POLITECNICO DI TORINO
Repository ISTITUZIONALE

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

Original
Knowledge spillovers, green entrepreneurship and the demand for sustainability: evidence from Italian innovative startups / Colombelli, A., D'Ambrosio, A., Le Masle, B., Ravetti, C., Tubiana, M.. - In: THE JOURNAL OF TECHNOLOGY TRANSFER. - ISSN 0892-9912. - (2026). [10.1007/s10961-025-10224-8]

Availability:
This version is available at: 11583/2997927 since: 2025-02-27T01:40:26Z

Publisher:
Springer

Published
DOI:10.1007/s10961-025-10224-8

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

08 November 2026

ONLINE APPENDIX

Appendix A

A.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 A 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.

A.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. Figure A1 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.

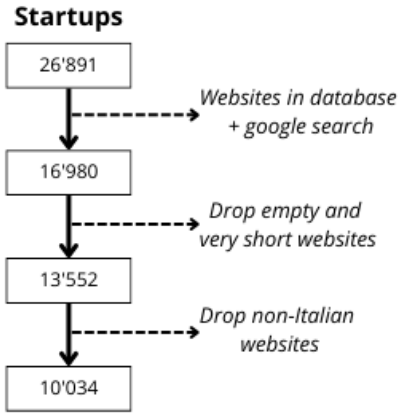


Figure A1. Data cleaning steps.

The resulting sample has a geographical distribution equivalent to that of the whole population at the regional and provincial levels (Figure A2 and Figure A3).

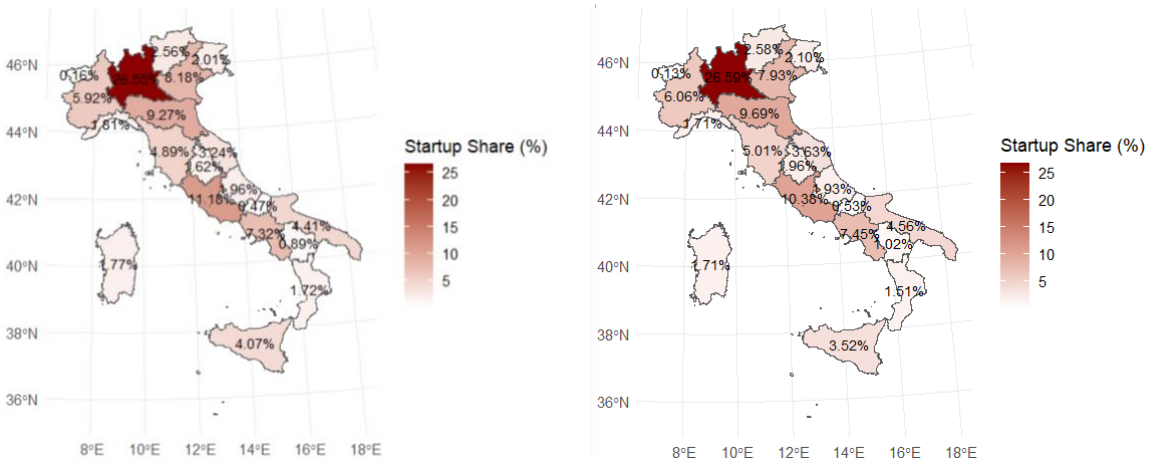


Figure A2. Share of innovative startups at the regional level, whole population (left) and scrapped sample (right).

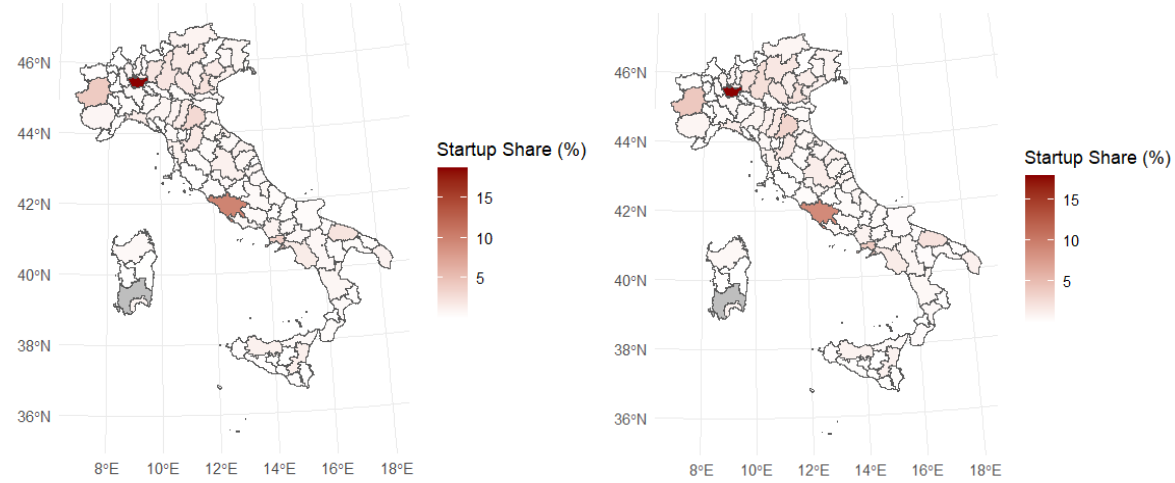


Figure A3. Share of innovative startups at the provincial level, whole population (left) and scrapped sample (right). For readability, we omit the percentages within each province.

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 Figure A4 for their frequency distribution).

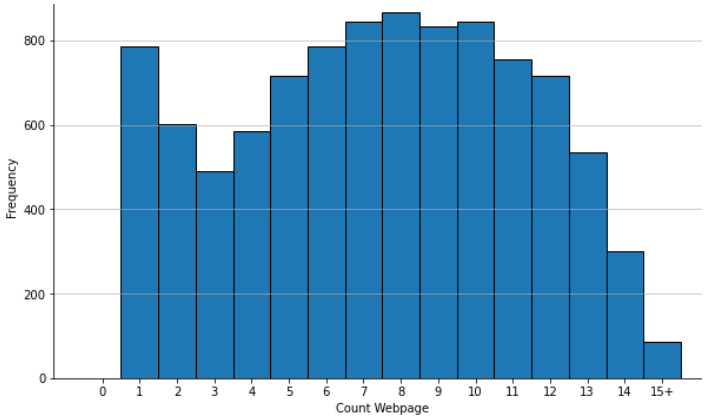


Figure A4. Distribution of the webpage counts per website.

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).

A.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 Figure A5. The following sections explain in detail each step of the application of BERTopic.

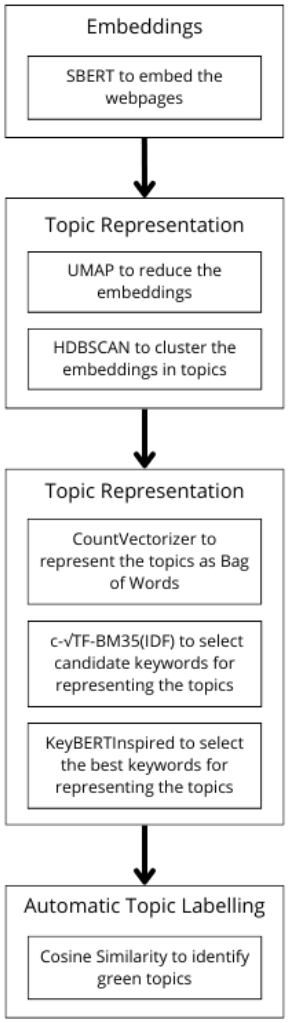


Figure A5. Step-by-step implementation of the BERTopic algorithm.

i. 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. Figure A6 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.

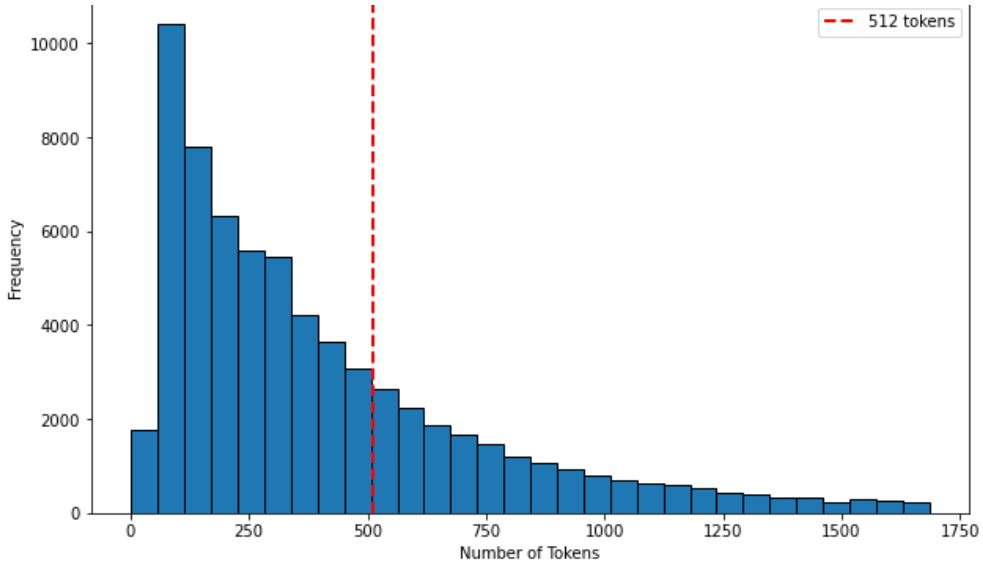


Figure A6. Distribution of token count per webpage after the cleaning steps. We represented only 95% of the distribution as a few outliers made the plot unreadable. Only the first 512 tokens of each webpage are processed by SentenceBERT.

ii. 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.

iii. Generating topic representation

We now create a topic representation for each topic so we can understand them. Figure A7 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 $w_{x,c} = |tf_{x,c}| * log\left(1 + \frac{A}{f_x}\right)$,

with

$$tf_{x,c} = frequency\ of\ word\ x\ in\ class\ c,$$

$$f_x = frequency\ of\ word\ x\ across\ all\ classes,$$

$$A = average\ number\ of\ words\ per\ class.$$

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 $w_{x,c} = |\sqrt{tf_{x,c}}| * log\left(1 + \frac{A - f_x + 0.5}{f_x - 0.5}\right)$, 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.

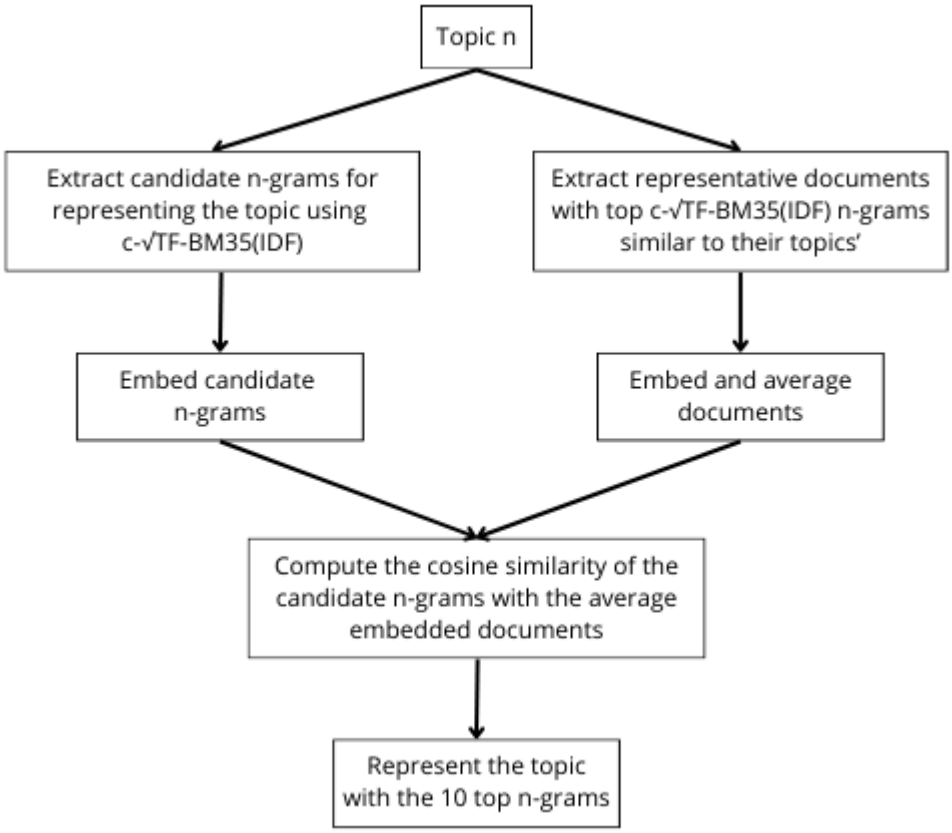


Figure A7. Infographic of the KeyBERTInspired topic representation algorithm.

iv. 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 Figure A8). 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.

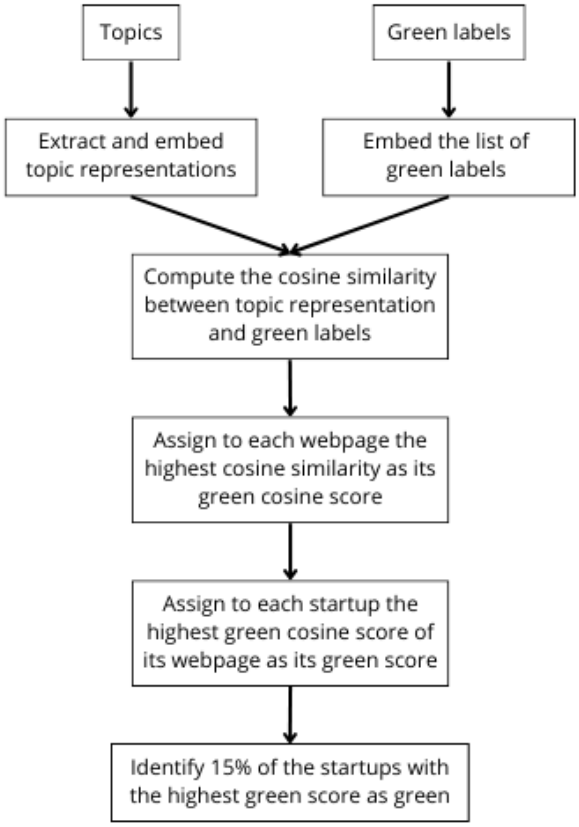


Figure A8. Infographic of the automatic labelling using cosine similarity.

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.[1] Figure A9 shows the distribution of green votes across green startups.

[1] We also tried using the green score average for identifying green startups, but it led to worse results as some webpages had high standard deviation in their green cosine score. This is due to the fact that UMAP, reducing the dimensions of the embeddings, sometimes hide their green dimensions.

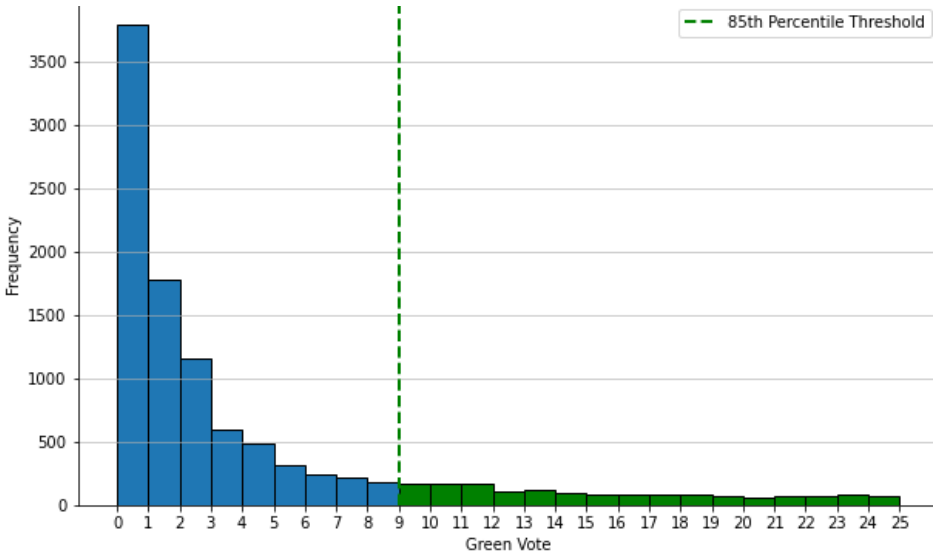


Figure A9. Distribution of startups' green vote on 25 iterations. Over 3500 startups are never voted green, while very few startups are labelled green on each iteration. We identify startups as green when they are labelled so at least in nine iterations.

Figure A10 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.

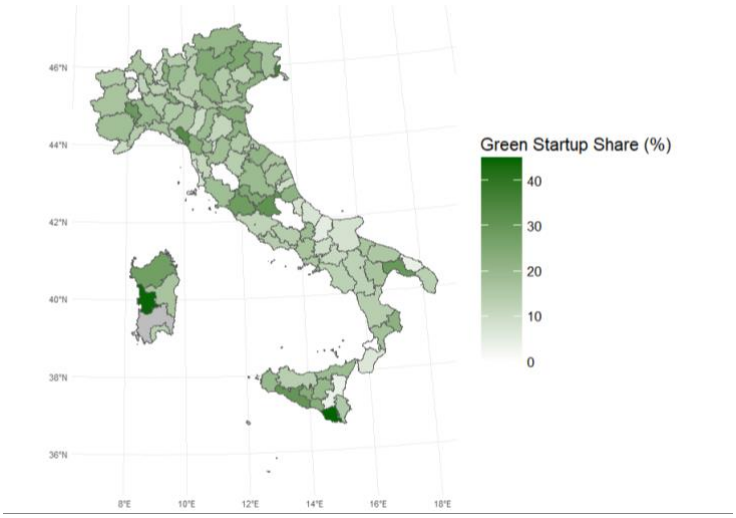


Figure A10. Green startup share in Italian provinces.

As a sanity check, in Table A1, 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 A2. 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 (Figure A11). 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 A1. 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 A2. 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%

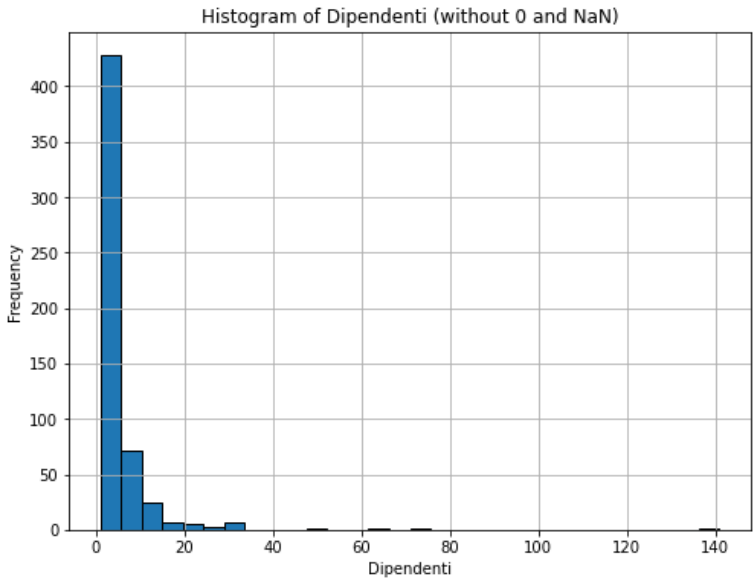


Figure A11. Histogram of the number of employees for firms with more than 0 employees for readability.

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

Appendix B

B.1 Principal component analysis (PCA): factor loadings

Table B1. 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. Figure 5 in the main text shows the green demand scores in Italian provinces in 2021.

B.2 Variables' description, summary statistics and correlation matrix

Table B2. 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 B3. 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 B4. 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

B.3 Robustness checks

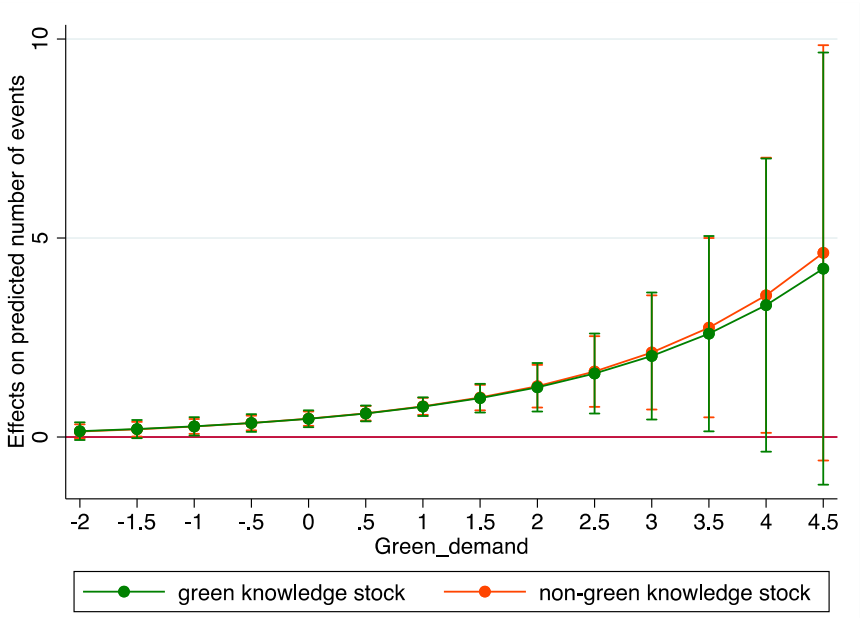
In Table B5, we report the estimates from two separate models: (i) one that regresses green firm formation on green knowledge stocks, green demand, their interaction, and controls, and (ii) one that regresses green firm formation on non-green knowledge stocks, green demand, their interaction, and controls. In Figure B1, 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 B5. 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.

Figure B1. Average marginal effects of green and non-green knowledge stocks at different levels of green demand on the generation of green startups.



In Table B6, 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 B6. 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 B7). For both estimators, the results of the graphical analysis of marginal effects are very similar to the ones reported in Figure 6 and Figure 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 B7. 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 B8). The table shows that the main text results are robust.

Table B8. 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.