Center for
Big Data Statistics
Working paper

Using Website texts to detect Innovative Companies

Working paper no.: 01-21

Piet J.H. Daas
Suzanne van der Doef

January 2021

Contents

Abstract

Producing an overview of innovative companies in a country is a challenging task. Traditionally, this is done by sending a questionnaire to a sample of companies. This approach, however, usually only focuses on large companies and completely misses small companies, such as startups. We therefore investigated an alternative approach: determining if a company is innovative by studying the text on its website. For this task a model was developed based on the texts of the websites of companies included in the Community Innovation Survey of the Netherlands. The latter is a survey carried out every two years that focuses on the detection of innovative companies with 10 or more working persons. We found that the text-based model developed was able to reproduce the result from the Community Innovation Survey and was also able to detect innovative companies with less than 10 employees, such as startups. Model stability, model bias, the minimal number of words extracted from a website and companies without a website were found to be important issues in producing high quality results. How these issues were dealt with and the findings on the number of innovative companies with large and small numbers of employees are discussed in the paper.

1 Introduction

In our modern world, more and more data are generated on the web and produced by sensors in the ever growing number of electronic devices surrounding us. These data have very interesting potential applications such as providing novel insights on the activities of companies (Gökk et al. 2015), to inform policy makers (Höchtl et al. 2015) and also for official statistics (Florescu et al. 2014), especially when performed at large scale. However, extracting relevant and reliable information from Big Data sources in a reproducible way is not an easy task (Kitchin 2015, Daas et al. 2015).

In this paper we describe the results of our research on the development of a method that is able to determine if a company is innovative or not, based on the text on its website. Here, innovation is defined as all activities aimed at renewal within a company (Statistics Netherlands 2019a), where renewal is the act or process of making changes in order to improve the company so that it becomes more successful. Detecting innovative companies is traditionally done by a survey in many European countries: the Community Innovation Survey (CIS). However, this survey only focuses on a sample of 10,000 companies[1)] in the Netherlands, it only includes companies with 10 or more working persons, it is carried out once every two years, and it takes considerable time to produce the final outcomes (Eurostat 2019). The goal of the work described in this paper was the development of an approach that could be applied to all companies in a country, with special attention to small companies (companies with less than 10 working persons), and could be produced in a relative short time without putting any administrative burden on the companies involved.

1.1 Related work

A number of studies have been performed looking at company webpages that very much resemble the approach described in this paper. The work of Gökk et al. (2015) focussed on a set of nearly 300 company websites in the United Kingdom, of both small and medium-sized companies, to determine the potential of using web derived data to detect clues on their research and development activities. They conclude that webpages are a useful complement and may offer new insights not easily obtained from other sources. Baudry et al. (2016) report on a webpage study of Canadian companies focussed on detecting innovation. In this work, 133 websites were investigated and compared to survey results. Here, fairly low correlations were found but their conclusion was that some of the data extracted could be used as proxy for specific variables obtained from more classical methods. Mirończuk and Protasiewicz (2016) have investigated the relation between the content of documents available on a company's web domain and its innovative nature for companies in Poland. They developed a combination of Bayesian models (an ensemble) for this task which provided acceptable results. Kinne and Lenz (2019) and Kinne and Axenbeck (2018) have published studies in which large numbers of company webpages were investigated. In the 2019 paper, the focus was specifically on the development of web-based innovation indicators. By using the German CIS survey as a

[1)] The CIS-survey is send to enterprises (Eurostat 2019). However, because of the focus on websites in our work and because of the challenging relation between a website and a statistical 'business' unit, especially for large businesses, we have decided to use the more general term 'company' in this document to indicate the business organization on which the content of a website focusses. The reader should be aware that for small businesses the relation between a website and the statistical unit of interest is usually one-on-one.

starting point a model was developed to detect product innovation based on company texts; the average F1-score of the model on both classes was 0.8. Several hundreds of thousands company websites were classified with the model, revealing reliable predictions with a high regional granularity.

1.2 Goal of the study

With the findings of the study of Kinne and Lenz (2019) as inspiration, we investigated the possibility to use the webpages of companies in the Netherlands to detect innovation for both large and small companies, focused on the stability of the model developed over time and looked at the validity of the results. These topics and their application to official statistics have not been described in similar work by others. In section 2, the results of the CIS survey are combined with website data for the companies included. This required obtaining the corresponding Uniform Resource Locator (URL) of the website for each company. Next, it is described how the websites were scraped and how the text was extracted and processed. Section 3 subsequently focuses on the development of a model that enables the detection of innovation based on these texts. In addition, applying this model to other, smaller companies and stability of the model over time are discussed. These issues were found to be the major challenges in this work and are essential to assure future use. In section 4 the percentage of innovative companies without a website is estimated and the numbers of large and small innovative companies in the Netherlands are estimated by the approach developed. Final conclusions are drawn in section 5, which also includes a discussion on the application and future research.

2 Combining and processing data

2.1 Survey data

The CIS survey was the starting point for our work. The focus of this electronic questionnaire is on product, process, organizational and marketing innovations. In this biennial survey, companies self-report by indicating whether they are innovative or not in these areas. In the Netherlands, very large companies, those with 200 or more working persons, are always included in the survey while a stratified sample is drawn of companies with 10 or more working persons. In total around 10,000 companies are included in the CIS survey. Let it be clear here that companies with less than 10 working persons are not included in this survey nor are the small innovative companies estimated in any other way. For our work we used the response of the 2016 CIS survey, which covered the reporting period 2014-2016, and focused on the detection of technological innovative companies (Statistics Netherlands 2019b). Technological innovation is about renewing or greatly improving products or services or the processes that produce products and services (Statistics Netherlands 2019c). Of the respondents, all technological innovative companies were selected and a random sample of around 3,000 non-innovative companies was drawn. This resulted in a selection of 3,340 innovative and 3,002 non-innovative companies which were used to develop a model.

2.2 URL finding

At the start of our study, the URL of the website of a limited set of companies was known at Statistics Netherlands. We therefore had to search for this link for the total of 6,342 companies selected with the URL-finding approach developed in the first ESSnet Big Data (Stateva et al. 2017). This means that the Google search API in combination with the name and location information of each company was used to collect possible links to websites. The best matching URLs were selected and were subsequently manually checked to assure they referred to the correct company. More specifically, the name of the company, the address and the Chamber of Commerce number (if available on the website) were used to confirm this relation. In the end, the corresponding websites were found for 3,338 innovative (99.9%) and for 2,876 non-innovative companies (95.8%) in the CIS survey. During the study, for increasing number of companies URLs became available.

2.3 Webscraping and processing

Next, the websites were scraped. The page the URL referred to was scraped and -optionally-all pages referred to by that page, if located in the same domain, were additionally scraped. The urllib.request function in Python 3.7 was used to download the pages. We compared this scraping method with several alternative approaches, such as those using Node.js or Selenium. URLs that could not be scraped during the first attempt were visited at least four times -at later points in time- to deal with websites that were temporarily unavailable. The raw-html file obtained was parsed with the Beautiful Soup 4 library (v 4.7.1), backed up on the local machine, and processed in several stages. In the first stage, all script and style sections were removed, followed by language detection of the visible text with the langdetect library (v 1.07). Since the majority of the pages were either written in Dutch or English, we only discerned between those languages here; i.e. any non-Dutch page was considered written in English. Subsequently, all words were converted to lower case and all punctuation marks, numbers, and all words below a specific number of characters, either 2 or 3, were removed. Next, depending on the language detected and needs, any words included in the NLTK stop words list (v 3.4.1) were removed followed by mapping of the different morphological variants of the remaining words to their base form; i.e. stemming. For this the SnowbalStemmer library (v 1.2.1) was used. This resulted in a list of stemmed words for each webpage which were subsequently stored.

3 Model development

Since the goal of our study was to determine if the text on a webpage of a company could be used to detect whether they are technologically innovative or not, a model was specifically developed for this task. We used a data-driven modelling approach (Breiman 2001, Hand 2019). Not all of the 3,338 and 2,876 websites of the innovative and non-innovative companies could be accessed. A total of 2,581 (77.3%) and 2,299 (79.9%) webpages for each class could be scraped. After processing, 2,529 (75.8%) and 2,236 (77.7%) of the files contained 10 or more words for the innovative and non-innovative companies class, respectively. Since the outcome is known for each company and only two classes exist, innovative and non-innovative, this is a supervised binary classification task. Prior to any of the subsequent variable selection and model development steps, the dataset was randomly split in a 80% training and a 20% test (holdout) set.

3.1 Variable selection

To enable the development of a model based on the words in text files, the well-known representation in the form of frequency-annotated bag-of-words was used (Aggarwal 2016, chap. 13). A document-term matrix was created in which the rows corresponded to the company webpages and the columns to the words selected from the text collection. For each word in the processed webpage text, the term frequency-inverse document frequency (tf-idf) was included in each cell. Tf-idf is generally considered a good way to identify words that characterize the topics in a text (Aggarwal 2016, Gentzkow et al. 2019). Many of the tf-idf values were zero. During this work we found that taking the log of the tf-ifd+1 value increased the accuracy of the model by 1%. The scikit-learn library (v 0.21.2) in Python was used for this task (Pedregosa et al. 2011). To the document-term matrix, the language of the webpage was included as a binary feature. Word embeddings, as implemented in the word2vec skip gram and Continuous Bag Of Words algorithms of the gensim library (v 3.4.0), could also be included. Word embeddings are based on word co-occurrences and enable an improved extraction of topic information by encoding the semantic and syntactic information of words (Allen and Hospedales 2019, Li and Yang 2018). Various numbers of word embedding based features could be added to the document-term-matrix. This matrix, dimensions 3,904 by 257,976, formed the start for model development.

3.2 Classification

A wide variety of algorithmic classification methods, included in the scikit learn library, were applied to select the most appropriate one for the task at hand. However, their results are obviously affected by the decisions made during data collection and subsequent processing steps. The most important ones considered in this study are i) the type of webscraping used, ii) the level of webpages scraped (only main or main and subsequent pages referred to), iii) the pre-processing steps applied (e.g. remove stop words or not, stem words or not and the minimum number of characters of the words included), iv) the minimum number of words needed for model development, v) the minimum document frequency of the words in the document-term matrix, vi) the number of word embeddings based features added, and vii) the need to use language specific models. During model development, the following general findings were observed. i) When different types of webscraping methods were compared it became clear that models based on pages collected by the urllib.request function slightly outperformed those based on pages collected by Node.js and Selenium. The latter both convert JavaScript code to text which is subsequently included in the page stored. Such code is ignored by the urllib.request function. Apparently the JavaScript part provided no additional information related to innovation. ii) The findings of only scraping the page the URL referred to or to combine it with all additional pages, within the same subdomain, to which this page refers resulted in hardly any improvement of the model. This observation and the fact that it was foreseen that a huge number of company webpages would be studied, let us decide to focus on the findings based only on the page the URL referred to. Downside of this decision is that pages that only contain a limited number of words may not be classified (see vi). iii) When comparing the effect of various preprocessing steps it was found that stop word removal and stemming of words both improved the classification results. The effect of including 2 or 3 characters words and more was found to be very important and is discussed separately below. iv) The minimum number of words required for model development was determined by comparing the effect of a range of values; i.e. 5, 10, 15 and 20. It was found that 10 words was the absolute minimum number of words needed. Models based on fewer words performed poorly. v) The minimum number of times a word needs to occur in all documents to be included in the document term matrix, i.e. the minimum document frequency, is a very important selection criteria. Most important effect is that this value filters out all less occurring words and -as a consequence- seriously reduces the document term matrix. Values between 50 and 300, in steps of 50, were compared and 100 was found the most appropriate value. Models with lower and higher values usually performed less. vi) Word embeddings based features could be included or not. The effect of including various numbers of dimension based features, between 100 and 400 dimensions in steps of 100, were compared. It was found that using 200 features sufficed. Including more had no additional effect. vii) The language of a webpage was classified as either Dutch or English. Comparing the findings of a single model with those of two language specific models demonstrated that the first performed better than the combined results of the latter. This is likely due to the fact that the majority of the webpages was written in Dutch (87%) and that these usually also contained some English words.

Based on the general observations described above, the findings of various classification algorithms were compared. The results of various performance metrics, as derived from the confusion matrix of the predicted and actual class results on the test set, were compared (Olson and Delen 2008). We found that accuracy provided the most useful metric. The accuracy of the models developed are shown in Table 1 in which the effect of including words of minimal 2 and 3 character lengths and the additional effect of 200 word embeddings based features on words of 3 and more characters -on model development- are listed. The average accuracy and standard deviation of a 1,000 independent models, developed on the training set and evaluated on the test set, are shown. Methods were tested with their default settings unless indicated. Including combinations of two adjacent words (bigrams) had no additional effect on the results shown except for a slight improvement of the results when only words of 3 and more characters were included. Various characteristics of the companies, such as the economic activity code and number of working persons, were also added but we found no improvement on the text-based classification results.

From Table 1 it is clear that when words of 2 or more characters are used (column 2) a lot of the methods performed extremely well. Compared to results based on words of only 3 or more characters (column 3), this indicated that the 2 character words had a high information content for the topic studied. Inspecting the variables and their coefficients in the logistic regression model confirmed this observation. Of the 20 words with the highest coefficients in

Table 1. Accuracy of the various classification algorithms tested.

Classification algorithm	Words of 2 and more characters (%)	Words of 3 and more characters (%)	Words of 3 and more characters + word embeddings (%)
Bernoulli Naïve Bayes	87 ± 1	60 ± 1	61 ± 2
Logistic Regression (L1 regularization)	94 ± 1	60 ± 2	93 ± 1
Nearest Neighbors (k = 2)	61 ± 2	52 ± 1	58 ± 1
Support Vector Machine (linear)	93 ± 1	60 ± 1	81 ± 1
Support Vector Machine (radial basis)	53 ± 1	53 ± 1	57 ± 3
Stochastic Gradient Decent	93 ± 1	58 ± 2	79 ± 3
Quadratic Discriminant Analysis	77 ± 8	57 ± 2	56 ± 2
Neural Network (multi-layer perceptron)	92 ± 1	62 ± 2	74 ± 3
Decision Tree	94 ± 1	54 ± 1	61 ± 2
Random Forests	94 ± 1	56 ± 2	64 ± 1
Gradient Tree Boosting	94 ± 1	59 ± 1	71 ± 1

Default settings were used. The average and standard deviation of 1,000 tries are shown.

that model, 18 had a length of 2 characters; such as ‘nl’, ‘de’, and ‘en’. Although it is great to find so many well performing models, these findings prompted us to inspect the origin of the 2 character words more carefully. It was found that many of these words originated from email and web links displayed on webpages, suggesting that innovative companies are more focused on displaying contact information and links. This was, however, not the relationship we were looking for as this finding suggests that a website solely filled with email addresses and links would be classified as innovative. We therefore decided to ignore the 2 characters words and focused on developing a model based on words of 3 characters and more. It is important to realize that because the words are stemmed after character length selection this can still result in the occurrence of (a few) 2 character words which may end up in the model. As mentioned above, the accuracy of the models based on words of 3 and more characters are also shown in Table 1 (column 3). From this it is clear that the overall performance decreased considerably, dropping from around 90% to around 60% for many of the classification methods. Because of this, we tested the effect of additionally including word embeddings. Within the Natural Language Processing (NLP) community it is a well-known fact that word embeddings often improve the extraction of topic information from texts (Li and Yang 2018). Including word embeddings indeed had a positive effect on a considerable number of the approaches tested; compare column 3 and 4 in Table 1. However, only one of them reached a very high accuracy, viz. logistic regression with L1 (lasso) regularization had an accuracy of 93% ± 1. Since a high accuracy and an interpretable model are preferred, certainly within the context of official statistics, logistic regression was clearly the classification method of choice. This model had a precision of 99% ± 1, a recall of 87% ± 1 and a F1-score of 93% ± 1. A total of 350 features, including 180 unique stemmed words, were included in the model. All work subsequently described below is based on logistic regression based models.

3.3 External validation and application to small companies

To test the external validity of the model, we first applied it to the websites of startups. Although it is not an officially defined term, startups are usually described as young innovative companies active in technology or an associated area with a small number of employees (Robehmend 2016). Links to the webpages of 1,078 Dutch startups were obtained from the dutchstartupdatabase.com website; this website is unfortunately no longer active at the moment of writing. A total of 955 webpages (88.6%) could be accessed and scraped and 933 contained 10 or more words after processing (86.5%). Of the 933 webpages, 855 were found innovative according to our model. This is 79.3% (855/1078) of all startup links obtained and 91.6% (855/933) of all webpages classified. The 933 classified websites were manually judged which confirmed the findings of the model in 921 cases (98%). These results indicated that the model developed on websites of companies with 10 or more employees also performed well on webpages of companies with (very likely) much less employees. It also confirmed our expectation on the innovative character of many startups.

The model was subsequently applied to the webpages of companies with less than 10 employees for which a URL was available in the Business Register of Statistics Netherlands. A total of 542,066 webpages were attempted to be scraped which resulted in 490,413 pages (90.5%) actually being collected. After processing, 466,523 webpages contained 10 or more words (86.1%). Of these 194,508 were classified as innovative (35.9%). The pages of a random sample of 1,000 classified companies were subsequently manually checked, in a double-blind test, which revealed that 95% of them were indeed correctly classified. The confusion matrix of this classification check was: True Positives 313, False Positives 28, True Negatives 638 and False Negatives 21. This value is comparable to the accuracy of the model on large company webpages. This demonstrated that the model developed could be applied to both large (10 or more employees) and small (less than 10 employees) company websites. Comparison of the Likelihood ratios for the detection of innovative and non-innovative companies (McGee, 2002), which are 22 and 15 respectively, indicate that the former are somewhat better identified.

3.4 Model long-term stability

However, not everything worked out well. After several months of applying the model to the freshly scraped webpages of Dutch companies, including some of those selected for the 2016 CIS-survey (see section 2.1), we noticed a deterioration of the models performance. This resulted in less and less innovative companies being classified as innovative. This indicated that the texts and the distributions on which the model was based were not static. Reduction of the long-term stability of a model is not uncommon for text based classifiers and has, for instance, also been observed in news topic classification (Kim and Hovy 2006) and event detection and tracing (Atefeh and Khreich 2015). We therefore looked in the scientific literature for ways to deal with this issue. In the Machine Learning community, especially by those that study so-called data streams, this phenomenon is generally described as ‘concept drift’ (Lu et al. 2018). The two most common suggested solutions are: i) retraining the model on new data or ii) adding new data to the original data followed by model retraining (Janardan 2017, Gama et al. 2014). Both approaches were investigated.

First we ‘freshly’ scraped and processed the websites of the original, CIS included, companies. On website data scraped 6 months after the original model was developed, the accuracy of the new model was found to be 76% while after one year it was as low as 63%. Clearly, retraining the model on new data did not work. These observations also suggested a change in the content of the CIS-included websites of one or both groups discerned. This, obviously, could be caused by companies updating their website during this period. Comparison of the first, original dataset, and the data set scraped a year later confirmed this difference and additionally revealed that around 285 websites were absent in the latter data set because they were no longer available online. These included 156 innovative and 129 non-innovative companies. The texts of these websites, which were only included in the first data set, were found to be essential for highly accurate model development; see Daas and Jansen (2020) for more details. This explained why the decrease in accuracy could not be regained by retraining. From these findings it is obvious that sticking to the original data set of companies would not solve the model stability problem.

The second suggestion was to add newly classified data to the original data set and retrain the model. We found that this worked when the classified websites of innovative startups and those of a large number of websites from companies (both large and small) in the Business Register were jointly included. Best results were obtained when a sample of 20,000 classified and manually checked websites from companies in the Business Register and all of the startup webpages classified as innovative, 855 in total, were added. The document term matrix produced had the following dimensions, 20,588 by 426,270. Here, an accuracy of 88% ± 1 on the test set was obtained for the new model. By comparing the results of the updated model on webscraped data for various samples of CIS survey selected companies that were not included in model development, scraped at 6 and 12 month after original scraping, it was demonstrated that the model trained on the combination of the three datasets produced the best overall results. Increasing the number of Business Register websites to 30,000 and more did not additionally improve the accuracy of the model and model development took longer to complete. As an additional preprocessing step, many of the common words included in website texts that are obviously not related to innovation, such as the day of the week and the month of the year, were removed prior to creating the document-term matrix which accelerated model development without negatively affecting the accuracy. The precision, recall and F1-score of the new model were 88% ± 1, 84% ± 1 and 86% ± 1, respectively, for the 6 and 12 month scraped data. A total of 584 unique stemmed Dutch and English words were included in the model. To provide an overview of the features included, the words with the 20 most positive and negative standardized coefficients in the logistic regression model are shown in Table 2. For Dutch words an English translation is added.

Table 2. The 20 stemmed words with the highest and lowest standardized coefficients in the new logistic regression model developed.

Positive			**Negative**		
Stemmed word	English translation	Model coefficient	Stemmed word	English translation	Model coefficient
com	com	13.399	aanbied	sale	-15.311
system	system	13.090	kop	buy	-15.035
inspiratie	inspiration	12.073	creat	create	-14.720
data	data	11.225	powered	powered	-10.421
technologie	technology	11.158	vorm	shape	-10.263
doe	do	10.465	exclusiev	exclusive	-9.889
agenda	agenda	10.327	activiteit	activity	-9.382
analys	analysis	9.872	tijd	time	-8.756
trot	proud	9.559	set	set	-8.713
check	check	9.423	us	us	-8.562
complet	complete	9.327	facebok	facebook	-8.523
softwar	software	9.224	schad	damage	-8.085
detail	details	8.933	kijk	look	-8.033
film	movie	8.788	winkelwag	shopping cart	-7.946
dutch	dutch	8.726	afsprak	appointment	-7.796
contactformulier	contact form	8.325	wer	again	-7.354
markt	market	8.299	zak	business	-7.314
innov	innovative	8.206	dienst	service	-7.238
eenvoud	simplicity	8.172	sale	sale	-6.845
rendement	efficiency	8.159	werkplat	workplace	-6.824
search	search	8.126	opdrachtgever	client	-6.702

Both Dutch and English words are included. To improve readability, each word is translated into English (if needed) and shown as its most occurring non-stemmed version.

4 Detecting Innovation

4.1 Innovative companies without a website

The approach described above focuses on companies with websites and implicitly assumes that nearly all innovative companies have a website. This does not have to be the case and begs answering the question “How many companies without a website are innovative?” Let’s start answering this question by looking at the data obtained from the CIS survey. Here, we did our utmost best to find the accompanying website for our sample of 3,340 innovative and 3,002 non-innovative companies. For two of the innovative companies no website could be found, suggesting that 0.06% of the innovative companies with 10 or more working persons has no website. For the non-innovative companies in the sample, this was much higher. Here, no website was found for 126 companies, indicating that 4.2% of the non-innovative companies had no website.

But what about the small innovative companies? To get an indication of the number of innovative companies without a website with less than 10 working persons we took a random sample of 1,000 small companies from a complete list of all companies in the city of Eindhoven. We selected this city because a high quality list of small companies and URLs was available, it is the fifth-largest city in the Netherlands, and a considerable number of technological companies are located there. For the companies in the sample we carefully checked the URLs manually. A total of 107 small companies remained for which no website could be found. The data available in the Business Register of Statistics Netherlands including information available at the Chamber of Commerce website were used to check the characteristics of these companies, such as the description of its activities, so any that were certainly not innovative could be removed. As a result of this exercise 13 small companies without a website remained that could potentially be innovative. These were contacted by phone. In a short structured interview, the owner of the company was asked: i) if the company was still active, ii) if it had a website and iii) what kind of products or services the company provided. In the end one company was identified as innovative; it produces tailor-made software. Combined with the CIS derived findings, this indicates that -in general- a maximum of 1 in a 1,000 innovative companies in the Netherlands (0.1%) may be missing because they have no website.

4.2 Large and small innovative companies

The new model was used to classify a huge dataset of Dutch company websites. Since the aim was to determine the total number of small and large innovative companies in the Netherlands as accurately as possible, an as complete as possible list of Dutch companies and their associated website was constructed. By aiming for completion, a census like approach was applied. We therefore combined the results of all projects in our office that included URL-finding initiatives (for details see Oostrom et al. 2016, Ortega and Heerschap 2019, Statistics Netherlands 2019d, and Ten Bosch 2018) and added this to our own list. This resulted in a set of 824,972 Dutch company and website combinations that had a valid number of working persons assigned. These websites were subsequently scraped, processed and classified. The findings for large and small companies are separately discussed below.

4.2.1 Large companies

Of the 40,957 large companies for which a website was found, 38,601 could be accessed and scraped (94.2%). After processing, 37,576 of the websites contained 10 or more words (91.7%) and were classified. A total of 17,783 innovative large companies were found (43.4%). The model was additionally used to predict the innovation probabilities for the large companies to reveal its distribution (Fig. 1). The mean was 0.467, the median was 0.405 and the lowest and highest predicted probabilities were 0.00001 and 0.99998, respectively. The U-shape distribution in Fig. 1 reveals that the model is able to predict well separated probabilities for a large number of companies.

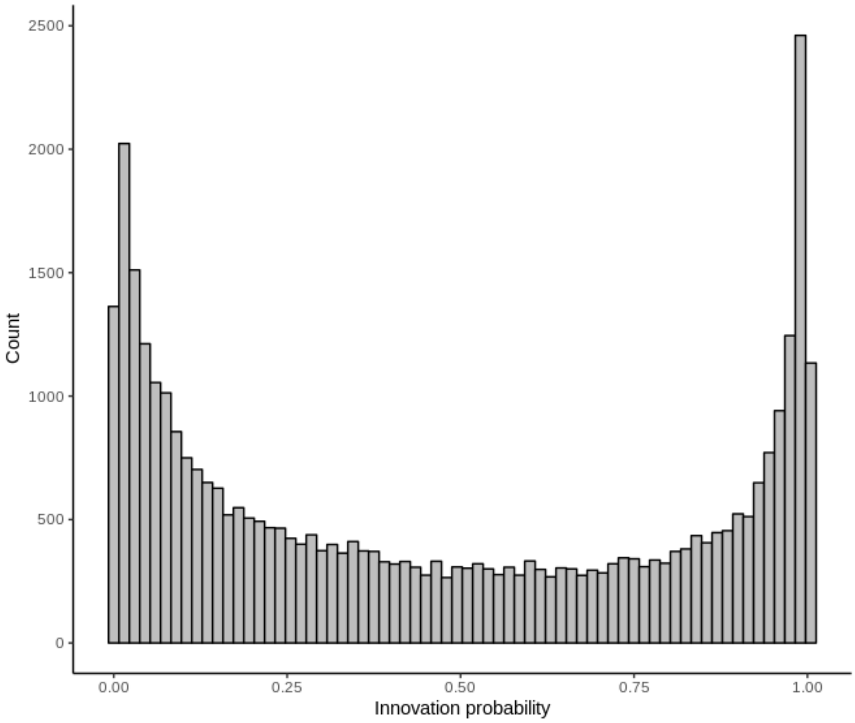

Figure 1. Innovation probability distribution for large companies. Histogram of predicted innovation probabilities for 37,576 company websites are shown

The estimate of 17,783 innovative large companies is not definitive as it needs to be corrected because: i) a model-based approach is used which may introduce a bias, ii) not all websites contained sufficient words to be classified, and iii) not all innovative companies have a website. In our opinion, a correction for the number of websites that could not be accessed and scraped is not required. Since these websites were repeatedly attempted to be scraped and failed, they were clearly not active anymore and hence do not belong to the target population.

Correction was performed as follows. First, the bias in the model-based estimation was corrected. This was needed as the precision (88%) and recall (84%) differed for the model applied (Meertens et al. 2019b). In the confusion matrix results precision and recall (Olsen and Delen 2008) are defined as:

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

tp, number of true positives; fp, number of false positives; fn, number of false negatives.

By comparing (1) and (2) it becomes clear that this difference indicates an imbalance in the number of false positives and false negatives in the classification results. The fact that the recall is lower demonstrates that the number of false negative is higher than the number of false positives. Since the number of innovative companies is the result of the combination of the amount of true positives and false positives provided by the model, the number of innovative companies is underestimated. The number of false positives and false negatives were therefore estimated by a Bayesian approach, one that imposes constraints on the model parameters (Meertens et al. 2019a), to produce a new, corrected, estimate of the number of innovative companies. This resulted in an increase to 18,745 innovative large companies. A second correction was needed as not all websites contained sufficient number of words to enable classification after processing; the limit for classification is 10 words or more. Inspection of the (limited number of) words in these websites did not reveal any subgroup that needed to be dealt with differently. We therefore assumed that the distribution for innovative and non-innovative companies for scraped websites with less than 10 words was similar to those of websites with 10 or more words. This resulted in an increase of the estimate to 19,257 innovative companies. As a last step, the estimate needs to be adjusted for the number of innovative companies without a website; this was estimated to be 0.1% (see above). This results in a final estimate of 19,276 for the number of large innovative companies in the Netherlands. The whole procedure -including model development- was repeated a 1,000 times revealing a 95% confidence interval of 190 for the final estimate. The percentile bootstrap confidence interval procedure described in Davison and Hinkley (1997) was used. Interestingly, the final estimate is very close to the most recently available official number published for the total number of large technological innovative companies in the Netherlands; which is 19,916, with a confidence interval of 680 as obtained by the CIS survey (Statistics Netherlands 2019e). This demonstrates that the text-oriented approach developed produced valid findings, supports the model stability correction method used and the bias correction steps applied, and also suggests that the census-based approach worked well for the large companies.

4.2.2 Small companies

A similar approach was applied to the websites of small companies, with the exception that a number of companies were removed prior to scraping. Of the 784,015 small companies with a website, 992 were found to be exclusively associated with a single large company according to the data in our Business register. This reduced the population of small companies to 783,023 (99.9%). Since we are only interested in companies that have at least a part-time worker assigned, those with zero employees were also excluded. This resulted in 699,332 companies remaining (89.2%). From this set, 547,237 websites could be accessed and scraped (69.8%). After processing, 497,738 of these websites contained 10 or more words (63.5%) which resulted in a total of 212,216 that were classified as innovative (27.1%). Compared to the large companies, it is clear that the process of scraping and processing of small company websites, even after excluding companies with zero employees, resulted in a relative high loss during each step. This suggests that fewer websites were active and that more websites contained relative few words. What is also clear is that the overall number of small innovative companies found is very high. However, before discussing this in more detail, the estimates of the number of innovative companies were first corrected for by the three-step procedure developed for large companies and repeated a 1,000 times to determine their confidence intervals. These results are shown in Table 3 for the various groups of small companies discerned. In this table, the number of websites scraped, the bias-corrected estimate of the number of innovative companies and their confidence intervals are shown for the part-time self-employed, self-employed and small companies with 2 to 9 working persons. A particular large number is observed for companies of self-employed, indicating that a lot of self-employed have a website that is classified as innovative. This is also the case for the websites of part-time self-employed people. When this number is compared to the number of websites scraped for these groups, it becomes clear that in both cases a relative high amount of websites are classified as innovative; around 47%. For companies with 2 and more working persons these values are considerably lower; values between 34-39% are observed. This suggests a difference in the behaviour of both groups. The probability distributions for the various groups of small companies were all U-shaped and highly resembled the one found for the large companies (shown in Fig. 1). The only exceptions were those of the self-employed and part-time self-employed which additionally revealed a small spike around 0.8.

Table 3. Classification findings for small companies over the ranges of working persons studied.

Range of working persons	Number of websites scraped	Estimated number of innovative companies	Confidence interval	Percentage innovation detected (%)	Description
0.1 - 0.9	65,522	31,973	450	48.8	part-time self-employed
1 - 1.9	387,936	180,979	554	46.7	self-employed
2 - 2.9	43,857	15,544	260	35.4	2 working persons
3 - 3.9	16,4	5,657	108	34.5	3 working persons
4 - 4.9	10,329	3,573	96	34.6	4 working persons
5 - 5.9	6,964	2,5	77	35.9	5 working persons
6 - 6.9	5,209	1,963	65	37.7	6 working persons
7 - 7.9	4,11	1,613	59	39.2	7 working persons
8 - 8.9	3,435	1,366	56	39.8	8 working persons
9 - 9.9	3,475	1,383	52	39.8	9 working persons

4.3 New statistical products

The number of innovative companies for the Netherlands with 2 or more working persons, after bias correction, was aggregated at zip-code 4-level and at the municipality level. This enabled the production of some interesting visualizations. It is important to realize that the visualizations show the number of innovative companies, and therefore do not indicate how many people are employed at these companies. For companies with multiple establishments, the location of the headquarter is used. Fig. 2 provides an overview of the estimated number of innovative companies at the municipality level in the Netherlands. From this figure it’s clear that the Dutch capital, Amsterdam, is the city with the largest number of innovative companies; viz. 4,438. The other 4 large Dutch cities, from high to low: Rotterdam, The Hague, Utrecht and Eindhoven, also have relative large numbers of innovative companies; viz. 1,844, 1,251, 1,392 and 1,075, respectively. For the rest of the country, the data indicates that in municipalities where universities and universities of applied sciences are located relative large numbers of innovative companies can be found (see also Statistics Netherlands 2019d).

Because detailed location information is available, maps of the estimated number of innovative companies with 2 or more working persons can be created at the zip-code 4-level, the neighbourhood level. These maps are particularly interesting for municipalities as they reveal the areas were innovative companies reside. Maps for the city of Amsterdam and Eindhoven are shown in Figures 3 and 4. In Amsterdam (Fig. 3), the area with the highest numbers of innovative companies is the one below the city centre. But the left side of the city centre and the

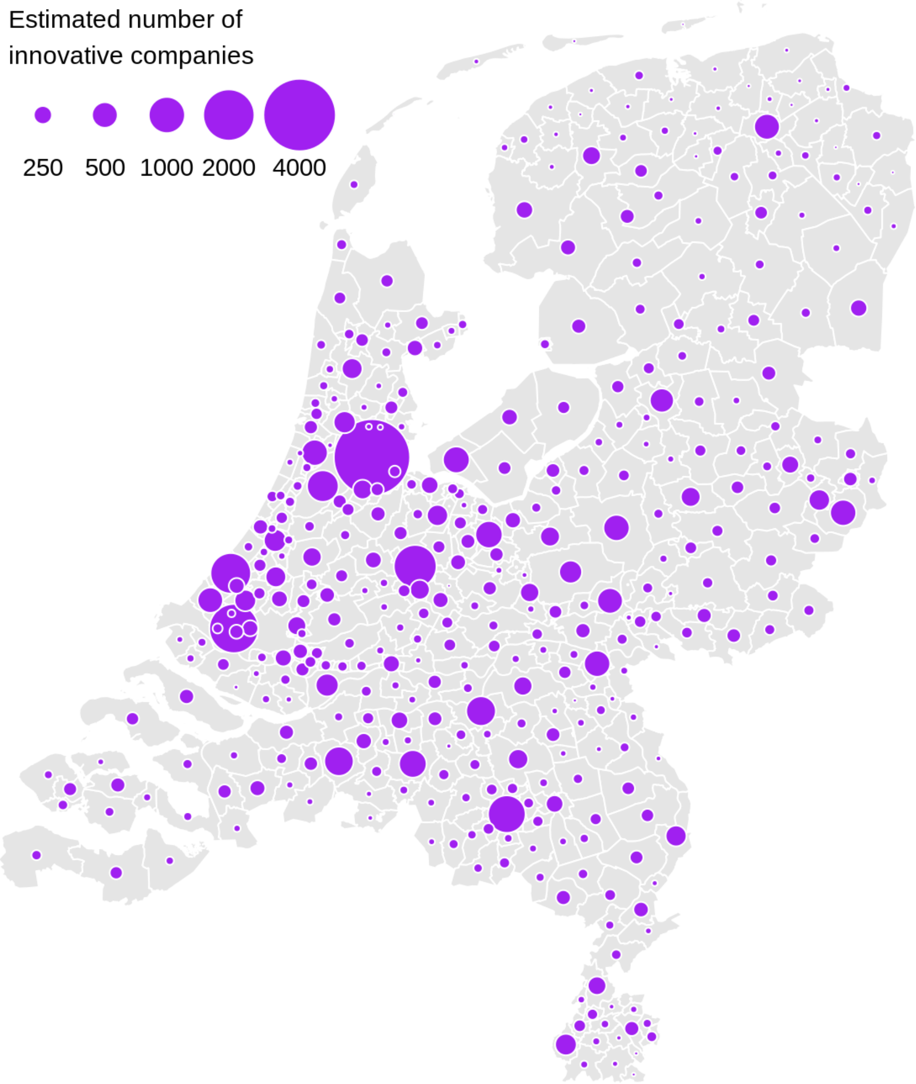

Figure 2. Estimated number of innovative companies with 2 or more working persons in the Netherlands at the municipality level. A total of 52,875 companies are shown.

adjacent areas on the left and right also contain many innovative companies. These are areas where a known number of Amsterdam startup initiatives reside. In the south-east area, the upper left part also contain considerable numbers of innovative companies. The plot for the city of Eindhoven (Fig. 4) reveals that the most innovative area is the city centre, where a large startup incubator is located. The industrial area's left of the city centre also contain a considerable number of innovative companies. In addition, the former industrial region Strijp-S, an area known for creative and innovative young companies, and the High Tech Campus where a number of cooperation's between industry and university are located, also contain considerable numbers of innovative companies.

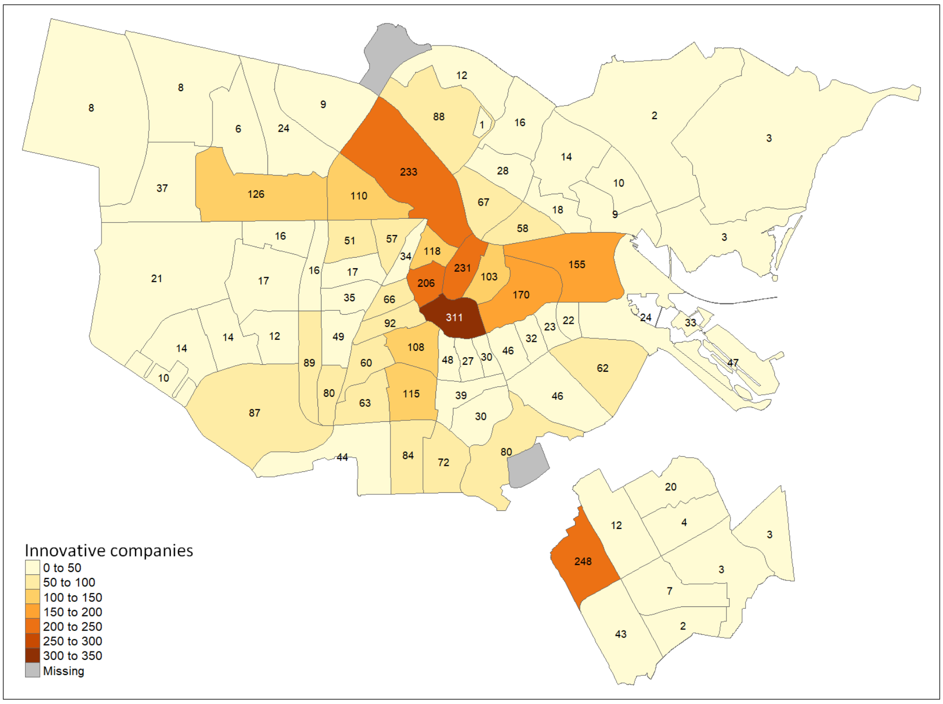

Figure 3. Estimated number of innovative companies of 2 and more working persons at the zip-code 4-level (neighbourhood) for Amsterdam. A total of 4,438 companies are shown.

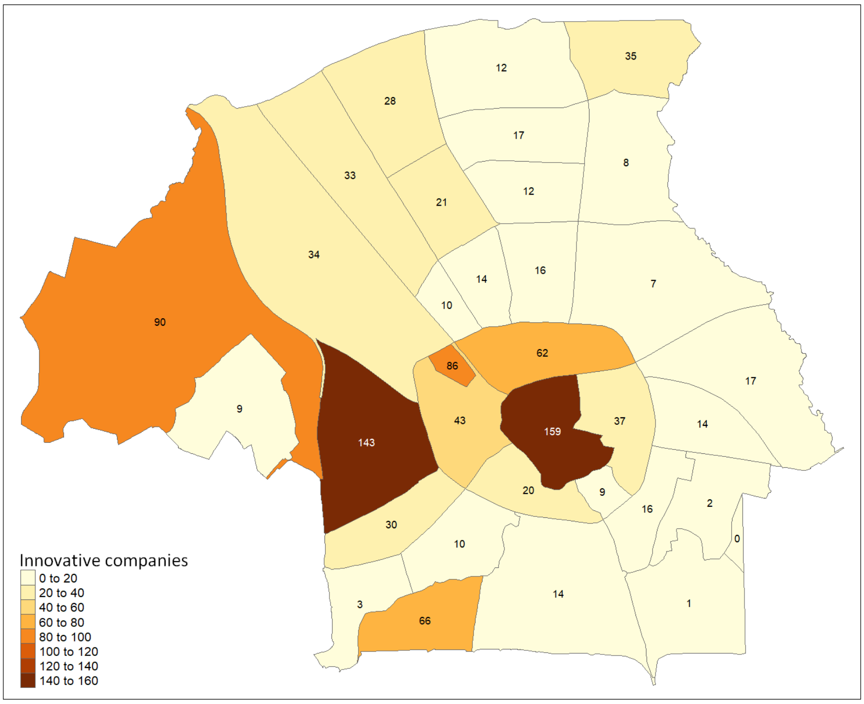

Figure 4. Estimated number of innovative companies of 2 and more working persons for the city of Eindhoven at the zip-code 4-level (neighbourhood) level. A total of 1,075 companies are shown.

5 Discussion

The results described above are promising and reveal that it is possible to provide a reasonably correct estimate of whether a Dutch company is technological innovative or not by using the text displayed on its website. For this purpose a logistic regression based model was selected from a series of other candidate models that contained 584 stemmed words. Inspection of the words revealed that many of them are positively or negatively associated with innovation (see Table 2). The method performed well on the webpages of both large (10 or more working persons) companies, e.g. the target population of the CIS survey, and those of small (less than 10 working persons) companies. The overall accuracy of the model was 88% with an F1-score of 86%. The latter is higher than the value reported by Kinne and Lenz (2019) for their model, and so our model provides some improvement. The U-shaped innovation probability distributions were found to be very similar for small and large companies indicating that the model was well able to discern between both classes in the majority of cases (see Fig. 1). Interestingly, this distribution suggests looking at the innovative character of a company on a more sliding scale. Our model is able to do this which could make the detection of innovation more realistic.

Model stability over time was found to be an issue; e.g. the performance of the model deteriorated. In retrospect, this is not unexpected as many of the words associated with innovation can logically be expected to change over time. Model stability was improved by adding large amounts of words extracted from the websites of newly classified companies and innovative startups, followed by retraining the model. This is an approach commonly used in the analysis of data streams that suffer from concept drift (Janardan 2017, Gama et al. 2014). In the context of text-based models this approach seems to work well because the new cases provide i) additional synonyms for words already included in the model and ii) add new upcoming words related to the phenomenon studied (Aggarwal 2016, Gentzkow et al. 2019). Because concept drift may occur, it is required to regularly check the validity of the model-based findings. We suggest checking the findings for a set of around 6,000 websites, preferably once every 6 months. If the findings start to deviate, the model needs to be updated with new cases, retrained and the results verified. This needs to be done until the results for the new CIS survey become available.

While the results are promising, implementation in the area of official statistics may also require more research on the explainability of the results. Table 2 presents the words with the strongest positive and negative evidential strength. While some of them can easily be given a (post-hoc) explanation, this does not hold for all the words. It is also recognized that measuring a broad concept as ”innovation” is quite challenging; definitions and interpretations may vary between business sector areas and the concept may change over time. The use of expert knowledge and economic theory on this concept may shed light on this. A model that is the result of the combination of a data driven approach and expert knowledge and theory may also be less prone to concept drift.

To illustrate the possibilities of the approach developed, the total numbers of large and small innovative companies in the Netherlands were estimated with our method. For this a census oriented approach was used for which an as complete as possible list of Dutch companies and their associated URL was the starting point. The model based estimate of the number of large innovative companies was, after bias correction, similar to those reported by the CIS survey. This suggests that the way large companies profile themselves -in the Netherlands- on their website is comparable to the way such companies report their innovation activities in the CIS survey. The findings for the number of small innovative companies were new and could therefore not be compared with other data. Here, it was observed that websites of (part-time) self-employed were more often classified as innovative, compared to the websites of companies with 2 to 9 working persons. This suggests that self-employed people are more inclined to use words associated with innovation on their website and, hence, may be deliberately promoting themselves in a more innovative way. Because of this finding and the large amounts of innovative self-employed companies found, we suggest that the number of innovative companies of 2 to 9 working persons is -according to the current state of knowledge- the best indicator for the number of small innovative companies in the Netherlands. The cumulative number of those companies was found to be 33,599 (from Table 3). Because location information is available, detailed maps of the Netherlands can be created displaying the distribution of innovative companies over the country and within neighbourhoods of cities.

Let it be clear that the newly developed method only focuses on estimating the number of technological innovative companies and not on the other variables collected by the CIS survey, such as the amount of money invested in innovation. As such, each approach has its own pros and cons. Because the CIS survey is a harmonized European survey conducted in all European member states, it would be very interesting to check if the approach described in this paper is also effective in other European member states. The work of Kinne and Lenz (2019) demonstrates that this is certainly the case for Germany. Unfortunately no findings on the stability of the model were reported in that study. A study performed in cooperation with our colleagues from the Swedish statistical office did not find a strong relation between the words on a website and the innovative character of the companies included in the CIS survey for that country[2)]. At the moment of writing, a similar study is being performed for Flanders; i.e. the Dutch-speaking northern part of Belgium. Future work in the Netherlands will focus on model interpretability, robustness and stability over time, on the way companies, startups and self-employed people profile themselves on their website, on detecting other forms of innovation and on the location and economic activity of innovative companies for various Dutch cities.

Acknowledgements

Part of this research was supported by the Dutch Ministry of Economic Affairs and Climate Policy. We thank the many colleagues at Statistics Netherlands and students that contributed to the work on which this paper is based, provided feedback and especially thank those that assisted us in our search for company URLs. We also thank the employees of the Dutch company InnovatieSpotter for their valuable insights and expertise provided. The reviewers are gratefully acknowledged for their insightful comments and suggestions that have considerably improved the paper.

[2)] An accuracy of 71% was found (Jansen and Daas, internal report)

References

Aggarwal, CC. (2016). Mining Text Data. In: Aggarwal, CC, editor. *Data Mining: the Textbook*, New York, Springer, pp. 429-455.

Allen C, Hospedales T. (2019). Analogies Explained: Towards Understanding Word Embeddings. Proceedings of the 36th International Conference on Machine Learning, June 10-15, Long Beach, USA. Available at: https://arxiv.org/pdf/1901.09813.pdf.

Atefeh F, Khreich W. (2015). A survey of techniques for event detection in twitter. *Comput. Intell.* 31(1), 132-164. doi:10.1111/coin.12017.

Beaudry C, Heroux-Vaillancourt M, Rietsch C. (2016). Validation of a web mining technique to measure innovation in the Canadian nanotechnology-related community. International Conference on Advanced Research Methods and Analytics (CARMA) 2016, Valencia, Spain, pp. 100-115, doi:10.4995/CARMA2016.2016.3140.

Breiman, L. (2001). Statistical Modelling: The Two Cultures. *Stat. Sci.* 16(3), 199-231. doi:10 .1214/ss/1009213726.

Daas PJH, Jansen J. (2020). Model degradation in web derived text-based models. International Conference on Advanced Research Methods and Analytics (CARMA) 2020, Valencia, Spain. pp 77-84. doi.org/10.4995/CARMA2020.2020.11560

Daas PJH, Puts MJH, Buelens B, van den Hurk, PAM. (2015). Big Data and Official Statistics. *J. Off. Stat.* 31(2), 249-262. doi:10.1515/jos-2015-0016.

Davison AC, Hinkley DV. (1997). *Bootstrap Methods and Their Application*. Cambridge, Cambridge University Press. Chapter 5, Confidence Intervals, pp. 191-255.

Eurostat (2019). Website of the Community Innovation Survey. October 16, 2019. Available at https://ec.europa.eu/eurostat/web/microdata/community-innovation-survey.

Florescu D, Karlberg M, Reis F, Rey Del Castillo P, Skaliotis M., Wirthmann A. (2014). Will 'big data' transform official statistics? Quality in Official Statistics Conference, June 2-5, Vienna, Austria. Available at: http://www.q2014.at/fileadmin/user_upload/ESTAT-Q2014-BigDataOS-v1a.pdf.

Gama J, Zliobaite I, Bifet A, Pechenizkiy M, Bouchachia A. (2014). A survey on concept drift adaptation. *ACM Comput. Surv.* 46(4), 1-37. doi:10.1145/2523813.

Gentzkow M, Kelly B, Taddy M. (2019). Text as Data. *J. Econ. Lit.* 57(3), 535-574. doi:10.1257 /jel.20181020.

Gökk A, Waterworth A, Shapira P. (2015). Use of web mining in studying innovation. *Scientometrics* 102(1), 653-671. doi:10.1007/s11192-014-1434-0.

Hand D. (2019). What is the purpose of statistical modelling? *Harvard Data Science Review*. doi:10.116299608f92.4a85af74.

Höchtl J, Parycek P, Schöllhammer R. (2015). Big Data in the Policy Cycle: Policy Decision Making in the Digital Era. *J. Org. Comp. Elec. Com.* 26(1-2), 147-169. doi:10.1080/10919392.2015 .1125187.

Janardan SM. (2017). Concept drift in Streaming Data Classification: Algorithms, Platforms and Issues. Procedia. *Comput. Sci.* 122, 804-811. doi:10.1016/j.procs.2017.11.440.

Kim S-M, Hovy E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. Proceedings of the Workshop on Sentiment and Subjectivity in Text, Association for Computational Linguistics, pp. 1–8. Available at: https://www.isi.edu/natural-language/people/hovy/papers/06ACL-WS-opin-topic-holder.pdf.

Kinne J, Axenbeck J. (2018). Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany. ZEW Discussion paper no 18-033, Mannheim, Germany. Available at: http://ftp.zew.de/pub/zew-docs/dp/dp18033.pdf

Kinne J, Lenz D. (2019). Predicting Innovative Firms using Web Mining and Deep Learning. ZEW Discussion paper no 19-001, Mannheim, Germany. doi:10.13140/RG.2.2.22526.84809.

Kitchin R. (2015). The opportunities, challenges and risks of big data for official statistics. *Statistical Journal of the IAOS* 31(3), 471-481. doi:10.3233/SJI-150906.

Li Y, Yang, T. (2018). Word Embedding for Understanding Natural Language: A Survey. In: Srinivasan S. editor. *Guide to Big Data Applications*. New York, Springer, pp. 83-104.

Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G. (2018). Learning under Concept Drift: A Review. *IEEE T. Knowl. Data En.* 31(12), 2346-2363. doi:10.1109/TKDE.2018.2876857.

Meertens QA, Diks CGH, van den Herik HJ, Takes FW. (2019a). A Bayesian Approach for Accurate Classification-Based Aggregates. Proceedings of the 2019 SIAM International Conference on Data Mining, May 2-4, Calgary, Canada, pp. 306-314. doi:10.1137/1.9781611975673.35.

Meertens QA, van Delden A, Scholtes S, Takes FW. (2019b). Bias Correction for Predicting Election Outcomes with Social Media Data. Fifth International Conference on Computational Social Science IC2S2, Amsterdam, The Netherlands. Available at: http://tiny.cc/hgqcdz.

McGee S. (2002). Simplifying Likelihood Ratios. *J. Gen. Intern. Med.* 17(8), 647–650. doi:10.1046 /j.1525-1497.2002.10750.x.

Mirończuk M, Protasiewicz J. (2016). A diversified classification committee for recognition of innovative internet domains. In: Kozielski S, Mrozek D, Kasprowski P, Małysiak-Mrozek B, Kostrzewa D, editors. *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery* 613. Springer, Switzerland, pp. 368–383.

Olson DL, Delen D. (2008). Performance Evaluation for Predictive Modelling. In: Olson DL, Delen D, editors. *Advanced Data Mining Techniques*. New York, Springer, pp.137-147.

Oostrom LAN, Walker AN, Staats B, Slootbeek-Van Laar M, Ortega-Azurduy S, Rooijakkers, B. (2016). Measuring the internet economy in The Netherlands: a big data analysis. Discussion paper 2016-14, Statistics Netherlands, The Hague/Heerlen, The Netherlands. Available at: https://www.cbs.nl/-/media/_pdf/2016/40/measuring-the-internet-economy.pdf.

Ortega S, Heerschap N. (2019). Using social media: Exploratory research on the possibilities of using social media for business statistics (in Dutch). Research paper 2019-16, Statistics Netherlands, The Hague/Heerlen, The Netherlands. Available at: https://www.cbs.nl/-/media/_pdf/2019/16/gebruik-van-sociale-media.pdf.

Pedregosa F, Vaoquaux G, Gramfort A, Michel V, Thirion B, Grisel O. et al. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825-2830. Available at: https://arxiv.org/pdf/1201.0490.pdf

Robehmed N. (2013). What Is A Startup? Forbes, Dec. 6. Available at: https://www.forbes.com/sites/natalierobehmed/2013/12/16/what-is-a-startup/.

Stateva G, ten Bosch O, Maślankowski J, Barcaroli G, Scannapieco M, Summa D. et al. (2017). Methodological and IT issues and Solutions. Deliverable 2.2 of Workpackage 2 of the ESSnet on Big Data. Appendix 7.1. Available at: https://webgate.ec.europa.eu/fpfis/mwikis/essnet-bigdata/images/6/66/ WP2_Deliverable_2.2_2017_07_31.pdf.

Statistics Netherlands (2019a). Definition of the concept Innovation. Sept. 20, 2019. Available at: https://www.cbs.nl/en-gb/our-services/methods/definitions?tab=i#id=innovation.

Statistics Netherlands (2019b). Results of the 2016 Community Innovation Survey (in Dutch). June 2, 2019. Available at: https://www.cbs.nl/-/media/cbs%20op%20maat/microdata bestanden/documents/2019/15/cis%202016.pdf.

Statistics Netherlands (2019c). Definition of the concept Technological innovation (in Dutch). Sept. 20, 2019. Available at: https://www.cbs.nl/nl-nl/onze-diensten/methoden/begrippen?tab=i#id=innovatie.

Statistics Netherlands (2019d). Innovation in small businesses, CBDS beta product. June 2, 2019. Available at: https://www.cbs.nl/en-gb/our-services/innovation/project/innovation-in-small-businesses.

Statistics Netherlands (2019e). Enterprises with innovation, key figures. July 19, 2019. Available at: https://opendata.cbs.nl/statline//CBS/en/dataset/80066eng/table?ts= 1569225569076.

Ten Bosch O, Windmeijer D, van Delden A, van den Heuvel G. (2018). Web scraping meets survey design: combining forces. BigSurv18 conference, Oct. 25-27, Barcelona, Spain. Available at: https:// www.bigsurv18.org/conf18/uploads/73/61/20180820_BigSurv_WebscrapingMeets SurveyDesign.pdf.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
`www.cbs.nl`

Prepress
Statistics Netherlands

Design
Edenspiekermann

Enquiries
Telephone: +31 88 570 70 70
Via contact form: `www.cbs.nl/infoservice`