JOURNAL ARTICLE

Topic modeling applied on innovation studies of Flemish companies

Annelien Crijns[a,*], Victor Vanhullebusch[a,*], Manon Reusens[a,**], Michael Reusens[b], Bart Baesens [a,c]

[a] *Department of Economics and Business, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium.*

[b] *Statistics Flanders, Havenlaan 88 bus 100, 1000 Brussels, Belgium.*

[c] *Department of Decision Analytics and Risk, University of Southampton, 12 University Road, Highfield, Southampton SO17 1BJ, United Kingdom*

[*] First author

[**] Corresponding author: manon.reusens@kuleuven.be

ARTICLE HISTORY
Compiled March 31, 2023

ABSTRACT

Mapping innovation in companies for the purpose of official statistics is usually done through business surveys. However, this traditional approach faces several drawbacks like a lack of responses, response bias, low frequency, and high costs. Therefore, possible solutions like text-based models have been developed to complement or substitute traditional business surveys. Web scraped company websites are used as input texts for these models. Previous research often focuses on the classification of companies into innovative or non-innovative through these models. This paper makes use of web scraping and text-based models in order to map the business innovation in Flanders. What differentiates this research from previously published work, is the special focus on the different types of innovation, discovered through topic modeling. More specifically, the scraped web texts are used to identify innovative economic sectors or topics, and to classify firms into these topics. The Flemish firms considered in this research are those that participated in the CIS 2019. It was found that the Top2Vec model can discover topics related to innovation within a large unstructured text corpus. Consequently, the Lbl2Vec model can use the Top2Vec output as an input to classify firms into the discovered topics. Therefore, this paper shows the potential of combining Top2Vec and Lbl2Vec model for discovering topics (or sectors) and classifying companies into these topics which results in an additional parameter for mapping innovation in different regions.

KEYWORDS
Text Analysis, Web Scraping, Topic Modeling, Text Classification, Innovation, Top2Vec, Lbl2Vec

1. Introduction

Over the past few years, many methodologies have been introduced to extract information about companies' innovativeness, and subsequently classify them as innovative or non-innovative. A well-known, traditional approach relies on sending out a business survey, the Community Innovation Survey (CIS), analyzing the responses, and classifying the firms. This method only focusses on businesses with 10 or more employees, and is carried out every two years by the EU member states since 1992 (Eurostat, n.d.). Currently, the CIS is the main tool for mapping innovation and for identifying Flemish companies as innovative or non-innovative for the purpose of official statistics.

However, this traditional business survey approach faces some challenges. A common business survey problem is that not all companies are motivated to respond, while responding companies may suffer from a self-reporting bias. Moreover, the entire process starting from sending out the surveys until having the data processed can take a significant amount of time. A lengthy process like this often entails a large administrative cost. Consequently, the survey cannot be repeated frequently, resulting in rather low-frequency data (Bavdaž et al., 2020).

Therefore, a text-based model has been developed to expand upon the results obtained from the CIS while including companies with less than 10 employees. This new approach gathers data by web scraping business websites and has been carried out and described by different researchers in different countries and regions, including the Netherlands and Flanders. However, a drawback of this approach is that it is a black-box model, while transparency is important in official statistics (Daas & van der Doef, 2020; Ipek, 2020).

This paper builds further on this new approach, but focuses on the use of topic modeling to enhance the mapping of innovation in Flemish companies to complement the CIS. Topic modeling analyzes the text and searches for hidden patterns that are related to the characteristics of innovative firms. These patterns are meaningful features that can be used to detect innovation. Moreover, it provides an efficient approach for analyzing a large corpus of text. Scraped homepages of Flemish company websites are used as input data. On this large text corpus, topic modeling is applied using the Top2Vec algorithm, which discovers topics related to innovative companies in the dataset. Afterwards, these topics are used as input for the classification model, Lbl2Vec, which classifies the firms into the discovered topics. This method thus provides an additional parameter for mapping innovation in different regions.The Top2Vec and Lbl2Vec models are both recently released, and rely on a multidimensional vector space, where distance reflects semantic similarity, and cosine similarity is used as a measure to quantify this similarity (Angelov, 2020; Schopf, Braun, & Matthes, 2021). The methodological contribution of this research entails the innovative combination of both models, where the Top2Vec output topics and keywords serve as input for the Lbl2Vec model for classification. The structure of this paper is as follows: firstly, the research questions are stated, afterwards relevant previous research is described. Consequently the research methodology is explained, and finally, the results, and conclusions are specified.

2. Research Questions

The main goal of this paper is to describe a method for mapping companies' innovativeness, based on web scraping business websites followed by a text-based model. With mapping innovation meaning three things: describing the main topics innovative companies revolve around, classifying companies into these discovered topics, and discovering the words most closely related to company innovation.
Therefore, the main research question is:
"To what extent can topic modeling enhance detection of innovation in Flemish com-

panies?".
The side research questions are:
"What are the main topics or sectors represented in Flemish, innovative companies?"
"How can companies be classified into the discovered topics or sectors, based on their scraped business websites?"
"What words on company websites are most related to innovation?"

3. Related Research

Several papers already described applications of gathering and using data from the web in an automated manner. Cothey (2004) discussed whether the use of web crawled information is reliable enough to make strategic decisions. It is concluded that web crawling can be reliable given complete reporting of the crawl policy. Afterwards, Katz and Cothey (2006) focused on the development of a robust web crawler to identify the presence of higher education institutions in Europe and Canada. Furthermore, Yang, Wilson, and Wang (2010) developed a climatic data scraper to keep track of real-time data in a database. This database gave opportunity to a wide range of possible innovative applications such as the prediction of rice growth stages and the best conservation of water in rivers.

These examples show how web scraping and crawling showed their relevance as a source of information already more than a decade ago. Some more recent papers dive deeper into the topic at hand, being the mapping of company innovation through web scraping and text mining. Gök, Waterworth, and Shapira (2015) recognized the shortcomings of the traditional data sources of company information. Their research is based on web scraping content of 296 UK-based firms for innovation studies. They conclude that website data offers valuable complementary insights next to traditional sources. Moreover, Mironczuk and Protasiewicz (2016) classified innovative companies by analyzing their web pages using a Naïve Bayes (NB) Classifier. Furthermore, Daas and van der Doef (2020) described the development of a text-based technique to detect innovation in Dutch companies. They achieved to reproduce the same results as the CIS while including companies with less than 10 employees. However, a major challenge was the model stability, as the accuracy of the model dropped from 93% to 63% after one year. Therefore, Daas and Jansen (2020) dove deeper into this model stability issue. They explained how disappearing websites on one hand, and updated websites on the other hand contribute to this model degradation. Moreover, Kinne and Lenz (2021) described the use of web mining and deep learning to distinguish product innovative firms from non-innovative firms in Germany. Their predictions were reliable, and can therefore become a cost-saving addition to the traditional methods for mapping innovation.

However, the current literature still faces some shortcomings regarding the mapping of innovation through web scraping and text mining. Most studies solely focus on the binary classification of companies into innovative or non-innovative. Another interesting angle to explore is the classification of firms into different innovative sectors or topics. Therefore, this paper delves into this alternative angle through the use of topic modeling. Additionally, this is one of the first innovation studies based on, Top2Vec and Lbl2Vec, which are relatively recently published models.

4. Research Methodology

4.1. Data

The data used to conduct this research is the CIS 2019[1] dataset provided by ECOOM KU Leuven in which a sample was drawn from the population of Flemish businesses, mainly based on size, sector as well as the presence of ongoing research, development activities and receipt of support from the government for R&D and innovation. This dataset includes information about the 3,179 Flemish firms that participated in the CIS 2019. The companies' data points given, include the company name, BTW number, address, email, website, and the inno5 variable. This inno5 binary variable indicates whether the company is classified as innovative or non-innovative according to the CIS 2019 results.

To provide a view on the sectors which are most represented among all Flemish companies in general, a sector division is given by Statistics Flanders. This division is twofold: on one hand it is based on the employment per sector, and on the other hand it is based on the export markets. The first division shows that Healthcare and social services, followed by Industry are the two largest sectors, employing respectively 16.0% and 13.2% of the total working population aged between 20 and 64 in the Flemish region in 2021. The following sectors in order of size include: Wholesale, retail and vehicle reparation (12.8%); Education (9.8%); Public administration, military and social security (7.1%) (Statistics Flanders, 2023). The second division identifies Chemistry and pharmaceuticals as the biggest sector, representing 27.6% of the Flemish export in terms of Euro value in 2021. The following sectors in the top 5 are: Mineral products (9.9%); Machines, mechanics, electronic equipment and components (9.9%); Transportation materials (9.9%); Synthetic Materials (8.4%) (investment & trade, 2022).

However, there are several attention points to consider when using this dataset. Firstly, there is a class imbalance as 75.3% of the companies is classified as innovative. Secondly, some extra preprocessing had to be done like deleting firms without inno5 variable, and deleting firms without company website URL. This resulted in an increased class imbalance of 80.4%, since many non-innovative firms had no given URL. Therefore, the missing URLs of the non-innovative companies were searched for manually on the internet using other available data points, which resulted in a total number of usable businesses of 2,535, and a decreased class imbalance of 74.6% innovative firms. The next issue to keep in mind is that the dataset contains just a small sample of all Flemish companies. On December 31, 2020, already more than 645,000 SMEs subject to value-added tax (VAT) were located in Flanders (Economie Belgische overheid, 2022).

Moreover, since innovation is a broad concept, distinguishing innovative from non-innovative companies highly depends on the interpretation of innovation (Kahn, 2018). Therefore, Eurostat has communicated the definition of innovation in context of the CIS. In this definition, a distinction is made between product or service innovation, and process innovation. These concepts are defined as a new or significantly improved good, service, or process. The innovation must be new for the firm, not necessarily new in the market or new in all companies' processes. Not included in

[1] The CIS 2021 results were not processed yet.

the concept of innovation are only selling things which are produced and invented by other firms and purely aesthetic changes (Eurostat, 2012). The CIS questionnaire focusses on five types of innovation: product innovation, business process innovation, abandoned innovation activities, on-going innovation activities, and in-house or external R&D activities (Ipek, 2020).

4.2. Web Scraping and Data Storage

Before scraping the URLs from the CIS 2019 dataset(Eurostat, n.d.), they were transformed to http prefix URLs, and their validity was checked through the status code. Websites giving a redirection or client error status code were removed from the dataset.

Afterwards, the remaining websites were scraped. Since Daas and van der Doef (2020) showed that scraped sub-domains do not enhance predictive models, only the main domains or homepages were scraped. Different web scraping methods were tested, using Requests, Beautiful Soup 4 (BS4), and Selenium. Three databases were created to test the different scraping approaches, the first with all URLs scraped using Requests and BS4, the second using Selenium and BS4, and the third using a combination of both. The difference in these scraping methods is that Selenium decodes JavaScript included in the page source, while Requests does not (vanden Broucke & Baesens, 2018). As a result, texts scraped with Selenium are often longer, while the scraping process takes more time. However, in previous research conducted by Daas and van der Doef (2020), the different scraping models were tested, and results indicated that extra text included in JavaScript did not display new perspectives regarding innovation (Daas & van der Doef, 2020). Therefore, further analysis in this research is based on the database where websites' homepages were scraped using only Requests and BS4. An advantage of this approach is that the web scraping process goes relatively fast, resulting in a more scalable model for large datasets. The scraped web texts were stored into ArangoDB (community edition 3.8.4) a free open-source database system, accessible and modifiable through Python, and able to store big, unstructured data.

4.3. Preprocessing and Language Detection

To prepare the web scraped texts for the topic modeling and classification, preprocessing and language detection were performed.

Firstly, web texts without relevant content were removed. As checking all texts manually would not be scalable, a sample of 30% of the companies was taken to identify the main characteristics of irrelevant web texts. Consequently, two types of irrelevant texts were identified. The first type entailed too short text, not including relevant words. As the sample showed that a relevant text required a bare minimum of 20 words, consequently all web texts with less than 20 words were removed. The second type entailed web texts describing that the domain name was for sale. Therefore, only the web texts including words like “domain” and “sale” were manually checked and the ones actually describing a domain name for sale were removed. Afterwards, duplicates were removed from the dataset.

Subsequently, the scraped texts' languages were detected with the langdetect library (Danilak, 2021) because the applied text models must be adapted to the text's language. A major part of more than 95% of all the scraped texts were written in Dutch or English. Therefore, only this part of the web texts was considered, and texts in other languages were removed from the dataset. This resulted in a Dutch and an English corpus with class imbalances of respectively 71.6% and 82.3% innovative companies. This difference in class imbalance already shows that in general, Flemish companies with their websites written in English are more likely to be innovative.

Furthermore, standard stopwords were removed using the NLTK library (Bird, Klein, & Loper, 2009). Additionally, some dataset-specific stopwords were manually listed and removed. This additional stopwords list included words concerning cookie policies, languages and opening hours, in both Dutch and English. Next, all words were converted to lower case, and punctuation marks were removed.

After preprocessing, the average word count per document for innovative and non-innovative firms has been respectively 343 and 314 words per landing page.

No stemming and/or lemmatization is conducted since the text-models in this research (Top2Vec and Lbl2vec) do not require this type of preprocessing (further explanation in section 4.4 and 4.5). As a result of these limited preprocessing requirements, time and costs are saved compared to other text models, resulting in higher scalability.

After all data cleaning and preprocessing 2,096 companies, or 65.9% of all firms in the total CIS 2019 datasets, were left for further analysis. A more detailed overview of all steps can be found in Table 1. The remaining data can be split up into four groups based on language and inno5 variable: Dutch Innovative, Dutch Non-innovative, English Innovative, and English Non-Innovative. The number of documents in each of these four datasets are 927, 367, 660, and 142 respectively.

Description	Number of firms	% innovative firms	% of total CIS dataset
Total CIS 2019 dataset	3,179		100%
- Firms without inno5 variable	37		
=	3,142	75.3%	99.0%
- Firms without given URL	790	60.4%	
=	2,352	80.4%	74.0%
+ Manually found URLs of non-innovative firms	183	0.0%	
=	2,535	74.6%	79.7%
- Firms with invalid URL or error, and deleted firms in preprocessing	439		
=	**2,096**	**75.7%**	**65.9%**

Table 1. From total CIS dataset to total number of firms for analysis

4.4. Topic Modeling

4.4.1. General

Topic modeling is used for finding topics, or more specifically latent semantic structure in a corpus of documents. The main methods used for this purpose are Latent Dirichlet Allocation (LDA) and Probabilistic Semantic Analysis. However, a more sophisticated algorithm called Top2Vec has been developed recently to create topic models using vector embeddings in a semantic space followed by clustering. The increased complexity of Top2Vec is compensated by several major advantages. The algorithm is able to capture the meaning of words and phrases. Meanwhile, this is impossible for the previous models as they make use of bag-of-words, and therefore ignore the ordering and semantics of words. Also, several preprocessing steps like stop-word removal, stemming, and lemmatization are not required before the use of Top2Vec. Another unique charachteristic is that the model can determine the number of topics in the text corpus, while this number must be given as an input to other models like LDA. More specifically, Top2Vec assumes that the number of dense areas in the semantic space are equal to the number of topics in the corpus. This is a major advantage because giving a number of topics which is too small can lead to bad coverage of the corpus, while selecting a number of topics which is too big can lead to redundancy (Angelov, 2020; Lu & Chesbrough, 2021).

4.4.2. Top2Vec

As described by Angelov (2020), the Top2Vec algorithm generally performs several steps. In the first stage, semantic word and document embeddings are created in a semantic space where semantic similarity is represented by the distance between vectors. Doc2Vec is used to learn jointly embedded word and document vectors. In this semantics space, the number of topics is determined through clustering. More specifically, dense areas of document vectors in the semantic space represent groups of documents about the same topic. Consequently, the number of dense document vector areas gives the number of topics. However, due to the "curse of dimensionality", document vectors are sparse and finding dense clusters is difficult. Therefore, dimensionality reduction, using Uniform Manifold Approximation and Projection (UMAP), is performed before clustering. Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is used as a clustering algorithm to find the dense document vector areas. The actual topic vectors are calculated in the original dimensional space by calculating the centroids of the dense areas. The centroid, or arithmetic mean of the document vectors, is calculated for each dense cluster. The word vectors close to this topic vector represent the words that can best describe the underlying topic. Therefore these words are called the topic words. However, training this model on a small dataset can cause overfitting and can cause the model to not be able to effectively capture rare words and/or infrequencies. Top2Vec generates many outliers, because every document needs to be assigned to a topic. However, the outliers detected by HDBSCAN are not taken into account when calculating the centroid(Angelov, 2020).

In context of this research, multiple Top2Vec iterations were executed because of the stochastic characteristic of the model. The number of iterations performed was determined upfront, and all results are saved and discussed in section 5.1. The reason behind Top2Vec's stochasticity is that the model is based on Doc2Vec as an embedding model. Since Doc2Vec is trained using neural networks, which is a stochastic process, randomness is added to the model. Consequently, Top2Vec is a non-deterministic model where new topics can occur when regenerating the model. Moreover, half of the iterations were trained on unigrams only and the other half on bigrams, since the two methods revealed different topics.

The model was used for two different tasks. The first task consisted of discovering topics related to innovation within the data groups. This was achieved by creating word clouds representing different topics around the keyword 'innovation'. Subsequently, the most prominent words of these word clouds were listed as topic words describing the topics at hand. Afterwards, these topics were linked to economic sectors or themes.

The second task was to identify the words most closely related to innovation within each data group. The cosine similarity (1) was used as a metric to evaluate which words are semantically most related to innovation. It is a widely used metric in text classification and many other fields, which measures the similarity between two or more vectors (Alake, 2020; Li & Han, 2013). In context of this research, the vectors under consideration are word, document, topic, or label vectors in the semantic space, created through the applied text models. Mathematically, it is the cosine of the angle between the vectors under consideration. The highest possible value is 1, indicating that the vectors point in the exact same direction. The lowest possible value is 0, meaning that the vectors are perpendicular to each other (Alake, 2020).

$$\cos(\mathbf{t}, \mathbf{e}) = \frac{\mathbf{te}}{\|\mathbf{t}\| \|\mathbf{e}\|} = \frac{\sum_{i=1}^{n} \mathbf{t}_i \mathbf{e}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{t}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{e}_i)^2}} \tag{1}$$

Altogether, 20 iterations of the Top2Vec model were performed of which 10 on the Dutch Innovative dataset, and 10 on the English Innovative dataset. The non-innovative datasets were not used because performing the model as described above was impossible on the Dutch Non-innovative dataset. In order to remain consistent, the English Non-innovative dataset was not used either. The Dutch Non-innovative dataset could not be used because the word 'innovatie' (innovation in Dutch) did not occur enough in this dataset. Consequently the word 'innovatie' was not learned by the model, and since both tasks of the Top2Vec model revolve around that word, it was impossible to run the model.

4.5. Classification Model

4.5.1. Lbl2Vec

Next to finding topics related to innovative firms, the focus of this research is on the classification of companies into these discovered topics, using a text-based model applied on the scraped company web pages. The classification of companies into economic sectors or themes can generally be achieved using supervised learning techniques. However, these models need a large, labeled training dataset, while labeled data is often not available or requires a lot of manual work. Lbl2Vec was used as a classification method in this research because on one hand, it is an unsupervised model which does not require any labeled data. On the other hand, because this model can make use of the Top2Vec model's output, resulting in the combination of two models complementing each other.

Recently published by Schopf et al. (2021), Lbl2Vec is an unsupervised approach able to retrieve certain documents, related to a given topic, out of a large set of documents. More specifically, a large corpus of unlabeled documents is given, and certain topics are known to occur in this corpus. Lbl2Vec only needs keywords related to a topic in order to retrieve the documents related to this topic. Only limited preprocessing is required, and there is no need to annotate any data, resulting in time and cost savings. The best results are achieved when the topic keywords have high intratopic similarity and low intertopic similarity (Schopf et al., 2021).

The approach described by Schopf et al. (2021) relies on jointly embedded word, document, and label embeddings in the same semantic space where distance represents semantic similarity, and the cosine similarity is used as a metric to quantify semantic similarity. In the first step, word and document vectors from the document corpus are simultaneously learned in the same space through iterative training on interleaved distributed bag of words version of paragraph vector (PV-DBOW) and Skip-gram architectures. In the second step, the topic keyword embeddings are used to find the label embeddings. More specifically, for each topic, the centroid of its keyword embeddings is calculated. Afterwards, the cosine similarity of the topic keyword centroid to each document in the corpus is calculated. The documents with the highest cosine similarity to a topic keyword centroid are selected. Next, the outliers from this selected group of documents are removed by local outlier factor (LOF) cleaning. Finally, the centroid of the remaining selected documents is calculated to represent the topic's label embedding. When the embeddings are created, Lbl2Vec can retrieve documents related to predefined topics with high precision (Schopf et al., 2021).

In this research, the Lbl2Vec model is used to classify the Dutch firms into the topics discovered through Top2Vec. The most prominent topic keywords belonging to these topics discovered through Top2Vec are used as Lbl2Vec input topic keywords. The model is tested on an unseen test set of 292 firms, 22.6% of all Dutch firms. The results are discussed in section 5.2.

5. Results

5.1. Top2Vec

5.1.1. Topics discovered in innovative firms' web texts

As mentioned before, ten Top2Vec iterations were performed on each of both innovative datasets. An overview of the resulting topics derived from the word clouds is given in Table 2 and 3. In these tables, the 'top relevant terms' column represents prominent words in the topics' word clouds. The 'frequency' column shows in how many of the ten iterations the topic was clearly represented in a word cloud. This percentage was constructed as follows, 10% was counted every time a topic was represented as one clear word cloud. Additionally, 5% was counted every time a topic was found as one of the topics described in a word cloud that represented two topics at the same time. In this case, a word cloud included two different topics which were related to each other, like engineering and wood- and furniture industry. Only 5% was counted instead of 10% because the topic was represented, but not completely distinct from the other ones.
Both the Dutch and English Innovative datasets generated five to nine clear word clouds in every conducted iteration. The topics found in the word clouds obviously described sectors in which the underlying firms operated. To summarize the topics, they were linked to sectors described by the Flemish government institution Maatschap-

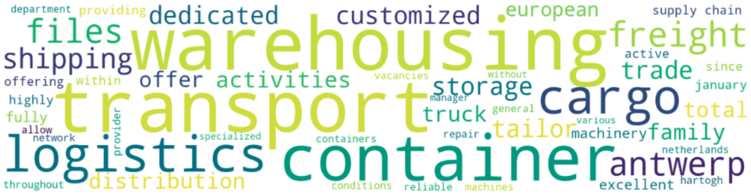

Figure 1. Example of an output word cloud representing the Transport & Logistics sector

pelijk Verantwoord Ondernemen (MVO) Vlaanderen (MVO Vlaanderen, n.d.). The discovered topics are explained below. From the nine discovered topics, five occurred in both the Dutch and English Innovative datasets. These topics will be described first.

The first topic clearly described the wood and furniture industry, and occurred in every Dutch iteration, but only in one English iteration. This means that in the Dutch semantic space, a clear, dense area consists of vectors representing words like interior, architects and materials. From this can be derived that a significant part of the innovative Flemish companies are active in this industry and often they have a Dutch website,while not always providing an English version.

The second topic was present in most Dutch and English iterations and obviously represented a broad, engineering related topic. Since MVO Flanders describes many distinct sectors in this area, while the topics in the word clouds are more general, this topic was linked to four sectors: Common industry, Energy and water, Chemistry and metal, and Construction. The high frequency of this topic indicates that a significant part of the innovative Flemish companies is active in one of these four sectors.

The next topic, is linked to two related sectors: IT services and companies, and Consultancy and other services for enterprises. As this topic also occurred in most iterations, both in Dutch and English, a significant part of the innovative Flemish companies is active in the IT and consultancy industry.

The fourth topic is labeled as the food industry, and can be linked to the Food and tobacco sector. This topic occurred in most iterations on the Dutch dataset, but only in a few iterations on the English dataset. This means that a significant part of the innovative Flemish companies is active in the food industry, and mostly their websites are written in Dutch and not in English.

The last topic present in both the Dutch and English dataset is linked to the Transport and logistics sector. Also this topic occurred in most iterations in both datasets and is therefore well represented among the innovative Flemish companies.

Subsequently, the topic only occurring in the Dutch word clouds is labeled as the banking and insurance industry and is a subset of the Banking, insurance, mail, and telecom sector. As it only occurs in less than half of the iterations, it is a relatively smaller sector among the innovative Flemish companies.

Finally, the three topics only occurring in the English word clouds will be described in order of high to low frequency. The first topic cannot be linked to any economic sector. However, its keywords typically occur on websites of firms listed on the stock exchange and therefore subject to shareholder reporting. This indicates that a significant part of the innovative Flemish companies is listed, and reporting to its shareholders about financial performance, ESG initiatives etc.

The next topic was linked to the textile and clothing sector. This industry is clearly represented among the innovative Flemish companies, but only in the English dataset.

The final topic was linked to the Health services and institutions sector. This topic only occurred in one iteration, and is therefore not represented by an important, dense area in the semantic space. However, it indicates that a part of the innovative Flemish companies is active in the health sector.

In conclusion, the sectors most represented among the innovative Flemish companies are Engineering, Wood- and furniture, IT and consultancy, Food, and Transport and logistics. Some smaller sectors among the innovative Flemish companies include Banking and insurance, Textile and clothing and Health services and institutions.

Topic number	Topic	Top relevant terms (translated in English)	Frequency
1	Wood- and furniture industry	Interior, placement, customization, architect(s), materials, kitchen, bathroom, design, craftsmanship, finishing, realizing, plan, home, showroom, window, doors, renovation, project, quote, comfort, home, inspire, workshop, collections, finish, result, beautiful	100%
2	Engineering	Engineering, CNC, machines, construction, machine park, machinery, machine building, techniques, part, steel, installation, technical, assembly, prefab, industrial plastic, milling, knowhow, button, heating, ventilation, warm air, home, energy, battery, comfort, warranty, installation, dealer, sustainable, promotions, water, engineers, building, experience, renovation, specialized	95%
3	IT and consultancy	Software, cloud, security, digital, cases, management, experts, data, ICT, knowledge, helpdesk, strategy, business, efficient, ERP, CRM, hardware, consulting, consultancy, case	90%
4	Food	Bread, cookies, delicious, recipes, fresh, bakery, taste, passion, promotions, shop, order, points of sale, made, belgian, package, natural, delicious, assortment, meat, chocolate, webshop, packaging, pigs, family business, story, healthy, orders, craftsmanship, organic, enjoy, retails, porc, flemish, quality, consumer	90%
5	Transport and logistics	Logistics, transports, goods, storage, sustainable, flexibility, efficiency, warehouse, containers, family business, container, central, our customers	75%
6	Banking and insurance	Insurance, savings, loans, investments, card, home, app, safe, professional, individuals damage, bank, corona, term	40%

Table 2. Dutch Innovative dataset: topics related to innovation

Topic number	Topic	Top relevant terms	Frequency
1	Transport and logistics	Logistics, warehousing, bulk, transport, supply chain, shipping, container, storage, deliver, distribution, fleet port, terminal, cargo, activities, dedicated, antwerp, truck, delivery, service, supply, freight, warehousing, navigation, goods	100%
2	Investor relations and ESG	Sustainability report, sustainability, investor relations, statement, reports, progress, update, annual, report, latest news, awards	75%
3	IT and consultancy	Communications, cloud, platform, integration, digital transformation, accelerate, integration, organizations, data, connected, model, digital, connectivity, users, growth, secure, manage, insights, client, strategy security, infrastructure, insurance, provider, Microsoft, consulting, teams, banking, expertise, increase, enable, fast, enhance, solution, apps	70%
4	Engineering	Machinery, machine, high quality, manufacturer, plastics, customized, material, innovative, years, experience, tailor made, raw materials, wide range, production, hydraulic, parts, extensive, certified, specialized, unique, dedicated, plastics, molding, engineered, seals, compounds, rubbers, coatings, glass, chemistry, chemical, packaging, constructions, polymers, profiles, aerospace, filtrations, custom, components	70%
5	Textile and clothing	Fashion, fabrics, textiles, textile, creative brands, passion, trends, plants, coatings, quality, sports, activities, variety, sports, activities, variety, message, production, innovative, highest, best, unique innovative	40%
6	Food	Fruit, taste, ingredients, fresh, meat, food, love, made, selected, label, grow, growing, passion, natural, farm, rice, premium, quality, IQF, production, processes	35%
7	Wood- and furniture industry	Tailor made, extensive, wide range, taste, house, premium, high quality, specialized, years, experience, designed, design, customized, state art, innovative, collection, inspiration, lighting, black, sale, catalogue, outdoor, luminaires, hand, pergola, wall, warranty, shop, furniture	10%
8	Health services and institutions	Patients, patient, clinical, risk, improve, deliver, regulatory, expertise, professionals	10%

Table 3. English Innovative dataset: topics related to innovation

5.1.2. Evaluation of Top2Vec cosine similarities on Dutch Innovative dataset

Cosine similarity is used as the main measure to quantify semantic similarity in this research. However, it is not clear which value of this measure must be met at minimum to indicate a significant similarity. No standard cut-off value is known because this heavily depends on the vector space. Therefore, a sensitivity analysis was conducted, where the performance of different cosine similarity cutoff values was analyzed. For this analysis, Top2Vec was used as some kind of "classifier".

As explained before, the Top2Vec algorithm relies on a semantic space filled with word vectors, document (web text) vectors, and topic (sector) vectors. In this multidimensional space, semantic similarity between vectors is quantified through the cosine similarity (Angelov, 2020). Consequently, the model offers the possibility to extract all documents which document vectors have the highest cosine similarity to a certain topic vector. In other words, it is possible to select a topic generated by the model, and get a list of the documents most related to this specific topic. As a result, Top2Vec can be used to link documents to the semantically most related topics, which were generated by the model before. In other words, web texts (documents) can be classified into sectors (topics).

For this analysis, the Dutch Innovative dataset and its five most frequent topics were considered. All 380 firms which document vector had a cosine similarity of 0.2 or higher to one of the five topic vectors, were manually labeled. In order to select the right label, their websites have been checked, followed by a manual classification into sectors using the sector definitions described by MVO Flanders. Nevertheless, some overlap between different industries is possible.

Topic number	Topic	CS $\geq$ 0.2	CS $\geq$ 0.3	CS $\geq$ 0.4	CS $\geq$ 0.5
1	Wood- and furniture industry	0.61	0.95	1	1
2	Industry (common)	0.91	0.95	1	1
3	IT services and consultancy	0.89	0.97	1	1
4	Food	0.67	0.96	1	1
5	Logistics	0.64	0.91	1	1

Table 4. Dutch Innovative dataset: Sensitivity analysis of Top2Vec cosine similarity values
Classification accuracy per cosine similarity (CS) and sector

As can be derived from Table 4, if the cosine similarity of the document and the topic vector is equal to or higher than 0.4, the classification accuracy reaches the maximum value of 1. Moreover, if the cosine similarity is equal to or higher than 0.3, the model performs well with an accuracy of 0.91 or higher on all tested topics. However, only 26% of all companies are classified with a cosine similarity of 0.3 or higher. A majority of 60% of all web texts are classified to a sector with a cosine similarity of 0.2 or higher. In this part, the performance highly depends on the sector since the accuracies range from 61% to 91%.

To be clear, this analysis was not performed on an unseen test set, but on documents that were used to create the semantic space. The Top2Vec model does not work with test sets, since it can only work with data which is trained, and placed into the semantic space. For additional clarification, Top2Vec is not used in this research as a proper classifier, but only in this section with the goal of putting cosine similarity values into perspective.

5.1.3. Words most closely related to innovation

Besides discovering topics, the Top2Vec model can present the words most closely related to another word. This relation is measured through the cosine similarity between word vectors in the created semantic space, and its maximum value is one. The words most closely related to innovation were analyzed through this Top2Vec application. More specifically, during each iteration, the 30 words with the highest cosine similarity to innovation (or 'innovatie' in the Dutch iterations) were listed. Next, the words with a cosine similarity of 0.30 or greater in one or more iterations were selected. Afterwards, the number of times that the cosine similarity was 0.30 or greater, out of the ten iterations was counted for every selected word in Dutch and English. The results are shown in Table 5 and 6 for both innovative datasets.

Selected words in Dutch	Translation in English	Number of iterations with cosine similarity to 'innovatie' >= 0.30 out of ten iterations
nieuws	news	7
life	life	5
nieuwste	newest	5
technologie	technology	5
ontwikkeling	development	4
blog	blog	3
oplossing	solution	3
aandacht	attention	2
dankzij	thanks to	2
gaat	goes	2
onderzoek	research	2
teams	teams	2
zet	put	2
biedt	offers	1
grootste	biggest	1
organisatie	organisation	1
retail	retail	1
samenwerking	cooperation	1
trends	trends	1
verschil	difference	1

Table 5. Dutch Innovative dataset: words most closely related to innovation according to Top2Vec model

Overall, the cosine similarities of the words most closely related to innovation were higher in the English dataset. This is also reflected in the number of selected words with a cosine similarity equal to or greater than 0.30 in one or more of the Top2Vec iterations. In the English Innovative set, 57 words were selected, while only 21 words made the Innovative Dutch Selection.

The following five words were selected both in the English and Dutch Innovative datasets: technology, life, research, news, and solutions.

The highest cosine similarity in the Dutch Innovative case (0.38) was reached by the word 'nieuwste' (newest), while the word 'nieuws' (news) reached the selection in the highest number of iterations. Two striking words were 'retail' and a company name. The word retail in the selection reflects an entire sector being closely related to innovation instead of just one word. On the other hand, the selected company name was an innovative company which was mentioned 391 times on its website's homepage, and therefore reached a cosine similarity of 0.32 in one of the iterations. This company name was deleted from the table because of confidentiality reasons.

Selected words in English	Number of iterations with cosine similarity to innovation >=0.30 (out of ten iterations)
sustainability	10
strategy	10
future	10
technology	9
sustainable	9
energy	8
environment	7
investors	7
reports	6
report	7
life	6
governance	6
business	5
approach	5
glance	5
stories	5
vision	5
investor	4
diversity	4
challenges	4
annual	4
impact	4
human	4
history	4
collaboration	3
global	3
research	3
locations	3
careers	3
inclusion	3
learn	3
news	3
climate	3
releases	2
transparency	2
expertise	2
engagement	2
shareholders	2
electronics	2
solutions	2
top	2
progress	1
suppliers	1
investor relations	1
corporate	1
corporate governance	1
esg	1
supplier	1
people	1
ethics	1
financial	1
create	1
businesses	1
science	1
presentations	1
natural	1
newsroom	1

Table 6. English Innovative dataset: words most closely related to innovation according to Top2Vec model

The highest cosine similarity in the English Innovative case (0.56) was reached by the word ‘sustainability’, while this was also one of the three words reaching the selection in every iteration. Additionally, many other words related to sustainability were selected, including sustainable, diversity, ESG, and ethics. Another recurring theme within the selected words is investor relations, with the words annual, reports, vision, shareholders, transparency, investor relations, and financial. This topic also occurred in the word clouds generated by the Top2Vec model and indicates that many innovative firms are listed on a stock exchange, and therefore obliged to keep their shareholders up to date through annual reports etc. A third recurring theme seems to revolve around research and new technologies, including the words technology, research, learn, electronics, releases, expertise, progress, create, science.

5.2. Lbl2Vec

5.2.1. Sector classification tested on Dutch companies

As explained in section 4.5.1, the Top2Vec output topics and related prominent topic words are used as input for the Lbl2Vec model. The six classes discovered through Top2Vec in the Dutch Innovative dataset and its key words are stated in Table 2. To evaluate this classification model, a labeled test set is required. Because the sector labels were not provided in the CIS 2019 dataset, a random set of 350 Dutch companies was manually classified into one of the six sectors or ”other sector”. This last class is required because not all economic sectors are covered with the six output sectors from the Top2Vec model on the Dutch Innovative dataset. For running the Lbl2Vec model, input keywords are needed for every class. However, for the ”other sector” class, no keywords are available from the Top2Vec output. As a result, the Lbl2Vec model could only classify the firms in one of the six sectors mentioned above. Therefore, the companies classified as ”other sector” were removed because they could never be classified correctly in one of the other six sectors. Consequently, 292 companies (22.6%) remained as unseen test set.

The result of the test is shown in Table 7. With an overall classification accuracy of 0.75, the model performs well. The highest F1-score (0.88) is reached for the Food sector, probably because its keywords are very sector-specific resulting in a high intratopic similarity, and low intertopic similarity. The lowest F1-score (0.60) is obtained for the Banking and insurance sector. This low F1-score is caused by its low precision score. This means that a relatively large amount of companies were wrongly classified into this topic. The keywords within the topic show a low intratopic similarity, ranging from ”corona” to ”insurance” and ”professionals”, making this classification task more challenging, resulting in lower performance.

Topic	Precision	Recall	F1
Banking and Insurance	0.44	0.94	0.60
Engineering	0.92	0.60	0.73
Food	0.82	0.94	0.88
IT and Consultancy	0.76	0.89	0.82
Transport and Logistics	0.71	0.77	0.74
Wood- and Furniture Industry	0.62	0.70	0.66
Accuracy			0.75

Table 7. Dutch dataset: Lbl2Vec classification model tested on unseen test set of 292 companies

6. Conclusion

This research focused on the use of topic modeling for mapping innovation in Flemish companies, complementing the CIS. Flemish company websites were scraped, and Top2Vec and Lbl2Vec models were applied. Both models rely on a multidimensional vector space, where distance reflects semantic similarity, and cosine similarity is used as a measure to quantify this similarity.

In conclusion, the Top2Vec and Lbl2Vec models can be valuable for mapping company innovation and can be used in a complementary way. Top2Vec is an ideal model for the detection of topics or trends in a large unstructured text corpus. The model successfully identified several economic sectors in which Flemish innovative firms are active. Subsequently, Lbl2Vec can classify the companies into the detected topics (or sectors), using the output topics and related key words of the Top2Vec model.

Further research in this area would be interesting to further explore the possibilities of topic detection and classification in innovation studies. For example, conducting this research on a larger scale including more companies would probably result in more topics and related keywords in Top2Vec. Consequently, more exhaustive inputs would be available for Lbl2Vec, resulting in a more complete classification including classes representing all economic sectors.This would also lead to a higher intratopic similarity and lower intertopic similarity, further improving the results of this classifier.

These results could be used for further research on developing more complete statistics. One example of an interesting possibility would be a company database with the possibility to filter firms based on economic sector, geographic location etc., all based on their scraped web texts. A database like this could help companies in finding partners for realizing new projects, or for analyzing the competitive landscape.

7. Additional Information

The Python code that was written for this research can be accessed through the following GitHub repository: `https://github.com/VictorVanhulle/Topic_modeling_applied_on_innovation_studies_of_Flemish_companies`

References

Alake, R. (2020). *Understanding Cosine Similarity And Its Application.* Retrieved 2022-04-15, from https://towardsdatascience.com/understanding-cosine-similarity-and-its-application-fd42f585296a

Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. Retrieved from http://arxiv.org/abs/2008.09470

Bavdaž, M., Snijkers, G., Sakshaug, J. W., Brand, T., Haraldsen, G., Kurban, B., . . . Willimack, D. K. (2020). Business data collection methodology: Current state and future outlook. *Statistical Journal of the IAOS*, *36*(3), 741–756.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Cothey, V. (2004). Web-Crawling Reliability. *Journal of the American Society for Information Science and Technology*, *55*(14), 1228–1238. Retrieved from http://www.scit.wlv.ac.uk/{~}in7803/publications/cothey{_}2004.pdf

Daas, P. J., & Jansen, J. (2020). Model degradation in web derived text-based models. In (pp. 1–8). Retrieved from http://ocs.editorial.upv.es/index.php/CARMA/CARMA2020/paper/viewFile/11560/5699

Daas, P. J., & van der Doef, S. (2020). Detecting Innovative Companies via their Website. *Statistical Journal of the IAOS*, *36*(4), 1239–1251.

Danilak, M. (2021). *langdetect.* https://github.com/Mimino666/langdetect. GitHub.

Economie Belgische overheid. (2022). *Statistieken over kmo's in België.* Retrieved 2022-04-07, from https://economie.fgov.be/nl/themas/ondernemingen/kmos-en-zelfstandigen-cijfers/statistieken-over-kmos-belgie{#}:{~}:text=Enkelecijfers{&}text=Op31december2020{%}2Czijn,isdelocatieervanonbekend.

Eurostat. (n.d.). *Website Eurostat - Community Innovation Survey (CIS).* Retrieved 2021-10-19, from https://ec.europa.eu/eurostat/web/microdata/community-innovation-survey

Eurostat. (2012). *Glossary: Innovation.* Retrieved 2021-10-19, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Innovation

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*(1), 653–671. Retrieved from https://doi.org/10.1007/s11192-014-1434-0

investment, F., & trade. (2022). *Flanders Trade.* Retrieved from https://www.flandersinvestmentandtrade.com/export/node/14568

Ipek, N. (2020). *Detecting Flemish Innovative Companies Using Web Scraping.*

Kahn, K. B. (2018). Understanding innovation. *Business Horizons*, *61*(3), 453–460. Retrieved from https://doi.org/10.1016/j.bushor.2018.01.011

Katz, J. S., & Cothey, V. (2006). Web indicators for complex innovation systems. *Research Evaluation*, *15*(2).

Kinne, J., & Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. *PLoS One*, *16*(4 April), 1–18.

Li, B., & Han, L. (2013). Distance weighted cosine similarity measure for text classification. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 8206 LNCS, pp. 611–618).

Lu, Q., & Chesbrough, H. (2021). Measuring open innovation practices through topic modelling : Revisiting their impact on firm financial performance. *Technovation.* Retrieved from https://doi.org/10.1016/j.technovation.2021.102434

Mironczuk, M., & Protasiewicz, J. (2016). A Diversified Classification Committee for Recognition of Innovative Internet Domains. In *Beyond databases, architectures and structures. advanced technologies for data mining and knowledge discovery* (pp. 368–383). Springer.

MVO Vlaanderen. (n.d.). *Sectoren MVO Vlaanderen.* Retrieved from https://www.mvovlaanderen.be/sectoren

Schopf, T., Braun, D., & Matthes, F. (2021). Lbl2vec: An embedding-based approach for unsupervised document retrieval on predefined topics. In *Proceedings of the 17th international conference on web information systems and technologies - webist,* (p. 124-132). SciTePress.

Statistics Flanders. (2023, may). *Tewerkstelling per sector.* Retrieved from `https://www.vlaanderen.be/statistiek-vlaanderen/arbeid/tewerkstelling-per-sector`

vanden Broucke, S., & Baesens, B. (2018). *Practical Web Scraping for Data Science: Best Practices and Examples with Python.*

Yang, Y., Wilson, L. T., & Wang, J. (2010). Development of an automated climatic data scraping, filtering and display system. *Computers and Electronics in Agriculture*, *71*(1), 77–87. Retrieved from `http://dx.doi.org/10.1016/j.compag.2009.12.006`