AI-Supported Innovation Monitoring

Barteld Braaksma[1], Piet Daas[1], Stephan Raaijmakers[2,3],
Amber Geurts[2], André Meyer-Vitali[2]

[1] CBS, Statistics Netherlands, The Hague/Heerlen, NL
[2] TNO, Netherlands Research Organization, The Hague, NL
[3] Leiden University Centre for Linguistics (LUCL), Leiden University, NL

Abstract. Small and medium enterprises (SME) are a driving force for innovation. The stimulation of innovation in these SME shall be the target for political interventions, both regionally and nationally. Which technical areas should be in the focus? In this position paper, we propose AI-supported methods to combine innovation monitoring using natural language processing (NLP) and a dynamic knowledge graph that combines learning, reasoning and knowledge sharing in collaboration with innovation experts (Hybrid AI).

Keywords: Natural Language Processing, Knowledge Graph, Hybrid AI, Innovation, Policy-Making, Human-Machine Interaction.

1 Introduction

1.1 Innovation Scanning: Research Questions

Companies utilize innovation scanning services, such as Google Trends, TIM and Quid, to provide insights into innovation and technology trends. Such scans often use keywords, irrespective of human expert involvement, to measure real-time and emerging trends. However, such approaches are often inadequate when dealing with innovative novel phenomena such as emerging technologies or innovations as the terminology of some of the trends of interest does not yet exist or the meaning of existing terms in the area of interest changes over time [1]. Combined with the accelerating pace of technological innovation and social change [2], policy makers face the continuing challenge to navigate an increasingly complex space of strategic policy options [3].

In this position paper, we therefore explore ways to embed an AI-driven innovation monitor in the policy making process. To do so, we propose to detect trends in the innovation ecosystem using an AI-supported innovation monitor in which there is close collaboration between an AI-based system and an expert that is able to explain its statistical findings, specifically in text. Using such an evidence-based, hybrid approach in the policy making process, it is essential for the AI system to be able to monitor the ecosystem, present innovations and to be embedded in the policy-making process. Therefore, the first research question is of socioeconomic nature: *how to embed an AI-driven innovation monitor in the process of policy making?*

A complicating factor for such an AI-supported innovation monitor is that the language used in innovative fields changes quite rapidly [1, 4]. The extraction of relevant terminology and definitions of emerging technologies in these fields is crucial, and – once linked to a semantic domain representation like a knowledge graph – can help to detect upcoming trends upon which policy makers can base their decisions. Therefore, the second research question is: *how to extract concepts from technical and socio-economic domains and associate them with a dynamic knowledge graph*?

We address our research questions in two pilots. The results and insights from these two pilot studies will be merged into a planned, principled approach for AI-supported innovation monitoring.

1.2 Pilot Study 1: Detecting Small Innovative Companies

In pilot 1, Statistics Netherlands (CBS) has developed a novel approach to detect innovative companies. The method detects them by studying the text on the main page of a company's website. Traditionally, innovation is determined at CBS by sending a questionnaire to a sample of companies. This approach, however, only focuses on a random sample of the large and medium-sized companies and completely misses the small enterprises, such as start-ups, that commonly develop and commercialize innovative technologies and products. As a consequence, innovation monitoring usually anticipates change in the environment, interpret its consequences and develop future courses of action for responding to such change, but tends to miss the emergence of novel trends [5]. To enable this, a logistic regression bag-of-words based model was developed by CBS based on the webpage texts of the companies included in the innovation survey.

The final model had an accuracy of 88% and contained around 600 stemmed words including word embeddings [10]. The text-based method enables the detection of large, medium-sized and small innovative companies as long as they have a website. The latter was found to be the case for 99.9% of all innovative companies in the Netherlands. With the text-based approach the officially estimated number of large innovative companies by CBS, i.e. 19,916 ± 680, could be reproduced, i.e. 19,276 ± 190 [10]. This clearly demonstrated the high-quality results of the approach developed. For the small companies, a total of 33,599 innovative companies were detected in the Netherlands. Because location information is available, detailed maps of the Netherlands can be created displaying the distribution of innovative companies over the country (see Fig. 1) and within neighbourhoods of cities (not shown). The numbers of innovative companies are shown, at the municipality level, for the Netherlands. In a similar way, more specific models can be developed to detect innovative companies active in specific areas, such as 5G, hydrogen and mobility, for instance. This information can be embedded in the policy making process to enable policy makers to make decisions regarding how to spur early innovative trends in local or regional settings.

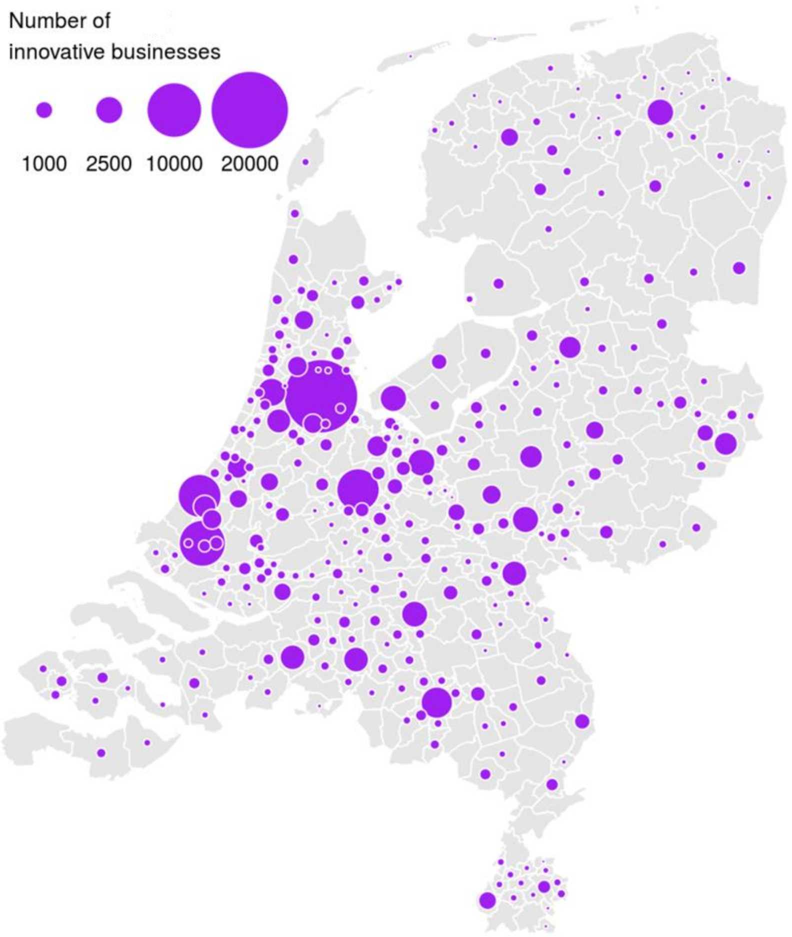

Fig. 1. Estimated number of innovative companies, excluding those of the (semi-)self-employed, in the Netherlands at the municipality level. A total of 52,875 innovative companies are shown.

1.3 Pilot Study 2: A Hybrid Intelligence Approach

When trying to capture emerging trends in innovation and technologies for which the terminology of these trends of interest does not yet exist or the meaning of existing terms in the area of interest changes over time [1], ontologies play an important role. To capture both domain knowledge and the emergence of new trends, a *hybrid intelligence* method has been developed in which domain experts and machine-driven insights are reiterated to construct and refine an ontology and knowledge graph via a cyclic process [2][7]. This hybrid method has been applied in a case study addressing Mobility-as-a-Service (MaaS). MaaS brings together the planning, booking and payment of all possible transportation options via apps. Such transportation options are not limited to trains, trams and taxis, but also includes shared bikes, cars and scooters. In our case study, we defined an ontology for MaaS, as follows.

First, domain experts created a ‘seed’ domain ontology through a structured knowledge elicitation process called MARVEL [8]. MARVEL, short for the Method to Analyse Relations between Variables using Enriched Loops, helped to structure and map relevant concepts and possible (inter-)relations between those concepts in a knowledge graph. This knowledge graph could then be transformed into web ontology language (OWL) for AI-supported foresight. Second, state-of-the-art AI has been used to gather data with so-called crawlers and scrapers: tools for obtaining data from websites and for cleaning up data. Data was gathered from a number of different sources, including ArXiv, Reuters, online journals and offline documents. The cleaned-up data was assigned to ontology nodes based on textual similarities with the data describing the ontology nodes, and stored in an index for retrieval purposes. The obtained insights are used to further refine the ontology. Based on human-machine interactions, missing knowledge is added or identified errors are removed from the ontology. This way, path dependencies, domain language or the over-reliance on past situations or experiences can be circumvented. Such an approach, whereby the expert knowledge base is combined with the machine’s database (i.e. a hybrid approach between experts and machines) can significantly advance the quality of the outcomes [1]. Still, concept drift – or the degrading quality of ontologies over time – remains an issue. To address this issue, we propose the use of spatio-temporal entropy that takes into account both the spatial structure of an ontology and its evolution through time in terms of data coverage. The practical merits of this approach need to be assessed in an experimental, user-centric study we aim for as follow up of this work [9]. The outcomes can be used in the policy making process to enable policy makers to make decisions regarding early innovative trends.

2 Discussion of Results

When comparing the findings from the two pilot studies, we see opportunities to expand the model-based approach.

First, it was observed in both pilot studies that the text-based model developed suffered from what is known as ‘concept-drift’; e.g. degrading quality of classification over-time (see [11] for more on this). This is not an unexpected finding as one can expect that the words associated with innovation may change over time. However, this was solved in pilot 1 by retraining the model on huge numbers of classified webpages, but still required regular checking of the model’s findings [10]. In Pilot 2, a statistical, spatio-temporal approach for analysing probability distributions of external data (specifically, topics) over internal knowledge representations of domains has been produced. Topic drift, under such an approach, can be seen as a dynamic pattern within topics: the probabilities of words associated with a topic changes over time, and the difference between two word distributions at different time steps can be computed with measures like the Kullback-Leibler divergence [12], a method that underlies the spatio-temporal metrics developed in pilot 1 [9].

Next, the results from both pilots also indicate that more attention should go to international comparison. In pilot 1 an international comparison revealed that a comparable web-based innovation detection approach also worked for companies in Germany [13]. A study performed in Sweden, based on the pilot 1 approach developed for the Netherlands, incidentally revealed that this approach did not perform well in this country. The latter could be due to a behavioural difference of companies in this country or may be the result of the need for a more (language specific) advanced syntactical and NLP-based analysis. All in all, these findings indicate a need for a more dynamic approach which enables the inclusion of new topics in the original innovation detection model developed, a mechanism that can build upon our findings in pilot 1, where low entropy scores of ontology topics covered with external data indicate necessary ontology repair, following pilot 2.

Next, our findings also indicate the sensitivity of our model. That is, pilot 1 excluded the findings for the (semi-)self-employed in the study [10]. This group of companies was found to be challenging to classify as it was observed that a very large percentage of them, compared to other small businesses, were classified as innovative. This suggested that self-employed people are more inclined to use words associated with innovation on their website and, hence, may be deliberately promoting themselves in a more innovative way.

Finally, our findings indicate the relevance of AI-supported innovation monitoring as an effective early warning system for policy makers. That is, an AI-supported innovation monitor helps to spot emerging topics and trends at a higher level of automation than before. Its hybrid and adaptive nature further allows the updating of information available to policy makers. This suggests promising avenues for the embeddedness of an AI-supported innovation monitor in the policy making process.

3 Conclusions and Outlook

In this position paper, we described two pilots addressing the analysis of innovation trends in society in order to address the overarching question how an AI-driven innovation monitor could be embedded in the policy-making process while addressing the complicating factor that the language used in innovative fields changes quite rapidly. Our first pilot applies various NLP techniques to analysing company information to detect innovative ones, and identifies the need for addressing concept drift. Our second pilot develops a 'human-in-the-loop" approach to innovation mining, where human experts seed a domain ontology, which subsequently becomes enriched with externally harvested and automatically analysed data. The two pilots, although addressing different use cases and largely non-overlapping techniques, appear to pave the way for a combined set of techniques for future approaches to innovation analysis: advanced text mining (NLP) techniques combined with statistical methods for mapping out concept drift across the topics relevant to an innovation domain. As such, the results show the promise of an AI-supported, modular approach to innovation monitoring for policy makers. How to deal with the notion of ‘concept drift’ is an important issue when discussing innovation. We have observed that the indicator words we use to identify websites, belonging to innovative companies, change over time. This is a recurring issue with text-based methods. Policy makers are interested in following trends over time and they want to be sure that the quantitative evidence, used for proposing policy measures, is reliable and accurate. Our two pilots have produced metrics for analysing the coverage of the ontology and data, and we believe these methods can be used for addressing concept drift.

A question for future research is how to transfer methods developed in one language domain to another. This question is particularly relevant in the European multilingual context and a recurring theme for Policy makers in the European Commission. How to deal with such linguistic issues, especially if we want to zoom in to specific topics like 5G-related innovations or the hydrogen-based economy in international, national and regional settings? Questions related to this are: What are internationally emerging trends? What are the primary topics in our region? Where are the geographic hotspots in our region?

A second question is if and how methods developed from an innovation monitoring perspective can be reused in other areas of interest or for other types of trends. For instance, is it possible to identify companies that try to move to a more sustainable (e.g circular economy) based production method? Can we pinpoint the emergence of new types of business activities (e.g. servitisation) or production methods (e.g. AI-based)?

An important third question in this process is; what can be automated and what has to be left to human insight? And how can we make findings insightful for policy makers? Aiming for a hybrid intelligence approach – as illustrated in this study – sounds like a sensible way forward. By answering such questions, we will be able to create an AI-driven, evidence-based approach to help policymakers in their policy-making for stimulating innovative trends.

3.1 Outlook

In our future research endeavours, we therefore aim to work with policy makers as well as stakeholders from the user side in order to better understand how to make the scientific results applicable and actionable in diverse settings. We have engaged national, regional and local governments, as well as investment companies, to find out what is the best way to present information. We listen to these stakeholders to understand their concrete questions, to get feedback on our results and to receive guidance on the way forward. We further wish to identify when and how policy makers wish to introduce AI-supported innovation monitoring into the policy making process, and how that introduction affects the policy making process in practice. We invite potential partners to collaborate with us to address these and other issues, e.g. in the context of the H2020 project TAILOR.

References

1. Geurts et al. (2020). Data Supported Foresight. Creating a new foundation for future anticipation by leveraging the power of AI and Big Data to go beyond current practice. TNO/Frauenhofer ISI whitepaper.
2. Geurts, A. (2018). A critical review of Alex Ross's The Industries of the Future. Tech-nological Forecasting and Social Change 128 (1), 311-313.
3. J. M. Utterback (1994). Mastering the Dynamics of Innovation: How Companies Can Seize Opportunities in the Face of Technological Change. Cambridge, MA, USA: Harvard Univ. Press, 1994
4. Cozzens, S., Gatchair, S., Kang, J., Kim, K., Lee, H.J., Ordonez, G., Porter, A., (2010). Emerging technologies: quantitative identification and measure-ment. Technology Analysis and Strategic Management 22, 361–376.
5. Mühlroth, C. & Grottke, M. (2020). Artificial Intelligence in Innovation: How to Spot Emerging Trends and Technologies. IEEE Transactions on Engineering Management.
6. Himanen, L., Geurts, A., Foster, A. & Rinke, P. (2019). Data-driven materials science: Status, challenges, perspectives. Advanced Science, 6 (21).
7. Smith, B. (2003). Ontology. In L. Floridi, editor, Blackwell Guide to the Philosophy of Computing and Information, 155–166. John Wiley & Sons, Incorporated.
8. Zijderveld, E. J. A. (2007). MARVEL - principles of a method for semi-qualitative system behaviour and policy analysis. TNO paper.
9. Raaijmakers et al. (2020). AI-supported Foresight and bias: Towards a hybrid approach. TNO working paper.
10. Daas, P.J.H., van der Doef, S. (2020) Detecting Innovative Companies via their Website. Paper accepted for publication in the Statistical Journal of the IAOS.
11. Daas, P.J.H., Jansen, J. (2020) Model degradation in web derived text-based models. International Conference on Advanced Research Methods and Analytics (CARMA), Valencia, Spain. Accepted for publication.
12. Kullback, S. (1987). "Letter to the Editor: The Kullback–Leibler distance". The American Statistician. 41 (4): 340–341
13. Kinne, J., Lenz, D. (2019) Predicting Innovative Firms using Web Mining and Deep Learning. ZEW Discussion paper no 19-001, Mannheim, Germany. doi:10.13140/RG.2.2.22526.84809.