Sõna, Sander; Masso, Jaan; Sharma, Shakshi et al.

Book

Predicting company innovativeness by analysing the website data of firms : a comparison across different types of innovation

University of Tartu
School of Economics and Business Administration

PREDICTING COMPANY INNOVATIVENESS BY ANALYSING THE WEBSITE DATA OF FIRMS: A COMPARISON ACROSS DIFFERENT TYPES OF INNOVATION

Sander Sõna, Jaan Masso, Shakshi Sharma, Priit Vahter, Rajesh Sharma

Tartu 2022

ISSN-L 1406-5967
ISSN 1736-8995
ISBN 978-9985-4-1328-9 (pdf)
The University of Tartu FEBA
https://majandus.ut.ee/en/research/workingpapers

Predicting company innovativeness by analysing the website data of firms: a comparison across different types of innovation

Sander Sõna, Jaan Masso, Shakshi Sharma, Priit Vahter, Rajesh Sharma*

Abstract

This paper investigates which of the core types of innovation can be best predicted based on the website data of firms. In particular, we focus on four distinct key standard types of innovation – product, process, organisational, and marketing innovation in firms. Web-mining of textual data on the websites of firms from Estonia combined with the application of artificial intelligence (AI) methods turned out to be a suitable approach to predict firm-level innovation indicators. The key novel addition to the existing literature is the finding that web-mining is more applicable to predicting marketing innovation than predicting the other three core types of innovation. As AI based models are often black-box in nature, for transparency, we use an explainable AI approach (SHAP - SHapley Additive exPlanations), where we look at the most important words predicting a particular type of innovation. Our models confirm that the marketing innovation indicator from survey data was clearly related to marketing-related terms on the firms' websites. In contrast, the results on the relevant words on websites for other innovation indicators were much less clear. Our analysis concludes that the effectiveness of web-scraping and web-text-based AI approaches in predicting cost-effective, granular and timely firm-level innovation indicators varies according to the type of innovation considered.

JEL Classification: C30, C84, C88.

Keywords: Innovation, Marketing Innovation, Community Innovation Survey (CIS), Machine learning, Neural network, Explainable AI, SHAP

* Sander Sõna, Junior Research Fellow of International Business and Innovation, School of Economics and Business Administration, University of Tartu, Estonia. E-mail: sander.s6na@gmail.com.
Jaan Masso, Associate Professor of Applied Econometrics, School of Economics and Business Administration, University of Tartu, Estonia. E-mail: jaan.masso@ut.ee.
Shakshi Sharma, Junior Research Fellow of Information Systems, Institute of Computer Science, University of Tartu, Estonia. E-mail: shakshi.sharma@ut.ee.
Priit Vahter, Professor of Applied Economics, University of Tartu, School of Economics and Business Administration; Priit.Vahter@ut.ee
Rajesh Sharma, Associate Professor in Information Systems, Institute of Computer Science, University of Tartu, Estonia. E-mail: rajesh.sharma@ut.ee
The research was written partly as the master's thesis of Sander Sõna in the School of Economics and Business Administration at the University of Tartu supervised by Rajesh Sharma, Jaan Masso and Priit Vahter. The authors acknowledge financial support from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 822781 GROWINPRO– Growth Welfare Innovation Productivity, and the Estonian Research Council project PRG791 "Innovation Complementarities and Productivity Growth". Rajesh Sharma and Shakshi Sharma also acknowledge funding from the EU H2020 programme under the SoBigData++ project (grant agreement No. 871042), CHIST-ERA grant CHIST-ERA-19-XAI-010, FWF (grant No. I 5205), EPSRC (grant No. EP/V055712/1), NCN (grant No. 2020/02/Y/ST6/00064), ETAG (grant No. SLTAT21096), BNSF (grant No. КП-06-ДОО2/5). We owe thanks to Statistics Estonia for their indispensable help in supplying data.

1. INTRODUCTION

Traditional innovation indicators, such as firm-level innovation survey variables – R&D indicators or patent statistics – each have their limitations in terms of either timeliness in providing the information needed for policy making, cost-effectiveness in terms of expenditure on collecting information, or limitations related to granularity and representativeness of these indicators, such as limited coverage of some sectors, size-groups and types of firms (Mairesse and Mohnen, 2010; Kinne and Lenz, 2019; OECD, 2015: 26–28; Kinne and Axenbeck, 2018).

Data mining from the web for the measurement of innovation could potentially help address some of the shortcomings of these standard innovation indicators (Kinne and Lenz, 2019). Web data on the internet is mostly freely available and with good coverage, as most firms own a website. One of the main communication channels of firms is their homepage on the internet. The homepage is like an organisation's business card and the first place where clients and partners go to get more information.

However, previous research on the use of web data in predicting innovation is still relatively scarce and in the early stages of development. Prior studies on this issue suggest as a general lesson that web-content mining (text on websites), web structure mining (hyperlinks), or web-search terms could potentially be used to predict firm-level innovation (Fisher et al., 2007: 253; Gök et al., 2015: 654; Kinne and Lenz, 2019; Axenbeck and Breithaupt, 2021, Krüger et al., 2020). In particular, the most related prior studies to ours are by Kinne and Lenz (2019, 2020) and Axenbeck and Breithaupt (2021) that use German firm-level survey data on innovation (Mannheim Innovation Panel) combined with data from the websites of firms. They find clear evidence that one can construct new indicators based on data from the websites of firms that are reasonably accurate proxies for firm-level innovation.

First, we add to this literature by showing which of the core types of innovation output in firms (as defined by the Oslo Manual 2005) can best be predicted based on website data and using machine learning methods. We focus on four distinct key standard types of innovation – product, process, organisational, and marketing innovation. The first two are technological innovations, and the latter two are non-technological innovations. Innovation output indicators – product, process, organisational and marketing innovation – reflect innovation that is new to the firm but may not be new to the market.[1]

To date, it remains still largely unclear which types of innovation can be predicted well based on web data and which ones not, and which textual features of the website matter for accurate predictions of different innovation proxies at firm level. A recent exception is a paper by Axenbeck and Breithaupt (2021) that predicts different types of innovation. They confirm that web-mining and machine learning based approaches are more successful in predicting product innovation and non-innovators than process innovation in firms. This is a natural result to expect, as much of the website's information would be targeted toward clients and would be

[1] Process innovation is defined as the "application of new or significantly improved methods for the production or delivery/distribution of a good or service" (OECD/Eurostat 2005, p. 48). Product innovation is the provision of new or significantly improved goods or services. Organisational innovation is "new or significantly changed business practises in the organisation of work, business structure and decision-making or in ways to manage external relations" (OECD/Eurostat 2005, p. 49). Marketing innovation is "the implementation of a new marketing method involving significant changes in product design or packaging, product placement, product promotion or pricing" (OECD/Eurostat 2005, p. 49).

about the firm's product portfolio. Communicating the adoption of process or organisational innovation may be less frequent on a website. It may simply be less crucial information to publish for achieving market success and informing clients. Reporting process innovation may vary a lot depending on the type and size of the technology investments, with more minor process innovations with less immediate implications on the commercialisation of products remaining less visible in this type of data. Another relevant study is Ashouri et al. (2021), which retrieves the data of almost 100,000 medium-high and high-technology firms in the European Union and United Kingdom and identifies from that data information on products, collaborations and ISO codes. However, that was done without studying how that information helps predict the innovativeness of firms, as reported in sources such as the Community Innovation Survey (CIS).

The angle we add here compared to Axenbeck and Breithaupt (2021) is the prediction of non-technological innovation, such as marketing and organisational innovation. In general, we would expect website data to predict product and marketing innovation with higher accuracy. In contrast, organisational innovation and process innovation are less likely to be visible on a firm's website and thus more difficult to predict based on web data.

Our second contribution to the literature concerns the role of the features of websites in predicting innovation. We add to the prior literature by analysing the textual data (words) on websites that help predict innovation. We show that models that could predict marketing innovation use words that were clearly related to marketing activities in firms, but the models predicting other innovation indicators lacked such a clear set of relevant words. AI approaches, specifically deep learning models, considered excellent for dealing with textual data, are also considered black-box in nature. In other words, it is not easy to identify what features these models banked on for concluding their predicted outcome. To overcome the non-transparency of these advanced models, we used explainable AI method, specifically, SHAP (SHapley Additive exPlanations*)* to understand the feature (specific words) importance. The outcome of SHAP indicates that identifying and predicting marketing innovation were clearly related to marketing-related words on the websites of the firms. In contrast, there were much less clear results on relevant words on websites in the case of other innovation indicators. Our results confirm that the effectiveness of web scraping combined with deep learning approaches to provide cost-effective, granular and timely firm-level information on innovation indicators varies according to the type of innovation.

We used the Community Innovation Survey (CIS) 2016 and 2018 (hereinafter CIS2016 and CIS2018) as the input data. For all the firms that participated in either of these two waves of the CIS surveys, we scraped their web pages and combined the text from their web pages with the innovation labels appointed by CIS. To collect the text from the websites of firms, we used a program called ARGUS (Automated Robot for Generic Universal Scraping).

We collected data on as many firms as possible (1,290 firms). We tested various AI techniques (specifically machine learning and neural network methods), namely, RandomForest, XGBoost, Long-Short-Term-Memory (LSTM), Gated Recurrent Unit (GRU) etc[2]. The most prominent results were achieved when we filtered out the web pages of firms with less than 10,000 overall words (meaning they also contained duplicated words). With a dataset of 388 web pages, we could obtain acceptable models for predicting marketing and overall innovation (companies reporting any of the four main types of innovation). But because innovation is such a complex phenomenon, we also needed to test their feature usage

[2] From here on we use the term AI and Machine Learning interchangeably.

(out of 162,296 unique words in combination). We needed to ensure that the company being labelled innovative was not determined by some webpage structural tag but rather by words related to a firm and its innovation type. To decrypt the features of the model, we used SHAP because it gave stable and repeatable results. Therefore, an additional contribution of our study is showing whether investigating company innovativeness using machine learning and company websites is also applicable in the case of small countries with a much smaller number of active companies; however, in addition to the mere number of companies available, also the amount of the text on the websites is expected to matter.

Based on data from Estonia, we confirm that web-mining textual data from the websites of firms combined with the application of machine learning methods is suitable to predict innovation in firms with reasonable accuracy compared to the responses of firms on innovation in CIS 2016 and 2018.[3] Compared to the existing literature, the first key novel finding is that web-mining appears to be especially suitable for predicting marketing innovation compared to the other three core types of innovation, even compared to product innovation.

Second, we show that the set of words used by the algorithms in the identification and prediction of marketing innovation was clearly related to marketing activities in firms, with a lack of such a clear set of relevant words in the case of other innovation indicators. Our results confirm that the effectiveness of web scraping combined with machine learning approaches to provide cost-effective, granular, and timely firm-level information on innovation indicators varies according to type of innovation.

The rest of the paper has been structured as follows. The following section reviews the literature on the earlier uses of website data, machine learning and data mining in economic research, particularly concerning innovation studies. The third section describes in detail the data collection process via web scraping, the innovation data used and the learning mechanism. The fourth section summarises the performance of the used models in predicting firm innovativeness using website data and the SHAP analysis to further understand the words used in the prediction. The final section concludes with implications and suggestions for future research, particularly using similar indicator building exercises.

2. LITERATURE REVIEW

In social science, words and text have a lot of meaning. Decoding a text can be an opportunity to get far richer explanations for phenomena in comparison to the use of more structured kinds of data. In recent years there has been a rise in empirical studies in economics and business fields that use text as data (Gentzkow et al., 2019). For example, in macroeconomics, there have been studies using neural networks on unemployment and inflation (Wanto et al., 2018; Choudhary 2012). In marketing, there have been studies with neural networks to understand the drivers of consumer decision-making (Baesens et al., 2002). In general, web scraping and text mining have begun to rise in various fields as novel tools and sources of insights (Levenberg et al., 2014).

[3] We tested our overall dataset (1,290 companies) using simple neural network methods (e.g. LSTM, GRU Conv1D) and deep learning methods, such as combinations of the above with more than 3 layers, BERT with text summaries. Embeddings were TF-IDF and Word2vec. The best results were still obtained using machine learning methods.

There are substantial potential advantages of using web text as data and a complementary source of information on innovation, in addition to traditional innovation indicators from questionnaire-based surveys, patent-based studies and R&D indicators (Nagaoka et al., 2010; Kinne and Lenz, 2019). The use of web-based text as data can have advantages in terms of coverage and granularity concerning sectors, firms and regions. It can also improve the timeliness of the indicators for their application and reduce data collection costs (Kinne and Axenbeck, 2018).

Most of the data is free and openly available on the internet. In the case of web data, there is no need to contact and interview someone to collect their responses or focus only on a small and limited sample of respondents due to the costs of collecting the information (Kinne and Axenbeck, 2018). Website data enables us to gather information on particular sectors and types of firms that are under-represented in CIS-type surveys. Unlike patent statistics that cover only patentable knowledge, and thus refer to a limited set of firms engaging in patenting (Smith, 2009; Mairesse and Mohnen, 2010), website data could help predict broader new-to-firm level innovation for a more comprehensive set of firms (Kinne and Lenz, 2019). Web data would also make it possible to cover small and resource-poor firms that are much less likely to engage in patenting and would also help collect innovation data on novelties that are not directly patentable (Axenbeck and Breithaupt, 2021).

A further advantage is the timeliness of web data compared to many other sources of innovation data. Website data may also benefit from collection speed, which could mean the faster application of new collected knowledge in policy making and research (Kinne and Lenz, 2019, Pukelis and Stanciauskas 2019). In comparison, applying firm-level innovation data from the CIS survey from a particular year takes several years until the data becomes available, seriously limiting its application in timely policy decision-making or innovation research on timely issues such as, for example, the Covid-19 shock on firms. Collecting real-time or most recent information and doing that regularly and automatically could make a significant difference in the impact of the collected information.

At the same time, web-data applications have their limitations. The classical problem with analysing text using machines is its inherently high dimensionality (Gentzkow et al., 2019: 535). Usually, raw data is represented as a numerical array because the computer does not understand words. Then it is mapped with the predicted values of unknown outcomes. Later, the outcome is used in subsequent descriptive or causal analyses (Gentzkow et al., 2019: 536). So, if the problem/phenomenon needs a lot of input data, it also requires a lot of computational power.

Another fundamental limitation is the self-reported nature of data on firms' websites, which limits the confidence that what is reported is necessarily directly comparable across firms or across time (e.g., as discussed in the context of web data in Pukelis and Stanciauskas, 2019). However, the self-reported nature is similarly a problem in the case of innovation survey data. The collected information reflects each respondent's (manager's) perceptions of innovation and may or may not reflect the reality in the firm; see, for example, Bloom et al. (2012) for a discussion of the limitations of using manager's perceptions of management quality in the firm.

Further, web-data use means reusing existing self-reported information, which was often created to boost sales and communicate a favourable image of the firm to its clients and the public. The data was not created as an accurate measurement or with policy making in mind (Pukelis and Stanciauskas, 2019). Finally, the frequency of text changes on websites can vary

a lot across firms, which can make the accurate measurement of change in innovation indicators over time difficult.

Despite the high expectations and the need to complement standard innovation indicators, few studies still apply web data or use machine learning algorithms to learn about firm innovation activities from their homepages. Differently, research on the use of web data to predict innovation is still in the early developmental stages. As a general lesson, prior studies suggest that web-content mining (text on websites), web structure mining (hyperlinks) or Google-search terms could potentially be used to predict firm-level innovation (Fisher et al., 2007: 253; Gök et al., 2015: 654; Kinne and Lenz, 2019; Axenbeck and Breithaupt, 2021; Krüger et al., 2020). In particular, those studies most related to ours are by Kinne and Lenz (2019, 2020) and Axenbeck and Breithaupt (2021), which use German firm-level survey data on innovation (Mannheim Innovation Panel) combined with website data and web-scraping with the ARGUS web-scraper. They find clear evidence that one can construct new innovation indicators based on data from websites that are reasonably accurate proxies for firm-level innovation.

Kinne and Lenz (2019) show that it is possible to train algorithms with firm-level website data to classify firms as innovative or non-innovative. They used CIS-type data from the Mannheim Innovation Panel (MIP) to categorise the innovativeness of firms and neural network modelling to train it. The result was similar to CIS, so the neural network understood what was on the webpages of innovative firms, and it could predict with reasonable accuracy whether a firm was innovative or not just by the text on its website. A further related study by Pukelis and Stanciauskas (2019) also applies a machine learning approach using website data and artificial neural networks, confirming the results of Kinne and Lenz (2019) that web data can predict innovation in firms. A study by Mirtsch et al. (2020) applies a web-mining based approach to analyse the adoption of an information security management system that complies with an ISO/IEC standard as indicating the adoption of an innovative practice. Krüger et al. (2020) apply large-scale web scraping to the structure of the websites, their text content and hyperlinks between websites to investigate the vital network-related characteristics of innovative and non-innovative firms.

However, how well web scraping and deep learning models can produce new data on different types of innovation remains the area with limited attention in this field of research. A recent study investigating this issue was conducted by Axenbeck and Breithaupt (2021). Based on web-mining of data on German firms and Random Forest classification models, they show that web data is better for predicting product innovation and non-innovative firms than predicting process innovation. To the best of our knowledge, there is still a shortage of studies that study how well web data and machine learning based approaches work for predicting marketing innovation, which is one innovation type that could possibly be more visible in internet data compared to information on the adoption of process innovation or organisational innovation.

Various research focuses on how to make black-box models trustworthy when comprehending and interpreting the decisions made by the models (Bejger and Elster, 2020; Hoepner et al., 2021). In this regard, AI explainability tools like SHAP and LIME[4] are widely employed in various domains, including economics (Bussmann et al., 2021). For example, explainable

[4] LIME (Local Interpretable Model-agnostic Explanations) is an explainable AI method that helps to explain the classifier for a specific single instance and is therefore suitable for local explanations.

models have been used to analyse the risks associated with credit cards (Hadji Misheva et al., 2021) or to predict mortgage defaults (Bracke et al., 2019).

3. DATA COLLECTION PROCESS

The main goal of our research is to understand whether the webpage texts published by innovative firms have anything common. Our study follows the two German studies by Jan Kinne and David Lenz (2019) and Jan Kinne and Janna Axenbeck (2018). Both articles used the Mannheim Enterprise Panel (hereinafter MUP) database in Germany combined with the Mannheim Innovation Panel (MIP).[5] Kinne and Lenz (2019) used 8,080 firms from all over Germany, selected based on the criteria of having a constant innovation tag over the three years and clean plot-free webpages (Kinne and Lenz, 2019). In Estonia, the number of firms is significantly smaller than in the German study due to the small size of the Estonian economy. Overall, there are about 235,000 registered firms (legal entities), but the number of economically active firms (those with positive sales or number of employees) is much lower; in 2020, 126,000 (e-Äriregister, 2021; own calculations using the Estonian Business Registry data). The number of unique firms participating in CIS 2016 and CIS 2018 and having Estonian language homepages is about one thousand. Not every firm in CIS has a webpage, and the number of unique firms present in either CIS2016 or CIS2018 is reduced by some firms being included in both surveys. With that in mind, we did not set any constraints on firms for inclusion in our list, except that the only criteria was that the firm must have a web page. The webpage URL information was retrieved from the Estonian Business Register (Estonian *Äriregister*) and Amadeus (currently Orbis Europe) databases. The website data was collected from November 2020 until January 2021.

There were 1,964 firms present in CIS 2016 or 2018 waves. We considered only the Estonian language texts of the web pages for our analysis. There were also web pages in English, Russian, or Finnish only (and with no equivalent version in Estonian), and these were discarded from the analysis. The main reason is that conducting machine learning solely on the text of one language creates enough complexity because we have millions of words in just one language (here: Estonian). Letting machines study different languages within the same study is computationally challenging, raising additional needs for computer hardware and time. We managed to identify and use 1,290 firms with Estonian language web pages. The firms were filtered by their business registry number, so if one firm had two or more web pages, it was still counted as one distinct firm. The innovativeness of the firms, as reported in CIS, was an essential input for our analysis because we used that to label the training set, where later the machine tries to learn and imitate the innovation of the firm in the test set.

We used the Python web scraping program for collecting data from webpages, called Automated Robot for Generic Universal Scraping (ARGUS) that was also used in the German studies (Kinne and Lenz, 2019: 3; Kinne and Axenbeck, 2018: 9–11; ARGUS: Automated...,

[5] MUP database is the most comprehensive database of companies in Germany available for the public (the official business registry is not) based on the data of the largest German credit rating agency Creditreform e.V. MIP database is the innovation survey commissioned by the Federal Ministry of Education and Research. It contains a random sample of German companies with five or more employees (ZEW, 2022). As the data to the MIP database is collected every year, they can look at companies who have been three years straight innovative or non-innovative. There were 2.52 million companies in Germany, and 1.15 million of these (roughly 46%) had web pages (in both studies).
Companies that had less than five workers and were not active before 2018 (study conducted 2019) were removed from the sample (Kinne and Axenbeck, 2018: 9-12; Kinne and Lenz, 2019: 2-3, 6).

2020). ARGUS makes it possible to scrape all web page texts and their links (on-site and external links). Before scraping, the user must set the number of pages ARGUS should scan (more information is provided in Appendix 1). Scanning was done following a structural order, where the first scanned pages were in the main menu and subsequent pages in submenus, and so on. Kinne and Lenz (2019) argued that a 250-page limit should cover 90% of the firms, but they set the limit at 100 pages for their analysis. The idea behind that choice was that company information (e.g. "About us") should be in the first 100 pages, and there should only be product or service information on the subsequent pages. With 100 pages, scanning and filtering data is expected to be faster. Because Germany has many more firms than Estonia, it is understandable why Kinne and Lenz (2019) set a relatively low page limit. We raised this limit to 500 pages because we have fewer firms in Estonia. Theoretically, this should help machines understand product innovation better because the additional web pages are typically used for product and service descriptions.

The next stage had to be conducted before data can be fed to ML algorithms for training the model. This step, which is called as pre-processing phase, involves various steps, and its goal was converting the text to a machine-readable format. For that purpose, text was converted into a word vector in lowercase (as in Kinne and Lenz, 2019). Lowercase helps keep the word count lower and equalises words that computers may think are different (e.g. the words "paper" and "Paper" are two different words for the computer). Raw data was cleaned from elements that do not communicate anything, such as punctuation, numbers (dates were not needed), HTML tags, etc. Finally, there is a need to filter out common words called “stopwords”. Common words such as “I”, “we”, and “how” have very high frequency (also in Estonian), but they have no value because they do not add meaning to the phenomenon being searched. To lower the number of distinct words in the analysis, the text was lemmatised – every word was converted to its root form as in a dictionary. This is a natural language process that converts words to their root form (Gentzkow et al., 2019). For lemmatising, we used the ESTnltk Python package. After all the processes mentioned above were completed, we combined all the data in one file; see Table 1 for the sample data.

Table 1. Sample data after completing the data collection process (Python Pandas)

Registry number	Innovation indicators					Text from website
	Overall	Product	Process	Organi–sational	Marketing	
ABC0177	1	1	1	1	0	oskama pakkuma lahendus valik leiduma huvitav ...
ABC0684	1	1	1	0	0	aitama kodu ehitama katus vahetama vana uus te...
ABC0934	1	1	1	1	1	arveldus seonduv info iseteenindus kõnekeskus
ABC1615	1	0	1	1	0	teadma parool unustama võima meiliaadress saat...
ABC1941	1	1	1	1	0	sada aasta kondiitritoode tootja pakkuma toote...

Sources: Statistics Estonia CIS 2016 and 2018; compiled by the authors.
Note: Column "REG" has been changed in this view because we are not allowed to reveal the identities of the firms.

4. TESTED MODELS

Overall, there were usable web pages belonging to 1,290 firms. Their distribution on the basis of the firms' innovation indicators can be seen in Appendix 2. We observe the same number of innovative firms in the sample based on organisational innovation and marketing innovation. However, the correlation between the two non-technological innovation indicators is not exceptionally high (see Appendix 3). To counter problems due to unbalanced data; for example, in the case of overall innovation, we needed to use data sampling techniques in machine learning. SMOTE (Synthetic Minority Oversampling Technique) was the best method to implement in every test. This method creates values for minority classes and reduces the bias observed in the analyses, which helps achieve an equal sampling size. For example, for overall innovation, the new data size after using SMOTE would be 1,850 firms.[6] The data used for the models has been split into training and testing samples in the following ratios: 90/10, 80/20, and 70/30 (training/test) samples.

We mentioned earlier the basic fact that computers do not understand text as humans do. Thus, the text must be decoded for machines to understand it. We categorised these processes into two types that we label frequency-based and dimensional data methods. Frequency-based methods were considered the methods where word libraries were created, and machines replaced words with numbers in the sentences. The numbers could be an index from the library or a calculated value. The used frequency-based methods were term frequency (TF) and term frequency-inverse document frequency (TF-IDF). The outcome of these embeddings will be a matrix, where one side is a company's number, and the other side is words in the whole corpus. The dimensional data methods were Word to vector (Word2Vec) and Document to vector (Doc2Vec). The idea of those embeddings was to give every word a vector of values and compare them with each other to see what words are alike and what are not. The firm's "dimensional values" are calculated from the summary value of words. The main advantage of this method is that values are generated from the dimensional value of all words from the corpus. The output of those embeddings was a matrix with the companies' numbers in the rows and the dimensional values in the columns. The last value may vary depending on the number of dimensions; in our case, these were respectively 50, 100 and 300. Differences between these two embedding methods are the output of matrix size. The TF-IDF matrix size is determined by the number of firms and unique words, such as 1,290 firms times 126,000 unique words. This means it takes a short time to set up (about half a minute) but a lot of time to calculate values in the models (in the neural network, one epoch could take 10 to 20 minutes). In contrast, word2vec had a matrix output of 1,290 firms times 300 dimensional values, where the embedding part took some time (about 2 hours). However, during modelling phase, calculating was fast (in the neural network, one epoch could take 60 milliseconds).

To improve the scores in machine learning, stratified k-fold cross-validation was applied for the models with a *k* value of 10. Based on standard machine learning practices, we kept the value of *k* at 10, and we used it for further tests. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample (Kohavi, 1995). The parameter *k* refers to the number of the groups into which the analysed data sample is divided. After that, the groups will be tested one by one and put back in the sample. Other groups at the same time will be included in the training data. So, in the end, every group will be in the

[6] The number of 1,850 firms was reached by creating 560 additional companies artificially and adding these to the original sample of 1,290 firms. The number 560 comes from the difference in the number of innovative and non-innovative firms, respectively 925 and 365 (see Appendix 2.)

test data once and in the training data at other times. Stratified means that the sample is split in balance so that innovative (1) and non-innovative (0) would be in the same ratio in different groups.

An AUC-ROC score (Area Under the Curve - Receiver Operating Characteristics) analysis was conducted with every model. This analysis aims to find out the best validation model that would be a baseline for the future. An AUC-ROC is a probability curve, and it is one of the most crucial evaluation metrics for checking the performance of any classification model. It describes the model's capability to distinguish classes at various thresholds.

We employed various AI approaches (machine learning and neural network models) for the prediction tasks, some of which are black boxes in nature. Specifically, traditional and ensemble-based machine learning models were run to validate the model. These models were Logistic regression, Naive Bayes (Bernoulli, Multinomial, and Gaussian), AdaBoost, Support vector machine (SVM), XGBoost, LightGBM, and Random Forest. We tried two neural network models, namely, LSTM and GRU models, but these did not give us better results. Later, to test the Bidirectional Encoder Representations from Transformers (BERT) model, we needed to cut the text sample down for each firm (in some cases from over 100,000 words) to 512 words.[7] For that, we used text summarisation methods (first, frequency, and second, auto summarisation), where we used the 512 most important words for every company. From the OPUS (open parallel corpus), we added the Estonian subtitle corpus of subtitles collected from opensubtitles.org for the language model. But even with that, we did not obtain AUC-ROC scores above the threshold of 0.8 in any of the machine learning and neural network models. These tests were conducted with the web texts we had from all 1,290 firms. The overall process of our analyses is depicted in Figure 2.

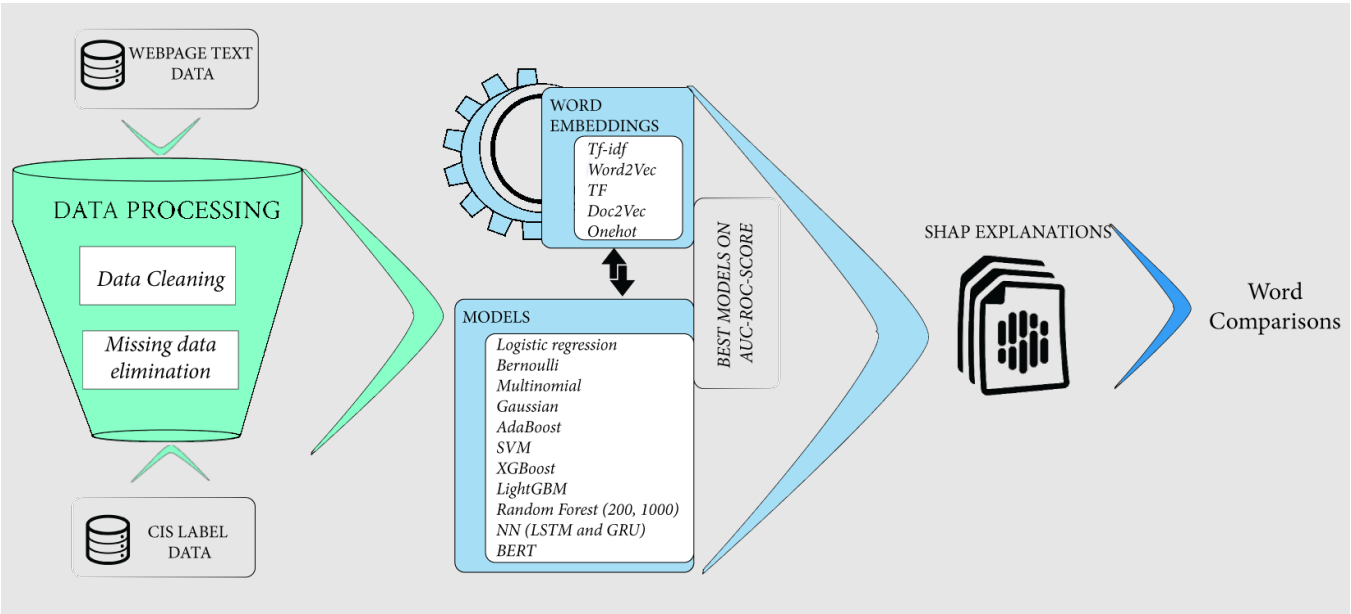


Figure 1. Overall process flowchart presenting our analysis

[7] BERT models are models that are trained two times. First, the model is trained like a language model, and second, it focuses on the phenomenon. In the language model, what is trained first classically has two variants: 512 and 1024 number of input words. This is a limitation because the phenomenon model must have the same maximum number of words as input.

We discovered that innovative firms have more words on average on their web page than non-innovative firms in our dataset (see Appendix 4). This gives machines more opportunity to understand innovation better because there are more words to remember in relation to innovation. To be sure about that conclusion, we also looked into unique word count, which gave the same results. This gave us the idea to look further at the size of firms and their webpage differences and whether firm size might be driving these differences. Firm size was measured here by the number of employees. Our data did not show that firms with more employees had larger web pages. One explanation for that is that we had many firms from the retail sector, and they had very large websites but few employees. On the other hand, there were manufacturers who had a lot of employees but because they were mainly operating in one or another very specific area, they did not have many words on their web pages. However, our data showed that if a firm had more than 1,000 employees, then they had more than 10,000 words on their web page. If we look at the web pages with very high word counts, such as those with 100,000 or more words (in 500 pages on the web page), this group included 22 innovative and 14 non-innovative firms. Furthermore, we looked at web pages with more or less than 10,000 words; these distributions are in Table 2.

Filtering our usable data by the number of words, we discovered that firms with more than 10,000 words on their web pages are giving better results compared to firms with less than 10,000 words. We separated the web pages into two categories: pages with less than 10,000 words and pages with more than 10,000 words. The first category had an AUC-ROC score near 0.60 in the mentioned models, but when testing only web pages with more than 10,000 words, the AUC-ROC scores were above 0.80 for the overall innovation and marketing innovation categories. Alternatively, we tried to filter data using the 5,000, 7,500, 15,000, and 20,000-word count points, but the best results were obtained with the 10,000-word count. The number of firms that had more than 10,000 words was 388 (see

Table **2**), split into training and testing samples with shares of 90% and 10%, respectively. We also ran tests with 80/20% and 70/30%, but their AUC-ROC scores were lower than when using a 90/10% split.

In Table 3, we report best AUC-ROC scores for each innovation type. Embeddings that achieved the top results were tf-idf and word2vec. Product, process, and organisational innovation had too low AUC-ROC scores (0.7388, 0.7555 and 0.7309), and we had to discard them. In theory, we had expected that scraping more pages on the company's webpage would give us better product innovation results. But in the end, it gave us the second-worst results (the worst was for predictions of organisational innovation). One explanation for that could be that it was sufficient to scan just 500 pages in the case of manufacturing companies. On the other hand, retailers had a lot of products described on their web pages; still, their lack of innovativeness may have signalled a large number of words associated with non-innovation, which is why it was hard for the machine to recognise patterns in words associated with product innovation.

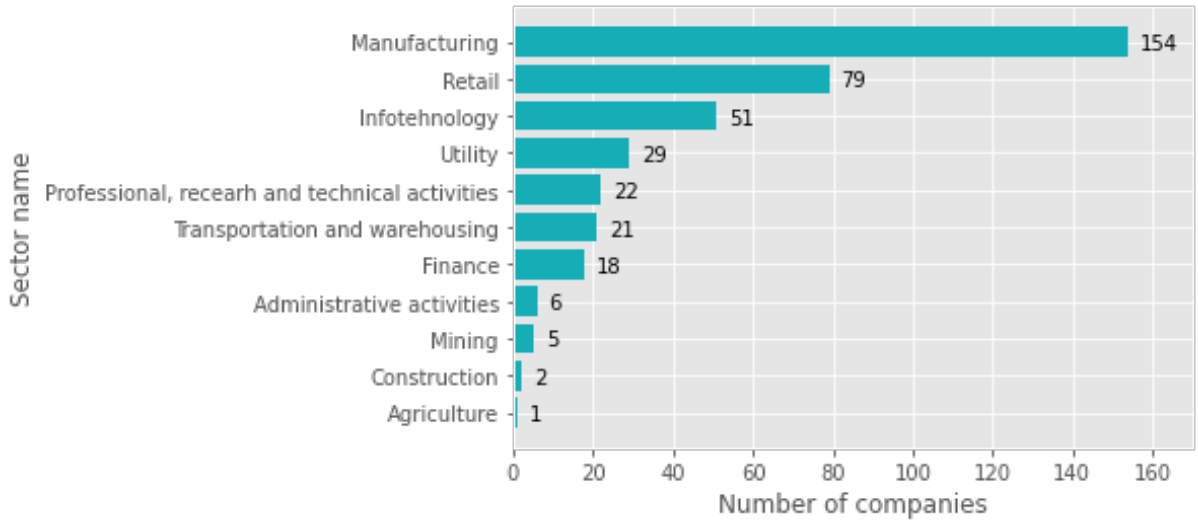


Figure 2. Sectoral distribution of 388 firms
Sources: Statistics Estonia CIS 2016 and 2018; compiled by the authors.

Table 2. Distribution of companies by innovation type and number of words on their web pages

Innovation type	Innovative (Yes/No)	Less than 10,000 words on webpages		More than 10,000 words on webpages	
		No. of companies	Share	No. of companies	Share
Overall innovation	Yes	601	66.6%	324	83.5%
	No	301	33.4%	64	16.5%
Product innovation	Yes	335	37.1%	240	61.9%
	No	567	62.9%	148	38.1%
Process innovation	Yes	529	58.6%	290	74.7%
	No	373	41.4%	98	25.3%
Organisational innovation	Yes	313	34.7%	204	52.6%
	No	589	65.3%	184	47.4%
Marketing innovation	Yes	313	34.7%	241	62.1%
	No	589	65.3%	147	37.9%
No. of firms with web pages		902		388	

Source: compiled by the authors.

Table 3. Best models for different types of innovation

Model	Split	Best AUC-ROC-score
Overall Innovation		
XGBoost Tf-idf with min_word_count=5 and ngram= 2:	90/10	0.8131
XGBoost with Tf-idf default:	90/10	0.8080
LightGBM with Tf-idf with min_word_count=5 and ngram= 2:	90/10	0.8030
Product Innovation		
Random Forest with estimators 1000 with default Tf-idf: **	90/10	0.7388
Process Innovation		
XGBoost with Tf-idf with min_word_count=5 and ngram= 2:	90/10	0.7555
Organisational Innovation		
Bernoulli with word2vec with dimensions 300 and window=2*:	90/10	0.7309
Marketing Innovation		
Random Forest with estimators 4150, with default Tf-idf:	90/10	0.8285
Random Forest with estimators 2000, Tf-idf with min_word_count=5 and ngram= 2:	90/10	0.8257

Note. *window=2 in word2vec means that we used neighbouring words combinations. ** dataset with 1290 companies. All models with stratified k-fold=10 and SMOTE).
Source: Compiled by the author.

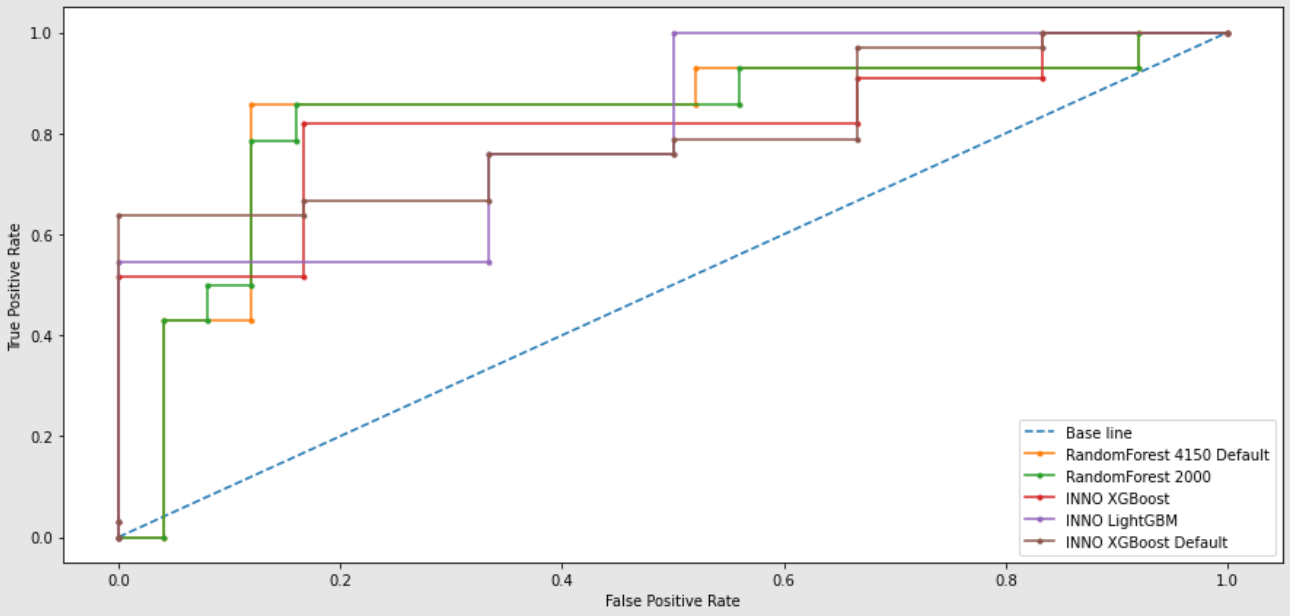


Figure 3. ROC-AUC table of evaluated models
Source: Compiled by the authors.

In the end, there were five evaluated machine learning models with AUC-ROC scores higher than 0.8. All results used the 90/10 training/test distribution. The first three models were with overall innovation: two XGBoost models with tf-idf embeddings (default and with embedding limitations) and a LightGBM model. The limitations with the tf-idf models were that the minimum word count in the corpus had to be five, and we looked at the combinations of words with their neighbours (ngram=2). The best of them was the XGBoost model with limited embeddings. Better results were obtained with marketing innovation, and there we found two Random Forest optimised models with acceptable results. The first was the Random Forest model with 4,150 n-estimators and default tf-idf embeddings. The second in marketing innovation was the Random Forest model with 2,000 n-estimators and limited embeddings, likewise for overall innovation. We present the AUC-ROC results for these evaluated models in Figure 4, where we can see that marketing innovation models perform slightly better than the overall innovation models.

Our machine learning models only predicted innovation well in the case of firms with more than 10,000 words on their first 500 web pages. Therefore, we decided to learn more about the words that algorithms considered related to overall innovation (technological or non-technological) and marketing innovation. The main goal was to confirm that machines have captured the most essential words related to innovation and that the models explain the correct phenomenon. At first sight, the models seem to have good prediction performance, but that might not be because the machines have "understood" the innovation phenomenon. Furthermore, in our used dataset, the manufacturing sector was over-represented (see Figure 3), which may have affected the outcomes of our models.

5. DECRYPTING MODELS WITH SHAP

Many ML models although being efficient in prediction suffer from the "the black box" problem. In other words, interpreting which features helped the model in concluding a decision is often not clear. In many real-world applications, it is important to understand the inner working of the model or how the model has concluded a certain decision. For instance,

if a doctor uses machine learning to diagnose a patient, the doctor also wants to know why the machine gave a particular prediction. Understanding essential features for predicting the output can help reinforce human trust in machine learning approaches as humans can verify if machine learning approaches rely on (logical) words in prediction tasks. In our case, every word is a feature (in total, there are about 70,000–120,000 features). Therefore, the weights for each word are much lower. In our case, the best words in the models had about 2% weighting, which was sufficient to change the whole model.

In this study, for model's explainability we used SHAP to explain the output of machine learning models. SHAP can be used to explain the output of any machine learning model; it uses the classic Shapley values from game theory and their related extensions (hence the name SHAP) (see for further explanations Lundberg and Lee, 2017). In simple terms, it is an algorithm that intersects the model at its feature level and looks at how they are the same in every sample. Results are usually represented on easy-to-read figures (plots).

Without going further into the technical details, the Shapley values are obtained when the predicted probability of the particular type of innovation is decomposed into contributions from particular predictors (in our case, particular words). One calculates here, how much a particular variable (word) adds to the predicted probability of innovation; that is, the predicted probability calculated with the particular word included minus the probability without using the particular word, and so sequentially over all the possible subsets of the words (or in more technical terms, the average of the marginal contributions of individual words across all permutations of words). Therefore, Shapley values measure how much individual predictors or variables (in our case, words) drive the prediction of the machine learning model. Regarding examples of the application of SHAP in economics, Bluwstein et al. (2020) when applying machine learning to predict financial crises used the Shapley value framework to understand which individual variables helped to predict the crises, and they argued that at least in this literature their study was the first to address “the black box” critique of machine learning approaches. They also emphasized that the latter is also important in ensuring transparency and accountability if using the indicators generated by machine learning in policy making. For a further intuitive explanation of the value of SHAP, see Kuo (2019). We used the Python’s SHAP package to investigate at every evaluated model's top impact words.

First, we tested overall innovation models with AUC-ROC scores over 0.80, but SHAP showed a different view (see Appendix 5, 6 and 7). We found words that we could not relate to "overall innovative" companies. We evaluated the models and their top words in 4 categories, see Table 4. We tried to set rules to read everything, positive and negative cases. Firstly, regarding the concentration of samples: if a word is in the text, it is highlighted in red (high), and if not, it is indicated by a blue circle (Appendix 5, 6, 7). So, if reds are easily separated, it is distinguishable (checkmark); if not, it is marked unnoticeable (cross-mark). Second, we studied the direction of the word-effect on model (positive or negative). Third, we studied the distance between the positive and negative cases. If the SHAP value was close to 0, the word is considered to have a weak impact in the model, and in other cases, the impact is considered to be strong.

Table 4. Word analyses for overall innovation models

Words 3 time

Model		LightBGM				XGBoost, Tf-idf with limits				XGBoost, default Tf-idf			
Eng words	*Est words*	*Concentration*	*Direction*	*Distance*	*Value*	*Concentration*	*Direction*	*Distance*	*Value*	*Concentration*	*Direction*	*Distance*	*Value*
burn	põlema	✓	−	strong	-1,5	✗	−	strong	-0,5	✗	−	strong	-0,5
unsurpassable	ületamatu	✓	−	strong	-1,5	✗	−	weak	-0,2	✗	−	weak	-0,2
taking	võtmine	✓	+	weak	0,5	✓	+	strong	0,4	✓	+	strong	0,4
confirm	kinnitama	✓	+	weak	0,25	✓	+	strong	0,4	✓	+	strong	0,4
private client	eraklient	✓	−	weak	0,5	✓	+	weak	0	✓	+	weak	0
within	piires	✗	−	weak	-0,5	✗	+	weak	0	✓	+	weak	0,1

Words 2 time

Eng words	*Est words*	*Concentration*	*Direction*	*Distance*	*Value*	*Concentration*	*Direction*	*Distance*	*Value*	*Concentration*	*Direction*	*Distance*	*Value*
fast delivery	*kiire tarne*	✓	−	*strong*	*-0,75*	✗	−	*strong*	-				
secure device	*tagama seade*	✓	−	*strong*	*-1,5*	✗	−	*strong*	-				
ambition	*ambitsioon*	✗	+	*weak*	*0,5*	✗	+	*weak*	-				
history	*ajalugu*	✗	+	*weak*	*0,5*	✗	+	*weak*	-				
behavior	*käitumine*	✓	+	*strong*	*1,75*					✓	+	*strong*	*0,5*
designing	*projekteerimine*	✓	−	*strong*	*-1*					✓	−	*strong*	*-0,3*
run for	*kandideerima*	✓	+	*strong*	*0,3*					✓	+	*weak*	*0,5*
delay visit	*viibima*					✓	+	*weak*	*0,1*	✓	+	*weak*	*0,1*
call	*helistama*					✓	+	*weak*	*0,1*	✓	+	*weak*	*0,2*

Words 1 time

Eng words	*Est words*	*Concentration*	*Direction*	*Distance*	*Value*	*Concentration*	*Direction*	*Distance*	*Value*	*Concentration*	*Direction*	*Distance*	*Value*
help	*aitama kaasa*	✓	+	*strong*	*0,5*								
planning	*planeering*	✓	−	*weak*	*-0,5*								
first stage	*esimene etapp*	✓	+	*strong*	*1*								
bottom	*põhi*	✓	+	*weak*	*0,25*								
measure	*meede*	✓	+	*weak*	*0,25*								
programming	*programmeerimine*	✓	−	*strong*	*-1*								
account	*konto*	✓	−	*weak*	*-0,25*								
net	*võrk*					✓	+	*weak*	*0,1*				
juniper	*kadakas*					✗	+	*weak*	*0,1*				
execute necessary	*teostama vajalik*					✗	−	*strong*	*-0,2*				
may use	*võima kasutama*					✗	−	*strong*	*-0,2*				
goal	*siht*					✓	+	*strong*	*0,2*				
glasses	*prill*					✓	+	*strong*	*0,05*				
fuel	*kütus*					✓	+	*weak*	*0,05*				
comfortable	*mugavam*					✓	+	*weak*	*0,1*				
pump	*pump*									✗	−	*weak*	*0,2*
two-sided	*kahepoolne*									✓	+	*weak*	*0,1*
equipment	*varustus*									✗	-	-	-
made	*made*									✗	−	-	-
dining room	*söögituba*									✗	−	-	-
measuring point	*mõõtepunkt*									✗	−	-	-
living room	*elutuba*									✗	−	-	-
maintenance contract	*hooldusleping*									✗	−	-	-
maximum	*maksimaalne*									✓	+	*weak*	*0,1*

Source: Compiled by the authors.
Notes. The analysis is conducted using the overall innovation models from Table 3.

In Table 4, we report words that were among the top 20 words in three innovation models. Six words were found in all of the models. Words like "*burn*" and "*unsurpassable*" were clearly distinguishable and had a negative effect on the innovation model. Words like "*taking*",

"*confirm*", and "*private client*" had a positive effect on the innovation model. Interestingly, these words had the same direction in every model. A word such as *"within"* was weakly distinguishable only in one model. In overall innovation, some of the top words are more likely related to manufacturing companies, like "*burning"* (põlema). The problem we may encounter here (and what we expected) is that these models may describe and predict which company is a manufacturing company and which one is not. Of course, overall innovation is the combination of all innovation types (two technological and two non-technological), and it can be a difficult phenomenon to describe. Despite that, there is also the possibility that these words are related to overall innovation.

Second, when predicting marketing innovation, we found a model using many words related to innovation and marketing. Most of the marketing innovation words were distinguishable, whereby a word had a positive effect if it was a top word on the web pages and a negative effect if not. Only Random Forest with 2,000 n-estimators and tf-idf limits had one word in the top 20 that had a negative or a positive effect in some cases. This word was "*requirement claim*" (in Estonian, *"nõue"*). Words that were in both models were "*customer support*" (in Estonian, "*klienditugi*"), "*development*" ("*arendus"*), "*marketing*" ("*turundus"*), "*export*" ("*eksport"*), "*automatically*" ("*automaatselt"*), "*launch/start*" ("*käivitama"*), "*trying*" ("*proovima"*) and "*tool*" ("*vahend"*) (see figures 5 and 6). We can see that these words affect models positively if included and negatively if not. That is expected as marketing innovation can be a new product/service packaging, new media channels, or marketing methods. These last two models' SHAP plots confirm that these models indeed prioritised words related to marketing or innovation. With overall innovation, we did not see a clear understanding from the model, while in the case of marketing innovation, we achieved a model that understands innovation. We find the latter result rather encouraging, as that was achieved despite the relatively small sample size.

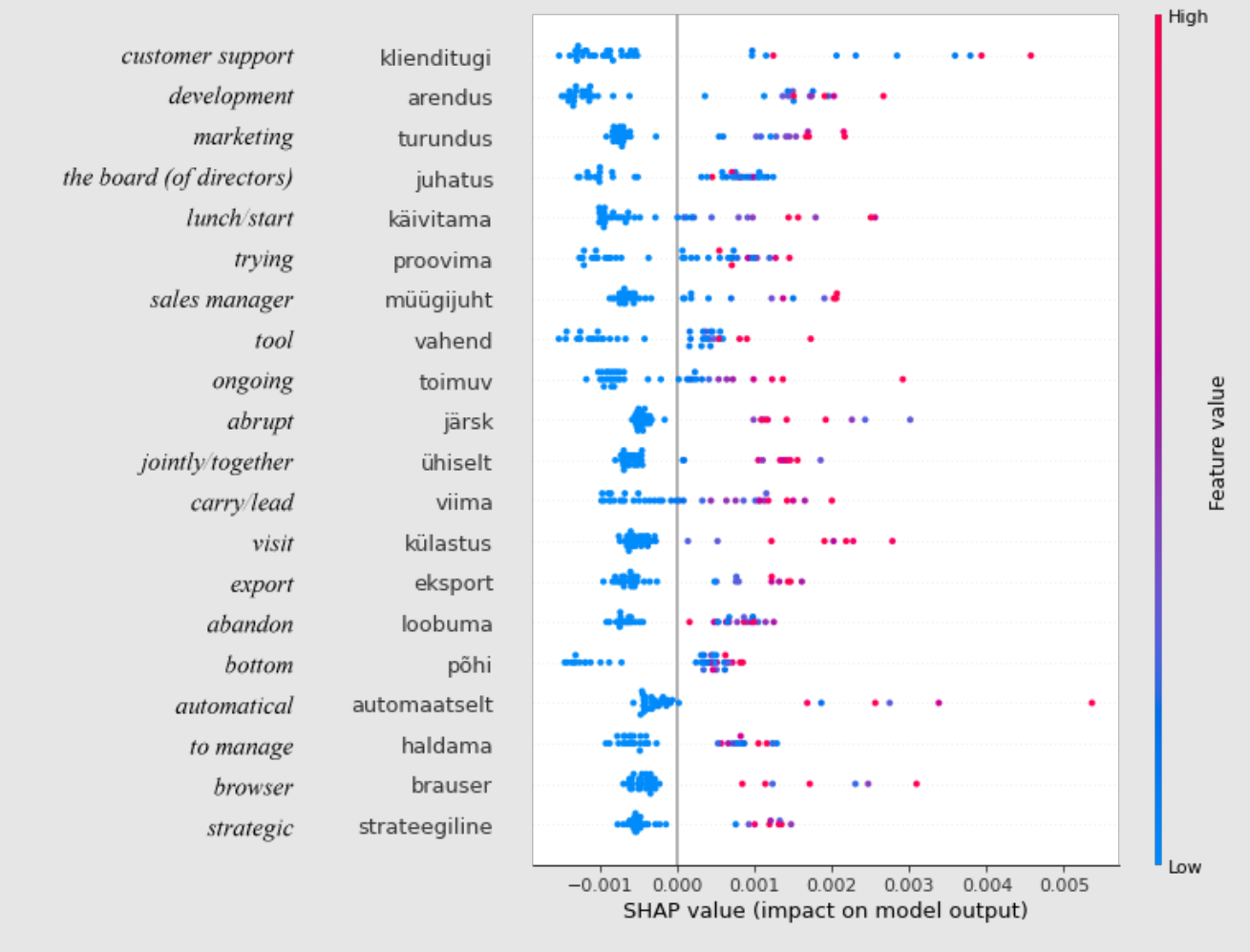


Figure 4. SHAP summary plot with marketing innovation (1st acceptable model)

Note. Random Forest 4,150 with TF-IDF default, unique word count 162,296.

The plot has been made so that variables (in our case, words) are ranked from above in descending order, the first being the most important etc. The horizontal location indicates whether the particular word is associated with a lower or higher prediction of innovation. The colours indicate low (blue) or high (red) values for the observation; that is, for these words we have many observations with a small negative contribution, but also observations with much larger positive associations. Therefore, "customer support" has a high and positive impact on the predicted marketing innovation, the "high" indicated by the red colour and "positive" indicated by its position on the X axis (see also Kuo 2019).

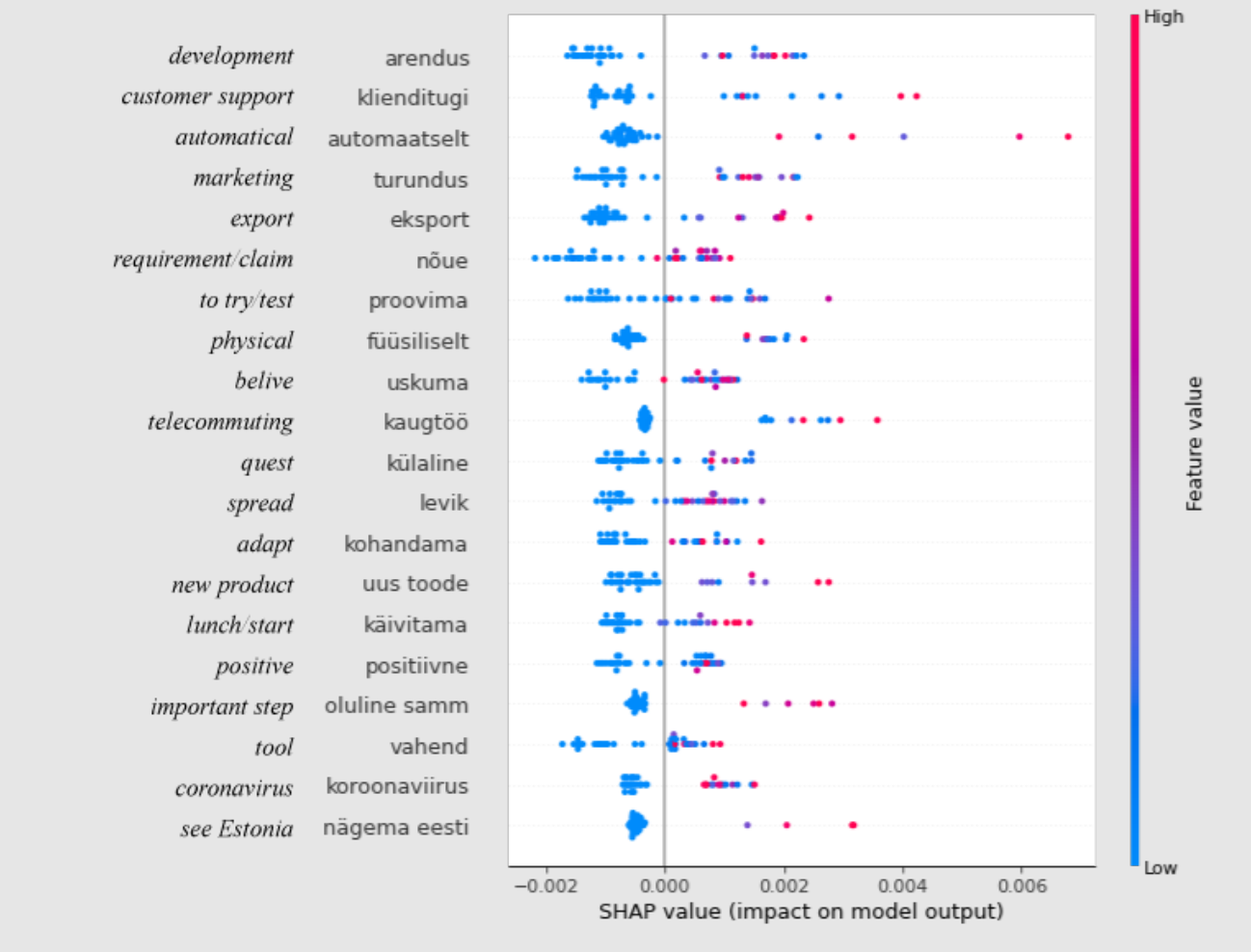


Figure 5. SHAP summary plot with marketing innovation (2nd acceptable model)
Note. Random Forest 2000, TF-IDF with min_word_count=5 and ngram= 2, unique word count 76,043.

In marketing innovation, the top SHAP words were indeed related to marketing, and it is possible to teach the machine to learn about marketing innovation. These are the top 20 words from models with over 162,296 unique words. We must remember that these companies' web pages have a lot of text, and these features (words) are still selected to be the best in the models. We also looked at overall innovation and found the top words in these models. But because these words were those that affected the model positively/negatively, and some words did not make sense, then we concluded that overall innovation as a phenomenon might be too complex to describe using the top 20 words from our models. It was even possible that our models tried to find and predict a company belonging to the manufacturing sector because many words were related to this sector. Our marketing innovation results were good examples of how machines understand the words that companies use when communicating on their web pages. In their communication, some words were more common for innovative companies.

6. CONCLUSIONS

This paper explores the possibility of predicting innovation in firms based on their internet website texts and using the machine learning techniques. We use website data from firms in Estonia, combined with firm-level innovation survey data, to label firms as innovative or non-

innovative. The key novel addition to the literature, and in particular to Kinne and Lenz (2019) and Axenbeck and Breithaupt (2021), is the finding that web mining is more applicable for predicting marketing innovation compared to the other three core types of innovation (product, process and organisational innovation) or innovation in general (any of the four types of innovation).

More specifically, our key contribution to the literature concerns the role of the features of websites in predicting innovation. We add to the literature by analysing textual data (words) on websites that help predict innovation. AI approaches, specifically deep learning models, considered excellent for dealing with textual data, are also black-box in nature. This means that it is not easy to identify what features these models rely on for concluding their predicted outcome. To overcome the non-transparency of these models, we used explainable AI methods. Specifically, we use SHAP (*SHapley Additive exPlanations)* to understand the importance of the feature (specific words). The results of SHAP show that the identification and prediction of marketing innovation were related to marketing-related words on firms' websites: such as, for example, the words "development", "testing", and "marketing". This suggests that our models explain innovation broadly in a manner similar to how humans would. We also find it to be an encouraging result that the approach we used can be applied to small-sized countries that have relatively smaller samples of firms available for testing. However, limitations were applied to obtain the positive results. For example, firms must have a substantial amount of text on their web pages: in our case, more than 10,000 words per 500 web pages.

As a significant result and in contrast to the prediction of marketing innovation, there were much less clear results on relevant words on websites in the case of predicting the other key innovation indicators. Our results confirm that the suitability of web scraping combined with deep learning approaches for providing cost-effective, granular and timely firm-level information on innovation indicators varies on the basis of the type of innovation.

We further note that compared to the results from SHAP, for the prediction of innovation indicators based on the standard machine learning and neural network approaches, the size of the sample turned out to be small. This is not necessarily surprising, as Estonia is one of the smallest countries in the European Union and has, accordingly, a relatively smaller number of firms compared to larger countries. One particular problem in the analysis was also that firms' webpages were different from each other, some web pages had about 1,000 words, and some had more than 100,000 words. Running our algorithms on a sample that included both firms with a relatively small number and a large number of words did not give many useful results. Filtering out firms with more than 10,000 words on their web page – there were 388 such companies – gave us for the first time AUC-ROC scores above 0.8. However, these results were sensitive (non-robust) to how the sample of firms was split between the training and test samples; in most cases, the AUC-ROC scores were below 0.8, and well-performing models were obtained only in one case.

The results of our study could be of interest to researchers and organisations that study firm-level innovation activities, their determinants and effects. Our research managed to find at least one working method out of several machine learning approaches that tried to predict different innovation output indicators. Therefore, we can conclude that it is possible to predict marketing innovation in companies using the contents of their websites.

As text data is very rich and has very high dimensionality, there is much room for additional research in this area. For example, our study has shown that data from different countries may

contribute to the literature by investigating how the results might be sensitive to local and national contexts. In our case, the latter concerned the limited size of the website contents available for the analysis. Therefore, in an ideal case, our analysis should be replicated in other countries to isolate the possible effects of the national contexts on the results. One natural suggestion following that could be adding information from the social media accounts of firms, as, for example, the smaller companies may use these more instead of websites. Yet, as we have shown in our website-based analysis, that might not help if the amount of text available for analysis in a firm is relatively limited.

More general recommendations on using website data to predict firm innovativeness (and other aspects of firm performance) should rely on more extensive evidence than is currently available. While our study was exclusively based on website contents, the mining and analysing of web structure (hyperlinks) could provide many possibilities for exploring various aspects of innovation. The list of such research topics may include the use of information for innovation from different sources (other firms, customers, suppliers), innovation cooperation (e.g. with universities), the choice between closed and open modes of innovation, and the complementarities of different sources. Given the above, we hope that the literature will develop significantly over the following years.

REFERENCES

ARGUS (2020. Argus: Automated Robot for Generic Universal Scraping, datawizard1337, Github, [https://github.com/datawizard1337/ARGUS] Last accessed: 11.06.2020.

Ashouri, S., Suominen, A., Hajikhani, A., Pukelis, L., Schubert, T., Türkeli, S., Van Beers, C., Cunninghum, S. (2021). "Indicators on firm level innovation activities from web scraped data" Available at SSRN: https://ssrn.com/abstract=3938767 or http://dx.doi.org/10.2139/ssrn.3938767. Last accessed: 8.10. 2021.

Axenbeck, J., Breithaupt, P., (2021). "Innovation indicators based on firm websites—Which website characteristics predict firm-level innovation activity?." *PLOS ONE* 16(4): e0249583. https://doi.org/10.1371/journal.pone.0249583

Baesens, B., Viaene, S., Van den Poel, D., Vanthienen, J., Dedene, G. (2002). "Bayesian neural network learning for repeat purchase modelling in direct marketing". European *Journal of Operational Research*, 138(1), pp.191-211.

Bejger, S., Elster, S. (2020). "Artificial Intelligence in economic decision making: how to assure a trust?". *Economics and Law*, 19(3), pp.411-434.

Bluwstein, K., Buckmann, M., Joseph, A, Kang, M., Kapadia, S., Şimşek, Ö. (2020). "Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach". Bank of England Staff Working Paper No. 848

Bracke, P., Datta, A., Jung, C. Sen, S. (2019). "Machine learning explainability in finance: an application to default risk analysis". Bank of England Staff Working Paper No. 816.

Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J. (2021). "Explainable machine learning in credit risk management". *Computational Economics*, 57(1), pp.203-216.

Choudhary, M.A., Haider, A. (2012). "Neural network models for inflation forecasting: an appraisal". *Applied Economics*, 44(20), pp.2631-2635.

e-Äriregister. (2021) Statistika. [https://ariregister.rik.ee/est/statistics/charts/chart_company_all/2021]. Last accessed: 04.10.2021.

Fisher, J., Craig, A., Bentley, J. (2007). "Moving from a web presence to e-commerce: The importance of a business—Web strategy for small-business owners". *Electronic Markets*, 17(4), pp.253-262.

Gentzkow, M., Kelly, B., Taddy, M. (2019). "Text as data". *Journal of Economic Literature*, 57(3), pp. 535-574.

Gök, A., Waterworth, A., Shapira, P. (2015). "Use of web mining in studying innovation". *Scientometrics*, 102(1), pp. 653-671.

Hadji Misheva, B., Hirsa, A., Osterrieder, J., Kulkarni, O., Fung Lin, S., (2021). "Explainable AI in Credit Risk Management". Credit Risk Management. https://doi.org/10.48550/arXiv.2103.00949

Hoepner, A.G., McMillan, D., Vivian, A. Wese Simen, C. (2021). "Significance, relevance and explainability in the machine learning age: an econometrics and financial data science perspective". *The European Journal of Finance*, 27(1-2), pp. 1-7. DOI: 10.1080/1351847X.2020.1847725

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.

Kinne, J., Lenz, D. (2019). "Predicting innovative firms using web mining and deep learning". ZEW Discussion Paper No. 19-01.

Kohavi, R. (1995). "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", International Joint Conference on Artificial Intelligence (IJCAI), Stanford University, 1995, pp. 1-7.

https://www.researchgate.net/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection

Krüger, M., Kinne, J., Lenz, D., Resch, B. (2020). "The Digital Layer: How Innovative Firms Relate on the Web". ZEW Discussion Paper No. 20-003. https://ssrn.com/abstract=3530807

Kuo, C. (2019), "Explain Your Model with the SHAP Values". Towards Data Science. https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d

Levenberg, A., Pulman, S., Moilanen, K., Simpson, E., Roberts, S. (2014). "Predicting economic indicators from web text using sentiment composition". *International Journal of Computer and Communication Engineering*, 3(2), pp. 109-115.

Lundberg, S. M., Lee, S-I. (2017) "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. https://arxiv.org/abs/1705.07874

Mairesse, J., Mohnen, P. (2010). "Using innovation surveys for econometric analysis". CIRANO-Scientific Publication, (2010s-15), pp. 1–40.

Mirtsch, M., Kinne, J., Blind, K. (2021), "Exploring the Adoption of the International Information Security Management System Standard ISO/IEC 27001: A Web Mining-Based Analysis". IEEE Transactions On Engineering Management, Vol. 68, No. 1, pp. 87-100.

Nagaoka, S., Motohashi, K., Goto, A. (2010). "Patent statistics as an innovation indicator". In Handbook of the Economics of Innovation, North-Holland, Vol. 2, pp. 1083-1127.

OECD (2015), The Future of Productivity, OECD Publishing, Paris, https://doi.org/10.1787/9789264248533-en.

OECD/Eurostat (2005), Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data, 3rd Edition, The Measurement of Scientific and Technological Activities, OECD Publishing, Paris, https://doi.org/10.1787/9789264013100-en.

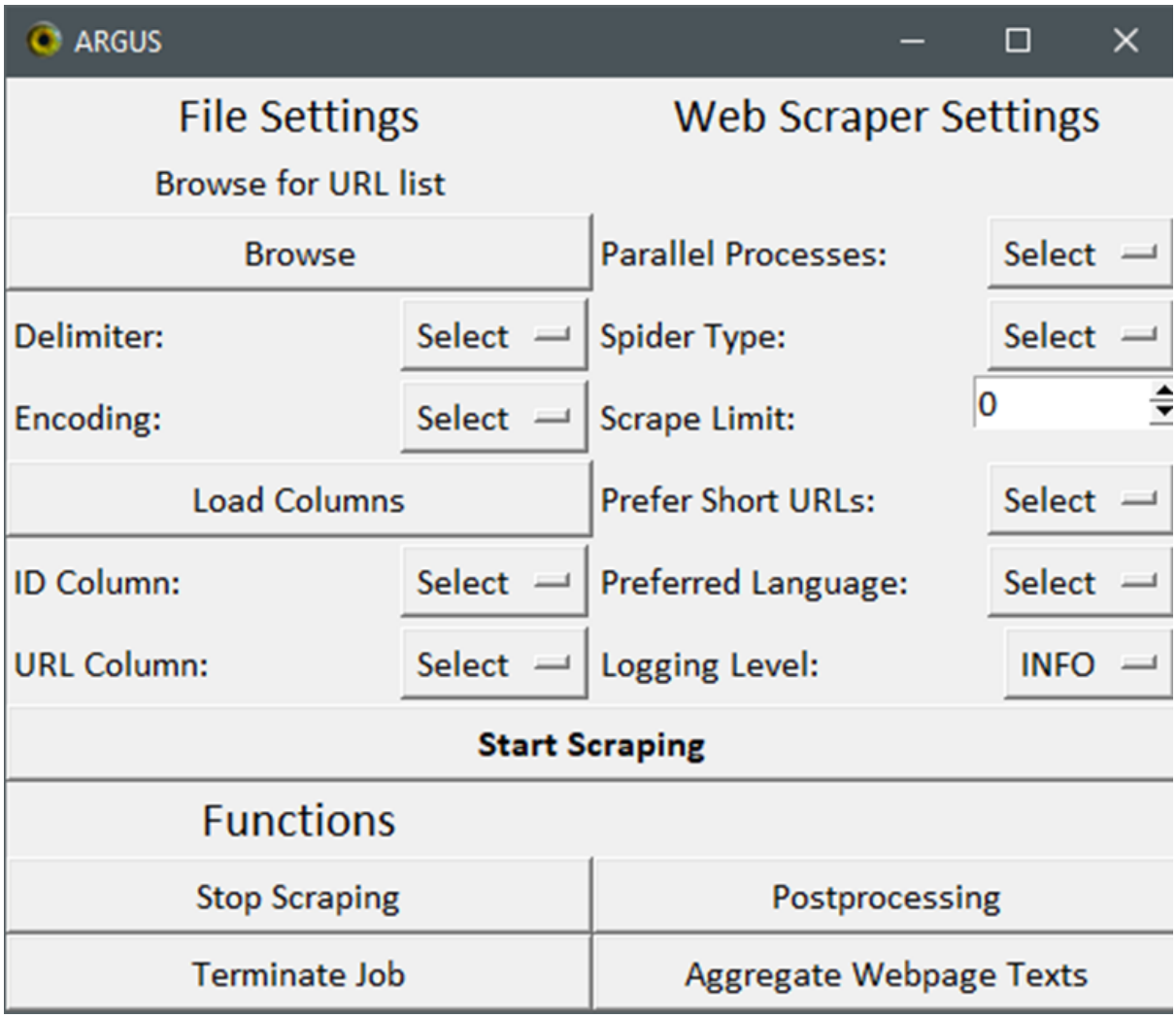
Pukelis, L., Stančiauskas V. (2019). "The Opportunities and Limitations of Using Artificial Neural Networks in Social Science Research". Politologija, 94(2), pp. 56-80. https://doi.org/10.15388/Polit.2019.94.2

Smith, K. (2009). "Measuring Innovation". In The Oxford Handbook of Innovation, pp. 148–173.

Wanto, A., Damanik, I.S., Gunawan, I., Irawan, E., Tambunan, H.S., Sumarno, S. and Nasution, Z. M., (2018). "Levenberg-Marquardt Algorithm Combined with Bipolar Sigmoid Function to Measure Open Unemployment Rate in Indonesia". 10.31227/osf.io/u5fhm

ZEW (2022). Mannheim Innovation Panel: Innovation Activities of Enterprises in Germany. [https://www.zew.de/en/research-at-zew/mannheim-innovation-panel-innovation-activities-of-enterprises-in-germany]. Last accessed: 18.02.2022.

Appendix 1. ARGUS user interface



Sources: ARGUS: Automated...: 2020

Appendix 2. The distribution of companies across groups of innovation within combined CIS2016 and CIS2018.

Group of companies	Innovative (Yes/No)	No of companies	Share of companies
Overall innovation	Yes	925	71.7%
	No	365	28.3%
Product innovation	Yes	715	55.4%
	No	575	44.6%
Process innovation	Yes	819	63.5%
	No	471	36.5%
Organisational innovation	Yes	773	60%
	No	517	40%
Marketing innovation	Yes	773	60%
	No	517	40%
The overall no. of companies		**1,290**	**100%**

Source: Compiled by the author

Appendix 3. Correlation coefficients of the main innovation indicators of the Estonian CIS2016 and CIS2018 data

	Overall innovation	**Product innovation**	**Process innovation**	**Organisational innovation**	**Marketing innovation**
Overall innovation	1.000				
Product innovation	0.566	1.000			
Process innovation	0.765	0.411	1.000		
Organizational innovation	0.468	0.402	0.481	1.000	
Marketing innovation	0.347	0.307	0.295	0.444	1.000

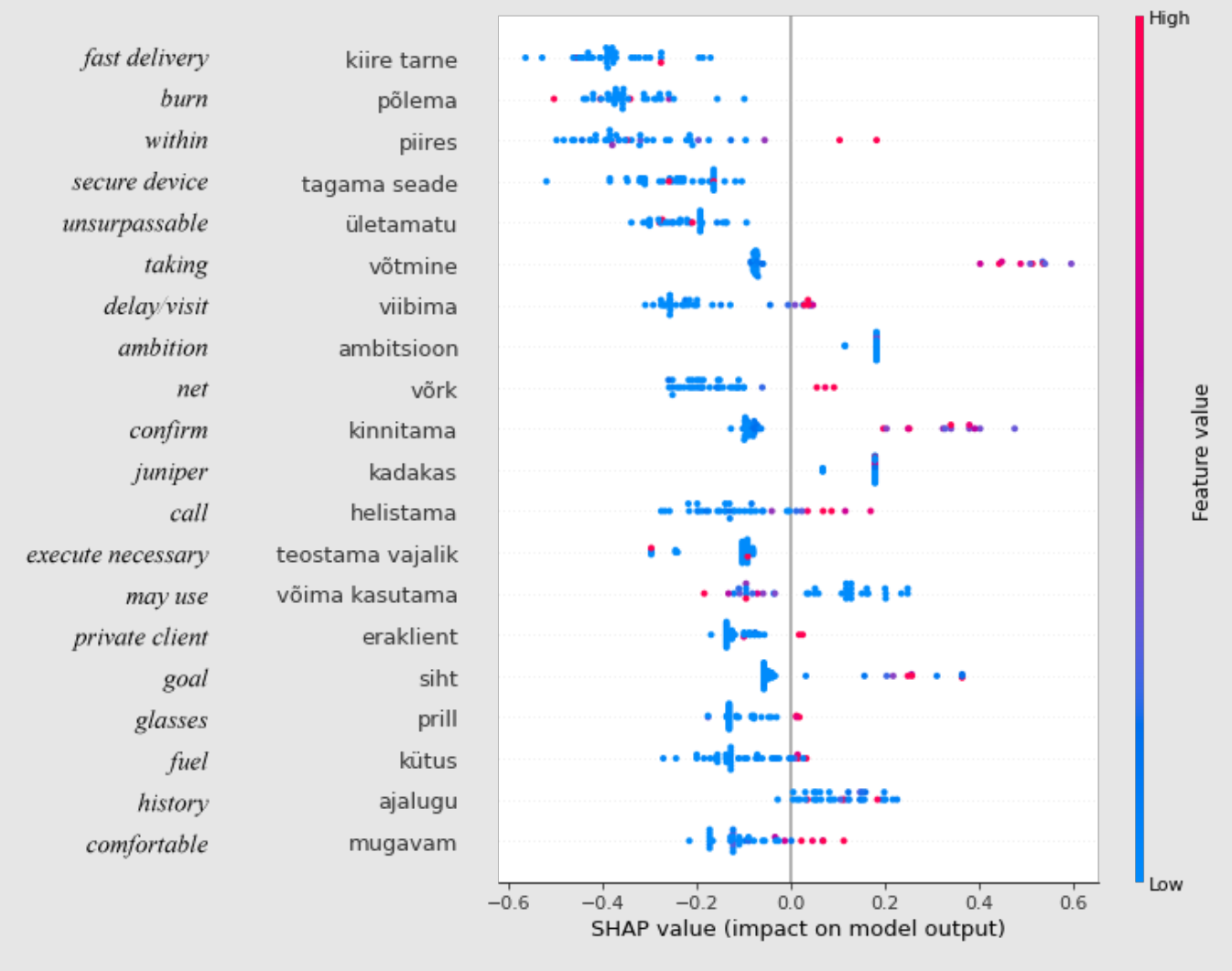
Source: Compiled by the author

Appendix 4. Word count averages on companies' web pages by innovation type

Innovation type	Innovative (Yes/No)	All companies (1,290 companies)		Companies with at least 10,000 words on web pages (388 companies)	
		Average	St. Dev.	Average	St. Dev.
Overall innovation	Yes	18,588	52,945	41,092	37,008
	No	9,696	36,401	34,742	31,520
Product innovation	Yes	23,046	62,771	41,701	36,527
	No	10,463	33,074	37,357	35,622
Process innovation	Yes	18,639	54,728	40,204	37,405
	No	11,608	36,564	39,990	35,849
Organisational innovation	Yes	20,944	55,907	41,402	36,461
	No	12,813	43,486	38,539	35,997
Marketing innovation	Yes	20,944	55,907	47,758	40,850
	No	12,813	43,486	35,339	32,237

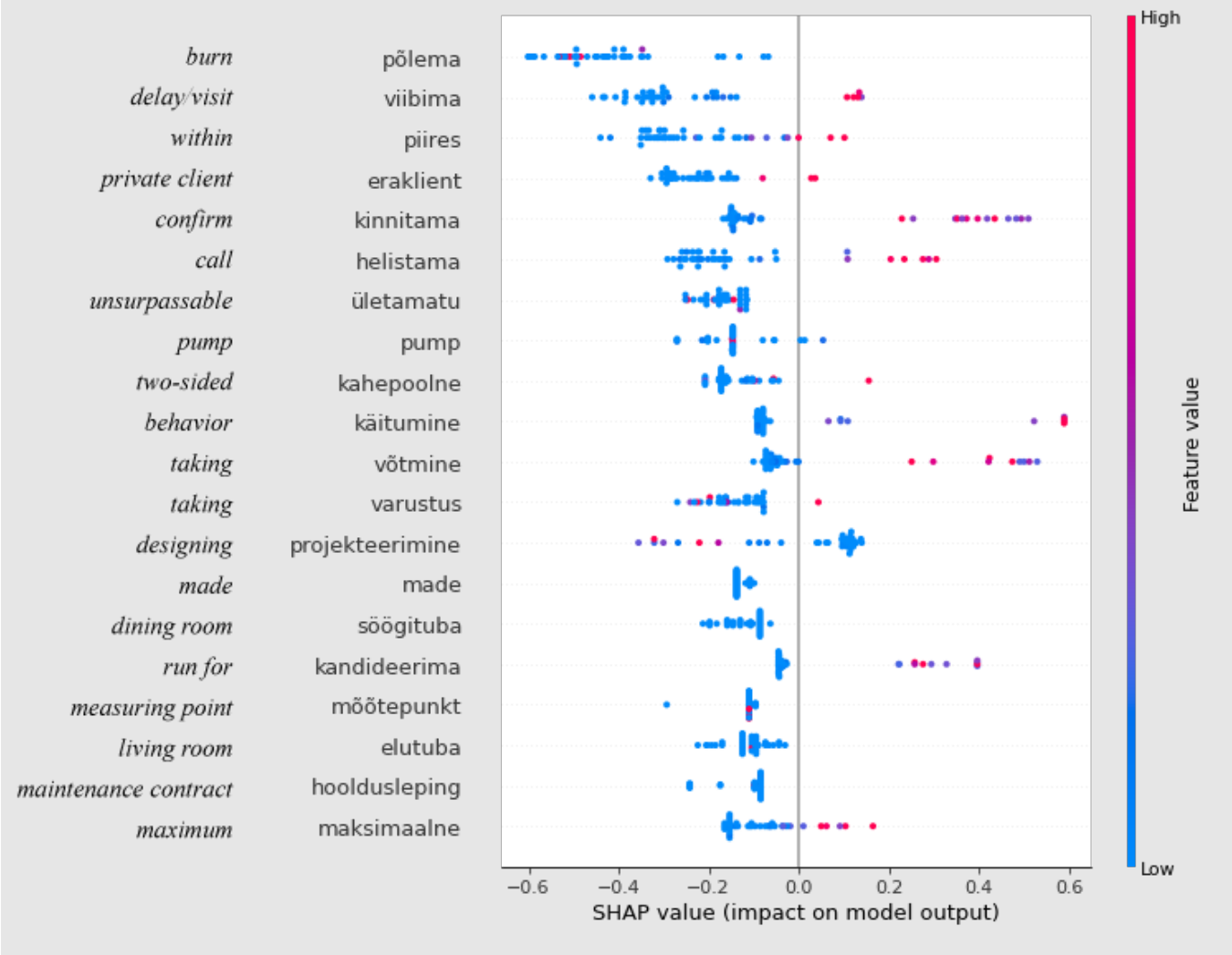
Source: Compiled by the authors.

Appendix 5. SHAP summary plot with overall innovation (1st model)



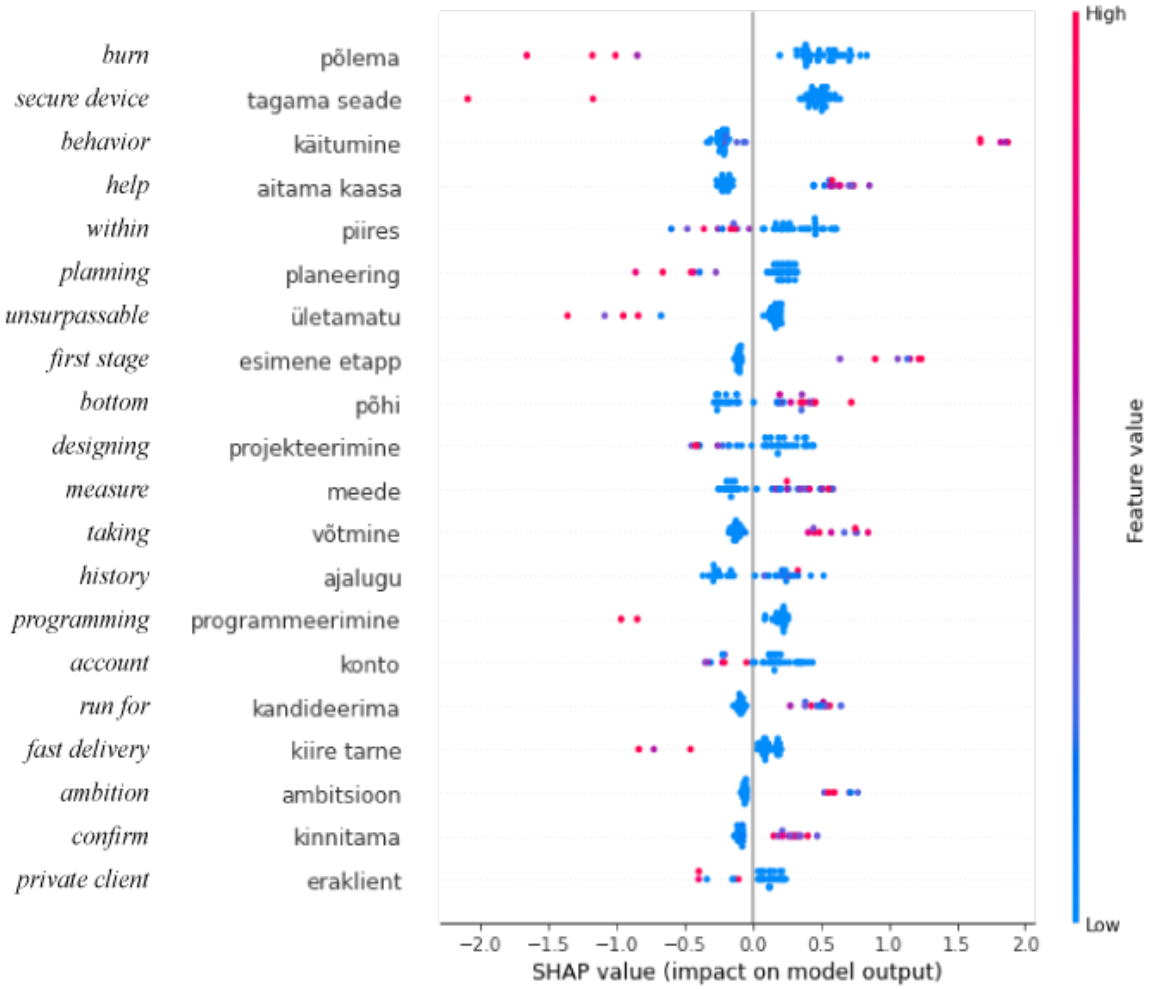
Note. (XGBoost, Tf-idf with min_word_count=5 and ngram= 2, unique word count 85,069)

Appendix 6. SHAP summary plot with overall innovation (2nd model)



Note. (XGBoost, default, unique word count 162,430)

Appendix 7. SHAP summary plot with overall innovation (3rd model)



Note. LightGBM, default, unique word count 85,069

KOKKUVÕTE

Ettevõtete innovaatilisuse ennustamine nende veebilehtede analüüsi põhjal

Käesolev artikkel uurib, milliseid põhilistest ettevõtete innovatsiooni (uuendustegevuse) tüüpidest (toote, protsessi-, turundus - ja organisatsiooniline innovatsioon) saab kõige paremini ennustada ettevõtete veebilehtede sisu põhjal. Kasutades analüüsis Eesti ettevõtete veebilehtede tekstiandmeid ja ettevõtete 2016 ja 2018 aasta innovatsiooniuuringu andmeid kombinatsioonis erinevate tehisintelligentsi (masinõppe) meetoditega osutus, et selline lähenemine on tõepoolest kasutatav ettevõttetaseme innovatsiooninäitajate ennustamiseks. Oluline uudne panus kirjandusse on saadud tulemus, et ettevõtete veebilehtede tekstid aitavad paremini ennustada ettevõtte poolt turundusinnovatsiooni kasutamist võrreldes kolme teise põhilise innovatsioonitüübiga. Kuna tehisintelligentsil põhinevad mudelid on oma olemuselt nn musta kasti tüüpi (mudeli sisemine toimimine pole selge), siis mudelitest aru saamiseks kasutasime SHAP lähenemist, tuvastamaks millised veebilehtedel kasutatud sõnad on mudelites kõige olulisemad ettevõtete klassifitseerimiseks innovaatiliseks või mitteinnovaatiliseks. Mudelid kinnitasid, et innovatsiooniuuringu andmetest saadud turundusinnovatsiooni indikaatorit aitasid tõepoolest ennustada turundusega seotud terminid (sõnad), samas muude innovatsiooni tüüpide korral nii selgeid ja kergesti tõlgendatavaid tulemusi ei õnnestunud saada. Meie analüüs näitab niisiis, et kuluefektiivsete, detailsete ja ajakohaste innovatsiooni indikaatorite konstrueerimise veebitekstide ja tehisintelligentsi meetodite kasutamisega on võimalik, kui selle lähenemise kasutatavus varieerub üle erinevate innovatsiooni tüüpide.