Capturing innovation in the Netherlands using website texts: a concept drift-resilient approach.

Peereboom, Sanne

Universiteit
Leiden
The Netherlands

Capturing innovation in the Netherlands using website texts: a concept drift-resilient approach.

Sanne Peereboom

Thesis advisor: Prof. dr. P. J. H. Daas, Statistics Netherlands
Thesis advisor: Prof. dr. M. J. De Rooij, Leiden University

Defended on August 28th, 2023

MASTER THESIS
STATISTICS AND DATA SCIENCE
UNIVERSITEIT LEIDEN

Contents

Abstract

Estimating the number of innovative companies in a country can be beneficial for policymakers and the production of official statistics. Currently, innovation activity is estimated by administering a survey to a stratified sample of companies. However, this method is costly, and small companies are not sampled. Previous research utilized company website texts with word embeddings to detect innovation activity, allowing for the inclusion of small companies. The model was initially highly accurate, but suffered from concept drift due to the words on websites changing over time. This paper proposes an alternative method of detecting innovation in website texts, using semantically meaningful sentence embeddings. We hypothesized that website texts stay semantically similar over time, although they may use different words, and that the use of sentence embeddings will provide a classification model with more stability over time. These hypotheses were confirmed, although the external validation of the model is inconclusive. Points of note and suggestions for further research are discussed.

1 Introduction

Innovation is an important concept indicative of a healthy economy in a continuously growing modern-day society. Its definition implies the introduction of novel products or business processes to adapt to the ever-evolving stream of supply and demand. Capturing innovation activity within a country can be a helpful tool to determine the state of the economy and to track emerging applications and technologies, which can in turn allow policymakers to focus their resources to support further growth and development of the economy.

The common practice to measure innovation in Europe is to send a biennial Community Innovation Survey (CIS) to a stratified sample of 10.000 companies with 10 or more employees. This survey measures to what extent these companies have contributed to innovation through their products or business processes as well as the amount of money spent on research, during the three years prior to the survey. A shortcoming of this method is that smaller, possibly innovative companies (with fewer than 10 employees) are not included. As a result, information on this very important group of small businesses is missed.

In contrast, big data sources such as public websites offer an interesting opportunity to include much more information on businesses than what

would be available through traditional survey methods. Websites have the additional advantage that they introduce the option of a census-like approach without requiring the time and resources necessary for a census-based survey. In other words, the use of big data sources has potential in the detection of innovative companies that otherwise would not be surveyed through the CIS. There is potential for these sources to be used in official statistics, provided that they can be reliably used and analyzed. It is for this reason that an alternative approach for detecting innovative companies was developed: establishing whether a company is innovative through their website text [1] [2].

1.1 Previous research

In their study, [1] scraped the public websites of companies that were included in the CIS, and used a data-driven modeling approach to predict whether a company is innovative or not by studying the text on the main page of their website. Texts in both the Dutch and English language were found. [1] included word embeddings as features in their analyses. Word embeddings are obtained by calculating co-occurrences and relative positions of different words, where words that are observed closely together are seen as more similar. Initially, the inclusion of these word embeddings allowed the model to detect innovative companies with an accuracy of 93%, whereas excluding word embeddings resulted in a classification accuracy of 60%. [1, Table 1, p. 5]. This implies that there is information to be found in the syntactic structure of website texts, that is, specific combinations and positions of words in the text. However, the trained model was found to classify new data with deteriorating accuracy over time, implying that the patterns learned by the model change over time [3]. This is known as concept drift.

1.1.1 Concept drift

Concept drift is described as "a phenomenon in which the statistical properties of a target domain change over time in an arbitrary way" [4, p. 2347]. Formalized, this implies that $P_t\left(X, y\right) \neq P_{t+1}$, where $P\left(X, y\right)$ is some joint probability distribution of feature vector X and labels y at timepoint t. Subscript t refers to the timepoint right before which the joint distribution of $P(X, y)$ changes. Subscript $t + 1$ then denotes the timepoint at which the joint distribution of $P(X, y)$ has changed, and concept drift has arisen.

The joint probability $P(X, y)$ is equal to $P(X)P(y|X)$: it is the product of the probability of features X and the probability of label y given features X. With these components, three possible probabilistic sources of drift can be defined [4]:

- $P_t(X) \neq P_{t+1}(X)$ while $P_t(y|X) = P_{t+1}(y|X)$, where the input features change while the probabilities of the target labels conditional on the features remain unchanged, known as virtual drift. One subcategory of this type of drift is feature evolution, where the set of input features dynamically changes over time [5].
- $P_t(y|X) \neq P_{t+1}(y|X)$ while $P_t(X) = P_{t+1}(X)$, where the probabilities of the target labels conditional on the feature vector change, while the distribution of the features itself remains the same. This type of drift is known as decision boundary drift. In short, the model employs the optimal decision boundary it learned from the original data, while the new data have a different optimal decision boundary, decreasing its accuracy.
- $P_t(X) \neq P_{t+1}(X)$ and $P_t(y|X) \neq P_{t+1}(y|X)$, a combination of former two sources of drift. In literature, this is the type of drift usually referred to as concept drift. Here, both the features and the probabilities of the target labels conditional on the features change over time, leading to a more dramatic decision boundary change. The model decays, because its learned decision boundary does not automatically adapt to the different feature probability distribution and decision boundary in the new data.

A common solution to concept drift (the third category) is to retrain the model on the newly scraped data for the websites of the innovative and non-innovative companies [4]. [1] retrained the model on more recent scrapes of the same companies, but this did not improve the deterioration of the model. In fact, classification accuracy on newly scraped websites of the same companies one year after model development declined from 93% to 63% even when including word embeddings [1]. Further investigation into this phenomenon revealed that this was the result of changing website texts over time, and the disappearance some websites that appeared to be highly

informative for the initial model [1, 3]. In short, the original model did suffer from concept drift, but retraining the model on new data did not remedy this problem.

Because the common solution of retraining the model did not restore model performance, [1] instead opted for "model resurrection", defined as the addition of a large number of newly classified observations to the data, increasing the amount of training and validation data. They found that this did improve model stability, seemingly because it allowed the classifier to learn a larger number of words that were positively or negatively related to innovation. Many words included in the resurrected model were synonyms to the words that were highly predictive of innovation (and non-innovation) in the original model [1]. However, this solution may only work temporarily - synonyms that are not present in the larger dataset are not automatically included, so the model may deteriorate again if companies adjust their phrasing after the model resurrection. In addition, adding a large number of new observations to the training set required a lot of manual checking and classification. It undermines a large advantage of using big data sources: any time and resources saved by using these sources are now spent on manual labor. A method that would automatically include information that is semantically similar to the training data could allow for a classification model to learn in a more efficient manner.

1.2 Limiting concept drift

The outcomes of the previous study call for a stable alternative method that is able to measure innovation by means of website texts, while limiting the amount of manual labor required. The foundation of this study emerges from two findings in the original study [1].

First, the model found information in word positions and co-occurrences, resulting in a 33% increase in classification accuracy when including the embeddings as features. Since words that occur positionally close together are seen as similar words, one could consider word embeddings as a very low-level semantic analysis. If the order of the words in a text changes over time, specific words may not be positioned close to one another anymore. Even though the text may remain very similar on a semantic level, word embeddings are sensitive to changes in word positions, and could have difficulties capturing syntactically different but semantically similar information over time.

Second, the resurrected model included a large number of synonyms to

the words that were found to be important in the original model. This is additional evidence that the original model has a limited capability to capture semantically similar information. Sentences may be semantically identical, but if a synonym does not occur in the initial data, the model using word embeddings cannot recognize the semantic similarity between the original word and its synonym. This is a large limitation of bag-of-words based models.

Consequently, we believed that a classifier may benefit from a more semantically oriented approach than the bag-of-words method with (or without) word embeddings. We hypothesized that the text on company websites – in general - remains semantically similar over time, even though they may differ syntactically. Our subsequent goal was to develop a method that could automatically recognize syntactically different sentences that are semantically identical, via which it would be possible to automatically recognize semantically similar information not present in the training data.

1.3 Sentence embedding models

The use of multilingual pre-trained semantically meaningful embeddings at the sentence level could potentially be more robust against changes in website texts than using a bag-of-words representation with word-level embeddings. Sentence-level embedding models split sentences and paragraphs into so-called tokens, which are words, word parts, and special characters. They then map them to a high dimensional vector representing the semantics of the input text. Making use of so-called pre-trained embedding models additionally allow for transfer learning. The latter models are trained on a very large amount of structured and labeled text, resulting in a large and high dimensional semantic vector space to which the text is mapped. Semantically similar text will be grouped close together in the vector space, while semantically different text will be positioned further away. This existing vector space can then inform the calculation of semantically meaningful embeddings for new texts. Instead of inferring the position of an input sentence in the vector space through the surrounding text, the embedding model can compare it to similar sentences that it has previously learned. By comparing the words and their positions of the input sentence to the previously learned sentences, not unlike looking something up in a dictionary, the model will generate an embedding vector that more accurately represents the semantics of a sentence as compared to embedding from scratch. In short, they should be capable of

recognizing semantically similar information of sentences that do not occur in the initial corpus of website texts. These semantically meaningful sentence embeddings are also expected to be more robust against changes in word positions within a sentence, provided the sentence remains semantically similar. This would allow for a more refined analysis of semantic differences between company websites than the (fairly basic) bag-of-words method.

1.3.1 Large Language Models

Large language models from the BERT family are a common, state-of-the-art choice of language representation models. BERT stands for Bidirectional Encoder Representations from Transformers, and was developed to incorporate context bidirectionally (i.e., on both the left and right side of the tokens) [6]. BERT was trained on token prediction using a masked language modeling (MLM) approach: it masks a proportion of random input tokens, and tries to predict the masked tokens using the remaining tokens. It was then also trained on a next sentence prediction task to learn common relationships between types of sentences. In other words, BERT was pre-trained to consider contextual information both within and between sentences. It splits input text into tokens, and generates a fixed length embedding vector of 768 elements for each input token. The training corpora for BERT models consist of the BooksCorpus [7] and English Wikipedia. The authors emphasize that they pre-train on full documents instead of separate sentences, so that the model can learn from long sequences of sentences [6].

Sentence-BERT (SBERT) models are an extension of BERT models. SBERT models add a mean pooling operation to the output of a BERT model to represent entire sentences or paragraphs as a single semantically meaningful feature vector [8]. The pre-trained SBERT models were fine-tuned through a siamese or triplet network structure. This method generated higher quality sentence embeddings than simply averaging the BERT embeddings, and proved to be very hardware efficient. Multilingual frameworks are available, automatically processing over 100 languages (including Dutch and English) [9]. SBERT models truncate input text to 128 tokens, disposing of any tokens that exceed this limit. In other words, SBERT can handle multi-sentence input, but it is less suitable for long texts.

1.4 The present study

In this study, we hypothesized that the use of pre-trained sentence embedding models would be more robust against concept drift as a result of syntactic changes in website texts over time as compared to the original approach [1]. In other words, this approach was focused towards handling the changes in input features and minimizing the resulting decision boundary changes between classes. The sentence embedding method would be considered an improvement (to the initial logistic regression model with word embeddings) if it could achieve two goals. First, it should classify the initially scraped texts with an accuracy comparable to the old model. Second, it should classify later scraped texts with a higher accuracy than the original model, more in range with the accuracy of the initial scrape. The second goal was essentially the most important one, as this would be evidence for increased robustness against concept drift, even if the classification on the initially scraped text was less accurate than in the old model. It would also be a starting point in understanding the underlying mechanisms at play during the original innovative company study.

2 Methods

2.1 Data and data collection

The available data were a combination of survey sources and big data sources. Survey data came from the CIS results from 2016. The reference period for most indicators ranges from the beginning of 2014 to the end of 2016 [10]. For indicators on innovation expenditures, the reference period is 2016 only [10]. We assumed that a company that was found to be innovative in the survey will remain innovative during the period studied, and vice versa.

2.1.1 Website scraping

The big data sources were the website texts from companies involved in the survey scraped by [1], in addition to a sample of website texts from companies in the Dutch Business Register. The first study [1] used a Google search API to find potential links that matched the companies in the CIS, and proceeded to manually check the best matching URLs for each company. This assured that the correct website was found for each company in the sample. The main

page of each website was scraped, and raw HTML-files were extracted and parsed using Beautiful Soup 4 (version 4.12.2)[11] in Python (version 3.9.16). Any style or script information was removed, reducing the HTML-files to the visible text on the websites. Both the original innovation detection model [1] and the proposed innovation detection model have used these data, after preprocessing the text, as a starting point.

2.1.2 Available website texts

The available data from website scrapes are sumarized in Table 1. The use of sentence based approaches required the original HTML-files or the extracted text files, but full texts of the websites that were used to develop the original model (scraped at timepoint $t = 0$) were no longer available: only processed website texts in bag-of-words format had been stored. Raw HTML-files for the scrape at $t = 3$ months after the original model was developed were available, but only a subset of the HTML-files could be matched to company IDs that were present in the original study [1]. Consequently, a subset of the companies used to develop the original model, scraped at timepoint $t = 3$, were used to build the SBERT classification model. A (larger) subset of HTML-files from companies in the original study was available for timepoint $t = 12$. The data at $t = 12$ were used to assess model robustness against concept drift resulting from changes in website texts over time. Performance of the original model and the SBERT model has been compared on these subsets of company websites.

Of the 1,488 companies in the dataset at $t = 3$, 1,440 (97%) were also in the dataset at $t = 12$. 418 out of 1,858 companies in the dataset at $t = 12$ (22%) were not present in the data at $t = 3$.

Finally, a sample from the Dutch Business Register scraped at $t = 3$ was used to assess the external validity of the classification model. The labels for this sample were not sourced from survey data - they were labels predicted by the original model [1]. 101 companies that were included in the model training data were removed from the external validation dataset.

Table 1. Summary of company data used for model building, concept drift assessment, and external validation.

Analysis	Timepoint	Data source	Total n companies	Share of innovative companies
Classification model building	t = 3	Companies in CIS[a]	1,488	54%
Concept drift analysis	t = 12	Companies in CIS[a]	1,858	54%
External validation	t = 3	Independend sample of companies from Dutch Business Register[b]	35,790	54%

[a] Class labels originated from CIS results
[b] Class labels predicted by original model [1]; companies occurring in the model building dataset removed

2.2 Preprocessing

The goal of the proposed model was to limit any preprocessing steps to retain as much of the available information as possible. Using regular expressions, any email addresses were replaced with "email" and any phone numbers were replaced with "phonenumber" to standardize that information. Full website texts were stored as a single paragraph. Classification using full website texts has been assessed despite the aforementioned input token truncation - if the model can accurately classify companies using the first 128 tokens in the text, it would seriously reduce computation time for larger future samples.

In order to enable the classifier to take *all* sentences on a website into account, full company website texts were also split into separate sentences using Punkt Sentence Tokenizer (version 3.7) [12]. Some sentences only consisted of one or two words, but they will still be referred to as sentences for simplicity. If a sentence originated from an innovative company website, it received the label "innovative", and vice versa. A sentence with the label "innovative" then does not necessarily mean the sentence itself implies in-

novation - rather, it implies that that sentence occurred on the website of an innovative company. This allowed for a classification analysis at sentence level: when a classifier can accurately predict whether a sentence stemmed from the website of an innovative company, its classification results can be aggregated to predict innovation status at company level. In other words, the sentence-level classification analysis allowed the model to predict company innovation status using *all* the sentences on their website. Table 2 describes some characteristics of the datasets after splitting full website texts into sentences.

Both full website texts and separate sentences were fed to the pre-trained SBERT model to obtain embeddings at full text level and sentence level. The specific SBERT model used to calculate the embeddings is paraphrase-multilingual-mpnet-base version 2. This SBERT model was trained on paraphrase data, making it very suitable for detecting semantically similar sentences that are phrased differently. The SBERT model was imported with the associated sentence transformers module (version 2.2.2).

Table 2. Summary of sentence-level data used for model building, concept drift assessment, and external validation. Each sentence received the same label as the company from whose website the sentence originated.

Analysis	**Timepoint**	**Data source**	**Total n sentences**	**Share of sentences from innovative company websites**
Classification model building	t = 3	Companies in CIS[a]	183,010	56%
Concept drift analysis	t = 12	Companies in CIS[a]	248,308	55%
External validation	t = 3	Sample of companies from Dutch Business Register[b]	2,900,467	52%

[a] Class labels originated from CIS results
[b] Class labels predicted by original model [1]; companies occurring in the model building dataset removed

2.3 Exploratory analyses

To confirm the assumption that website texts stay semantically similar over time, similarity of company websites at the $t = 3$ and $t = 12$ was checked by the proportion of common words (as in [3]), by manual inspection, and by cosine similarity. Cosine similarity is a common measure to numerically compare the semantics of text embeddings. The manual check was performed by taking a sample of 50 innovative and 50 non-innovative companies and noting any large differences in website content over time. Cosine similarity over time was calculated using the embedding vectors generated by SBERT. We looked at cosine similarity between embeddings using the full website texts as input, but since the embedding model truncates input text that exceeds a given number of tokens, checking semantic similarity this way might not be ideal. Therefore, the embeddings of all separate sentences on each individual company website were also averaged into a single embedding vector representing the overall content of the website. Then, the cosine similarity between timepoints $t = 3$ and $t = 12$ was calculated for each company website as a measure of semantic similarity over time. The downside of this method is that websites with many sentences may result in very "diluted" embedding vectors, since small differences between texts are averaged out.

2.4 Classification

The original study tested a number of supervised binary classification algorithms with default settings [1, Tab. 1]. The algorithms stem from the scikit-learn module (version 1.2.2) [13]. The embedding vectors functioned as features. Most classification algorithms in [1] were tested using the sentence embedding models, enabling the comparison of the performances of the original method and new method. The support vector machine algorithms were not used - these algorithms were exceptionally slow to train at sentence level, which would be impractical for even larger datasets. The algorithm that performed sentence classification with the highest accuracy was selected, and some fine-tuning steps were performed to determine if model performance could be easily optimized. Finally, the sentence-level classification results were aggregated to predict innovation status at company level. Company level classification accuracy was also assessed using full website text embeddings, and using predictions of the original model [1] on the current subset of data. Finally, the selected model was retrained on the full

dataset at $t = 3$, and tested on a sample of 35,790 companies from the Dutch Business Register (scraped at the same timepoint) to assess whether it could generalize to a large, independent sample of company website texts. Because the labels did not stem from survey data (see section 2.1.2), we also assessed classification performance on this sample using the original model trained on the original data at $t = 0$ [1]. Seeing as that model and those data were used to predict the labels in the sample, this analysis served as a measure of quality of the labels.

An online repository containing Python code for all the analyses can be found in Appendix A.

3 Results

3.1 Descriptive and exploratory analyses

3.1.1 Website similarity over time

In [3], website similarity over time was calculated as the proportion of common words to unique words for each website at each timepoint. The conclusion in that paper was that many website texts were notably different over time, with an average similarity of 0.5. We have replicated this analysis for the subset for which we were able to find the raw HTML-files (Fig. 1). The distribution of this measure of similarity was very similar to that in [3, Fig. 3]. For this subset, the average proportion of common words to unique words was 0.5 as well. It appeared that the subset of data used in this study exhibited similar behavior over time to the data examined in [3].

In contrast to the findings in Fig. 1 and [3], the sample of websites that was manually checked for content changes seemed to be quite consistent in website content over time. About 12% of innovative and 18% of non-innovative company websites differed slightly because of company news posts or similar updates. The website texts that differed significantly between timepoints (4% of innovative companies and 2% of non-innovative companies) seemed to be the result of an error in scraping the page: the scrape at one timepoint would contain the homepage, but the scrape at the other timepoint looked like a subpage or dropdown menu content of the company website instead. Overall, the general content of this sample did not change much over time.

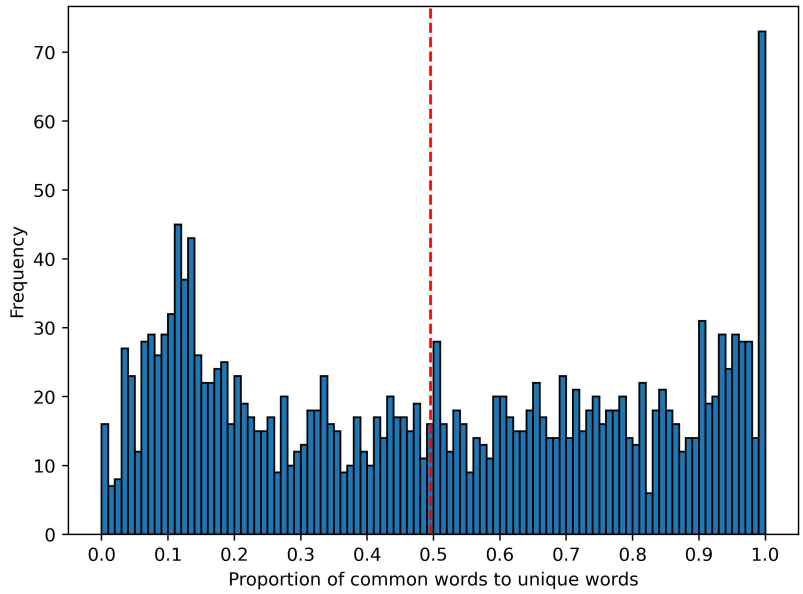

Figure 1. Frequency distribution of the proportion of common words to unique words in website texts at timepoints $t = 3$ and $t = 12$. A proportion of 1 indicates that all words at each timepoint are identical. The red dotted line represents the mean proportion of common words over time.

Analysis of the cosine similarity over time confirmed these observations. Despite the truncation of the input tokens, using full website texts as input yielded higher quality embedding vectors than averaging the embedding vectors of each separate sentence on a website did. SBERT embeddings had high cosine similarity between timepoints, showing that the (truncated) general content of nearly all company websites were highly semantically similar over time (Fig. 2). The median cosine similarity between websites was 0.99 (Fig. 2). Even though these similarities are based on truncated text, it is additional evidence that most company websites remain semantically similar over time even though they may differ syntactically. Summarized, websites seemed to differ syntactically, but not semantically.

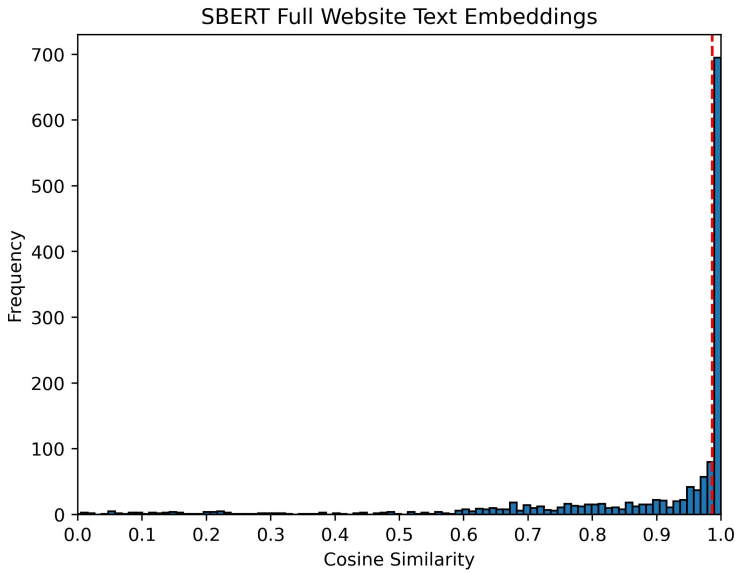

Figure 2. Frequency distribution of cosine similarities of company websites at timepoints $t = 3$ and $t = 12$. A cosine similarity of 1 indicates that the (truncated) company website text was identical between these timepoints. The red dotted line represents the median cosine similarity over time.

3.2 Classification model building

3.2.1 Classification of sentences at $t = 3$

The dataset at $t = 3$ was divided into a train set (80%) and a test set (20%). After training the algorithms on the train set, each algorithm predicted labels for the validation set. The resulting accuracies are shown in Table 3. For sentence level classification, the sentence label was compared to the predicted label for that sentence. Note that sentences received the label of the company in whose website text they occurred, as described in section 2.2.

With an initial classification accuracy of 69%, the random forest was the most accurate in predicting whether a sentence stemmed from the website of an innovative company or not (Table 3). 69% accuracy is not extremely high, but given that these performance metrics result from using default settings for each classifier, we investigated whether it could be further improved by tuning the hyperparameters of the algorithm.

Table 3. Innovative sentence classification accuracy of predicted labels using various classifiers. Higher accuracy represents a greater ability to recognize whether a sentence occurred on an innovative company's website or not.

	Accuracy (%)
Bernoulli Naive Bayes	55
Logistic Regression (L1 regularization)	59
Nearest Neighbors (k = 2)	64
Stochastic Gradient Descent	59
Quadratic Discriminant Analysis (QDA)	61
Neural Network (multi-layer perceptron)	68
Decision Tree	66
Random Forest	69
Gradient Tree Boosting	62

3.2.2 Hyperparameter tuning

The RandomizedSearchCV function from scikit-learn [13] was used to explore the effect of different combinations of hyperparameters for the random forest algorithm. The hyperparameters that were tested were the following:

- Number of decision trees in the random forest. The default value is 100, and adding trees will combat potential underfitting of the model, at the cost of computation time. The number of trees ranged from 100 to 1000, in steps of 100.
- Maximum number of features considered at each split. The default value is the square root of the number of features (resulting in 27 features for SBERT vectors). Larger values allow the classifier to test

more combinations of features. We tried the default setting as well as 10%, 20%, and 30% of the total number of features.

- Maximum tree depth. This controls the complexity of each decision tree in the random forest. Lower values result in simpler trees, capturing simpler patterns in the data than larger trees. This can help reduce potential overfitting. By default, the classifier does not restrict tree depth. The maximum tree depth restriction ranged from 100 to 1000, in steps of 100. The default setting (no restriction) was also tried.

- Minimum number of samples required for a note to be split. The default value is 2, meaning that nodes containing two samples can be split. Increasing this number results in less complex trees with fewer nodes, reducing potential overfitting. The default value was tried, as well as a minimum of 5, 10, and 100.

RandomizedSearchCV does not utilize all combinations of the hyperparameters to be tested - rather, it takes random samples from the possible combinations, thereby reducing computation time. The number of iterations was set to 50.

Different combinations of the set hyperparameters only led to a 1%-2% increase at most, whereas computation time increased to several hours (as compared to several minutes on default settings). In other words, fine-tuning did not reasonably increase sentence classification accuracy.

3.2.3 Predicted class probabilities of sentences at $t = 3$

The predicted class probabilities of the validation set were investigated in a further attempt to improve classification accuracy. The distribution of predicted class probabilities are displayed in Fig. 3. Sentences to the right of the classification boundary were predicted to stem from innovative company websites by the algorithm. The blue bars then show the predicted class probabilities of correctly predicted sentences, while the red bars show predicted class probabilities of sentences that were predicted to be innovative, but originated from a non-innovative company website. Sentences to the left of the decision boundary were predicted to be non-innovative by the algorithm. Similarly, the blue bars represent the predicted class probabilities of correct classifications, while the red bars represent predicted class probabilities of misclassifications.

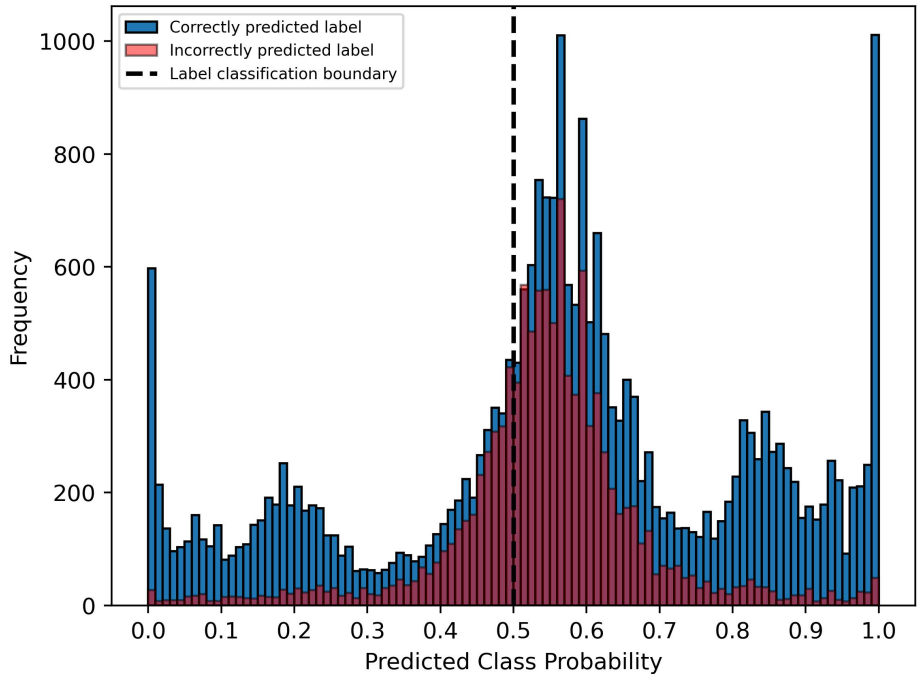

Figure 3. Predicted class probabilities for sentences in the validation set. Sentences to the right of the classification boundary were predicted to be innovative by the model, and vice versa. The blue bars depict predicted class probabilities for which sentence labels were predicted correctly. The red bars depict the predicted class probabilities for which sentence labels were predicted incorrectly.

Interestingly, the bulk of the misclassified sentences had a predicted probability between 0.5 and 0.6. This is likely the result of common sentences in both innovative and non-innovative company websites. For example, if the sentence "Welcome to our website" commonly occurs in both groups, that sentence will have low discriminative power. That is, the sentence is not "typically innovative" or "typically non-innovative", and the probability that it originates from an innovative or a non-innovative company website is approximately equal.

The distribution of predicted class probabilities for correct label clas-

sifications exhibited distinct peaks at both tails of the histogram, around probabilities of 0.2 and 0.85 (Fig. 3). Misclassification in these tails was minimal. This phenomenon suggests that there are sentences with characteristic or "typical" topics for innovative and non-innovative company websites, with certain sentences predominantly occurring in one class and rarely in the other. The frequent occurrence of correctly classified sentences with predicted class probabilities of 0 and 1 indicates the existence of sentences that are highly specific to each class. This phenomenon called for further investigation into class-exclusive topics.

3.2.4 Class-specific topics

Sentences that were correctly predicted to originate from an innovative or non-innovative company website were manually investigated. We were particularly interested in sentences with predicted class probabilities below 0.35, and above 0.75 (i.e., at the approximate edges of the peaks in the tails of the histogram (Fig. 3)). With minimal misclassification within these probability ranges, these sentences seemed to have high discriminative power, and the ability of the model to discern class-specific sentences or topics could inform further model improvement.

NLTK's KMeansClusterer [12] was used within each group for exploratory purposes - it provided some structure to the large number of sentences, making it easier to recognize topic patterns upon manual examination. We performed no optimization for the number of clusters or the number of clustering trials. The function generated 20 clusters for each group using 10 randomized clustering trials, and using cosine distance as the distance metric. For each cluster, we manually examined the 50 sentences with the smallest cosine distance to the cluster centroid, and the 50 sentences with the largest cosine distance to the centroid, to infer the general content of each cluster.

Table 4 describes the results for sentences from innovative websites: it shows the sentence that was closest to the centroid for each cluster, and common general topics within that cluster. The sentence closest to its cluster centroid is not necessarily an accurate representation of the general content in the cluster.

Some of the clusters contained information that does not seem indicative of innovation, generally speaking (Table 4). For example, cluster 12 mostly contained abbreviations and some company names. Many of the abbreviations seemed to be abbreviated names, so they are likely to be out of SBERT's

Table 4. "Typical sentences" for innovative company websites. The left column contains (sometimes one-word) sentences with the smallest distance to the centroid of their respective cluster. The second column contains common topics within each cluster.

	Correctly classified sentences with smallest distance to centroid	**Common terms and topics in cluster**
1.	Videoverdeler.	Video and audio; cameras and equipment; conferencing; film and television.
2.	Digitaal Vastgoed Beheer.	Digital services; software; technology; internet and online; data.
3.	Grondstoffen Nieuws.	Names of Dutch towns, regions, or companies.
4.	Telefoonnummer.	Contact-related terms (phone, email, chat, contact form)
5.	Rouwtransportauto's.	Industrial transportation; logistics; chauffeur; car; bicycle.
6.	Countries.	Country names; languages; geographical terms.
7.	Schapenvoer.	Food and food products.
8.	Projectontwikkeling.	Innovation; development; ICT; solutions; engineering; technology.
9.	Vraaggestuurde website.	Website-related terms; blog post, profile, "click here".
10.	meer nieuws >>.	Read more; more information; news; dates.
11.	Maritieme toepassingen.	Maritime-, fishing-, and water-related terms.
12.	SPAR MVO.	Abbreviations and company names.
13.	Lagertechniek.	Electromechanics; climate control; cooling and heating; fuel; sustainable energy; environment.
14.	ONDERSTEUNING.	Support, solutions, and advice.
15.	Onderhoudsprijzen.	Safety and security; legal and financial.
16.	Randapparatuur.	Technology; electronics; machinery.
17.	Bedrijfsprofiel.	Work, company, and industry; jobs; vacancies; entrepeneurship.
18.	Prodir.	Products and goods. Purchase; delivery; price quotes.
19.	Witgoed accessoires.	Accessories (clothing; interior decoration; etc.).
20.	Siertuin.	*No clear pattern. Very short texts.*

vocabulary, rendering them semantically virtually meaningless. Cluster 20 did not seem to have a coherent pattern in sentence content besides that nearly all sentences were one or two words, possibly the result of the number of clusters that was chosen.

However, the clusters also contain topics related to the words that were most important for predicting innovation in the original study [1]. Examples are cluster 1 (film), cluster 2 (data, technology, software), cluster 4 (contact and contact forms), and cluster 8 (innovation). The model seemed to pick up patterns relating to innovation in a way that was comparable to the original model.

The same cluster analysis was performed on correctly classified sentences from non-innovative company websites (Table 5). These results also reveal a cluster containing mostly abbreviations and names (cluster 12), and two clusters with no discernible pattern in their content (clusters 11 and 20).

In these clusters, terms that were found to be related to non-innovation in [1] emerged as well. Cluster topics include, "Facebook" (cluster 5), "service" (clusters 7 and 19), "sale" and "shopping cart" (cluster 10), "appointment" (cluster 17), and "workplace" (cluster 19). In other words, patterns relating to non-innovation are picked up as well.

The innovative clusters and non-innovative clusters do seem to share some topics. For example, transportation and logistics emerge in both groups. In the non-innovative group, taxis are mentioned more frequently, but the general topic remains approximately the same. Similarly, both groups contain a cluster pertaining to food products and the food industry. In the innovative group, this is mostly limited to products and animal feeders, and in the innovative group there are also sentences referring to kitchens and the hospitality industry. The phrasing or details of these sentences could be distinct between the groups by chance, however, that would mean that some patterns are learned from very small semantic differences between sentences. In other words, the model was able to recognize patterns that were found in the original study [1], but possibly also found patterns that might not generalize to a population. In the end, however, company website texts are aggregates of all the separate sentences. For that reason, classification on company level might be resilient against small amounts of nongeneralizable learned patterns.

Table 5. "Typical sentences" for non-innovative company websites. The left column contains (sometimes one-word) sentences with the smallest distance to the centroid of their respective cluster. The second column contains common topics within each cluster.

	Correctly classified sentences with smallest distance to centroid	**Common terms and topics in cluster**
1.	van EUR 325.000.	"EUR", prices.
2.	EUR phonenumber.	"Phone numbers" (prices)[a], fax.
3.	Millenaar & van Schaik Transport bv.	Transportation; taxis; lifts; names of companies and regions.
4.	Wagenrenk - Diergeneeskundig Specialistisch Centrum Nederland.	Plants and animals; names of cities
5.	Startpagina.	Homepage; Instagram; website; header; navigation; Facebook.
6.	Diepvriespizza.	Food products; catering and hospitality; kitchen.
7.	Online administratie.	ICT (e.g., hosting and tech support); service; software; digital.
8.	Koeltransport.	Industrial transportation; logistics; taxis; automotive brand names.
9.	Binnenland.	Distances; travel and recreation; sports; country names.
10.	Naar de webshop.	Retail related terms. Webshop; retail; "brands"; sale; shopping cart; clothing.
11.	HEKSLUITINGEN.	*No clear pattern. Very short texts.*
12.	Storingsoverzicht.	Abbreviations and names.
13.	Meer weten.	More information; view more; learn; question; FAQ.
14.	Uitvulplaatje kunststof.	Plastic and other materials; civil and electrical engineering; construction; project.
15.	Dieptereiniging.	Cleaning services and supplies; garbage and recycling; asbestos; health protection.
16.	Huurwoning.	Homes and rental; building and construction; hotel.
17.	Voorzieningen.	Appointment; private (particulier); request quote; pay; dates; family-related terms.
18.	Alle producten.	See all; products; categories; brands; outlet.
19.	Bedrijfprofiel.	Company; franchise; organization; administration; service and HR; professional; recruitment; workplace; financial advice; management.
20.	Trui.	*No clear pattern. Very short texts.*

[a] Apparently, some price information and other large numbers were erroneously considered phone numbers during preprocessing. Due to time limitations, this was not corrected for the analyses.

3.2.5 Classification of companies at $t = 3$

The sentence-level classifications were used to predict labels on company level (Table 6). Predicted labels on company level were calculated in several ways. First, companies were predicted to be innovative if the average of predicted positive sentence labels ($\overline{\text{Predicted label}_{\text{Innov}}}$) was above 0.5. In short, the majority of sentences on a website should be predicted to stem from an innovative company website for the predicted label to be "innovative". Second, the average predicted class probabilities of all sentences within each website were used. Companies were predicted to be innovative if the average predicted class probability of all sentences on their website ($\overline{p}_{\text{Innov}}$) exceeded a given threshold.

In addition, the original logistic regression model with word embeddings [1] was trained and tested on the subset of websites for which the full texts were available. For classification using full text embeddings, classification accuracy was also assessed using the predicted labels by the algorithm as well as using thresholds for the predicted probability of innovation status (p_{Innov}), similar to the analysis at sentence level. This enabled us to check whether full text embeddings (as opposed to separate sentence embeddings) would be of sufficient quality for classification, seeing as it greatly reduced computation time.

Classification accuracy was mediocre for the original logistic regression model with word embeddings [1] on the current subset, and for full website text embeddings (Table 6). The original model is much less accurate on this subset than it was on the full data, reaching an accuracy of only 60% as compared to an accuracy of nearly 90% at $t = 3$ [3, Fig. 1]. [1] mentioned that their model contained a small number of highly informative website texts that seemed to be essential for model performance - these websites likely do not occur in this subset. The labels calculated using full website text embeddings were, at most, 62% accurate at company level. This finding implies that informative sentences occur beyond the point of input token truncation.

Label prediction using the aggregated sentence embedding classification results produced much more accurate class labels at company level. Using the average of positive labels predicted by the algorithm resulted in 70% accuracy. However, using the average predicted class probability resulted in a classification accuracy of up to 86% at a threshold of $p = 0.55$ - the point at which the probability distributions of both correctly classified and

misclassified sentences are approximately symmetrical. Sentences that occur on the websites of both classes should be predicted to have around a 50% chance of belonging to either class. Setting the threshold at $p = 0.55$ ensures that, on average, sentences of ambiguous origin dominate less when predicting the innovative status of a company. The resulting classification accuracy of 86% is very close to the classification accuracy after model resurrection in the original study [1].

In summary, the best classification model at sentence level was the random forest. Tuning hyperparameters did not substantially improve model performance, likely due to a large number of sentences occurring in websites of both classes. The optimal way of predicting labels at company level was to classify a company as innovative if the average of predicted class probabilities for the sentences exceeded a threshold of 0.55. After these findings, the random forest was trained on the entire data at $t = 3$ and used to predict

Table 6. Classification accuracy at company level for innovative companies using predicted labels and predicted class probabilities.

	Accuracy (%)
Original model[1] on current subset	60
Full text embedding	
Predicted label	58
$p_{\text{Innov}} > 0.5$	58
$p_{\text{Innov}} > 0.55$	62
$p_{\text{Innov}} > 0.55$	62
Sentence embedding	
$\overline{\text{Predicted label}_{\text{Innov}}} > 0.5$	70
$\overline{p}_{\text{Innov}} > 0.5$	79
$\overline{p}_{\text{Innov}} > 0.55$	86
$\overline{p}_{\text{Innov}} > 0.6$	78

the data at $t = 12$ in the concept drift analysis.

3.3 Concept drift

3.3.1 Classification at $t = 12$

As described in section 2.1, there was a large overlap in companies in the datasets at $t = 3$ and $t = 12$, with the 88% of the companies in the latter dataset occurring in the former. The model was trained on the full dataset at $t = 3$, after which the full dataset at $t = 12$ was classified, according to the best procedure found in section 3.2.

The model trained on the data at $t = 3$ was able to classify sentences and companies with high accuracy (Table 7). Classification accuracy was highly similar for the datasets at $t = 3$ and $t = 12$, likely due to the overlap of companies contained in both datasets. However, the fact that classification accuracy does not decrease over these timepoints is a confirmation of the assumption that company website texts do not change considerably over time, at least on a semantic level.

Table 7. Classification accuracy at sentence-level and company-level at $t = 3$ and $t = 12$.

	Accuracy $t = 3$ (%)	**Accuracy** $t = 12$ (%)
Sentence-level prediction	69	72
Company-level prediction ($\overline{p}_{\text{Innov}} > 0.55$)	86	87

3.3.2 External validity

External validity was assessed by training the model on the dataset at $t = 3$, and predicting sentence-level and company-level labels for 35,790 companies sampled from the Dutch Business Register that were scraped at the same timepoint. Table 8 shows that there are issues in predicting the labels for the Dutch Business Register sample. The SBERT model was only able to classify sentences with 53% accuracy, and companies with 59% accuracy. However, the original model with the original data used to develop that model classified these companies with 62%, despite this model being used to predict these labels in the first place.

Around 54% of the companies in the sample were labeled as innovative; the SBERT model predicted 58% of companies to be innovative. The original model only predicted 20% of companies in the sample to be innovative. In 2016, Eurostat estimated the share of innovative companies in the Netherlands to be 60% [14], which is very close to the results of the classification model using SBERT. However, it was unclear whether the cause of mediocre prediction accuracy lay within the model(s) or within the data labels.

Table 8. Classification accuracy for the Dutch Business Register sample at company level for the original model and the SBERT model, and at sentence level for the SBERT model.

	Original model [1] accuracy (%)	**SBERT model accuracy (%)**
Sentence-level prediction	-	53
Company-level prediction	62	59

4 Discussion

The goal of this study was to build a classification model that could accurately classify companies as innovative by using the text on their websites. The model would be considered an improvement to the original model developed for this purpose [1], if it could classify companies with similar accuracy at timepoint $t = 3$, and if classification accuracy remained relatively stable for data scraped at $t = 12$. Given the previous findings, we hypothesized that websites changed syntactically over time while remaining semantically similar, and that analysis using SBERT embeddings would be robust against the syntactic changes.

The content of company websites was indeed found to be semantically similar over time. While there were changes in the specific words that occurred on these websites over time (Fig. 1), manual inspection and a cosine similarity analysis of the SBERT embeddings revealed that the difference was mainly syntactic and not semantic.

During model building, the random forest was found to be the best performing classifier at sentence level, with an initial classification accuracy of 69%. The distribution of predicted class probabilities revealed a large number of sentences that had low discriminative power between classes, in addition to potentially "typical" innovative or non-innovative sentences with high discriminative power. Sentences with particularly low or high probabilities of belonging to an innovative company website were manually investigated. An exploratory cluster analysis revealed some clusters that seemed typical to each class, with topics corresponding to words that were found to be important for class prediction in the original study [1] (Table 4 and 5). However, some types of sentences with a high predicted probability of originating from an innovative company website did not seem to be indicative of innovation. It is possible that these types of sentences occurred mainly on innovative company websites by chance, which was then recognized as a pattern by the model. In addition, some topics were predicted to have both low and high probabilities of originating from innovative website companies (e.g., transportation). The occurrence of these clusters of sentences in both classes was likely the result of very small semantic differences between them. This introduces the risk of overfitting - the model essentially learns that very specific sentences are related to innovation status, when in reality, it should find patterns in general topics so that it is generalizable to the population.

Using the aggregated results of this analysis, we were able to predict

company innovation status with 86% accuracy, similar to the accuracy of the original model at $t = 3$ [3]. However, at $t = 12$, classification accuracy for (mostly) the same companies was 87%, whereas the original study only reached 63% accuracy at $t = 12$. In other words, the SBERT model is more robust to concept drift as a result of syntactically changing website texts, and can be considered an improvement to the original model [1]. However, it can be expected that this model is not robust against concept drift resulting from the emergence of new topics. Should this happen over time, this method could be combined with concept drift detection and adaptation methods.

The external validation of the model, however, requires a more detailed analysis. The original model and data that were used to predict the very labels of the independent sample, do not predict the same labels with high accuracy. It is unclear whether the SBERT model falls short classifying an independent population sample, or whether the labels in the dataset are incorrect. The SBERT model did predict a number of innovative companies similar to the published estimation for that year [14], but we cannot say that external validity was achieved.

External validation can be further assessed by use of an independent dataset for which the labels are certainly correct. This would give additional insight into whether the performance on the external validation set in this paper resulted from the model overfitting during training, or whether it resulted from erroneous labels in the validation set. The newly classified companies from the Dutch Business Register used for model resurrection in [1] could be a starting point: the labels were determined after manual inspection of a large sample of company websites, so they are likely to be highly accurate. The resurrected model in [1] was able to detect innovative companies with 88% accuracy - classification using the SBERT model should be at similar accuracy for that data. Due to time constraints, however, an external validity analysis using those data was not performed in this paper.

Another suggestion for future research on detection innovation in the Netherlands is to develop a model trained on characteristic sentences - this may be a start in estimating innovation with consistent accuracy over time. The cluster analysis may serve as inspiration, but the question of what is "typically innovative" remains. Perhaps specific country names and geographical terms are indeed related to innovation, perhaps they occurred only in innovative company websites by chance. Employing field experts may support the creation of such a model, but it would require a lot of manual work. This could help ensure the external validity of the model, if that does prove

to be an issue upon further investigation. Creating a model based on "typical" sentences would then prevent the model from learning extremely specific patterns in the training data, making it more flexible in its classification of new data.

Overall, the findings of this paper provide a starting point in concept-drift resilient estimation of innovation activity in the Netherlands. SBERT embeddings provide extra robustness against model degradation as a result of syntactic website text changes. In due time, proper assessment and optimization of model generalizability may even help pave the way towards the use of big data sources in official statistics.

Acknowledgments

I would like to express my sincere gratitude to Piet Daas, who provided excellent supervision and guidance throughout this project. His encouragement to explore different approaches independently was both greatly enjoyable and very valuable.

I would also like to extend an honorable mention to my mother, Nienke Peereboom, for her insights when brainstorming solutions to the research problem. Her observations inspired a large part of the approaches used in this paper.

References

[1] P. Daas and S. van der Doef, "Detecting innovative companies via their website," *Statistical Journal of the IAOS*, vol. 36, no. 4, pp. 1239–1251, Nov. 25, 2020, ISSN: 18747655, 18759254. DOI: 10.3233/SJI-200627. [Online]. Available: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SJI-200627 (visited on 01/22/2023).

[2] J. Axenbeck and P. Breithaupt, "Innovation indicators based on firm websites—which website characteristics predict firm-level innovation activity?" *PLoS ONE*, vol. 16, no. 4, 2021. DOI: 10.1371/journal.pone.0249583.

[3] P. Daas and J. Jansen, "Model degradation in web derived text-based models," in *CARMA 2020 - 3rd International Conference on Advanced Research Methods and Analytics*, Universitat Politècnica de València, Jul. 8, 2020, pp. 1–8, ISBN: 978-84-9048-832-4. DOI: 10.4995/CARMA2020.2020.11560. [Online]. Available: http://ocs.editorial.upv.es/index.php/CARMA/CARMA2020/paper/view/11560 (visited on 01/22/2023).

[4] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2018, ISSN: 1041-4347, 1558-2191, 2326-3865. DOI: 10.1109/TKDE.2018.2876857. [Online]. Available: https://ieeexplore.ieee.org/document/8496795/ (visited on 02/15/2023).

[5] F. Bayram, B. S. Ahmed, and A. Kassler, "From concept drift to model degradation: An overview on performance-aware drift detectors," *Knowledge-Based Systems*, vol. 245, p. 108 632, Jun. 2022, ISSN: 09507051. DOI: 10.1016/j.knosys.2022.108632. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0950705122002854 (visited on 02/15/2023).

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019. arXiv: 1810.04805 [cs.CL].

[7] Y. Zhu, R. Kiros, R. Zemel, *et al.*, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 19–27. DOI: 10.1109/ICCV.2015.11.

[8] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Nov. 2019. [Online]. Available: https://arxiv.org/abs/1908.10084.

[9] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Nov. 2020. [Online]. Available: https://arxiv.org/abs/2004.09813.

[10] *Community innovation survey 2016 (cis2016) (inn$_c$is10)*, Jan. 2020. [Online]. Available: https://ec.europa.eu/eurostat/cache/metadata/en/inn_cis10_esms.htm.

[11] L. Richardson, "Beautiful soup documentation," *April*, 2007.

[12] S. Bird, E. Klein, and E. Loper, *Natural language processing with python*, ed. by O. Media, Jun. 2009. [Online]. Available: http://nltk.org/book/.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[14] *Community innovation survey: Latest results - products eurostat news - eurostat*. [Online]. Available: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/DDN-20190312-1.

A Online Repository

The code used for the analyses described in this paper can be found at `https://github.com/sannepeereboom/SBERTInnovativeCompanies`.

The data that were used in the analyses contained identifiable information. As a result, the repository does not contain the data files used for the analyses.