2020

Using machine learning for detection of innovative companies in Lithuania

Technical report

Author: Vitalijus Klincevičius

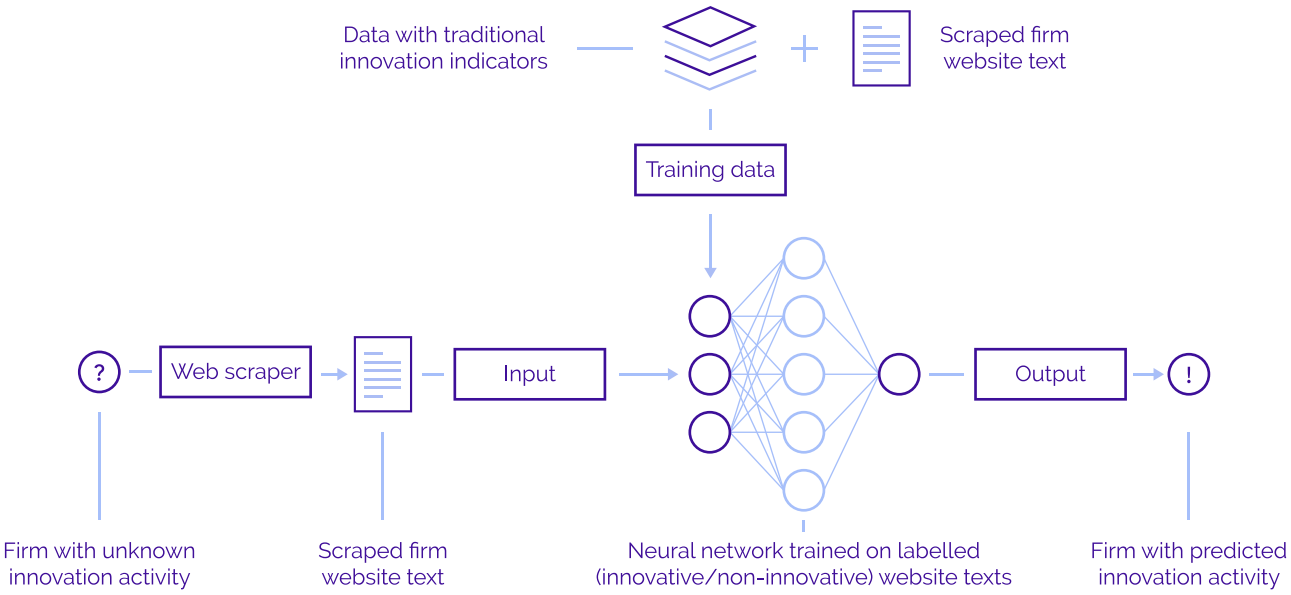
Strata

1. The need for an experiment

Innovative companies play an important role in promoting economic development at the national level. The attempt to identify such enterprises stems from statistical incentives and a desire to design intervention measures in a targeted way. However, the process of trying to identify such companies is costly in terms of the costs involved in obtaining and processing the data. Traditional indicators, such as R&D expenditure or the number of patents, are widely used to assess the innovativeness of companies however these methodological approaches also have significant shortcomings. Businesses may be inclined to avoid declaring R&D expenditure or to declare it incorrectly due to significant administrative costs. The identification of patenting firms also has limitations in assessing their potential for innovation, as not all inventions are patentable; this is due to differences in sectors as companies may use other ways to protect their inventions.

Existing constraints encourage the search for alternative methodological approaches to identify potentially innovative companies. One such approach is machine learning (ML) models that demonstrate their potential in performing text classification tasks. The aim is to identify potentially innovative companies using ML solutions. This study analysed websites of companies whereas data obtained from the websites were presented to a trained artificial neural network (Fig. 1).

Fig. 1 Research process



Source: Adopted from Kinne, Jan & Lenz David

Tests carried out in Germany showed that neural networks are able to identify business innovation with 80 percent accuracy based on the website content (Kinne, Jan & Lenz David, 2019). This study assumes that most companies these days have websites where companies seek to present themselves. This information may include indicators to identify the status of business innovation. The main advantage of this method over traditional data collection and processing methods is that the development of the model allows for the rapid and relatively accurate collection of information on the market situation. However, this method has significant limitations. First, businesses in the agricultural sector may be less likely to have websites. Some start-ups may not have a website at all making them inaccessible to this model. It should be noted that a large amount of data is required for the construction of the neural network in order to 'teach' this model to classify texts according to certain features. A test carried out in Germany showed that alternative machine learning models, such as the naive Bayes classifiers, logistical regression and decision trees had poorer classification results (Kinne, Jan & Lenz David, 2019).

Methodology

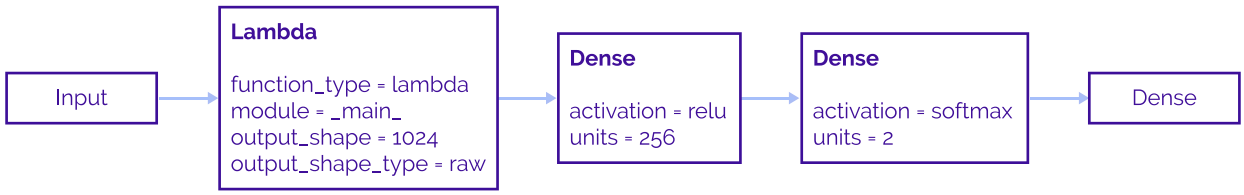
Data on companies from the Lithuanian 2014-2020 Operational Programme Priority 1 *Promotion of Research, Experimental Development and Innovation* and companies that have registered patents with the Lithuanian Patent Bureau were used for the training of the model. In order to form a larger sample for model training, the study used foreign innovative companies. These companies were selected using the *Crunchbase* database.

The ARGUS tool was used to scan websites of the companies, which allows to scrape websites. A limit of 25 links per domain was set when scraping websites of the companies. Such a number was chosen based on the experience of similar studies related to website scraping. Websites that were redirected to another domain were removed from the sample and the possible reasons of such redirection were the following: the company has changed its domain, the company is collaborating with other companies, or an error has been made in specifying the company's website. It has been observed that companies with different legal entity codes have provided the same website address that may be related to the fact that some companies belong to associations and/ or have subsidiaries, such cases will be analysed separately.

Other text standardization techniques were used in the data preparation process: 1) uppercase letters were changed to lowercase letters, 2) numbers, punctuation marks, and other ambiguous symbols were removed, 3) insignificant grammatical words such as 'and', 'or', 'but', etc. were removed. These data preparation techniques controlling all other variables increased the accuracy of the model by 2-3 percent.

A sequential model which consisted of four layers two of which were educational (Fig. 4) was used in this study. The first layer corresponds to the input point which indicates that the model analyses one sentence at a time. Since punctuation marks were removed during the data preparation process, one sentence corresponded to one case. The next layer is Lambda which consists of 1024 neurons in which the text is converted into vectors (is digitized) using a deep contextualized dictionary - ELMo. The training process takes place in layers 3 and 4 indicated by the number of parameters in each layer. The parameters correspond to the number of training elements in each layer. The model contains 262,914 training parameters of which 262,400 are in the third layer of the model that contains 256 neurons. There are two neurons in the fourth layer that corresponds to the predicted categories, in this case innovative and non-innovative. There are 514 training parameters in this layer. The tests performed showed that the change in the number of neurons in the layers did not significantly improve the predictions of the model. It has been observed that the addition of additional hidden layers lengthens the training process however the learning progress of the model does not improve significantly or even in some cases the model overfits.

Fig. 2 Model structure



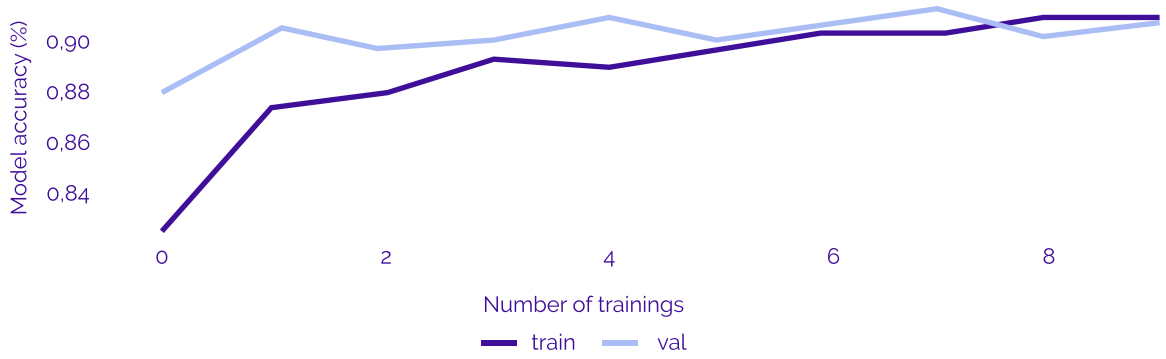
Source: STRATA

Model performance and results

Learning results showed relatively high accuracy (Fig. 3) - the model correctly classified over 90 percent cases after ten trainings. It was decided to discontinue the model learning process after ten trainings as further training progress was insignificant. The model did not overfit as the validation curve remained at a similar level in the course of training and the gap from the training curve did not deviate significantly. The variations in the validation curve are also insignificant and therefore can be treated as falling within the margin of error. When analyzing the model learning process in detail, it is important to note that out of 2671 predicted non-innovative companies, 2378 of them were truly non-innovative.

This represents about 89 percent of the group of non-innovative companies, however, among the predicted innovative companies; the number of erroneous predictions was less than 6 percent. Therefore, it can be argued that the model is more prone to type II error. A possible reason is the different degree of innovation between Lithuanian and foreign companies used to train the model.

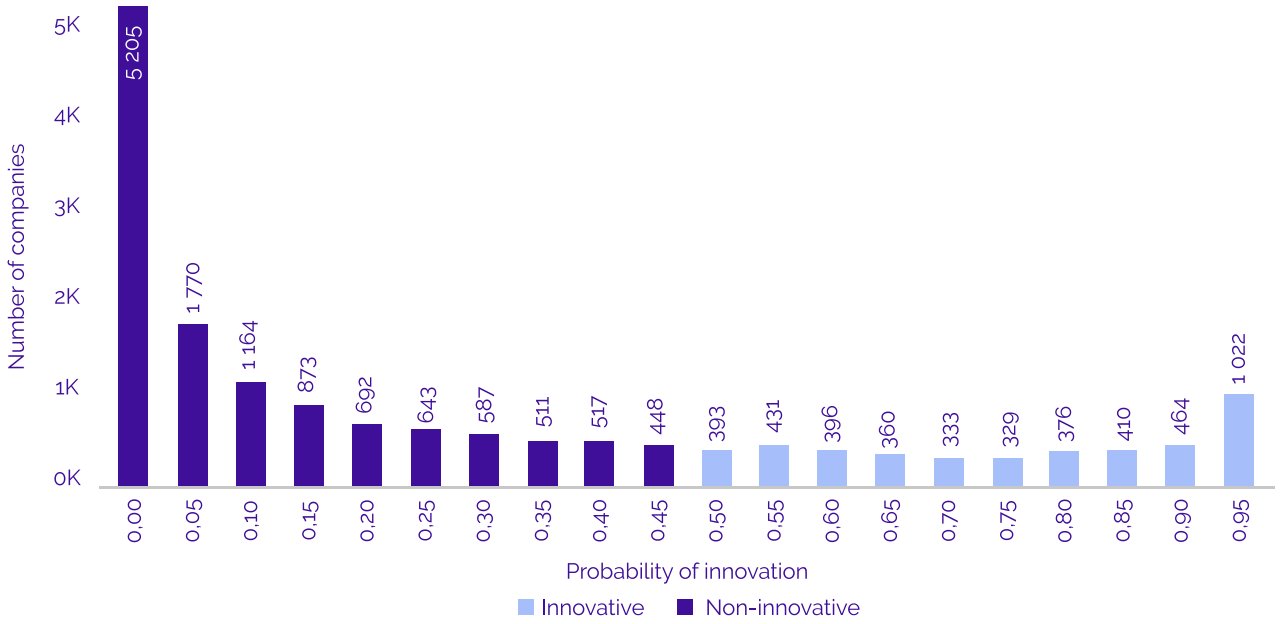
Fig. 3 Model learning progress



Source: STRATA

The trained artificial neural network was adapted for the analysis of websites of the remaining companies operating in Lithuania. A total of 16924 companies operating in Lithuania, which had a unique domain of the website, were included in the sample. The distribution of the probability of innovation of these companies is illustrated in Figure 4. Based on the predictions of this model, 4514 companies were identified as corresponding to the profile of innovative companies. This represents about 27 percent of all companies that were analysed.

Fig. 4 Predicted distribution of the probability of innovation of companies



Source: STRATA

The results of the surveys of companies conducted by the Lithuanian Department of Statistics (hereinafter referred to as 'LDS') which covered the periods 2014-2016 and 2016-2018 were used for data validation. Before performing data validation, it is important to emphasize that there are methodological differences between this study and surveys conducted by LSD.

Table 1. Comparison of methodological differences between STRATA and LSD studies

	STRATA study	**LSD survey**
Approach	Individual analysis of every company	Representative survey of the companies
Selection of companies	Companies with an official website	Companies with 10 or more employees
Difference between populations	All companies with available website (16,925)	Use of a sampling method covering 2,500 companies

Source: STRATA

However, despite methodological differences, the sample of companies analysed in this study overlaps with the survey conducted by LSD in 938 cases in 2016 - 2018 period and 877 cases in 2014 - 2016 period (Tables 2 and 3). The accuracy of the model judging by the results of both surveys was 46-47 percent. However, the true positive rate of the model, i.e. how many predicted innovative companies were actually innovative, was 71 percent compared to the 2016-2018 period survey and 79 percent compared to the 2014-2016 period survey. It is observed that among the predicted non-innovative companies, model predictions are confirmed by 36-39 percent compared to the

surveys conducted by LSD in both periods. These trends persist in controlling the regional distribution of companies. Possible reasons are related to the fact that companies operating abroad were used in the training process and therefore the predictions of the model may not correspond to the trends prevailing in the Lithuanian market.

Table 2. Model validation with a survey of companies conducted by LSD in 2016-2018

		True condition		
	n = 938	Innovative	Non-innovative	
Predicted condition	Innovative	163	65	228
	Non-innovative	434	276	710
		597	341	

Table 3. Model validation with a survey of companies conducted by LSD in 2014-2016

		True condition		
	n = 877	Innovative	Non-innovative	
Predicted condition	Innovative	155	42	197
	Non-innovative	429	251	680
		584	293	

While analyzing the model predictions by sectors of activity, it is observed that model predictions are most true for predicting innovative companies engaged in transport and storage activities (compared to surveys - 87 percent of cases). Among the predicted innovative companies engaged in activities related to wholesale and retail trade and motor vehicle repair, the model predictions proved to be the least correct; in both surveys conducted by LSD, the model forecasts were confirmed by 66-71 percent.

Lesson learned:

Conducted study provides valuable insights about ML model implementation in public policy research field. Model performance in training process shows that model is able to correctly classify up to 91 percent of cases. Out-of-sample validation during training process conforms this result. Additional model prediction validation using LSD surveys shows model tendency to misclassify non-innovative companies (according LSD surveys results). It is observed that among the predicted non-innovative companies, model predictions was confirmed only by 36-39 percent. However model underperformance in identifying non-innovative companies might be result of small validation sample, only up to 938 (out of 16924) cases were used for validation. Likewise, companies operating abroad were used in the training process, because of this "knowledge", the model might not correspond to tendencies in local environment. Furthermore, in order to improve model prediction, quality of data could be improved by implementing web page spam classifiers. Potentially it might impact the accuracy of model positively. Despite some ambiguous performance in predicting non-innovative companies, model shows decent result in predicting innovative companies, where up to 79 percent of cases were confirmed. Main advantage of ML related approach in predicting innovators is low-cost, compared to survey-based approach. Costs mainly incur in initial development process while further costs of application of the model are significantly lower.