a region. The KSTE predicates that large incumbent firms and organisations generate knowledge spillovers that constitute a unique business opportunity for

potential entrepreneurs aiming to establish new innovative enterprises within a territory (Audretsch et al., 2015; Audretsch & Keilbach, 2007). However, available knowledge in a location is necessary but insufficient for innovation because knowledge transfer is not frictionless. Transforming knowledge stocks into economically valuable innovations requires enabling conditions at the local level, such as supporting institutions, knowledge intermediaries, regulatory frameworks and dedicated financial markets. The absence of these conditions creates a barrier or a cost – the Knowledge Filter (KF) – that hinders the transformation of broad knowledge into economic knowledge à la Arrow. To appropriate knowledge spillovers and transform them into commercially useful knowledge, potential entrepreneurs must (be set in the position to) penetrate the KF (Acs et al., 1994, 2004).

The KSTE is a story of knowledge opportunities: incumbents, mostly unwillingly, supply knowledge resources and thus create local spillovers and opportunities for innovative entrepreneurship. The abundance of knowledge spillovers directly increases the propensity to start a new venture in the same territory, as aspiring entrepreneurs attempt to exploit the opportunities that incumbent firms do not appropriate. New enterprises can develop more agile and flexible solutions to leverage the existing knowledge stock, possibly at a lower cost than large incumbents (Antonelli, 2013). Knowledge opportunities allow latent entrepreneurs and innovators to emerge, provided that they can recognise the chance, absorb the knowledge, and take risks to initialise and operationalise a new venture (Caiazza et al., 2020; Qian & Jung, 2017).

As such, the KSTE qualifies as a supply-side theory of innovation. It highlights the importance of the provision and accessibility of knowledge for entrepreneurship to emerge. It serves as a *liaison* between individual-level analyses on entrepreneurship determinants and studies about contextual-level factors. Specifically, it bridges the literature on individual motivations to take entrepreneurial risks and the research on the resources, institutions and culture that give the entrepreneurial effort a chance to start and perform (Audretsch et al., 2017; Autio et al., 2014). To further understand the role of knowledge spillovers in creating specific entrepreneurial opportunities, we test the KSTE on an interesting subset of nascent firms: innovative green startups.

Green innovative entrepreneurship can be challenging because it faces a double externality issue, taking place both at the invention and diffusion stages: inventors cannot entirely appropriate the returns from knowledge generation (the starting point of the KSTE), and the technology they produce positively affects a public good, the environment. Indeed, public support is particularly relevant for green knowledge creation and diffusion (Orsatti, 2024). Inventors can partially protect their potential returns from green innovations like any other innovation. However, the demand for their innovative green product or service may be weak because the innovation does not focus on generating private value but contributes to the maintenance of a public good. Hence, green innovations often result in knowledge creation and environmental benefits that are non-excludable and subject to market imperfections (Cohen & Winn, 2007). Therefore, prospective entrepreneurs face uncertainty about expected returns when starting a new “green” venture (Horbach et al., 2013; Rennings, 2000).

From a KSTE perspective, the relationship between knowledge availability/accessibility and innovation through firm creation should also hold true for green startups (Cojoianu et al., 2020; Colombelli & Quatraro, 2018; Giudici et al., 2019). The reliance on knowledge spillovers – a necessary but insufficient condition for innovative startup creation – could even be more relevant for green startups (Horbach et al., 2013). Indeed, green innovations show different, more demanding characteristics than non-green ones, as they are typically more complex and novel[9] (Barbieri et al., 2020). Thus, the spillovers from the existing stock of knowledge should be especially relevant for the creation of new green startups. Measuring the stock of knowledge with local patent applications, we test the following hypothesis:

H1: The larger the size of the stock of knowledge — proxied by patent applications — the higher the number of green innovative startups in a province

Recent literature advocates that a specific type of knowledge spillovers, namely those arising from green knowledge, matters especially for green entrepreneurship (Cojoianu et al., 2020; Colombelli & Quatraro, 2019; Giudici et al., 2019; Vedula et al., 2019). The argument is that, while general knowledge stocks offer a broad basis for new enterprise creation, the presence of green knowledge –R&D, innovative ecosystems, incubators and research centres focused on renewable technologies, circular economy, decarbonisation solutions, and so on – can boost green entrepreneurship in a targeted way. Green technologies and knowledge often display homophily[10] and path dependence and, therefore, might need a consolidated and specific set of skills, data, equipment and resources that can contribute to new green knowledge, green technologies and, ultimately, green ventures (Nomaler & Verspagen, 2019). The link between a subset of the knowledge stock and the equivalent type of entrepreneurial activity has been posited before, beyond the context of green startups: Bonaccorsi et al. (2013), for instance, identify an association between the specialisation of the knowledge created locally and the kind of innovative startups funded, and Colombelli et al., (2023) found evidence of this link for artificial intelligence startups.

Drawing from the diversity and innovation literature, we argue that the knowledge stock can be viewed as the sum of two components: a size and a composition component (Alesina et al., 2016; Antonelli et al., 2017, 2022). Both are expected to foster knowledge creation, whereas some ambiguity remains regarding the correlation between knowledge composition and performance (Antonelli et al., 2022). We focus on the green composition of the overall knowledge stock to assess the role of environmentally friendly knowledge. Whereas H1 explores the direct correlation

[9] One could question whether green knowledge is intrinsically different from non-green knowledge or if such characteristics derive from green knowledge being the turbulent frontier of the knowledge space. Fusillo (2023) marks a step in this direction, but further research on the topic is warranted.

[10] High homophily in the context of knowledge creation means that green patents mostly cite green patents in prior art (Nomaler & Verspagen, 2019), but it does not necessarily correlate with measurement of knowledge recombination such as complexity, variety and novelty, mostly because green knowledge often is general and broad. Indeed, new knowledge can recombine creatively and unexpectedly knowledge from various domains within the same area, especially if the latter is general and broad.

with the size component, our second hypothesis tackles whether the stock's composition correlates with innovative green startup emergence.

H2: The larger the share of green knowledge — proxied by green patent applications — in the composition of the knowledge stock, the higher the number of green innovative startups in a province.

A crucial contribution of the present paper is to suggest that the KSTE, as a supply-side innovation theory, would benefit from incorporating demand-side economic forces. Theoretically and empirically, the interplay between supply and demand in innovation dynamics is a neglected research avenue (Kalcheva et al., 2018). The debate about whether demand or supply factors are the most influential in driving innovations peaked in the Seventies and was followed by a substantial body of research on the supply side or technology-push perspective, claiming that the provision of knowledge through R&D investments and science is a key engine of the gales of innovation and growth. On the contrary, the research on demand for innovations focused more on pull effects, directing "the trajectory towards the right economic venues" (Di Stefano et al., 2012, pp. 1291). Recent works have started to look at the interplay between demand and supply: for instance, Kalcheva et al., (2018) show that, in the medical industry, shocks to the demand for new products significantly impact the innovation process, particularly when they take place in territories with a high supply of knowledge through a multiplicity of channels.

The interplay between supply and demand side forces is especially topical in the context of eco-innovations and green entrepreneurship. The literature is currently exploring a wide range of push and pull factors, but no universal consensus has been reached, and plenty of gaps remain (Abdesselam et al., 2024; Del Río et al., 2016; Rehfeld et al., 2007). One of the most significant hurdles to the successful diffusion of green products and services is achieving a competitive price relative to non-green alternatives (Rehfeld et al., 2007). Hence, demand-pull forces should play a crucial role in determining the decision to invest and enter the market, even though they might have a milder impact on investment intensity (Kesidou & Demirel, 2012).

Hence, we want to test the theoretical argument that the demand side is particularly relevant in the emergence of innovative new firms in the green domain. Since green innovations tackle a wide range of environmental externalities – from pollution to waste, greenhouse gas emissions to biodiversity loss, resource depletion to ecosystem degradation – they partially suffer from the same market inefficiencies of public goods. Consequently, they struggle to find a clear local demand. The first step for entrepreneurial emergence is the recognition of a positive pay-off leading to a high expected value of the innovation investment (Caiazza et al., 2020). Hence, in the presence of a strong demand signal, the willingness to start a risky green enterprise is expected to increase as the uncertainty over future pay-offs lowers.

Moreover, corporate stakeholder theory asserts that consumer pressure is a crucial incentive for green innovation investments because it signals consumers' willingness to pay and enacts organisational learning, particularly for product innovation (Wong, 2013; Zhang & Zhu, 2019).

Indeed, product innovation, being more explorative and uncertain than process innovation, requires considering consumers' preferences and coordinating with them. The clearer and more intense the consumers' demand, the stronger the opportunities to profit from innovative capabilities, provided enough and immediate firm responsiveness (Roome & Wijen, 2006). At the same time, green innovations can address the unmet demand for sustainability of local stakeholders and feed the creation and exploitation of new market niches (Schaltegger & Wagner, 2011). Overall, we expect a positive link between local demand and innovative green entrepreneurship. As discussed in detail in the data section, we operationalize local green demand with an index combining different local green behaviours to proxy for sustainability-oriented demand in a territory, leading us to our third hypothesis.

H3: The higher the green demand — proxied by local green behaviours — the higher the number of green innovative startups in a province.

Building on the previous arguments, we posit that demand forces affect and shape the KF. The above reasoning regarding knowledge spillovers and demand should hold for every kind of innovative entrepreneurship. Still, we propose that the interplay between demand and knowledge spillover availability in a territory matters particularly for green startup emergence. Green innovation often arises from recombining multiple local knowledge bases, not necessarily belonging to green-specific domains (Fusillo, 2023; Orsatti et al., 2024; Quatraro & Scandura, 2019), which makes green knowledge particularly susceptible to knowledge spillovers. Moreover, green innovations' intrinsic market and knowledge characteristics constitute a disincentive to embark on a risky endeavour, strengthening the KF due to high uncertainty and low expected short-term returns (York & Venkataraman, 2010). However, if demand for environmental sustainability manifests, it constitutes a counterbalancing force and softens the KF because it increases short-term expected returns on the innovation investment. In other words, the green arena is one of the best contexts to provide new evidence of the supply-demand nexus in entrepreneurial research.

The literature has extensively studied the impact of non-market forces, such as pro-environmental institutions, on local absorptive capacity in the form of opportunity recognition (Vedula et al., 2019). Some studies have shown that environmental sensitivity informs policymakers' actions towards supporting regulations (Cojoianu et al., 2020; Coll-Martínez et al., 2022; Giudici et al., 2019), a crucial antecedent of green innovation (Santoalha & Boschma, 2021). The entrepreneurial ecosystems literature stresses the importance of contextual factors (such as institutions, culture, infrastructure, demand, networks, finance, talent, and support services) in determining startup emergence (Spigel, 2017; Stam, 2015). However, this literature does not directly address the moderating role of contextual factors on the appropriation and commercialisation of knowledge spillovers. Importantly, even though previous studies have not enlisted local demand among the key contextual factors that should matter for the KSTE (Qian, 2018), we argue that the current

form of the KSTE as a supply-side theory of entrepreneurial emergence would greatly benefit from an approach that also accounts for demand-side forces.

Overall, a natural implication of the KSTE is that the size of the stock of knowledge positively interacts with green demand. If the knowledge stock captures the availability of marketable knowledge opportunities, and green demand reduces the risk of converting such opportunities into entrepreneurial ventures, a combination of the two is expected to increase green entrepreneurship.

H4a: Green demand — proxied by local green behaviours — positively moderates the relationship between the size of the stock of knowledge and green startups.

A related question concerns whether the specific composition of the knowledge base, in this case, the "greenness" of such knowledge stock, will play a particular role in interaction with local demand forces. If innovative green entrepreneurship is particularly reliant on green knowledge, a context marked by a "greener" stock of knowledge will produce "greener" spillovers and give rise to opportunities with a more marked green content, making demand more effective in weakening the KF.

H4b: Green demand — proxied by local green behaviours — positively moderates the relationship between the share of green knowledge and green startups.

3.4 EMPIRICAL FRAMEWORK

3.4.1 Data

3.4.1.1 Innovative startups

For our analysis, we rely on the population of italian innovative startups detailed in Chapter 2. These startups represent a subset of companies with high innovation capacity and considerable growth potential, and therefore, they constitute a suitable sample to study applications of the KSTE in terms of novel green value creation. From this initial sample, we used AIDA, an Amadeus-Bureau Van Dijk database, to access relevant information about the startups, such as firm characteristics and location. Innovative startups in our sample concentrate in Italy's most dynamic economic centres, such as in Lombardy and Lazio's regions, particularly in large provinces like Milan and Rome. Fig. 3.1 maps their distribution in Italian territory.

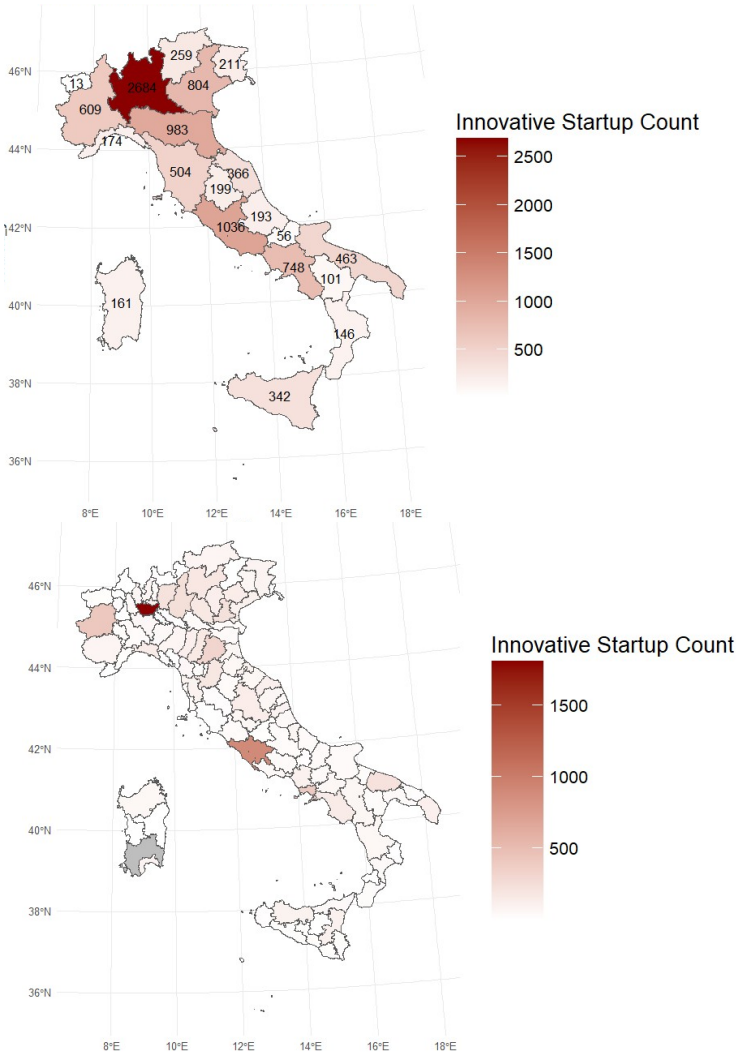


Fig. 3.1 Count of innovative startups at the regional and province level as of May 2023

Source: authors' own elaboration on data from the Italian Chamber of Commerce.

As a first step, we must identify green startups in the sample. However, this task is not straightforward: there is no consensus on the definition of a "green startup" in the literature (see for example the discussions in Colombelli et al., 2024; Gast et al., 2017; Purvis et al., 2019), nor a single empirical approach to tag them (Colombelli & Quatraro, 2019, Tiba et al., 2021, Jha & Pande, 2024, Gidron et al., 2023). We develop a new approach that leverages the recent developments of Natural Language Processing, using BERTopic (Grootendorst, 2022) to extract topics from the startups' website, and in parallel, we identify topics related to SDG environmental targets as green (see Fig. 3.2 for a summary of the procedure, and Appendix B for a detailed overview of the whole algorithm). Ultimately, we tag 1'529 Italian innovative startups as "green".

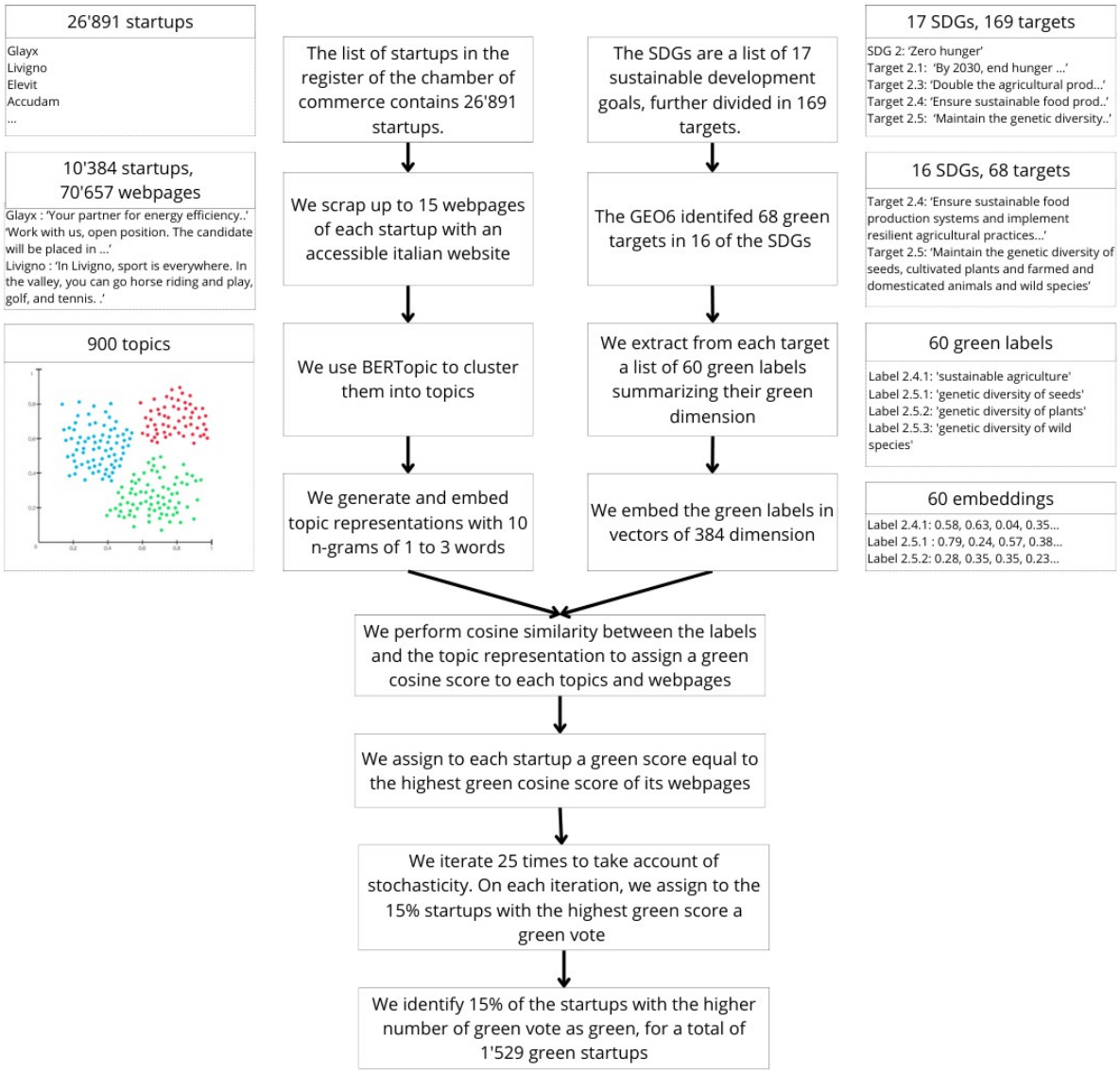


Fig. 3.2 Summary of the green startup identification process through BERTopic, comparing it with targets from green sustainable development goals (SDGs)

Fig. 3.3 shows the share of our green innovative startups per region and province. The distribution shows a high concentration of green entrepreneurial activity in specific provinces. Still, it should be noted that these are also locations with few total startups (Oristano in Sardinia has nine innovative startups, among which four are green, and Ragusa in Sicily has 20, among which nine are green). Milan and Rome, while having the highest number of total innovative startups, as shown previously in Figure 1, have a relatively low rate of green startups, with 13% and 12%, respectively.

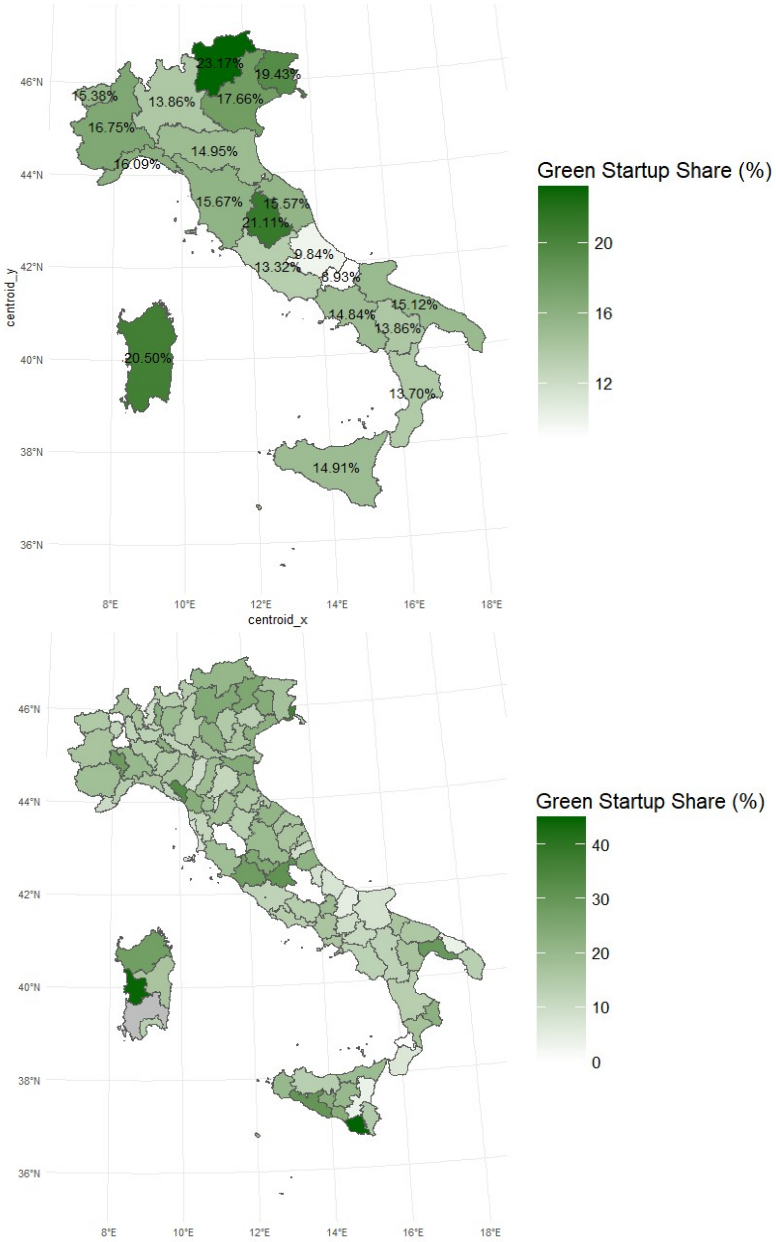


Fig. 3.3 Share of green startups identified by BERTopic with green SDG targets in Italian regions and provinces

3.4.1.2 Operationalising the KSTE: knowledge stock and its "greenness"

The presence of a vibrant knowledge ecosystem qualified by the availability of useful knowledge has long been operationalised with the stock of patent applications at the territorial level whenever the focus of the analysis was explicitly on innovation outcomes (Acs et al., 2009; Colombelli & Quatraro, 2018, 2019; Mokyr, 2011). As discussed in the previous section, we depart from prior literature on green entrepreneurship (Cojoianu et al., 2020; Colombelli & Quatraro, 2019; Giudici

et al., 2019; Vedula et al., 2019) and measure two characteristics of such knowledge stock: its size and composition (Alesina et al., 2016; Antonelli et al., 2017, 2022).

To measure the size component of the knowledge stock at the highest possible level of granularity (NUTS3), we measure it from the output side based on patent applications (Acs et al., 2009; Hall et al., 1986). Patent data capture codified and legally recognized knowledge outputs, which are particularly suitable for approximating the stock of economically relevant knowledge within regions. Unlike broader innovation indicators, patents reflect knowledge that is sufficiently novel, non-obvious, and industrially applicable to pass examination procedures. A key advantage of patent data is their precise geographic attribution to inventors, enabling consistent measurement at highly disaggregated territorial levels such as NUTS3 regions. Alternative indicators (e.g., R&D expenditure, innovation surveys) are typically unavailable or unreliable at this level of spatial resolution. Patent statistics offer standardized definitions and harmonized collection procedures across countries and over time, which ensures comparability in longitudinal and cross-regional analyses—an essential requirement for this study. We are aware of the limitations of patents as a measure of innovation – not all patents represent innovation, many innovations are not patented, the propensity to patent differs across industries, and patents downplay innovation in service industries. Yet, patents remain the most widespread used proxy for measuring innovation due to their availability at different levels of aggregation and over time. Despite their limitations, patent-based measures have been extensively validated and widely adopted in empirical studies of regional knowledge production and innovation (e.g., Acs et al., 2009; Hall et al., 1986), supporting their continued use as a reliable proxy.

To measure the composition of the total knowledge stock, we assess the extent to which the knowledge stock is green. We identify green patents based on their Cooperative Patent Classification (CPC) code as all patents falling in the Y02 class and, for each year and province, compute the share of the patent stock that is attributable to green patents as

$$\mathrm{GPSH}_{i,t} = \frac{\mathrm{GREEN\_KSTOCK}_{i,t}}{\mathrm{KSTOCK}_{i,t}}$$

Conditional on a given patent stock, this simple share captures the extent to which local knowledge is more or less intensive in green knowledge and, thus, is a reasonable proxy of the green composition of local knowledge. In a robustness check, we study the robustness of our results to the inclusion of $GREEN\_KSTOCK$ directly in the estimation, in line with existing literature (Audretsch et al., 2020; Cojoianu et al., 2020; Colombelli & Quatraro, 2019; Giudici et al., 2019; Vedula et al., 2019). Fig. 3.4 shows the patent stock in Italian provinces for 2022, the last year of the analysis.

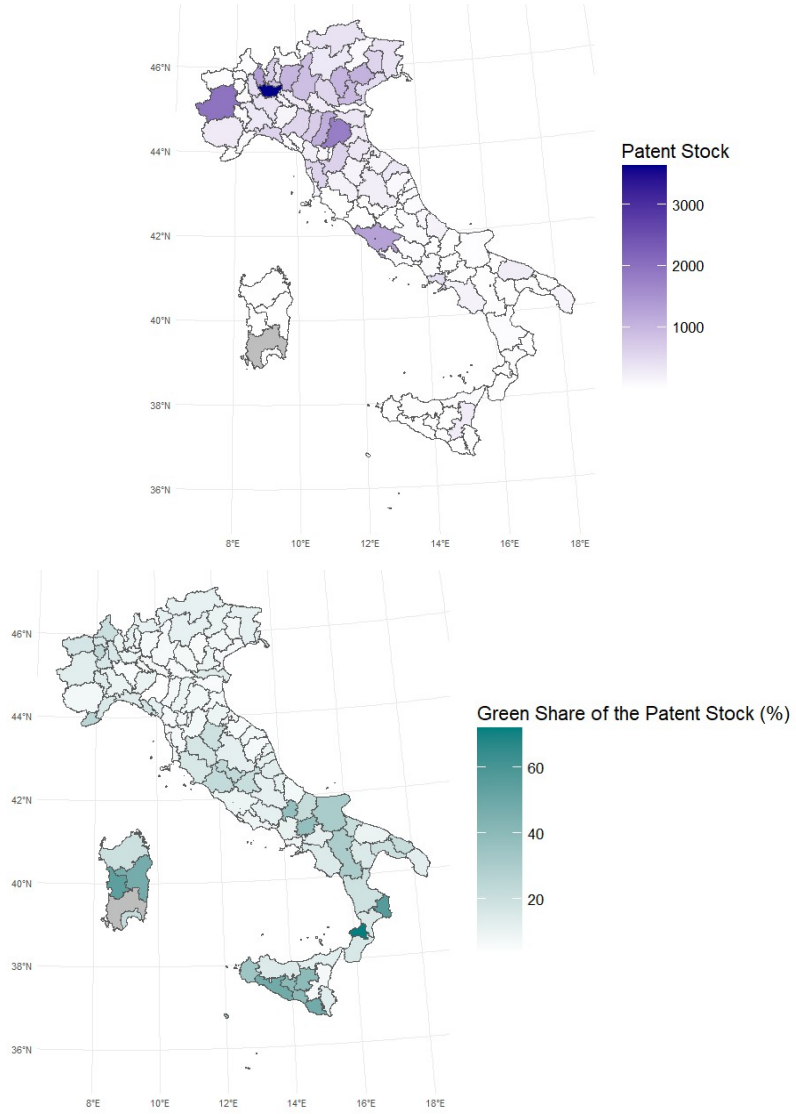


Fig. 3.4 Patent stocks and green share of the patent stocks in Italian provinces as of 2022

3.4.1.3 Green demand index from local green behaviours

As argued in Section 2, we want to test whether the demand side plays a role in the emergence of innovative new firms in the green domain. Operationalising the demand for environmental sustainability is challenging due to the complex interplay of environmental sensitivity, actual behaviours oriented to protect the environment, and disposable income. Previous studies have confirmed the role of self-assessed environmental sensitivity in green startup creation (Giudici et al., 2019). However, subjectively stated environmental sensitivity does not necessarily correlate with environmentally sustainable behaviours. The expected success of green startups critically relies on environmentally sustainable behaviours and not simply self-declared green sensitivity: these startups can only expect to thrive if their target consumers enact pro-environmental behaviours and choose their products and services against alternative, more polluting ones. Thus,

we conceptualize green demand as the revealed preference of urban populations for environmentally sustainable goods, services, and practices, as expressed through mobility choices, waste behaviours, and environmental pressure indicators. Drawing on these considerations, our operationalisation focuses on proxies for environmentally sustainable behaviours at the highest possible granularity level, which is once again the NUTS3 level.

The Italian National Institute of Statistics (ISTAT) data on the quality of urban environment appear ideal to this end.[11] A broad range of data relating to the quality of urban mobility, waste collection, and pollution are collected yearly for province capitals. The set of variables has evolved over time, but information about the core variables is available yearly for about one decade. In choosing the variables that we use to construct our measure for green demand, we must balance the need to ensure a sufficiently long time coverage with the aim to represent the variety of green behaviours with a reasonably broad set of variables. The result of this process is the following selection of variables, all available yearly from 2012 to 2022: availability of bike-sharing services; availability of car-sharing services; bike lane density; demand for local public transportation services; days exceeding the acceptable limit of PM10 concentration in urban air; share of recycled/sorted waste on total waste; motorisation rate (i.e., motor-vehicles per inhabitant). While some indicators (e.g., availability of bike- and car-sharing services) reflect supply conditions, they are included insofar as they emerge in response to local demand for sustainable mobility options, and therefore indirectly capture demand intensity. Although this measure does not directly capture environmental attitudes and may partly reflect infrastructural constraints, it provides a consistent proxy for revealed green demand at the urban level.

To obtain a proxy for green demand, we combine these variables in a single index through a principal component analysis (PCA, Afifi et al., 2019). PCA identifies linear combinations of the variables that explain the largest component of the covariance. Applying this methodology to our variables, we obtain seven principal components, the first of which explains 37.4% of the overall variance. The relative factor loadings are reported in the Appendix C (Table C1).

[11] See, e.g., https://www.istat.it/it/archivio/291918

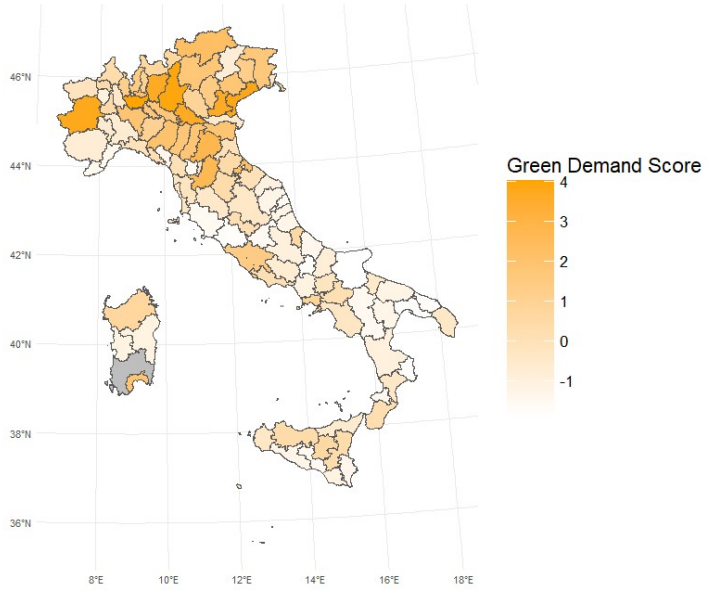


Fig. 3.5 Green demand score in Italian provinces as of 2021

In addition to the key variables of interest and consistent with previous literature, we also include a set of controls in the empirical analysis that capture local socio-economic conditions: firm density, value added, the share of manufacturing, the share of graduates, and the unemployment rate. Our variables' description, summary statistics and correlation matrix are reported in the Appendix C (Tables C2, C3, and C4, respectively).

3.4.2 Empirical model

Our dependent variable's discrete and non-negative nature motivates the adoption of count data models (Cameron & Miller, 2015; Hausman et al., 1984) to estimate our relationship of interest. Our dependant variable display remarkable overdispersion. The negative binomial model offers several advantages in this context. First, it relaxes the equidispersion assumption of the Poisson model by introducing an additional parameter that allows the variance to exceed the mean. Second, it can be interpreted as a Poisson model with unobserved heterogeneity, thereby accounting for latent differences across observational units. As a result, it provides a better fit to overdispersed count data and yields more reliable standard errors, avoiding the downward bias and overstated significance typically associated with Poisson estimates under overdispersion. At the same time, the negative binomial model preserves the interpretability of count models, allowing coefficients to be expressed as incidence rate ratios. Thus, we expect the negative binomial estimator to be more efficient than the Poisson estimator (Greene, 2003). Our baseline model is:

$$GREEN\_STARTUPS_{i,t}=exp(\alpha + \beta_1 KSTOCK_{i,t-1} + \beta_2 GPSH_{i,t-1} + \beta_3 GREENDEMAND_{i,t-1} + Z\gamma + \rho_r\delta + \psi_t\vartheta + \epsilon_{i,t})$$

where Z represents a vector of control variables, ρ_r is a vector of NUTS2-level regional dummies, ψ_t a vector of time dummies, $\epsilon_{i,t}$ is a random error component, and $\alpha,\beta_1,\beta_2,\beta_3,\gamma,\delta,\vartheta$ are parameters or parameter vectors to be estimated. We lag all our regressors by one year to account for the time necessary for knowledge creation to translate into knowledge diffusion and appropriation. While the inclusion of both temporal and geographical dimensions allows us to control for unobserved heterogeneity across regions and over time, it does not fully address potential endogeneity concerns. Accordingly, we refrain from attributing a causal interpretation to our estimates and instead interpret them as strong associations (Bellemare et al., 2017).

The sign and significance of the marginal effects of *KSTOCK* and *GPSH* provide evidence addressing Hypothesis H1 and H2, respectively. The sign and significance of the marginal effects of *GREENDEMAND* test hypothesis H3.

Hypotheses H4a and H4b concern interaction effects. In the presence of non-zero interaction effects, the model equation becomes:

$$\text{GREEN\_STARTUPS}_{i,t}=\exp(\alpha+\beta_1 KSTOCK_{i,t-1}+\beta_2 GPSH_{i,t-1}+\beta_3 GREENDEMAND_{i,t-1}+\beta_4 KSTOCK_{i,t-1}\times GREENDEMAND_{i,t-1}+\beta_5 GPSH_{i,t-1}\times GREENDEMAND_{i,t-1}+Z\gamma+\rho_r\delta+\psi_t\vartheta+\epsilon_{i,t})$$

The econometric literature on non-linear models specifies that the coefficient of the interaction term is not equal to the marginal effect of the interaction (Ai & Norton, 2003; Mize, 2019). Although the problem is less severe for exponential models like the Negative Binomial and Poisson, we follow Mize (2019) in analysing the presence of moderating effects with a graphical inspection of how marginal effects change for different levels of another variable of interest. Finally, in a set of robustness checks, we show the robustness of our results to alternative count model estimators (Poisson and Zero-Inflated Negative Binomial).

3.5 RESULTS

In Table 1, we test our hypotheses about green startup formation. In line with the KSTE and hypothesis H1, the results consistently indicate that a higher stock of knowledge is significantly associated with higher firm formation (column 1)[12]. However, when we exploit the availability of data on both the size (total stock of patent applications, *KSTOCK*) and the composition (share of green stock over total knowledge stock, *GPSH*) of the knowledge stock (from column 2 onward), we obtain rather unexpected insights. Our results do not indicate that the green component of the knowledge stock has a specific effect on green entrepreneurship, contradicting hypothesis H2. In

[12] We test and confirm that stylized facts about knowledge spillovers apply to total startups, too (results available upon request).

other words, we observe a significant correlation between green startup emergence and the *size* of the knowledge stock, but we do not witness a statistically meaningful relationship with its *composition*. Altogether, results about H1 and H2 indicate that green startup emergence requires a vibrant local knowledge ecosystem rather than a specialised knowledge trajectory.[13]

Turning to hypothesis H3, the evidence in column 4 of Table 3.1 confirms the expectation of a positive and statistically significant association between the green demand index *GREENDEMAND* and green entrepreneurship. In line with our theoretical reasoning, more pro-environmental behaviours in a territory go hand in hand with greener entrepreneurship. Column 5 addresses hypotheses 4a and 4b by introducing the interaction between green demand and knowledge stocks. Coherently with the above results, the size component of the knowledge stock turns out positive and significant, while the green composition component is not statistically significant. For the sake of completeness, in column 6 of Table 3.1, we also consider the results of a model that excludes knowledge stocks. The estimates confirm the significance of green demand for green entrepreneurship.

Table 3.1 Green firm formation

	(1)	(2)	(3)	(4)	(5)	(6)
KSTOCK	0.744***	0.801***	0.584***	0.487***	0.450***	
	(0.115)	(0.116)	(0.106)	(0.091)	(0.078)	
GPSH		2.246	0.934	0.743	-2.113	
		(1.426)	(1.211)	(1.174)	(1.644)	
PC_VADDED			0.977	0.618	0.238	4.433***
			(1.368)	(1.137)	(1.140)	(1.077)
GRADSH			0.132	0.105	0.076	0.191
			(0.139)	(0.134)	(0.144)	(0.123)
MANUF_SH			-3.638	-3.612	-3.938*	-3.162
			(3.500)	(2.341)	(2.213)	(2.089)
FIRM_DENS			0.006***	0.006***	0.005**	0.008***
			(0.002)	(0.001)	(0.002)	(0.001)
UNEMPL			-0.024	-0.028*	-0.020	-0.023
			(0.016)	(0.016)	(0.014)	(0.019)

[13] We check that the mean variance inflation factors corresponding to the linear version of our specifications are all below the conventional value of 10, reassuring that multicollinearity should not be a major issue for our estimates.

GREENDEMAND				0.257***	-0.608	0.492***
				(0.080)	(0.380)	(0.096)
KSTOCK ´ GREENDEMAND					0.145**	
					(0.061)	
GPSH ´ GREENDEMAND					-0.752	
					(0.947)	
_cons	-5.742***	-6.419***	-15.467	-10.898	-6.281	-50.008***
	(0.901)	(1.008)	(14.647)	(12.122)	(12.269)	(11.907)
/						
lnalpha	-1.156***	-1.214***	-1.756***	-1.907***	-2.320***	-1.397***
	(0.345)	(0.313)	(0.400)	(0.348)	(0.312)	(0.346)
N	952	952	952	952	952	952
Region FE	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES

Negative binomial regression coefficients. Dependent variable: number of green startups. Standard errors clustered at the regional level are in parentheses. All regressors lagged one year. *p<0.10, **p<0.05, ***p<0.01.

Table 3.2 reports the marginal effects corresponding to the key specifications in columns 4 and 5 of Table 3.1 to more immediately convey the magnitude of the estimated effects on the predicted number of new startups. In the case of a negative binomial model, coefficients capture the effect of a unit change in the explanatory variable on the log of the expected counts of the response variable, holding other factors constant. To make the interpretation more straightforward, we thus include the marginal effects – which, in our context, can be interpreted in the standard sense as the effects of a unit increase in the regressor on the predicted number of startups. With regards to the interaction effects, we study the moderating role of green demand on knowledge stock size and composition graphically (Ai & Norton, 2003; Mize, 2019). Fig. 3.6 and Fig. 3.7 show how the marginal effects of the size of the total knowledge stock and its green composition change for different levels of green demand.

Table 3.2 Green firm formation (marginal effect)

	(1)	(2)
KSTOCK	0.663***	0.781***
	(0.129)	(0.147)
GPSH	1.011	-3.745
	(1.598)	(3.114)

GREENDEMAND	0.349***	0.435***
	(0.110)	(0.146)
PC_VADDED	0.841	0.327
	(1.547)	(1.568)
GRADSH	0.143	0.105
	(0.182)	(0.197)
MANUF_SH	-4.911	-5.410*
	(3.206)	(3.044)
FIRM_DENS	0.009***	0.007**
	(0.002)	(0.003)
UNEMPL	-0.037*	-0.028
	(0.021)	(0.020)
N	952	952
Region FE	YES	YES
Year FE	YES	YES

Negative binomial regression. Marginal effects (predicted number of startups) corresponding to the specifications in columns 4 and 5 in Table 3.1. Delta method standard errors are in parentheses. All regressors lagged one year. *p<0.10, **p<0.05, ***p<0.01.

Overall, our results imply rather sizeable effects of our variables of interest. A 1% increase in the knowledge stock (more precisely, a unit increase in the inverse-hyperbolic-sine-transformed stock of patent applications *KSTOCK*) implies an increase in the predicted number of green startups by 0.663, holding other factors constant (Column 1 of Table 3.2). A 1% increase in the knowledge stock is non-trivial, given the uneven distribution in the untransformed patent stocks across provinces. For illustration, considering that the median number of green startups per province in our sample is 1, this implies that increasing the knowledge stock from the level of the province of Palermo (whose knowledge stock corresponds to the 25th percentile of the distribution in 2020) to the level of the province of Lodi (corresponding to the 55th percentile) would double the predicted number of green startups. However, when interpreting the results, we should note that in our empirical analysis, we cannot make any claim of causal interpretation.

For further context, the median number of patent applications per year is 19, so increasing the number of patent applications filed by 1% implies one or two additional patents but, over the years, about 20% of the provinces in our sample have filed zero patent applications. Therefore, this estimate refers to an average effect, but we expect these relationships' economic size to be highly heterogeneous across the Italian territory.

In turn, increasing the green demand index by 1% (a unit increase in the inverse-hyperbolic-sine-transformed green demand index *GREENDEMAND*, corresponding to about one standard deviation in the index) leads to about 0.349 additional green startups, everything else equal (Column 1 of Table 3.2).

In line with the considerations about hypotheses H1 and H2, the graphical analysis reveals a positive and statistically significant moderation effect for the knowledge stock's size but not for composition. As the green demand index increases, the effect of local knowledge stock gets magnified (Figure 3.6). The effect is most precisely estimated for small-to-moderate levels of green demand due to numerosity in this range of values (95% of the observations have values of the green demand index below 3.5), yet it is unambiguously increasing. This suggests that the local demand for environmental sustainability activates the potential for local knowledge stocks to promote green entrepreneurship. Instead, Figure 3.7 shows that higher shares of green knowledge do not significantly interact with our proxy for the local demand for environmental sustainability. The estimated impact of the share of green knowledge appears to be declining with green demand, pointing to the possibility of substitution between the two. Overall, we interpret these results as evidence that what matters for green entrepreneurship is the presence of knowledge rather than *green* knowledge and that green demand acts as a trigger to activate local knowledge for green entrepreneurship.

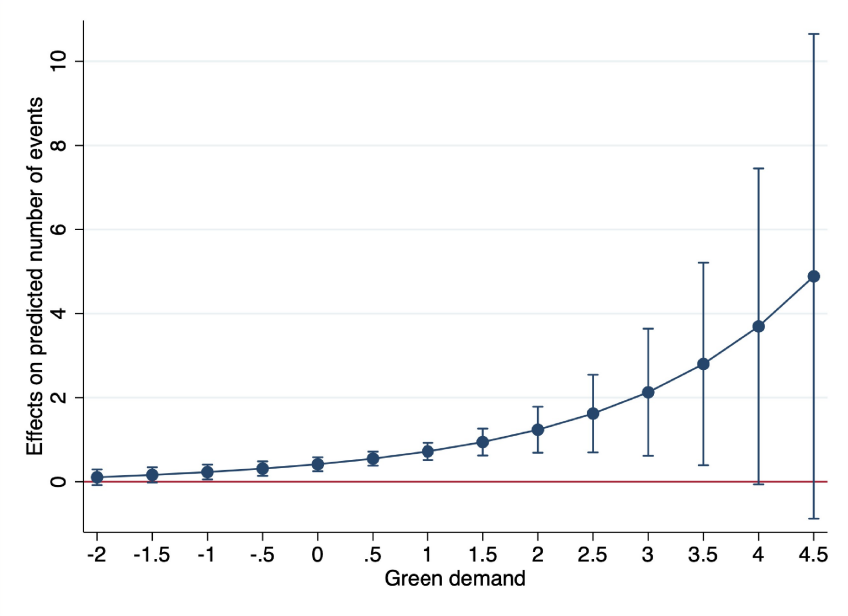


Fig. 3.6 Estimated marginal effects of knowledge stocks on green firm formation at different levels of green demand

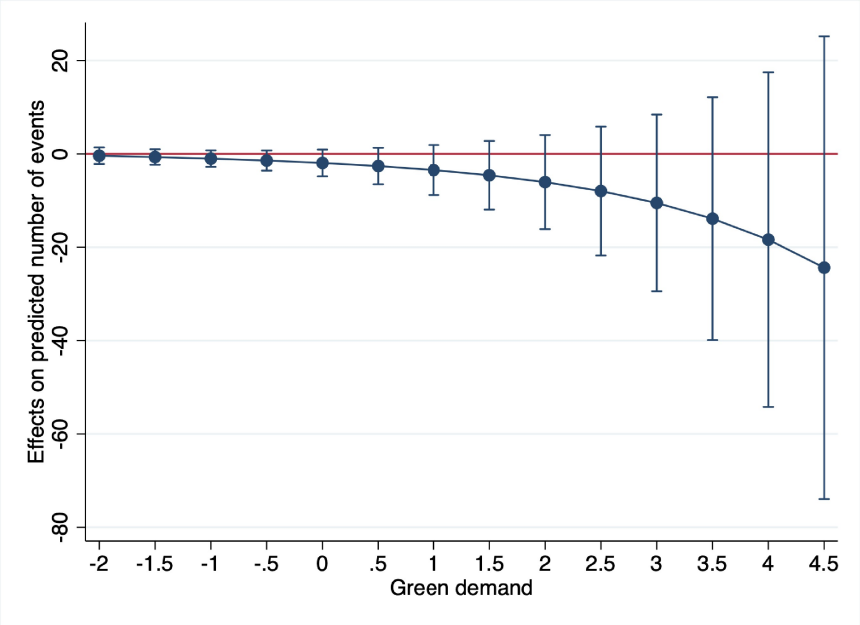


Fig. 3.7 Estimated marginal effects of the share of green knowledge on green firm formation at different levels of green demand

We apply four approaches to test the robustness of our results. First, we show that our findings match the existing literature when we restrict the analysis to green knowledge stocks. Second, we consider an alternative operationalisation of green startups. Third, we consider alternative estimators. Fourth, we introduce a measurement capturing the tacit component of knowledge better and refine our estimates' precision.

As a first robustness check, we compare an estimation considering only green knowledge stocks, versus one with only non-green knowledge stocks – and the usual baseline variables for green demand, interaction with knowledge stocks, and controls. We find that the estimated impact of green knowledge stocks of patent applications on green entrepreneurship, while positive and significant and significantly mediated by green demand, is empirically indistinguishable from that of non-green knowledge (Table C5 and Figure C1 in the Appendix C). This result confirms that the size of knowledge matters rather than its composition.

In a second robustness check, we apply our specification to a different dependent variable: energy startups. The latter is a more established, although more restrictive, measure of green entrepreneurship (e.g., Colombelli & Quatraro, 2019). We see that the estimated marginal effects of the knowledge stock and green demand on energy entrepreneurship are smaller but still significant with this measure (Table C6 in the Appendix C). Again, we do not find significant effects of the green composition of the knowledge stock.

Furthermore, we check the robustness of our results by employing different estimators, the Poisson and Zero-Inflated negative binomial regressions. The Poisson estimates largely confirm the above results (Table C7 in the Appendix C). Again, the local knowledge stocks and green

demand index correlate significantly with entrepreneurship, while the effects of the share of green stocks are not robust on their own or in interaction.

A remaining concern may be that our measure of knowledge stock captures mostly codified knowledge and neglects the role of tacit knowledge in green firm formation (Polanyi, 1966). To address this challenge, we note that STEM graduates possess technical expertise, analytical skills, and problem-solving abilities that are critical for developing innovative solutions to environmental challenges (Audretsch et al., 2002). This knowledge base should support green startup emergence, which often requires advanced scientific and technological insights to create sustainable products or processes, such as renewable energy technologies, green manufacturing, or eco-friendly materials. STEM graduates frequently develop this knowledge through hands-on experience in labs, research projects, or internships (Amoroso et al., 2018).

Hence, we expect more STEM graduates in a region to proxy for the presence of a concentrated pool of individuals who can address local environmental challenges. To gain insights into this aspect, we collect province-level data about the graduates in STEM fields from the European Commission ETER database on higher education institutions[14]. This variable, *STEMGRAD*, can be computed for a shorter period than our main database, and is highly collinear with our graduate share measure. Hence, we only add it as a control in a further robustness check, eliminating the graduate share from the control variables. The results are reported in Table C8 in the Appendix C. The results generally confirm our main findings about the complementarity of the knowledge stocks with green demand, even controlling for this further measure of tacit knowledge.

3.6 DISCUSSION AND CONCLUSIONS

Based on theoretical and empirical arguments, in this paper we have suggested that the Knowledge Spillover Theory of Entrepreneurship may be fruitfully expanded by integrating the role of demand-side factors, especially in relation to the analysis of green entrepreneurship and its innovative expressions. Indeed, our findings are consistent with the established result that knowledge spillovers are more likely to arise in areas rich of knowledge, but, in addition, they confirm a crucial role of demand-side factors. Proxying green demand with sustainability-oriented behaviours, we not only find a role of green demand per se, but also a specific interplay of green demand with local knowledge. The results suggest that green demand makes it easier for firms to effectively leverage knowledge spillovers to start new businesses. Green demand acts as a factor that softens the Knowledge Filter because it increases short-term expected returns on the innovation investment, ultimately allowing knowledge spillovers to be converted into new innovative businesses more swiftly. Moreover, regions with more widespread green behaviours that lead to higher green demand are more likely to have pro-environmental institutions and greater local absorptive capacity in the form of opportunity recognition (Cojoianu et al., 2020; Coll-Martínez et al., 2022; Giudici et al., 2019; Vedula et al., 2019). Finally, we find some indications that both tacit and codified knowledge play a role in sustaining green startup emergence, although we cannot fully disentangle the tacit and codified components and invite future research to focus on this.

[14] For more information, see https://eter-project.com/.

Somewhat unexpectedly, our results suggest that the specific environmental content of the knowledge spillovers matters relatively little. Indeed, when explicitly comparing the size and composition of the knowledge stocks, we find that only the former matters for green entrepreneurship. Accordingly, only the total stock of knowledge significantly interacts with green entrepreneurship. These results indicate that innovative green entrepreneurship relies on total rather than green knowledge, which can be interpreted in the light of the recombinant knowledge theory.

Similarly to green technologies, green businesses may be viewed as standing at the technological frontier and representing the commercialisation of complex and radically innovative business ideas (Barbieri et al., 2023; De Marchi, 2012; Fusillo, 2023; Fusillo et al., 2022). Rather than the outcome of a specifically green pattern of spillovers, green innovation is often the result of the hybridisation of diverse and heterogeneous technologies and knowledge sources and the creative recombination of green and non-green technologies (Colombelli & Quatraro, 2019; Dechezlepretre et al., 2017; Messeni Petruzzelli et al., 2011; Zeppini & Van Den Bergh, 2011). In particular, non-green complementary technologies play a crucial role in green innovation (Barbieri et al., 2023; Fusillo, 2023; Orsatti et al., 2020). These arguments clarify why green demand interacts significantly with the total stock of knowledge but not with its composition. At the same time, the availability of green demand emerges as a leading factor enabling the conversion of diversified knowledge opportunities into innovative green entrepreneurship.

We invite future research to develop this intriguing finding further. Further studies may investigate whether green demand interacts in specific ways with the relatedness or unrelatedness of knowledge (Frenken et al., 2007; Saviotti, 1988) and with the breadth and depth of the local knowledge stock (D'Ambrosio et al., 2017; Laursen & Salter, 2006) to foster green entrepreneurship. Also, future research may investigate the role of green demand in conjunction with market concentration, particularly the possible opposition of large incumbent green firms (Akcigit & Van Reenen, 2023).

While endeavouring to enrich the KSTE with novel insights, our study still suffers from some limitations. First, as discussed, our empirical measure of demand is based on pro-environmental behaviours. Extensions of this work could include more direct measurements of the demand for green products or services, such as consumer surveys, choice experiments, field experiments or lab experiments in the field, willingness-to-pay elicitation methods, and so on.

Second, our results rely on the novel AI-powered web-based methodology that we have developed to identify green startups. On a positive note, this makes our identification of green startups substantially more granular and suitable for embracing the green business phenomenon at large. By construction, we encompass a broader range of firms than previous literature, and our results pertain to a wide spectrum of firms developing novel products or process and service solutions in a B-to-B or B-to-C environment. On the other hand, the same arguments imply that our application does not necessarily compare with previous operationalisations, such as in the case of studies of energy firms. Still, the robustness of our results to the subset of energy startups and their alignment with previous literature confirm that the drivers we identify for green startups at large are broadly relevant, hence reassuring our results' external validity.

Moreover, using startups' websites as the basis for defining a green business is not free from limitations. We expect that most sustainability-oriented startups will highlight these priorities on their websites. However, not all startups have the resources or expertise to create comprehensive websites, and some may not have the necessity to communicate with their stakeholders through a website (for instance, B-to-B companies), resulting in their being underrepresented in our analysis. Furthermore, as information on the websites is self-reported by companies, there is the risk that some companies, due to the increasing interest in sustainability in society, will overstate their green initiatives with exaggerated claims akin to greenwashing (Montgomery et al., 2023). Our approach could, therefore, capture some false positives and tag them as green startups. However, we observe that green startups without a website are equally distributed in Italian territory. Since our analysis is at the province level rather than for the individual firm, we expect that the prevalence of greenwashing should not vary dramatically between locations (Tiba et al., 2021).

Another limitation is that our empirical analysis cannot distinguish green companies based on innovations at the level of products, services, processes or business models. Future development of large language models could explore ways to identify such information from companies' websites and categorise startups accordingly. We suppose that the relevance of knowledge spillovers and local demand is not equal across all these typologies of green businesses. Thus, it would be interesting to test if some of these startups do need to rely on green-specific knowledge or if, instead, our result holds.

Notwithstanding the limits intrinsic to our analysis, we believe our results bear essential implications. Theory-wise, our proposition to include demand-side factors directly into the KSTE domain significantly enlarges the scope of the theory and helps explain what factors help enable knowledge spillover exploitation (Morris et al., 2024; Qian, 2018). The implications of adding the demand side to the KSTE are not limited to new startups but could also involve those large companies that generate a significant share of the local knowledge stock and its spillovers. Green demand, in fact, creates business opportunities for both large firms and startups. While large firms may struggle to adapt quickly to this new demand due to inertia, shortsightedness, or slower decision-making processes, startups can step in quickly to capture this demand with innovative solutions. As large companies attempt to respond to growing green demand, they require adapting specific technologies and processes, which startups are uniquely positioned to provide. Consequently, other than the direct and indirect effects described and tested in this paper within the classical KSTE framework, rising green demand indirectly supports green innovative startups through a derived demand for their technological solutions exerted by large firms. Our work cannot disentangle the two indirect effects arising from green demand (lessening the KF and inducing derived demand), but we hope that future research will look deeper at the interplay between large companies and startups in responding to new demand pressures.

Policy-wise, our results show that environmentally sustainable behaviours instantiating green demand foster green entrepreneurship; hence, sustainable mobility and waste and pollution management could be promoted in their own right and as factors that trigger effects on green entrepreneurship. Such regional and local policies can be a crucial pillar of entrepreneurial

ecosystems (Morris et al., 2024). Moreover, our results suggest that rather than targeting specific environmental technologies, policymakers could more effectively promote entrepreneurship if they supported R&D and knowledge development in general, allowing the inherently unpredictable patterns of innovation to unfold.

3.7 DECLARATION OF INTEREST STATEMENT

No conflict of interest

3.8 REFERENCES

Abdesselam, R., Kedjar, M., & Renou-Maissant, P. (2024). What are the drivers of eco-innovation? Empirical evidence from French start-ups. *Technological Forecasting and Social Change*, *198*, 122953. https://doi.org/10.1016/j.techfore.2023.122953

Acs, Z. J., Audretsch, D. B., Braunerhjelm, P., & Carlsson, B. (2004). The Missing Link: The Knowledge Filter and Entrepreneurship in Endogenous Growth. *CEPR Discussion Papers 4783*.

Acs, Z. J., Audretsch, D. B., & Feldman, M. P. (1994). R & D spillovers and recipient firm size. *The Review of Economics and Statistics*, 336–340.

Acs, Z. J., Braunerhjelm, P., Audretsch, D. B., & Carlsson, B. (2009). The knowledge spillover theory of entrepreneurship. *Small Business Economics*, *32*(1), 15–30. https://doi.org/10.1007/s11187-008-9157-3

Afifi, A., May, S., Donatello, R. A., & Clark, V. A. (2019). *Practical Multivariate Analysis* (6th ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315203737

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters*, *80*(1), 123–129. https://doi.org/10.1016/S0165-1765(03)00032-6

Akcigit, U., & Van Reenen, J. (Eds.). (2023). *The economics of creative destruction: New research on themes from Aghion and Howitt*. Harvard University Press.

Alesina, A., Harnoss, J., & Rapoport, H. (2016). Birthplace diversity and economic prosperity. *Journal of Economic Growth*, *21*(2), 101–138. https://doi.org/10.1007/s10887-016-9127-6

Amoroso, S., Audretsch, D. B., & Link, A. N. (2018). Sources of knowledge used by entrepreneurial firms in the European high-tech sector. *Eurasian Business Review*, *8*(1), 55–70. https://doi.org/10.1007/s40821-017-0078-4

Antonelli, C. (2013). Knowledge governance: Pecuniary knowledge externalities and total factor productivity growth. *Economic Development Quarterly*, *27*(1), 62–70.

Antonelli, C., Crespi, F., Mongeau Ospina, C. A., & Scellato, G. (2017). Knowledge composition, Jacobs externalities and innovation performance in European regions. *Regional Studies*, *51*(11), 1708–1720. https://doi.org/10.1080/00343404.2016.1217405

Antonelli, C., Crespi, F., & Quatraro, F. (2022). Knowledge complexity and the mechanisms of knowledge generation and exploitation: The European evidence. *Research Policy*, *51*(8), 104081. https://doi.org/10.1016/j.respol.2020.104081

Armington, C., & Acs, Z. J. (2002). The Determinants of Regional Variation in New Firm Formation. *Regional Studies*, *36*(1), 33–45. https://doi.org/10.1080/00343400120099843

Audretsch, D. B., & Belitski, M. (2020). The role of R&D and knowledge spillovers in innovation and productivity. *European Economic Review*, *123*, 103391. https://doi.org/10.1016/j.euroecorev.2020.103391

Audretsch, D. B., Belitski, M., & Caiazza, R. (2021). Start-ups, Innovation and Knowledge Spillovers. *The Journal of Technology Transfer*, *46*(6), 1995–2016. https://doi.org/10.1007/s10961-021-09846-5

Audretsch, D. B., Bozeman, B., Combs, K. L., Feldman, M., Link, A. N., Siegel, D. S., Stephan, P., Tassey, G., & Wessner, C. (2002). The Economics of Science and Technology. *The Journal of Technology Transfer*, *27*(2), 155–203. https://doi.org/10.1023/A:1014382532639

Audretsch, D. B., Colombelli, A., Grilli, L., Minola, T., & Rasmussen, E. (2020). Innovative start-ups and policy initiatives. *Research Policy*, *49*(10), 104027. https://doi.org/10.1016/j.respol.2020.104027

Audretsch, D. B., & Keilbach, M. (2007). The Theory of Knowledge Spillover Entrepreneurship. *Journal of Management Studies*, *44*(7).

Audretsch, D. B., Lehemann, E. E., & Hinger, J. (2015). From knowledge to innovation: The role of knowledge spillover entrepreneurship. In *Routledge Handbook of the Economics of Knowledge*.

Audretsch, D. B., Obschonka, M., Gosling, S. D., & Potter, J. (2017). A new perspective on entrepreneurial regions: Linking cultural identity with latent and manifest entrepreneurship. *Small Business Economics*, *48*(3), 681–697. https://doi.org/10.1007/s11187-016-9787-9

Autio, E., Kenney, M., Mustar, P., Siegel, D., & Wright, M. (2014). Entrepreneurial innovation: The importance of context. *Research Policy*, *43*(7), 1097–1108. https://doi.org/10.1016/j.respol.2014.01.015

Barbieri, N., Marzucchi, A., & Rizzo, U. (2020). Knowledge sources and impacts on subsequent inventions: Do green technologies differ from non-green ones? *Research Policy*, *49*(2), 103901. https://doi.org/10.1016/j.respol.2019.103901

Barbieri, N., Marzucchi, A., & Rizzo, U. (2023). Green technologies, interdependencies, and policy. *Journal of Environmental Economics and Management*, *118*, 102791. https://doi.org/10.1016/j.jeem.2023.102791

Bellemare, M. F., Masaki, T., & Pepinsky, T. B. (2017). Lagged Explanatory Variables and the Estimation of Causal Effect. *The Journal of Politics*, *79*(3), 949–963. https://doi.org/10.1086/690946

Bonaccorsi, A., Colombo, M. G., Guerini, M., & Rossi-Lamastra, C. (2013). University specialization and new firm creation across industries. *Small Business Economics*, *41*(4), 837–863. https://doi.org/10.1007/s11187-013-9509-5

Bonfanti, A., De Crescenzo, V., Simeoni, F., & Loza Adaui, C. R. (2024). Convergences and divergences in sustainable entrepreneurship and social entrepreneurship research: A systematic review and research agenda. *Journal of Business Research*, *170*, 114336. https://doi.org/10.1016/j.jbusres.2023.114336

Brouhle, K., & Khanna, M. (2012). Determinants of participation versus consumption in the Nordic Swan eco-labeled market. *Ecological Economics*, *73*, 142–151. https://doi.org/10.1016/j.ecolecon.2011.10.011

Caiazza, R., Belitski, M., & Audretsch, D. B. (2020). From latent to emergent entrepreneurship: The knowledge spillover construction circle. *The Journal of Technology Transfer*, *45*(3), 694–704. https://doi.org/10.1007/s10961-019-09719-y

Cameron, A., & Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, *50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Caravella, S., & Crespi, F. (2021). The role of public procurement as innovation lever: Evidence from Italian manufacturing firms. *Economics of Innovation and New Technology*, *30*(7), 663–684. https://doi.org/10.1080/10438599.2020.1761591

Carree, M., & Thurik, R. (1996). Entry and Exit in Retailing: Incentives, Barriers, Displacement and Replacement. *Review of Industrial Organization, 11*(2), 155–172.

Chapman, G., & Hottenrott, H. (2022). Green start-ups and the role of founder personality. *Journal of Business Venturing Insights*, *17*, e00316. https://doi.org/10.1016/j.jbvi.2022.e00316

Cillo, V., Petruzzelli, A. M., Ardito, L., & Del Giudice, M. (2019). Understanding sustainable innovation: A systematic literature review. *Corporate Social Responsibility and Environmental Management*, *26*(5), 1012–1025. https://doi.org/10.1002/csr.1783

Cohen, B., & Winn, M. I. (2007). Market imperfections, opportunity and sustainable entrepreneurship. *Journal of Business Venturing*, *22*(1), 29–49. https://doi.org/10.1016/j.jbusvent.2004.12.001

Cojoianu, T. F., Clark, G. L., Hoepner, A. G. F., Veneri, P., & Wójcik, D. (2020). Entrepreneurs for a low carbon world: How environmental knowledge and policy shape the creation and financing of green start-ups. *Research Policy*, *49*(6), 103988. https://doi.org/10.1016/j.respol.2020.103988

Coll-Martínez, E., Kedjar, M., & Renou-Maissant, P. (2022). (Green) Knowledge spillovers and regional environmental support: Do they matter for the entry of new green tech-based firms? *The Annals of Regional Science*, *69*(1), 119–161. https://doi.org/10.1007/s00168-022-01111-3

Colombelli, A., D'Ambrosio, A., & Ravetti, C. (2024). Women in innovative start-ups and regional inclusiveness: 'Green' and socially-responsible companies. *Regional Studies*, 1–14. https://doi.org/10.1080/00343404.2024.2340999

Colombelli, A., D'Amico, E., & Paolucci, E. (2023). When computer science is not enough: Universities knowledge specializations behind artificial intelligence startups in Italy. *The Journal of Technology Transfer*, *48*(5), 1599–1627. https://doi.org/10.1007/s10961-023-10029-7

Colombelli, A., Krafft, J., & Vivarelli, M. (2016). To be born is not enough: The key role of innovative start-ups. *Small Business Economics*, *47*(2), 277–291. https://doi.org/10.1007/s11187-016-9716-y

Colombelli, A., & Quatraro, F. (2018). New firm formation and regional knowledge production modes: Italian evidence. *Research Policy*, *47*(1), 139–157. https://doi.org/10.1016/j.respol.2017.10.006

Colombelli, A., & Quatraro, F. (2019). Green start-ups and local knowledge spillovers from clean and dirty technologies. *Small Business Economics*, *52*(4), 773–792. https://doi.org/10.1007/s11187-017-9934-y

Corradini, C. (2019). Location determinants of green technological entry: Evidence from European regions. *Small Business Economics*, *52*(4), 845–858. https://doi.org/10.1007/s11187-017-9938-7

Costantini, V., Crespi, F., Martini, C., & Pennacchio, L. (2015). Demand-pull and technology-push public support for eco-innovation: The case of the biofuels sector. *Research Policy*, *44*(3), 577–595. https://doi.org/10.1016/j.respol.2014.12.011

D'Ambrosio, A., Gabriele, R., Schiavone, F., & Villasalero, M. (2017). The role of openness in explaining innovation performance in a regional context. *The Journal of Technology Transfer*, *42*(2), 389–408. https://doi.org/10.1007/s10961-016-9501-8

De Marchi, V. (2012). Environmental innovation and R&D cooperation: Empirical evidence from Spanish manufacturing firms. *Research Policy*, *41*(3), 614–623. https://doi.org/10.1016/j.respol.2011.10.002

Dechezlepretre, A., Martin, R., & Mohnen, M. (2017). *Knowledge Spillovers from clean and dirty technologies*. GRI Working Papers 135, Grantham Research Institute on Climate Change and the Environment.

Del Río, P., Peñasco, C., & Romero-Jordán, D. (2016). What drives eco-innovators? A critical review of the empirical literature based on econometric methods. *Journal of Cleaner Production*, *112*, 2158–2170. https://doi.org/10.1016/j.jclepro.2015.09.009

Di Stefano, G., Gambardella, A., & Verona, G. (2012). Technology push and demand pull perspectives in innovation studies: Current findings and future research directions. *Research Policy*, *41*(8), 1283–1295. https://doi.org/10.1016/j.respol.2012.03.021

Dong, S., Gong, H., & Liu, T. (2022). Environmental technology spillovers and green start-up emergence: The moderating role of patent commercialization policy and patent enforcement. *Environmental Science and Pollution Research*, *29*(46), 70070–70083. https://doi.org/10.1007/s11356-022-20791-0

Farrow, K., Grolleau, G., & Ibanez, L. (2017). Social Norms and Pro-environmental Behavior: A Review of the Evidence. *Ecological Economics*, *140*, 1–13. https://doi.org/10.1016/j.ecolecon.2017.04.017

Fichter, K., Lüdeke-Freund, F., Schaltegger, S., & Schillebeeckx, S. J. D. (2023). Sustainability impact assessment of new ventures: An emerging field of research. *Journal of Cleaner Production*, *384*, 135452. https://doi.org/10.1016/j.jclepro.2022.135452

Frenken, K., Van Oort, F., & Verburg, T. (2007). Related Variety, Unrelated Variety and Regional Economic Growth. *Regional Studies*, *41*(5), 685–697. https://doi.org/10.1080/00343400601120296

Fritsch, M., & Falck, O. (2007). New Business Formation by Industry over Space and Time: A Multidimensional Analysis. *Regional Studies*, *41*(2), 157–172. https://doi.org/10.1080/00343400600928301

Fusillo, F. (2023). Green Technologies and diversity in the knowledge search and output phases: Evidence from European Patents. *Research Policy*, *52*(4), 104727. https://doi.org/10.1016/j.respol.2023.104727

Fusillo, F., Quatraro, F., & Usai, S. (2022). Going green: The dynamics of green technological alliances. *Economics of Innovation and New Technology*, *31*(5), 362–386. https://doi.org/10.1080/10438599.2020.1799143

Gast, J., Gundolf, K., & Cesinger, B. (2017). Doing business in a green way: A systematic review of the ecological sustainability entrepreneurship literature and future research directions. *Journal of Cleaner Production*, *147*, 44–56. https://doi.org/10.1016/j.jclepro.2017.01.065

Gebhardt, L., & Bachmann, N. (2023). Entrepreneurial contributions to sustainability transitions—A longitudinal study of their representation and enactment through topic modeling and thematic analysis. *Journal of Cleaner Production*, *420*, 138255. https://doi.org/10.1016/j.jclepro.2023.138255

Gidron, B., Bar, K., Finger Keren, M., Gafni, D., Hodara, Y., Krasnopolskaya, I., & Mannor, A. (2023). The Impact Tech Startup: Initial Findings on a New, SDG-Focused Organizational Category. *Sustainability*, *15*(16), 12419. https://doi.org/10.3390/su151612419

Giudici, G., Guerini, M., & Rossi-Lamastra, C. (2019). The creation of cleantech startups at the local level: The role of knowledge availability and environmental awareness. *Small Business Economics*, *52*(4), 815–830. https://doi.org/10.1007/s11187-017-9936-9

Gorovaia, N., & Makrominas, M. (2024). Identifying greenwashing in corporate-social responsibility reports using natural-language processing. *European Financial Management*, eufm.12509. https://doi.org/10.1111/eufm.12509

Greene, W. H. (2003). *Econometric analysis* (5th Edition). Prentice Hall.

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2203.05794

Hall, B. H., Griliches, Z., & Hausman, J. A. (1986). Patents and R and D: Is There a Lag? *International Economic Review*, *27*(2), 265. https://doi.org/10.2307/2526504

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market Value And Patent Citations. *Rand Journal of Economics*, *36*(1), 16–38.

Hausman, J., Hall, B. H., & Griliches, Z. (1984). Econometric Models for Count Data with an Application to the Patents-R & D Relationship. *Econometrica*, *52*(4), 909. https://doi.org/10.2307/1911191

Haws, K. L., Winterich, K. P., & Naylor, R. W. (2014). Seeing the world through GREEN-tinted glasses: Green consumption values and responses to environmentally friendly products. *Journal of Consumer Psychology*, *24*(3), 336–354. https://doi.org/10.1016/j.jcps.2013.11.002

Hoogendoorn, B., Van Der Zwan, P., & Thurik, R. (2020). Goal heterogeneity at start-up: Are greener start-ups more innovative? *Research Policy*, *49*(10), 104061. https://doi.org/10.1016/j.respol.2020.104061

Horbach, J., Oltra, V., & Belin, J. (2013). Determinants and Specificities of Eco-Innovations Compared to Other Innovations—An Econometric Analysis for the French and German Industry Based on the Community Innovation Survey. *Industry & Innovation*, *20*(6), 523–543. https://doi.org/10.1080/13662716.2013.833375

Hörisch, J. (2015). Crowdfunding for environmental ventures: An empirical analysis of the influence of environmental orientation on the success of crowdfunding initiatives. *Journal of Cleaner Production*, *107*, 636–645. https://doi.org/10.1016/j.jclepro.2015.05.046

Horne, J., & Fichter, K. (2022). Growing for sustainability: Enablers for the growth of impact startups – A conceptual framework, taxonomy, and systematic literature review. *Journal of Cleaner Production*, *349*, 131163. https://doi.org/10.1016/j.jclepro.2022.131163

Horne, J., Recker, M., Michelfelder, I., Jay, J., & Kratzer, J. (2020). Exploring entrepreneurship related to the sustainable development goals—Mapping new venture activities with semi-automated content analysis. *Journal of Cleaner Production*, *242*, 118052. https://doi.org/10.1016/j.jclepro.2019.118052

Jha, V. K., & Pande, A. S. (2024). Making sustainable development happen: Does sustainable entrepreneurship make nations more sustainable? *Journal of Cleaner Production*, *440*, 140849. https://doi.org/10.1016/j.jclepro.2024.140849

Joshi, Y., & Rahman, Z. (2015). Factors Affecting Green Purchase Behaviour and Future Research Directions. *International Strategic Management Review*, *3*(1–2), 128–143. https://doi.org/10.1016/j.ism.2015.04.001

Kalcheva, I., McLemore, P., & Pant, S. (2018). Innovation: The interplay between demand-side shock and supply-side environment. *Research Policy*, *47*(2), 440–461. https://doi.org/10.1016/j.respol.2017.11.011

Kesidou, E., & Demirel, P. (2012). On the drivers of eco-innovations: Empirical evidence from the UK. *Research Policy*, *41*(5), 862–870. https://doi.org/10.1016/j.respol.2012.01.005

Koellinger, P. (2008). Why are some entrepreneurs more innovative than others? *Small Business Economics*, *31*(1), 21–37. https://doi.org/10.1007/s11187-008-9107-0

Kuckertz, A., Berger, E. S. C., & Gaudig, A. (2019). Responding to the greatest challenges? Value creation in ecological startups. *Journal of Cleaner Production*, *230*, 1138–1147. https://doi.org/10.1016/j.jclepro.2019.05.149

Laursen, K., & Salter, A. (2006). Open for innovation: The role of openness in explaining innovation performance among U.K. manufacturing firms. *Strategic Management Journal*, *27*(2), 131–150. https://doi.org/10.1002/smj.507

Lee, S. S., Kim, Y., & Roh, T. (2023). Pro-environmental behavior on electric vehicle use intention: Integrating value-belief-norm theory and theory of planned behavior. *Journal of Cleaner Production*, *418*, 138211. https://doi.org/10.1016/j.jclepro.2023.138211

Li, Z., Shang, W., & Yan, M. (2016). News text classification model based on topic model. *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, 1–5. https://doi.org/10.1109/ICIS.2016.7550929

Messeni Petruzzelli, A., Maria Dangelico, R., Rotolo, D., & Albino, V. (2011). Organizational factors and technological features in the development of green innovations: Evidence from patent analysis. *Innovation*, *13*(3), 291–310. https://doi.org/10.5172/impp.2011.13.3.291

Mio, C., Panfilo, S., & Blundo, B. (2020). Sustainable development goals and the strategic role of business: A systematic literature review. *Business Strategy and the Environment*, *29*(8), 3220–3245. https://doi.org/10.1002/bse.2568

Mize, T. (2019). Best Practices for Estimating, Interpreting, and Presenting Nonlinear Interaction Effects. *Sociological Science*, *6*, 81–117. https://doi.org/10.15195/v6.a4

Mokyr, J. (2011). *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton University Press.

Montgomery, A. W., Lyon, T. P., & Barg, J. (2023). No End in Sight? A Greenwash Review and Research Agenda. *Organization & Environment*, 108602662311689. https://doi.org/10.1177/10860266231168905

Morris, A. K., Fiedler, A., & Audretsch, D. B. (2024). Enablers of knowledge spillover entrepreneurship in entrepreneurial ecosystems: Synthesis and future directions. *The Journal of Technology Transfer*, *49*(5), 1737–1761. https://doi.org/10.1007/s10961-023-10056-4

Mrkajic, B., Murtinu, S., & Scalera, V. G. (2019). Is green the new gold? Venture capital and green entrepreneurship. *Small Business Economics*, *52*(4), 929–950. https://doi.org/10.1007/s11187-017-9943-x

Nguyen, T. N., Lobo, A., & Greenland, S. (2016). Pro-environmental purchase behaviour: The role of consumers' biospheric values. *Journal of Retailing and Consumer Services*, *33*, 98–108. https://doi.org/10.1016/j.jretconser.2016.08.010

Nomaler, Ö., & Verspagen, B. (2019). Greentech homophily and path dependence in a large patent citation network. *UNU-MERIT Working Papers*, *#2019-051*.

Orsatti, G. (2024). Government R&D and green technology spillovers: The Chernobyl disaster as a natural experiment. *The Journal of Technology Transfer*, *49*(2), 581–608. https://doi.org/10.1007/s10961-023-10000-6

Orsatti, G., Quatraro, F., & Pezzoni, M. (2020). The antecedents of green technologies: The role of team-level recombinant capabilities. *Research Policy*, *49*(3), 103919. https://doi.org/10.1016/j.respol.2019.103919

Orsatti, G., Quatraro, F., & Scandura, A. (2024). Green technological diversification and regional recombinant capabilities: The role of technological novelty and academic inventors. *Regional Studies*, *58*(1), 120–134. https://doi.org/10.1080/00343404.2023.2176476

Paul, J., Modi, A., & Patel, J. (2016). Predicting green product consumption using theory of planned behavior and reasoned action. *Journal of Retailing and Consumer Services*, *29*, 123–134. https://doi.org/10.1016/j.jretconser.2015.11.006

Polanyi, M. (1966). The Logic of Tacit Inference. *Philosophy*, *41*(155), 1–18.

Purvis, B., Mao, Y., & Robinson, D. (2019). Three pillars of sustainability: In search of conceptual origins. *Sustainability Science*, *14*(3), 681–695. https://doi.org/10.1007/s11625-018-0627-5

Qian, H. (2018). Knowledge-Based Regional Economic Development: A Synthetic Review of Knowledge Spillovers, Entrepreneurship, and Entrepreneurial Ecosystems. *Economic Development Quarterly*, *32*(2), 163–176. https://doi.org/10.1177/0891242418760981

Qian, H., & Jung, H. (2017). Solving the knowledge filter puzzle: Absorptive capacity, entrepreneurship and regional development. *Small Business Economics*, *48*(1), 99–114. https://doi.org/10.1007/s11187-016-9769-y

Quatraro, F., & Scandura, A. (2019). Academic Inventors and the Antecedents of Green Technologies. A Regional Analysis of Italian Patent Data. *Ecological Economics*, *156*, 247–263. https://doi.org/10.1016/j.ecolecon.2018.10.007

Rehfeld, K.-M., Rennings, K., & Ziegler, A. (2007). Integrated product policy and environmental product innovations: An empirical analysis. *Ecological Economics*, *61*(1), 91–100. https://doi.org/10.1016/j.ecolecon.2006.02.003

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1908.10084

Rennings, K. (2000). Redefining innovation—Eco-innovation research and the contribution from ecological economics. *Ecological Economics*, *32*(2), 319–332. https://doi.org/10.1016/S0921-8009(99)00112-3

Reynolds, P., Storey, D. J., & Westhead, P. (1994). Cross-national Comparisons of the Variation in New Firm Formation Rates. *Regional Studies*, *28*(4), 443–456. https://doi.org/10.1080/00343409412331348386

Roome, N., & Wijen, F. (2006). Stakeholder Power and Organizational Learning in Corporate Environmental Management. *Organization Studies*, *27*(2), 235–263. https://doi.org/10.1177/0170840605057669

Santoalha, A., & Boschma, R. (2021). Diversifying in green technologies in European regions: Does political support matter? *Regional Studies*, *55*(2), 182–195. https://doi.org/10.1080/00343404.2020.1744122

Saviotti, P. P. (1988). Information, variety and entropy in technoeconomic development. *Research Policy*, *17*(2), 89–103. https://doi.org/10.1016/0048-7333(88)90024-8

Schaltegger, S., & Wagner, M. (2011). Sustainable entrepreneurship and sustainability innovation: Categories and interactions. *Business Strategy and the Environment*, *20*(4), 222–237. https://doi.org/10.1002/bse.682

Schøtt, T., & Jensen, K. W. (2016). Firms' innovation benefiting from networking and institutional support: A global analysis of national and firm effects. *Research Policy*, *45*(6), 1233–1246. https://doi.org/10.1016/j.respol.2016.03.006

Shane, S. (2009). Why encouraging more people to become entrepreneurs is bad public policy. *Small Business Economics*, *33*(2), 141–149. https://doi.org/10.1007/s11187-009-9215-5

Shepherd, D. A., & Patzelt, H. (2011). The New Field of Sustainable Entrepreneurship: Studying Entrepreneurial Action Linking "What is to be Sustained" with "What is to be Developed". *Entrepreneurship Theory and Practice*, *35*(1), 137–163. https://doi.org/10.1111/j.1540-6520.2010.00426.x

Spigel, B. (2017). The Relational Organization of Entrepreneurial Ecosystems. *Entrepreneurship Theory and Practice*, *41*(1), 49–72. https://doi.org/10.1111/etap.12167

Stam, E. (2015). Entrepreneurial Ecosystems and Regional Policy: A Sympathetic Critique. *European Planning Studies*, *23*(9), 1759–1769. https://doi.org/10.1080/09654313.2015.1061484

Sutaria, V., & Hicks, D. A. (2004). New firm formation: Dynamics and determinants. *The Annals of Regional Science*, *38*(2), 241–262. https://doi.org/10.1007/s00168-004-0194-9

Tiba, S., Van Rijnsoever, F. J., & Hekkert, M. P. (2021). Sustainability startups and where to find them: Investigating the share of sustainability startups across entrepreneurial ecosystems and the causal drivers of differences. *Journal of Cleaner Production*, *306*, 127054. https://doi.org/10.1016/j.jclepro.2021.127054

Umamaheswaran, S., Dar, V., Sharma, E., & Kurian, J. S. (2023). Mapping Climate Themes From 2008-2021—An Analysis of Business News Using Topic Models. *IEEE Access*, *11*, 26554–26565. https://doi.org/10.1109/ACCESS.2023.3256530

UNEP. (2019). *Global Environment Outlook 6 (GEO6)*. United Nations Environment Programme. https://www.unep.org/resources/global-environment-outlook-6

Vedula, S., York, J. G., & Corbett, A. C. (2019). Through the Looking-Glass: The Impact of Regional Institutional Logics and Knowledge Pool Characteristics on Opportunity Recognition and Market Entry. *Journal of Management Studies*, *56*(7), 1414–1451. https://doi.org/10.1111/joms.12400

Wagner, J., & Sternberg, R. (2004). Start-up activities, individual characteristics, and the regional milieu: Lessons for entrepreneurship support policies from German micro data. *The Annals of Regional Science*, *38*(2), 219–240. https://doi.org/10.1007/s00168-004-0193-x

Wöhler, J., & Haase, E. (2022). Exploring investment processes between traditional venture capital investors and sustainable start-ups. *Journal of Cleaner Production*, *377*, 134318. https://doi.org/10.1016/j.jclepro.2022.134318

Wong, S. K. S. (2013). Environmental Requirements, Knowledge Sharing and Green Innovation: Empirical Evidence from the Electronics Industry in China. *Business Strategy and the Environment*, *22*(5), 321–338. https://doi.org/10.1002/bse.1746

York, J. G., Vedula, S., & Lenox, M. J. (2018). It's Not Easy Building Green: The Impact of Public Policy, Private Actors, and Regional Logics on Voluntary Standards Adoption. *Academy of Management Journal*, *61*(4), 1492–1523. https://doi.org/10.5465/amj.2015.0769

York, J. G., & Venkataraman, S. (2010). The entrepreneur–environment nexus: Uncertainty, innovation, and allocation. *Journal of Business Venturing*, *25*(5), 449–463. https://doi.org/10.1016/j.jbusvent.2009.07.007

Yun, J., & Geum, Y. (2020). Automated classification of patents: A topic modeling approach. *Computers & Industrial Engineering*, *147*, 106636. https://doi.org/10.1016/j.cie.2020.106636

Zeppini, P., & Van Den Bergh, J. C. J. M. (2011). Competing Recombinant Technologies for Environmental Innovation: Extending Arthur's Model of Lock-In. *Industry & Innovation*, *18*(3), 317–334. https://doi.org/10.1080/13662716.2011.561031

Zhang, F., & Zhu, L. (2019). Enhancing corporate sustainable development: Stakeholder pressures, organizational learning, and green innovation. *Business Strategy and the Environment*, *28*(6), 1012–1026. https://doi.org/10.1002/bse.2298

CHAPTER 4

Legitimacy to attract attention and develop networks: exploring entrepreneurial legitimacy claims on LinkedIn

4 LEGITIMACY TO ATTRACT ATTENTION AND DEVELOP NETWORKS: EXPLORING ENTREPRENEURIAL LEGITIMACY CLAIMS ON LINKEDIN

4.1 ABSTRACT

Narratives and legitimacy claims are strong tools for entrepreneurs to face the liability of newness and establish legitimacy. By sharing legitimacy claims on social networking sites, entrepreneurs present their ventures' values, achievements, and purpose to broad audiences, fostering cognitive, normative, and pragmatic legitimacy while gathering attentional ressources. This task is particularly difficult and crucial for green ventures, which must convince varied and often skeptical stakeholders of both their economic viability and environmental impact. We introduce a novel methodology leveraging Large Language Models to identify legitimacy claims in narratives shared in 61,395 LinkedIn posts by 1,703 Italian entrepreneurs. Our results show that legitimacy claims positively affect attentional ressources and on network size. Moreover, attentional ressources mediate the effect of legitimacy claims on network growth. Our research advances the quantitative study of entrepreneurial narratives and demonstrates how legitimacy-building communication contributes to the development of social networks of green and non-green entrepreneurs.

4.2 INTRODUCTION

To start their ventures, entrepreneurs have to overcome the liability of newness (Stinchcombe, 1965). Being unknown to their audience and having few network connections, this is a hard task. A first step is to convince their stakeholders that their business is going to be beneficial to society and that they will be profitable; in other words, they must establish their legitimacy (Suchman, 1995). Attaining legitimacy gives access to resources and builds connections with the customer base (Gordo Molina et al., 2022). This is particularly difficult for ventures facing complex and often challenging audiences, such as green ventures or startups (Castelló et al., 2016). In this context, legitimacy is critical, as it enables higher levels of environmental performance and enhances firms' capacity for green innovation (Ge et al., 2016; Soewarno et al., 2019).

The legitimacy work, gathering all the actions contributing to build legitimacy, can have multiple forms (Lefsrud et al., 2020): it can be offline, for instance attending startup event, making partnership with well-known firms, attract media attention, or leveraging network connections (Stuart et al., 1999); or online, for instance describing the company on social media (Antretter et al., 2019). Entrepreneurs are increasingly using the internet and Social Networking Sites (SNS) not only to build, develop, and maintain their network (Ferro, 2015), but also to present their venture, attract attention, and develop legitimacy (Zhao et al., 2023). These platforms increase the potential for interaction and engagement with audiences and facilitate the mobilization of resources from entrepreneurial networks.

The driver of legitimacy most explored by researchers is narratives (Taeuscher et al., 2021), which are information or stories about a firm that can induce emotional or cognitive responses by the narratee (Gordo Molina et al., 2022). Entrepreneurs embed legitimacy claims within the narratives they share on social networking sites (SNSs) to establish and reinforce their legitimacy. By making

the firm understandable and taken for granted, a narrative contributes to its cognitive legitimacy; by portraying values that align with those of its audience, a narrative enhances the firm's normative legitimacy; and by highlighting the direct benefits delivered to stakeholders, the firm's narrative strengthens its pragmatic legitimacy (Suchman, 1995).

Prior research has identified networks as a key source of legitimacy for new ventures. Strong ties—such as endorsements, partnerships, or investment relationships—with already legitimate actors can transfer legitimacy to entrepreneurial firms (Hallen & Eisenhardt, 2012; Higgins & Gulati, 2003, 2006). However, the effect of legitimacy on networks has not been explored. We aim to fill this gap by combining legitimacy theory with the literature on social networking sites. Accordingly, we propose that entrepreneurs strategically share legitimacy claims on social networks to build online legitimacy and attract attention, which in turn allow them to expand and strengthen their networks. Further, we expect the effectiveness of legitimacy to differ between green and non-green ventures, given their distinct audiences, values, and institutional contexts.

We explore those propositions on 61,395 LinkedIn posts of 1,703 Italian entrepreneurs. While most of the literature explored narratives qualitatively on relatively small samples, we leverage recent advances in NLP (Chew et al., 2023) and large language models (LLM) to develop a novel method to automatically perform qualitative data analysis. This method enables quantitative analysis of narrative data, addressing a methodological gap in entrepreneurial and legitimacy research. (Audretsch & Lehmann, 2023; Suddaby et al., 2017). Our results show that both cognitive and pragmatic legitimacy claims attract attention on green and non-green entrepreneurs' Social Networking Sites (SNS) posts. Normative legitimacy claims only attract attention to non-green firms, suggesting that green firms already appear as normatively legitimate. Further, the attention gathered by the post has a significant impact on network size. On the other hand, only the cognitive legitimacy claims have a direct impact on network size, and only for non-green firms.

This paper makes multiple contributions to the existing literature. First, it enriches the understanding of the links between legitimacy and network by showing that entrepreneurs' legitimacy claims have a positive impact on network size, mostly mediated by the attention they gather. Second, it contributes to the literature on green entrepreneurship by underlining the different impact of legitimacy claims for green and non-green firms. Third, it introduce a new method for quantifying narratives and legitimacy claims, opening the opportunity for much quantitative research on legitimacy and entrepreneurship (Audretsch & Lehmann, 2023; Suddaby et al., 2017)

4.3 LITTERATURE REVIEW

4.3.1 Legitimacy

All new ventures suffer from the liability of newness (Stinchcombe, 1965): to be successful, they have to gather resources, create networks, and hire workers while creating a new and by definition unknown activity. One step to overcome this is to establish legitimacy (Aldrich & Fiol, 1994). This construct has had multiple definitions with minor variations. A widely accepted definition of legitimacy is "a generalized perception or assumption that the actions of an entity are desirable,

proper, or appropriate within some socially constructed system of norms, values, beliefs, and definitions." (Suchman, 1995).

Early theoretical works have defined legitimacy as a multidimensional construct (Aldrich & Fiol, 1994; Scott, 1995; Suchman, 1995). In this work, we will rely on the framework of Suchman (1995), the most used in the literature. He identified three main dimensions of legitimacy: the cognitive, the normative (or moral), and the pragmatic legitimacy.

The cognitive legitimacy is the dimension that received the most research interest: the acceptance that an organization or activity makes sense within a shared system of beliefs and cultural frameworks. A new venture achieves full cognitive legitimacy when its business is well known, easily understood, and taken for granted as a natural part of the market (Suchman, 1995). One way for new ventures to develop cognitive legitimacy is to show their actions and the products and services they propose (Fisher et al., 2017). Another approach is to emphasize similarities to established firms and to the broader industry, which helps audiences recognize new ventures as legitimate market actors. Early research framed cognitive legitimacy as opposing firm distinctiveness; however, later studies show that new ventures benefit from balancing similarity and distinctiveness (Navis & Glynn, 2011; van Werven et al., 2015). By aligning with a recognized market category while simultaneously differentiating themselves within that category, new ventures can achieve optimal distinctiveness, thereby enhancing their legitimacy (Taeuscher et al., 2021; Vossen & Ihl, 2020).Thus, both differentiation and conformity contribute to cognitive legitimacy (Zamantılı Nayır & Shinnar, 2020). This is especially true for ventures such as startups, who introduce innovative products on the market and are often challenged on their innovativeness.

In a similar way, the normative legitimacy has received a lot of interest. By other authors, this construct has been labelled the socio-political or moral legitimacy (Aldrich & Fiol, 1994; Scott, 1995). In the rest of the paper, we will use the term normative legitimacy, since it is the most used today. It refers to the normative evaluation of whether an organization's actions are appropriate, desirable, or "the right thing to do" according to social values and norms. Leaders of organisations, such as founders of a new venture, can build normative legitimacy using their personal influence and charisma and their identity to show their values and beliefs (Navis & Glynn, 2011; Suchman, 1995). They can display isomorphism, showing that they conform to regulatory pressure and social norms (Suddaby et al., 2017). With the growing importance that society places on sustainability, marked by the adoption of the Sustainable Development Goals (UN, 2015), actions that promote sustainability are becoming central for organizations to attain normative legitimacy (Ellerup Nielsen & Thomsen, 2018). To do so, entrepreneurs may engage in symbolic actions, signalling commitment to sustainability primarily through external communication, or in substantive actions, where disclosure is reinforced by concrete organizational changes, policies, and performance targets (Lodhia et al., 2022).

The pragmatic legitimacy refers to the self-interested evaluation by stakeholders, based on whether an organization serves their interests or delivers them tangible benefits. Startups are young and innovative ventures; their consumers are keen on innovative products and often early adopters (Ebrahimi et al., 2022). Thus, showing innovativeness is key to interest their consumers and to achieve pragmatic legitimacy. Since startups face the liability of newness and a low rate of survival (Stinchcombe, 1965), a second challenge is to convince stakeholders that they have gathered sufficient resources for overcoming it. They can do so by displaying financial legitimacy

(Rutherford et al., 2016), which can be considered a subdimension of pragmatic legitimacy, by showing their financial performance and the financial support they receive.

Those three dimensions of legitimacy have been extensively explored in the literature, but they have never been linked to online legitimacy. We propose that building legitimacy in those three dimensions, in particular with legitimacy claims, contributes to online legitimacy.

4.3.2 Legitimacy for green and non-green new ventures

While legitimacy gives benefits to every firm, it is crucial for young ventures (Fisher, 2020; Zimmerman & Zeitz, 2002). In line with the majority of research on the subject (Suddaby et al., 2017), we view legitimacy as an intangible asset, which new ventures can strategically develop (Castelló et al., 2016; Suchman, 1995; Tornikoski & Newbert, 2007; Zimmerman & Zeitz, 2002) to access resources, stakeholders, new markets, and construct identity, image, and customer loyalty (Gordo Molina et al., 2022). Founders, by sharing narratives and their identity, play a key role in establishing the legitimacy of their venture (Navis & Glynn, 2011). Legitimation, the process of establishing legitimacy, is inherently challenging. It is context-dependent (Kibler et al., 2018), it must target different audiences with different standards, norms, and values and often involves pursuing various levels of legitimacy simultaneously (Fisher et al., 2017). To build legitimacy, firms must engage in legitimacy work (Riandita et al., 2022), for example, by sharing success stories, creating links with legitimizing actors, or receiving media attention. It must be continuously sustained to maintain, and if needed, to recover legitimacy (Suchman, 1995; Suddaby et al., 2017).

Legitimacy is particularly crucial and challenging for green new ventures. These ventures often communicate their environmental values and beliefs to external audiences, yet such claims are frequently questioned or challenged (O'Neil & Ucbasaran, 2016). Legitimation for sustainable development issues requires an active engagement and establishing a trust-based relationship with a wide variety of stakeholders (Castelló et al., 2016). Since green entrepreneurship is multifaceted, defending commercial and environmental values, which are sometimes contradictory, their legitimation claims might be ambiguous, and they could require more legitimacy work (Riandita et al., 2022). Legitimacy enables green proactive entrepreneurs to achieve higher environmental performance and positively influences their capacity for green innovation (Ge et al., 2016; Soewarno et al., 2019). Further, green initiatives have a positive effect on the legitimacy of all types of firms (Truong & Nagy, 2021), but the literature did not explore whether this effect is different between green and non-green firms. Setting to enrich the green entrepreneurship literature, we state the following hypothesis:

Hypothesis 1. The role of legitimacy differs for green and non-green entrepreneurs

4.3.3 Legitimacy and networks

On a parallel ground, research on networks, broadly defined as sets of actors and the links between them (Alvedalen & Boschma, 2017; Hoang & Antoncic, 2003), showed that networks are also key to face the liability of newness (van Burg et al., 2022). Indeed, entrepreneurs can leverage their networks to access crucial resources, such as knowledge, skill, talent, new markets, customers, and suppliers (Alvedalen & Boschma, 2017), and have a positive impact on

performance (Stam et al., 2014). Thus, building and developing a network is crucial at the initial, but also later stages, of a venture's life cycle. In the strategic view of the firm, entrepreneurs pursue strategies to form helpful networks (Stuart & Sorenson, 2007), such as building legitimacy. These arguments lead to the following research questions: What role does legitimacy play in the strategic development of entrepreneurial networks ?

Research on legitimacy has primarily focused on the role of networks in legitimacy building (Fisher et al., 2017; Suddaby et al., 2017). Engagement with stakeholders has been identified as crucial for building legitimacy (Castelló et al., 2016), and it enables firms to access and mobilize resources within their networks (Gebert-Persson & Káptalan-Nagy, 2016; Martens et al., 2007). Research highlights that different types of strong ties, such as endorsement (Higgins & Gulati, 2003, 2006; Stuart et al., 1999), customer-supplier relationship (Crespin-Mazet & Dontenwill, 2012), and investment ties (Hallen & Eisenhardt, 2012) with legitimate organisations, generate positive spillovers (Haack et al., 2014). A few qualitative studies suggest that actors engage in strategic networking by sharing legitimacy claims (Connelly et al., 2020; Hallen & Eisenhardt, 2012). We argue that this is particularly relevant on Social Networking Sites (SNS), used by entrepreneurs primarily to share narratives and to network (Ferro, 2015).

To make the firm easier to understand and taken-for-granted, entrepreneurs can share legitimacy claims consisting of information about their firms products and services, their activities, and the similarity and distinctiveness to incumbent firms (Suchman, 1995). Cognitive legitimacy reduces uncertainty (Shepherd et al., 2003)), which lowers the barriers to networks. This lead us to the following hypothesis:

Hypothesis 2: Cognitive legitimacy has a positive impact on entrepreneurial networks.

Bjornali et al. (2017) showed that startups state their environmental engagement and leverage their normative legitimacy to attract customers, investors and governmental supports, and even to build partnerships. Accordingly, we will explore the following hypothesis:

Hypothesis 3: Normative legitimacy has a positive impact on entrepreneurial networks.

By building reputation, startups contribute to stakeholder engagement (Chen et al., 2019). Further, potential investors and partners are looking for opportunities and would like to network with the firms with the greatest payoff (Prashantham & Madhok, 2023), which underline the importance of communicating on successes and achievements and more generally to establish pragmatic legitimacy. Thus, we formulate the following hypothesis:

Hypothesis 4: Pragmatic legitimacy has a positive impact on entrepreneurial networks.

4.3.4 Legitimacy, narratives, and social networking sites

Researchers explored the key role of entrepreneurs' claims and narratives in building the legitimacy of their venture (Garud et al., 2014; Gordo Molina et al., 2022; Lounsbury & Glynn, 2001; Taeuscher et al., 2021). From the narrowest to the broadest, a narrative has been defined as (1) a story composed of at least an original state, an action or an event, and a consequent state (Czarniawska, 1998); a simple story or explanation of events (Shiller, 2017); any information that can be interpreted as a narrative by a beholder (Rudrum, 2005). Many sources of narratives have been explored, such as media accounts (Navis & Glynn, 2010), interviews (O'Neil & Ucbasaran,

2016), Corporate Sustainability Report (Ellerup Nielsen & Thomsen, 2018), websites, blogs (O'Neil & Ucbasaran, 2016), or social media posts (Seigner et al., 2023). Narratives can include legitimacy claims, which are effective tools for building legitimacy and gathering resources (Lounsbury & Glynn, 2001; Martens et al., 2007; Taeuscher et al., 2021, 2022).

Narratives do not have a predefined form (Czarniawska, 1998). While the most studied are texts, for instance shared in company brochures, websites, or social media posts (Czarniawska, 1998), they can also be oral (Martens et al., 2007), include images (Rudrum, 2005), and videos (Zhao et al., 2023). Much research explored narratives shared in interviews (Castelló et al., 2016) or in new venture documents, such as corporate sustainability reports (Dobers & Springett, 2010) or crowdfunding campaigns (Antretter et al., 2019). With the increasing use of social media and Social Networking Sites (SNS), entrepreneurs share more and more narratives in a "communicative stream", composed of short posts and stories published regularly to build legitimacy with a large audience (Castelló et al., 2016; Fischer & Rebecca Reuber, 2014; Zhao et al., 2023) and to contribute to developing online networks (Olanrewaju et al., 2020; Wang et al., 2020)

By enabling multi-stakeholder access and fostering two-way communication, SNS are particularly effective in helping sustainable new ventures legitimize their sustainable development efforts (Castelló et al., 2016). It allows entrepreneurs to build online legitimacy—defined as the social appreciation and/or desirability constructed and measured in the online world—which is a key indicator of firm survival (Antretter et al., 2019). Although very few studies explored online legitimacy so far, its importance for entrepreneurs is growing with the increasing use of the internet and SNS, such as Facebook or LinkedIn. For instance, (Banerji & Reimer, 2019) show that entrepreneurs well-connected on LinkedIn raise funds more easily.

Early research shows that legitimacy is crucial to attract attention and support (Ashforth & Mael, 1989). By sharing posts that provide information about their identity, business, and legitimacy, entrepreneurs attract audience attention (Ferro, 2015). Thus, we state the following hypotheses:

Hypothesis 5: Cognitive legitimacy has a positive impact on attentional resources.

Hypothesis 6: Normative legitimacy has a positive impact on attentional resources

Hypothesis 7: Pragmatic legitimacy has a positive impact on attentional resources.

Startups attract attention from their audiences and stakeholders to facilitate their engagement and interaction (Ferro, 2015). Such activity goes beyond a marketing effort of making the firm known to their audience: with social media posts, entrepreneurs can persuade investors to invest, improve knowledge management (Crammond et al., 2018) and develop new business models (Ketonen-Oksi et al., 2016). Thus, attention can be considered as a rare and valuable resource (Zhao et al., 2023), which improves the perceived value and importance of the firm in the eyes of stakeholders and improves its ability to network (Petkova et al., 2013). Thus, we state the following hypothesis:

Hypothesis 8. Attentional resources mediate the effect of legitimacy on entrepreneurial networks.

Fig. 4.1 presents our conceptual framework and summarizes the hypothesis.

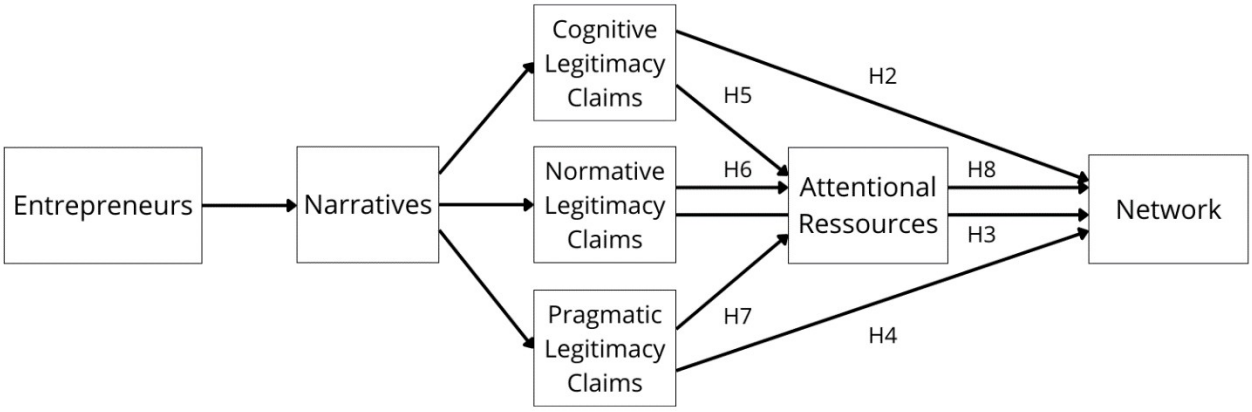


Fig. 4.1 Conceptual framework

4.4 METHOD

Research on legitimacy used mosly qualitative inductive thematic analysis. With inductive thematic analysis, researchers review the data, identify its key themes and patterns without relying on a pre-established codebook (Zhang et al., 2024). This enriched the understanding of legitimacy but also increased its complexity, producing a proliferation of dimensions with unclear links between them (Suddaby et al., 2017). Despite the central role of narratives and legitimacy claims in the legitimation process, researchers have struggled to move beyond qualitative exploration toward systematic quantification (Audretsch & Lehmann, 2023; Suddaby et al., 2017).

Attempts to quantify how narratives contribute to legitimacy using traditional Natural Language Processing (NLP) tools remain limited. (Taeuscher et al., 2022) studied crowdfunding campaigns and measured cognitive legitimacy claims through topic distinctiveness and normative legitimacy claims with a dictionary. (Antretter et al., 2019) measured online legitimacy claims in Twitter posts by applying dictionary-based sentiment analysis, Naïve Bayes word classification, and simple language dummies. (Castelló et al., 2016) measured legitimacy claims with the number of Twitter posts, their topic diversity, and their interaction with their followers. These approaches remained constrained to surface-level word counts and associations and were unable to capture the nuanced semantic and contextual dimensions of legitimacy.

Still, most researchers agree that qualitative analysis remains the best tool to explore legitimacy and rely on case studies to develop the field. Unfortunately, qualitative research is very time-consuming and requires high experience from multiple coders (Zhang et al., 2024), which do not allow for large-scale quantitative studies. (Navis & Glynn, 2010) present an interesting attempt to link qualitative and quantitative analysis of legitimacy by manually coding entrepreneurial narratives in satellite radio to see their contribution to the legitimacy over time. Unfortunately, due to the considerable cost of manual coding, this research is limited to a single case study.

With recent advances in Natural Language Processing, particularly the emergence of Large Language Models (LLMs), researchers are beginning to explore LLM-guided qualitative research (Barros et al., 2025). (Goyanes et al., 2025) used LLM as a coding assistant to speed up and improve the first steps of qualitative analysis. (Chew et al., 2023) pushed the use of LLM further and offered a methodology for deductive analysis. They developed a list of prompts to use with an LLM API to code texts with a level of agreement with human coders as high as human coders. (Qiao et al., 2025) extended the use of LLM for inductive coding, using multiple LLM agents to create a first-order theme codebook, aggregate them in a second-order theme, and then code the texts. Those examples show that LLM offer the potential to bridge the gap between traditional qualitative approaches and scalable text analysis, opening new possibilities for studying legitimacy at larger scales while retaining conceptual nuance.

To answer our research questions, we will run two sets of regression. In particular, we run a first set at the post level to see whether legitimacy claims attract attentional resources. The second set is at the entrepreneur level, to see whether legitimacy claims and attentional resources contribute to bigger networks. In the next parts, we will present the collection of our data, the different variables, and the models.

4.4.1 Data sample

We leverage the sample of italian innovative startups discussed in 3.4.1, using AIDA to access firm-level information and the scrapping procedure to gather the full text of the websites of 10,939 start-ups. To identify the green startups in this sample, we followed the method detailed in Chapter 1 the websites of the startups. To begin with, we established a list of green labels using topic modelling applied to the green SDG identified in the Global Environement Outlook 6 (UN, 2019). Then, we applied the three NLP algorithms of Chapter 1, leveraging respectively a Machine Learning (ML) dictionary approach, Latent Dirichlet Allocation (LDA), and BERTopic. For each methods, we tag the 15% of startups with the highest green score. Finally, we label the startups tagged by at least two methods as green.

To enrich our dataset with entrepreneur-level data, we bought LinkedIn data from BrightData, accessing rich employee profile information and the LinkedIn ID of 6,219 startups and 52,592 employees. We identified entrepreneurs' profiles searching for the keywords 'ceo' and 'founder', identifying 1703 entrepreneur profiles. Using their LinkedIn ID, we scraped up to 100 of their latest posts (including reposts) using Apify, for a total of 61,395 posts. We removed the top 1% of the posts with the highest number of posts, identifying them as outliers, for a final number of 60,781 posts (Figure 2).

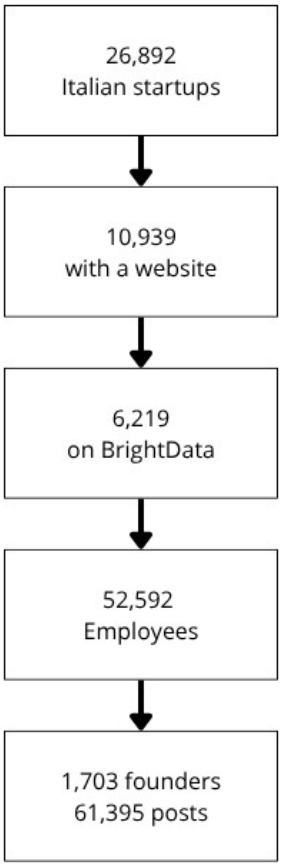


Fig. 4.2 Data pipeline

To determine whether this subsample of startups with a linkedin account is representative of the population of innovative startups, we conducted several analyses. First, we explored the distribution by ATECO 2-digit code, the classification of economic activities used by the Italian National Statistical Institute (ISTAT). We plot the 10 most frequent ATECO codes for each sample and a category 'others' for readability in Fig.4.3. Both the registered population and subsample are empirically concentrated in software, ICT, and other knowledge-intensive services[15], and are very similar across the other sectors.

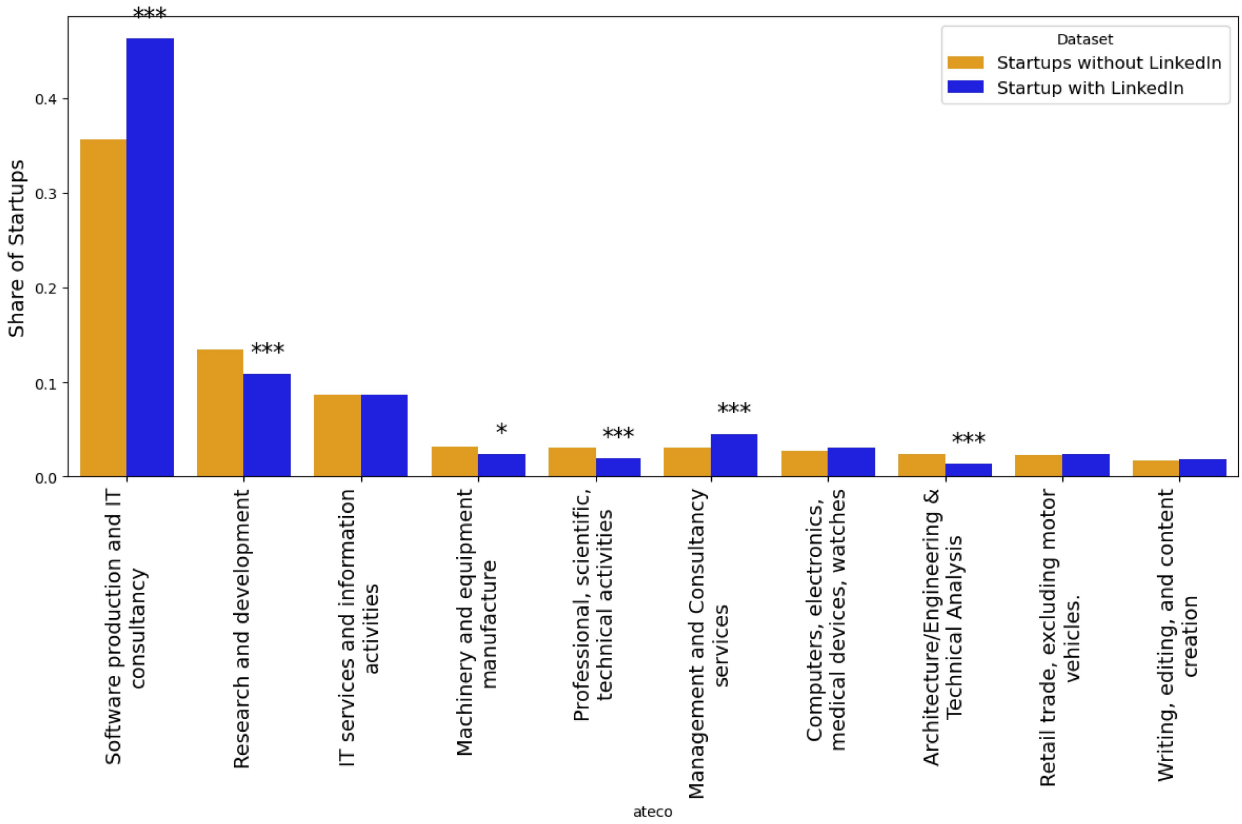


Fig. 4.3 Comparison of areas of activities (measured by ATECO codes) between startups without and with LinkedIn account. Only the top 10 ATECO are represented, covering three-fourths of the start-ups.

Further, we compare the population and subsample on a few economic characteristics. Table 4.1 shows that startups with websites are slightly younger, larger (as measured by assets or employees), and have a higher relative growth rate. Differences are statistically significant, but their magnitude is not economically worrisome.

Table 4.1 Comparison of statistics of startups without and with Linkedin account

Statistic	Startups without LinkedIn	Startups with LinkedIn
Age	6.34	6.15***
	(3.04)	(3.00)
Assets	136.61	168.36***
	(195.37)	(211.56)
Number of employees	1.67	3.33***
	(6.26)	(10.32)
Relative growth	0.02	0.03***
	(0.06)	(0.07)

In Fig 4.4 and Fig 4.5, we plot the geographical distribution of startups without and with websites in Italy, showing a very similar distribution. However, due to the size of the sample, the regions with few startups have very few representants.

Fig. 4.4 Number of startups with LinkedIn account in Italian regions

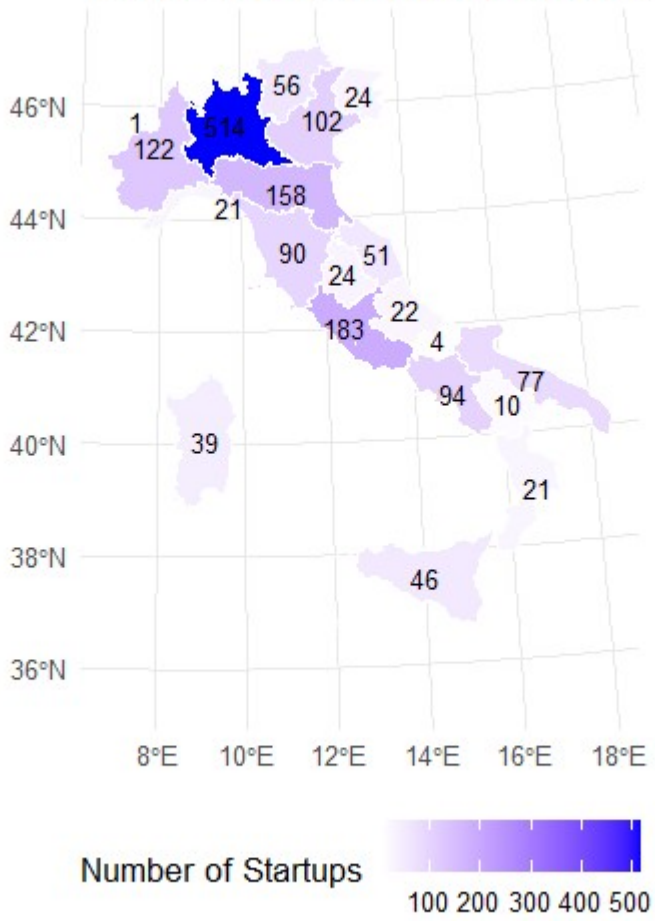
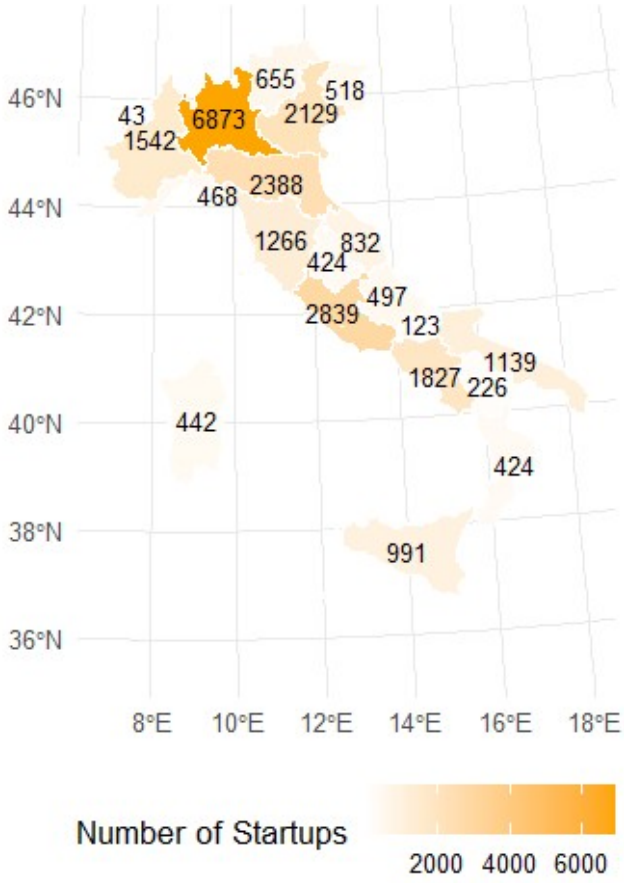
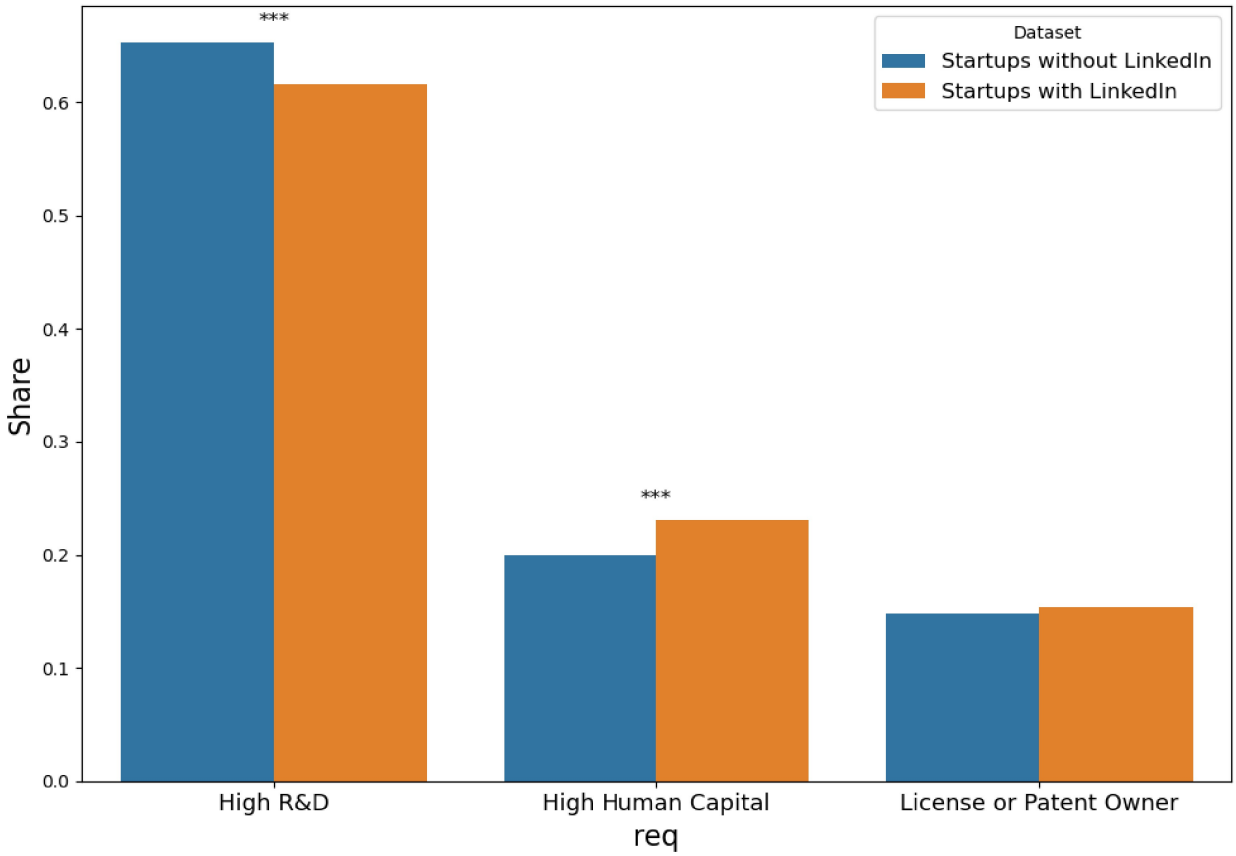


Fig. 4.5 Number of startups without LinkedIn account in Italian regions



Finally, we plot in Fig. 4.6 the innovation criteria that startups with and without websites met. Startups without websites are slightly more likely to meet the high R&D criteria and less likely to meet the High Human Capital and Patent Owner criteria, but, again, the magnitude of the difference is not relevant for our methodological application.

Fig. 4.6 Barplots of innovation criteria met by the startups with and without LinkedIn account



4.4.2 Dependant variables

Our first dependent variable, which we will use for the regression at the post level, is attentional resources, measured with the number of reactions to a post. A high number of reactions shows that the entrepreneur is accepted by his audience, suggesting a potential for audience involvement, and is crucial for startup success (Zhao et al., 2023). The distribution of this variable displays a very long tail, so we log-transform it. The transformation appears normally distributed.

The second, which we will use for the regression at the entrepreneur level, is network size, measured by the number of followers on LinkedIn. Although such ties are often relatively weak, their benefits have strong theoretical roots (Granovetter, 1983) and empirical validations, among others on LinkedIn (Rajkumar et al., 2022). A large network of weak ties allows for informational dissemination and leveraging resources. For similar reasons as for attentional resources, we log-transform it.

4.4.3 Independent variables

We have three sets of independent variables: normative legitimacy claims, cognitive legitimacy claims, and pragmatic legitimacy claims.

Normative legitimacy refers to the extent to which a firm's actions align with societal norms and values. Enterprises, and especially new ventures, are increasingly expected to make efforts toward sustainability, for instance by contributing to the Sustainable Development Goals (SDGs).

We categorize a post as a normative legitimacy claim when it relates to one of the 17 SDGs. Although such posts do not necessarily imply commitment of the startup, they are still indicative of the entrepreneur's engagements, interests, and beliefs. Our sample being too big to use manual qualitative analysis, we leverage the LLM OSS-20B, an open LLM published by OpenAI with strength on par with GPT 4o. Following (Chew et al., 2023), we build the codebook with the 17 SDGs and their description, and write a prompt asking the LMM which of the codes are related to the post. We plot their repartition in Fig. 4.7. Then, for readability and for mitigating SDGs with few related posts, we group SDG 6 (clean water and sanitation), SDG 7 (affordable and clean energy), SDG 12 (responsible consumption), SDG 13 (climate action), SDG 14 (life below water) and SDG 15 (life on land) under the label green SDG; SDG 1 (no poverty), SDG 2 (zero hunger), SDG 3 (good health), SDG 5 (gender equality), SDG 10 (reduced inequality) and SDG 16 (peace and justice) under the label social sdg; and we group the SDG 8 (decent work), SDG 9 (industry and innovation), SDG 11 (sustainable cities) and SDG 17 (partnerships) as economic. The categorisation of some of those SDGs might be discussed, for instance the SDG 2 (zero hunger), since many startups are proposing services related to precision agriculture, which might relate more to the green than to the social pillar of sustainability, but changing it did not lead to a significant change in results.

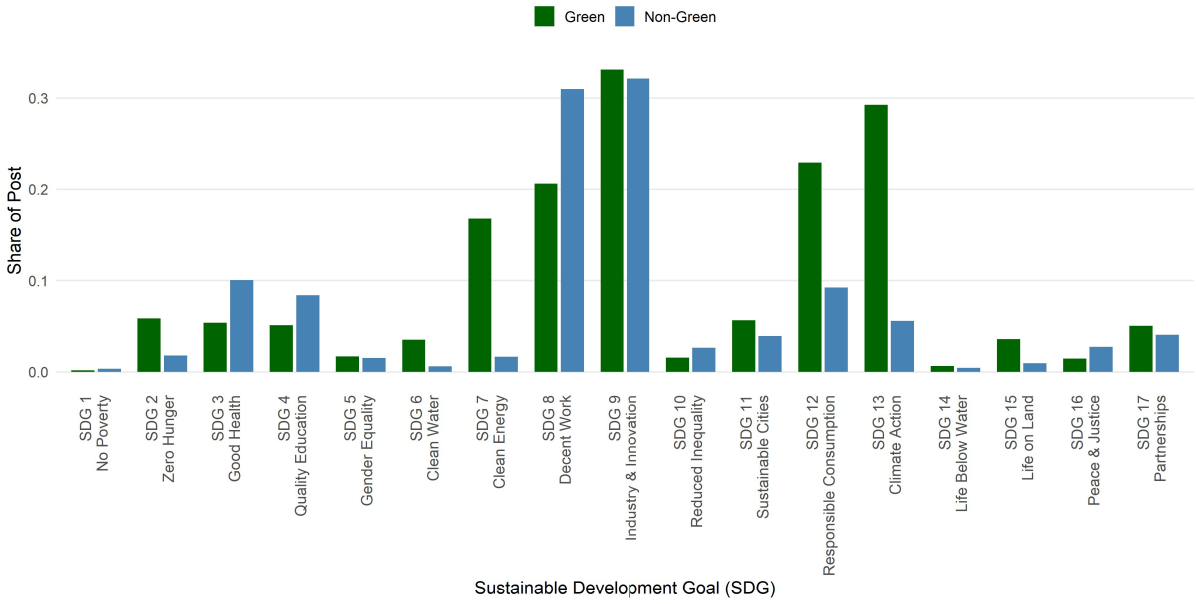


Fig. 4.7 Repartition of SDG legitimacy claims in posts

Cognitive legitimacy refers to the extent to which a firm is understandable to its audience. A firm establishes cognitive legitimacy by clearly articulating what it does, what it offers, and how it compares to incumbent firms. Accordingly, we classify a post as a cognitive legitimacy claim when

it describes the startup's products or services, its activities, or its similarities to or differences from the market.

A firm establishes pragmatic legitimacy by emphasizing the practical benefits it provides to its stakeholders. Startup stakeholders are typically eager for innovation and strongly invested in the firm's success. Accordingly, we classify a post as making a pragmatic legitimacy claim if it highlights technological innovation, firm achievements, financial performance, or financial support.We prompted the LLM to know which of the posts are related to one of those categories. Although we could have run two different prompts, we did only one for both cognitive and pragmatic legitimacy to save computing time since the number of categories remains small. The full prompt and codebooks are in Appendix D, and we show their repartition in Fig. 4.8

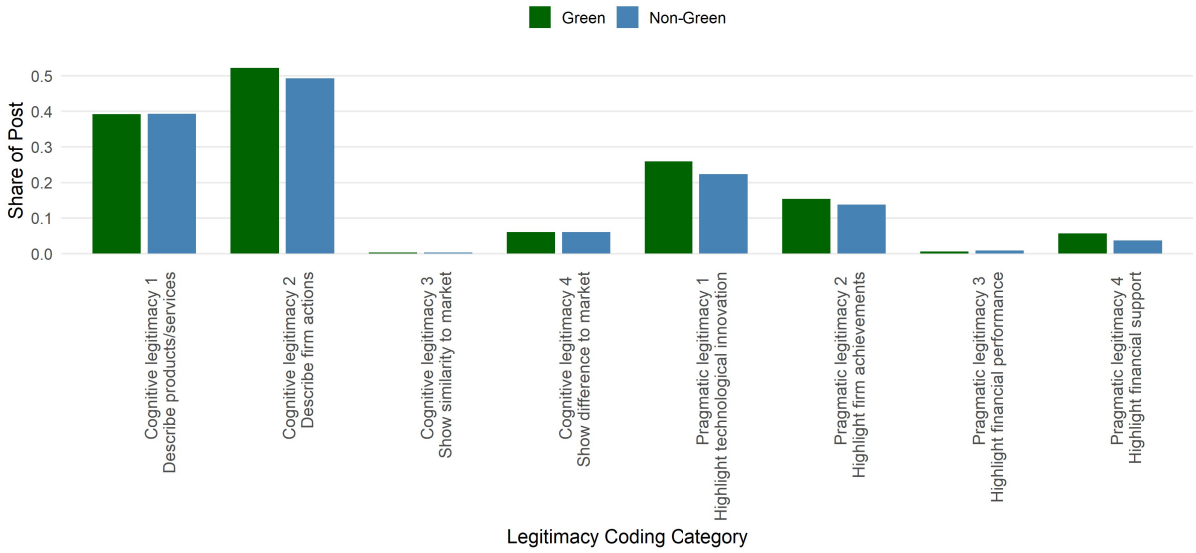


Fig. 4.8 Repartition of cognitive and pragmatic legitimacy claims in posts

4.4.4 Control variables

Startup/entrepreneur level controls: We include control variables at the level of the startup and of the entrepreneur. We control for the age of the company and its size using the number of employees of the company on LinkedIn.

Since entrepreneurs might leverage their network established in prior experiences, we use the BrightData dataset to control for it. It contains the data that entrepreneurs filled on their LinkedIn profiles, including a field with the start and end dates. We sum all the experiences' start and end dates to get the total length. To account for a few extremely high and unrealistic values, we cap the maximum experience length at 50 years. In a similar way, we use the education field to get the education level of the entrepreneurs. To do so, we explore the most frequent words in both of those fields, and we extract the relevant keywords to build dictionaries for PhD, master's, bachelor and sub-degree categories. We proceed in a similar way to identify their educational field, building a dictionary for the STEM and Economics categories. Further, we proxy for the entrepreneur's gender. Since the gender variable is not readily available in our databases, we apply the Python

package genderguesser to the names of the entrepreneurs, extracting the most frequent gender associated with a name.

Post control: Finally, we follow other studies on online legitimacy to control for variables that might have an influence on the success of the posts (Antretter et al., 2019; Zhao et al., 2023). The post's length, as the number of characters, is informative of the quantity of information in the post. We proxied for the reading ease using the Flesh Reading Ease score. We also control for sentiment, which has been shown to have an important impact on posts' reaction in many studies, using the multilingual model cardiffnlp/twitter-xlm-roberta-base-sentiment. We also use the data from Apify to control whether the post is a repost.

4.4.5 Endogeneity concerns

The main source of endogeneity in our study stems from simultaneity and reverse causality. While our theoretical framework assumes that legitimacy claims embedded in narratives drive attentional resources (measured through reactions), the reverse relationship is also plausible. Entrepreneurs may adapt their posting behavior based on the attention their previous posts receive, publishing more—or differently—depending on past engagement. This creates a feedback loop between content and attention that may bias our estimates.

In addition, attentional resources are closely intertwined with network size, giving rise to a second source of endogeneity. On the one hand, a larger network mechanically increases the potential reach of posts and thus the number of reactions they can generate. On the other hand, higher attentional resources can contribute to network growth by increasing visibility and attracting new followers. As a result, attentional resources and network size are both drivers and outcomes of each other, reinforcing simultaneity concerns across our dependent variables.

Our empirical strategy does not fully resolve these endogeneity issues due to its observational and time-invariant caractere. Therefore, our findings should be interpreted as associational rather than strictly causal. We highlight this limitation explicitly and encourage future research to address it using longitudinal variables.

4.5 RESULTS

Descriptive statistics

Table 4.2 provides descriptive statistics of the variables at the post levels. Very few posts consist of legitimacy claim C3 (show similarity to market) and P3 (highlight financial support), so we do not use them in further analysis. We explore the correlation matrix for the remaining variables. Most of the correlations remain low, but the highest, between C1 (Describe products/services) and P1 (Highlight technological innovation) is 0.48, so we run a VIF test on our regression.

Table 4.2 Descriptive statistics of the posts

Statistic	N	Mean	St. Dev.	Min	Max

Attentional ressources (log number of reactions)	60,781	2.39	1.35	0.00	7.20
N1: Green SDG	60,781	0.17	0.38	0	1
N2: Social SDG	60,781	0.24	0.43	0	1
N3: Economic SDG	60,781	0.52	0.50	0	1
C1: Describe products/services	60,781	0.39	0.49	0	1
C2: Describe firm actions	60,781	0.50	0.50	0	1
C3: Show similarity to market	60,781	0.003	0.06	0	1
C4: Show difference to market	60,781	0.06	0.24	0	1
P1: Highlight technological innovation	60,781	0.23	0.42	0	1
P2: Highlight firm achievements	60,781	0.14	0.35	0	1
P3: Highlight financial performance	60,781	0.01	0.09	0	1
P4: Highlight financial support	60,781	0.04	0.19	0	1
Negative sentiment	60,781	0.07	0.26	0	1
Positive sentiment	60,781	0.41	0.49	0	1
Is a repost	60,781	0.40	0.49	0	1
Post length	60,781	542.72	539.81	1	3,000
Flesch reading ease score	60,781	27.08	34.34	-1,486.18	121.22
Number of links	60,781	0.27	0.65	0	40
Number of hashtags	60,781	3.41	4.74	0	99

Similarly, Table 4.3 resents the descriptive statistics at the entrepreneur level. For the same reason as above, we log the average number of reactions and the number of followers. We set all the post-related variables at 0 for the entrepreneurs who did not publish any post. We ran a correlation matrix, which shows correlations higher than 0.8 between the variables N3, C1, C2, and P1. Those high correlations suggest possible issues of multicollinearity, which we investigate using the VIF test on our regressions.

Table 4.3 Descriptive statistics of the entrepreneurs

Statistic	N	Mean	St. Dev.	Min	Max

Network size (log number of followers)	1,703	6.51	1.38	0.00	11.59
N1: Green SDG (count)	1,703	6.23	13.18	0	100
N2: Social SDG (count)	1,703	8.54	14.28	0	100
N3: Economic SDG (count)	1,703	18.49	21.83	0	97
C1: Describe products/services (count)	1,703	14.07	17.13	0	92
C2: Describe firm actions (count)	1,703	17.82	20.09	0	87
C4: Show difference to market (count)	1,703	2.17	3.81	0	38
P1: Highlight technological innovation (count)	1,703	8.15	12.75	0	92
P2: Highlight firm achievements (count)	1,703	4.98	6.63	0	62
P4: Highlight financial support (count)	1,703	0.04	0.08	0.00	1.00
Attentional ressources (log number of reactions)	1,703	2.41	1.35	0.00	6.57
Share of positive sentiment	1,703	0.37	0.26	0.00	1.00
Share of negative sentiment	1,703	0.05	0.10	0.00	1.00
Share of repost	1,703	0.37	0.31	0.00	1.00
Average number of links	1,703	0.23	0.27	0.00	4.00
Average number of hashtags	1,703	2.56	2.43	0.00	18.13
Average Flesch Reading Ease Score	1,703	24.62	18.02	-132.58	97.03
Average post length	1,703	428.82	315.78	0.00	2,267.00
Startup age	1,703	6.15	2.96	1.00	15.00
Number of employees	1,703	77.84	591.66	0	6,487
Entrepreneur is a male	1,703	0.73	0.44	0	1
Education level	1,703	1.71	1.31	0	4
Education STEM	1,703	0.27	0.44	0	1
Education economics	1,703	0.31	0.46	0	1
Experience length	1,703	20.00	13.70	0.00	50.00

Regressions

Table 4.4 presents our first analysis, in which we ran an ordinary least squares (OLS) regression at the post level, with fixed effects since we recorded up to 100 posts for each entrepreneur. The dependent variable that we use is attentional resources, proxied by the number of reactions to a post. We run one regression for the whole set of startups, one for the green startups, and one for the non-green startups to compare the effect of the independent variables for each sample. We do it for four sets of regression: the first only with the normative legitimacy claims, the second with the cognitive legitimacy claim variables, the third with the pragmatic legitimacy claims variables, and finally with all legitimacy claims as independent variables.

The complete estimated equation is:

$$Reactions_e = \beta_0 + \beta_1 Normative_i + \beta_2 Cognitive_i + \beta_3 Pragmatic_i + \gamma \boldsymbol{X}_{ij} + \alpha_j + \varepsilon_{ij}$$

where i indexes posts and e indexes entrepreneurs, $Normative_e$, $Cognitive_e$ and $Attention_e$ are vectors of boolean variables respectively assessing wether a post constitutes a normative, cognitive or pragmatic legitimacy claims along their different dimensions, Xi is a vector of control variables at the post level, αj denotes entrepreneur fixed effects, and εij is the error term. β_1 corresponds to hypothese H6, β_2 to H5 and β_3 to H7.

The Breusch-Pagan test rejects the null hypothesis of homoscedasticity for some of the regressions, so we use robust standard errors. For each regression, we also run an OLS regression without fixed effects to perform a VIF test. All VIF values are lower than 2, which suggests no multicollinearity issues (Stock & Watson, 2007).

Hypothesis 5 stated that cognitive legitimacy claims attract more attentional resources. This is partly supported for both green and non-green firms: C2 (describe firm actions) has a positive and significant impact on the number of reactions, but C1 (describe product/services) has no significant impact, and C4 (show difference to maket) loose its significance when adding the other dependant variables.

Hypothesis 6 stated that normative legitimacy claims attract more attentional resources. This claim is supported for the non-green firms: the three types of normative legitimacy claims have a positive and significant effect on the number of reactions, but only weakly significant for green SDG claims. In contrast, none of those variables are significant for green startups. It suggests that their normative legitimacy is already established, and that legitimacy claims on LinkedIn are redundant.

Hypothesis 7 stated that pragmatic legitimacy claims attract more attention. This is supported by our models for both green and non-green firms, with a positive and significant impact for P2 (highlight firms achievement) and P4 (highlight financial support).

As an illustration, regression (12) shows that, among non-green startups, including a legitimacy claim related to a social SDG is associated with an increase of approximately 3.6% in the number of reactions to a post, holding other factors constant; a post describing the firm actions would lead to an increase of 6.4% of reactions; highlighting firms achievement would have a much stronger impact of 27.3% on the number of reactions.

Table 4.4 OLS regressions with fixed effect at the post level[16]

	Dependent variable: Attentional ressources											
	all	green	nongreen	all	green	nongreen	all	green	nongreen	all	green	nongreen
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
N1: Green SDG	0.026**	0.008	0.028**							0.017	-0.023	0.023*
	(0.013)	(0.028)	(0.014)							(0.013)	(0.029)	(0.014)
N2: Social SDG	0.035***	0.028	0.036***							0.035***	0.033	0.036***
	(0.011)	(0.035)	(0.011)							(0.011)	(0.035)	(0.011)
N3: Economic SDG	0.051***	0.023	0.055***							0.027***	-0.004	0.032***
	(0.009)	(0.025)	(0.009)							(0.009)	(0.025)	(0.009)
C1: Describe products/services				-0.007	0.040	-0.012				-0.012	0.045	-0.018*

[16] To understand how much the independent variables contributed to the R^2, we ran the list of regressions without the control variables. The values of the adjusted R^2 in each regression ranged between 0.379 and 0.430. However, running regressions only with the controls shows a very strong explanatory power, with R^2 between 0.545 and 0.556. This is mostly driven by the length of the posts and the repost dummy.

	(0.009)	(0.025)	(0.009)				(0.009)	(0.027)	(0.010)
C2: Describe firm actions	0.096***	0.107***	0.095***				0.065***	0.082***	0.064***
	(0.008)	(0.025)	(0.009)				(0.008)	(0.025)	(0.009)
C4: Show difference to market	0.046***	0.079*	0.042**				0.024	0.062	0.020
	(0.016)	(0.047)	(0.017)				(0.016)	(0.047)	(0.017)
P1: Highlight technological innovation				0.005	0.023	0.002	-0.008	-0.013	-0.008
				(0.010)	(0.028)	(0.011)	(0.011)	(0.031)	(0.012)
P2: Highlight firm achievements				0.282***	0.257***	0.285***	0.269***	0.243***	0.273***
				(0.011)	(0.031)	(0.012)	(0.012)	(0.031)	(0.012)
P4: Highlight financial support				0.108***	0.139***	0.101***	0.096***	0.137***	0.087***
				(0.022)	(0.050)	(0.024)	(0.022)	(0.050)	(0.024)

Negative sentiment	0.033*	-0.048	0.043**	0.048**	-0.015	0.056***	0.045**	-0.025	0.054***	0.056***	-0.006	0.064***
	(0.019)	(0.058)	(0.020)	(0.019)	(0.058)	(0.020)	(0.018)	(0.058)	(0.019)	(0.018)	(0.058)	(0.019)
Positive sentiment	0.317***	0.237***	0.327***	0.315***	0.238***	0.325***	0.288***	0.214***	0.297***	0.288***	0.217***	0.297***
	(0.009)	(0.025)	(0.009)	(0.009)	(0.025)	(0.009)	(0.009)	(0.026)	(0.009)	(0.009)	(0.026)	(0.009)
Is a repost	-1.050***	-1.058***	-1.048***	-1.047***	-1.051***	-1.047***	-1.055***	-1.058***	-1.054***	-1.048***	-1.052***	-1.048***
	(0.010)	(0.029)	(0.011)	(0.010)	(0.029)	(0.011)	(0.010)	(0.028)	(0.011)	(0.010)	(0.029)	(0.011)
Flesh reading ease score	-0.0001	0.001*	-0.0002	-0.0002	0.001*	-0.0003**	-0.0001	0.001**	-0.0002*	-0.0001	0.001*	-0.0002*
	(0.0001)	(0.0004)	(0.0001)	(0.0001)	(0.0004)	(0.0001)	(0.0001)	(0.0004)	(0.0001)	(0.0001)	(0.0004)	(0.0001)
Number of links	-0.019*	-0.010	-0.021*	-0.019*	-0.010	-0.022*	-0.020*	-0.009	-0.023*	-0.021*	-0.010	-0.024*
	(0.011)	(0.021)	(0.012)	(0.011)	(0.022)	(0.012)	(0.011)	(0.021)	(0.013)	(0.011)	(0.021)	(0.013)
Count of hashtags	-0.002**	-0.005*	-0.002*	-0.002**	-0.006**	-0.002*	-0.002*	-0.005*	-0.001	-0.002**	-0.005**	-0.002*
	(0.001)	(0.003)	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)	(0.003)	(0.001)
Length of the posts	0.0005***	0.001***	0.0005***	0.0005***	0.0005***	0.0005***	0.0004***	0.0005***	0.0004***	0.0004***	0.0004***	0.0004***

	(0.00001)	(0.00003)	(0.00001)	(0.00001)	(0.00003)	(0.00001)	(0.00001)	(0.00003)	(0.00001)	(0.00001)	(0.00003)	(0.00001)
Observations	60,781	7,053	53,728	60,781	7,053	53,728	60,781	7,053	53,728	60,781	7,053	53,728
R^2	0.566	0.556	0.567	0.567	0.558	0.568	0.571	0.561	0.572	0.572	0.562	0.572
Adjusted R^2	0.556	0.545	0.556	0.556	0.546	0.557	0.560	0.549	0.561	0.561	0.550	0.562

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 4.5 presents our second analysis, in which we ran OLS regression at the level of the entrepreneur, with network size as the dependent variable. We started by running a first set of regression with all the legitimacy claims and the attentional ressources as dependent variables. However, the VIF test returned values higher than 10, suggesting multicollinearity issues. We ran a second set of regression after removing the variable N3 (economic SDG), which had the highest VIF value and which already had many high correlations in the correlation matrix. This improves the test of the VIF tests, which show values under 10 for each remaining variable, except for C1 (Describe products/services). We tried to remove C1 in a third set of regressions, and the results led to the same conclusion than the second set. For clarity, we show the second set of regressions in Table 4.4 so that the variables are as similar as possible as the ones that we used at the post level.

The estimated equation is:

$$NetworkSize_e = \beta_0 + \beta_1 Normative_e + \beta_2 Cognitive_e + \beta_3 Pragmatic_e + \beta_4 Attention_e + \gamma \mathbf{X}_e + \varepsilon_e$$

where e indexes entrepreneurs, $Normative_e$, $Pragmatic_e$, and $Cognitive_e$ are vectors of variables counting respectively the number of normative, cognitive and pragmatic legitimacy claims along their different dimensions, $Attention_e$ captures attentional resources, $\mathbf{X}_e$ is a vector of entrepreneur-level control variables, and ε_e is the error term. As in the first analysis, we estimate separate specifications including subsets of legitimacy claims (normative, cognitive, and pragmatic) as well as a full model. β_1 corresponds to hypothese H3, β_2 to H2, β_3 to H4 and β_4 to H8.

Hypothesis 2 stated that entrepreneurs sharing many cognitive legitimacy claims have bigger networks. This is supported in regression (4) to (6): the coefficients for C2 is significant and positive for the all firms, while C1 is significant only for the green firms. C4 (show difference to market) is insignificant.

Hypothesis 3 stated that entrepreneurs sharing many normative legitimacy claims have bigger networks. This is supported by regression (1) to (3): claims related to green SDGs (N1) have a positive impact for all startups, and claims related to social SDGs have a positive impact for non green firms.

Hypothesis 4 stated that entrepreneurs sharing many pragmatic legitimacy claims have bigger networks. This is supported for the non-green startups, with P1 (highlight technical innovation) and P2 (highlight firm achievements) and P4 (highlight financial support) positive and significant for non-green firms in regression (7) to (9).

Regression (13) to (15) show a positive impact of attentional ressources on legitimacy claims, and lower effects of all legitimacy claims on network size (with the exeption of C1). Thus, hypothesis 8, stating attentional resources mitigate the effect of legitimacy claims on networks, is supported. As an illustration, regression (11) indicates that a 1% increase in reactions to a startup's posts is associated with an approximate 0.43% increase in the number of its followers, holding other factors constant.

Finally, hypothesis 1 stated that the impact of legitimacy claims on attentional resources and network differs between green and non-green firms. We showed that normative legitimacy claims had a positive and significant impact on attentional resources for non-green firms but not for green firms, and similarly, that cognitive legitimacy claims have a positive impact only on non-green firms network size. Those result support the hypothesis.

Table 4.5 OLS regressions at the entrepreneurial level

	Dependent variable: Network size														
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
N1: Green SDG	0.014***	0.016***	0.016***										-0.001	0.010**	-0.003
	(0.002)	(0.004)	(0.003)										(0.002)	(0.005)	(0.003)
N2: Social SDG	0.019***	0.005	0.020***										0.001	0.003	0.002
	(0.002)	(0.008)	(0.002)										(0.002)	(0.007)	(0.002)
C1: Describe products/services				0.010***	0.001	0.011***							0.019***	0.012	0.019***
				(0.003)	(0.012)	(0.004)							(0.004)	(0.013)	(0.004)
C2: Describe firm actions				0.018***	0.023**	0.018***							0.009***	0.008	0.008***
				(0.003)	(0.009)	(0.003)							(0.003)	(0.009)	(0.003)
C4: Show difference to market				-0.001	-0.020	0.002							0.001	-0.001	0.005
				(0.009)	(0.020)	(0.010)							(0.009)	(0.021)	(0.010)
P1: Highlight technological innovation							0.012***	0.006	0.013***				-0.009***	-0.020	-0.008**
							(0.003)	(0.011)	(0.003)				(0.003)	(0.013)	(0.004)
P2: Highlight firm achievements							0.045***	0.029	0.047***				0.006	-0.009	0.007
							(0.005)	(0.022)	(0.006)				(0.006)	(0.018)	(0.007)
P4: Highlight financial support							0.019*	0.017	0.025*				0.011	0.017	0.014
							(0.012)	(0.031)	(0.013)				(0.011)	(0.029)	(0.012)

Attentional ressources										0.446***	0.476***	0.476***	0.350***	0.424***	0.337***
										(0.035)	(0.095)	(0.095)	(0.033)	(0.097)	(0.035)
Share of positive sentiment	0.535***	0.963***	0.522***	0.494***	0.952**	0.460***	0.410***	1.007**	0.357***	-0.179	0.006	0.006	-0.113	0.085	-0.118
	(0.126)	(0.373)	(0.132)	(0.120)	(0.380)	(0.126)	(0.124)	(0.410)	(0.128)	(0.132)	(0.426)	(0.426)	(0.123)	(0.397)	(0.128)
Share of negative sentiment	0.909***	3.473***	0.732**	1.147***	3.397***	1.002***	1.351***	3.794***	1.174***	0.994***	3.153***	3.153***	0.868***	2.845***	0.752***
	(0.312)	(0.977)	(0.316)	(0.299)	(0.941)	(0.304)	(0.324)	(0.967)	(0.329)	(0.298)	(0.852)	(0.852)	(0.266)	(0.908)	(0.273)
Share of repost	0.007	0.081	-0.023	0.045	0.091	0.030	0.062	0.108	0.041	0.214**	0.176	0.176	0.206**	0.147	0.198**
	(0.099)	(0.319)	(0.102)	(0.095)	(0.312)	(0.098)	(0.097)	(0.331)	(0.099)	(0.097)	(0.308)	(0.308)	(0.091)	(0.285)	(0.095)
Average number of links	0.183	0.252	0.180	0.159	0.333	0.152	0.191	0.341	0.185	0.058	0.326	0.326	0.044	0.203	0.036
	(0.134)	(0.401)	(0.138)	(0.118)	(0.381)	(0.121)	(0.128)	(0.377)	(0.132)	(0.123)	(0.335)	(0.335)	(0.105)	(0.308)	(0.110)
Count of hashtags	0.010	0.007	0.011	-0.015	-0.005	-0.014	0.012	0.011	0.015	0.015	0.041	0.041	-0.025*	0.013	-0.025*
	(0.016)	(0.047)	(0.018)	(0.015)	(0.048)	(0.016)	(0.016)	(0.050)	(0.017)	(0.015)	(0.044)	(0.044)	(0.014)	(0.044)	(0.015)
Average flesch reading ease score	0.008***	0.006	0.009***	0.006***	0.004	0.006***	0.007***	0.004	0.008***	0.002	0.003	0.003	0.003	0.004	0.003
	(0.002)	(0.004)	(0.002)	(0.002)	(0.004)	(0.002)	(0.002)	(0.005)	(0.002)	(0.002)	(0.005)	(0.005)	(0.002)	(0.004)	(0.002)
post length	0.0004***	0.0003	0.0003**	0.0002*	0.0003	0.0002	0.0002	0.0003	0.0001	-0.0002*	-0.001	-0.001	-0.0004***	-0.001**	-0.0004***
	(0.0001)	(0.0004)	(0.0001)	(0.0001)	(0.0004)	(0.0001)	(0.0001)	(0.0004)	(0.0001)	(0.0001)	(0.0003)	(0.0003)	(0.0001)	(0.0003)	(0.0001)
Startup age	-0.012	-0.029	-0.010	-0.017*	-0.031	-0.014	-0.011	-0.022	-0.009	-0.009	-0.014	-0.014	-0.011	-0.022	-0.009
	(0.010)	(0.027)	(0.011)	(0.010)	(0.028)	(0.010)	(0.010)	(0.028)	(0.011)	(0.010)	(0.025)	(0.025)	(0.009)	(0.027)	(0.010)
Number of employees	0.00005	-0.003	0.00004	0.0001*	-0.002	0.0001*	0.0001*	-0.002	0.0001*	0.0001	-0.001	-0.001	0.0001*	-0.002	0.0001*
	(0.00005)	(0.004)	(0.00005)	(0.00004)	(0.004)	(0.00004)	(0.00005)	(0.005)	(0.00005)	(0.00004)	(0.003)	(0.003)	(0.00004)	(0.003)	(0.00004)
Entrepreneur is a	0.067	0.015	0.085	0.008	-0.022	0.014	-0.035	-0.084	-0.024	0.032	-0.044	-0.044	0.018	0.019	0.022

male															
	(0.063)	(0.181)	(0.067)	(0.060)	(0.176)	(0.064)	(0.062)	(0.189)	(0.067)	(0.062)	(0.170)	(0.170)	(0.059)	(0.167)	(0.063)
Education level	0.135***	0.114*	0.139***	0.140***	0.106	0.140***	0.129***	0.088	0.129***	0.123***	0.060	0.060	0.124***	0.102	0.124***
	(0.023)	(0.066)	(0.025)	(0.022)	(0.067)	(0.023)	(0.023)	(0.066)	(0.025)	(0.023)	(0.061)	(0.061)	(0.021)	(0.062)	(0.023)
Education STEM	0.082	0.296*	0.048	0.076	0.216	0.054	0.024	0.266	-0.007	-0.018	0.210	0.210	0.060	0.217	0.039
	(0.061)	(0.160)	(0.066)	(0.058)	(0.159)	(0.062)	(0.059)	(0.163)	(0.064)	(0.060)	(0.164)	(0.164)	(0.056)	(0.167)	(0.061)
Education economics	0.461***	0.201	0.483***	0.468***	0.253	0.493***	0.518***	0.319*	0.543***	0.491***	0.434***	0.434***	0.445***	0.292*	0.459***
	(0.061)	(0.177)	(0.065)	(0.058)	(0.174)	(0.061)	(0.060)	(0.177)	(0.064)	(0.060)	(0.161)	(0.161)	(0.056)	(0.172)	(0.060)
Experience length	0.019***	0.025***	0.018***	0.019***	0.025***	0.018***	0.021***	0.025***	0.020***	0.024***	0.027***	0.027***	0.020***	0.026***	0.019***
	(0.002)	(0.007)	(0.002)	(0.002)	(0.007)	(0.002)	(0.002)	(0.007)	(0.002)	(0.002)	(0.007)	(0.007)	(0.002)	(0.007)	(0.002)
Constant	4.829***	4.668***	4.837***	4.857***	4.708***	4.877***	4.882***	4.724***	4.899***	4.555***	4.385***	4.385***	4.633***	4.363***	4.667***
	(0.145)	(0.423)	(0.155)	(0.142)	(0.422)	(0.152)	(0.145)	(0.426)	(0.155)	(0.145)	(0.422)	(0.422)	(0.143)	(0.435)	(0.153)
Observations	1,703	200	1,503	1,703	200	1,503	1,703	200	1,503	1,703	200	200	1,703	200	1,503
R^2	0.311	0.384	0.309	0.370	0.395	0.371	0.340	0.368	0.343	0.340	0.420	0.420	0.420	0.465	0.419
Adjusted R^2	0.305	0.331	0.301	0.363	0.338	0.364	0.333	0.308	0.336	0.334	0.372	0.372	0.412	0.395	0.410

Note: *p<0.1; **p<0.05; ***p<0.01

4.6 DISCUSSION AND CONCLUSION

Contributions

The contribution of legitimacy to the entrepreneurial process has sparked high interest in the entrepreneurial literature (Lounsbury & Glynn, 2001; Martens et al., 2007; Navis & Glynn, 2010, 2011; Taeuscher et al., 2021, 2022), leading to a proliferation of qualitative studies and resulting in a rich yet fragmented field (Suddaby et al., 2017). To reconnect different facets of the literature, this paper set out to examine how cognitive, normative, and pragmatic legitimacy claims—the most frequently studied dimensions of legitimacy—contribute to the establishment of networks and online legitimacy.

Our study contributes mainly to the legitimacy literature. Our first contribution is methodological. By introducing the use of LLM to perform qualitative analysis, we quantify the impact of different dimensions of legitimacy, answering the call of researchers (Audretsch & Lehmann, 2023; Suddaby et al., 2017). This method can and should be applied to other contexts to test quantitatively the impact of the many established dimensions of legitimacy on different entrepreneurial success measures. Further, the method could be adapted to study narratives and other type of texts outside of the legitimacy field.

Our second contribution comes from the application of this method in our research context. While we show that normative, cognitive, and pragmatic legitimacy (Suchman, 1995) contribute to online legitimacy, not all their facet have an impact, challenging results established by qualitative research. For instance, a large part of the literature suggests that startups should balance familiarity and distinctiveness to reach legitimacy (Navis & Glynn, 2011). Researchers used narrative topic diversity using LDA to proxy for distinctiveness, finding a positive impact on legitimacy (Taeuscher et al., 2021, 2022). However, we argue that such an indirect measure might rather proxy for narrative's characteristics rather than venture's characteristics. Our research shows that entrepreneurs rarely share posts stating the similarity of their startup to other firms, and that stating the distinctiveness of their venture does not have a significant impact on online legitimacy. Nevertheless, this effect might hold in other contexts; therefore, our results call for quantitative testing across different settings.

Our third contribution is at the intersection of the legitimacy and network literature. Although networks, and specifically strong links with legitimizing actors, have been shown to have a positive impact on legitimacy, we show that the attention gathered by all types of legitimacy claims leads to a bigger LinkedIn network, as well as a direct effect of cognitive legitimacy claims. Thus, startups appear to leverage legitimacy to build weak links with other actors. This is an important finding for entrepreneurs, since network size and weak links open access to financial resources and contribute to entrepreneurial success (Banerji & Reimer, 2019).

Our fourth contribution relates to the green entrepreneurship literature. Legitimacy claims on LinkedIn appear to have a lesser impact on green entrepreneurs than on non-green entrepreneurs. For green entrepreneurs, only cognitive and pragmatic legitimacy claims have a positive impact on attentional ressources, while non-green entrepreneurs also profit from normative legitimacy claims. Further, for green entrepreneurs, only green legitimacy claims have a direct positive impact on network size, while cognitive legitimacy positively influences the network size of non-green entrepreneurs. This result might be due to the distinct expectations held by the audience of green startups (Castelló et al., 2016; O'Neil & Ucbasaran, 2016).

Generalizability and limitations

Our study focused on startups. Due to the lack of significance of the legitimacy items related to innovation and market differentiation, we expect it to generalize well to all types of young firms. However, we expect that cognitive legitimacy is less important for older firms and bigger firms, since they are established and might not need to introduce their business as thoroughly (Stinchcombe, 1965).

We leveraged LinkedIn data. Although it is the most used social network with a professional aim, exploring how entrepreneurs build legitimacy on other widely used social networks, such as Facebook, would be of interest. Different social networks have different audiences with different expectations and uses (Auxier & Anderson, 2021).

Further, we only used textual data from SNS posts. More and more narratives shared on SNS are multimodal (Gordo Molina et al., 2022). Also, our study is only rooted in the legitimacy literature, but other literature, such

as rhetoric, exploring the emotional responses to posts, could have been explored.

We gave insights into the process of entrepreneurs building online legitimacy online on social networking cites. It would be interesting to compare it to the building of offline legitimacy, for instance at startup networking events such as VivaTech or incubators. However, collecting data for quantitative analysis for such a context is very challenging.

A key limitation of this study lies in the absence of a systematic human validation of the LLM-based classification of legitimacy claims. While the proposed approach introduce the use of LLM for qualitative analysis, it does not incorporate inter-rater reliability assessment or expert annotation to verify classification consistency. Conducting such an evaluation would require a carefully designed coding protocol and substantial human effort, which was beyond the scope of the present study. Future research should address this limitation by implementing human validation procedures and reporting inter-rater reliability metrics (e.g., Cohen's kappa), thereby strengthening the robustness and credibility of the classification framework.

We compared the impact of legitimacy for green startups to non-green startups. It would have been interesting to explore other types of startups, such as social startups. Although research sometimes uses the terms social, sustainable, and green startups interchangeably, the literature firmly establishes that they should be studied separately (Bonfanti et al., 2024; Halberstadt et al., 2024). Since the SDGs that we used to build our normative legitimacy variable have an environmental, a social, and an economic pillar, it would have been interesting to explore the impact of social normative legitimacy claims on social startups.

4.7 REFERENCES

Stinchcombe, A. L. (1965). Social structures and organisation: Handbook of Organization (James G. March, pp. 142–193). Rand McNally.

Suchman, M. C. (1995). Managing Legitimacy: Strategic and Institutional Approaches. The Academy of Management Review, 20(3), 571–610. https://doi.org/10.2307/258788

Gordo Molina, V., Díez Martín, F., & Del Castillo Feito, C. (2022). Legitimacy in entrepreneurship. Intellectual structure and research trends. Cuadernos de Gestión, 22(1). https://doi.org/10.5295/cdg.211471vg

Castelló, I., Etter, M., & Årup Nielsen, F. (2016). Strategies of Legitimacy Through Social Media: The Networked Strategy. Journal of Management Studies, 53(3), 402–432. https://doi.org/10.1111/joms.12145

Ge, B., Jiang, D., Gao, Y., & Tsai, S.-B. (2016). The Influence of Legitimacy on a Proactive Green Orientation and Green Performance: A Study Based on Transitional Economy Scenarios in China. Sustainability, 8(12), 1344. https://doi.org/10.3390/su8121344

Soewarno, N., Tjahjadi, B., & Fithrianti, F. (2019). Green innovation strategy and green innovation: The roles of green organizational identity and environmental organizational legitimacy. Management Decision, 57(11), 3061–3078. https://doi.org/10.1108/MD-05-2018-0563

Lefsrud, L., Graves, H., & Phillips, N. (2020). “Giant Toxic Lakes You Can See from Space”: A Theory of Multimodal Messages and Emotion in

Legitimacy Work. Organization Studies, 41(8), 1055–1078. https://doi.org/10.1177/0170840619835575

Stuart, T. E., Hoang, H., & Hybels, R. C. (1999). Interorganizational Endorsements and the Performance of Entrepreneurial Ventures. Administrative Science Quarterly, 44(2), 315–349. https://doi.org/10.2307/2666998

Antretter, T., Blohm, I., Grichnik, D., & Wincent, J. (2019). Predicting new venture survival: A Twitter-based machine learning approach to measuring online legitimacy. Journal of Business Venturing Insights, 11, e00109. https://doi.org/10.1016/j.jbvi.2018.e00109

Ferro, T. (2015). The Importance of Publicly Available Social Networking Sites (SNSs) to Entrepreneurs. Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW '15, 917–928. https://doi.org/10.1145/2675133.2675300

Zhao, C., Liu, Z., & Zhang, C. (2023). Real or fictional? Digital entrepreneurial narratives and the acquisition of attentional resources in social entrepreneurship. Journal of Innovation & Knowledge, 8(3), 100387. https://doi.org/10.1016/j.jik.2023.100387

Taeuscher, K., Bouncken, R., & Pesch, R. (2021). Gaining Legitimacy by Being Different: Optimal Distinctiveness in Crowdfunding Platforms.

Academy of Management Journal, 64(1), 149–179. https://doi.org/10.5465/amj.2018.0620

Hallen, B. L., & Eisenhardt, K. M. (2012). Catalyzing Strategies and Efficient Tie Formation: How Entrepreneurial Firms Obtain Investment Ties. The Academy of Management Journal, 55(1), 35–70.

Higgins, M. C., & Gulati, R. (2003). Getting off to a Good Start: The Effects of Upper Echelon Affiliations on Underwriter Prestige. Organization Science, 14(3), 244–263.

Higgins, M. C., & Gulati, R. (2006). Stacking the Deck: The Effects of Top Management Backgrounds on Investor Decisions. Strategic Management Journal, 27(1), 1–25.

Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023). LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding (ArXiv:2306.14924). arXiv. https://doi.org/10.48550/arXiv.2306.14924

Audretsch, D. B., & Lehmann, E. E. (2023). Narrative entrepreneurship: bringing (his)story back to entrepreneurship. Small Business Economics, 60(4), 1593–1612. https://doi.org/10.1007/s11187-022-00661-2

Suddaby, R., Bitektine, A., & Haack, P. (2017). Legitimacy. Academy of Management Annals, 11(1), 451–478. https://doi.org/10.5465/annals.2015.0101

Aldrich, H. E., & Fiol, C. M. (1994). Fools Rush in? The Institutional Context of Industry Creation. The Academy of Management Review, 19(4), 645–670. https://doi.org/10.2307/258740

Scott, W. R. (1995). Institutions and organizations (Sage, Vol. 2).

Fisher, G., Kuratko, D. F., Bloodgood, J. M., & Hornsby, J. S. (2017). Legitimate to whom? The challenge of audience diversity and new venture legitimacy. Journal of Business Venturing, 32(1), 52–71. https://doi.org/10.1016/j.jbusvent.2016.10.005

Navis, C., & Glynn, M. A. (2011). Legitimate Distinctiveness and the Entrepreneurial Identity: Influence on Investor Judgments of New Venture Plausibility. The Academy of Management Review, 36(3), 479–499.

van Werven, R., Bouwmeester, O., & Cornelissen, J. P. (2015). The power of arguments: How entrepreneurs convince stakeholders of the legitimate distinctiveness of their ventures. Journal of Business Venturing, 30(4), 616–631. https://doi.org/10.1016/j.jbusvent.2014.08.001

Vossen, A., & Ihl, C. (2020). More than words! How narrative anchoring and enrichment help to balance differentiation and conformity of entrepreneurial products. Journal of Business Venturing, 35(6), 106050. https://doi.org/10.1016/j.jbusvent.2020.106050

Zamantılı Nayır, D., & Shinnar, R. S. (2020). How founders establish legitimacy: A narrative perspective on social entrepreneurs in a developing country context. Social Enterprise Journal, 16(3), 221–241. https://doi.org/10.1108/SEJ-10-2019-0073

UN. (2015). Transforming our world: the 2030 Agenda for Sustainable Development | Department of Economic and Social Affairs. https://sdgs.un.org/2030agenda

Ellerup Nielsen, A., & Thomsen, C. (2018). Reviewing corporate social responsibility communication: a legitimacy perspective. Corporate Communications: An International Journal, 23(4), 492–511. https://doi.org/10.1108/CCIJ-04-2018-0042

Lodhia, S., Kaur, A., & Kuruppu, S. C. (2022). The disclosure of sustainable development goals (SDGs) by the top 50 Australian companies: substantive or symbolic legitimation? Meditari Accountancy Research, 31(6), 1578–1605. https://doi.org/10.1108/MEDAR-05-2021-1297

Ebrahimi, P., Salamzadeh, A., Soleimani, M., Khansari, S. M., Zarea, H., & Fekete-Farkas, M. (2022). Startups and Consumer Purchase Behavior: Application of Support Vector Machine Algorithm. Big Data and Cognitive Computing, 6(2), 34. https://doi.org/10.3390/bdcc6020034

Rutherford, M. W., Tocher, N., Pollack, J. M., & Coombes, S. M. T. (2016). Proposing a Financial Legitimacy Threshold in Emerging Ventures: A Multi-Method Investigation. Group & Organization Management, 41(6), 751–785. https://doi.org/10.1177/1059601116669632

Fisher, G. (2020). The Complexities of New Venture Legitimacy. Organization Theory, 1(2), 2631787720913881. https://doi.org/10.1177/2631787720913881

Zimmerman, M. A., & Zeitz, G. J. (2002). Beyond Survival: Achieving New Venture Growth by Building Legitimacy. The Academy of Management Review, 27(3), 414–431. https://doi.org/10.2307/4134387

Tornikoski, E. T., & Newbert, S. L. (2007). Exploring the determinants of organizational emergence: A legitimacy perspective. Journal of Business Venturing, 22(2), 311–335. https://doi.org/10.1016/j.jbusvent.2005.12.003

Kibler, E., Salmivaara, V., Stenholm, P., & Terjesen, S. (2018). The evaluative legitimacy of social entrepreneurship in capitalist welfare systems. Journal of World Business, 53(6), 944–957. https://doi.org/10.1016/j.jwb.2018.08.002

Riandita, A., Broström, A., Feldmann, A., & Cagliano, R. (2022). Legitimation work in sustainable entrepreneurship: Sustainability ventures' journey towards the establishment of major partnerships. International Small Business Journal, 40(7), 904–929. https://doi.org/10.1177/02662426211056799

O'Neil, I., & Ucbasaran, D. (2016). Balancing "what matters to me" with "what matters to them": Exploring the legitimation process of environmental entrepreneurs. Journal of Business Venturing, 31(2), 133–152. https://doi.org/10.1016/j.jbusvent.2015.12.001

Truong, Y., & Nagy, B. G. (2021). Nascent ventures? green initiatives and angel investor judgments of legitimacy and funding. Small Business Economics, 57(4), 1801–1818.

Alvedalen, J., & Boschma, R. (2017). A critical review of entrepreneurial ecosystems research: towards a future research agenda. European Planning Studies. (world). https://www.tandfonline.com/doi/full/10.1080/09654313.2017.1299694

Hoang, H., & Antoncic, B. (2003). Network-based research in entrepreneurship. Journal of Business Venturing, 18(2), 165–187. https://doi.org/10.1016/S0883-9026(02)00081-2

van Burg, E., Elfring, T., & Cornelissen, J. P. (2022). Connecting content and structure: A review of mechanisms in entrepreneurs' social networks. International Journal of Management Reviews, 24(2), 188–209. https://doi.org/10.1111/ijmr.12272

Stam, W., Arzlanian, S., & Elfring, T. (2014). Social capital of entrepreneurs and small firm performance: A meta-analysis of contextual and methodological moderators. Journal of Business Venturing, 29(1), 152–173. https://doi.org/10.1016/j.jbusvent.2013.01.002

Stuart, T. E., & Sorenson, O. (2007). Strategic networks and entrepreneurial ventures. Strategic Entrepreneurship Journal, 1(3–4), 211–227. https://doi.org/10.1002/sej.18

Gebert-Persson, S., & Káptalan-Nagy, E. (2016). Legitimacy in the Business Network Context. In P. Thilenius, C. Pahlberg, & V. Havila (Eds.), Extending the Business Network Approach: New Territories, New Technologies, New Terms (pp. 301–314). Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-53765-2_17

Martens, M. L., Jennings, J. E., & Jennings, P. D. (2007). Do the Stories They Tell Get Them the Money They Need? The Role of Entrepreneurial Narratives in Resource Acquisition. The Academy of Management Journal, 50(5), 1107–1132.

Crespin-Mazet, F., & Dontenwill, E. (2012). Sustainable procurement: Building legitimacy in the supply network. Journal of Purchasing and Supply Management, Sustainable Procurement, 18(4), 207–217. https://doi.org/10.1016/j.pursup.2012.01.002

Haack, P., Pfarrer, M. D., & Scherer, A. G. (2014). Legitimacy-as-Feeling: How Affect Leads to Vertical Legitimacy Spillovers in Transnational Governance. Journal of Management Studies, 51(4), 634–666. https://doi.org/10.1111/joms.12071

Connelly, S., Bryant, M., & Sharp, L. (2020). Creating Legitimacy for Citizen Initiatives: Representation, Identity and Strategic Networking. Planning Theory & Practice, 21(3), 392–409. https://doi.org/10.1080/14649357.2020.1776892

Shepherd, D. A., Zacharakis, A., Shepherd, D. A., & Zacharakis, A. (2003). A New Venture's Cognitive Legitimacy: An Assessment by Customers. Journal of Small Business Management. (world). https://www.tandfonline.com/doi/abs/10.1111/1540-627X.00073

Bjornali, E. S., Giones, F., & Billstrom, A. (2017). Reveal or Conceal? Signaling Strategies for Building Legitimacy in Cleantech Firms. Sustainability, 9(10), 1815. https://doi.org/10.3390/su9101815

Chen, Y.-S., Wang, C., Chen, Y.-R., Lo, W.-Y., & Chen, K.-L. (2019). Influence of Network Embeddedness and Network Diversity on Green Innovation: The Mediation Effect of Green Social Capital. Sustainability, 11(20), Article 20. https://doi.org/10.3390/su11205736

Prashantham, S., & Madhok, A. (2023). Corporate-startup partnering: Exploring attention dynamics and relational outcomes in asymmetric settings. Strategic Entrepreneurship Journal, 17(4), 770–801. https://doi.org/10.1002/sej.1475

Garud, R., Schildt, H. A., & Lant, T. K. (2014). Entrepreneurial Storytelling, Future Expectations, and the Paradox of Legitimacy. Organization Science, 25(5), 1479–1492.

Lounsbury, M., & Glynn, M. A. (2001). Cultural entrepreneurship: stories, legitimacy, and the acquisition of resources. Strategic Management Journal, 22(6–7), 545–564. https://doi.org/10.1002/smj.188

Czarniawska, B. (1998). A narrative approach to organization studies. Sage Publ.

Shiller, R. J. (2017). Narrative Economics (Working Paper No. 23075). National Bureau of Economic Research. https://doi.org/10.3386/w23075

Rudrum, D. (2005). From Narrative Representation to Narrative Use: Towards the Limits of Definition. Narrative, 13(2), 195–204.

Navis, C., & Glynn, M. A. (2010). How New Market Categories Emerge: Temporal Dynamics of Legitimacy, Identity, and Entrepreneurship in Satellite Radio, 1990-2005. Administrative Science Quarterly, 55(3), 439–471.

Seigner, B. D. C., Milanov, H., Lundmark, E., & Shepherd, D. A. (2023). Tweeting like Elon? Provocative language, new-venture status, and audience engagement on social media. Journal of Business Venturing, 38(2), 106282. https://doi.org/10.1016/j.jbusvent.2022.106282

Taeuscher, K., Zhao, E. Y., & Lounsbury, M. (2022). Categories and narratives as sources of distinctiveness: Cultural entrepreneurship within and across categories. Strategic Management Journal, 43(10), 2101–2134. https://doi.org/10.1002/smj.3391

Dobers, P., & Springett, D. (2010). Corporate social responsibility: discourse, narratives and communication. Corporate Social

Responsibility and Environmental Management, 17(2), 63–69. https://doi.org/10.1002/csr.231

Fischer, E., & Rebecca Reuber, A. (2014). Online entrepreneurial communication: Mitigating uncertainty and increasing differentiation via Twitter. Journal of Business Venturing, 29(4), 565–583. https://doi.org/10.1016/j.jbusvent.2014.02.004

Olanrewaju, A.-S. T., Hossain, M. A., Whiteside, N., & Mercieca, P. (2020). Social media and entrepreneurship research: A literature review. International Journal of Information Management, 50, 90–110. https://doi.org/10.1016/j.ijinfomgt.2019.05.011

Wang, W., Liang, Q., Mahto, R. V., Deng, W., & Zhang, S. X. (2020). Entrepreneurial entry: The role of social media. Technological Forecasting and Social Change, 161, 120337. https://doi.org/10.1016/j.techfore.2020.120337

Banerji, D., & Reimer, T. (2019). Startup founders and their LinkedIn connections: Are well-connected entrepreneurs more successful? Computers in Human Behavior, 90, 46–52. https://doi.org/10.1016/j.chb.2018.08.033

Ashforth, B. E., & Mael, F. (1989). Social Identity Theory and the Organization. The Academy of Management Review, 14(1), 20–39. https://doi.org/10.2307/258189

Crammond, R., Omeihe, K. O., Murray, A., & Ledger, K. (2018). Managing knowledge through social media: Modelling an entrepreneurial approach for Scottish SMEs and beyond. Baltic Journal of Management, 13(3), 303–328. https://doi.org/10.1108/BJM-05-2017-0133

Ketonen-Oksi, S., Jussila, J. J., & Kärkkäinen, H. (2016). Social media based value creation and business models. Industrial Management & Data Systems, 116(8), 1820–1838. https://doi.org/10.1108/IMDS-05-2015-0199

Petkova, A. P., Rindova, V. P., & Gupta, A. K. (2013). No News Is Bad News: Sensegiving Activities, Media Attention, and Venture Capital Funding of New Technology Organizations. Organization Science, 24(3), 865–888.

Zhang, H., Wu, C., Xie, J., Rubino, F., Graver, S., Kim, C., Carroll, J. M., & Cai, J. (2024). When Qualitative Research Meets Large Language Model: Exploring the Potential of QualiGPT as a Tool for Qualitative Coding (ArXiv:2407.14925). arXiv. https://doi.org/10.48550/arXiv.2407.14925

Barros, C. F., Azevedo, B. B., Neto, V. V. G., Kassab, M., Kalinowski, M., Nascimento, H. A. D. do, & Bandeira, M. C. G. S. P. (2025). Large Language Model for Qualitative Research -- A Systematic Mapping

Study (ArXiv:2411.14473). arXiv. https://doi.org/10.48550/arXiv.2411.14473

Goyanes, M., Lopezosa, C., & Jordá, B. (2025). Thematic analysis of interview data with ChatGPT: designing and testing a reliable research protocol for qualitative research. Quality & Quantity. https://doi.org/10.1007/s11135-025-02199-3

Qiao, T., Walker, C., Cunningham, C., & Koh, Y. S. (2025). Thematic-LM: A LLM-based Multi-agent System for Large-scale Thematic Analysis. Proceedings of the ACM on Web Conference 2025, WWW '25, 649–658. https://doi.org/10.1145/3696410.3714595

UN. (2019, April 3). Global Environment Outlook 6. UNEP - UN Environment Programme. http://www.unep.org/resources/global-environment-outlook-6

Li, K., Mai, F., Shen, R., & Yan, X. (2021). Measuring Corporate Culture Using Machine Learning. The Review of Financial Studies, 34(7), 3265–3315. https://doi.org/10.1093/rfs/hhaa079

Mansouri, S., & Momtaz, P. P. (2022). Financing sustainable entrepreneurship: ESG measurement, valuation, and performance. Journal of Business Venturing, 37(6), 106258. https://doi.org/10.1016/j.jbusvent.2022.106258

Guo, L., Vargo, C., Pan, Z., Ding, W., & Ishwar, P. (2016). Big Social Data Analytics in Journalism and Mass Communication: Comparing Dictionary-Based Text Analysis and Unsupervised Topic Modeling. Journalism & Mass Communication Quarterly, 93. https://doi.org/10.1177/1077699016639231

Blei, D. M., Y. Ng, A., & I. Jordan, M. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.

Tiba, S., Van Rijnsoever, F. J., & Hekkert, M. P. (2021). Sustainability startups and where to find them: Investigating the share of sustainability startups across entrepreneurial ecosystems and the causal drivers of differences. Journal of Cleaner Production, 306, 127054. https://doi.org/10.1016/j.jclepro.2021.127054

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure (ArXiv:2203.05794). arXiv. https://doi.org/10.48550/arXiv.2203.05794

Granovetter, M. (1983). The Strength of Weak Ties: A Network Theory Revisited. Sociological Theory, 1, 201–233. https://doi.org/10.2307/202051

Rajkumar, K., Saint-Jacques, G., Bojinov, I., Brynjolfsson, E., & Aral, S. (2022). A causal test of the strength of weak ties. Science, 377(6612), 1304–1310. https://doi.org/10.1126/science.abl4476

Stock, J. H., & Watson, M. W. (2007). Econometrics. In Econometrics.

Auxier, B., & Anderson, M. (2021). A majority of Americans say they use YouTube and Facebook, while use of Instagram, Snapchat and TikTok is especially common among adults under 30.

Bonfanti, A., De Crescenzo, V., Simeoni, F., & Loza Adaui, C. R. (2024). Convergences and divergences in sustainable entrepreneurship and social entrepreneurship research: A systematic review and research agenda. Journal of Business Research, 170, 114336. https://doi.org/10.1016/j.jbusres.2023.114336

Halberstadt, J., Schwab, A.-K., & Kraus, S. (2024). Cleaning the window of opportunity: Towards a typology of sustainability entrepreneurs. Journal of Business Research, 171, 114386. https://doi.org/10.1016/j.jbusres.2023.114386

CHAPTER 5

GENERAL CONCLUSION

5 GENERAL CONCLUSION

This thesis makes a substantive contribution to the research on green entrepreneurship. It is structured around three interrelated research articles, each addressing a key gap of the green entrepreneurship literature. The first article aims to establish a robust and automated method for identifying green startups, thereby enabling a more rigorous study of green entrepreneurship. The second article builds on this classified sample to examine the dynamics underlying the emergence of green startups, exploring the factors and processes that drive their creation. The third article focuses on a subsample of these ventures to analyze how entrepreneurs establish legitimacy for their new ventures to attract attention and expand their networks.

In the first article, we apply state-of-the-art NLP methods to construct a green classification framework based on the environmental topics articulated in the Sustainable Development Goals (SDGs). We then adapt two approaches commonly used in the sustainability literature—a machine learning–based dictionary method and Latent Dirichlet Allocation (LDA)—to identify sustainable startups from the textual content of their websites. In addition, we introduce a third method leveraging BERTopic to capture more nuanced thematic structures. We show that these three approaches identify overlapping yet distinct subsets of startups, highlighting the methodological implications of different classification strategies. As a showcase application, we compare the population of Italian startups with the priorities of the National Recovery and Resilience Plan (NRRP), assessing the extent to which entrepreneurial activity aligns with policy efforts to address the previously identified green topics. This article introduce new tools for researcher exploring green entrepenreurship.

In the second article, we examine the emergence of green startups through the lens of the Knowledge Spillover Theory of Entrepreneurship (KSTE). We extend the KSTE framework by incorporating a green demand component, arguing and showing that environmentally oriented demand acts as a catalyst for entrepreneurial activity by softening the knowledge filter and facilitating the commercialization of existing knowledge. Our findings further indicate that the size of the knowledge base matters more than its specific composition, suggesting that entrepreneurs recombine both green and non-green knowledge to generate green innovations. By highlighting the critical role of demand within the KSTE framework, this article broadens the theory’s

explanatory scope and opens new avenues for research on the demand-side drivers of entrepreneurship.

In the third article, we examine the process and effects of entrepreneurial legitimacy work. We develop a novel methodology leveraging large language models (LLMs) to systematically identify legitimacy claims embedded in entrepreneurial narratives. We apply this approach to LinkedIn posts published by entrepreneurs and analyze how different forms of legitimacy claims shape audience responses. Our findings show that such claims help entrepreneurs attract attentional resources, which in turn facilitate the expansion of their professional networks. Beyond its substantive insights, the article makes a methodological contribution by demonstrating how LLM-based tools can be used to study narratives and legitimacy at scale, thereby opening new avenues for research on entrepreneurial communication and legitimation processes.

Overall, this thesis contributes novel methods and insights to the green entrepreneurship literature. The second article introduces and compares three NLP methodologies, while the third article demonstrates the use of LLMs for qualitative analysis, highlighting approaches that can be readily adapted to other research domains. In the context of the rapid development of NLP, future research can increasingly leverage these tools to both open new avenues of inquiry and accelerate the process of knowledge creation.

The second and third articles highlight the complexity of green entrepreneurship and the importance of studying it as a distinct field of research. Our findings show that green and conventional startups benefit differently from green demand and legitimacy claims. Moreover, some results are counterintuitive: startup emergence appears to benefit equally from green and non-green knowledge stocks, and non-green startups attract more attention from normative legitimacy claims than their green counterparts. These findings underscore the need for further research to deepen our understanding of the mechanisms shaping green entrepreneurship.

One particularly important direction for future research concerns the role of high-growth ventures in green entrepreneurship. Although green entrepreneurship aims to improve society's environmental footprint, conventional entrepreneurship may not be sufficient to address the escalating climate crisis. Most startups struggle to overcome the liability of newness and establish themselves as viable, legitimate enterprises. Only a

small fraction survive, and an even smaller subset achieves rapid scaling and substantial market impact. Among these high-growth ventures—often referred to as “unicorns” when they reach valuations above one billion dollars—are the firms with the capacity to transform industries, reshape consumption patterns, and drive systemic change. Given the scale and urgency of climate challenges, further works could devote attention toward understanding and supporting the emergence and success of high-growth green startups, as their ability to scale quickly may generate disproportionately large environmental benefits.

In the introduction, we emphasized the importance of leveraging tipping points to accelerate the green transition. Recent research highlights the rapidly expanding role of Artificial Intelligence (AI) in accelerating scientific discovery, enhancing innovation processes, and reshaping industries—developments that may amount to a broader technological paradigm shift. AI has the potential to reduce experimentation costs, optimize resource allocation, improve energy efficiency, and unlock new materials and clean technologies, thereby lowering barriers to green innovation and scaling. Future research could therefore examine the interaction between AI and green entrepreneurship within the context of the so-called “twin transition,” where digital and environmental transformations reinforce one another. At the same time, it will be important to assess potential trade-offs, including the environmental footprint of digital infrastructures and the risk of rebound effects. Understanding how AI can be steered to complement and amplify green entrepreneurial activity—rather than merely accelerate overall consumption—represents a promising and timely avenue for advancing both theory and practice in sustainable innovation.

This work is not free from limitations. Research on green entrepreneurship faces important limitations stemming from the persistent difficulty of defining what qualifies as “green.” The concept is multidimensional and encompasses a wide range of activities, from environmentally oriented innovations and clean technologies to broader sustainability-driven business models, yet there is no universally accepted definition or clear boundary. This conceptual ambiguity generates measurement challenges, as different studies rely on divergent criteria—such as industry classifications, self-reported missions, environmental certifications, or textual analyses—leading to inconsistent samples and limited comparability across findings. Advancing the field requires greater conceptual clarity. In this work, we develop a classification framework based on the Sustainable Development Goals (SDGs). This

approach provided a solid and legitimate foundation for identifying green activities. At the same time, however, it adds another framework to an already fragmented landscape, further contributing to to the myriad of conceptual approaches obscuring the field.

Moreover, firms may adopt partial or incremental environmental practices, making it difficult to distinguish genuinely green ventures from those engaging in symbolic or marginal sustainability efforts. In this thesis, we developed methods to identify green startups based on the content of their websites, an approach that enables research in data-scarce contexts but remains vulnerable to greenwashing and strategic self-presentation. Similarly, our analysis of entrepreneurial legitimacy work relied on posts published on social networking sites, which capture communicative efforts but do not allow us to assess more substantive or operational forms of legitimacy building. More broadly, accurately evaluating the true environmental impact of firms would require access to detailed and reliable performance data, which are often difficult to obtain due to limited disclosure and measurement complexity, especially for new ventures.

6 APPENDIX A

6.1 DATA: SCRAPPING PROCEDURE

As of May 2023, 26'892 Italian startups were registered as innovative startups. We used AIDA, an Amadeus-Bureau Van Dijk database, to access relevant firm-level information for these companies, such as a link to their websites and their localization. We used a Python-based, open-source web-crawling framework, Scrapy, to extract the html content of 15 webpages of each website. It manages websites with an excessive number of webpages and covers the whole website of more than 95% of the innovative startups (Fig. 5.1). When the links from our data were missing or leading to a malfunctioning website, we used Google's Programmable Search Engine to perform an automated search with the name of the startup. When a link was found whose domain included the startup name, we checked manually for a reasonable correspondence and added it as the startup website.

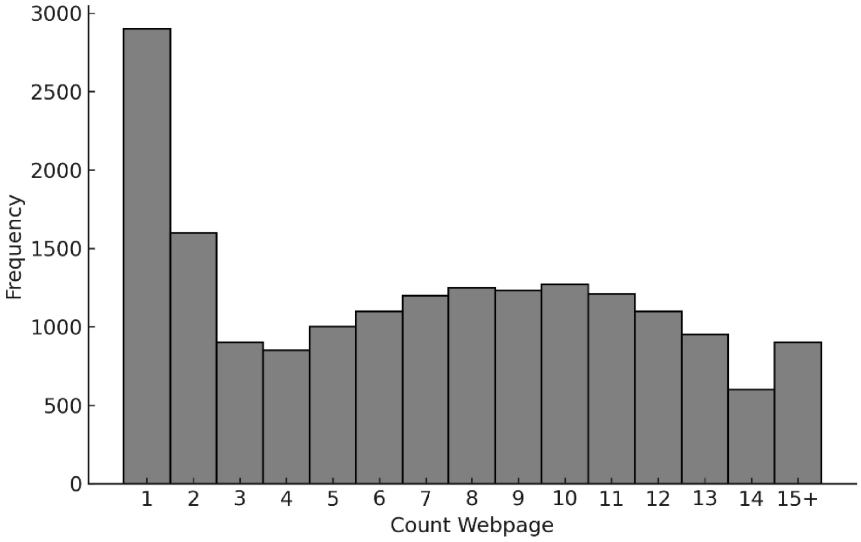


Fig. 5.1 Distribution of the Webpage Counts per Website

Then, to remove noise and irrelevant content, we extracted the text from the html using the program Trafilatura (Barbaresi, 2021), which selects only meaningful text from html content, removing so-called "boilerplate" meaningless content, such as footers, addresses, etc. (Bevendorff et al., 2023). We translated the text of the websites into English using Google Translate API. From the 26'892 startups from the initial dataset, we could

extract and translate the website of 15’024 companies. We removed 4’085 startups with empty or very short websites that provided no useful information. Our final sample is composed of 10’939 startups with websites longer than 250 characters and translated into English.

6.2 METHODS: DETAILED DESCRIPTION

The field of economics increasingly uses NLP methods to use text as data (Gentzkow et al., 2019). Three approaches are commonly used to classify text based on their topics: supervised classification, unsupervised topic modelling, and dictionary approaches (Arseneau et al., 2022). Supervised classification and traditional dictionary approaches require labelled data or established dictionaries, which come with high costs of implementation and subjectivity.

In this work, we start by applying advanced NLP methods to build a green classification framework from the SDG. Then, we implement three approaches that use this framework to identify green startups: an ML dictionary that builds its dictionary directly from the data (Li et al., 2021), and two unsupervised topic modelling approaches, using LDA (Blei et al., 2003) and BERTopic (Grootendorst, 2022) approach.

For interested readers, we compare them to two well-known NLP approaches that use external frameworks or labelled data for identifying green texts: a traditional dictionary using DicoEnviro (L’ Homme et al., 2020) and a BERT-variant approach using climateBERT (Webersinke et al., 2022). Those algorithms use external dictionaries and classification, and we do not present them in the main paper.

6.2.1 Green classification framework

While numerous sustainable taxonomies exist (IPSF Taxonomy Working Group, 2021), we build our classification scheme from the SDGs framework, as it is the most known and used in research. As the SDGs 17 goals with their 169 targets encompass economic, social and environmental sustainability goals, we need to refine them to identify green startups. In the Global Environment Outlook- GEO6 (UN, 2019), the UN identified 70 targets in 16 of the SDG related to environmental sustainability. Yet, those targets remain multidimensional and encompass economic, social and

environmental dimensions. Moreover, different SDG targets mention similar green goals. For instance, target 4.7 includes ‘education on sustainable development’, while target 13.3 includes ‘education on climate change’.

To group the different green topics mentioned in the green SDGs at the target level in a classification scheme, we manually extracted every green noun phrase (group of words working as a noun in a sentence) from each green SDG target, for a total of 220 noun phrases. As an example, we show the noun phrases that we extracted in bold from two target: 1.4 : ‘By 2030, ensure that all men and women, in particular the poor and the vulnerable, have equal rights to economic resources, as well as access to basic services, ownership and control over land and other forms of property, inheritance, *natural resources*, appropriate new technology and financial services, including microfinance’; and 7.a: ‘By 2030, enhance international cooperation to facilitate access to *clean energy research* and technology, including *renewable energy, energy efficiency* and advanced and *cleaner fossil-fuel technology*, and promote investment in *energy infrastructure* and *clean energy technology*’.

Then, we use SentenceBERT (Reimers & Gurevych, 2019) all-MiniLM-L6-v2, a language model trained on English texts to represent the noun phrases as embeddings. Embeddings are mathematical representations of texts that take into account the contextual complexity of language. For instance, they can distinguish the meaning of the word ‘environment’ in ‘work environment’ versus ‘natural environment’. Embedding the noun phrases allows us to perform mathematical operations, such as clustering.

To understand how many topics are present in our list of 220 noun phrases, we use HDBSCAN (McInnes et al., 2017), a clustering algorithm that groups similar items and identifies outliers. HDBSCAN considers how closely the noun phrase embeddings are grouped together in space. It suggests that the noun phrases can be grouped into 14 topics. However, since we want the topics to represent all the noun phrases, we use the KMeans clustering algorithm to group them into 14 green topics without outliers (Nigel & Britt, 1966). It ensures that the topics reflect all the green noun phrases present in the green SDG. KMeans assigns noun phrases to randomly selected centroids and then iteratively updates the centroids and reassigns the noun phrases to the nearest centroid until the centroids stabilize. To confirm that

the number of topics is adequate, we tried KMeans with more and fewer topics, which led to worse clustering. Finally, we use ChatGPT 4.0[17] to assign 14 green SDG labels to the 14 green topics. The final result is to be found in Table 1. Those 14 green topics will be key for identifying startups as “green” in the next parts.

Table 5.1 The 14 green topics extracted from the green SDGs

Topic Label	Green Noun Phrases	N. Noun Phrases	Targets contributing to the topic
Air Quality and Pollution Management	air pollution, hazardous chemicals, reducing pollution, hazardous materials, ...	11	3.9, 6.3, 8.4, 11.6, 12.c, 13.3, 14.1
Biodiversity Conservation and Protection of Species	genetic diversity of seeds, genetic diversity of cultivated plants, genetic diversity of farmed and domesticated animals, seed and plant banks, small-scale artisanal fishers, ...	17	2.5, 14.b, 15.5, 15.7, 15.8, 15.9, 15.c
Climate Change Adaptation and Mitigation Strategies	climate change, green spaces, adaptation to climate change, mitigation to climate change, sustainable and resilient buildings, ...	10	2.4, 11.7, 11.b, 11.c, 13.3, 13.a, 13.b
Disaster Resilience and Climate-related Risk Management	environmental shocks, environmental disasters, climate-related extreme events, flooding, drought, ...	15	1.5, 2.4, 11.5, 11.b, 13.1, 15.3
Environmentally Sound Waste and Resource Management	safe reuse, eliminating duping, recycling, recycling, reuse technologies, ...	21	6.3, 6.a, 9.4, 11.6, 11.c, 12.3, 12.4, 12.5, 12.c, 14.4, 17.7

[17] We chose ChatGPT 4.0 because it was the best LLM on Chatbot Arena (Chiang et al., 2024) for the overall category and many others at the time of the analysis. We used the following prompt “Here is a list of green noun phrase: {green noun phrases}. Please provide a concise topic label that effectively represents and captures the overall theme shared by all these noun phrases."

Forest Conservation and Sustainable Management	protect and restore forests, conservation of forests, sustainable use of forests, restoration of forests, reforestation, ...	11	6.6, 15.1, 15.2, 15.b
Freshwater Ecosystem Conservation and Restoration	supply of freshwater, protect and restore wetlands, protect and restore rivers, water-related ecosystems, protect and restore aquifers, ...	15	6.4, 6.6, 15.1
Land and Ecosystem Conservation and Restoration	improve soil quality, improve land quality, soil pollution, protect and restore mountains, conservation of mountains, ...	20	2.4, 3.9, 6.6, 15.1, 15.3, 15.4, 15.5, 15.9, 15.a
Marine Conservation and Sustainable Ocean Management	marine debris, marine pollution, restoration of marine ecosystem, protect coastal ecosystem, protect marine ecosystems, ...	21	14.1, 14.2, 14.3, 14.4, 14.5, 14.7, 14.a, 14.b, 14.c
Sustainable Development and Practices	sustainable food production, education for sustainable development, culture's contribution to sustainable development, sustainable lifestyles, Sustainable Consumption and Production, ...	25	2.4, 4.7, 8.4, 8.9, 11.2, 11.3, 11.4, 12.1, 12.6, 12.7, 12.8, 12.a, 12.b, 14.7, 15.4, 17.9, 17.14
Sustainable Energy Services and Efficiency	affordable energy services, reliable energy services, modern energy services, renewable energy, improvement in energy efficiency, ...	14	7.1, 7.2, 7.3, 7.a, 7.b, 12.c
Sustainable Fisheries Management and Illegal Fishing Prevention	end unreported fishing, end destructive fishing, end unregulated fishing, end illegal fishing, end overfishing, ...	10	14.4, 14.6
Sustainable Resource Management and Efficiency	natural resources, resource efficiency in consumption, resource efficiency in production, ...	9	1.4, 5.a, 8.4, 11.b, 12.2, 15.2, 15.6
Water Quality Management and	water pollution, affordable safe drinking water, untreated wastewater, water quality,	13	3.9, 6.1, 6.3, 6.4, 6.5,

Sustainable Use	water-use efficiency, ...	6.a, 6.b

6.2.2 Machine Learning (ML) Dictionary Approach

To build a dictionary that is specific to our dataset and green framework, we developed a method close to Li et al. (2021), which has been applied by Mansouri & Momtaz (2022) to assign ESG scores to start-ups. For each of the dimensions of the ESG, Mansouri & Momtaz (2022) initialize a dictionary with the most used words in newspaper articles mentioning the ESG dimension as seed words. Then, they extend each dictionary with words from the CSR report that are semantically close to the seed words. Finally, they assign each startup an ESG score as the share of ESG words in their CSR report.

We build a dictionary for the 14 green topics that we derived in the previous section using their labels as seed words. To diminish the size of the vocabulary of the startups' websites, we remove all stopwords (highly frequent words that convey little meaning, such as 'the' or 'at') and lowercase all the words using NLTK (Loper & Bird, 2002). Next, we perform lemmatization to reduce each word to its base form. For example, the lemmas of 'polluting' and 'wastes' are 'pollute' and 'waste,' respectively. To lemmatize the text on our websites, we utilize NLTK's lemmatizer.

After those preprocessing steps, we extract all uni-, bi-, and trigrams (groups of 1, 2 and 3 words found next to another) from our text. Then, we compute the embeddings of the seed words and the n-grams using SentenceBERT. Finally, we compute the cosine similarity between the seed word of each dictionary and the n-grams. The cosine similarity of two embeddings is the cosine of the angle between their vectors and represents well the semantic similarity of the words. Upon manual check, we set a threshold at 0.5, extending the dictionaries with all n-grams with a cosine similarity higher than 0.5 as green, identifying between 5'000 and 75'000 n-grams as green for the 14 dictionaries. The large size of the dictionaries is explained by the spelling and formatting mistakes in the websites, the words that are only used rarely (such as the name of the companies), and the n-gram approach, which causes each word to create multiple n-grams.

Finally, we compute for each webpage a green dictionary score as the number of green n-grams present in the text is divided by the total number of n-grams. We assign to each startup the highest green score of its webpages, again with the goal of identifying a startup as green when at least one of its webpages is about a green topic. We identify the 15% of startups with the highest score as green, in total 1'643. Fig. 5.2 shows the distribution of the scores. All the steps of this methodology are presented alongside those of a traditional dictionary approach in Fig 5.9.

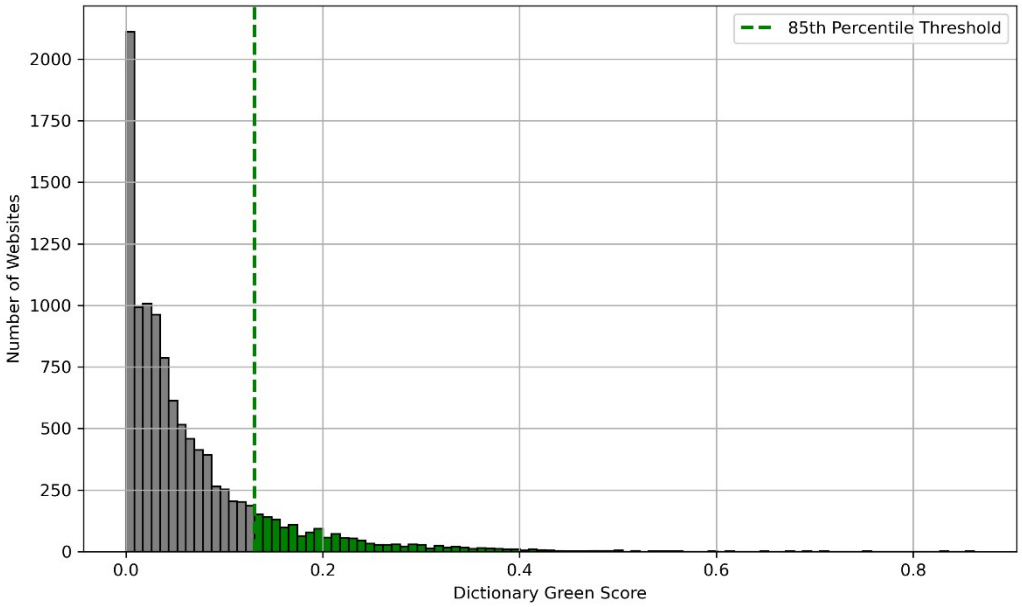


Fig. 5.2 Distribution of websites' green score of the machine-learning dictionary approach

6.2.3 Latent Dirichlet Allocation (LDA)

Second, we implement a well-established unsupervised topic model for natural language processing: Latent Dirichlet Allocation (LDA). This is a generative probabilistic model used in unsupervised learning to discover latent semantic structures in unstructured documents (Blei et al., 2003). Using the bag-of-words representation of documents, in which each document is represented by a list of words associated with their counts, the model assumes that documents are generated by drawing words from a set of *k* topics according to a specific probability distribution. LDA infers the hidden topic structures in a corpus to maximize the likelihood of observing the given set of documents. Even though the bag-of-word representation of the text ignores the contextual complexity of text, LDA is still widely used today. It is mostly used for identifying topics in a corpus of text, it has also been used for text classification (Yun & Geum, 2020).

LDA needs a few more cleaning steps to generate meaningful topics. We start by removing email addresses, urls, special characters, numbers, and one-letter words from the webpages. We then remove the English stopwords using NLTK's stopword lists. Then, we lemmatize the texts using NLTK's lemmatizer.

Before running LDA on our clean corpus of text from the websites, we have to set the number *k* of topics, which has been subject to several rounds of discussion among researchers. The diversity of the startups does not allow us to assume the number of topics a priori. While researchers have set the number of topics *k* for optimizing statistical indicators, such indicators have been shown to be uncorrelated with human judgement (Chang et al., 2009). Thus, we followed Tiba et al. (2021), trying different numbers of topics and finally setting it at 50, as more topics did not show noticeable improvement. We use the Python package Gensim (Řehůřek & Sojka, s. d.) to apply LDA to our data.

To identify which of the 50 LDA topics are green, we rely on the framework that we built in 3.2. We start by embedding each LDA topic and the 14 SDG green topics, again using SentenceBERT (Reimers & Gurevych, 2019) all-MiniLM-L6-v2. Then, we compute a green cosine score for each LDA topic as its maximal cosine similarity with an SDG green topic. Following a set of

manual checks, we set a cosine similarity threshold at 0.5, which systematically captured all topics that would be labelled as green by human classification. In any given run of LDA, between 0 and 4 topics are typically identified as green. Finally, we assign to each webpage its green score as the probability of being generated by green topics. We identify green websites as those that have at least one green webpage tagged by the LDA algorithm. Thus, we assign to each website the highest green score of its webpages, and, on each run, we identify the top 15% of startups with the highest green score as green.

Classifying the LDA topics automatically as green allows us to account for the stochasticity of LDA. We compute LDA models for seeds running from 0 to 25 and assign to each webpage its average green score. On some iterations, no green topics and no green startups are identified, suggesting a conservative cosine similarity threshold. Finally, we estimate the probability of each website to be identified as green on a single iteration of LDA and identify the 15% of startups with the highest probability as green, identifying 1520 startups (Fig. 5.3).

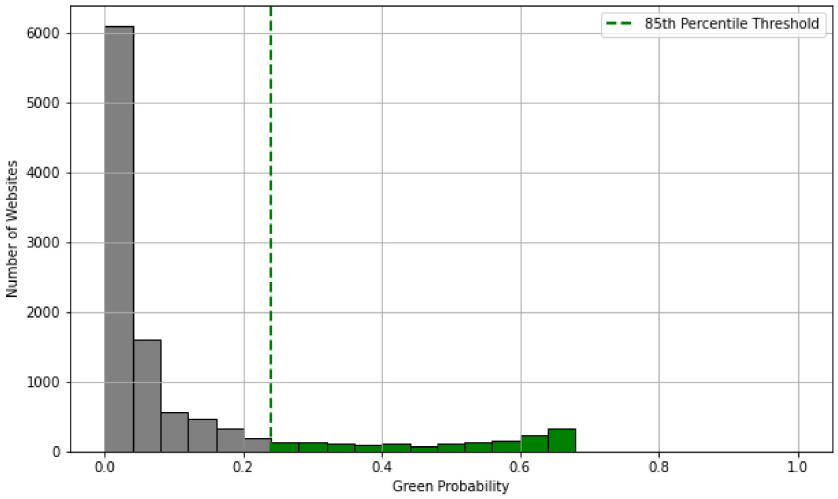


Fig. 5.3 Distribution of websites' green probabilities estimated by 25 iterations of LDA

6.2.4 BERTopic

Third, we examine BERTopic, a state-of-the-art unsupervised topic model introduced by Grootendorst (2022), which leverages embeddings to identify topics in a corpus of text. It starts by embedding the texts, then clusters them into topics, and finally generates a topic representation for each topic cluster. This approach has already been compared in other contexts to older unsupervised modelling techniques and has shown many advantages (Egger & al, 2022, Umamaheswaran & al, 2023), such as taking account of the semantic context of the words, automatically determining the number of topics, and requiring fewer data cleaning steps. While it has already been widely used in research, this paper is the first to apply it to identify green startups.

The procedure for the implementation of BERTopic can be divided into three steps. First, the algorithm embeds the text of our webpages using a pre-trained language model. Second, it reduces the dimensionality of the embeddings and clusters them. Third, it generates a topic representation for each cluster. As BERTopic is modular, we will present the algorithm that we used for each step. While little scientific discussion exists on the first two steps, a range of algorithms have been proposed for the final step of topic representation.

For embedding the webpages, we use again the SentenceBERT (While other models could have been tried, Borčin & Jose (2024) suggest that BERTopic achieves similar results with different embedding models. One should note that SentenceBERT has a token limit of 512 tokens. While webpages can be longer than that, we believe that it is not an issue as Trafilatura (Barbaresi, 2021) shortens the webpages by cleaning them of most boilerplate content, and the important information on a webpage are located at its beginning. 73% of the webpages have less than 512 tokens (Fig. 5.4). Those with more are truncated at 512 tokens, keeping the first and most informative tokens of the webpages.

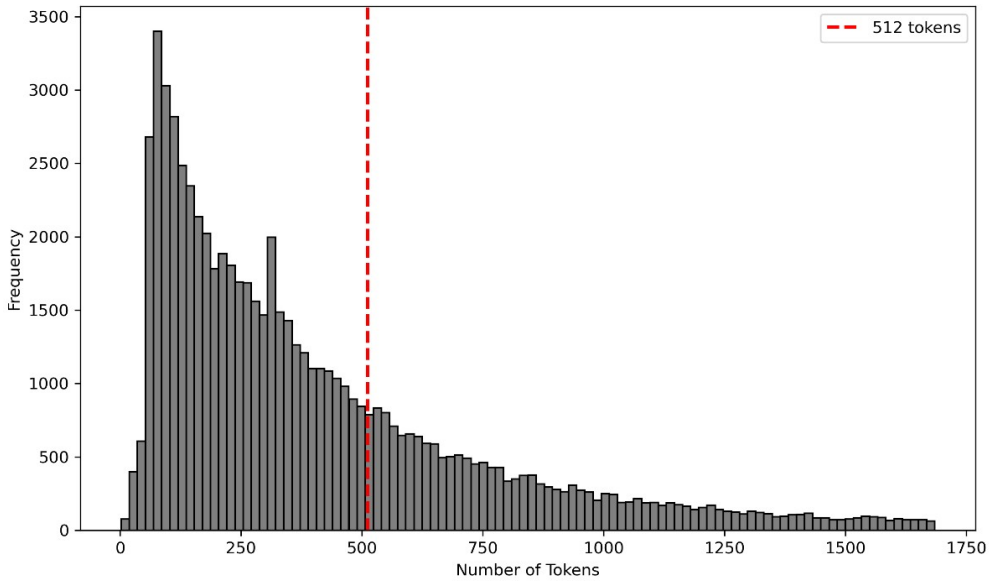


Fig. 5.4 Distribution of number of tokens per webpages

BERTopic reduces the dimensionality of the webpages' embeddings using UMAP (McInnes et al., 2020) to optimize the clustering process and deal with the curse of dimensionality. UMAP is a state-of-the-art, widely used dimensionality reduction algorithm. UMAP is a stochastic algorithm making use of randomness to approximate the best dimensionality reduction. It allows a low computing cost but introduces randomness into BERTopic. Then, we cluster the webpages using HDBSCAN (McInnes et al., 2017). It is based on DBSCAN, extending it to identify clusters of varying densities. These clusters are groups of webpages about similar topics, henceforward called topics. Each topic is formed by a varied number of startups with a minimum of 10. An outlier topic is also created from the webpages that were not clustered into any other topics. The number of topics is set by the algorithm.

We extract the 10 n-gram topic representation from each topic using KeyBERTInspired on c-√TF-BM35(IDF) candidates' representation n-gram. We used a count vectorizer to identify uni- bi- and tri-grams to increase the interpretability and richness of the topic representations. While BERTopic's default algorithm is the simpler c-TF-IDF, which identifies words that are more prevalent within a topic compared to the rest of the corpus, Borčin &

Joemon (2024) found that it generates topic representation with a relatively high rate of stopwords, and recommends using c-√TF-BM35(IDF). Upon inspection, while there were no stopwords using c-√TF-BM35(IDF), some n-grams were still meaningless. Thus, we used the KeyBERTInspired representation model to ensure that the n-grams are coherent and represent the topic well. It starts by computing 100 c-√TF-BM35(IDF) n-grams. Then, it embeds them and computes their cosine similarity to the embeddings of representative documents in the cluster. The 10 n-grams with the highest similarity scores are extracted to build the topic representation. This led to better topic representation. The topic representations returned for our texts are unambiguous to a human reader.

We identify the green topics needed to compute the cosine similarity with a procedure similar to LDA, computing the cosine similarity of each topic representation to our list of SDG green topics and assigning each webpage a green cosine score as the maximal cosine similarity of its topic with an SDG green topic. We group those webpages at the website level, assigning to each website the green cosine score of its highest-scoring webpage.

We compute the green cosine score of the startups' websites with 25 iterations of BERTopic, to mitigate the stochasticity of BERTopic. In a single run, we identify the websites with the top 15% green cosine scores as green. Then, we use the outcome of all the iterations to estimate each website's green probability, its probability to be identified as green by BERTopic on one iteration. This led to better results than using the average green cosine score. Setting the threshold for identifying of green startups at the top 15% of the distribution, we consider startups with the highest green probability as green, identifying 1'524 green startups (Fig. 5.5). The steps of the BERTopic methodology are described alongside those of the LDA methodology in Fig. 6.

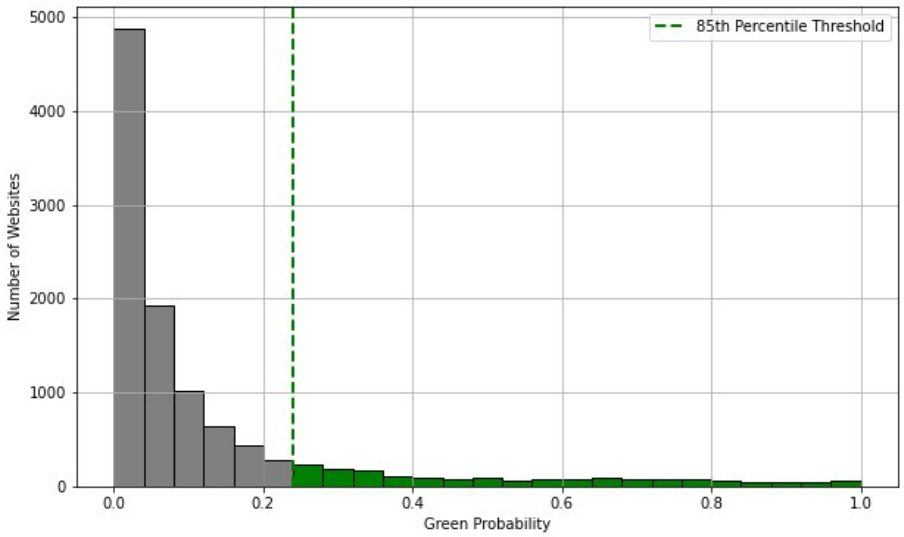


Fig. 5.5 Distribution of website found green by BERTopic for 25 iterations

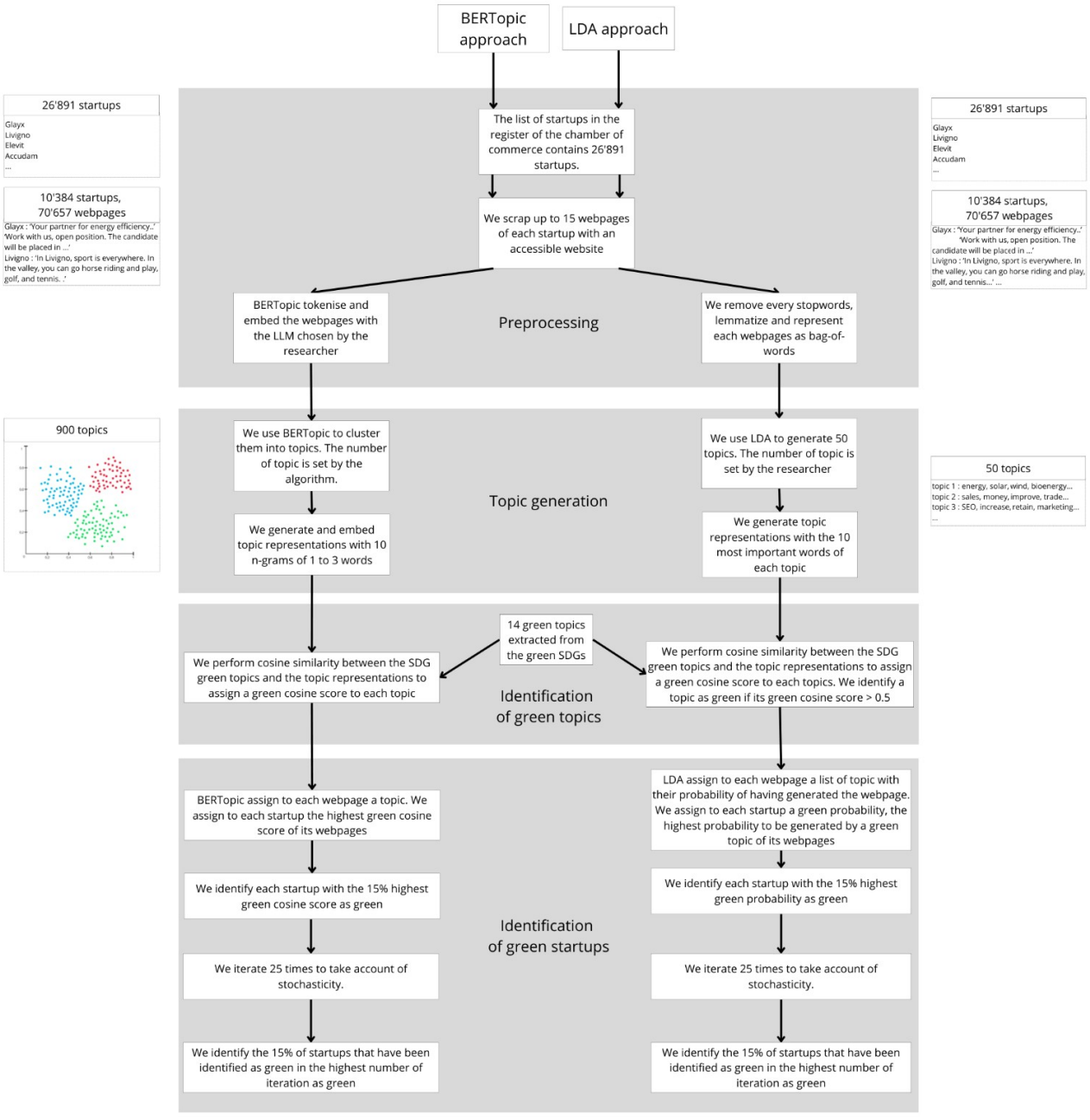


Fig. 5.6 Infographic of LDA and BERTopic

6.2.5 Traditional dictionary approach

The best-known and conceptually simplest NLP text classification approach is the traditional dictionary. Its first use dates to 1968 due to its simplicity: researchers start by building dictionaries containing lists of words corresponding to classes and then count the frequencies of presence of those words in the text they wish to classify (Guo et al., 2016). While this method does not consider the syntax, semantics, and context complexity of language, it is still widely used and provides valuable results as of today. To identify green firms, Gorovaia & Makrominas (2024) have used DicoEnviro (L' Homme et al., 2020), a lexical resource built for applying NLP to text related to the environment. Even if such an external dictionary does not perfectly use our framework, we implement this method and compare its precision to the ML dictionary method to show its advantages.

DicoEnviro is a dictionary of around 1414 words for French and 1200 words for English. Dictionaries for other languages are being developed but are yet too small to be used, for instance, with only 50 words for the Italian language. The dictionaries were built by comparing the content of a standard corpus of text to an environmental corpus of text and extracting the terms more prevalent in the environmental corpus as candidate green terms. Then, terminologists analyze the candidate terms to keep the relevant ones and complete it with information on argument structure, annotation of context, and semantic frames (L' Homme et al., 2020). In this work, we use the English dictionary to count green words and compute the green score of startups' websites.

To properly count the number of DicoEnviro's term instances in the websites, we need to lemmatize their words to ensure that plural forms or other modifications of the DicoEnviro terms are counted. Once the texts are lemmatized, we count the number of times a DicoEnviro keyword appears in each webpage, and we compute a green DicoEnviro score as the share of DicoEnviro keywords over the total number of words in the webpage. We assign to each startup the highest green DicoEnviro score of its webpages, thus identifying a startup as green if at least one webpage of its website has a high green DicoEnviro score. This score is between 0 and 1, a score of 0.1 indicating that 10 % of the words present on the website are in DicoEnviro.

We identify the 15% of startups with the highest green DicoEnviro score as green (Fig. 5.7).

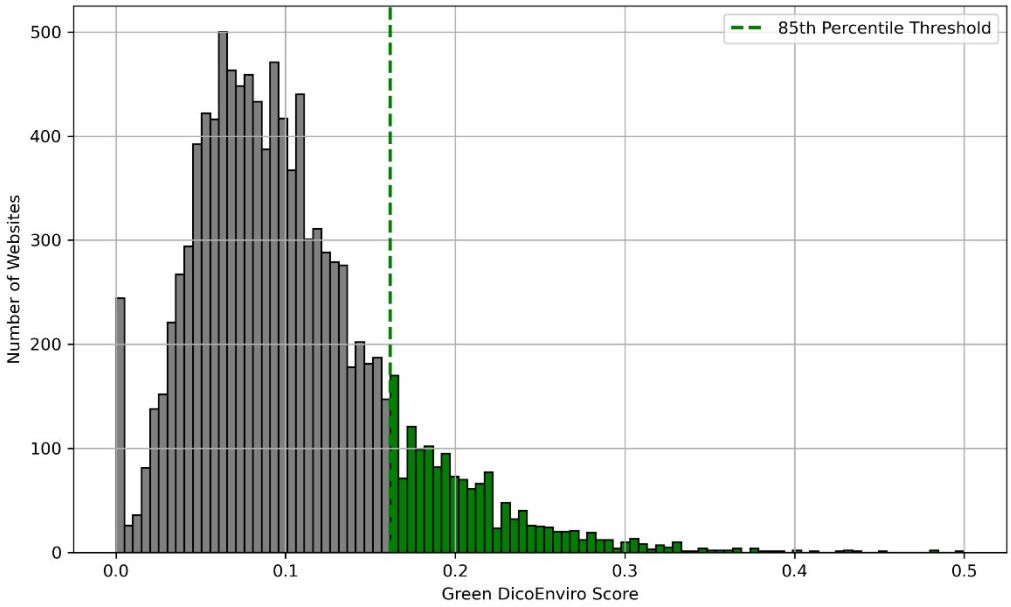


Fig. 5.7 Distribution of startup's green DicoEnviro score

Now, we compare the traditional and the ML dictionary approaches in Fig. 5.8. They identify about half the same green startups. We evaluate the precision of the traditional dictionary approach by randomly drawing 100 startups and inspecting their website. We could link 81 websites with one of the green dimensions of the SDGs, thus reaching a precision of 81%, a little lower than the machine learning dictionary approach (83%).

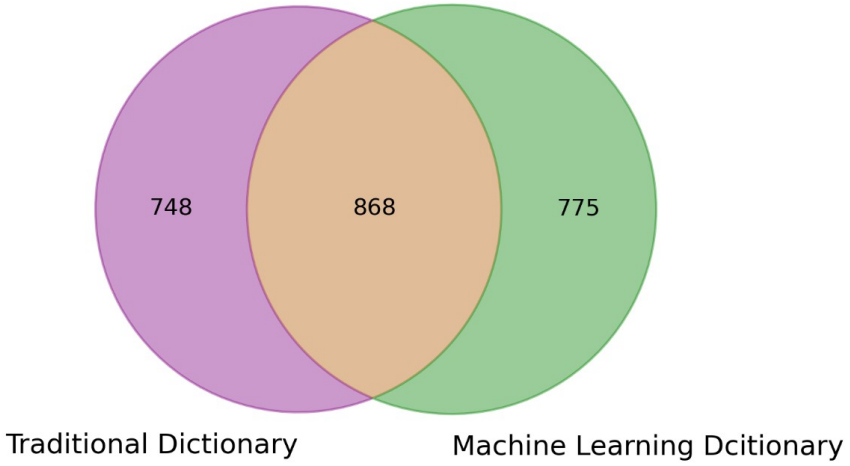


Fig. 5.8 Venn Diagram of the identification of startups with the ML and traditional dictionary approach

Since the traditional dictionary approach is based on a single external dictionary, it does not allow us to identify which green topic is discussed on the websites. Overall, the ML dictionary approach leads to marginally better results than the traditional dictionary approach. However, it allows us to construct dictionaries for any framework, even when dictionaries are not available. Fig. 5.9 shows each step of the traditional and ML dictionary approach alongside.

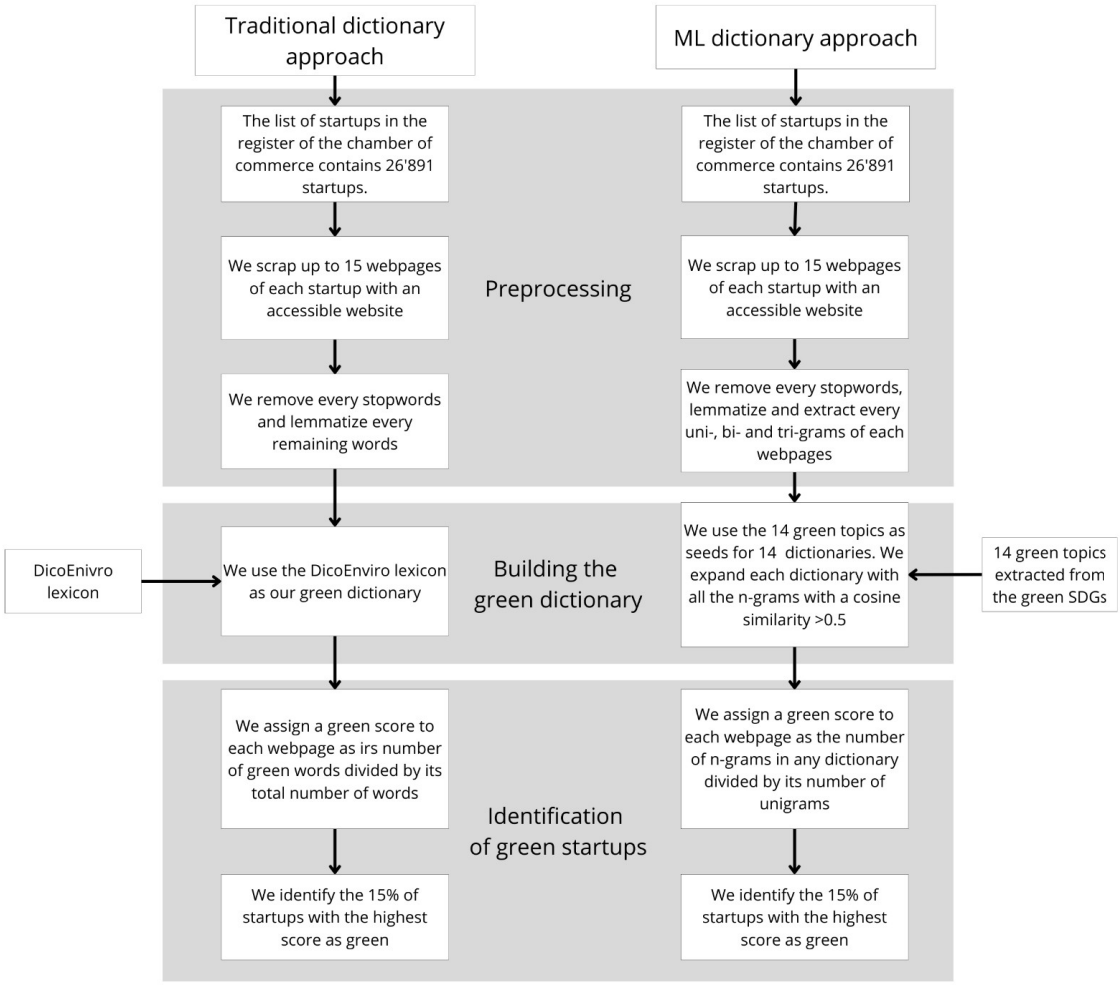


Fig. 5.9 Infographic of the traditional and ML dictionary approach

7 APPENDIX B

7.1 RELEVANCE AND OVERVIEW OF THE METHODOLOGY

To test our hypotheses on the role of knowledge and demand for innovative green startup creation, we first have to identify green startups. The task is not straightforward due to the diverse definitions of sustainability or greenness adopted in the literature (see, for example, the discussions in Colombelli et al., 2024; Gast et al., 2017; Purvis et al., 2019). To date, the literature has not reached a consensus on how to identify green startups, which has led to a variety of empirical approaches to tag them. Operationalizations range from narrower measures, such as only looking at energy startups (Colombelli & Quatraro, 2019), to broader definitions, including all sustainability-oriented startups (Tiba et al., 2021). While the term “sustainability” is multidimensional and also encompasses a social dimension, we focus on the environmental side of sustainability since social and green entrepreneurship have different characteristics (Bonfanti et al., 2024; Colombelli et al., 2024; Schaltegger & Wagner, 2011). At the same time, to understand the role of systemic local knowledge and demand factors, we need a comprehensive measure that spans beyond specific fields, such as the energy sector, to capture the emergence of green entrepreneurship across all sectors of the economy.

Therefore, a key contribution of the present work is the development of a new methodology to identify specifically green startups. To define “greenness” and environmental sustainability, we select the targets within the UN Sustainable Development Goals (SDGs) that refer specifically to environmental goals. While the SDGs have been used extensively to understand business sustainability strategy and sustainable entrepreneurship (Horne et al., 2020; Jha & Pande, 2024; Mio et al., 2020), prior research has often employed broader definitions at the SDG level. Our more granular, target-level approach seeks to improve on previous studies to capture dimensions hidden at the broader SDG level, such as waste management, resilience to disaster or air quality. In this endeavour, we take stock of the United Nations Environment Programme (UNEP)’s recent publication of guidelines to identify among the SDGs the targets that refer specifically to environmental sustainability. The UNEP 6th Global Environment Outlook (GEO6) identifies 16 goals, with 70 targets and 93 indicators that are environmentally related, building a granular green taxonomy (UNEP, 2019). Hence, we define green entrepreneurship as the

discovery, creation, evaluation, and exploitation of opportunities to create future goods and services consistent with the subset of green SDG targets identified by the GEO6 (see also Shepherd & Patzelt, 2011 for a similar, broader definition based on SDGs).

Next, we develop a new AI-based protocol that analyses the text of the companies' websites to determine their alignment with the green SDG targets. Assessing the alignment of startups to SDG targets may be challenging when drawing solely on sectoral, product and patent classifications. In many cases, green products and services are introduced as sustainable variants of existing, non-green ones; core innovations may not be patented; and startup companies typically lack detailed documentation about their activities, such as sustainability reports. In contrast, startups' websites are a valuable source of publicly available information about their products, services, values, and mission. They offer a succinct but informative depiction of their operations on a few web pages. Several studies have applied web-based approaches to identifying green startups, but in most cases, they have classified the content of websites manually or through keywords associated with sustainability (Colombelli et al., 2024; Horne et al., 2020; Kuckertz et al., 2019; Mrkajic et al., 2019). However, thanks to the latest AI-powered text analysis tools, it is possible to fully automate the classification process in a web-based approach, making it scalable and time-efficient even with large numbers of companies with heterogeneous website contents.

Our AI-based approach significantly improves classic approaches to identifying green startups. To date, numerous studies have relied on rare or expensive data and time-consuming methodologies to tag green startups: for example, they have used third-party data providers (Cojoianu et al., 2020; Dong et al., 2022; Jha & Pande, 2024), questionnaires (Abdesselam et al., 2024; Chapman & Hottenrott, 2022; Hörisch, 2015), or manual classifications (Wöhler & Haase, 2022). Others have identified green startups based on application for green patents (Coll-Martínez et al., 2022) or from press articles mentioning their green engagement (Gebhardt & Bachmann, 2023). All of these methods are difficult and costly to scale up to large populations of companies, especially startups.

A few recent studies moved beyond the limitations of such approaches and implemented Natural Language Processing (NLP) algorithms to tag new sustainable ventures using text data. Some studies have proposed dictionary approaches that classify text based on keywords, but their main limitation is

the need to rely on external dictionaries or dictionaries written by the researcher team (Gorovaia & Makrominas, 2024; Horne et al., 2020). Further, dictionary approaches often struggle to capture the contextual nuances of the text since similar words can have different meanings depending on their context, such as the word “environment”, referring to the natural environment and a work environment.

More advanced solutions have integrated some degree of machine learning to aid the classification process: for example, unsupervised topic models, in particular Latent Dirichlet Allocation (LDA), have been used for identifying SDG-related startups (Tiba et al., 2021). While this method is more advanced than dictionary approaches as it focuses on word combinations and not on single keywords, a significant limitation is that it represents text through lists of words, disregarding their order and thus still losing contextual nuances. Fully automated supervised classification of texts has been proposed but relies again on labelled data to train the classifier (Gidron et al., 2023; Li et al., 2016; Yun & Geum, 2020). Moreover, current implementations of LDA to select sustainability-related topics have relied on manual adjustments after text processing, which reduces the reproducibility and scalability of the procedure and exposes the method to biases.

Exploiting the recent advancements in Natural Language Processing, we propose a methodology to classify green startups from their websites automatically, based on a state-of-the-art unsupervised topic model that does not require any pre-labelled data or dictionaries: the BERTopic (Bidirectional Encoder Representations from Transformers Topic) algorithm by Grootendorst (2022). The first advantage of BERTopic is that it leverages Large Language Models (LLM) and can understand words in context since it relies on the embedding representation of texts, a mathematical representation considering texts’ semantic complexity (Reimers & Gurevych, 2019). For example, BERTopic can distinguish between terms like ‘natural environment’ and ‘work environment’, capturing the context-specific meanings of each. The second advantage is that the algorithm selects the number of topics, while earlier methods require the researcher to set it (Tiba et al., 2021). Research has shown further technical advantages of BERTopic over earlier methods in many contexts (Schaltegger & Wagner, 2011; Umamaheswaran et al., 2023).

In short, our method to identify green firms consists of the following steps: starting from a sample of startups, we scrape their web pages and apply

BERTopic to identify topics in that text. BERTopic processes webpages by embedding their content using a multilingual language model. Then, the dimensionality of the webpages is reduced with UMAP before they are clustered with HDBSCAN. Finally, the algorithm extracts topic representations with KeyBERTInspired. In parallel, we define green labels from the Sustainable Development Goals that GEO6 defined as “green targets”. Finally, we compare the topics extracted from the startups’ websites and the green labels from the SDGs and consider the highest levels of cosine similarity to capture green startups. Using cosine similarity as a selection criterion allows for a computationally inexpensive, fully automatized classification of the websites. The rest of Appendix B presents the building blocks of the machine learning protocol developed to identify green startups through BERTopic, starting from data mining from companies’ websites to tagging environmentally sustainable companies.

7.2 SAMPLE OF INNOVATIVE STARTUPS AND THEIR WEBSITES

First, we select startups from the Italian Registry of Innovative Startups with active websites with useful content. Fig 6.1 summarizes the different steps that define our initial sample from the population of Italian innovative startups. From the 26’891 startups available in the dataset as of May 2023, we could extract the webpages of 16’980. We removed 3’338 startups with empty or very short websites, providing no useful information. Those were typically websites under construction. Finally, we dropped non-Italian websites. While BERTopic works for multilingual websites, having only one language simplifies its interpretation.

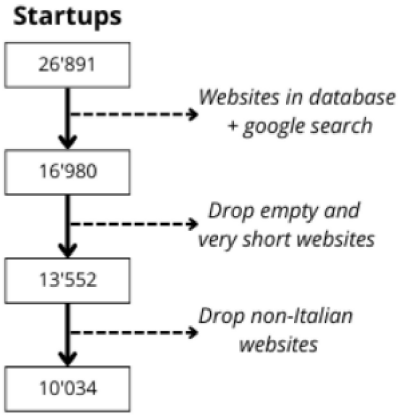


Fig. 6.1 Data cleaning steps.

The resulting sample has a geographical distribution equivalent to that of the whole population at the regional and provincial levels (Fig. 6.2 and Fig. 6.3).

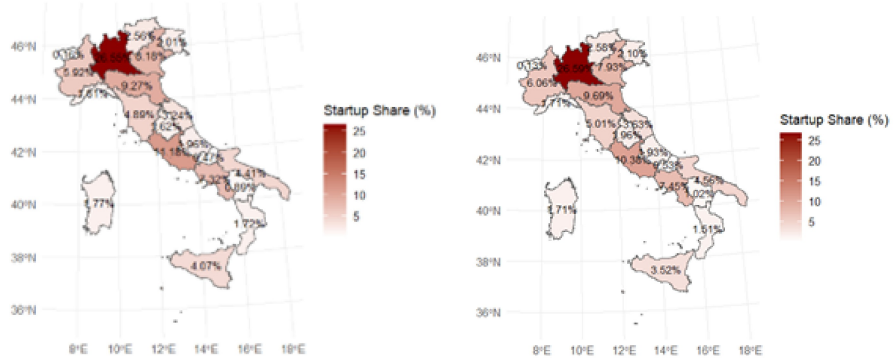


Fig. 6.2 Share of innovative startups at the regional level, whole population (left) and scrapped sample (right)

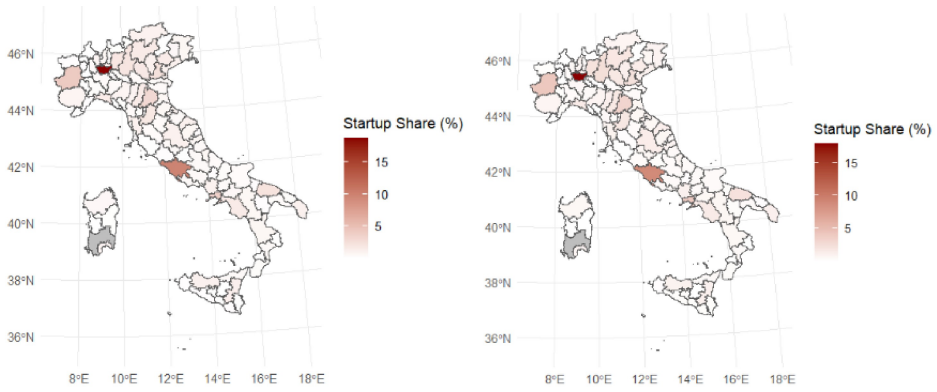


Fig. 6.3 Share of innovative startups at the provincial level, whole population (left) and scrapped sample (right)

Then, we used Scrapy to extract the HTML of the websites. We extract the HTML of the home webpage and up to 15 pages sharing the exact domains. We limited the number of webpages to avoid scraping all the pages of a few

large websites, which would make the process computationally intensive and significantly slower. Ultimately, this limitation does not affect our textual data significantly, as 96.56% of the websites have less than 15 pages (see Fig. 6.4 for their frequency distribution).

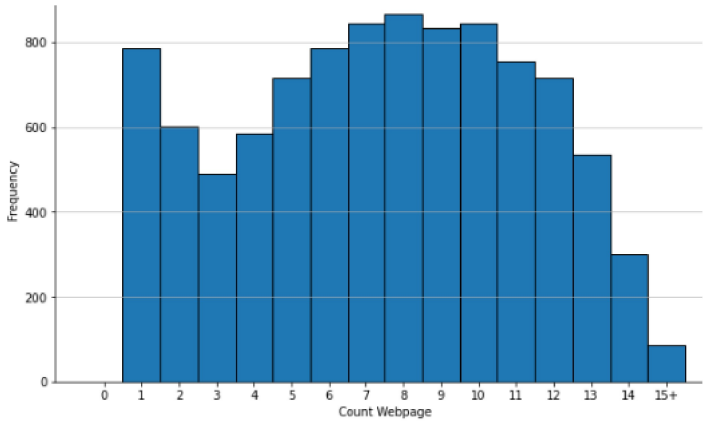


Fig. 6.4 Distribution of the webpage counts per website

When the links from our data were missing or leading to a misfunctioning website, we used Google's Programmable Search Engine to perform an automated Google search with the startup's name. When a link was found whose domain included the startup name, we added it as the startup website. A large share of text on webpages, such as footers, cookies, URLs, etc., is made up of boilerplates: they do not provide useful information but add noise for further analysis. To improve our data quality, we extracted the meaningful text from the HTML using Trafilatura (Barbaresi, 2021), a state-of-the-art boilerplate removal that works well on complex webpages (Bevendorff et al., 2023).

7.3 IDENTIFICATION OF GREEN WEBSITE WITH BERTOPIC

BERTopic is a state-of-the-art unsupervised topic model introduced by Grootendorst (2022). BERTopic identifies and assigns one topic to each text given a corpus of texts. It leverages the so-called embeddings, mathematical representations of texts. BERTopic is built on three steps: it starts by embedding the corpus texts, then clusters them into topics, and finally generates a topic representation for each topic. BERTopic is modular, allowing the researcher to choose which algorithm to use for each step. BERTopic has been compared to other unsupervised modelling techniques

and has shown many advantages, among which are its ease of use, its lesser need for human intervention, and its better results on statistic measure as well as on human interpretation (Egger & Yu, 2022, Umamaheswaran et al., 2023). While it has been widely used in research, this paper is the first to use it to identify green startups. To fully automatize the identification of green startups, we add a new step to label green topics automatically, allowing us to assign a green score to each startup. We identify the top 15% of startups with the highest green score as green startups. For a quick overview of the different algorithms used at each step, see Fig. 6.5. The following sections explain in detail each step of the application of BERTopic.

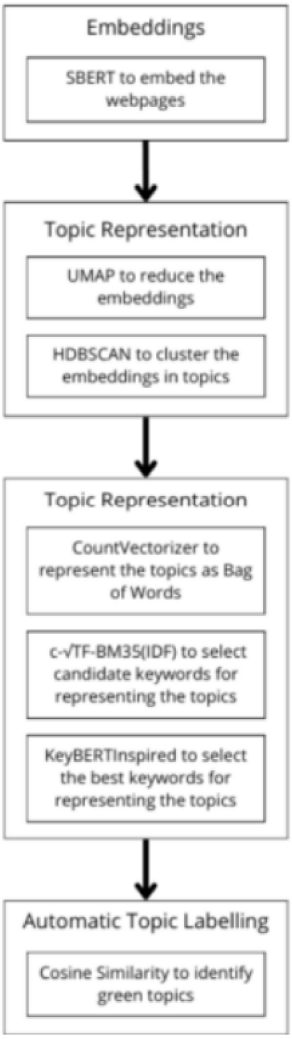


Fig. 6.5 Step-by-step implementation of the BERTopic algorithm

7.3.1 Embedding the webpages

For embedding the webpages, we use SentenceBERT, or SBERT (Reimers & Gurevych, 2019), a language model trained to compute similar sentence or paragraph embeddings to semantically similar sentences or paragraphs. A range of pre-trained models exists for different use cases. Many models

are trained only on English data and can only compute meaningful embeddings for English texts. As our texts are in Italian, we use the paraphrase-multilingual-MiniLM-L12-v2 model, trained in more than 50 languages, including Italian. We embed each of our webpages into a 384-dimension vector.

As computing embeddings has a quadratic complexity with respect to the sequence length (O(n^2)), SentenceBERT has a token limit of 512 per text. Fig. 6.6 shows the distribution of tokens per website. If a webpage has more than 512 tokens, only the first 512 tokens are considered by BERTopic. While 32% of the startups have more than 512 tokens, we believe it is not an issue, as crucial information is usually written at the websites' beginning. While other multilingual language models could be considered, Borčin & Jose (2024) suggest that BERTopic achieve similar results for different models.

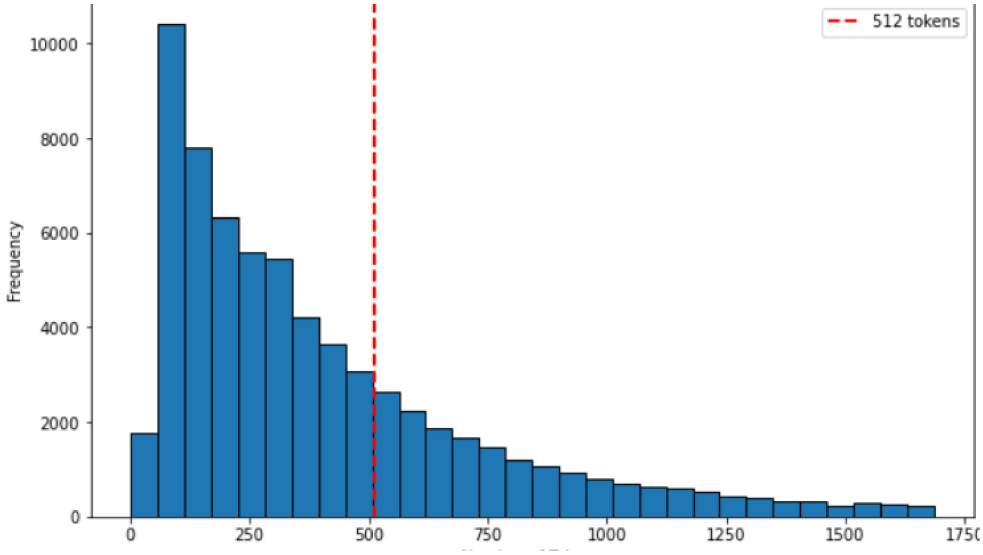


Fig. 6.6 Distribution of token count per webpage after the cleaning steps

7.3.2 Clustering webpages in topics

Before clustering our webpage embeddings, we have to consider the curse of dimensionality: distances between different data points converge for high-dimensional data, which creates difficulties for clustering. Our approach to deal with this issue is to start by using UMAP (Uniform Manifold Approximation and Projection) to reduce the dimensionality of the webpages' embedding (McInnes et al., 2020), reducing the dimensions of the embeddings from 384 to 5. Although other dimensionality-reducing algorithms exist, such as PCA or t-SNE, UMAP is a state-of-the-art stochastic algorithm that preserves the original characteristics. Furthermore, it improves the time and accuracy performance of clustering algorithms such as HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), which we use to cluster our webpage embeddings. HDBSCAN (McInnes et al., 2017) finds regions of high density in the data and uses these to build a hierarchy of clusters of varying shapes and sizes. It extracts the most stable cluster and identifies outliers. We set the minimum cluster size to 10 and use the Euclidean distance metric. On one iteration of BERTopic on our data, 850 to 950 topics are identified. Those topics are clusters of webpages with similar embeddings and semantically similar text content. The same number of topics are not found on each iteration as UMAP introduces stochasticity in BERTopic. At this point, all webpages are clustered in a topic, but the topics are not yet interpretable by a human reader.

7.3.3 Generating topic representation

We now create a topic representation for each topic so we can understand them. Fig. 6.7 presents a quick overview of this process. To do so, we start by representing our topic in Bag-of-Words (BoW) representation. We use a Count Vectorizer so that our Bag-of-Words includes a count for all groups of n words for $1<n<3$, called n-grams. Indeed, some words meaning change depending on their context: for example, 'green energy' refers to energy that has a low impact on the environment, rather than energy of the green colour. Bag-of-Words, although very simple, allows for cost-efficient topic representation.

BERTopic's default topic representation algorithm is c-TF-IDF, a variation of TF-IDF (Term Frequency – Inverse Document Frequency) that identifies words that are relatively highly present in each topic compared to the others.

Formally, c-TF-IDF assigns a score w to each n-gram x in every class c such as

$$w_{x,c}=|tf_{x,c}|\ |*\log(1+Af_x) w_{x,c} = (|tf_{x,c}|) * log\left(1+\frac{A}{f_x}\right)$$

with

$tf_{x,c}$=frequency of word x in class c $tf_{x,c} = frequency\ of\ word\ x\ in\ class\ c$

f_x=frequency of word x across all classes $f_x = frequency\ of\ word\ x\ across\ all\ classes$

A=average number of words per class $A = average\ number\ of\ words\ per\ class$

The score $w_{x,c}$ increases with the frequency of n-gram x in class c and decreases with its relative frequency in all classes. The n-grams with the highest score can be used to represent each topic.

Such a procedure already allows for interpretable topic representation, but Borčin & Jose (2024) recently showed that it tends to include commonplace words that convey little meaning (so-called stopwords), and found that the variant c-√TF-BM35(IDF) leads to a better result. Formally, c-√TF-BM35(IDF) assign a score w to each word x in every class c such as

$$w_{x,c}=|\ |tf_{x,c}-\sqrt{|}\ |\ |*\log(1+A-f_x+0.5f_x-0.5) w_{x,c} = (|\sqrt{tf_{x,c}}|) * log\left(1+\frac{A-f_x+0.5}{f_x-0.5}\right),$$

which gives a lower score to highly frequent words. While this variant creates topics with no stopwords, the topics occasionally include words unrelated to the topic.

To deal with the latter issue, we use KeyBERTInspired to select only the most relevant n-grams. Using KeyBERTInspired to generate the representation of a given topic, we start by extracting 100 candidate n-grams of each document using the c-√TF-BM35(IDF) procedure. In parallel, we compare the candidate n-grams of each document to candidate n-grams extracted from the whole topic. Documents with the most similar candidate n-grams are identified as the most representative document of the topic. Then, we embed those documents and collapse them in a representative centroid. We embed the candidate n-grams and compute their cosine

similarity to the representative centroid. Cosine similarity measures the cosine of the angle between two embeddings and their semantic similarity. Finally, we generate the topic representation from the ten n-grams with the highest cosine similarity. Such a method results in the most interpretable topic representations, with no stopwords and rarely words that seem unrelated to the topics.

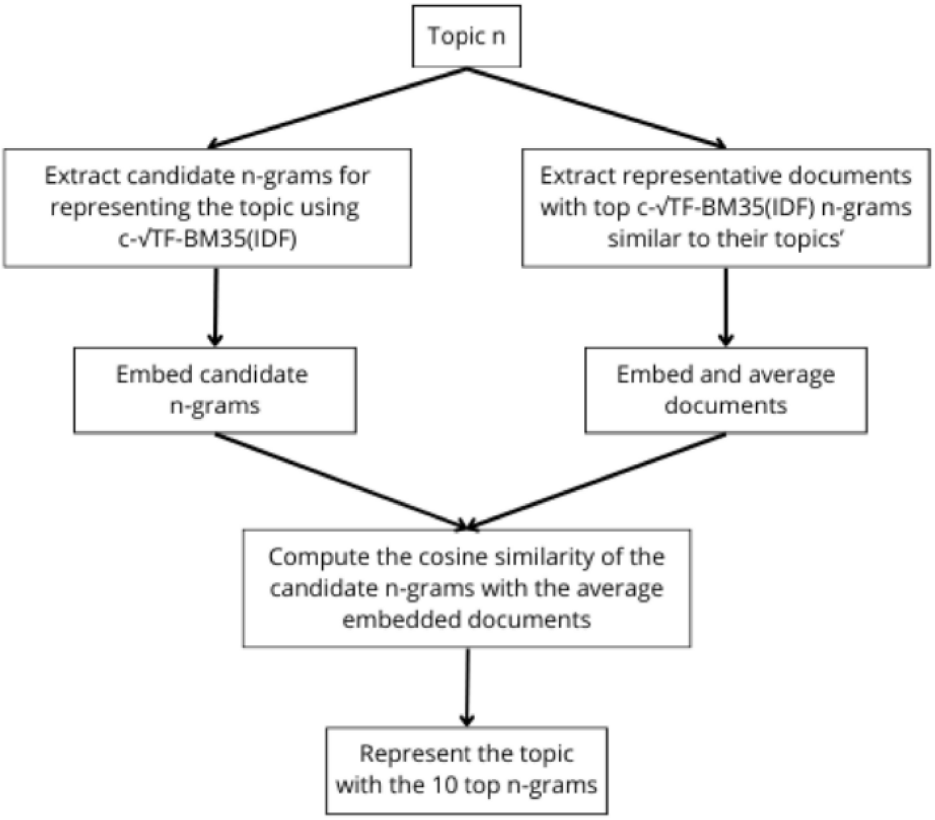


Fig. 6.7 Infographic of the KeyBERTInspired topic representation algorithm

7.3.4 Labelling automatically green topics

At this point, the startups' webpages are clustered into approximately 900 topics. We introduce a method based on cosine similarity to identify which are green (see Fig. 6.8). We start by embedding every topic representation.

In parallel, we embed the list of green labels extracted from the GEO6, accurately representing the SDGs' green dimensions. Then, for a given topic, we compute the cosine similarity between its representations and each label. We attribute the highest cosine similarity to the topic and all its startups as a green cosine score. We continue by grouping the webpages at the startup level, assigning to each startup the highest green cosine score of its webpage. Hence, we will be able to identify startups with at least one green webpage as green.

Researchers using topic models have traditionally labelled their topics manually, reading them individually and assigning a label based on human interpretation. Such practice is work intensive, prone to biases and mistakes, and reduces the method's reproducibility. An automatic labelling method has recently been introduced on BERTopic's website, leveraging zero-shot classification. However, in our use case, it performed worse and had a higher computational cost than our approach.

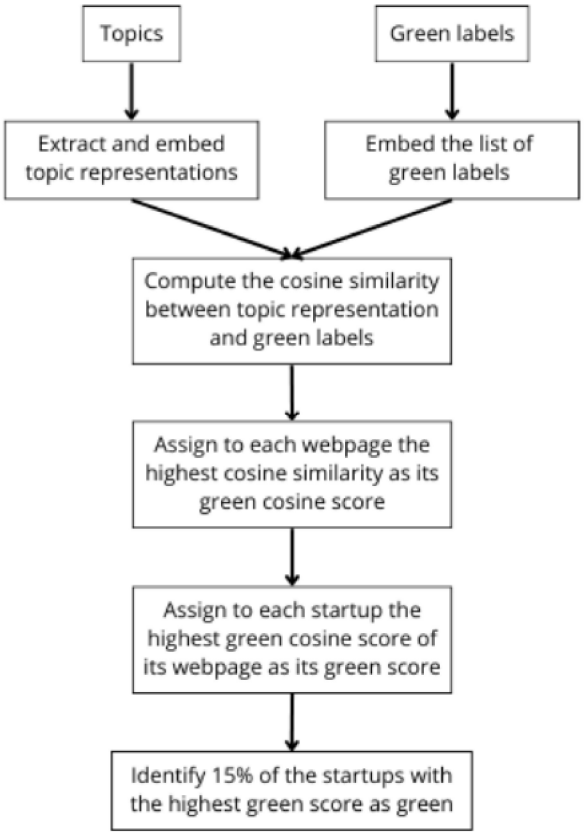


Fig 6.8 Infographic of the automatic labelling using cosine similarity

Our implementation of BERTopic can now assign a green score to startups without human intervention. To take account of the stochasticity introduced by UMAP, we iterate over BERTopic 25 times. We assign 15% of startups with the highest green score a green vote on each iteration. Finally, we identify the 15% of the startups with the highest number of green votes as green, identifying 15'019 startups as green. Fig. 6.9 shows the distribution of green votes across green startups. Over 3500 startups are never voted green, while very few startups are labelled green on each iteration. We identify startups as green when they are labelled so at least in nine iterations.

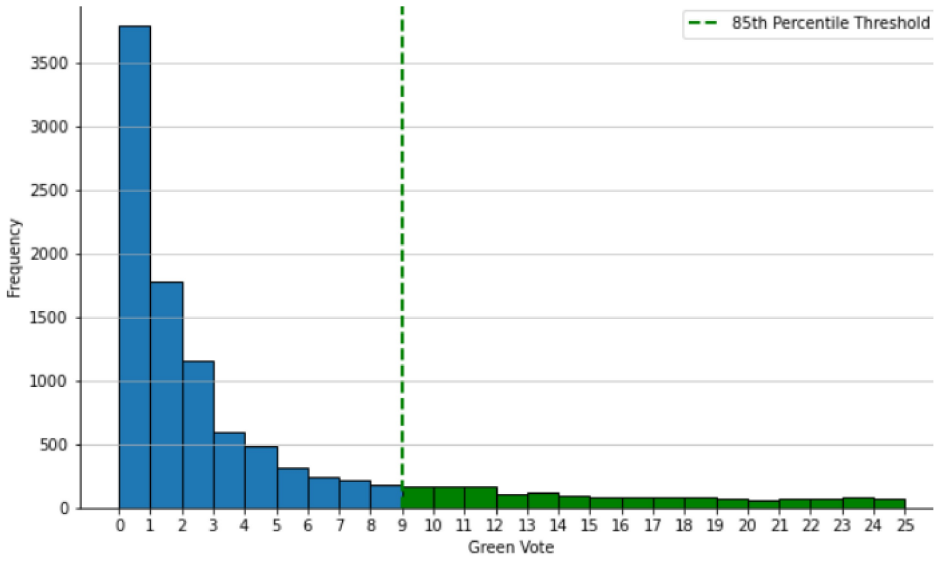


Fig. 6.9 Distribution of startups' green vote on 25 iterations

Fig. 6.10 reports the share of green startups in Italian provinces, as reported in the main manuscript. The map shows the geographic distribution of green innovative startups out of all innovative startups in a province. No green startups are found in four provinces: Vercelli and Vibo Valentia, in which our

sample only contains four and three startups, and Siena and L'Aquila, with 25 and 34 startups.

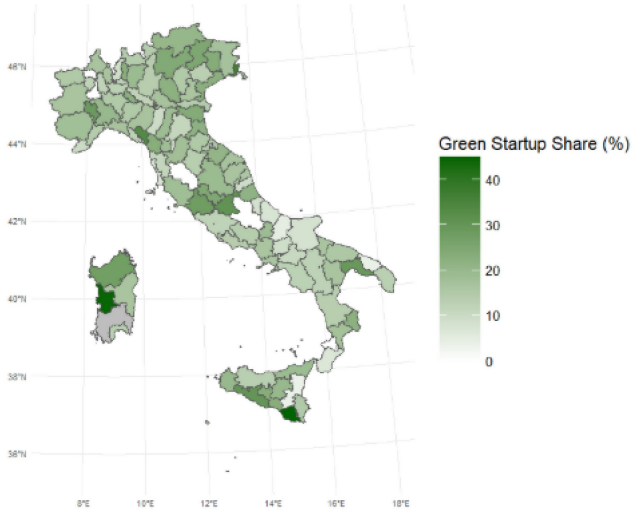


Fig. 6.10 Green startup share in Italian provinces

As a sanity check, in Table 6.1, we report the websites of 20 green startups randomly chosen and their green dimensions. Eighteen startups are linked to the following green dimensions: sustainable products, green mobility, renewable energy and energy efficiency, sustainable agriculture, sustainable mobility, and waste management. Two startups were arguably non-green. UPSKILL4.0 organizes training and workshops for developing innovative skills. While they list some green workshops they recently took part in on their website, it is an occasional occurrence. ACME21 produces devices to protect hives from theft and regulate humidity and temperature. The topics identified in this website have been linked to biodiversity protection and sustainable agriculture. Such false positives can arise from two steps of the algorithm. The first is the dimensionality-reduction step: UMAP is a state-of-the-art dimension-reducing algorithm; it includes stochasticity and does not lead to the exact same clusters on each iteration. We mitigate such an inconvenience by iterating the procedure as described above. The second is the cosine-similarity identification of green topics. While it allows identifying green topics automatically, some errors can occur. For instance, the cosine similarity of 'sustainable tourism' and 'tourism in the mountains' is 0.64, high enough to be identified as a green topic. We mitigate such a potential issue by exploring different algorithms at the topic representation step, ensuring

each topic is characterized by a diverse and comprehensive set of ten representative uni- to three-grams.

Further, we explored the ATECO 2-digit sectors of green startups in Table 6.2. The table shows that there are no startups identified as green that belong to industries related to fossil fuel production – ATECO number 05 (extraction of coal), 06 (extraction of oil and natural gas), 07 (extraction of metal minerals) and 19 (production of coke and products derived from oil); nor to airplane companies – ATECO n. 51 (air transport); nor to agricultural production and beef farming – ATECO number 01 (farming and production of animal products, hunting and related services). The startups primarily focus on a few sectors, as more than one third of the startups fall in two ATECO sectors, Software production, IT consultancy and related activities and Scientific research and development, and the top ten ATECO sectors cover 75% of all the startups.

Finally, as expected, we note that startups in our sample have, on average, a very small size (Fig. 6.11). Only four green startups report having more than 50 employees. Since larger firms may tend to invest more in greenwashing, we explore their websites. It turns out they are active in wastewater management solutions, precision agriculture, a company working towards obtaining the ISO14001 (Environmental Management System) certification and one selling windows with superior insulation properties, so we consider them all appropriately tagged. There are only 16 startups with more than 20 employees, which we checked manually, and none appeared at high risk of greenwashing. Therefore, we consider it unlikely that some large companies like airlines, fossil fuel producers, or beef farms are escaping our analysis.

Summing up, we introduced a methodology that allows a fully automated identification of green startups and that can be run locally. The researcher only has to input a green framework (in our case, the green targets from the SDGs); there is no need for manual intervention or external dictionaries, which ensures its replicability. The method appears to tag startups with green websites with satisfactory precision.

Table 6.1 Green startups and their green dimension.

Company name	Website URL	Green Dimension
NOICE	https://www.belowzero.it/prodotto/	Sustainable products
METAL CARBON DESIGN	https://www.metalcarbondesign.it/index.php/it/	Green mobility
GLIVEE	https://glivee.com	Sustainable products
OMEGAWIND	http://www.omegawind.com/chisiamo.html	Renewable energy
TRIENERGIA	https://www.trienergia.com	Renewable energy
IFLY	https://www.acquanoleggio.it/top-table/	Sustainable products
MR ENERGY SYSTEMS	https://www.mrenergy.it/consulting/	Energy efficiency
HYPERION	https://www.hyperionsrl.eu	Renewable energy
ACME21	https://antifurtoarnia.it/it/antifurto-e-sensori/	GPS to protect bee hives from theft and monitor heat and humidity (non-green)
UPSKILL 4.0	https://www.upskill40.it/blog/	Training and workshops, with a few projects on green sustainability (non-green)
ALMALANA	https://www.almalana.it	Sustainable agriculture
BIOVERDISSIMO	https://bioverdissimo.it/la-tecnologia/	Sustainable agriculture
IDEA ENERGIA	https://www.ideaenergia.it	Energy efficiency
UFARMER	https://www.ufarmer.it/assistenza/	Sustainable agriculture
GREEN RECYCLING	https://www.green-recycling.it	Sustainable agriculture
UP2GO	https://www.up2go.it/servizi/mobility-management/	Sustainable mobility
GO GREEN	http://www.gogreensrl.it/	Waste management
IKINOVA	https://www.ikinova.com/inovafarm.html	Sustainable agriculture
SMART SUN	https://www.smartsunsrl.it	Renewable energy

I.P.M.	https://www.systemipm.com/buildings/	Sustainable building

Table 6.2 ATECO distribution of the green startups sorted from most frequent to least frequent

ATECO 2 Digits	Title (Translated from Italian)	N. of startups	Cumulated Frequency (%)
62	Software production, IT consultancy and related activities	306	19.91%
72	Scientific research and development	292	38.91%
63	Information services activities and other computer services	109	46.00%
74	Other professional, scientific and technical activities	101	52.57%
28	Manufacture of machinery and equipment not elsewhere classified	89	58.36%
71	Activities of architecture and engineering firms; technical tests and analyses	70	62.91%
27	Manufacture of electrical equipment and non-electrical equipment for household use	57	66.62%
70	Corporate management and management consulting activities	48	69.75%
43	Specialized construction work	40	72.35%
26	Manufacture of computers and electronics and optics products; electromedical equipment, measuring equipment and clocks	29	74.24%
41	Construction of buildings	28	76.06%
46	Wholesale trade (excluding motor vehicles and motorcycles)	27	77.81%
35	Supply of electricity, gas, steam and air conditioning	25	79.44%
20	Manufacturing of chemicals	24	81.00%
82	Support activities for office functions and other business support services	23	82.50%
47	Retail trade (excluding motor vehicles and motorcycles)	23	83.99%
32	Other manufacturing industries	19	85.23%
79	Activities of travel agencies, tour operator services and booking services and related activities	18	86.40%
10	Food industry	14	87.31%

22	Manufacture of rubber and plastic items	13	88.16%
25	Manufacturing of metal products (excluding machinery and equipment)	13	89.00%
33	Repair, maintenance and installation of machines and equipment	11	89.72%
11	Beverage industry	11	90.44%
77	Rental and operational leasing activities	10	91.09%
30	Manufacture of other means of transport	10	91.74%
16	Wood and wood and cork products industry (excluding furniture); manufacturing of straw items and weaving materials	10	92.39%
38	Waste collection, treatment and disposal activities; recovery of materials	9	92.97%
12	Tobacco industry	8	93.49%
73	Advertising and market research	8	94.01%
29	Manufacture of motor vehicles, trailers and semi-trailers	8	94.53%
23	Manufacture of other non-metallifier mineral processing products	7	94.99%
15	Manufacture of leather and similar items	6	95.38%
58	Editorial activities	6	95.77%
85	Instruction	6	96.16%
81	Service activities for buildings and landscapes	5	96.49%
13	Textile industries	5	96.81%
61	Telecommunications	5	97.14%
31	Furniture manufacturing	5	97.46%
55	Accommodation	4	97.72%
68	Real estate activities	4	97.98%
96	Other personal service activities	3	98.18%
14	Packaging of clothing items; packaging of leather and fur items	3	98.37%
42	Civil engineering	3	98.57%
88	Non-residential social assistance	2	98.70%
66	Activities auxiliary to financial services and insurance activities	2	98.83%
90	Creative, artistic and entertainment activities	2	98.96%
17	Manufacturing of paper and paper products	2	99.09%
64	Financial services activities (excluding insurance and pension funds)	2	99.22%
59	Film, video and television program production activities, music and sound recordings	2	99.35%

56	Restaurant services activities	2	99.48%
95	Repair of computers and goods for personal and home use	1	99.54%
86	Health care	1	99.61%
52	Warehousing and transport support activities	1	99.67%
36	Collection, treatment and supply of water	1	99.74%
39	Recovery activities and other waste management services	1	99.80%
53	Postal services and courier activities	1	99.87%
49	Land transport and transport by pipelines	1	99.93%
45	Wholesale and retail trade and repair of motor vehicles and motorcycles	1	100.00%

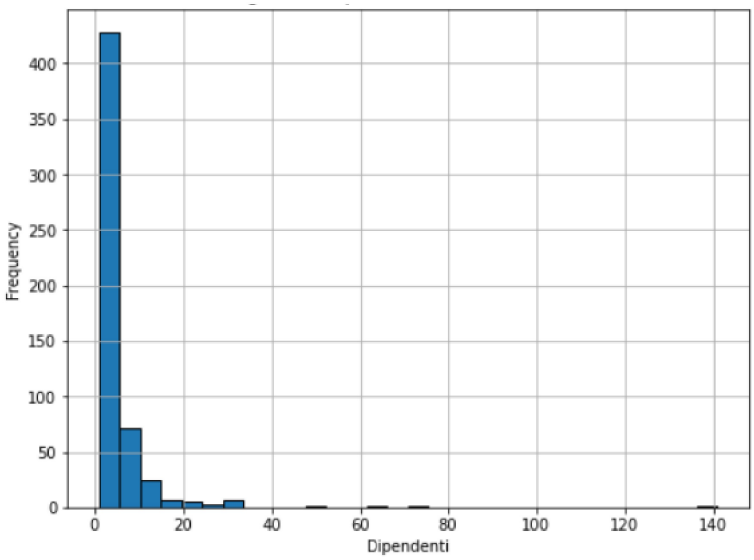


Fig. 6.11 Histogram of the number of employees for firms with more than 0 employees for readability

7.4 ADDITIONAL REFERENCES

Abdesselam, R., Kedjar, M., & Renou-Maissant, P. (2024). What are the drivers of eco-innovation? Empirical evidence from French start-ups. Technological Forecasting and Social Change, 198, 122953. https://doi.org/10.1016/j.techfore.2023.122953

Barbaresi, A. (2021). Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In H. Ji, J. C. Park, & R. Xia (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations (pp. 122–131). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-demo.15

Bevendorff, J., Gupta, S., Kiesel, J., & Stein, B. (2023). An Empirical Comparison of Web Content Extraction Algorithms. Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2594–2603. https://doi.org/10.1145/3539618.3591920

Bonfanti, A., De Crescenzo, V., Simeoni, F., & Loza Adaui, C. R. (2024). Convergences and divergences in sustainable entrepreneurship and social entrepreneurship research: A systematic review and

research agenda. Journal of Business Research, 170, 114336. https://doi.org/10.1016/j.jbusres.2023.114336

Borčin, M., & Jose, J. M. (2024). Optimizing BERTopic: Analysis and Reproducibility Study of Parameter Influences on Topic Modeling. In N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, & I. Ounis (Eds.), Advances in Information Retrieval (pp. 147–160). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-56066-8_14

Chapman, G., & Hottenrott, H. (2022). Green start-ups and the role of founder personality. Journal of Business Venturing Insights, 17, e00316. https://doi.org/10.1016/j.jbvi.2022.e00316

Cojoianu, T. F., Clark, G. L., Hoepner, A. G. F., Veneri, P., & Wójcik, D. (2020). Entrepreneurs for a low carbon world: How environmental knowledge and policy shape the creation and financing of green start-ups. Research Policy, 49(6), 103988. https://doi.org/10.1016/j.respol.2020.103988

Coll-Martínez, E., Kedjar, M., & Renou-Maissant, P. (2022). (Green) Knowledge spillovers and regional environmental support: Do they matter for the entry of new green tech-based firms? The Annals of Regional Science, 69(1), 119–161. https://doi.org/10.1007/s00168-022-01111-3

Colombelli, A., D'Ambrosio, A., & Ravetti, C. (2024). Women in innovative start-ups and regional inclusiveness: 'Green' and socially-responsible companies. Regional Studies, 1–14. https://doi.org/10.1080/00343404.2024.2340999

Colombelli, A., & Quatraro, F. (2019). Green start-ups and local knowledge spillovers from clean and dirty technologies. Small Business Economics, 52(4), 773–792. https://doi.org/10.1007/s11187-017-9934-y

Dong, S., Gong, H., & Liu, T. (2022). Environmental technology spillovers and green start-up emergence: The moderating role of patent commercialization policy and patent enforcement. Environmental Science and Pollution Research, 29(46), 70070–70083. https://doi.org/10.1007/s11356-022-20791-0

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Frontiers in Sociology, 7. https://doi.org/10.3389/fsoc.2022.886498

Gebhardt, L., & Bachmann, N. (2023). Entrepreneurial contributions to sustainability transitions—A longitudinal study of their representation and enactment through topic modeling and thematic analysis. Journal of Cleaner Production, 420, 138255. https://doi.org/10.1016/j.jclepro.2023.138255

Gidron, B., Bar, K., Finger Keren, M., Gafni, D., Hodara, Y., Krasnopolskaya, I., & Mannor, A. (2023). The Impact Tech Startup: Initial Findings on a New, SDG-Focused Organizational Category. Sustainability, 15(16), 12419. https://doi.org/10.3390/su151612419

Gorovaia, N., & Makrominas, M. (2024). Identifying greenwashing in corporate-social responsibility reports using natural-language processing. European Financial Management, eufm.12509. https://doi.org/10.1111/eufm.12509

Gast, J., Gundolf, K., & Cesinger, B. (2017). Doing business in a green way: A systematic review of the ecological sustainability entrepreneurship literature and future research directions. Journal of Cleaner Production, 147, 44–56. https://doi.org/10.1016/j.jclepro.2017.01.065

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure (arXiv:2203.05794). arXiv. https://doi.org/10.48550/arXiv.2203.05794

Horne, J., Recker, M., Michelfelder, I., Jay, J., & Kratzer, J. (2020). Exploring entrepreneurship related to the sustainable development goals—Mapping new venture activities with semi-automated content analysis. Journal of Cleaner Production, 242, 118052. https://doi.org/10.1016/j.jclepro.2019.118052

Jha, V. K., & Pande, A. S. (2024). Making sustainable development happen: Does sustainable entrepreneurship make nations more sustainable? Journal of Cleaner Production, 440, 140849. https://doi.org/10.1016/j.jclepro.2024.140849https://doi.org/10.1016/j.jclepro.2024.140849

Kuckertz, A., Berger, E. S. C., & Gaudig, A. (2019). Responding to the greatest challenges? Value creation in ecological startups. Journal of Cleaner Production, 230, 1138–1147. https://doi.org/10.1016/j.jclepro.2019.05.149

Li, Z., Shang, W., & Yan, M. (2016). News text classification model based on topic model. 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), 1–5. https://doi.org/10.1109/ICIS.2016.7550929

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. The Journal of Open Source Software, 2(11), 205. https://doi.org/10.21105/joss.00205

McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (arXiv:1802.03426). arXiv. https://doi.org/10.48550/arXiv.1802.03426

Mio, C., Panfilo, S., & Blundo, B. (2020). Sustainable development goals and the strategic role of business: A systematic literature review. Business Strategy and the Environment, 29(8), 3220–3245. https://doi.org/10.1002/bse.2568

Mrkajic, B., Murtinu, S., & Scalera, V. G. (2019). Is green the new gold? Venture capital and green entrepreneurship. Small Business Economics, 52(4), 929–950. https://doi.org/10.1007/s11187-017-9943-x

Purvis, B., Mao, Y., & Robinson, D. (2019). Three pillars of sustainability: In search of conceptual origins. Sustainability Science, 14(3), 681–695. https://doi.org/10.1007/s11625-018-0627-5

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (arXiv:1908.10084). arXiv. https://doi.org/10.48550/arXiv.1908.10084

Schaltegger, S., & Wagner, M. (2011). Sustainable entrepreneurship and sustainability innovation: Categories and interactions. Business Strategy and the Environment, 20(4), 222–237. https://doi.org/10.1002/bse.682

Tiba, S., Van Rijnsoever, F. J., & Hekkert, M. P. (2021). Sustainability startups and where to find them: Investigating the share of sustainability startups across entrepreneurial ecosystems and the

causal drivers of differences. Journal of Cleaner Production, 306, 127054. https://doi.org/10.1016/j.jclepro.2021.127054

Umamaheswaran, S., Dar, V., Sharma, E., & Kurian, J. S. (2023). Mapping Climate Themes From 2008-2021—An Analysis of Business News Using Topic Models. IEEE Access, 11, 26554–26565. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3256530

UNEP. (2019). Global Environment Outlook 6 (GEO6). United Nations Environment Programme. https://www.unep.org/resources/global-environment-outlook-6

Wöhler, J., & Haase, E. (2022). Exploring investment processes between traditional venture capital investors and sustainable start-ups. Journal of Cleaner Production, 377, 134318. https://doi.org/10.1016/j.jclepro.2022.134318

8 APPENDIX C

8.1 PRINCIPAL COMPONENT ANALYSIS (PCA): FACTOR LOADINGS

Table 7.1 Factor loadings for PCA

Variable	**Component 1: *GREENDEMAND***
Availability of bike-sharing services	0.4064
Availability of car-sharing services	0.4358
Bike lane density	0.4899
Local public transportation demand	0.3790
Days exceeding pm10 limit	0.4297
Share sorted waste	0.2494
Motorisation rate	-0.1207

The factor loadings of the first principal component align with our expectations about the relationships of the underlying behavioural variables with green demand. Indeed, the first principal component is increasing with the availability of sustainable urban transportation services, such as bike-sharing and car-sharing services and bike lane density; it increases with the demand for public local transportation services and decreases with higher motorization rates. Moreover, our proxy for green demand increases with another key proxy of environmentally sustainable behaviours: the share of recycled/sorted waste. Finally, in line with our interpretation of the first component as a proxy for green demand, higher levels of pollution (days exceeding the PM10 limit) have a positive factor loading. Overall, we conclude that the first component of the PCA provides a reasonable approximation of green demand. Fig. 3.5 in the main text shows the green demand scores in Italian provinces in 2021.

8.2 VARIABLES' DESCRIPTION, SUMMARY STATISTICS AND CORRELATION MATRIX

Table 7.2 Variables' description

Variable	Description
KSTOCK	Annual patent application stocks for each NUTS3 region, KSTOCK, with the permanent inventory method, with an obsolescence rate of 15% per annum (Hall et al., 2005). To mitigate the high right-skewness of the variable, which is typical of patent data, we apply an inverse hyperbolic sine transformation to the estimated stock when we include it in the analysis.
GPSH	Based on Cooperative Patent Classification (CPC) code, all patents falling in the Y02 class and, for each year and province, compute the share of the patent stock that is attributable to green patents as GPSH_(i,t)= 〖GREEN_KSTOCK〗_(i,t)/KSTOCK_(i,t)
GREENDEMAND	Principal component analysis of: availability of bike-sharing services; availability of car-sharing services; bike lane density; demand for local public transportation services; days exceeding the acceptable limit of PM10 concentration in urban air; share of recycled/sorted waste on total waste; motorisation rate (i.e., motor-vehicles per inhabitant).
Control variables	
FIRM_DENS	*Firm density.* Many studies include the density of population or firms to control for the size of the economy and the probability that agglomeration economies arise. Density facilitates the interactions between economic agents, the exchange of knowledge and the diffusion of information about business opportunities. Accordingly, previous studies have found that it positively affects firm formation (Armington & Acs, 2002; Fritsch & Falck, 2007). We include firm density and define it as the ratio between the total number of registered firms and the regional land-use area.
PC_VADDED	*Per capita value-added.* We include the log of per-capita value added in our analysis, measured as the ratio between value-added and population, to control for the general role of demand and purchasing power as an incentive for prospective entrepreneurs (Carree & Thurik, 1996). Importantly, ceteris paribus, higher disposable income can be expected to allow consumers to choose from a broader range of products and services with differing degrees of greenness and price; hence, it is likely to interact in interesting ways with our measure of green demand.
GRADSH	*Share of graduates.* Several studies find that higher education is positively related to innovation, entrepreneurship, and green entrepreneurship (Colombelli & Quatraro, 2019; Giudici et al., 2019; Hoogendoorn et al., 2020; Koellinger, 2008; Schøtt & Jensen, 2016). Hence, we include the share of graduates over

	the total population.
MANUF_SH	*Manufacturing share.* Startup rates, knowledge stocks and green demand may correlate with the composition of the local economy. To control for this correlation, we include the share of manufacturing firms over total firms.
UNEMPL	*Unemployment rate (UNEMPL$_{i,t}$)*. The effects of unemployment on entrepreneurship rates are debated in the literature. Some authors argue that more employment proxies for more opportunities (Reynolds et al., 1994; Sutaria & Hicks, 2004), while others suggest that entrepreneurship emerges as a survival strategy against unemployment (Wagner & Sternberg, 2004). Regarding green firm formation, we expect the first mechanism to prevail: due to the high degree of uncertainty highlighted in Section 2 concerning firm formation, we do not expect green entrepreneurship to emerge as survival entrepreneurship. Hence, we expect a negative relationship with green startup rates.

Table 7.3 Summary statistics

Variable	Mean	Std. dev.	Min	Max	Median	Skewness	Kurtosis
GREEN_STARTUPS	1.39	2.97	0	37	1	6.03	52.55
ENERGY_STARTUPS	0.55	1.23	0	12	0	4.18	27.61
KSTOCK	5.40	1.80	0.43	9.16	5.49	-0.38	2.55
GPSH	0.16	0.14	0.03	1	0.11	3.05	15.28
GREENDEMAND	-.062	1.11	-1.56	2.31	-0.25	0.41	1.89
PC_VADDED	10.95	0.13	10.65	11.29	10.96	-0.08	2.37
GRADSH	5.01	1.02	1.89	8.52	4.88	0.26	2.89
MANUF_SH	0.09	.029	0.04	.26	0.09	2.12	11.31
FIRM_DENS	20.09	29.08	1.72	203.24	12.40	4.29	22.97
UNEMPL	11.49	5.69	2.88	31.46	9.45	0.96	3.06

Observations: 952.

Table 7.4 Correlation matrix

	1	2	3	4	5	6	7	8	9	10
1.GREEN_STARTUPS	1.00									
2.ENERGY_STARTUPS	0.68	1.00								
3.KSTOCK	0.41	0.35	1.00							
4.GPSH	-0.17	-0.17	-0.74	1.00						
5.GREENDEMAN$_i$	0.37	0.30	0.75	-0.48	1.00					
6.PC_VADDED	0.35	0.25	0.75	-0.54	0.69	1.00				
7.GRADSH	0.03	0.03	-0.36	0.35	-0.38	-0.52	1.00			
8.MANUF_SH	-0.11	-0.10	0.18	-0.20	0.19	0.10	-0.19	1.00		
9.FIRM_DENS	0.60	0.50	0.42	-0.22	0.38	0.35	-0.05	0.07	1.00	
10.UNEMPL	-0.13	-0.05	-0.60	0.51	-0.56	-0.73	0.50	-0.32	-0.11	1.00

8.3 ROBUSTNESS CHECKS

In Table 7.5, we report the estimates from two separate models: (i) one that regresses green firm formation on green knowledge stocks, green demand, their interaction, and controls, and (ii) one that regresses green firm formation on non-green knowledge stocks, green demand, their interaction, and controls. In Fig 7.1, we report the corresponding interaction effects on the same graph for comparison. If the greenness of knowledge mattered more, the estimated impact of the green stock should be greater than that of the non-green stock. Instead, both the table and the figure clearly show that the estimated impact of green knowledge on green entrepreneurship, while positive and significant and significantly mediated by green demand, is empirically indistinguishable from that of non-green knowledge. This result confirms that the size of knowledge matters rather than its composition.

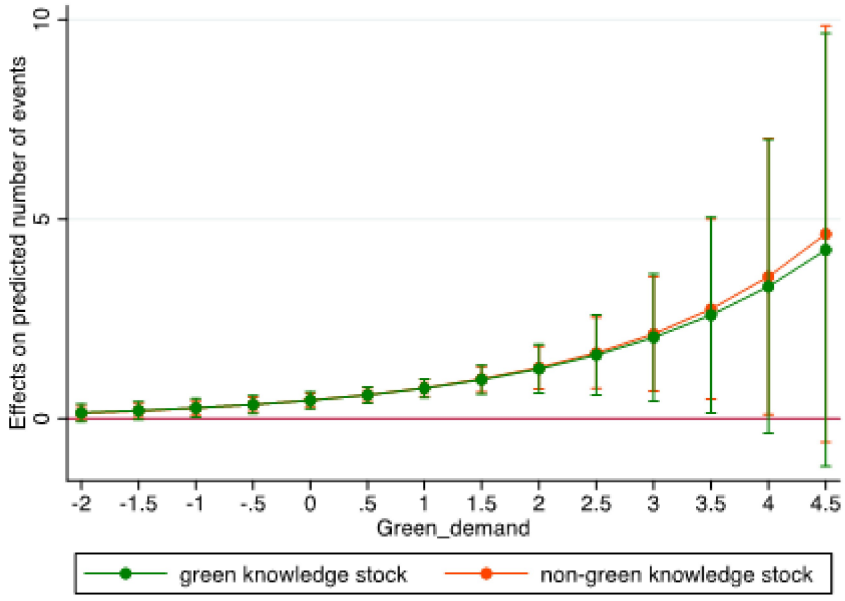
Table 7.5 Green firm formation as a function of green and non-green knowledge stocks

	Dependent variable: green innovative startups							
	Neg. Bin. regression coefficients				Marginal effects			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Patent stock measure	**Green stock**		**Non-green stock**		**Green stock**		**Non-green stock**	
Patent stock	0.488***	0.457***	0.434***	0.487***	0.664***	0.740***	0.589***	0.823***
	(0.089)	(0.086)	(0.098)	(0.086)	(0.125)	(0.133)	(0.137)	(0.158)

GREENDEMAND	0.291***	-0.112	0.266***	-0.629**	0.396***	0.485***	0.361***	0.427***
	(0.081)	(0.185)	(0.084)	(0.316)	(0.111)	(0.167)	(0.114)	(0.147)
Patent stock *GREENDEMAND*		0.100*		0.138***				
		(0.0576)		(0.0530)				
N	952	952	952	952	952	952	952	952
Controls	YES	YES	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES	YES	YES

Columns 1-4: Negative binomial regressions coefficients. Standard errors clustered at the regional level in parentheses. Columns 5-6: Marginal effects (predicted number of startups) corresponding to the specifications in columns 1-4. Delta method standard errors are in parentheses. All specifications include regional dummies, year dummies, and control variables (log value added per capita, the share of graduates, the share of manufacturing firms, firm density, and unemployment rate). All regressors lagged one year. *p<0.10, **p<0.05, ***p<0.01.

Fig. 7.1 Average marginal effects of green and non-green knowledge stocks at different levels of green demand on the generation of green startups



In Table 7.6, we test an alternative definition of green startups: startups with high value-added in the energy sector. Results for the patent stock, green demand and its interaction are confirmed.

Table 7.6 Energy firm formation

	Dependent variable: energy innovative startups			
	Neg. Bin. regression coefficients		Marginal effects	
	(1)	(2)	(3)	(4)
KSTOCK	0.544***	0.526***	0.298***	0.330***
	(0.0779)	(0.0663)	(0.0428)	(0.0452)
GPSH	-1.047	-2.508	-0.574	-1.339
	(0.962)	(2.061)	(0.529)	(1.621)
GREENDEMAND	0.214***	-0.491	0.118***	0.136***
	(0.0807)	(0.419)	(0.0444)	(0.0504)
KSTOCK ´ GREENDEMAND		0.107*		
		(0.0574)		
GPSH ´ GREENDEMAND		0.109		
		(1.436)		
N	952	952	952	952
Controls	YES	YES	YES	YES
Region FE	YES	YES	YES	YES
Year FE	YES	YES	YES	YES

Columns 1-2: Negative binomial regressions coefficients. Standard errors clustered at the regional level in parentheses. Columns 3-4: Marginal effects (predicted number of startups) corresponding to the specifications in columns 1-2. Delta method standard errors are in parentheses. All specifications include regional dummies, year dummies, and control variables (log value added per capita, the share of graduates, the share of manufacturing firms, firm density, and unemployment rate). All regressors lagged one year. *p<0.10, **p<0.05, ***p<0.01.

Next, we test two alternative estimators for our regressions: Poisson and Zero Inflated Negative Binomial regressions (Table 7.7). For both estimators, the results of the graphical analysis of marginal effects are very similar to the

ones reported in Fig. 3.6 and Fig. 3.7 and confirm the magnifying effect of green demand on the local knowledge stock. We do not report these figures for brevity, but they are available upon request.

Table 7.7 Green firm formation – Alternative estimators

	Dependent variable: green innovative startups							
	Regression coefficients				Marginal effects			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Poisson		**Zero-inflated NB**		**Poisson**		**Zero-inflated NB**	
KSTOCK	0.490***	0.448***	0.434***	0.406***	0.680***	0.798***	0.593***	0.725***
	(0.0936)	(0.0778)	(0.0934)	(0.0839)	(0.130)	(0.145)	(0.129)	(0.153)
GPSH	0.860	-2.125	1.359	-1.714	1.194	-3.699	1.857	-3.181
	(1.221)	(1.623)	(1.357)	(1.739)	(1.696)	(3.117)	(1.857)	(3.203)
GREENDEMAND	0.256***	-0.680*	0.247***	-0.651	0.356***	0.439***	0.338***	0.414***
	(0.0847)	(0.381)	(0.0843)	(0.397)	(0.118)	(0.149)	(0.115)	(0.147)
KSTOCK GREENDEMAND		0.153**		0.148**				
		(0.0622)		(0.0631)				
GPSH GREENDEMAND		-0.649		-0.731				
		(0.954)		(0.940)				
N	952	952	952	952	952	952	952	952
Controls	YES	YES	YES	YES	YES	YES	YES	YES
Region FE	YES	YES	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES	YES	YES

Columns 1-2: Poisson PML regression coefficients; Columns 3-4: Zero-inflated negative binomial regression coefficients; Zero-inflating factor included: log number of firms at time *t-1*. Standard errors clustered at the regional level in parentheses. Columns 5-6: Marginal effects (predicted number of startups) from the corresponding Poisson and ZINB regressions reported. Delta method standard errors in parentheses. All specifications include regional dummies, year dummies, and control variables (log value added per-capita, share of graduates, share of manufacturing firms, firm density, unemployment rate). All regressors lagged one year. *p<0.10, **p<0.05, ***p<0.01.

As a final robustness check, we add a control measure of local STEM graduates to capture some of the tacit knowledge available in each location that our patent stocks may not capture fully (Table 7.8). The table shows that the main text results are robust.

Table 7.8 Robustness check – Additional measure of knowledge

	Dependent variable: green innovative startups			
	Neg. Bin. regression coefficients		**Marginal effects**	
	(1)	(2)	(3)	(4)
KSTOCK	0.281***	0.292***	0.477***	0.688***
	(0.098)	(0.091)	(0.166)	(0.173)
GPSH	-1.790	-3.505	-3.041	-5.924
	(1.438)	(2.167)	(2.446)	(5.265)
GREENDEMAND	0.236**	-0.579**	0.401**	0.517***
	(0.110)	(0.225)	(0.187)	(0.200)
STEMGRAD	0.050***	0.035**	0.085***	0.059**
	(0.018)	(0.017)	(0.031)	(0.030)
KSTOCK ´ GREENDEMAND		0.123***		
		(0.043)		
GPSH ´ GREENDEMAND		0.035		
		(1.382)		
N	710	710	710	710
Region FE	Yes	Yes	YES	YES
Year FE	Yes	Yes	YES	YES

Columns 1-2: Negative binomial regressions coefficients. Standard errors clustered at the regional level in parentheses. Columns 3-4: Marginal effects (predicted number of startups) corresponding to the specifications in columns 1-2. Delta method standard errors are in parentheses. All specifications include regional dummies, year dummies, and control variables (log value added per capita, the share of manufacturing firms, firm density, and unemployment rate). The STEM graduates *STEMGRAD* variable is computed as the inverse-hyperbolic sine transformed yearly number of graduates. All regressors lagged one year. *p<0.10, **p<0.05, ***p<0.01.

9 APPENDIX D

9.1 DETAILS ON THE DEPENDANT VARIABLES

To identify automatically legitimacy claims, we tasked ChatGPT OSS20B to perform qualitative analysis by identifying which codes were relevant to each post. In the next parts, we detail the prompts and the codebooks.

9.1.1 Prompts

Below are the prompts used to identify the normative, cognitive and pragmatic legitimacy claims in the linkedin posts. Note that only one prompt was used for the two legitimacy and pragmatic legitimacy claims to save computational time.

SDG Coding Prompt

You are a qualitative coder annotating LinkedIn posts written by entrepreneurs.

Below you will find:

A codebook containing the 17 United Nations Sustainable Development Goals (SDGs).

A LinkedIn post.

Your Task

Carefully read the SDG codebook.

Analyze the LinkedIn post.

Identify which SDG(s), if any, apply to the content of the post.

For each SDG selected, provide a brief explanation of why it is relevant.

If no SDG applies, state that clearly and explain why.

Codebook

{sdg_codebook}

Post

{post_text}

Cognitive and Pragmatic Legitimacy Coding Prompt

You are a qualitative coder annotating LinkedIn posts written by entrepreneurs.

Below you will find:

A codebook.

A LinkedIn post.

Your Task

Carefully read the codebook.

Analyze the LinkedIn post.

Identify which code(s), if any, apply to the content of the post.

For each selected code, briefly explain why it is relevant.

If no code applies, clearly state that and explain why.

Codebook

{cognitive_pragmatic_codebook}

Post

{post_text}

9.1.2 Codebooks

Table 8.1 present the codebook used to identify posts related to SDGs and their grouping into Social, Green and Economic SDG.

Table 8.1 SDG codebook and grouping

SDG 1	End poverty in all its forms everywhere	Social SDG
SDG 2	End hunger, achieve food security and improved nutrition, and promote sustainable agriculture	Social SDG
SDG 3	Ensure healthy lives and promote well-being for all at all ages	Social SDG

SDG 4	Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all	Social SDG
SDG 5	Achieve gender equality and empower all women and girls	Social SDG
SDG 6	Ensure availability and sustainable management of water and sanitation for all	Green SDG
SDG 7	Ensure access to affordable, reliable, sustainable and modern energy for all	Green SDG
SDG 8	Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all	Economic SDG
SDG 9	Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation	Economic SDG
SDG 10	Reduce inequality within and among countries	Social SDG
SDG 11	Make cities and human settlements inclusive, safe, resilient and sustainable	Economic SDG
SDG 12	Ensure sustainable consumption and production patterns	Green SDG
SDG 13	Take urgent action to combat climate change and its impacts	Green SDG
SDG 14	Conserve and sustainably use the oceans, seas and marine resources for sustainable development	Green SDG
SDG 15	Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss	Green SDG
SDG 16	Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels	Social SDG
SDG 17	Strengthen the means of implementation and revitalize the Global Partnership for Sustainable Development	Economic SDG

Table 8.2 present the codebook used to identify posts related to the cognitive legitimacy claims, and Table 8.3 to the pragmatic legitimacy claims. Those two codebooks were merged to save computational time.

Table 8.2 Cognitive legitimacy codebook

C1	Describe the products or services of the firm.
C2	Describe the actions of the firms.
C3	Show that the firm is similar to established firms in the market.
C4	Show that the firm is different to

	established firms in the market.

Table 8.3 Pragmatic legitimacy codebook

P1	Emphasize technological innovation or advanced capabilities of the firm.
P2	Emphasize the achievements of the firm.
P3	Highlight financial performance of the firm.
P4	Highlight financial support

9.2 DETAILS ON THE CONTROL VARIABLES

For the educational level and the education category, we explored the most frequent keywords present in the relevant fields, and we grouped them into the groups of Table 8.4 and Table 8.5. Some words are in Italian due to the high prevalence of Italian LinkedIn account. We identified only STEM and Economics educational categories because very few words were related to potential other fields such as humanities.

Table 8.4 Education level keywords

Level	Description	Included Keywords
4	Doctoral-level degree	Phd, doctor, dottor
3	Master-level degree	Master, msc, mba
2	Bachelor-level degree	Bachelor, laurea
1	Sub-degree/vocational	Diploma , certificate, liceo, istituto, formazione, degree, certificazione

Table 8.5 Education category keywords

Category	Included Keywords
STEM	Engineering, ingegneria, informatica
Economics	Business, management, marketing, economics managerial