I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15941

On the Validity of Using Webpage Texts to Identify the Target Population of a Survey: An Application to Detect Online Platforms

Piet Daas
Wolter Hassink
Bart Klijs

FEBRUARY 2023

I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15941

On the Validity of Using Webpage Texts to Identify the Target Population of a Survey: An Application to Detect Online Platforms

Piet Daas
Eindhoven University of Technology

Wolter Hassink
Utrecht University and IZA

Bart Klijs
Statistics Netherlands

FEBRUARY 2023

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

On the Validity of Using Webpage Texts to Identify the Target Population of a Survey: An Application to Detect Online Platforms[*]

A statistical classification model was developed to identify online platform organizations based on the texts on their website. The model was subsequently used to identify all (potential) platform organizations with a website included in the Dutch Business Register. The empirical outcomes of the statistical model were plausible in terms of the words and the bimodal distribution of fitted probabilities, but the results indicated an overestimation of the number of platform organizations. Next, the external validity of the outcomes was investigated through a survey held under the organizations that were identified as a platform organization by the statistical classification model. The response by the organizations to the survey confirmed a substantial number of type-I errors. Furthermore, it revealed a positive association between the fitted probability of the text-based classification model and the organization's response to the survey question on being an online platform organization. The survey results indicated that the text-based classification model can be used to obtain a subpopulation of potential platform organizations from the entire population of businesses with a website.

JEL Classification: C81, C83, D20, D83, L20

Keywords: online platform organizations, external validation, type-I error, machine learning, web pages

Corresponding author:
Wolter Hassink
Utrecht University
Utrecht University School of Economics
Kriekenpitplein 21-22
3584 EC Utrecht
The Netherlands

E-mail: W.H.J.Hassink@uu.nl

[*] The authors like to thank Kelby Çakim for her excellent assistance during the beginning of the project and Tim de Jong for his assistance in the hyperparameter optimization part of the study. Comments by Joep Burger, Javier Garcia Bernardo, Joram Vuik, and Ger Snijkers are gratefully acknowledged. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. Part of this paper is based on the Dutch Online Platform survey carried out on behalf of the Dutch Ministry of Economic Affairs and Climate Policy.

1. Introduction

Obtaining reliable information from a small or rare subpopulation is a challenging topic for survey researchers (Tourangeau et al. 2014; Snijkers et al. 2023), especially in an era where response rates continue to decline (Luiten et al. 2022; Wu et al. 2022). Approaches commonly used to find rare or so-called hard-to-identify groups are a screening survey, network sampling, area sampling, or a combination (Snijkers et al. 2013). Sometimes, lists of particular types of units are obtained from commercial organizations or they are constructed from administrative data sources (United Nations 2020). Unfortunately, these approaches do not always provide a good overview of the population of interest, especially when the topic of the study is new (Tourangeau et al. 2014; United Nations 2020, chap 8). However, the increasing availability of new data sources, so-called Big Data (Daas et al. 2015), may offer a solution to this problem. For example, such sources could be used to identify the relevant subpopulation, i.e. the target population of the survey, as completely as possible without contacting them. More specifically, it can be applied to identify businesses with an online platform (defined below).

The surge of internet technology in recent decades has enabled the rapid development of online platform organizations, and it has strongly altered the functioning of society. As a virtual digital meeting point, the intermediary platforms bring together persons and organizations, via which goods, services, or information can be exchanged. The Organisation for Economic Co-operation and Development (OECD 2019) defines online (digital) platforms as "a digital service that facilitates interactions between two or more distinct but interdependent sets of users (whether firms or individuals) who interact through the service via the Internet." Digital labour platforms can, for instance, be applied to a geographically dispersed crowd, e.g. "crowdwork", and by apps (Berg et al. 2018; Howcroft and Bergvall-Kåreborn 2019). Furthermore, online platforms have been applied to shopping (Ducci 2020) as well as to the sharing economy (Sutherland and Jarrahi 2018). A substantial number of online platform organizations – such as Airbnb, Greenwheels, and Uber – are profit-driven, and it may have implications for competition in two-sided markets (Ducci 2020; Cui et al. 2020; Rochet and Tirole 2003).

In recent years, national statistical institutes (NSIs) were lagging behind the phenomenon of the emergence and rapid growth of online platforms. Reliable statistics on the key components and dimensions of the online platform economy are still lacking (United Nations 2019). To fill this gap, in the past years' NSIs have debated a framework for measuring elements of the online economy for the Gross Domestic Product and the national accounts (OECD 2020). However, many empirical studies that investigate the (size of the) platform economy are solely based on surveys asking businesses or individuals about their use of online platforms (De Groen et al. 2017). Directly targeting online platform organizations instead of their users has proven to be more difficult (Heerschap et al. 2021; Klijs 2021). The main reason is that the identification of an online platform organization is far from straightforward. Online platform organizations cannot adequately be identified from Business Registers of NSIs, since the business classification system (NACE) used, classifies businesses according to their main economic activity; the system has no separate category for online platforms (NACE 2022). This means that these types of businesses cannot easily be approached with questionnaires assessing their economic activity. Also, alternative approaches, such as a generic list or a register of online platform organizations, are currently not available.

The inability to obtain a population of online platforms has hindered the direct research of those organizations using questionnaires or administrative data sources. However, because all online platform businesses have a website, it is of interest to investigate if the texts on those websites could be used to accurately identify them. In that way, a list of potential online platform organizations active in a country could be obtained. Text mining techniques could be used to do this (see Becue et al. 2004). In this paper, we describe the results of our study which aims to identify the online platform organizations in the Netherlands. Our empirical analysis is based on information obtained from the websites of about 600 thousand Dutch organizations; these are, in principle, all websites that have been assigned to the businesses in the Business Register of Statistics Netherlands (Daas and van der Doef 2020). For all these businesses, we have collected the textual content of the pages on their website. During the text mining analysis, we focus on combinations of words that tend to be associated with online platforms. The organizations will be ranked with respect to the likelihood of being an online platform. We demonstrate that the selection of platform organizations obtained is confirmed by an ex-post statistical analysis. The findings are subsequently validated using the information from the Dutch Online Platform survey conducted among the organizations identified as online platforms. To the best of our knowledge, only a limited number of empirical studies have assessed the validity of web-based text mining results through ex-post survey information (García Lozano et al. 2020).

Our study has two major implications. First, it demonstrates that text-based classification is a valid way to obtain a subpopulation strongly enriched with the target population of interest. Second, it demonstrates the advantages of combining text mining techniques and survey data for the study of the online economy.

The setup of this paper is as follows. Section 2 describes the general methods used and introduces text mining as a classification method. In Section 3, this method is applied to the texts extracted from the websites of Dutch organizations and the external validity of the text-based classification results is examined using survey information. Finally, in Section 4 the findings are discussed.

2. Material and Methods

2.1 Data collection and text processing

All scripts used were written in Python (v3.7). The Business Register of Statistics Netherlands (Ritzen 2007) is used to provide an overview of all businesses in the Netherlands. To this register, at the most detailed level possible, the corresponding websites are linked. The linking procedure, amongst others, compares the Chamber of Commerce number and address displayed on the website with those in the Business Register; for more details see Oostrom et al. (2016) and Daas and van der Doef (2020). To a total of 960,588 organizations, websites were assigned.

Web pages were scraped with the urllib.request function in Python. Each page was parsed with the Beautiful Soup library (v4.7.1) after which it was stored on the local machine. Pages that could not be scraped during the first attempt were visited at least four times – at later points in time – to deal with temporarily unavailable websites. Scraping started at the main page of the website, followed by all pages referred to that were located on the same website, up to a maximum of 1,000. Collecting all data took 3.5 weeks and resulted in a total of nearly 1 Terabyte of data. The locally stored files were processed in several steps. First, the script and style sections were removed followed by extracting the text inside the remaining Hyper Text Markup Language tags. Next, the language of the extracted text was determined with the langdetect (v1.0.7) library. Since the majority of the pages were either written in Dutch or English only those languages were discerned; e.g. any non-Dutch text was classified as English. Subsequently, the text was converted to lowercase and all numbers and punctuation marks were removed. This was followed by the removal of the language-dependent stop words, e.g. words that contain little information and occur often; such as articles. The Dutch and English stop word lists in the NLTK-library (v 3.4.1) were used for this. These are all standard text-processing steps (Aggerwal 2016). Next, optionally, the remaining words could be stemmed, e.g. reducing words to their root form. For this, the SnowbalStemmer library (v1.2.1) was used. Stemming has the advantage that it considerably reduces the number of variants of a word; e.g. the words ‘helpful’, ‘helpfully’, and ‘helping’ are all converted to ‘help’. Subsequently, words up to 2 character lengths could be removed. For model development, either only the text on the main webpage or the texts extracted from all pages collected on the website were used. The texts were combined into a single document in which the words were separated by a single space. Websites for which 10 or fewer words remained after processing, which is particularly relevant when only the text of the main page was studied, were excluded for further analysis as this has been demonstrated to hardly provide any relevant information (Daas and van der Doef 2020).

To enable model development, the well-known representation of the text extracted in the form of frequency-annotated bag-of-words was used (Aggerwal 2016). This started by creating a document-term matrix in which the rows correspond to the business webpages and the columns to the unique words included in all the text extracted. The natural logarithm of the term frequency-inverse document frequency (log(*tf-idf*) + 1) for each word was used as a feature value (Daas and van der Doef 2020). The *tf-idf* value indicates how important a word is in the texts as the term frequency increases proportionally to the number of times a word appears in it. The inverse document frequency offsets this number by the number of texts that contain the word. The latter adjusts for the fact that some words appear more frequently than others in website texts, of both platform and non-platform texts, which severely reduces the influence of often occurring, non-discriminating, words. In addition, the language of the text was added as a binary feature to the matrix, for which *English* = 1 and *Dutch* = 0. Word Embeddings, a technique focused on word co-occurrences that is often used to improve text classifications by encoding semantic and syntactic information (Allen and Hospedales 2019), were included by applying the gensim library (v3.4.0). Up to 300 vectors of either the word2vec skip-gram or Continuous Bag Of Words algorithms of the gensim library could be additionally added to the matrix. Machine Learning models were developed with the scikit-learn library (v0.21.2; Pedregosa et al. 2011).

2.2 Model development and classification

The overall process of text processing, that will be applied in the empirical analyses, consists of three stages. In the first stage, a data set with known examples of platform and non-platform websites is constructed by experts. All experts are employees of Statistics Netherlands with at least five years of experience in business statistics and have been involved in the study of online platform businesses for at least two years. To identify the platform and non-platform organizations, the experts review information on websites of a set of organizations. Based on the definition of an online platform (see the introduction) this leads to a set of

$$Platform = \{DPlatform_j = 1 | Text_j; j = 1, \dots., N\}$$

where the 0-1 variable $DPlatform$ has the value of one if the organization is characterized as an online platform according to the judgment of the experts. It is based on the multidimensional variable $Text$, which is composed of the combination of words included on the website of the organization. The elements in this data set are referred to by the subscript j. In total there are N platforms in the set $Platform$. The experts also assemble a second set of N non-platforms, for which the 0-1 variable $DPlatform$ has the value of zero

$$Non\_Platform = \{DPlatform_j = 0 | Text_j; j = 1, \dots., N\}$$

In the combined dataset, the sets $Platform$ and $Non\_Platform$ have an equal number of elements and do not overlap. The combined data set created by experts consists of

$$Combined\ data\ = \{\ Platform, Non\_Platform\} \qquad (1)$$

In the second stage of text processing, a supervised generative model-based approach (Genzkow et al. 2019) is applied to a large random sample of the combined data set; this is 80% in our case. This is referred to as the training data. The model-based approach reduces the multidimensional variable $Text$ to a lower dimensional variable Z (see next paragraph), such that the new variable Z discriminates the elements of platform versus non-platform organizations. There is a whole range of machine learning algorithms available that can be applied to computational efficiently perform this task (Pedregosa et al. 2011).

$$Z_j \in Text_j \mid j \in \{Platform, Non\_Platform\}$$

Z is a subset of the variable $Text$ for the organizations in the training set. The vector Z consists of variables – words – that characterize the elements that belong to the set $Platform$ versus those that are part of the set $Non\_Platform$. In addition, there usually is a vector of estimated weights $\hat{\theta}$, obtained through a machine-learning algorithm, which are used to predict the dichotomy $DPlatform_i = 1$ versus $DPlatform_i = 0$; see Gentzkow et al. (2019) for more details. It leads to the probability that an organization is an online platform, conditional on Z_j and $\hat{\theta}$

$$P_j = Prob\big(DPlatform_j = 1 \big| Z_j, \hat{\theta}\big) \qquad j \in \{Platform, Non\_Platform\} \qquad (2)$$

As is usual in machine learning, an independent sample, referred to as a test set (also known as a holdout set), is used to check the performance of (2). Here, the test set is the 20% part of the combined data that remained after selecting the training data.

In the third stage, for the entire population of organizations, the statistical model of equation (2) is used to predict $DPlatform_i = 1$ by using Z_j and $\hat{\theta}$.

$$\hat{P}_i = Prob\big(DPlatform_i = 1 \big| Z_i, \hat{\theta}\big) \;\; i \in \{population\ of\ organizations\} \qquad (3)$$

The elements of the population are referred to by the subscript i. Depending on the machine learning algorithm used, either a binary value or a value in the range 0-1 is produced. In the latter case, all organizations i for which the estimated probability $\hat{P}_i$ is above a specifically defined threshold are classified as online platform organizations. With q as the threshold, the set of organizations becomes

$$\{i | \hat{P}_i \geq q, i \in population\ of\ organizations\}$$

Usually, a threshold value of 0.5 is used for this purpose, but this is not always the case; for instance when dealing with highly imbalanced data, such as data with only a limited number of positive cases (Kuhn and Johnson 2013, chap. 16).

Following the three stages described above, there are two major outcomes regarding the identification of online platform organizations. First, there is the set of platform organizations, included in the test set, that have been classified by the statistical model (equation (2)). This set solely consists of organizations for which the correct classification is known as these have also been determined by experts. This result is used to determine the performance of the model developed, which, in the case of accuracy, refers to the correct identification of platform and non-platform businesses. Here, one wants to obtain a model with the highest accuracy possible. Second, there is the outcome of the classification of the unseen (new) organizations in the population. Some of them are identified as online platform organizations (equation (3)). How well the classification model performs on the unobserved organizations, e.g. the second outcome, affects the findings tremendously. Especially for the latter results, there is usually no information on the type-I and type-II errors. The type-I errors, i.e. the false positives, consist of the businesses that are non-platform organizations that are identified by the model as online platforms. The type-II errors, i.e. the false negatives, are organizations not classified as a platform by the model but are actually online platforms. Given that the number of platform organizations is expected to be relatively scarce (Heerschap et al. 2021), it will be hard to assess the type-II errors through a survey.

It would be possible, however, to estimate the size of the type-I error. We sent all organizations with a value above the threshold the Dutch Online Platform questionnaire. Businesses themselves will disclose in the survey whether they can be categorized as an online platform organization, according to the OECD definition additionally provided. We used the first two questions in this survey for this purpose. The questions are included in the Appendix. Thus there are two outcomes of the external validation. Either a business confirms the outcome of the text-based classification – “the model is right” – or it reports that it is not a platform organization, which leads to a type-I error. More formally, the external validation is based on the external measure $\hat{P}_i$ of equation (3) which is confronted with externally collected information on the latent 0-1 variable $Platform^*$ for the organizations above the threshold. This variable gets a value of 1 if the business confirms that it can be characterized as an online platform. Here, only the information is used of businesses that: i) indicate they are a platform and ii) are not included in the combined data set created by experts.

Overall, the external validation of the businesses above the threshold provides two important pieces of information that will be examined in the empirical analysis. First, it gives an estimate of the fraction of type-I errors in the estimated platform organizations obtained after applying the model. Second, it leads to an estimate of the statistical association between $\widehat{P}$ of equation (3) and the latent variable $Platform^{*}$. There will be an indication that the text-based classification leads to satisfactory estimation results if there is a positive association between $\widehat{P}$ and $Platform^{*}$.

3. Results

3.1 Analysis of text-based results

3.1.1 Step 1: Combined data set creation and data collection

In the first stage, we constructed a data set containing known examples of platform and non-platform organizations (equation (1)). The websites assigned to the organizations in the Business Register of Statistics Netherlands were an important starting point. Based on this register, the findings of some initial studies (Heerschap et al. 2021), and their expertise, three business statistics employees of Statistics Netherlands created a set of 590 online platform organizations and identified 303 non-platform organizations, with very similar characteristics, during this process. To the latter, a random sample of 287 non-platform organizations, from the websites linked to the Business Register, were additionally added. The websites in this sample were manually checked to assure they were active, were not already included, and were definitely of non-platform organizations. Let it be clear that the organizations and websites included in the combined data set were removed from the large Business Register linked dataset in subsequent analysis.

The combined data set of 1,180 websites was used as the training and test set to develop an online platform text-based classifier. All websites in the combined data set, e.g. all 1,180, and a total of 629,284 (66%) websites linked to the Business Register could be scraped. The websites that could not be scraped were found to be no longer active.

3.1.2 Step 2: Text processing and model development

In the second stage, the variable $Text$ is reduced to the variable Z of lower dimension, and the estimate $\hat{\theta}$ is obtained (equation (2)). This step starts with the texts extracted from the scraped webpage files of the businesses included in the combined data set. For the combined data set, the text in a total of 1,138 websites (96%), could be extracted and processed. These included 50% platform (positive) and 50% non-platform (negative) cases. A data set constructed in this way is very well suited to determine if a particular text-based classification method is able to differentiate between the texts in the positive and negative cases for the topic studied (Kuhn and Johnson 2013; Daas and van der Doef 2020). Of this dataset, an 80% random sample was used for model development; this is the so-called training set. The remaining 20% was used as the test set (see below). Model development required the creation of a document-term matrix in which the rows corresponded to the organizations webpages and the columns to all the unique words in the training set; see Materials and Methods for more details. Words that occurred less than 50 times in the training data were removed. The resulting document-term matrix had a dimension of 910 rows by 570 columns and 910 rows by 4,300 columns for those based on the text of the main page only and those based on the texts extracted from all pages collected on the website, respectively. Subsequently, a whole range of different machine learning methods was trained to discern platform from non-platform websites in the best possible way. The performance of those models was evaluated on the unseen 20% test set. Accuracy, e.g. the number of correctly classified cases of the total number of cases included, Precision, e.g. the number of positives correctly classified of the total number of cases classified as positives, and Recall, e.g. the number of positives correctly classified of the total number of positive cases included, were used as the most important evaluation metrics.

The metrics for various trained classification methods, such as Naive Bayes, Logistic Regression, Support Vector Machines, Regression Trees, and Neural Networks, were compared. During this comparison, the effect of various processing steps on the texts and the choice to use only the words on the main page or the words on all pages scraped for a website were compared. Hyperparameter tuning, via a Tree of Parzen Estimators (Bergstra et al. 2011) followed by 5-fold cross-validation, was used to assure the best possible outcome was obtained for each method. It was found that a trained Support Vector Machine (SVM) model with a linear kernel produced the best results when: i) the words on all pages collected from of a website were used, ii) the words were stemmed and iii) only words of 3 or more characters were included. Hyperparameter optimization revealed that the standard settings for this method already resulted in the best performance. For the SVM model, an accuracy of 82% (±2%) was obtained on the test set. The standard deviation was determined by repeating the entire procedure on resamples, with replacement, 1,000 times. The precision was 84% (±3%), and the recall was 79% (±4%). Even though this approach did not work perfectly, it produced the best statistical model to identify online platform websites of all options and combinations tested. Including Word Embeddings derived features did not improve the classification findings even after additional hyperparameter optimization. Applying deeply trained Bidirectional Encoder Representations from Transformers (BERT) or its Dutch version BERTje (Fialho et al. 2020) did also not produce better results. The creation of two language-specific models, one for Dutch and one for English websites, did also not improve the overall findings. The SVM model obtained provided a score of being an online platform website that was scaled to a probability via a five-fold cross-validation procedure (Platt 2000). That probability is, from here on, indicated as $PWebsite$. This is a value between 0 and 1 of which, generally, a value above a threshold value of 0.5 is considered a positive (platform) case.

The findings of the SVM model were additionally checked by studying the distribution of the probabilities on the test set. This revealed a, somewhat noisy, U-shape indicating that the two cases could be separated fairly well. In addition, the 10 words with the highest positive and highest negative coefficients used by the SVM model were inspected (see Table 1). The findings for the words with high positive coefficients indicate that the trained model picked up the intended classification topic. The words with high negative coefficients are indicative of a heterogeneous group of websites which is not an unexpected finding as there is a whole range of non-platform websites.

3.1.3 Step 3: Prediction

In the third stage, the statistical model is used to predict the online platforms in the population of organizations (equation (3)). Hence, the SVM model was applied to the texts extracted and processed from the huge set of web pages linked to the Business Register, while excluding those included in the combined data set. From the 629,284 websites linked a total of 10,964,998 pages could be scraped; indicating an average of slightly more than 17 pages collected per website. The web pages were processed according to the optimal procedure described above which resulted in 629,284 text files. Of those files, 594,574 (94.5%) contained 10 words or more. These files were classified with the SVM model developed and, for each case, the probability of being an online platform website was determined. This resulted in 41,811 (7%) websites being classified with a value above the threshold of 0.5. However, the histogram of the distribution of the probabilities for these websites indicated a strongly negatively skewed distribution (Figure 1). Because of this finding and the fact that the model was developed on an equal number of platform and non-platform websites, a situation that is expected to be far off from the platform-non-platform ratio of websites linked to the Business Register, made clear that the classification findings needed to be studied in more detail. This was done by drawing random samples of 50 websites in 9 probability ranges, each 0.1 wide, and manually inspecting the websites selected. This revealed that in these samples, online platform websites started to occur at probabilities values of 0.8 and higher. In the lower value ranges, none were detected in the samples drawn. From this, it is clear that the model obviously overestimates the number of platform websites in the Business Register when a value of 0.5 is used as the positive detection threshold. This is an important finding as it reveals that the model behaves differently on the Business Register data compared to the training and test data. This not only has interesting research applications, described in more detail in Puts and Daas (2020), but also suggests applying a higher threshold value for online platform identification. The fact that the lowest point in the probability distribution is somewhere located around 0.90 corroborates this observation (Figure 1). We found that the number of websites with a $\hat{P}$ value above 0.8 is 9,129 (1.5%). In addition, the probability distribution reveals a small peak around 0.99 (insert in Figure 1), which suggests the occurrence of a group with very high $\hat{P}$ results.

3.2. External validity

3.2.1 Type-I errors

To validate the results of the text-based classification, we examine the relative size of the type-I errors, which are the false positives of observing an online platform organization. There are two reasons for a type-I error. First, in the first stage of the estimation procedure, the expert opinion leads to a wrong assessment of some of the organizations that are considered platform organizations. Second, in the third stage of the estimation procedure, in which the empirical model is applied to the entire population, some of the organizations are mistakenly predicted as platform organizations.

To inspect the false positives, we make use of the organization's response to the two questions included in the Dutch Online Platform survey conducted by Statistics Netherlands; see Appendix. First, we report the selection procedure followed for the organizations given in Table 2. We start with the 9,129 websites identified as those of an online platform organization by the statistical model, all of which have a fitted probability $\hat{P}$ above a value of 0.8. Three subsequent selection steps were taken to construct the final survey population (see Table 2). First, websites with adult content were removed because these were not considered to belong to the target population. The URLs of these websites were checked for the occurrence of words or parts of words typical for adult content websites. Any sites that contained one (or more) of these words or word parts were removed. Next, the relation of the website with the units included in the Business Register was meticulously checked. For a considerable number of (legal) entities, multiple websites were found. For these businesses only one website was randomly selected to avoid sending multiple surveys to one business. In the final step, it was assessed whether information about the business unit to which the legal entities belonged could be retrieved from the business register. When this was not the case, the legal entity was excluded from the survey. In the end, a total of 4,385 organizations were approached to participate in the Dutch Online Platform survey. The response to the survey was 68% (2,997 organizations), which is relatively high for a business survey. Of all responding organizations 289 were excluded from further analysis since information on any of the essential variables required in the subsequent analysis procedures was missing in their response. This resulted in a selection of 2,708 organizations as a starting point for our further analysis.

First, we report on the percentage of false positives. Out of the 2,708 organizations, 2,064 organizations (76.2%) responded negatively to the question that their website is an online platform. This suggests a substantial percentage of (potential) false positives. So a total of 644 platform organizations were initially found.

Next, a comparison is made between the combined data (used for model development) and the classified websites linked to the Business Register; the latter is identified as the "other data" from here on. Of the 2,708 responding organizations, there are 77 organizations included in the combined data set. This means that there are 2,631 responding organizations in the other dataset; i.e. those retrieved by step 3 of the statistical procedure. The percentage of organizations that indicated they are *not* an online platform is 54.5% for the organizations in the combined data and 76.9% for the organizations in the other data; a total of 2,022. This brings us to the second result, that organizations in the combined data – obtained through expert assessment in step 1 – better reflect the platforms than the remaining ("other") organizations obtained by prediction in step 3.

Next, we focus on the relationship between the false positives and $\hat{P}$. To prevent any potential biases by the expert assessment (step 1), we restrict ourselves to the 2,631 responding organizations in the other dataset. Hence, only information on the 2,631 organizations is used in Tables 3 and 4. Table 3 gives an overview of the findings of the other data broken into several classes of $\hat{P}$. The average for $Platform$ seems to be positively related to the probability of the website being a platform, as indicated by the model, ranging from an average of 15.9% for the category 0.80 – 0.839 to 36.7% for the category 0.96 and higher. This brings us to the third result, namely that there seems to be a smaller fraction of false positives for organizations with a larger $\hat{P}$ value. This relationship will be explored in more detail in the next section. Table 3 also reports on some additional characteristics of the organizations included. A relatively large part of the organizations is comparatively small. About 25% of the organizations have one employee only; 14% of the organizations are large and have at least 50 employees.

Finally, we reassessed a sample of the false positive organizations in the other dataset by re-evaluating their website by the experts of Statistics Netherlands; a so-called second opinion. A random sample of 100 (false positive) organizations was drawn from the 2,022 organizations identified as false positive in the other data set. The sample was stratified by the $\hat{P}$ value categories used in Table 3. Table 4 reports the percentage positives for the false positives in each category according to the combined opinion of the experts. As a fourth result, it was found that – for some of the organizations – there is a disagreement between the opinion of the experts and the organization itself on whether their website is an online platform. According to the experts' reassessment, a weighted average of about 21% of the false positives in the other dataset is characterized as a positive (a platform); indicating that it is a "*false* false positive"; e.g. an actual platform organization. For organizations with a $\hat{P}$ value close to one, the disagreement is most strong. In the highest $\hat{P}$ value category, 65% of the false positives are online platforms according to the experts' reassessment (Table 4).

Based on these findings, we computed a lower and an upper bound for the percentage of false positives. There are two extreme situations. If we assume that the response to the survey gives a complete correct representation of whether or not the organizations surveyed are online platforms, the percentage of false positives is 76.9%. Alternatively, if we interpret the expert opinion as the gold standard, the percentage of false positives is only 60.8%; (76.9*(100-21)/100). Based on the information available, it is expected that the actual percentage of false positives is somewhere between 60.8 to 76.9%. This suggests that the population contains at least 1,121 to 1,455 online platform organizations after correcting for false positives. Since there is no information on the false negatives, we emphasize that this value is – very likely – an underestimation of the true number of platform organizations. The group of organizations not surveyed, i.e. those with a $\hat{P}$ below 0.8, could potentially still include some online platform organizations. However, because of the expected low numbers in this group, a costly and time-consuming survey is needed to reliably obtain information on the number of false negatives; i.e. the type-II errors among those organizations.

3.2.2 Statistical association

In this sub-section, we measure the statistical association between $\hat{P}$ of equation (3) – which was obtained through the empirical analysis of the text-based classification findings – and the latent variable $Platform^{*}$, obtained by the survey. The requirement to receive a questionnaire was that the $\hat{P}$ value of the text-based classification was at least 0.8 so the fitted probability value $\hat{P}$ has a range of 0.8 – 1.0. There is an indication that the text-based analysis leads to satisfactory estimation results if there is a positive association between $\hat{P}$ and $Platform^{*}$. We only used information on the selection of organizations that participated in the Dutch Online Platform survey conducted by Statistics Netherlands. Organizations that were part of the combined data set were excluded to ensure there was no contamination of the expert opinion from the first stage of the empirical analysis.

The starting point for the empirical specification is the response of the organizations to the survey questions about whether their website is an online platform. The dependent variable is a 0-1 indicator that gets the value of one if there is a positive response to the two questions shown in the Appendix. If that is not the case, the organization is characterized as a false positive. The variable is regressed on is the fitted probability value of the platform organization – reported in section 3.1.3 – as well as some control variables. The average of the dependent variable is 0.231 (Table 2), which makes it sufficiently large to specify the regression as a linear probability model

$$Y\_platform_i = \beta_0 + \beta_1 \hat{P}_i + \gamma' X_i + u_i \qquad i = 1, \ldots . N \tag{5}$$

where subscript i refers to the i-th organization; there are N organizations. $Y\_platform$ is a 0-1 indicator variable which is one, if the organization responds in the survey that it performs online platform activities. $\hat{P}$ is the fitted probability value that was obtained through the text-based classification; it may have a slight attenuation bias due to the uncertainty of $\hat{P}$. The vector X contains variables from the Business Register: firm size (4 categories), economic sector (3 categories), and legal status (2 categories). u is an idiosyncratic error term.

Although a valid probabilistic measure of $\hat{P}$ implies that the estimated β_1 has a positive sign, we can be more specific about its value. There is a strong indication of a valid empirical model for the platform probability since the change of $\hat{P}$ from 0.8 to 1.0 is associated with a change of $Y\_platform$ by 0.2 percentage points. In other words, there is an indication of a one-to-one relationship between $PWebsite$ and $Y\_platform$. This hypothesis will be tested below.

Next, we discuss the parameter estimates of equation (5) that are reported in Table 5. According to the various specifications, the marginal effect of the estimated parameter on the platform probability ranges from 0.011 to 0.013. It means that an increase of the probability $\hat{P}$ by 1 percentage point corresponds to an increase of the probability of being a platform by 1.1 to 1.3 percentage points. The estimated parameters presented in Table 6 confirm this outcome. The platform probability is distinguished into five categories. The estimates indicate that relative to the reference category of 0.8 – 0.834, the difference for the upper category is 19.5 percentage points. Thus, for a change of $\hat{P}$ from 0.8 to 1.0, there is also an increase of a positive survey response by about 20 percentage points. Remarkably, the categories 0.88 – 0.919 and 0.92 – 0.959 give similar parameter estimates (about a 10 percent difference relative to the reference category). For completion, using a Logit specification for equation (5) leads to the same estimated marginal effects.

4. Discussion

This paper describes an ex-post analysis of the validation of a subpopulation identified through the combination of web scraping and text-based classification. Such a statistical procedure is a useful tool in case it is hard to identify the target population of research when conventional sampling methods from a predefined population – such as stratified sampling or cluster sampling – cannot be applied. Here, it helps that the procedure can be easily applied to large amounts of data. Indeed, for the application that is described in this paper, it became clear that online platform activities are unevenly distributed economy-wide.

The estimates of the text-based classification procedure on online platform organizations lead to satisfactory outcomes after a number of additional selection steps. The most relevant and least relevant words identified by the statistical model are plausible. Furthermore, the distribution of fitted probabilities of being an online platform for the population of businesses gives a bimodal distribution. This work also confirmed the earlier observation that a model developed on 50% positive (platform) and 50% negative (non-platform) cases, behaves differently when applied to real-world ratios of these cases (Puts and Daas 2020). For online platform detection, which occurs much less than 50% in our "real-world" data, this results in an overestimation of the number of positive cases and makes the study, and consequently the reduction, of the number of false positive cases an important topic of the work described in this paper. Here, it becomes clear that the model is able to identify a rare occurring group of (potential) online platform organizations in a very large population. This group is surveyed.

Our results show that applying the classification model seriously reduces the initial population of around 600 thousand businesses to a set of a bit more than 9000 organizations; a more than 60-fold reduction. The latter data set is highly enriched in platform organizations and could be, after some additional checking and selection, almost completely surveyed. To statisticians that want to apply our approach, we strongly advise first making sure that the machine learning model is able to discern between the positive and negative cases of the topic studied in the best possible way. This requires not only a data set with typical positive and negative examples but also clear negative examples that, at first sight, resemble the positive cases reasonably well. In this way, one tries to make sure that only the relevant (and hence important) words are to be included in the model. When such a model is applied to the entire population, one subsequently needs to carefully check the external validity of the model by, for instance, manually inspecting websites.

We list a number of methodological learning points from our work. First, the quality of the data set used to build a classification model is important to get a bi-modal distribution of predicted outcomes. Second, the fraction of objects in the target population must be sufficiently large to enable adequate detection. Third, the external validation indicates a positive association between the fitted probability of the text-based model and a positive response to the survey question on being an online platform organization. Fourth, the fraction of false positives is large. Fifth, we observe a remarkable disagreement between the organization and the opinion of the experts on the question of whether the organization can be characterized as an online platform. The latter could, for instance, be caused by the fact that the first two questions in the Dutch Online Platform survey focus on determining this. Any organization that does not want to answer the remaining questions can simply end this by stating they are not a platform organization. Future research will focus on dealing with the last three topics as good as possible as this will greatly stimulate the application of Machine Learning in official statistics and the study of rare business subpopulations. In addition, the option to perform a more detailed study of the size of type-II errors will be investigated in more detail.

References

Aggarwal C.C. 2016. "Mining Text Data." In *Data Mining: the Textbook*, edited by C.C. Aggarwal: 429–455, New York: Springer.

Allen C. and T. Hospedales. 2019. "Analogies Explained: Towards Understanding Word Embeddings." In Proceedings of the 36th International Conference on Machine Learning: ICML, June 11-13, 2019, Long Beach: CA. Available at: http://proceedings.mlr.press/v97/allen19a/allen19a.pdf (accessed July 2022).

Becue, M., B. Fridlund, A. Fyhrlund, A. Prat, and B. Sundgren. 2004. "Text Mining in Official Statistics." In *Text Mining and its Applications: Results of the NEMIS Launch Conference*, edited by S. Sirmakessis: 189–204, Berlin: Springer.

Berg, J., M. Furrer, E. Harmon, U. Rani, and M. Silberman 2018. *Digital labour platforms and the future of work. Towards decent work in the online world*. Geneva: International Labour Organization. Available at: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_645337.pdf (accessed July, 2022).

Bergstra, J., R. Bardenet, Y. Bengio, and B. Kégl. 2011. "Algorithms for Hyper-Parameter Optimization." In *Advances in Neural Information Processing Systems 24*, edited by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, Curran Associates, Inc.: New York. Available at: https://proceedings.neurips.cc/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf (accessed July, 2022).

Cui, L., Y. Hou, Y. Liu, and L. Zhang. 2020. "Text mining to explore the influencing factors of sharing economy driven digital platforms to promote social and economic development." *Information Technology for Development* 27: 779–801. DOI: https://doi.org/10.1080/02681102.2020.1815636.

Daas, P.J.H., M.J. Puts, B. Buelens, and P.A.M. van den Hurk. 2015. "Big Data and Official Statistics." *Journal of Official Statistics* 31: 249–262. DOI: https://doi.org/10.1515/jos-2015-0016.

Daas, P.J.H. and S. van der Doef. 2020. "Detecting Innovative Companies via their Website." *Statistical Journal of IAOS* 36: 1239–1251. DOI: https://doi.org/10.3233/SJI-200627.

De Groen, W.P., Z. Kilhoffer, K. Lenaerts, and N. Salez. 2017. "The Impact of the Platform Economy on Job Creation." *Intereconomics* 52:345–351. DOI: https://doi.org/10.1007/s10272-017-0702-7.

Ducci, F. 2020. *Natural Monopolies in Digital Platform Markets.* Cambridge: Cambridge University Press.

Fialho, P., L. Coheur, and P. Quaresma. 2020. "To BERT or Not to BERT Dealing with Possible BERT Failures in an Entailment Task." In *Information Processing and Management of Uncertainty in Knowledge-Based Systems, Computer and Information Science Book 1237*, edited by M-J. Lesot, S. Vieira, M.Z. Reformat, J.P. Carvalho, A.Wilbik, B. Bouchon-Meunier, and R.R. Yager: 734–747. Cham: Springer International Publishing.

García Lozano, M., J. Brynielsson, U. Franke, M. Rosell, E. Tjörnhammar, S. Varga, and V. Vlassov. 2020. "Veracity assessment of online data," *Decision Support Systems* 129: 113132. DOI: https://doi.org/10.1016/j.dss.2019.113132.

Gentzkow, M., B. Kelly, and M. Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57: 535–574. DOI: https://doi.org/10.1257/jel.20181020.

Heerschap N., B. Klijs, A. Mares, M. van Rossum, and J. Vuik. 2021. "Getting a grip on the platform economy in the Netherlands." In the 36th IARIW Virtual General Conference, August 23-27, 2021, Online conference. Available at: https://iariw.org/wp-content/uploads/2021/07/vanRossum_Paper.pdf (accessed July, 2022).

Howcroft, D. and B. Bergvall-Kåreborn. 2019. "A Typology of Crowdwork Platforms." *Work, Employment and Society* 33: 21–38. DOI: https://doi.org/10.1177%2F0950017018760136.

Klijs, B. 2021. *Monitor online platforms 2020 (in Dutch)*. The Hague/Heerlen: Statistics Netherlands. Available at: https://www.cbs.nl/nl-nl/longread/rapportages/2021/monitor-online-platformen-2020. (accessed July 2022).

Kuhn, M. and K. Johnson. 2013. *Applied Predictive Modeling.* New York: Springer.

Luiten, A., J. Hox, and E. de Leeuw 2022. "Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys." *Journal of Official Statistics* 36: 469–487. DOI: https://doi.org/10.2478/jos-2020-0025.

NACE. 2022. Complete list of all NACE Code. Available at: https://nacev2.com/en (accessed December, 2022).

OECD. 2019. *Measuring the Digital transformation: A Roadmap for the Future*. Paris: Organization for Economic Co-operation and Development. DOI: https://doi.org/10.1787/9789264311992-en.

OECD. 2020. *A roadmap toward a common framework for measuring the digital economy*. Paris: Organisation for Economic Co-operation and Development. Available at: https://www.oecd.org/sti/roadmap-toward-a-common-framework-for-measuring-the-digital-economy.pdf (accessed July, 2022).

Oostrom, L.A.N., A.N. Walker, B. Staats, M. Slootbeek-Van Laar, S. Ortega-Azurduy, and B. Rooijakkers. 2016. *Measuring the internet economy in The Netherlands: a big data analysis*. The Hague/Heerlen: Statistics Netherlands. Available at: https://www.cbs.nl/-/media/_pdf/2016/40/measuring-the-internet-economy.pdf (accessed July, 2022).

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–2830. DOI: https://dl.acm.org/doi/10.5555/1953048.2078195.

Platt, J.C. 2000. "Probabilities for Support Vector Machines." In *Advances in Large Margin Classifiers*, edited by A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans: 61–74. Cambridge: MIT Press.

Puts, M.J.H. and P.J.H. Daas. 2021. "Unbiased Estimations Based on Binary Classifiers: A Maximum Likelihood Approach." In the 2021 Symposium on Data Science and Statistics, June 2-4, 2021, Online conference. Available at: https://arxiv.org/abs/2102.08659 (accessed July, 2022).

Ritzen, J.H.G. 2007. "Statistical Business Register: Content, Place and Role in Economic Statistics." In the 3rd International Conference on Establishment Surveys, June 18-21, 2007, Montréal: Canada. Available at: https://ec.europa.eu/eurostat/ramon/coded_files/RITZEN_Jean_SBR.pdf (accessed July, 2022).

Rochet, J-C. and J. Tirole. 2003. "Platform Competition in Two-Sided Markets." *Journal of the European Economic Association* 1: 990–1029. DOI: https://doi.org/10.1162/154247603322493212.

Snijkers, G., M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J.W. Sakshaug, K.J. Thompson, and A. van Delden. 2023. *Advances in Business Statistics, Methods and Data Collection.* New Jersey: Wiley.

Snijkers, G., G. Haraldsen, J. Jones, and D.K. Willimack. 2013. *Designing and Conducting Business Surveys*. New Jersey: Wiley.

Sutherland, W. and M.H. Jarrahi. 2018. "The sharing economy and digital platforms: A review and research agenda." *International Journal of Information Management* 43: 328–341. DOI: https://doi.org/10.1016/j.ijinfomgt.2018.07.004.

Tourangeau, R., B. Edwards, T.P. Johnson, K.M. Wolter, and N. Bates. 2014. *Hard-to-Survey Populations*. Cambridge: Cambridge University Press.

United Nations. 2019. *Digital Economy Report 2019, Value Creation and Capture: Implications for Developing Countries*. New York: United Nations. Available at: https://unctad.org/system/files/official-document/der2019_en.pdf (accessed July, 2022).

United Nations. 2020. *Guidelines on Statistical Business Registers*, Final draft. New York United Nations. Available at: https://unstats.un.org/unsd/business-stat/SBR/Documents/UN_Guidelines_on_SBR.pdf (accessed January, 2023).

Wu, M-J., K. Zhao, and F. Fils-Aime. 2022. "Response rates of online surveys in published research: A meta-analysis." *Computers in Human Behavior Reports* 7: 100206. DOI: https://doi.org/10.1016/j.chbr.2022.100206.

Appendix

Questionnaire items and definition included in the Dutch Online Platform survey to assess whether the organization's website is an online platform.

An online platform is a website or app where different people, organizations or companies come into contact with each other and can be linked to each other. Goods, services or information can then be exchanged via the online platform. The online platform usually does not supply these goods, services or information itself, but mainly acts as an intermediary.

1. Does your website support or mediate the exchange of goods, services of information between persons, firms or organizations?
   >>This can involve mediation or support in the sale of goods, bringing residents, patients and family together, crowdfunding, dating, new friendships, renting out accommodations, borrowing things etc.<<
   1. Yes
   2. No

2. Are you or is your organization the only provider of the goods, services of information on your website?
   1. Yes
   2. No, in addition to our own supply, there is also supply from other parties
   3. No, there is only offer from other parties

Platform confirmed: question 1='Yes', question 2='No, in addition..' or 'No, there is..'

Tables

Table 1. Words with the 10 highest and 10 lowest coefficients in the trained Support Vector Machine model

Word	Coefficient	Word	Coefficient
platform	2.526	Portfolio	-1.202
account	1.690	Phone	-1.113
help	1.653	Info	-1.098
crowdfunding	1.587	Skip	-1.047
register	1.551	Approach	-1.029
login	1.236	Year	-1.025
entrepreneur	1.175	Wordpress	-0.982
deal	1.152	Since	-0.981
ask	1.149	P.O. box	-0.964
neighborhood/vicinity	1.143	customization	-0.959

Table 2. The selection procedure followed*

	Number	Percentage selected	Percentage surveyed
Websites with a platform probability of at least 0.8	9,129	100.0%	
Removal of adult content websites	7,764	85.0%	
Distinct legal entities	6,057	66.3%	
Legal entities approached with the platform survey	4,385		100.0%
Response from the legal entities	2,997		68.3%
Usable response for empirical analysis	2,708		61.8%
Usable response excluding websites in the “combined data set”	2,631		60.0%

* Only websites with a platform probability of 0.8 or higher, as indicated by the model developed, were included

Table 3. Summary statistics

	Number of organizations	Percentage platform
Total	2,631	23.1
Platform probability		
0.800-0839	753	15.9
0.840-0.879	649	20.8
0.880-0.919	521	25.7
0.920-0.959	411	27.0
>=0.960	297	36.7
Number of employees		
<=1	1,366	24.1
1.1-4.9	594	24.9
5.0-19.9	328	22.6
20.0-49.9	144	23.6
>=50	199	12.1
Branch		
Wholesale and retail trade; repair of motor vehicles	420	12.1
Information and communication	745	31.7
Renting and trading real estate	189	15.9
Consultancy, Research and Other Specialist Business Services	444	24.8
Rental of movable property and other business services	166	34.9
Other branches	667	18.6
Legal form		
Proprietorship	971	20.9
Private company	1,192	24.0
General partnership	267	27.0
Other legal form	201	23.9

Table 4. Experts' second opinion on the false positives

	Number (%) no platform according to survey response	Percentage positive according to experts' check[a]
Platform probability		
0.800-0.839	633 (84,1)	15.0
0.840-0.879	514 (79.2)	10.0
0.880-0.919	387 (74.3)	25.0
0.920-0.959	300 (73.0)	20.0
>=0.960	188 (63.3)	65.0
Total	2,022 (76.9)	21.0[b]

[a] For a random sample of 100 organizations (20 in each stratum of platform probability).

[b] Weighed by number of organizations in the five strata.

Table 5. Estimates equation (5), part 1

	Estimate (Robust Standard Error)		
	Model 1	Model 2	Model 3
Intercept	-0.823 *** (0.133)	-0.744 *** (0.135)	-0.655 (0.512)
Platform probability	1.196 *** (0.152)	1.073 *** (0.150)	0.973 . (0.585)
Number of employees			
<=1		0.025 (0.026)	-0.088 (0.407)
5-19.9		-0.017 (0.030)	-0.135 (0.502)
20-49.9		-0.03 (0.041)	-1.05 . (0.623)
>=50		-0.128 *** (0.032)	1.026 . (0.542)
Branch			
Wholesale and retail trade; repair of motor vehicles		-0.104 *** (0.026)	0.297 (0.438)
Information and communication		0.059 * (0.026)	0.244 (0.422)
Renting and trading real estate		-0.093 ** (0.034)	0.205 (0.582)
Rental of movable property and other business services		0.107 * (0.042)	0.235 (0.627)
Other branches		-0.065 * (0.026)	0.682 . (0.404)
Legal form			
General partnership		0.086 * (0.036)	-0.097 (0.592)
Private company		0.064 ** (0.023)	-0.459 (0.359)
Other legal form		0.111 ** (0.036)	-1.595 * (0.641)
Interactions platform probability * employees			
PP * one or less			0.130 (0.465)
PP * more than five, less than twenty			0.137 (0.577)
PP * mode than twenty, less than fifty			1.164 (0.722)
PP * more than fifty			-1.320 * (0.619)
Interaction platform probability * Branch			
PP * Wholesale and retail trade; repair of motor vehicles			-0.458 (0.507)
PP * Information and communication			-0.211 (0.482)
PP * Renting and trading real estate			-0.339 (0.669)
PP * Rental of movable property and other business services			-0.140 (0.712)
PP * Other branches			-0.847 . (0.462)
Interactions platform probability * Legal form			
PP * General partnership			0.205 (0.677)
PP * Private company			0.590 (0.409)
PP * Other legal form			1.948 ** (0.739)
Goodness-of-fit			
R-squared	0.024	0.062	0.069
Percentage predicted platform probability between 0-1	100.0%	99.1%	99.7%

Dependent variable is the 0-1 variable $Y\_platform$.

*** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1

Table 6. Estimates equation (5), part 2

	Estimate (Robust Standard Error)	
	Model 1	Model 2
Intercept	0.159 *** (0.013)	0.137 *** (0.034)
Platform probability		
0.840-0.879	0.049 * (0.020)	0.042 * (0.021)
0.880-0.919	0.098 *** (0.023)	0.096 *** (0.023)
0.920-0.959	0.111 *** (0.026)	0.092 *** (0.025)
>=0.960	0.208 *** (0.031)	0.187 *** (0.030)
Number of employees		
<=1		0.024 (0.026)
5-19.9		-0.018 (0.030)
20-49.9		-0.034 (0.041)
>=50		-0.129 *** (0.032)
Branch		
Wholesale and retail trade; repair of motor vehicles		-0.104 *** (0.026)
Information and communication		0.062 * (0.027)
Renting and trading real estate		-0.091 ** (0.034)
Rental of movable property and other business services		0.111 ** (0.042)
Other branches		-0.063 * (0.026)
Legal form		
General partnership		0.085 * (0.036)
Private company		0.063 ** (0.023)
Other legal form		0.110 ** (0.036)
Goodness of fit		
Multiple R-squared	0.023	0.061

Dependent variable is the 0-1 variable $Y\_platform$.

*** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1

Figures

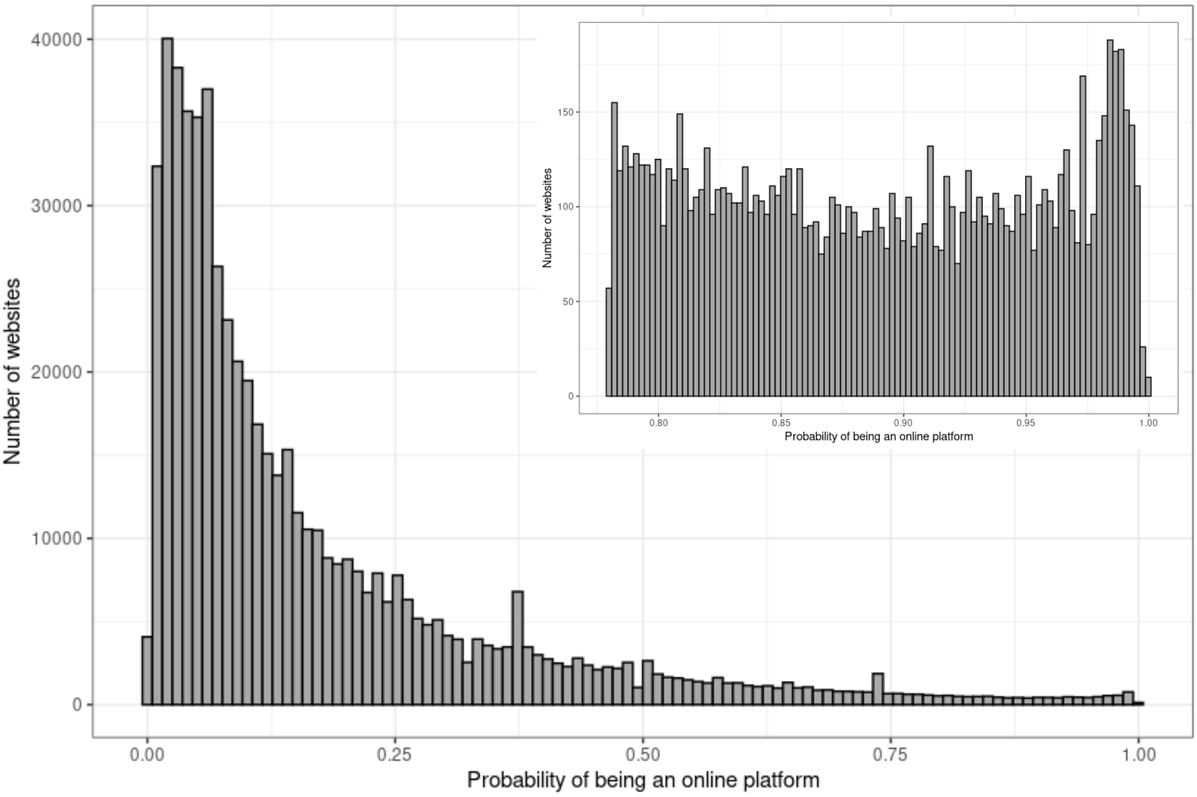

Figure 1. Histogram of the model-based probabilities of being an online platform website*

*Information is used for the 594,574 websites with 10 words or more linked to the Business Register. The insert shows the findings for websites with a probability above 0.78 and reveal a peak around 0.98-0.99.